Letras & Letras

Linguística de *Corpus*: abordagem e metodologia em pesquisas linguísticas de base empírica

Organização:
Prof. Dr. Ariel Novodvorski
Profa. Dra. Maria José Bocorny Finatto

2º Semestre 2014
Volume 30, número 2

ISSN: 1981-5239

Expediente

EDUFU – Editora e Livraria da Universidade Federal de Uberlândia
Av. João Naves de Ávila, 2121 - Bloco 1S - Térreo - Campus Santa Mônica - CEP: 38.408-144 - Uberlândia - MG
Telefax: (34) 3239-4293
Email : vendas@edufu.ufu.br | www.edufu.ufu.br

Editoração: Prof. Guilherme Fromm

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

Letras & Letras, v. 30, n. 2, jul/dez 2014, Uberlândia,
Universidade Federal de Uberlândia, Instituto de Letras e Linguística, 1985-

Semestral.
Modo de acesso: http://www.seer.ufu.br/index.php/letraseletras
Editoração: Guilherme Fromm.
Organização: Ariel Novodvorski; Maria José Bocorny Finatto.

ISSN: 1981-5239

1. Língua. 2. Literatura-Crítica, 3. Linguística.
1. Universidade Federal de Uberlândia. Instituto de Letras e Linguística.

CDU: 801(05)

Participaram dessa edição como pareceristas *ad hoc*:

Adriana Silvino Pagano (UFMG)

Alena Ciulla (UFRGS)

Cláudio Márcio do Carmo (UFSJ)

Cleci Regina Bevilacqua (UFRGS)

Dalia Maria Guerreiro (Univ. de Évora)

Elizamari Rodrigues Becker (UFRGS)

Gabriel de Avila Othero (UFRGS)

Giacomo Figueiredo (UFOP)

Heberth Paulo Souza (IPTAN)

Heloísa Orsi Koch Delgado (PUC/RS)

Luciane Leipnitz (UFPB)

Maria Filomena Candeias Gonçalves (Univ. de Évora)

Maria Mercedes Riveiro Quintans Sebold (UFRJ)

Patrícia Bértoli (UERJ)

Paulo Pinheiro-Correa (UFF)

Pedro Henrique Lima Praxedes Filho (UECE)

Roberta Rego Rodrigues (UFPel)

Valéria Neto de Oliveira Monaretto (UFRGS)

Vera Lúcia Santiago Araújo (UECE)

Vera Maria Araujo Pigozzi de Araújo (UFRGS)

Sumário

Apresentação

Linguística de *Corpus* no Brasil: uma aventura mais do que adequada

Giovanni Parodi (2010, p. 167), nas reflexões finais de seu livro *Linguística de Corpus: da teoria à empiria*, trazia, quatro anos atrás, esta impressão: "parecem tempos em que ser linguista de *corpus* é uma aventura adequada". Considerando o percurso da Linguística de *Corpus* (doravante LC) no nosso país desde 2004, época do lançamento do livro de Tony Berber Sardinha, nosso primeiro manual brasileiro de LC, cabe aqui refletir um pouco sobre essa combinação de palavras dirigida à Linguística de *Corpus*.

Aventuras podem ser mais ou menos adequadas? O que é ser adequado em termos de uma aventura no terreno dos Estudos da Linguagem no nosso país? Ao apresentar este volume da revista Letras & Letras, queremos também tratar dessas questões.

Neste ano de 2014, comemoramos também os cinquenta anos do *Corpus Brown* (1964), um ponto de referência inevitável em qualquer retrospectiva sobre a LC em nível mundial. Ainda que de modo bastante mais restrito, em termos de repercussão local, podemos destacar que, exatamente há um mês, em novembro de 2014, conseguimos completar mais uma aventura, realizamos o *XII Encontro de Linguística de Corpus (ELC)* e a *VII Escola Brasileira de Linguística Computacional (EBRALC)*, na Universidade Federal de Uberlândia (UFU), interior de Minas Gerais, que também recebe a organização deste número da Letras & Letras especialmente dedicado à LC.

A partir dessas trajetórias e de outras, desenhadas por toda uma comunidade de pesquisadores em LC do Brasil, as palavras do professor Parodi renovam-se em significância e não poderiam definir melhor o momento atual. Se, por um lado, essas efemérides trazem à memória alguns marcos históricos importantes e nos fazem pensar em nossos próprios percursos até aqui, conduzem também a uma reconstrução do próprio processo, na constituição da área que hoje reconhecemos como LC.

Ainda cabe repetir que a LC se coloca como uma nova perspectiva para a Linguística (BERBER SARDINHA, 2004, p. 35), mas não como um novo tipo de Linguística. Mostra-se, para aqueles que se aproximam da LC, tanto como uma metodologia quanto como uma abordagem teórica diferenciada dos Estudos da Linguagem. De quem queira se aproximar da LC, apenas por se interessar por seu instrumental ou por seus procedimentos, nada será

cobrado em termos de uma filiação teórica – ou epistemológica – ainda que insistamos que LC também é um modo de compreender a língua, que temos nosso modo de defini-la como objeto de estudo: a língua é um sistema probabilístico de combinatórias, no qual uma unidade se define pelas associações que mantém com outras unidades.

Ao ocupar-se da exploração de grandes extensões de *corpora* textuais em formato digital, criteriosamente reunidos para representar um dado estado de uso de língua e "minerados" com apoio informatizado, com destaque para as explorações estatísticas de elementos lexicais e observação das frequências de combinatórias de palavras, vemos toda uma trajetória de estudos realizados no Brasil. Esses estudos, considerados em uma perspectiva muito ampla, podem ser bastante aproveitados em diferentes tipos de pesquisas e servem hoje, no mínimo, para caracterização de gêneros textuais[1]. Ao longo do seu percurso investigativo entre nós, quase todos os gêneros textuais escritos foram objeto de algum estudo em LC, do literário ao jornalístico, manuais técnicos e textos de culinária, entre vários outros, sem esquecermos dos *corpora* especialmente dedicados aos registros orais. A esse respeito, Mello (2012, p. 34) destaca que

> Apesar de os *corpora* escritos ainda dominarem a produção na área, a compilação de *corpora* orais e multimodais tem se ampliado rapidamente. Os *corpora* orais têm encontrado crescente aplicabilidade não apenas nos estudos canônicos da Linguística (Sociolinguística, Dialetologia, Lexicografia, Morfossintaxe etc.), mas também no desenvolvimento de tecnologias da fala, tais como o reconhecimento e síntese da fala.

O questionamento sobre a importância da coleta de dados dos usos linguísticos para as pesquisas, recorrente nos inícios da década dos sessenta, em pleno contexto histórico de dominância de uma linguística gerativista, contrasta radicalmente com o panorama atual. Se considerada a perspectiva de uma época em que se presumia que os dados já estariam na mente do linguista, o surgimento do *Corpus Brown* exatamente nesse contexto teve um valor pioneiro incalculável e um efeito dinamizador dos estudos baseados em *corpora*, já apontados por diversos autores. Essa mudança de paradigma se traduz num caminho percorrido entre a idealização e a sistematização da observação de evidências.

Atualmente, a expansão do uso dos termos *corpus* e *corpora,* além da menção a muitas das ferramentas e princípios caros à LC, alcança áreas que poderiam parecer, num primeiro

[1] As concepções de gênero textual presentes em trabalhos sob a ótica de Linguística de *Corpus* são em geral derivações das ideias de SWALES (1990) e HALLIDAY (1991).

momento, incompatíveis ou inimagináveis. Assim, a alusão às terminologias típicas de LC (como *types*, *tokens* e concordâncias) vem se tornando cada vez mais recorrente. Em eventos científicos, em publicações, em nomes de disciplinas, teses e dissertações, a recorrência com que aparecem referências ou vestígios da LC denotam já uma presença marcada no plano acadêmico e servem como um bom termômetro do estado da arte.

Já é amplamente conhecida a afirmação de que todo *corpus* sempre traz questões novas ou questões que não se imaginava encontrar, ainda que – de acordo com o próprio Fillmore (1992) – nenhum *corpus* nos dê resposta para tudo. De tal modo, tanto as observações como os experimentos e hipóteses formuladas no âmbito de toda investigação nos conduzem a uma revisão à luz das comprovações e dos resultados. Um médico espanhol, Ramón y Cajal, já assinalava em 1899 a importância do exame direto dos fatos da natureza e o uso de métodos, na tentativa de reduzir o máximo possível fatores subjetivos. Com isso, toda observação de dados para sua posterior descrição demandaria, necessariamente, fundamentações teóricas e princípios metodológicos; mas, acima de tudo, exigiria o traçado de caminhos de ida e de volta para a própria revisitação dos dados e ajustes dos pressupostos iniciais.

Dessa maneira, a sistematização de dados e de observações chega a ser crucial, talvez ainda mais importante do que a simples aplicação e contraste de teorias. A descoberta e identificação de padrões a partir da observação são, para Hanson (1958), os problemas fundamentais. Assim, toda teoria deriva do resultado de um trabalho consciente sobre os dados, uma vez que a tarefa das teorias seria colocar fenômenos em sistemas. Todas essas observações conduzem nosso olhar para a compreensão da relevância dada aos processos de observação, etapa indispensável nas pesquisas com *corpora* e nas diferentes fases de descoberta.

As concepções da LC, conforme vemos, com base nas ideias de Stubbs (1996, p. 46; 2001) e de Sinclair (1991, p. xviii), são as seguintes:

a) Um *corpus* não é mera ferramenta de análise. É, sim, um importante conceito teórico;

b) A linguagem se mostra diferente quando examinada extensivamente.

Assim, a "aventura adequada" citada no início deste texto envolve toda uma trajetória e todo um empreendimento coletivo de uma comunidade de pesquisadores. Nela se colocam um conceito teórico diferenciado e um empreendimento que convida o linguista interessado a

apreciar o seu objeto de estudo sob um ângulo também diferenciado. Aqui, frisamos, *diferenciado* não deve ser compreendido como algo melhor ou contrário aos diferentes convites à Linguística da nossa atualidade.

Esta apresentação busca contextualizar o momento particular em que surge este número temático da revista Letras & Letras, já definido como um marco histórico em que se constitui uma nova área de pesquisa, com abordagens e métodos próprios. Por outro lado, este texto introdutório também procura enxergar o *corpus* como essa espécie de "caminho de ida e volta", amarrado à importância da observação empírica dos fatos linguísticos. Nessa perspectiva, os *corpora* se tornam um território vasto o suficiente e propício para a descoberta de evidências.

Este número da revista Letras & Letras, dedicado à LC, está composto por 19 artigos e uma entrevista com o linguista chileno Giovanni Parodi, realizada no âmbito do XII ELC e da VII EBRALC. A diversidade de assuntos que compõe os artigos aqui presentes vai desde aspectos metodológicos a estudos de caso muito específicos, passando por diferentes correntes e afiliações teóricas, dentre as quais destacamos a Linguística Sistêmico-Funcional, a Linguística Histórica, a Documentação e a Linguística Cognitiva.

Giacomo Figueredo, da Universidade Federal de Ouro Preto (UFOP), no artigo intitulado "Uma metodologia de perfilação gramatical sistêmica baseada em *corpus*", apresenta uma metodologia de investigação de funções gramaticais, embasado em princípios da LC. O autor propõe uma metodologia que possibilita a identificação e descrição de padrões, na análise do modo como a gramática é empregada na organização do texto. Com base na teoria sistêmico-funcional e utilizando um *corpus* composto por dez minibiografias escritas em português brasileiro, o pesquisador estabelece um mapeamento das funções gramaticais, da distância topológica entre as funções e do movimento do emprego dessas funções no espaço gramatical.

Cristiane Namiuti-Temponi e Aline Silva Costa, da Universidade Estadual do Sudoeste da Bahia (UESB), também abordam aspectos metodológicos em seu texto "Reflexões sobre anotação sintática e ferramentas de busca - Uso da linguagem XML para anotação sintática no *corpus* digital DOViC". No artigo é discutido o uso da linguagem XML como alternativa ao formato *Penn TreeBank* para anotação sintática no *corpus* digital denominado *Documentos Oitocentistas de Vitória da Conquista* (DOViC). Dentre as justificativas apresentadas, as autoras destacam a utilização da linguagem XML na anotação

de edições e de informações morfológicas do *corpus*, por favorecer a criação de recursos padronizados reutilizáveis que facilitam a extração de dados dos *corpora*.

Da Universidade de São Paulo (USP), com o texto "Linguística de *Corpus* e ensino: a compilação de um *corpus* de especialidade para preparação e implementação de um curso preparatório rápido para exame de proficiência", a professora Stella Esther Ortweiler Tagnin e Danilo Suzuki Murakami apresentam o processo de compilação de um *corpus* de especialidade na área de Relações Exteriores. Os autores buscam definir tanto o conteúdo programático como a preparação de material didático para candidatos a um exame de proficiência em inglês, pensando em interessados no preenchimento de um cargo público no âmbito do governo federal.

Também no âmbito do ensino, mas voltado para a escrita em língua espanhola de aprendizes brasileiros, Benivaldo José de Araújo Júnior, da Escola Superior de Propaganda e Marketing (ESPM), apresenta o artigo "As construções com SE na produção escrita de brasileiros aprendizes de espanhol como língua estrangeira: um estudo baseado em *corpus*". Com base na Gramática Cognitiva, o autor trata especificamente das construções reflexivas, médias, impessoais e passivas e estabelece uma comparação com dados observados em dois *corpora* de falantes nativos: um de espanhol na variedade peninsular e outro de português brasileiro.

Ainda no limiar dos pressupostos teóricos da Linguística Cognitiva, Heberth Paulo Souza, do Instituto de Ensino Superior Presidente Tancredo de Almeida Neves (IPTAN), aborda a metáfora no escopo da cognição humana, no âmbito das representações mentais, com aplicações para a descrição da articulação textual. No artigo intitulado "Metáforas e domínios narrativos numa perspectiva da Linguística de *Corpus*", o autor recorre à Teoria dos Espaços Mentais e à Teoria da Mesclagem Conceitual, para descrever o papel exercido pela metáfora na articulação textual, num *corpus* de redações de vestibulandos, observando uma forma de organização de elementos típica dos processos de narração.

Para este número da revista *Letras & Letras*, os trabalhos desenvolvidos em três artigos tomam como *corpus* de estudo a legendagem. Desse modo, Vera Lúcia Santiago Araújo e Ítalo Alves Pinto de Assis, da Universidade Estadual do Ceará (UECE), no texto "A segmentação na legendagem para surdos e ensurdecidos (LSE) de Amor Eterno Amor: uma análise baseada em *corpus*", analisam problemas na divisão linguística em diálogos de uma produção audiovisual legendada. Segundo os autores, os problemas mais recorrentes de

segmentação foram detectados na ordem de sintagmas verbal e nominal, em legendas de três linhas e com alta velocidade.

Também Ana Katarinna Pessoa do Nascimento e Stella Tagnin, da USP, recorrem ao estudo das legendas, considerando a tradição francesa na LSE. No texto "Efeitos sonoros na legendagem francesa para surdos e ensurdecidos", as autoras analisam a tradução dos efeitos sonoros de três filmes franceses, com subsídios do programa *WordSmith Tools 5.0*. Por sua vez, Amanda Verdan Dib e Paulo Pinheiro-Correa, da Universidade Federal Fluminense (UFF), analisam "A função pragmática Tópico na legendagem brasileira de um filme argentino em um estudo de *corpus* paralelo", na legendagem brasileira do filme argentino *O Segredo dos seus olhos*, com a fundamentação teórica da Gramática Discursivo-Funcional (GDF) e funções do programa para alinhamento de *corpora* paralelos *YouAlign* (Terminotix Inc.). Os autores analisaram dois tipos de construções de tópicos: as topicalizações e os deslocamentos à esquerda.

Mudando o foco para o âmbito musical, Patrícia Bertoli, da Universidade Estadual do Rio de Janeiro (UERJ), apresenta o artigo "Convergência Lexical entre Letras de música e Inglês Geral: Um estudo baseado em *Corpus*". A partir de um *corpus* de aproximadamente 1 milhão de palavras, resultante de 5.962 letras de músicas diferentes, a autora contrastou listas de palavras individuais e de trigramas do *corpus* de estudo com outras listas extraídas de dois *corpora* de referência do inglês geral. Os resultados da pesquisa apontam semelhanças entre a linguagem usada nas letras das músicas analisadas e o inglês coloquial.

Por sua vez, o trabalho de Fernanda Beatriz Caricari de Morais, da Instituição Nacional de Educação de Surdos (INES), toma como objeto de estudo artigos científicos de Linguística – um estudo baseado na Linguística Sistêmico-Funcional, com o auxílio da Linguística de *Corpus*. Nele, a autora analisa os participantes agentes dos verbos do *dizer* mais utilizados em artigos de Linguística, coletados através da plataforma de periódicos *Scielo*. A LC possibilitou-lhe o tratamento computacional e dados quantitativos e contextos de ocorrência de determinadas palavras. Cláudio Márcio do Carmo, da Universidade Federal de São João del Rei (UFSJ), também na linha da Linguística Sistêmico-Funcional, combinada à Análise Crítica do Discurso, examina os discursos sobre sustentabilidade e nos traz a *ecologia linguística* uma área de trabalho em Linguística de *Corpus* que se ocupa da análise de padrões lexicais de que um determinado item faça parte, visando descrever sentidos a que esse item se associe.

Demonstrando a afinidade metodológica da LC com estudos de Linguística Histórica, o trabalho de Simone Floripi (UFU) trata de mapear o artigo do português clássico ao português europeu moderno. Sua investigação diacrônica, por meio da quantificação, mobiliza pressupostos teóricos de vertente gerativista e o Modelo de Princípios e Parâmetros. O trabalho de Maria Mercedes Riveiro Sebold e Anne Katheryne Estebe Maggessy, ambos da Universidade Federal do Rio de Janeiro (UFRJ), explora o contexto de produtividade das perífrases de gerúndio e de particípio no português do Brasil e na variedade do espanhol do México. O aspecto verbal é o ponto para contrastar essas duas línguas ricas morfologicamente.

Silvana Maria de Jesus, da Universidade Federal de Uberlândia (UFU), percorrendo outra direção, nos traz um trabalho sobre *corpus* combinado e estudos de Tradução baseados em *corpora*. No seu artigo, a autora aborda as relações de tradução de say/dizer em textos ficcionais no par linguístico inglês-português. O *corpus* combinado apresentado é composto de três romances originais em inglês e suas traduções para o português e três romances originais em português e suas traduções para o inglês, sendo parte do CORDIALL (*Corpus* Discursivo para Análises Linguísticas e Literárias) desenvolvido pelos pesquisadores do LETRA (Laboratório Experimental de Tradução) da Faculdade de Letras da UFMG.

Na perspectiva do ensino de língua estrangeira, especificamente no que se refere a padrões de escrita de aprendizes brasileiros de língua inglesa, coloca-se o trabalho de Barbara Malveira Orfano (UFSJ), Ana Larissa Adorno (UFMG) e Adriana Tenuta (UFMG), intitulado "Epistemic modality through the use of adverbs: a corpus-based study on learners' written discourse" que trata da modalidade epistêmica concretizada pelo uso de advérbios. Esse artigo discute a forma como os falantes nativos e os aprendizes brasileiros diferem em sua produção escrita e as possíveis implicações pedagógicas dessas diferenças.

Voltando ao tema de estudos em *corpora* de Tradução, o trabalho de Adriana Silvina Pagano, Arthur de Melo Sá e Kícila Ferreguetti, todos da UFMG, aborda a equivalência tradutória de partículas modais, trazendo um dos trabalhos do grupo de pesquisa "Modelagem sistêmico-funcional da tradução e da produção textual multilíngue", do Laboratório Experimental de Tradução da UFMG. Os autores compilaram um *corpus* paralelo bilíngue português brasileiro – inglês, formado por histórias seriadas da *Turma da Mônica* e suas respectivas traduções para o inglês. O objetivo foi identificar quais partículas eram utilizadas

com mais frequência no *corpus*, como elas foram traduzidas para o inglês, e seria possível verificar um padrão para as opções tradutórias.

Pedro Henrique Lima Praxedes Filho e Cristiene Ferreira da Silva, ambos da Universidade Estadual do Ceará (UECE), por sua vez, nos trazem um estudo de caso baseado em um *corpus* de roteiros de audiodescrições francesas de filmes, abordando uma das modalidades da Tradução Audiovisual – a que diz respeito à acessibilidade sociocultural de pessoas com deficiência visual. Ao abordarem o parâmetro da neutralidade em Audiodescrição, os autores buscaram investigar, com o auxílio da Linguística de *Corpus* (LC), a presença ou ausência de interpretação por parte do tradutor/audiodescritor, segundo os fundamentos da Teoria da Avaliatividade, no escopo da Linguística Sistêmico-Funcional (LSF).

Sabrina Matuda e Stella Tagnin, ambas da USP, nos trazem um artigo sobre a terminologia do futebol em um estudo direcionado pelo *corpus*. É estudada a terminologia do futebol em inglês e português, mobilizando-se a Linguística de *Corpus*, a Terminologia Textual e a concepção da tradução técnica culturalmente condicionada.

Jorge Viana Santos e Giovane Santos Brito, ambos da Universidade Estadual do Sudoeste da Bahia (UESB), brindam-nos com um trabalho bastante diferenciado, pois tratam da fotografia técnica de documentos para formação de *corpora* digitais eletrônicos. Nele, os autores apresentam as etapas do método fotográfico que vem sendo desenvolvido e utilizado, desde 2008, no Lapelinc (Laboratório de Pesquisa em Linguística de *Corpus* – UESB). Apresenta-se o processo de transposição de documentos manuscritos históricos do tipo jurídico para formação de *corpora* linguísticos.

Enfeixando o volume, temos o relato de uma entrevista feita por Ariel Novodvorski e Ana Fritz Herrera, ambos da UFU, com o linguista de *corpus* citado no início desta apresentação: o Prof. Dr. Giovanni Parodi, diretor de Pós-Graduação em Linguística da Pontificia Universidade Católica de Valpararíso no Chile. Nessa entrevista, em novembro de 2014, durante o *XII Encontro de Linguística de Corpus (ELC)* e a *VII Escola Brasileira de Linguística Computacional (EBRALC)*, Parodi nos apresenta a plataforma de *corpora* do espanhol *El Grial*, mas também mostra-nos suas reflexões sobre o passado e o futuro da LC em seu país e em uma perspectiva global.

Em todos esses artigos, em especial na entrevista do nosso convidado, ecoam as bases teórico-metodológicas da LC, as quais remontam aos trabalhos do britânico J. R. Firth

(escritos de 1960 a 1980). Firth, lidando com um enorme computador dos anos 50, já pesquisava em textos autênticos a distribuição de palavras sócio-culturalmente relevantes e acreditava que o significado de uma palavra se configurava no contexto de uso. Sua tão repetida citação **"You shall know a word by the company it keeps"** desde então chama a atenção para a imensa rede de relações sintagmáticas e paradigmáticas que envolve léxico e gramática, apontando para o fenômeno que ele chamava de colocação. Observava Firth, como bem temos estudado em LC, que as palavras que um falante escolhe utilizar em meio a um todo de opções à sua disposição exibem um padrão de associação regular. Isto é, as palavras privilegiam um tipo de combinação ou, melhor dito, elas "preferem" determinadas associações e, ainda, "rejeitam" outras.

Assim inspirada, ao longo de sua trajetória brasileira, para finalizar este texto, cabe uma analogia com a célebre citação de Firth e com as colocações, situando a LC no nosso cenário de estudos linguísticos. Conforme acreditamos, a LC associou-se a diferentes aventuras de investigação e praticamente nada rejeitou em termos de parcerias de trabalho – o diálogo tem sido uma marca constante, mesmo com aqueles que encaram a LC apenas como um *modus operandi* computacional e quantitativo. A despeito dessa impressão, claro deve ter ficado nesses, pelo menos, primeiros 10 anos de percurso no Brasil, que vamos muito além de "contar palavras" e que já prestamos uma contribuição muito importante para toda uma comunidade de pesquisa nacional e globalmente conectada. Assim, a aventura tem sido, sim, adequada e, mais do que isso, já muito bem-sucedida.

Desejamos a todos uma ótima leitura dos trabalhos deste volume e agradecemos ao nosso colega Prof. Dr. Guilherme Fromm, pelo suporte sempre atento a tudo que precisamos durante a organização deste número da Letras & Letras.

Ariel Novodvorski[*]
Maria José Bocorny Finatto[**]
Editores

[*] Doutor em Estudos Linguísticos pela Universidade Federal de Minas Gerais (UFMG). Professor Adjunto no Curso de Graduação em Letras e no Programa de Pós-Graduação em Estudos Linguísticos (PPGEL) do Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia (UFU).
[**] Doutora em Letras (UFRGS, 2001). Docente do Departamento de Linguística, Filologia e Teoria Literária da UFRGS, orientadora de mestrado e de doutorado junto ao PPG-Letras da UFRGS na linha de pesquisa Lexicografia e Terminologia: Relações textuais, Especialidade: Teorias do Léxico. Coordenadora do GELCORP-SUL, Grupo de estudos em Linguística de Corpus do Sul, certificado pela UFRGS, pesquisadora do grupo TERMISUL, bolsista PQ-CNPq.

Referências Bibliográficas

BERBER SARDINHA, A. **Lingüística de Corpus**. Barueri, SP: Manole, 2004.

FILLMORE, C. J. Corpus linguistics or computer corpus linguistics. In: SVARTVIK, J. (org). **Directions in corpus linguistics. Proceedings of Nobel Symposium 82**, *Stockholm.* Berlim/Nova York, De Gruyter, 1992.

HALLIDAY, M. A. K. Corpus studies and probabilistic grammar. In: AIJMER, K. ; ALTENBERG, B. (Org.). **English corpus linguistics**: studies in honour of Jan Svartvik. London: Longman, 1991.

HANSON, N. **Patrones de descubrimiento**: observación y explicación. Madrid: Alianza Editorial, 1977 [1958].

PARODI, G. **Lingüística de Corpus**: de la teoría a la empiria. Madrid: Iberoamericana / Vervuert, 2010.

RAMÓN Y CAJAL, S. **Reglas y consejos sobre investigación biológica**. Segunda edição do seu discurso lido perante a R.A.C.E.F. e N. Madrid: Imprenta de Fortanet, 1899.

RASO, T.; MELLO, H. (org.). **C-ORAL-BRASIL I**: Corpus de referência do português brasileiro falado informal. Belo Horizonte: UFMG, 2012.

SÁNCHEZ, A. (org.). **Cumbre – Corpus Lingüístico del español contemporáneo**: fundamentos, metodología y aplicaciones. Madrid: SGEL, 1996.

SINCLAIR, J. M. **Corpus, Concordance, Collocation**. Oxford: Oxford University Press, 1991.

STUBBS, M. **Text and corpus analysis**: Computer-assisted studies of language and culture. Oxford: Blackwell, 1996.

SWALES, J. M. **Genre analysis**: English in academic and research settings. Cambridge: Cambridge University Press, 1990.

Artigos

Uma metodologia de perfilação gramatical sistêmica baseada em *corpus*

Towards a corpus-based methodology for grammatical systemic profiling

Giacomo Figueredo [*]

RESUMO: Embasado nos pressupostos da Linguística de *corpus* (LC), este artigo apresenta uma metodologia de investigação de funções gramaticais, a qual inclui as etapas de compilação, extração e análise de dados da gramática. Denominada perfilação gramatical, a metodologia apresentada neste artigo possibilita a identificação de padrões e subsequente descrição, principalmente da forma como a gramática é empregada na organização do texto – a dinâmica textual. A perfilação gramatical lança mão da teoria sistêmico-funcional devido ao fato de essa teoria abranger as relações entre a gramática, sua expressão e a organização do texto. Para a perfilação gramatical deste artigo, foi utilizado como Exemplo de Pesquisa um *corpus* composto por dez minibiografias produzidas em português brasileiro. Os passos da perfilação incluem um mapeamento das funções gramaticais, o estabelecimento da distância topológica entre as funções e, por fim, o movimento do emprego das funções no espaço gramatical – constituindo assim a perfilação. Como resultado, este artigo apresenta a forma pela qual a metodologia da perfilação gramatical pode contribuir para o desenvolvimento das pesquisas de *corpus*, em particular quando o foco é a gramática. Especificamente, a perfilação gramatical se mostra efetiva na compilação e análise de *corpora* para além da relação item lexical/*token*, compreendendo igualmente categorias teóricas/funções gramaticais. Demonstra ainda como é possível compilar e extrair dados de *corpora* inclusive para a forma como são empregados na dinâmica textual, o

ABSTRACT: This paper develops a methodology to investigate grammar functions based on corpus Linguistics. The methodology is grammar profiling and includes compiling, retrieving and analyzing data from the grammar. Grammar profiling is shown to be a useful methodology for handling grammar patterning, especially when dealing with such pattering in the development of text. Grammar profiling is grounded on systemic functional theory for it needs a theory comprehensive enough to deal with grammar metaredundancy in relation to both text organization and expression. The grammar profiling of this paper is applied to investigate a corpus of ten mini-biographies in Brazilian Portuguese as a Research Example. The steps taken during profiling identify grammar functions, establish their topological distances and places them within the grammar space. Results show that grammar profiling is a useful methodology for corpus research, especially when the need is for a non-lexical corpus – one made up of functions – is investigated. It is also useful for compiling corpora and retrieving data from text development, capturing not only corpus synopsis, but also text dynamics.

[*] Doutor em Linguística Aplicada. Universidade Federal de Ouro Preto.

que, de outra forma, não seria possível.

PALAVRAS-CHAVE: Perfilação gramatical. Metodologia baseada em *corpus*. Dinâmica de sistemas. Espaço gramatical.

KEYWORDS: Grammatical profiling. corpus-based methodology. Systems dynamics. Grammar space.

1. Introdução

Este artigo focaliza as relações entre a metodologia advinda da Linguística de *Corpus* (LC) e a análise dos sistemas gramaticais. Motivado pela investigação dos fenômenos linguísticos, em particular de sua organização paradigmática (ou sistêmica) de forma objetiva e replicável, este artigo pauta-se pelos pressupostos da Linguística de *Corpus* (LC) para apresentar uma metodologia de investigação de funções gramaticais. Toma-se aqui como base a constante preocupação da LC de investigar a língua para além da intuição do falante/analista (VIANA, 2011), bem como o estabelecimento de metodologias confiáveis de compilação, extração e análise de dados que busquem garantir resultados uniformizados. A metodologia apresentada neste artigo, denominada perfilação gramatical, possibilita a compilação de *corpora* de funções gramaticais que, posteriormente, podem ser analisadas para a busca de padrões e subsequente descrição da língua.

A necessidade da perfilação gramatical para os estudos afiliados à LC surge do fato de que a maior parte das funções gramaticais, diferentemente dos itens lexicais, não é realizada de forma aparente pela expressão (i.e., típica, aberta), mas sim pela relação entre itens linguísticos (i.e., criptotípica, encoberta) (WHORF, 1987). Diferentemente dos itens lexicais, que são identificados e recuperados com facilidade – uma vez que são realizados por palavras, codificadas como *tokens* – as funções gramaticais, por comparação, raramente são realizadas por palavras, especialmente quando estas cumprem o papel de organizar o texto (CONRAD, 2010). Por tal motivo, essas últimas requerem a anotação manual do analista e dificultam a tarefa do computador na extração de dados.

Na perfilação gramatical, a anotação do *corpus* para funções gramaticais torna-se um passo metodológico, a qual abre a possibilidade para a compilação de *corpora* de funções gramaticais (cf. PAGANO *et al*., 2014). Dessa forma, a perfilação gramatical contribui para o reconhecimento de padrões que, de outra forma, não seriam reconhecidos, o que tornaria a análise assim suscetível à variação conforme o analista (BIBER *et al*., 2004). Igualmente, a perfilação gramatical auxilia a busca automática pelo computador, convertendo as funções gramaticais em "*tokens* gramaticais".

Como aporte teórico, a perfilação gramatical baseia-se na Linguística Sistêmico-Funcional (doravante LSF) (cf. HALLIDAY e MATTHIESSEN, 2004). Na qualidade de teoria abrangente, a LSF proporciona um arcabouço de descrição e análise do sistema linguístico como um todo, em seus diferentes estratos, de forma que a gramática pode ser examinada tanto em relação aos elementos que a expressam quanto ao texto que organiza. Além disso, a LSF privilegia a organização paradigmática da língua, dando ênfase à disposição sistêmica das funções gramaticais, as quais são o foco do presente artigo.

Para a perfilação gramatical deste artigo, como Exemplo de Pesquisa, será utilizado um *corpus* composto por dez minibiografias de matemáticos importantes, extraídas de livros didáticos de matemática publicados em português brasileiro (PB). A metodologia de análise inclui um mapeamento das funções gramaticais, o estabelecimento da distância topológica entre as funções e, por fim, o movimento do emprego das funções no espaço gramatical relativamente à dinâmica textual (LEMKE, 1991).

Com isso, este artigo espera demostrar como a metodologia da perfilação gramatical pode contribuir para o desenvolvimento das pesquisas de *corpus*, em particular quando o foco é a gramática. Especificamente, a perfilação gramatical permite a compilação e análise de *corpora* para além da relação item lexical/*token*, compreendendo igualmente categorias teóricas/funções gramaticais.

2. Análise linguística e análise de dados linguísticos pelo computador

As metodologias advindas da LC possibilitam a compilação de amostras representativas da língua – i.e., *corpora* – bem como a identificação de padrões dos itens linguísticos (cf. VIANA, 2011). As metodologias da LC se relacionam ao uso do computador devido ao fato de os *corpora* serem armazenados em formato eletrônico e os procedimentos de investigação envolverem o uso de *softwares* (TOGNINI-BONELLI, 2001).

Nesse processo, contudo, cabe distinguir a análise linguística, por um lado, da análise de dados linguísticos pelo computador, por outro. Isso se faz necessário para evitar o emprego de *tokens*, concordâncias e frequência de itens computacionais como explicação para os fenômenos linguísticos (PAGANO *et al*., 2014).

O computador consegue acessar o *corpus* apenas indiretamente, por meio de uma versão dos itens da linguagem natural codificada em uma forma que possa ser lida por *softwares*. Por exemplo, a palavra (unidade gramatical constituída por morfemas) é

"traduzida" como um *token* (conjunto de caracteres que o *software* consegue ler). Por conseguinte, os itens da língua natural são etiquetados; e as etiquetas, juntamente com os padrões que formam, é que são analisadas. Uma vez que o computador processa dados linguísticos indiretamente, *tokens*, linhas de concordância e frequências de ocorrência, por si só, não explicam fenômenos linguísticos (cf. KE, 2012). Nesse ponto, a LC proporciona os meios pelos quais as etiquetas do computador podem ser "retraduzidas" para a linguagem natural.

Do ponto de vista dos estudos de *corpora*, Tognini-Bonelli (2001) detalha a distinção entre as pesquisas baseadas em *corpus* (*corpus-based*) e aquelas direcionadas pelo *corpus* (*corpus-driven*). Os estudos baseados em *corpus* servem como forma de validar/refutar princípios da teoria, ao passo que os estudos direcionados pelo *corpus* generalizam os resultados, contribuindo para a construção gradual da teoria. A busca pelos padrões linguísticos que não são típicos ou abertos, na forma de itens isolados ou estruturas, faz com que a teoria reflita mais de perto os dados (cf. HALLIDAY e JAMES, 1993).

A trajetória desde uma abordagem baseada em *corpus* para outra direcionada pelo *corpus* – na qual os resultados extraídos pelo *software* são reinterpretados em termos linguísticos – abre a possibilidade da construção de uma "microteoria" do *corpus* que permite sua análise. Nesse processo, os dados extraídos pelo *software* – linhas de concordância, listas de palavras, colocações, etc. – são reinterpretadas em termos linguísticos. Os *types* e *tokens* revelam padrões da classe gramatical 'palavra'; a colocação revela padrões da classe semântica 'colocação lexical'; as linhas de concordância revelam padrões da gramática criptotípica; e assim por diante.

Do ponto de vista da LSF, Halliday (1992) faz a distinção entre o estrato da organização e o estrato da expressão no sistema linguístico. Os padrões extraídos pelo *software* são relativos ao estrato da expressão da linguagem natural – em particular na forma grafológica. Com isso, o estrato da expressão é, por sua vez, a manifestação da organização linguística, a qual pode ser modelada desde os padrões sistêmicos mais gerais – a gramática – até itens específicos – o léxico. A investigação do *corpus*, por meio da perfilação gramatical revela a organização linguística codificada nos dados extraídos pelo *software*.

Com isso, (i) o intercâmbio entre teoria e dados (estudos baseados em *corpus*/estudos direcionados pelo *corpus*) e (ii) a mudança do estrato da expressão (conjunto grafológico arbitrário) para o estrato da organização (a gramática e o léxico) se tornam os passos adotados

para se extraírem dados linguísticos do *corpus* por meio do computador. Quando o interesse da análise se concentra na gramática, inclui-se um novo passo, (iii) passar das ocorrências de itens específicos, o léxico, para os padrões mais gerais de organização.

Em certo sentido, a investigação é mais "fácil" (HALLIDAY, 1992, p. 64) quando (i) parte da abordagem baseada em *corpus* para a direcionada pelo *corpus*; (ii) concentra-se no estrato da expressão; e (iii) limita-se ao léxico. Por outro lado, quando a investigação tem como objeto a gramática, outros aspectos devem ser levados em conta.

Os itens lexicais constituem a região mais delicada (i.e., detalhada, específica) dos sistemas gramaticais, e por isso abarcam os itens mais distintos. Como tal, podem ser, em geral, considerados itens únicos. Um item lexical é, portanto, um item específico da gramática, em geral realizado por um elemento da ordem da palavra (ou, com menor frequência, da ordem do grupo). Consequentemente, sua expressão também é única. Esta é, via de regra, a definição do item lexical: a sobreposição de um significado em um item gramatical em uma expressão (HALLIDAY, 2002).

Na investigação do *corpus*, o computador consegue identificar com maior facilidade os itens lexicais justamente por causa de sua forma particular. Por exemplo, uma lista de palavras demanda uma quantidade menor de etapas na análise lexical (*tokenização*) do que uma lista de Sujeito Indeterminado, ou de Tipo de Processo na análise sintática/gramatical (*parsing*) (cf. CÂNDIDO-JR, 2013).

Os itens gramaticais são generalizáveis, uma vez que constituem recursos empregados em uma quantidade maior de situações na produção de significado. Com a exceção de funções gramaticais específicas – como, por exemplo, a Pluralidade e o Tempo Verbal – os itens gramaticais não possuem forma ou expressão únicas. Ao contrário, as funções gramaticais tendem a ser realizadas por um agrupamento de formas, tais como estruturas, partículas, prosódias ou pela recorrência de itens em conjunto (HALLIDAY, 2002).

Diante dessas considerações, uma abordagem mais efetiva deve "selecionar os sistemas fundamentais para uma teoria gramatical probabilística"*[1] (HALLIDAY, 1992, p. 64). De determinada maneira, a análise gramatical pode ser tomada como um processo guia para a LC, tanto na etapa de compilação do *corpus*, como também na etapa de análise dos dados. A noção de '*corpus*' então se amplia, abrangendo a expressão e os itens lexicais

* Todas as traduções apresentadas são de autoria própria.
[1] […] select systems that could be critical for a probabilistic grammatical theory.

(conjuntos de caracteres no computador), bem como as funções gramaticais realizadas pela expressão (as etiquetas dos conjuntos de caracteres).

Com o objetivo de aplicar a perfilação gramatical ao *corpus* de pesquisa, passa-se aos pressupostos teóricos desta, a partir da LSF.

3. Organização do sistema linguístico e funções gramaticais

A arquitetura sistêmica da língua (HALLIDAY, 2002) é uma proposta abrangente na qual a descrição de cada fenômeno contribui para a descrição do sistema como um todo. A justificativa de um modelo sistêmico é dada por HALLIDAY e MATTHIESSEN (2004, p. 20):

> São muitas as razões para a adoção de uma perspectiva sistêmica. Uma delas é que as línguas não são projetadas, mas evoluem. Ademais, os sistemas evolutivos não podem ser explicados simplesmente como uma soma de partes. Nosso pensamento tradicional sobre a linguagem, que é composicional, deve ser, se não substituído, pelo menos complementado por um pensamento 'sistêmico', por meio do qual buscamos compreender a natureza e a dinâmica do sistema semiótico como um todo.[2]

Nesses moldes, a LSF propõe que o sistema seja concebido a partir das dimensões necessárias para que o mesmo seja explicado, bem como os princípios pelos quais se organiza. Dentre as dimensões, destacam-se pela relevância para o presente trabalho: estratificação, metafunção e eixo.

3.1 Dimensões do sistema linguístico

Estratificação: na condição de sistema semiótico, a língua possui dois estratos – o conteúdo e a expressão. No entanto, a divisão do estrato do conteúdo em dois – um gramático-organizacional e outro semântico-discursivo – possibilitou ao sistema não apenas transmitir, mas também criar significado. Pelo fato de o conteúdo possuir um estrato responsável pela organização (i.e., a gramática, que realiza os itens semânticos e o texto), qualquer escolha gramatical cria, necessariamente, uma mudança na produção do significado

[2] There are many reasons for adopting this systemic perspective; one is that languages evolve – they are not designed, and evolved systems cannot be explained simply as the sum of parts. Our traditional compositional thinking about language needs to be, if not replaced by, at least complemented by a 'systems' thinking whereby we seek to understand the nature and the dynamic of a semiotic system as a whole.

e no desenvolvimento do texto. Contudo, ressalta-se que essa criação não pode acontecer independente do contexto (retórico).

Halliday (2002) afirma que o fato de o sistema ser estratificado permite que a gramática comporte ao mesmo tempo mais de um modelo de realidade, podendo, com isso, equilibrar os recursos de cada modelo dependendo da experiência a ser concebida, o que se faz em ocasiões diferentes de formas diferentes entre os estratos.

Com isso, é possível encontrar padrões na utilização dos recursos; isso porque os eventos do mundo que são concebidos pelos falantes como fenômenos similares tendem a ser categorizados de forma similar e, consequentemente, seu significado ser formalmente produzido na gramática de forma similar. Esses padrões se formam pela pressão exercida pelo contexto e a necessidade de se produzirem novos significados para novas situações.

Por conseguinte, além de criar os significados, a língua cria também as variáveis contextuais relevantes a partir das quais deve ser interpretada. As variáveis compreendem (i) o estabelecimento e manutenção das relações entre os falantes, (ii) a organização da realidade e do conhecimento e (iii) a constituição de cada significado individual como parte da unidade básica da língua – o texto (discurso em contexto). A variação contextual promove uma diversificação dos recursos gramaticais, de forma que os sistemas se agrupem em conjuntos de funções distintos, denominados metafunções.

Metafunção: é a dimensão da língua responsável por agrupar os sistemas segundo a sua codependência para a produção de determinado tipo de significado, que pode ser (i) estabelecer e manter as relações entre os falantes, (ii) organizar a realidade e o conhecimento, e (iii) variar a frequência de itens segundo a situação para assim criar o discurso em contexto.

A metafunção interpessoal é responsável pelo tipo de relação ou interação que o falante (ou escritor) estabelece com seu ouvinte (ou leitor). Assim, relações de poder entre emissor e receptor, bem como as de polidez, subserviência, exigência, pedido, entre outras, são codificadas pelos sistemas desta metafunção. A metafunção ideacional representa nossa percepção das coisas que existem no mundo. É responsável pela transitividade, ou o tipo de interação existente entre as coisas (participantes) e os eventos (processos). Assim, os processos, isto é, o conjunto de realização dos eventos, os participantes e as circunstâncias nas quais estes processos acontecem compreendem esta metafunção. A organização das informações na construção do texto é de responsabilidade da metafunção textual. Segundo Halliday (1976), o texto é uma unidade operacional da língua, que pode ser falado ou

expresso através da escrita, independente de seu tamanho. É na composição do texto que o falante escolherá quais partes de seu texto são mais importantes e merecem destaque, quais ideias mantêm relação com outras, e assim por diante.

Eixo paradigmático: segundo a dimensão do eixo paradigmático, a língua é examinada como o conjunto de opções sistêmicas. Por um lado, isto implica em dar proeminência à escolha como forma principal de se produzirem significados; por outro lado, o valor de cada opção só pode ser conhecido na oposição às outras opções (i.e., na relação de agnação). O princípio que ordena o sistema linguístico do ponto de vista do eixo paradigmático é a delicadeza, segundo o qual a escolha de uma opção se dá em níveis cada vez mais refinados desde a condição de entrada mais geral (indelicada) até a escolha, no final do sistema, por um item lexical, que seria a opção mais delicada.

Os sistemas são organizados em redes e cada item de uma classe constitui um termo a ser selecionado. Dentre os itens de uma classe, os sistemas determinam a propriedade que permite a seleção de um item possível, e não de outro, igualmente possível. Como forma de notação do sistema, utilizam-se as redes dos sistemas. A título de ilustração, apresenta-se a seguir a rede do sistema de QUANTIFICAÇÃO do PB (Figura 1).

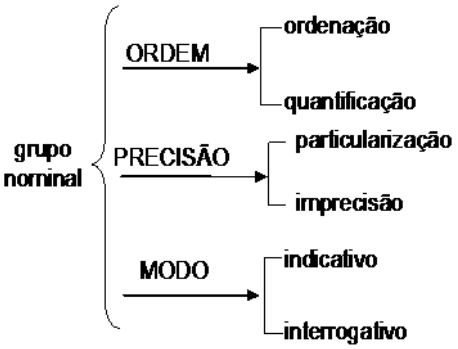


FIGURA 1 – O sistema de QUANTIFICAÇÃO em português brasileiro.

A unidade da condição de entrada para a QUANTIFICAÇÃO sistema é o grupo nominal, aberta para três subsistemas: ORDEM, PRECISÃO e MODO. Em cada subsistema, há duas opções e apenas uma delas deve ser selecionada. Por exemplo, se for selecionada uma opção da seguinte maneira: ORDEM → ordenação; PRECISÃO → particularização; MODO → indicativo, então a opção em PB é realizada pelos numerais ordinais, como em 'terceiro' ou 'décimo'.

Um passo importante da investigação gramatical, é relacionar as funções às outras opções que a ela são opostas, dispondo os dados da pesquisa formalmente. Logo, faz parte da metodologia desta pesquisa dispor formalmente os dados, apresentando-os como redes de sistemas.

Especificamente no caso deste trabalho, o foco serão os principais sistemas da oração, interpessoal, ideacional e textual (respectivamente, os sistemas de MODO, TRANSITIVIDADE e TEMA).

3.2 Dinâmica textual

A produção de significado é caracterizada como a atividade de seleções sucessivas de funções dos sistemas, de forma que, cumulativamente, gerem um texto. O acúmulo de significado no texto pode ser observado segundo dois pontos de vista (LEMKE, 1991, p. 25):

(i) Sinóptico: o sistema é visto como o conjunto geral de todas as funções, disposto de forma organizada pelo *valeur* e pela delicadeza. O sistema são todas as funções à disposição para selecionar dentre elas os inúmeros significados, conjuntos de significados e redes de relações diferentes. O termo 'sinóptico' implica em uma descrição geral de todas as possibilidades do sistema.

(ii) Dinâmico: o sistema é visto como o conjunto de relações entre as funções selecionadas durante a produção de um texto. O termo 'dinâmico' implica em observar, dentre todas as possibilidades, apenas aquelas que de fato foram selecionadas para produzir determinado texto.

Lemke (1991) caracteriza a relação entre os sistemas e as perspectivas afirmando que uma rede sistêmica é sempre sinóptica. Com isso, ela não consegue apreender o processo de produção de significado. Esse autor introduz a noção de sistema dinâmico, o qual descreve, para cada escolha no sistema, conjuntamente a opção "próxima escolha". Lemke (1991, p. 25-26) conclui:

> [...] é fundamental compreender "como" e "por que" se dá a variação. Dado que a descrição dos sistemas dinâmicos é, basicamente, uma questão de pesar a probabilidade de cada escolha, então cabe entender como as escolhas "até este ponto" condicionam as probabilidades para "próxima escolha".[3]

[3] […] it is very important for us to be able to say something about "how" and "why" it varies as it does. if we imagine the description of dynamic systems to be mainly a matter of the dynamic weightings of selection

A perspectiva da produção dinâmica altera o conceito de significado, que pode variar dependendo de quais outras funções se opõem em uma mesma rede, formando assim um fluxo de significados. Igualmente, a produção dinâmica cria uma história da rede. Na perspectiva dinâmica, a história da rede armazena o acúmulo de informação nos sistemas.

É justamente essa história da rede que a perfilação gramatical captura e, a partir daí, possibilita a compilação de *corpora* das histórias das redes. Esse processo tem início na dimensão do eixo e, subsequentemente, integra as outras dimensões segundo a relação entre as disposições tipológica e topológica da gramática.

3.3 Tipologia e Topologia

A tipologia pode ser definida como o conjunto das relações de oposição nos sistemas de acordo com as categorias que os diferenciam. Na descrição gramatical, a tipologia se associa à organização paradigmática, uma vez que captura o aspecto da escolha das funções, dividindo-as, assim, em tipos (i.e., classes).

A topologia, por sua vez, se define como a representação multiaxial que estabelece um espaço no qual as categorias são dispostas de forma gradual e em posição relativa (SOUZA, 2013). Na descrição gramatical, a topologia busca depreender as relações entre as funções que não são definidas pela oposição ou pela a escolha, mas que, de outra forma se associam por outras formas, tais como a coocorrência nos textos, a frequência, ou a proximidade semântica (MARTIN e MATTHIESSEN, 1991).

O modelo de organização sistêmica, construído a partir das relações paradigmáticas, dá ênfase ao aspecto tipológico do sistema. Contudo, o modelo tipológico não consegue apreender algumas relações entre as funções, por elas não se relacionarem como oposição em um determinado paradigma, mas por serem produzidas a partir de relações dinâmicas que variam com os textos. Com isso, é possível perceber que a organização tipológica se associa à visão sinóptica do sistema.

Dessa maneira, uma forma de complementar o modelo é tentar apreender tais relações. Se a necessidade se relaciona à visão dinâmica do sistema, faz sentido empregar um modelo topológico. É isto que a perfilação gramatical apresentada neste artigo busca fazer.

---

probabilities, then we wish to know how the selections “up to now” condition the probabilities for selections “now”.

4. Perfilação gramatical: uma metodologia baseada em *corpus*

Dado que o significado produzido pela gramática está no contraste das opções paradigmáticas, o trabalho realizado pela dimensão do eixo é alterar a organização dos sistemas para produzir variação e, assim, gerar textos para as diferentes situações (HALLIDAY, 1996). Nessas condições, faz-se necessária uma abordagem multidimensional, que integre estratificação, metafunção e eixo, segundo a modelagem das redes sistêmicas (MARTIN, 2013; QUIROZ, 2013). Além disso, é igualmente necessário explicar o funcionamento da dinâmica textual (LEMKE, 1991) – ou a seleção sucessiva de opções nos sistemas para a produção de um texto.

Tendo como base a LC e o aporte teórico da LSF, apresenta-se, a seguir, a forma como a metodologia para a perfilação gramatical se desenvolve. A perfilação parte da organização sistêmica e das redes para a construção de mapas topológicos e do espaço gramatical do qual serão extraídos os perfis gramaticais do *corpus*.

4.1 Mapas topológicos

O mapa topológico da gramática consiste em converter as oposições sistêmicas em espaços entre as funções (MARTIN e MATTHIESSEN, 1991). Como dito anteriormente, o mapa topológico abrange as seguintes etapas da perfilação gramatical: (a) o mapeamento das funções gramaticais, (b) o estabelecimento da distância topológica entre as funções e, por fim, (c) o movimento do emprego das funções no espaço gramatical.

Como ilustração do desenho dos mapas, será apresentado o sistema de MODO IMPERATIVO em PB, o qual possui quatro opções.

1) A primeira opção mais delicada do imperativo se refere ao Sujeito das orações imperativas. Este é sempre o ouvinte no Modo imperativo jussivo, mas pode também incluir o falante nas orações sugestivas. Por exemplo: "**Tomemos** como exemplo o segmento de frase".

2) Uma outra opção mais delicada é o recurso para aumentar a chance de se obter uma resposta esperada ao comando. Esta é realizada pela possibilidade de se expressar uma cópia do Predicador, igualmente com forma verbal imperativa, ao fim da proposta, funcionando como o Ecopredicador. Por exemplo: "**Conta** seu sonho pra ele, **conta**". Ressalta-se que a utilização do Ecopredicador se restringe às opções sugestivo ou jussivo: neutro no sistema de tipo de imperativo.

3a) Imperativo: singularizado: Ocorre quando o falante determina exatamente a quem será atribuído o comando. Nesse caso, a especificidade do comando é realizada pela posição na estrutura dos elementos do Negociador: Sujeito seguido pelo Predicador.

3b) Imperativo: negociado: este ocorre quando a responsabilidade por se obedecer a um comando é negociada entre os interlocutores. Nesse caso, o Sujeito segue o Predicador na estrutura Predicador^Sujeito. Por exemplo: "**Você abre** a porta." "Eu não, **abre você**".

Além do TIPO DE IMPERATIVO, o PB apresenta ainda um outro subsistema, denominado FORÇA. Em PB a FORÇA é realizada fonologicamente pelos movimentos tônicos 1, 1+ e 5. Uma oração imperativa pode codificar um comando (demanda de bens-e-serviços), que seria a opção não-marcada, opção esta realizada pelo Movimento Tônico 1(queda). Pode ainda codificar uma opção marcada: (i) codificando um pedido, realizada pelo Movimento Tônico 5 (subida; queda); ou (ii) codificando uma ordem, realizada pelo Movimento Tônico 1+ (queda forte).

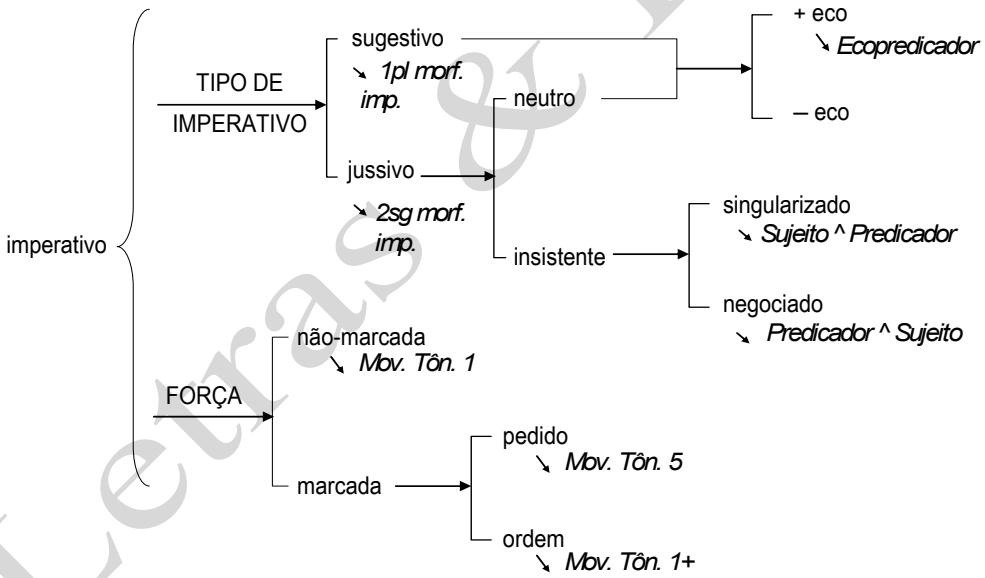


FIGURA 2 – A rede do sistema de MODO IMPERATIVO em PB.

A partir da rede, dividem-se as escolhas conforme seu nível de delicadeza, e cada uma delas é numerada. Cabe ressaltar que os números representam valores categóricos e, com isso, irão revelar a posição relativa das funções no mapa topológico (Figura 3).

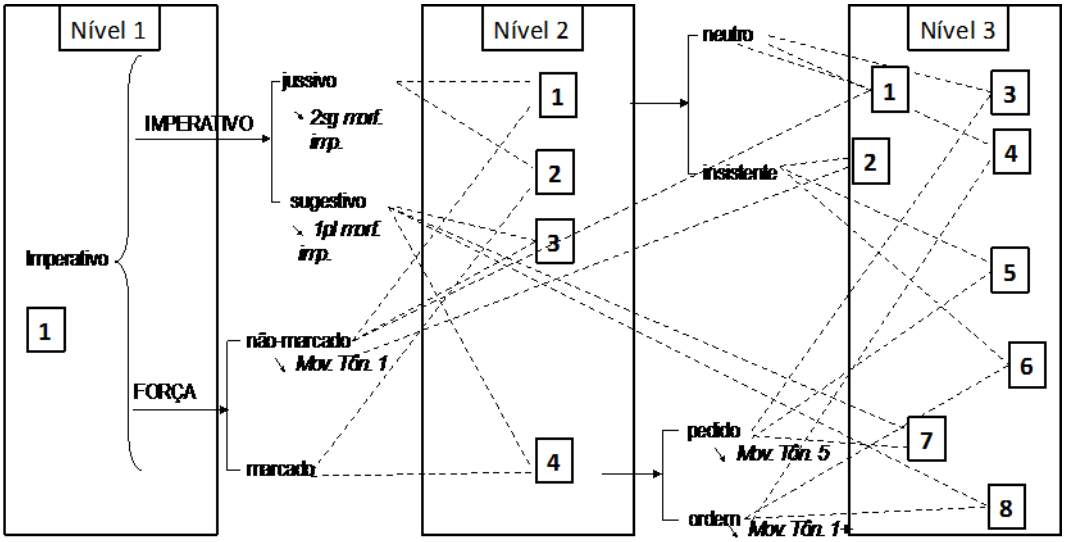


FIGURA 3 – Categorização por nível de delicadeza do IMPERATIVO.

A partir daí, é possível conferir a cada cosseleção uma etiqueta que representa o seu lugar no sistema, conforme um conjunto de escolhas. O Quadro 1 mostra as etiquetas de cada função, bem como a escolha feita para cada nível de delicadeza. Por exemplo, a função 1124 mostra que, no primeiro nível de delicadeza (N1), seleciona-se 1 para a função Imperativo. No segundo nível de delicaza (N2), seleciona-se 1 para a função Jussivo & Não-Marcado. Já no terceiro nível de delicaza (N3), seleciona-se 2 para a função Insistente. Por fim, para o nível mais delicado (N4), seleciona-se 4, a função Negociado. Por exemplo, "Assiste o filme você".

QUADRO 1 – Etiquetamento das cosseleções do MODO IMPERATIVO.

N1	função	N2	Função	N3	função	N4	#	exemplo
1	jussivo& não-marcado	1	neutro	1	não-eco	1	1111	Assiste o filme.
			insistente	2	eco	2	1112	Assiste o filme, assiste.
							1123	Você assiste o filme.
							1124	Assiste o filme você.
			ordem&neutro	3	singularizado	3	1231	Assiste o filme. (+1)
	jussivo& marcado	2	pedido&neutro	4	negociado	4	1232	Assiste o filme, assiste. (+1)
			ordem&insistente	5			1241	Assiste o filme. (+5)
			pedido&insistente	6			1242	Assiste o filme, assiste. (+5)
							1253	Vc assite o filme (+1)
							1254	Assiste o filme vc (+1)
			sugestivo&ordem	7			1263	Vc assiste o filme (+5)
	sugestivo& não-marcado	3	sugestivo&pedido	8			1264	Assiste o filme vc (+5)
							1311	Vamos ver o filme.
							1312	Vamos ver o filme, vamos.
							1471	Vamos ver o filme (+1)
	sugestivo& marcado	4					1472	Vamos ver o filme, vamos
							1481	(+1)
							1482	Vamos ver o filme (+5)
								Vamos ver o filme, vamos (+5)

O etiquetamento das funções é um recurso analítico importante por conseguir capturar as cosseleções entre os subsistemas, bem como o seu desenvolvimento na delicadeza. Ele permite desenhar um mapa topológico com as distâncias relativas entre as funções.

Para a perfilação, os mapas topológicos servem como uma "descrição dinâmica" da gramática, seguindo Lemke (1991), uma vez que permitem a visualização do panorama gramatical para um determinado fenômeno. A Figura 4 mostra o mapa topológico do MODO em PB. No caso da Figura 4, tem-se um panorama da gramática interpessoal da oração em PB. O mapa compreende o sistema de MODO e seus subsistemas, MODO INDICATIVO e MODO IMPERATIVO, incluindo todos os seus subníveis de delicadeza.

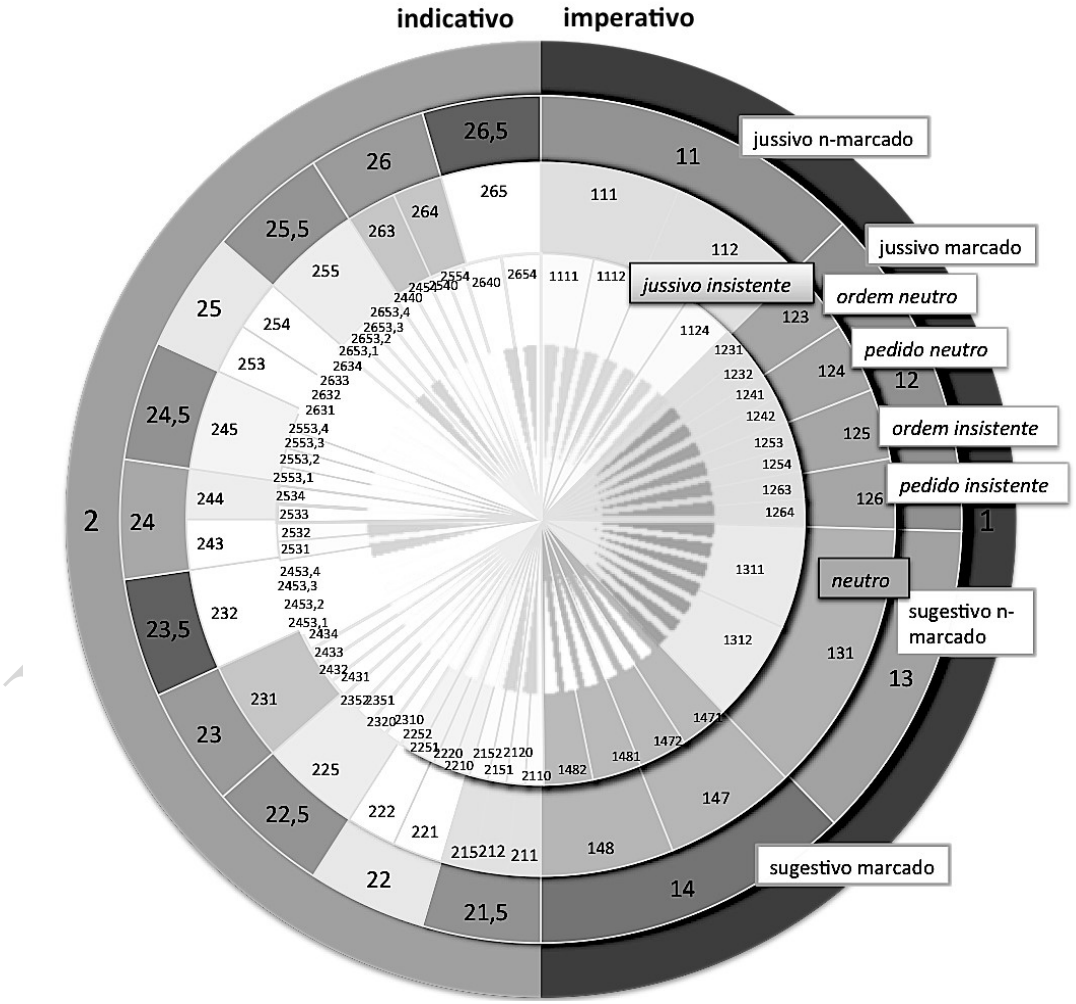


FIGURA 4 – Mapa topológico do MODO em PB.

4.2 Corpora gramaticais

Os mapas topológicos permitem que cada categoria – i.e., funções, ou cosseleções de funções, para qualquer nível de delicadeza – seja anotada por um código simples de apenas alguns dígitos (no caso do MODO, quatro ou cinco). Torna-se possível converter a anotação gramatical em um conjunto de *tokens*, sendo que cada etiqueta do código se torna um *token*. Por conseguinte, os mapas topológicos possibilitam a compilação de *corpora* gramaticais que, por sua vez, podem ser investigados pelas etiquetas, as quais representam "*tokens* gramaticais".

A Figura 5 traz um exemplo de um *corpus* de *tokens* gramaticais elaborado a partir do *corpus* da pesquisa. O número entre parênteses indica a sequência das orações e cada ponto final indica o início de uma nova oração. Cada código de seis dígitos indica as funções gramaticais da oração. Na primeira célula, as funções se apresentam na forma como foram empregadas no texto; na segunda célula, apresentam-se em um exemplo de busca por linhas de concordância.

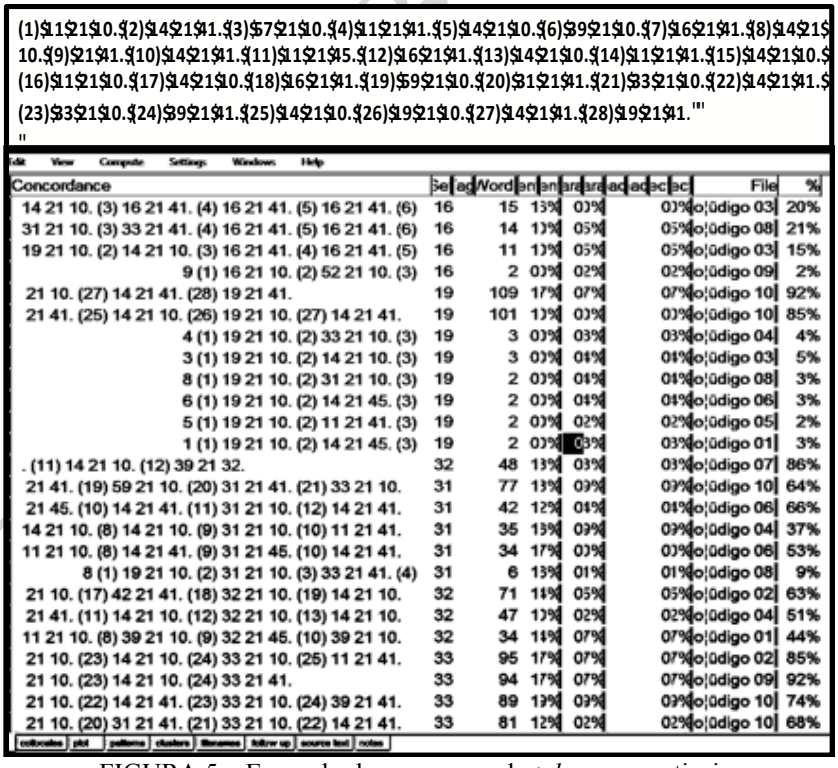


FIGURA 5 – Exemplo de um *corpus* de *tokens* gramaticais.

Quando o movimento do emprego das funções gramaticais é acrescentado à perfilação, faz-se necessário acompanhar a forma pela qual a dinâmica textual se desenvolve. Assim, não

basta apenas entender as distâncias topológicas entre as funções em um texto, mas, igualmente, cabe levar em conta a variável 'tempo' – ou a sequência em que ocorrem as mudanças, dinamicamente. Isso é feito estabelecendo a distância entre as funções gramaticais das orações, na ordem em que aparecem no texto.

Tomando como exemplo o TEXTO 10 do *corpus* da presente pesquisa, as quatro primeiras orações são classificadas conforme suas funções gramaticais para os sistemas de TRANSITIVIDADE, MODO e TEMA e recebem os seguintes códigos de acordo com sua localização nos mapas topológicos:

QUADRO 2 – Análise na perfilação gramatical

oração		FUNÇÃO	CÓDIGO
1	Évariste Galois nasceu nas proximidades de Paris, na aldeia de Bourg la-Reine, onde seu pai era prefeito.	TRANS.: Material: Intrasitivo (11) MODO: Declarativo (21) TEMA: Default (10)	112110
2	Aos 12 anos mostrava pouco interesse por Latim, Grego e Álgebra mas a Geometria de Legendre o fascinava.	TRANS.: Material: Transitivo (14) MODO: Declarativo (21) TEMA: Perspectiva (41)	142141
3	Aos 16 anos, julgando-se em condições, procurou entrar na Escola Politécnica mas foi recusado por falta de preparo e isto marcou seu primeiro fracasso.	TRANS.: Mental: Cognitivo (57) MODO: Declarativo (21) TEMA: Perspectiva (41)	572141
4	Aos 17 anos escreveu um artigo onde expunha suas descobertas fundamentais entregando-o a Cauchy para que o apresentasse na Academia.	TRANS.: Material: Transitivo (14) MODO: Declarativo (21) TEMA: Perspectiva (41)	142141

Relativamente a essas quatro orações, pode-se perceber que 1, 2 e 4 são próximas entre si e distantes de 3 na TRANSITIVIDADE; as quatro estão no mesmo ponto com relação ao MODO; e 1está distante de 2, 3 e 4 com relação ao TEMA. Quando esse tipo de medida é feito para o texto completo, é possível desenhar um gráfico que represente essas distâncias no espaço gramatical (Figura 6).

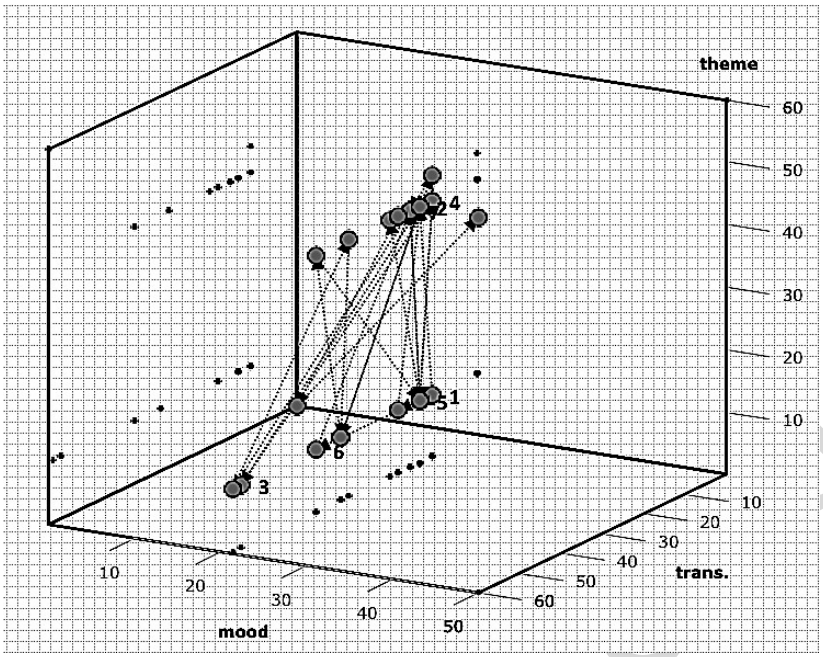


FIGURA 6 – Dinâmica no espaço topológico da gramática para o TEXTO 10

Como mostra a Figura 6, cada ponto representa uma oração e, com isso, um conjunto de coordenadas das funções oracionais relativas à TRANSITIVIDADE (eixo x), ao MODO (eixo y) e ao TEMA (eixo z). Esse tipo de gráfico que representa o espaço topológico da gramática consegue contribuir para a análise do texto por permitir que ela seja visualizada, por completo, de uma só vez. Além disso, consegue também estabelecer as distâncias, do ponto de vista gramatical, entre as orações e suas posições relativas. Por fim, as linhas que ligam as orações convertem-se, na verdade, em vetores. Consequentemente, depreendem movimento em função da variável tempo e, assim, capturam a dinâmica do desenvolvimento do texto por completo.

A coleção de espaços gramaticais de diversos textos também funciona como a compilação de *corpora*. Neste caso específico, não são *corpora* de tokens lexicais, ou gramaticais, mas sim de "tokens textuais".

Somado à compilação de *corpora* de "tokens textuais" está o movimento entre as orações e, por conseguinte, entre suas funções, extraindo dados relativos apenas à dinâmica do desenvolvimento do texto. A partir do gráfico, ainda tomando o TEXTO 10 como exemplo, cada vetor permite calcular a mudança gramatical entre as orações ao longo do tempo. Com isso, elabora-se uma lista da sequência de equações vetoriais – em outras palavras, compila-se um *corpus* de "tokens discursivos" (Quadro 3).

QUADRO 3 – *corpus* da dinâmica textual: equações vetoriais do desenvolvimento do texto.

Oração	trans.(x)	modo (y)	tema (z)	amostra do *corpus* de equações vetoriais	linhas de concordância do *corpus* de equações vetoriais
01	11	21	10		r = <11, **21, 10**> + t<3, **0, 31**>
				r = <11, 21, 10> + t<3, 0, 31>	r = <57, **21, 10**> + t<-46, **0, 31**>
02	14	21	41		r = <39, **21, 10**> + t<23, **0, 31**>
				r = <14, 21, 41> + t<43, 0, -31>	r = <14, **21, 10**> + t<7, **0, 31**>
03	57	21	10		
				r = <57, 21, 10> + t<-46, 0, 31>	
04	11	21	41		
				r = <11, 21, 41> + t<3, 0, -31>	
05	14	21	10		
				r = <14, 21, 10> + t<15, 0, 0>	
06	39	21	10		
				r = <39, 21, 10> + t<23, 0, 31>	
07	16	21	41		
				r = <16, 21, 41> + t<-2, 0, -31>	
08	14	21	10		
				r = <14, 21, 10> + t<7, 0, 31>	
09	21	21	41		
				r = <21, 21, 41> + t<-7, 0, 0>	
10	14	21	41		

O conjunto das análises e dos dados gerados a partir delas nessa seção compõe a perfilação gramatical. A próxima seção apresenta um Exemplo de Pesquisa da perfilação em um *corpus* de minibiografias.

4.3 Perfilação gramatical das minibiografias – Exemplo de Pesquisa

Esta seção tem por objetivo apresentar um exemplo de Pesquisa da perfilação. Especificamente, esta traz o perfil gramatical de um *corpus* de minibiografias em PB e, devido ao escopo reduzido, irá se concentrar apenas na função de maior número absoluto de ocorrências.

Para o desenho do perfil gramatical, fazem-se necessárias as seguintes etapas: (a) compilação do *corpus*; (b) análise sistêmica do estrato gramatical; (c) distribuição topológica das funções gramaticais; (d) constituição do *corpus* gramatical; (e) busca de padrões no espaço gramatical e no desenvolvimento do texto (LEMKE, 1991).

(a) Compilação do corpus: o *corpus* da pesquisa foi compilado a partir da seleção de dez minibiografias de matemáticos importantes, apresentadas no início dos capítulos de livros didáticos de matemática. Conforme a tipologia da língua no contexto de cultura (HALLIDAY e MATTHIESSEN, 2004), os textos do *corpus* são classificados como: Relatar/Escrito/Monólogo. Especificamente, foram retirados dos livros da coleção *Fundamentos de Matemática Elementar*, de Gelson Iezzi (IEZZI *et al*., 1977/78). A descrição do *corpus* pode ser vista no Quadro 4.

QUADRO 4 – O *corpus* da pesquisa.

TEXTO	NOME	TIPOLOGIA	VOL.	TOKENS
GAUSS 01	*De Plebeu a Príncipe*	Relatar/Escrito/Monólogo	1	391
ABEL 02	*Jovem Luta para Ser Ouvido*	Relatar/Escrito/Monólogo	1	335
FERMAT 03	*As Margens dos Livros Falam*	Relatar/Escrito/Monólogo		320
EULER 04	*Cego Enxerga Longe*	Relatar/Escrito/Monólogo	2	454
LEIBNIZ 05	*Autodidata Cria a Análise*	Relatar/Escrito/Monólogo	2	403
FOURIER 06	*Condução do Calor: Nova Teoria*	Relatar/Escrito/Monólogo	3	227
LAPLACE 07	*Napoleão Demite Ministro do Interior*	Relatar/Escrito/Monólogo	4	335
V.NEUMANN 08	*Computadores – As Máquinas com Memória*	Relatar/Escrito/Monólogo	5	279
NEWTON 09	*Nem só de Física Vive um Gênio*	Relatar/Escrito/Monólogo	5	477
GALOIS 10	*Intelectual Morre em Duelo*	Relatar/Escrito/Monólogo	6	366
TOTAL				3.587

(b) Análise sistêmica do estrato gramatical: A análise gramatical do presente estudo focalizou a oração como unidade de investigação, especificando seus três principais sistemas, a partir da diversificação metafuncional. Assim, analisaram-se as funções ideacionais do sistema de TRANSITIVIDADE; interpessoais do sistema de MODO; e textuais do sistema de TEMA. Para esta análise, foram contemplados os dois primeiros níveis de delicadeza. Por exemplo:

QUADRO 5 – Exemplo de análise sistêmica.

	Évariste Galois	nasceu	nas proximidades de Paris, na aldeia de Bourg la-Reine
TRANS.	*Ator*	*Material: Intransitivo e Criativo*	*Circ. Localização*
MODO	*Sujeito*	*Finito*	*Adj. Circunstancial*
	Modo: Indicativo: Declarativo		*Resto*
TEMA	*Tema: Default*	*Rema*	

(c) Distribuição topológica das funções gramaticais: De posse de todas as funções dos sistemas principais da oração, estas foram localizadas nos mapas topológicos. Com isso, foi

possível obter do *corpus* sua distribuição topológica. Como exemplo, apresentam-se os valores nos dois primeiros níveis de delicadeza para as funções do texto GAUSS 01.

QUADRO 6 – Distribuição topológica do texto GAUSS 01

oração	Trans. (X)		modo (y)		tema (z)	
GAUSS 01	Material: Intransitivo e Criativo	19	Indicativo: Declarativo Temporal	21	Default	10
GAUSS 02	Material: Transitivo e Transformativo	14	Indicativo: Declarativo Temporal	21	Proeminente: Intensivo	45
GAUSS 03	Material: Transitivo e Transformativo	14	Indicativo: Declarativo Temporal	21	Proeminente: Perspectiva	41
GAUSS 04	Mental: Fenômeno e Cognitivo	59	Indicativo: Declarativo Temporal	21	Default	10
GAUSS 05	Material: Transitivo e Transformativo	14	Indicativo: Declarativo Temporal	21	Proeminente: Perspectiva	41
GAUSS 06	Material: Intransitivo e Transformativo	11	Indicativo: Declarativo Temporal	21	Default	10
GAUSS 07	Material: Intransitivo e Transformativo	11	Indicativo: Declarativo Temporal	21	Default	10
GAUSS 08	Relacional: Identificativo e Intensivo	39	Indicativo: Declarativo Temporal	21	Default	10
GAUSS 09	Relacional: Atributivo e Possessivo	32	Indicativo: Declarativo Temporal	21	Proeminente: Intensivo	45
GAUSS 10	Relacional: Identificativo e Intensivo	39	Indicativo: Declarativo Temporal	21	Default	10
GAUSS 11	Material: Transitivo e Transformativo	14	Indicativo: Declarativo Temporal	21	Proeminente: Intensivo	45
GAUSS 12	Verbal: Semiose e Não-Recepção	41	Indicativo: Declarativo Temporal	21	Default	10
GAUSS 13	Material: Transitivo e Transformativo	14	Indicativo: Declarativo Temporal	21	Proeminente: Perspectiva	41
GAUSS 14	Material: Transitivo e Transformativo	14	Indicativo: Declarativo Temporal	21	Default	10
GAUSS 15	Material: Intransitivo e Transformativo	11	Indicativo: Declarativo Temporal	21	Default	10
GAUSS 16	Material: Transitivo e Criativo	16	Indicativo: Declarativo Temporal	21	Proeminente: Perspectiva	41
GAUSS 17	Material: Transitivo e Criativo	16	Indicativo: Declarativo Temporal	21	Proeminente: Perspectiva	41
GAUSS 18	Relacional: Identificativo e Intensivo	39	Indicativo: Declarativo Temporal	21	Default	10

(d) Constituição do corpus gramatical: A partir dos mapas topológicos, são realizadas as distribuições tais como aquela apresentada no Quadro 6. A partir desse trabalho feito para todos os textos do *corpus*, apresenta-se a seguir o "*corpus* gramatical" do presente estudo (Quadro 7).

QUADRO 7 – *corpus* gramatical da pesquisa

TEXTO	Posições topológicas
GAUSS 01	(1) 19 21 10. (2) 14 21 45. (3) 14 21 41. (4) 59 21 10. (5) 14 21 41. (6) 11 21 10. (7) 11 21 10. (8) 39 21 10. (9) 32 21 45. (10) 39 21 10. (11) 14 21 45. (12) 41 21 10. (13) 14 21 41. (14) 14 21 10. (15) 11 21 10. (16) 16 21 41. (17) 16 21 41. (18) 39 21 10.
ABEL 02	(1) 33 21 10. (2) 41 21 41. (3) 14 21 10. (4) 14 21 41. (5) 14 21 41. (6) 33 21 10. (7) 11 21 10. (8) 14 21 41. (9) 33 21 10. (10) 14 21 10. (11) 33 21 10. (12) 59 21 41. (13) 33 21 10. (14) 11 21 10. (15) 14 21 41. (16) 11 21 10. (17) 42 21 41. (18) 32 21 10. (19) 14 21 10. (20) 58 21 10. (21) 58 21 10. (22) 14 21 10. (23) 14 21 10. (24) 33 21 10. (25) 11 21 41. (26) 21 41. (27) 14 21 41.
FERMAT 03	(1) 19 21 10. (2) 14 21 10. (3) 16 21 41. (4) 16 21 41. (5) 16 21 41. (6) 33 21 41. (7) 14 21 41. (8) 16 21 45. (9) 33 21 10. (10) 39 21 10. (11) 16 21 45. (12) 39 21 41. (13) 33 21 10. (14) 16 21 41. (15) 14 21 41. (16) 33 21 10. (17) 59 21 10. (18) 16 21 41.
EULER 04	(1) 19 21 10. (2) 33 21 10. (3) 11 21 10. (4) 11 21 45. (5) 14 21 41. (6) 33 21 10. (7) 14 21 10. (8) 14 21 10. (9) 31 21 10. (10) 11 21 41. (11) 14 21 10. (12) 32 21 10. (13) 14 21 10. (14) 14 21 10. (15) 39 21 10. (16) 16 21 41. (17) 16 21 10. (18) 11 21 45. (19) 14 21 10. (20) 11 21 10. (21) 14 21 10. (22) 49 21 45.
LEIBNIZ 05	(1) 19 21 10. (2) 11 21 41. (3) 33 21 41. (4) 11 21 41. (5) 14 21 10. (6) 11 21 10. (7) 39 21 41. (8) 16 21 10. (9) 16 21 10. (10) 39 21 10. (11) 39 21 41. (12) 16 21 10. (13) 14 21 45. (14) 33 21 10. (15) 39 21 10. (16) 14 21 10. (17) 16 21 10. (18) 14 21 41. (19) 16 21 10. (20) 39 21 10. (21) 14 21 10. (22) 14 21 41. (23) 14 21 45. (24) 59 21 45.
FOURIER 06	(1) 19 21 10. (2) 14 21 45. (3) 14 21 41. (4) 14 21 45. (5) 42 21 45. (6) 21 21 41. (7) 11 21 10. (8) 14 21 41. (9) 31 21 45. (10) 14 21 41. (11) 31 21 10. (12) 14 21 41. (13) 14 21 41. (14) 21 21 41.
LAPLACE 07	(1) 11 21 10. (2) 11 21 45. (3) 14 21 41. (4) 14 21 41. (5) 33 21 10. (6) 39 21 10. (7) 41 21 41. (8) 16 21 41. (9) 16 21 41. (10) 39 21 10. (11) 14 21 10. (12) 39 21 32.
V. NEUM. 08	(1) 19 21 10. (2) 31 21 10. (3) 33 21 41. (4) 16 21 41. (5) 16 21 41. (6) 16 21 10. (7) 16 21 41. (8) 21 21 41. (9) 33 21 10. (10) 39 21 10. (11) 11 21 45. (12) 11 21 10. (13) 16 21 10. (14) 14 21 41.
NEWTON 09	(1) 16 21 10. (2) 52 21 10. (3) 14 21 41. (4) 11 21 41. (5) 14 21 10. (6) 14 21 41. (7) 42 21 10. (8) 14 21 10. (9) 14 21 41. (10) 16 21 41. (11) 14 21 41. (12) 16 21 41. (13) 59 21 41. (14) 14 21 41. (15) 11 21 10. (16) 14 21 41. (17) 41 21 41. (18) 39 21 10. (19) 14 21 45. (20) 41 21 10. (21) 16 21 41. (22) 41 21 10. (23) 14 21 10. (24) 33 21 41.
GALOIS 10	(1) 11 21 10. (2) 14 21 41. (3) 57 21 10. (4) 11 21 41. (5) 14 21 10. (6) 39 21 10. (7) 16 21 41. (8) 14 21 10. (9) 21 41. (10) 14 21 41. (11) 11 21 45. (12) 16 21 41. (13) 14 21 10. (14) 11 21 41. (15) 14 21 10. (16) 11 21 10. (17) 14 21 10. (18) 16 21 41. (19) 59 21 10. (20) 31 21 41. (21) 33 21 10. (22) 14 21 41. (23) 33 21 10. (24) 39 21 41. (25) 14 21 10. (26) 19 21 10. (27) 14 21 41. (28) 19 21 41.

1(e) Busca de padrões – a construção de uma microteoria do corpus: Tendo como origem os pressuposto de investigação apresentados anteriormente (TOGNINI-BONELLI, 2001; HALLIDAY, 1992), as pesquisas gramaticais realizas seguindo uma metodologia de *corpus* partem da abordagem baseada para a direcionada pelo *corpus*, construindo a partir dos dados uma microteoria da língua. Com isso, a pesquisa tem início no estrato da expressão e se move para o estrato da organização, e das ocorrências de itens específicos para os padrões mais gerais de organização.

A metodologia da perfilação gramatical permite a condução desse tipo de pesquisa porque extrai da expressão um *corpus* gramatical e outro do desenvolvimento do texto. Tendo

em mãos o *corpus* gramatical, torna-se possível realizar análises do funcionamento gramatical em conjunto com sua frequência e padrões de ocorrência. Dentro do escopo da LC, busca-se por "linhas de concordância", "listas de palavras", "colocações", dispersão e agrupamento, de forma análoga às pesquisas lexicais e da expressão.

Neste Exemplo de Pesquisa, começando a análise pela lista de concordâncias, esta mostra que a maior variação de funções no *corpus* se dá nos sistemas da TRANSITIVIDADE (11, 14, 57, 39, 16, 21, etc.), enquanto o sistema de MODO emprega sempre a mesma função (21), e o sistema de TEMA varia menos (10, 41 ou 45) (Figura 7).

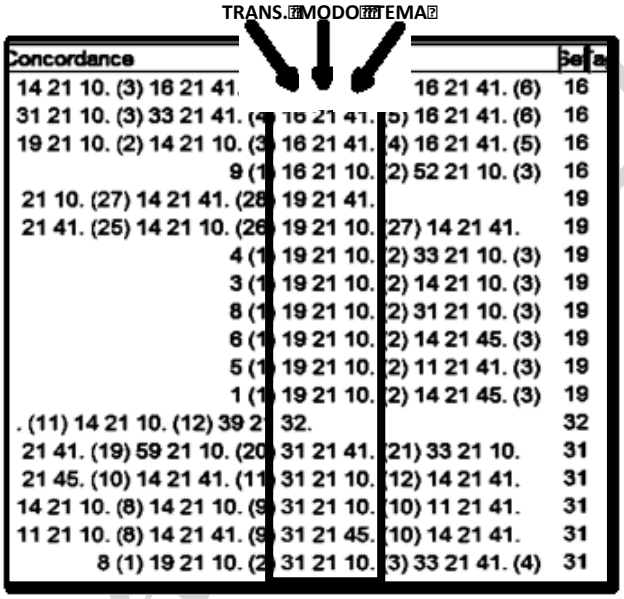


FIGURA 7 – Lista de concordâncias gramaticais.

A lista de concordâncias também mostra que o maior número de ocorrências é dos subtipos de Processo Material. Esse fato é corroborado por outros tipos de padrão que podem ser buscados.

A dispersão, efetuada pelo *plot* – com o auxílio da ferramenta Concord do *software* WordSmith Tools (SCOTT, 2007) – mostra que, no sistema de TRANSITIVIDADE, os Processos Materiais (Material: Transitivo e Transformativo [14]; Material: Transitivo e Criativo [16] e Material: Intransitivo e Transformativo [11]) apresentam maior dispersão, contribuindo assim para todas as fases do texto. De outra forma, os outros tipos de Processo, em particular os Relacionais (Relacional: Atributivo e Possessivo [32]; Relacional: atributivo e Intensivo [33] e Relacional: Identificativo e Intensivo [39]) ocorrem pontualmente, sugerindo a realização de um trabalho distinto no texto. Por exemplo:

(1) NEWTON 09 (4) 11 21 41. Por ocasião da peste, Newton **voltou <MATERIAL>** para casa
(2) VON NEUMANN 08 (2) 31 21 10. **Foi <RELACIONAL>** professor em Berlim e Hamburgo
(3) NEWTON 09 VERBAL (20) 41 21 10. Wallis lhe **comunica <VERBAL>** que, na Holanda, o Cálculo é considerado descoberta de Leibniz
(4) LEIBNIZ 05 (24) 59 21 45. Otimista ao extremo, sempre **acreditou <MENTAL>** numa universalização da linguagem

A Figura 8 mostra a dispersão das funções de TRANSITIVIDADE (Material: 11, 14, 16 e 19; Relacional: 33 e 39; Verbal: 42; e Mental: 52 e 59) no texto NEWTON 09.

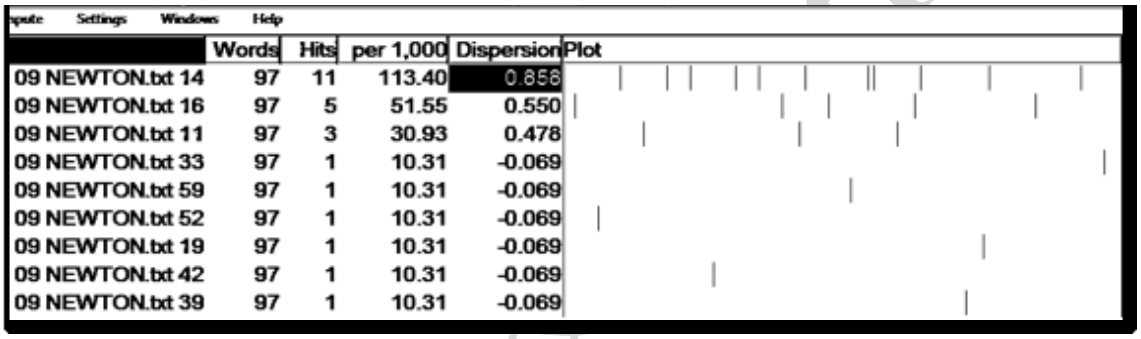
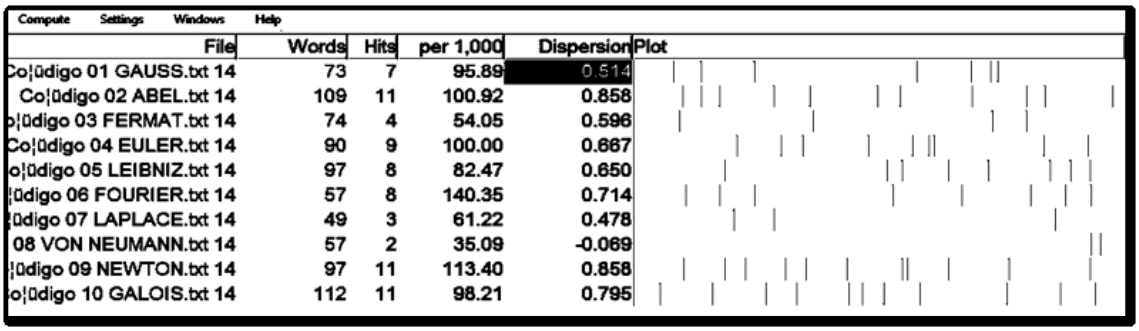


FIGURA 8 – Dispersão (*plot*) dos Tipos de Processo no texto NEWTON 09.

Dentre os subtipos de Processo Material, o mais frequente é o Transitivo e Transformativo [14]. A essa função cabe realizar uma mudança na Meta do Processo. Por exemplo:

(5) FOURIER 06 (2) 14 21 45. Órfão aos 8 anos, Fourier **foi colocado <MATERIAL>** no Colégio Militar
(6) GALOIS 10 (8) 14 21 10. Cauchy **perdeu <MATERIAL>** seu trabalho

Quando se observa o trabalho dessa função ao longo do *corpus*, encontra-se, de maneira semelhante, o trabalho importante desempenhado por ela nos textos. A seguir, a Figura 9 mostra como a função está dispersa.



Compute Settings Windows Help

File	Words	Hits	per 1,000	Dispersion	Plot
Co¦ûdigo 01 GAUSS.txt 14	73	7	95.89	0.514	
Co¦ûdigo 02 ABEL.txt 14	109	11	100.92	0.858	
¦ûdigo 03 FERMAT.txt 14	74	4	54.05	0.596	
Co¦ûdigo 04 EULER.txt 14	90	9	100.00	0.667	
o¦ûdigo 05 LEIBNIZ.txt 14	97	8	82.47	0.650	
ûdigo 06 FOURIER.txt 14	57	8	140.35	0.714	
ûdigo 07 LAPLACE.txt 14	49	3	61.22	0.478	
08 VON NEUMANN.txt 14	57	2	35.09	-0.069	
¦ûdigo 09 NEWTON.txt 14	97	11	113.40	0.858	
o¦ûdigo 10 GALOIS.txt 14	112	11	98.21	0.795	

FIGURA 9 – Dispersão (*plot*) Material: Transitivo e Transformativo no *corpus*.

A busca pelos agrupamentos do *corpus* revela ainda um outro aspecto importante dessa função, a saber, a maneira como ela forma padrões oracionais. Em primeiro lugar, os dois padrões oracionais mais frequentes empregam o Processo Material: Transitivo e Transformativo. O padrão 142141 é o mais frequente, com 106 ocorrências. Em seguida, 142110 aparece com 74 ocorrências (Figura 10).

N	Cluster	Freq.
1	14 21 41	106
2	14 21 10	74
3	16 21 41	69
4	33 21 10	47
5	11 21 10	39
6	39 21 10	34
7	11 21 41	25
8	16 21 10	25
9	14 21 45	19
10	19 21 10	14
11	11 21 45	14
12	39 21 41	13
13	33 21 41	11
14	41 21 41	10
15	59 21 10	9
16	31 21 10	8
17	21 21 41	8
18	59 21 41	8
19	41 21 10	7
20	16 21 45	6
21	32 21 10	6

FIGURA 10 – Agrupamentos (*clusters*) mais frequentes no *corpus*.

Na primeira linha (maior ocorrência) o código ‘142141’ significa orações Materiais Transitivas e Transformativas [14] Indicativas Declarativas Temporais [21] Tema Proeminente: Perspectiva [41].

O Modo Indicativo Declarativo, que ocorre em todas as orações do *corpus*, é empregado na biografia para estabelecer uma relação de especialista/leigo entre o produtor e o

leitor do texto, uma vez que não abre espaço para negociação. Soma-se a isso o fato de todas serem Temporais, não havendo modalização/possibilidade/dúvida. Além disso, O tempo empregado é sempre o passado (ou às vezes o "presente histórico", ou o presente simples em orações materiais tipicamente realizadas pelo presente-no-presente), o que contribui para a construção da narrativa, fundamental para o tipo de texto da biografia.

O Tema Proeminente: Perspectiva é realizado por Circunstâncias de Intensificação: Localização Espaço-Temporal, Modo, Causa e Contingência. Na biografia, as Circunstâncias fazem com que o texto tenha sempre como ponto de partida os lugares onde os matemáticos nasceram, estudaram, trabalharam, etc., bem como as datas. Do ponto de vista do discurso, criam uma linha do tempo para os eventos da vida dos matemáticos, que passa pelos lugares onde viveram.

O agrupamento com os Processos do tipo 14 mostra como as atividades dos matemáticos formam um padrão com a linha do tempo da biografia quando em conjunto com Modo 21 e Tema 41. Por exemplo:

> **(7)** *GAUSS 01 (5) 14 21 41. Ainda nesta obra* **<TEMA>** *Gauss apresenta* **<MATERIAL>** *a lei da reciprocidade quadrática* **<META>**
> **(8)** *GAUSS 01 (13) 14 21 41. No começo do séc. XIX* **<TEMA>** *abandonou* **<MATERIAL>** *a Aritmética* **<META>**
> **(9)** *ABEL 02 (4) 14 21 41. Nesta época* **<TEMA>**, *Abel conseguiu generalizar* **<MATERIAL>** *o teorema binomial* **<META>**

Retomando a ideia da construção de uma microteoria sobre o *corpus*, é possível, com este pequeno Exemplo de Pesquisa apresentar uma descrição de como as funções gramaticais operam. Devido ao seu número maior de ocorrências, o foco se deu sobre a função Processo Material: Transitivo e Transformativo [14]. Foi possível mostrar a sua frequência, a distribuição nos textos, a dispersão ao longo dos textos, bem como os agrupamentos que forma, nos padrões oracionais, com outras funções.

Apesar de limitada a esta função, determinada pelo escopo deste Exemplo de Pesquisa, a análise consegue revelar o potencial da perfilação gramatical. Ela aponta, assim, para a possibilidade desse tipo de análise ser feita para quaisquer funções gramaticais de interesse.

5. Conclusões

Demonstrando como as metodologias de *corpus* podem contribuir para o desenvolvimento das pesquisas do estrato da gramática, este artigo ofereceu uma metodologia de perfilação gramatical, a qual possibilita compilar e analisar *corpora* gramaticais para além do que podem oferecer as buscas e contagens computacionais, restritas às realizações por palavras, compreendendo igualmente funções encobertas, realizadas de forma criptotípica.

A justificativa para uma metodologia própria para os estudos de *corpora* gramaticais se deve ao fato de que a maior parte das funções gramaticais, diferentemente dos itens lexicais, não é realizada de forma aparente pela expressão, mas de forma encoberta a partir de um conjunto de reactâncias. Isto fica visível no caso das funções oracionais, como, por exemplo as funções dos sistemas de TRANSITIVIDADE, MODO e TEMA, objetos da presente pesquisa.

A perfilação gramatical aqui apresentada se mostrou eficaz ao atender as necessidades principais apontadas pelas pesquisas de *corpus* e gramática (cf. HALLIDAY, 1992; TOGNINI-BONELLI, 2001; CONRAD, 2010), a saber, a mudança de foco do léxico para a gramática; da expressão para o conteúdo; e da instância para o sistema.

No que diz respeito à mudança do léxico para a gramática, a perfilação não considera os itens lexicais e suas relações, mas, de outra forma, busca generalizar as funções desses itens para uma unidade determinada e, segundo essas generalizações, constitui o *corpus*. Neste trabalho a unidade selecionada foi a oração e as generalizações foram os sistemas de TRANSITIVIDADE, MODO e TEMA.

Quanto à mudança da expressão para o conteúdo, a perfilação se liberta da obrigatoriedade de a palavra gráfica/*token* ser a unidade de análise. Ao partir da distância topológica entre as funções e o emprego das distâncias como valores categóricos, estes se emparelham aos *tokens*, gerando um novo tipo de unidade de investigação de *corpora*, aqui denominada "*token* gramatical".

Por fim, ao tratar da relação entre instância e sistema, a perfilação gera resultados que podem alimentar a própria descrição gramatical, advindos da busca por padrões de ocorrência, coocorrência e agrupamentos de funções gramaticais que, sem as frequências não seria possível observar. No que diz respeito a esse tipo de frequência, Biber *et al*. (2004, p. 376) afirmam:

> A frequência, por si só, não possui qualquer poder explanatório. Na verdade, acontece o contrário: a frequência mostra padrões que, esses sim, precisam ser explicados. O valor da frequência, e da pesquisa em *corpus* como um todo, está no fato de ela identificar padrões que, de outra forma, escapariam ao analista.[4]

A perfilação gramatical segue exatamente o que afirmam Biber *et al*. (2004), por assim dizer, fazendo o caminho inverso. Ela mostra como a explicação já existente para as funções gramaticais pode ser vista na frequência. É nesse sentido que ela alimenta a descrição.

Ao realizar essas mudanças, a perfilação gramatical apresentada neste artigo aponta como, a partir da expressão, do léxico e da instância – ou o formato em que o *corpus* é coletado – é possível extrair os dados para compilar e anotar um *corpus* para funções gramaticais. A partir deste, retomando a ideia da construção de uma microteoria sobre o *corpus*, realizam-se descrições de como as funções gramaticais operam. Neste artigo foi possível oferecer um Exemplo de Pesquisa realizado com a metodologia em questão.

Focando-se na função com número maior de ocorrências, o artigo mostrou um conjunto de possibilidades de análise advindas da perfilação gramatical, tais como a frequência da função, a distribuição nos textos, a dispersão ao longo dos textos, bem como os agrupamentos que forma, nos padrões oracionais, com outras funções. Para as pesquisas de *corpus* de base gramatical, a perfilação gramatical mostrou como é possível compilar e extrair dados de *corpora*, inclusive para a forma como são empregados na dinâmica textual.

Referências:

BIBER, D.; CONRAD, S.; CORTES, V. **If You Look at...** : Lexical Bundles in University Teaching and Textbooks. Applied Linguistics, Oxford, n. 25(3), p. 371–405, 2004.

CÂNDIDO-JR, A. **Análise bidirecional da língua na simplificação sintática em textos do português voltada à acessibilidade digital**. 2013. 225f. Tese (Doutorado em Ciências de Computação e Matemática Computacional). Instituto de Ciências Matemáticas e da Computação, Universidade de São Paulo, São Carlos, 2013.

[4] We do not regard frequency data as explanatory. In fact we would argue for the opposite: frequency data identifies patterns that must be explained. The usefulness of frequency data (and corpus analysis generally) is that it identifies patterns of use that otherwise often go unnoticed by researchers.

CONRAD, S. What can a *corpus* tell us about grammar? In: O'KEEFFE, A.; McCARTHY, M. (Eds.). **The Routledge Handbook of** corpus **Linguistics**. Londres e Nova Iorque: Routledge, 2010, p. 227-240.

HALLIDAY, M. A. K. **System and function in language**. London: Oxford University Press, 1976.

HALLIDAY, M. A. K. Language as System and Language as Instance: the corpus as a Theoretical Construct. In: SVARTVIK, J. (Ed.) **Directions in** corpus **Linguistics**: Berlim e New York: Mouton de Gruyter, 1992, p.61-77. crossref http://dx.doi.org/10.1515/9783110867275.61

HALLIDAY, M.A.K. On Grammar and Grammatics. In: HASAN, R.; CLORAN, C.; BUTT, D. (Eds.) **Functional Descriptions**: Theory in Practice. Amsterdam/ Philadelphia: John Benjamins, 1996, p.1-38. crossref http://dx.doi.org/10.1075/cilt.121.03hal

HALLIDAY, M. A. K. **On grammar**. London: Continuum, 2002. (The collected works of M. A. K. Halliday, v. 1).

HALLIDAY, M. A. K.; JAMES, Z. A Quantitative study of polarity and primary tense in the English finite clause. In: SINCLAIR, J.; HOEY, M.; FOX, J. (Eds.). **Techniques of description**: spoken and written discourse. London: Routledge, 1993.

HALLIDAY, M. A. K.; MATTHIESSEN, C. **An introduction to functional grammar**. 3a ed. London: Edward Arnold, 2004.

IEZZI, G.; MURAKAMI, C.; DOLCE, O.; HAZZAN, S. **Fundamentos de Matemática Elementar**. São Paulo: Atual Editora, vol. 1-10, 1977/78.

KE, S-W. Clustering a translational corpus. In: OAKES, M.; MENG J. (Eds.). **Quantitative Methods in** corpus**-Based Translation Studies**: A Practical Guide to Descriptive Translation Research. Amsterdam: John Benjamins, 2012. crossrefhttp://dx.doi.org/10.1075/scl.51.06ke

LEMKE, J. Text Production and Dynamic Text Semantics. In: VENTOLA, E. (Ed.). **Functional and systemic linguistics**: approaches and uses. Berlim: Mouton de Gruyter. 1991. crossref http://dx.doi.org/10.1515/9783110883527.23

MARTIN, J. R.; MATTHIESSEN, C. Systemic typology and topology. In: CHRISTIE, F. (Ed.). **Literacy in Social Processes**: papers from the inaugural Australian Systemic Linguistics Conference, held at Deakin University, January 1990. Darwin: Centre for Studies in Language in Education, Northern Territory University. 1991.

MARTIN, J. R. **Systemic functional grammar**: a next step into the theory – axial relations. Pequim: Higher Education Press, 2013.

PAGANO, A.; FIGUEREDO, G.; LUKIN, A. **Modelling proximity in a** corpus **of literary retranslations**: a methodological proposal for clustering texts based on systemic-functional

annotation of lexicogrammatical features. In: International Quantitative Linguistics Conference. Olomouc: Univerzita Palackého v Olomouci, 2014, no prelo.

QUIROZ, B. **The interpersonal and experiential grammar of Chilean Spanish**: towards a principled systemic-functional description based on axial argumentation. 2013. 403f. Tese (Doutorado em Linguística). Universidade de Sydney, Sydney, 2013.

SCOTT, M. 2007. **WordSmith Tools**. Oxford: Oxford University Press.

SOUZA, A. Topologia geral de vários ângulos. 2013. Disponível em <http://topologia-geral.ourproject.org>. Acesso em: 10 set. 2014.

TOGNINI-BONELLI, E. corpus **linguistics at work**. Amsterdam e Philadelphia: John Benjamins, 2001. crossref http://dx.doi.org/10.1075/scl.6

VIANA, V. Linguística de *corpus*: conceitos, técnicas e análises. In: VIANA, V.; TAGNIN, S. (Orgs.). ***Corpora*** **no ensino de línguas estrangeiras**. São Paulo: HUB Editorial, 2011.

WHORF, B. L. **Language, thought, and reality**. Cambridge: MIT, 1987.

Artigo recebido em: 15.09.2014
Artigo aprovado em: 27.11.2014

Os dizentes nos artigos científicos de Linguística – um estudo baseado na Linguística Sistêmico-Funcional e com o auxílio da Linguística de *Corpus*

The Sayers in linguistics' scientific articles – a study based on Systemic functional grammar and Corpus Linguistics

Fernanda Beatriz Caricari de Morais*

RESUMO: Este artigo analisa os tipos de dizentes, participantes agentes dos verbos do dizer, mais utilizados em artigos da área de Linguística, coletados através da plataforma digitar Scielo. Como fundamentação teórica e metodologia qualitativa de análise, utilizou-se a Linguística Sistêmico-Funcional, formulada por Halliday (1985, 1994) e Halliday e Matthiessen (2004). A Linguística de *corpus* possibilitou o tratamento computacional por meio do programa WordSmith Tools (Scott, 2008), que fornece dados quantitativos e contextos de ocorrência de determinadas palavras. Ele foi utilizado na análise para o auxílio na busca dos dizentes mais usados nos artigos científicos da área estudada. Os resultados mostram que os dizentes são utilizados para expressar conhecimento de pesquisas anteriores ou de pressupostos da área de pesquisa e o uso de construções com o clítico se promove o afastamento do autor/pesquisador do artigo. Pretende-se contribuir para a elaboração de materiais didáticos e cursos instrumentais que visem à compreensão e produção escrita de artigos científicos.

PALAVRAS-CHAVE: Linguagem Acadêmica. Linguística Sistêmico-Funcional. Linguística de *Corpus*.

ABSTRACT: This article presents an analysis of the types of sayer, partipants agent of sayer verbs, used in linguistic scientific articles, collected in the digital plataform Scielo. As theory and qualitative methodology, we use the systemic functional linguistics, formulated by Halliday (1985, 1994) and Halliday e Matthiessen (2004). The corpus Linguistics enabled the computational treatment, through the WordSmith Tools (Scott, 2008) program, that provides quantitative data and the context of certain words. It was used in the analysis for helping the search of types of sayer in area studied. The results show that the sayers were used for expressing knowledge about preview researches or knowledge area and the use of constructions with clitic se promotes the dismissal of author/researcher of the scientific article. We intend to contribute for subsidizing the elaboration of course books and courses with emphasis on comprehension and production of scientific articles.

KEYWORDS: Academic language. Systemic functional Linguistics. Corpus Linguistics.

1. Introdução

Este artigo visa analisar os tipos de dizentes, isto é, os agentes dos verbos do *dizer*, em termos tradicionais, mais utilizados em artigos científicos da área de Linguística. Segundo a

* Doutora em Linguística Aplicada e Estudos da Linguagem (PUC-SP). Realizou estágio pós-doutoral (PNPD/Capes) na Universidade Federal de Uberlândia no período de dezembro de 2013 a Maio de 2014.

abordagem sistêmico-funcional, dizente é o participante (agente) principal do processo verbal que é representado pelos verbos do dizer.

Os processos verbais são verbos de ação verbal, tem relação com fatos da ordem do "*dizer*". Para Halliday (1994, p.143), esses processos não precisam necessariamente pressupor um participante humano, porém devido a sua alta frequência de uso, no gênero selecionado, julgou-se pertinente investigar suas ocorrências para compreender quem são os dizentes dos processos verbais mais utilizados em artigos científicos de Linguística, buscando seus padrões de usos e verificando se há uma tendência por construções com primeira pessoa do singular e do plural nessa área.

Como hipótese inicial, pensou-se que a maioria das ocorrências estariam ligadas ao uso de "nós" como dizentes dos processos, pois, embora os artigos sejam escritos, muitas vezes, por um único autor, é comum, na área de Linguística, o uso da primeira pessoa do plural indicando falsa modéstia, um recurso para o distanciamento do autor no texto.

Porém, muitas das ocorrências encontradas mostram o uso da terceira pessoa e de construções com o clítico *se* que permitem o afastamento do autor no texto, deixando o texto mais impessoal.

Esses usos impessoais são bastante utilizados em gêneros que exigem uma linguagem mais elaborada, termos de Bernstein (1971), como o artigo científico. A impessoalidade pode ser explicada como um fenômeno característico da linguagem científica, que prima em ser sintética e com foco nas ações, nos processos que envolvem as pesquisas e não em quem as fizeram. Isso explica a relação de modéstia em que o autor se coloca no texto, exigência do gênero e da linguagem. O desfocamento também ocorre quando não é importante mencionar pesquisadores da área, o que não prejudica a compreensão do texto.

O artigo científico, um importante meio de divulgação do trabalho de cientistas, é escrito por membros altamente letrados da comunidade, usando sua variante mais elaborada da linguagem, é escrito e revisado com cuidado, para ser avaliado por pares e aceito ou não para uma revista considerada de alto nível. É um gênero que exige clareza, objetividade e síntese e se caracteriza ainda pelo uso de passiva e de outros recursos para omissão de participantes. Esses aspectos são discutidos por ampla gama de pesquisa sobre esse gênero – alguns expoentes no exterior são Swales (1989, 1990), Swales e Feak (1999), Bazerman (1984), Bhatia (1993), Atkinson (1996) e no Brasil: Aranha (1996, 2002, 2004, 2007), Motta Roth (1995, 2006), entre outros.

Para analisar os dizentes dos artigos científicos de Linguística, foram coletados aleatoriamente 100[1] artigos científicos, retirados de duas revistas científicas que estão disponíveis no Scielo: Documentação de Estudos em Linguística Teórica e Aplicada (*D.E.L.T.A.)* e Trabalhos em Linguística Aplicada, ambas com Qualis A1.

Os artigos foram submetidos a um tratamento computacional possibilitado pela Linguística de *Corpus* (LC), que se faz presente metodologicamente, nesta pesquisa, através de ferramentas do programa *WordSmith Tools* (Scott, 2008).

Por meio da ferramenta *concord* do referido programa, foram obtidas listas de concordância com os processos verbais mais frequentes (*explicar, discutir* e *afirmar*), buscando observar seus contextos de ocorrência para analisá-las com a base teórico-metodológica da Linguística Sistêmico-Funcional de Halliday (1985, 1994) e Halliday e Matthiessen (2004).

A abordagem semântico-funcional se preocupa em explorar como a língua é estruturada para o uso em diferentes contextos. Uma das premissas básicas desta teoria é que o uso da língua é motivado pelas relações sociais e que as escolhas léxico-gramaticais realizadas pelos falantes não são aleatórias e estão condicionadas pelo contexto.

Assim como o gênero artigo científico, os gêneros do discurso são formas de como a linguagem é organizada para alcançar propósitos sociais. As situações específicas que envolvem os gêneros podem ser definidas como cadeias semióticas que estão ligadas aos três tipos de funções da linguagem chamadas por Halliday de metafunções (ideacional, interpessoal e textual), base de análise desta pesquisa, detalhada no item seguinte.

2. A abordagem Sistêmico-Funcional

Esta pesquisa tem como fundamentação teórica a Linguística Sistêmico-Funcional (doravante LSF) de Halliday (1985, 1994) e Halliday e Matthiessen (2004). A LSF tem como foco a linguagem em uso, por isso sua preocupação é explorar como a língua é estruturada para o uso em diferentes contextos. Halliday (1994) define que uma das premissas básicas da abordagem sistêmico-funcional é que o uso da língua é motivado pelas relações sociais e que as escolhas léxico-gramaticais realizadas pelos falantes/escritores não são aleatórias e estão condicionadas pelo contexto.

[1] Acredita-se que este número de artigos é representativo e possibilita encontrar padrões de usos do dizentes na área pesquisada.

Para a LSF, a análise do discurso compreende dois níveis de alcance: contribuir para a compreensão do texto, visando mostrar como e por que o texto transmite significado da maneira como o faz e relaciona-se com a avaliação do texto, procurando mostrar por que o texto é ou não efetivo para os seus propósitos.

Halliday (1994, p.16) argumenta que uma análise do discurso não baseada em gramática não é uma análise completa, mas um simples comentário sobre o texto. A realização de um texto acontece através das relações semânticas e gramaticais. A gramática é requerida por fornecer uma compreensão clara do sentido e da efetividade de um texto, por isso precisa ter esta orientação semântica e funcional.

Na LSF, funcionalidade significa ser baseada no significado e o fato de ser gramática é entendido como a interpretação das formas linguísticas. Por isso, a gramática separa as possíveis variáveis e aponta suas possíveis funções para podermos dar a nossa interpretação de um texto tanto pela sua descrição semântica, como pelas características linguísticas.

A linguagem é vista como prática social, cujo uso motiva-se por uma finalidade. Nessa perspectiva, a LSF estuda as maneiras pelas quais as pessoas utilizam a linguagem para atingir determinados objetivos em situações específicas dentro de uma sociedade (HALLIDAY, 1985, p. 4). A linguagem é vista como um recurso usado pelos seres humanos para criar significados.

De acordo com essa perspectiva teórica, quando produzimos um texto (oral ou escrito), estamos realizando três tipos de significado simultaneamente. Significados relativos à representação da experiência através da língua; significados relativos às representações de poder e solidariedade, atitudes em relação ao outro e aos papéis sociais assumidos e significados relativos à organização do conteúdo da mensagem, relacionando o que se diz ao que foi dito. Na LSF, cada um desses significados está relacionado a uma metafunção da linguagem, chamadas por Halliday (1985, 1994) de *ideacional*, *interpessoal* e *textual*.

Como o interesse desta pesquisa é investigar os dizentes em artigos científicos de Linguística, concentrou-se na metafunção ideacional da linguagem, também chamada experiencial, que expressa o que está acontecendo no mundo externo (eventos) ou interno (pensamentos). Esta metafunção estuda a oração como representação, ou seja, estuda seu aspecto como um meio de representar padrões de experiência e reflete como o usuário fala sobre as ações, a situação, estados, crenças e circunstâncias (HALLIDAY, 1994, p.107).

A oração tem um papel central, pois é nela que se incorpora um princípio geral de modelagem da experiência, que é o princípio de que a realidade é construída através dos processos, dos participantes e das circunstâncias.

Thompson (1996, p.76), com base em Halliday (1985, 1994), discute que a linguagem, na perspectiva experiencial, forma uma série de recursos para se referir às entidades no mundo de forma que essas entidades atuem ou se relacionem umas com as outras. O autor simplifica dizendo que a linguagem reflete a nossa visão de mundo, constituída por: processos, participantes e circunstâncias.

Na visão sistemicista, a impressão mais poderosa que temos da experiência é de que ela consiste de eventos (acontecer, fazer, sentir, significar, ser e tornar-se). Todos esses eventos estão organizados na gramática da oração e o sistema gramatical pelo qual isso é alcançado é o da transitividade. De acordo com Halliday (1994), é o sistema da transitividade que constrói o mundo da experiência em um conjunto manipulável de tipos de processo. O processo, os participantes e as circunstâncias constituem o sistema da transitividade. A oração, nesta perspectiva, possibilita ao falante, através das escolhas dos processos (ações), dos participantes (pessoas ou coisas) e das circunstâncias (advérbios), expressar-se perante o mundo. Os processos são divididos em: materiais (fazer), mentais (pensamento), verbais (dizer), comportamentais (comportamentos físicos e psicológicos), relacionais (ser) e existenciais (haver).

Como esta pesquisa analisa os dizentes utilizados com os processos verbais mais frequentes (*explicar, discutir* e *afirmar*), é apresentado, no item seguinte, as particularidades desses processos, segundo a abordagem Sistêmico-Funcional da Linguagem.

2.1 Processos verbais

Os processos verbais são os de *dizer*. Halliday e Matthiessen (2004, p. 252) apontam que seu uso é importante em vários tipos de discurso como nas narrativas e no discurso acadêmico, por exemplo. Eles permitem projeção através de (1) citação (discurso direto) ou (2) reportagem (discurso indireto) de pesquisadores antecessores e teóricos. Também são usados para mostrar o posicionamento do autor por meio de verbos como: *explicar, discutir e afirmar*, por exemplo.

Ao contrário dos processos mentais (ligados à representação do pensamento), os verbais não requerem um participante consciente. O participante, chamado de Dizente, pode ser qualquer coisa.

Ao projetar uma oração, pode-se ter uma proposição (troca de informações) como em: *Ele me disse que teria jogo à noite* ou uma proposta (troca de bens e serviços): *Ele me disse que irá me buscar para ir ao jogo à noite*.

Além de serem processos que podem projetar, os verbais possuem outros três tipos de participantes além do Dizente, são eles:

- Recebedor: representa o participante a quem é dirigida a mensagem, como em: *Não me disseram que você viria*[2].
- Verbiagem: é a função que corresponde àquilo que é dito, pode ser o conteúdo do que é dito, como *suas sugestões* em: *Ela explicou suas sugestões para o trabalho*, ou, ainda, o nome do dito, como *palavra* em: *Não diga mais uma palavra*.
- Alvo: é a entidade que é focalizada pelo processo de dizer. Nesse caso, é como se o Dizente estivesse agindo verbalmente sobre o outro, como *Maria* em: *Ele criticou Maria durante a reunião*[3].

Estes três últimos participantes têm papel importante na oração, porém, como o foco deste trabalho é compreender os tipos de participantes que têm papel de agente nas orações verbais, analisou-se apenas os dizentes.

3. O uso da Linguística de *Corpus*

A Linguística de *Corpus* (LC) se encontra presente metodologicamente, neste artigo, através das ferramentas computacionais utilizadas para analisar as escolhas linguísticas dos artigos científicos da área Linguística. A LC trabalha dentro de um quadro conceitual formado por uma abordagem empirista e uma visão da linguagem enquanto sistema probabilístico, no qual alguns traços linguísticos são mais frequentes que os outros, conforme discute Berber-Sardinha (2000, p. 349).

[2] Exemplo criado para fins explicativos.
[3] Exemplo criado para fins explicativos.

A LC fornece um mapeamento regular entre a frequência maior ou menor de um traço e o contexto de ocorrência, há uma relação entre as características linguísticas e as situacionais (os contextos de uso).

Para Biber et al. (1998, p. 9), a abordagem baseada em *corpus* é bastante útil, uma vez que "[...] quase todas as áreas da linguística podem ser estudadas a partir da perspectiva do uso, e a abordagem baseada em *corpus* fornece um conjunto de instrumentos particularmente eficaz para tais investigações".

Segundo Berber-Sardinha (2004, p. 34), a Linguística de *Corpus* fornece um suporte metodológico adequado às pesquisas que utilizam a Linguística Sistêmico-Funcional, por também trabalhar dentro de uma visão de linguagem enquanto sistema probabilístico.

A LC possibilita o estudo das regularidades lexicais, possibilitando estudos sistemáticos, a partir de *corpus*, descrevendo os tipos de associação frequentes encontrados na língua em uso.

3.1 O *corpus*

Para realizar esta análise, utilizamos 100 artigos de Linguística que fazem parte do Projeto SAL[4] coletados no período de 2000 a 2007. Os artigos foram coletados através da internet pelo *site Scielo* e gravados em arquivos individuais no formato *txt*, formato que permite seu uso no programa *WordSmith Tools* (SCOTT, 2008).

A seguir, apresentamos uma tabela com as principais características do *corpus*:

Tabela 1: Resultados estatísticos dos *corpora*

Estatísticas	Total
Total de palavras	873.523
Palavras diferentes[5]	35.841
Número de palavras do menor artigo	2.086
Número de palavras do maior artigo	15.454

O *corpus* foi submetido a um tratamento computacional possibilitado pela LC, que se faz presente metodologicamente, nesta pesquisa, através de ferramentas do programa *WordSmith Tools* (SCOTT, 2008).

[4] *O projeto SAL - Systemics Across Language*, é desenvolvido em parceria com pesquisadores da China, Argentina, México e Tailândia que procuram entender as características específicas e universais que partilham as línguas. No Brasil, o foco do projeto é estudar a linguagem acadêmica.
[5] Número de itens lexicais sem suas repetições.

Foram utilizadas duas das suas principais ferramentas para a análise: a lista de palavras (*wordlist*) e o concordanciador (*concord*). A primeira foi utilizada para organizar o *corpus* em listas das palavras. Elas podem ser ordenadas alfabeticamente ou pela frequência com que aparecem, começando pela palavra de maior frequência. Por meio dessa mesma ferramenta, foram obtidos dados estatísticos dos textos, como os dados da tabela acima. Ela ajudou tanto na organização dos dados estatísticos como na análise dos processos verbais mais frequentes utilizados nos artigos científicos.

Foram procurados, nas listas de palavras, os processos verbais mais utilizados nos artigos e observados os contextos em que eles ocorrem através do concordanciador. Nas concordâncias, a palavra de busca aparece destacada e no centro do cotexto em que ocorre - chamado de horizonte. Concentrou-se em um horizonte de cinco palavras à direita e cinco palavras à esquerda, mas, sempre que necessário, foi visto um contexto maior incluindo todo um parágrafo ou o texto como um todo.

Através das listas de concordância foi possível estudar o contexto de ocorrência das palavras de busca simultaneamente em todo o *corpus*. A análise, apresentada no item seguinte, está baseada na Linguística Sistêmico-Funcional, que é uma teoria de linguagem e um método de análise de textos em seus contextos de uso permitindo-nos entender como os indivíduos usam a linguagem e como a linguagem é estruturada em seus diferentes usos (HALLIDAY, 1994).

4. Análise dos dizentes utilizados em artigos científicos de Linguística

Para entender como os dizentes são utilizados nos artigos de Linguística, foi necessário analisar os participantes das ocorrências com os processos verbais mais frequentes – *explicar, discutir* e *afirmar*[6].

A tabela abaixo apresenta o número de ocorrências de cada processo, o número total de ocorrências analisadas e sua porcentagem no *corpus*:

[6] Outros processos verbais frequentes*: apontar, mostrar, referir, falar, explicitar, ressaltar, citar*, entre outros que não são o foco deste estudo.

Tabela 2: Dados dos processos verbais analisados.

Processo	No. de ocorrências	% no *corpus*
Afirmar	267	0,05%
Explicar	257	0,05%
Discutir	212	0,04%
Total	736	0,14%

Listas de concordâncias foram feitas e as ocorrências foram agrupadas de acordo com o tipo de dizente. Ao contabilizá-las, foi possível organizar os usos mais frequentes na figura abaixo:

	Explicar				Discutir				Afirmar			
	H	**N**	**A**	**S**	**H**	**N**	**A**	**S**	**H**	**N**	**A**	**S**
1ª p. sing. Pres. do ind.	1	-	-	-	4	-	-	-	**2**	**-**	-	-
1ª p. pl. Pres. do ind./perf. do ind.	1	-	-	-	15	-	-	-	8	-	-	-
1ª p. sing. Perf. do ind.	0	-	-	-	0	-	-	-	2	-	-	-
3ª p.sing. Pres. do ind.	97	13	-	2	22	17	-	18	115	13	-	2
3ª p. pl. Pres. do ind.	3	12	-	2	9	4	-	6	25	-	-	-
3ª p. sing. Perf. do ind.	2	1	-	-	3	-	-	2	6	-	-	3
3ª p. pl. Perf. do ind.	-	-	-	-	-	-	-	-	-	-	-	-
Infinitivo	**31**	**45**	-	20	49	7	-	9	53	6	-	26
Part. passado	1	18	8	-	7	0	40	-	1	-	5	-
Total	**136**	**89**	**8**	**24**	**109**	**28**	**40**	**35**	**212**	**19**	**5**	**31**

Figura 1: Formas verbais mais utilizadas com os processos *explicar, discutir* e *afirmar*.

As letras correspondem a: H: participante humano; N: participante não-humano; A: apagamento do agente (em construções passivas analíticas); e S: uso do clítico *se*. Como se pode notar, as formas que representam o autor no artigo, 1ª pessoa do singular e do plural, são menos usadas do que a 3ª pessoa, o que demonstra que, neste gênero acadêmico, as representações estão relacionadas à busca pela neutralidade, característica dos discursos científicos.

Como hipótese inicial, pensou-se que haveria muitas representações do "eu" (autor do artigo) ou "nós", conforme mencionado na introdução deste artigo, porém foram encontradas muitas ocorrências que representavam pesquisadores antecessores que são mencionados nos artigos para dar maior credibilidade ao trabalho ou serem contestados na argumentação.

Na seção revisão da literatura/fundamentação teórica, é esperado que o pesquisador faça menções aos pesquisadores/teóricos de sua área de estudo, situando sua pesquisa na área, ao mesmo tempo que a diferencia das já realizadas. A representação, segundo Rajagopalan (2003), não é algo que se dá automaticamente, ela passa por certas escolhas conscientes. A questão da escolha é uma questão-chave quando se discute política de representação. Dessa maneira, é importante analisar as representações feitas no gênero artigo científico, compreendendo-as como questão linguística e também política, questões estas consideradas indissociáveis para o autor citado.

Nesta análise, propõe-se a discutir as escolhas linguísticas feitas pelos autores dos artigos através dos dizentes, mostrando quais são as tendências de escrita da área.

Abaixo, algumas das ocorrências encontradas com 1ª pessoa do singular e do plural que mostram que parece haver um movimento na área que utiliza formas mais pessoais para a construção da identidade do pesquisador:

1) *As alterações sugeridas no nível pragmático-discursivo implicam em algumas modificações na materialidade textual, pois, como* ***afirmei*** *anteriormente, as categorias caracterizadoras de gêneros são semiotizadas via funções da linguagem no sistema linguístico.* (Ling003).

2) *Na segunda [parte],* ***discuto*** *os mecanismos alternativos de negação oracional no português do Brasil. (Ling054).*

3) *Na primeira parte, na medida em que* ***explico*** *os principais referenciais teóricos utilizados neste texto, analiso o jornal escolar enquanto um gênero híbrido. (Ling006).*

Ao contrário de algumas áreas da ciência (ciências duras) que repudiam a construção da identidade social do autor no discurso, na área da Linguagem há claramente uma tendência de mudança em que é permitido ao autor se representar no texto.

Muitas das ocorrências de 1ª pessoa do plural indicam o uso da modalidade, conforme sublinhado nas ocorrências abaixo:

1) *Apesar do destaque dado ao contexto, conforme se verifica no trecho acima, <u>podemos</u> **afirmar** que um limite da pragmática griceana é exatamente o não tratamento desta noção. (Ling032).*

2) *<u>Podemos</u> também **afirmar** que o uso das regras para implementação rítmica não ocorre no início do processo de aquisição. (Ling007).*

3) *Também <u>podemos</u> **explicar** a diferença entre as sentenças abaixo. (Ling049).*

4) *Desta maneira, <u>podemos</u> **explicar** o motivo de não haver violação do princípio Minimizar. (Ling006).*

5) *Assim como no caso das crianças em fase de construção de sua identidade social, <u>podemos</u> **discutir** as questões identitárias inerentes, por exemplo, à carta do guarda municipal: em que medida a incorporação total ou parcial do ethos interfere na identidade social do guarda municipal? (Ling012).*

As ocorrências acima são acompanhadas do adjunto modal de baixo grau *pode*, utilizado para indicar possibilidade, atenuando afirmações no discurso. Para Fairclough (2001, p. 180), a modalidade trata da relação entre os produtores e as proposições, de comprometimento ou, inversamente, do distanciamento entre produtores e proposições.

O sistema de modalidade, ligado aos significados interpessoais, revela o grau de engajamento, envolvimento e responsabilidade que um indivíduo assume diante de uma declaração, ou grau de polidez em propostas ou declarações em um evento de comunicação (HALLIDAY, 1994, p. 85-105). A modalidade pode ser representada de diversas formas tais como: adjuntos conjuntivos, adjuntos modais e operadores modais (HALLIDAY, 1994, p. 49).

Os usos das ocorrências acima indicam possibilidade, atribuindo baixo grau de responsabilidade aos sujeitos representados nas orações, evitando comprometimento das afirmações dos autores nos artigos.

Fairclough (2001, p. 203) discute o discurso acadêmico que possui tradição em evitar a modalidade categórica (X é Y), por motivos retóricos, motivados pelas projeções de uma subjetividade e um *ethos* cautelosos e circunspectos aprovados para acadêmicos e não por baixa afinidade com as proposições.

As ocorrências de 3ª pessoa são maioria (com os processos *explicar* e *afirmar*), conforme Tabela 2, no entanto, a escolha de seus participantes permite separar as ocorrências

em dois grupos. O primeiro deles é o de participantes humanos - autores/pesquisadores antecessores. Ao reportar o discurso/pesquisas de outros, o autor do artigo embasa sua pesquisa:

1) *Berber Sardinha (2004:42)* ***explica*** *que a colocação é uma "associação entre itens lexicais, ou entre o léxico e campos semânticos". (Ling029).*

2) *Ainda sobre a flexibilidade da pontuação em relação ao gênero, Halliday (1989:37-38)* ***explica*** *que há registros em que a pontuação é reduzida ao mínimo, como na linguagem legal. (Ling075).*

3) *Scarpa (1997, 2001) é a única que conhecemos que* ***discute*** *a aquisição de processos fonológicos que têm consequências para a estrutura rítmica. (Ling007).*

4) *O autor* ***afirma*** *que não há polêmica a respeito da existência destes mecanismos: a questão é a análise deste "conhecimento". (Ling032).*

O segundo grupo é formado por ocorrências em que o participante representado não é humano, mas sim uma teoria ou a própria pesquisa sendo referenciada:

1) *Ao contrário, a ADC* ***afirma*** *que o linguístico é social (Kress 1989). (Ling104).*

2) *No entanto, a linguística cognitiva* ***afirma*** *que as metáforas são de textura aberta (open-ended), não estando esse mapeamento sujeito a restrições. (Ling091).*

3) *Cabe, então, à teoria* ***explicar*** *como tais gramáticas são construídas e operadas pela própria mente humana (Ling018).*

4) *Este trabalho pretende* ***discutir*** *uma possível abordagem marxista da linguagem apresentando sumariamente a concepção de linguagem de alguns autores confessadamente marxistas(Ling061).*

5) *Este trabalho* ***discute*** *a legitimação do funcionamento do literário nesse tipo de ensino porque considera que os próprios conceitos genéricos tradicionais de literatura com que ainda se opera não são consistentes com o que exige a prática pedagógica. (Ling030).*

As escolhas feitas nas ocorrências acima, ao representar a teoria e não o teórico e ao representar o trabalho/a pesquisa ao invés do pesquisador mostram a característica do discurso

acadêmico em não enfatizar a representação humana por ser desnecessária, podendo ser retomada no contexto do artigo, ou, podendo ser inferida pelo leitor.

No discurso da ciência, é importante representar o mundo em forma de "coisas", incluindo entidades virtuais que podem ser trazidas se requeridas ao discurso. Algumas dessas entidades são construtos teóricos, enquanto outras funcionam como localizadas no argumento e depois desaparecem. Simbolicamente, o discurso da ciência representa o mundo em nomes, em coisas estáveis enquanto são observados em experimentos e medidos e pensados/compreendidos (HALLIDAY, 2004, p. 21).

Chama-se atenção no discurso acadêmico, o frequente uso do clítico *se*, utilizado para desfocar ou apagar os participantes do discurso. Em pesquisas anteriores, Morais (2013a, 2013b), artigos de diversas áreas da ciência foram analisados e constatou-se que há três categorias de usos do clítico *se* em artigos científicos: *se* em construções agnatas; *se* em construções com desfocamento de participante e *se* em construções médias[7]. As ocorrências encontradas nos artigos de Linguística se referem à segunda categoria, de desfocamento de participantes. O termo desfocamento foi mencionado pela primeira vez por Shibatani (1985) que, embora não tenha estudado a língua portuguesa, estudou as formas de *desfocamento de Agente* em muitas línguas como: ainu, chamorro, turco, quéchua, espanhol, francês e japonês, e discutiu a falta de atenção com as funções do *desfocamento.*

Para Shibatani (1985, p. 832), o desfocamento é a função primária e não uma mera consequência da promoção do objeto/paciente. Nas ocorrências encontradas nos artigos de Linguística, não temos o agente explícito, isto é, não se sabe ao certo quem é o dizente, mas as circunstâncias grifadas permitem pensar que o participante da ação é o autor/pesquisador do artigo:

1) *Na seção 1, apresenta-se a delimitação dos dados analisados; na seção 2,* ***explica-se*** *o status prosódico do prefixo.... (Ling057).*

2) *Com esta análise,* ***discute-se*** *a questão da (não) codificação gramatical do imperativo no português brasileiro por meio da sintaxe da negação e dos clíticos... (Ling 01).*

3) *O português tem sido considerado uma língua SVO, como já* ***se afirmou*** *no início deste trabalho. (Ling089).*

[7] Para maiores detalhes vide trabalho de Morais (2013a e 2013b).

> 4) *Neste artigo*, ***discute-se*** *a percepção de alguns estudiosos de que a pontuação demarca aspectos rítmicos da linguagem. (Ling84).*

Estes tipos de circunstâncias *na seção 1, com esta análise, no início deste trabalho* e *neste artigo* restringem a participação apenas ao autor que, nestes casos, é o Dizente dos verbos *explicar, afirmar* e *discutir*.

Apesar do dizente estar apagado na oração, as circunstâncias, destas ocorrências, indicam que a participação pode ser pressuposta com base na análise do contexto.

Sabe-se que muitos dos estudos sobre o clítico *se* em língua portuguesa estão ligados às discussões sobre a diferenciação entre índice de indeterminação do sujeito e partícula apassivadora.

Segundo as gramáticas tradicionais da língua portuguesa, o *se* deve ser classificado como partícula apassivadora, quando acompanhado de verbo transitivo direto, podendo ter sujeito definido simples, que deve concordar com o verbo que se encontra na voz passiva sintética, ou índice de indeterminação do sujeito, quando acompanhado de verbos intransitivos, transitivos indiretos ou de ligação, que devem ser empregados na terceira pessoa do singular.

Por se tratar de uma análise puramente formal, não possui explicações funcionais que levem em consideração questões semânticas. Construções como *Vende/compra-se casas* são consideradas incorretas gramaticalmente, segundo livros didáticos e gramáticas tradicionais. Estudiosos como Nunes (1991), Monteiro (1994), Bagno (2000) e Camacho (2002, 2003) constataram que elas ocorrem com frequência tanto em linguagem popular como culta.

Em outras línguas, espanhol, italiano e francês, há muitas discussões sobre impessoalidade e passividade. No espanhol, Suñer (2002, p. 211) discute que as construções com *se* impessoal acarretam a interpretação de um predicado como se aplicando a um conjunto não específico de seres humanos, representado pelo *se*. No italiano, Cinque (1988) analisa o papel do *si* impessoal, propondo variantes ligadas a seu uso como um sujeito genérico. Ruwet (1972) estudou a língua francesa, classificando alguns tipos de orações sem Agente como neutras.

Essa compreensão não pode ser baseada na análise do verbo principal, se é transitivo direto (partícula apassivadora) ou intransitivo, transitivo indireto ou de ligação (índice de indeterminação do sujeito). É preciso se atentar para as escolhas léxico-gramaticais e, principalmente para o contexto em que ocorrem, para assim compreender o significado do *se*

que pode permitir diferentes desfocamentos de participante – do autor/pesquisador ou de pesquisadores da área ou, ainda, de pessoas de modo geral. Os estudos citados, bem como outros, são discutidos em Morais (2013a e 2013b) que propõe uma nova reclassificação levando em conta os contextos das ocorrências.

5. Considerações Finais

Os dados analisados permitem dizer que as construções verbais, mais especificamente, as escolhas dos dizentes dos processos *explicar, discutir e afirmar* são utilizados para expressar conhecimento de pesquisas anteriores ou de pressupostos da área de pesquisa, dando maior credibilidade ao estudo.

A escolha dos dizentes também é acompanhada do uso da modalidade, através do uso de adjuntos modais, indicando possibilidade, atenuando afirmações no discurso. Muitas das ocorrências de 1ª pessoa do plural indicam o uso da modalidade, como discutido na seção de análise deste artigo, com ocorrências do adjunto modal de baixo grau *pode.*

O uso de construções com o clítico *se* favorece o desfocamento do autor, porém as circunstâncias e, principalmente, o contexto de ocorrência permitem pressupor o envolvimento do autor do artigo, revelando resquícios de sua participação.

Acredita-se que ficou demonstrado, neste artigo, que a função primária das construções chamadas indeterminadas e passivas não é a promoção do objeto a sujeito, mas sim a desfocalização do Agente (participante, em termos sistêmico-funcionais) que permite diferentes graus de desfocamento, conforme proposta de Morais (2013a). Como já propunha Said Ali (1908), essas construções com *se* são formas destinadas a calar o Agente.

É importante dizer que os dados também apontam uma tendência de mudança em que é permitido ao autor se representar no texto, permitindo a construção da identidade social do autor no discurso. Como estudo futuro, pretende-se analisar se esta tendência também ocorre em outras áreas da ciência ou se está presente apenas na área de humanas e/ou Linguística.

Referências

ARANHA, S. **A argumentação nas introduções de trabalhos científicos na área de Química**. Dissertação de Mestrado. PUCSP, 1996.

ARANHA, S. A otimização da escrita acadêmica através da conscientização textual. **Estudos Linguísticos** (São Paulo), São Paulo. v. 1, 2002.

ARANHA, S. A importância do domínio da língua inglesa e da linguagem acadêmica para a leitura e escrita de textos científicos. **Interciência** (Catanduva), Catanduva - SP, v. 1, p. 77-84, 2004.

ARANHA, S. A busca de modelos retóricos mais apropriados para o ensino da escrita acadêmica. **Revista do GEL** (Araraquara), v. 4, p. 97-114, 2007.

ATKINSON, D. The Philosophical Transactions of the Royal Society of London, 1675-1975: A sociohistorical discourse analysis. **Language in Society**, v.25, pp. 333-371, 1996. crossref http://dx.doi.org/10.1017/S0047404500019205

BAGNO, M. A "subversão herética" do ensino de língua. In: Bagno, M. **Dramática da Língua Portuguesa**: tradição gramatical, mídia e exclusão social. São Paulo: Loyola, 2000.

BAZERMAN, C. Modern evolution of the experimental report in Physics: spectroscopy articles in Physical Review, 1893-1980. **Social Studies in Science**, 14:163-196, 1984. crossref http://dx.doi.org/10.1177/030631284014002001

BERBER SARDINHA, T. Semantic prosodies in English and Portuguese: A contrastive study. **Cuadernos de Filología Inglesa**. V.9, 1:93-100. Murcia, Spain. 2000.

BERBER SARDINHA, T. **Linguística de Corpus**. Barueri-SP: Manole, 2004.

BERNSTEIN, B. **Class, code and control**. v.1. Londres: Routledge e Kegan Paul, 1971. crossref http://dx.doi.org/10.4324/9780203014035

BHATIA, V. K. **Analysing genre**: language use in professional settings. Longman, 1993.

BIBER, D. et al. corpus **Linguistics**: investigating language structure and use. Cambridge: Cambridge University Press, 1998. crossref http://dx.doi.org/10.1017/CBO9780511804489

CAMACHO, R. G. Construções de voz. In: Abarirre, M. B. e Rodrigues, S. C. A. (org). **Gramática do português falado**. v. 8, pp. 227-316. Campinas: Editora Unicamp, 2002.

CAMACHO, R. G. Em defesa da categoria de voz média no Português. **D.E.L.T.A.**, v. 19.1, pp. 91-122, 2003.

CINQUE, G. On si constructions and the theory of Arb. **Linguistics inquiry**. V. 19, 4.pp. 521-581, 1988.

FAIRCLOGH, N. **Discurso e mudança social**. Brasília: Editora Universidade de Brasília, 2001.

HALLIDAY, M.A.K. **An Introduction to Functional Grammar**. London: Edward Arnold, 1985.

HALLIDAY, M. A. K. **An introduction to Functional Grammar**. London: Edward Arnold, 1994.

HALLIDAY, M. A. K. **The language of science**. New York: Continuum, 2004.

HALLIDAY, M. A. K.; MATTHIESSEN, C. **An introduction to Functional Grammar**. London: Edward Arnold, 2004.

MONTEIRO, J. L. A questão do se. In: Monteiro, J. L. **Pronomes pessoais**: subsídios para uma gramática do português do Brasil. Fortaleza: EUFC, 1994.

MORAIS, F. B. C. **Entre alhos e bugalhos**: os diferentes usos do clítico SE na escrita acadêmica. Doutorado em Linguística Aplicada e Estudos da Linguagem. PUC-SP, 2013a.

MORAIS, F. B. C. As construções médias nos artigos científicos de linguística. **Cadernos de Linguagem e Sociedade**, v.14, n.2, 2013b.

MOTTA-ROTH, D. **Rethorical features and disciplinary cultures. A genre based study of academic book reviews in linguistics, chemistry and economics**. Tese de Doutorado. UFSC, 1995.

MOTTA-ROTH, D. Escrevendo no contexto: contribuições da Linguística Sistêmico-Funcional para o ensino de redação acadêmica. **Paper presented at the 33rd International Systemic Functional Congress**. PUC-SP, 2006.

NUNES, J. Se apassivador e se indeterminador: o percurso diacrônico no português brasileiro. **Caderno de Estudos Linguísticos**, V. 20, pp. 33-59, 1991.

RAJAGOPALAN, K. **Por uma linguística crítica – linguagem, identidade e a questão ética**. São Paulo: Parábola Editorial, 2003.

RUWET, N. **Les constructions pronominales neutres et moyennes. Théorie syntaxique et syntaxe du français**. Paris: Seuil, 1972.

SAID ALI, M. **Gramática histórica da língua portuguesa**. São Paulo: Melhoramentos, 1908.

SCOTT, M. R. **Wordsmith Tools v. 5**. Software for text analysis. Oxford: Oxford University Press, 2008.

SHIBATANI, M. Passives and related constructions: a prototype analysis. **Language** 61:4. pp. 821-848, 1985.

SUÑER, M. Las passives con se impessoal y la legitimación de las categorias vacías. In: Lopes, C. S. **Las construciones com *se***. Madri: Visor libros, 2002.

SWALES, J. M. Language and scientific communication: The case of the reprint request. **Scientometrics**, v.13, pp. 93-101, 1989. crossref http://dx.doi.org/10.1007/BF02017177

SWALES, J. M. **Genre analysis – English in academic and research settings**. Cambridge University Press, 1990.

SWALES, J. M. e FEAK, C. B. **Academic writing for graduate students.** Michigan: The University of Michigan Press, 1999.

THOMPSON, G. **Introducing functional grammar.** London: Arnold, 1996.

Artigo recebido em: 15.09.2014
Artigo aprovado em: 27.11.2014

As construções com *SE* na produção escrita de brasileiros aprendizes de espanhol como língua estrangeira: um estudo baseado em *corpus*[1]

Constructions with *SE* in the written production of Brazilian learners of Spanish as a foreign language: a corpus-based study

Benivaldo José de Araújo Júnior*

RESUMO: Este trabalho tem por objetivo analisar a produção de construções com o clítico *SE* (doravante construções-*SE*) em língua espanhola por aprendizes brasileiros de Espanhol como língua estrangeira (ELE). Nessa análise, cujo referencial teórico é a Gramática Cognitiva, serão tratadas em especial as construções reflexivas, médias, impessoais e passivas. Após breve introdução teórica, apresentamos nosso *corpus* de estudo, exibimos e discutimos os resultados do levantamento das construções-*SE* nesse *corpus*. Na discussão, comparamos os dados que obtivemos com aqueles observados em outros dois *corpora* de falantes nativos, um para o Espanhol na variedade peninsular e outro para o Português Brasileiro (doravante PB). Finalmente, com base nessa comparação, fazemos algumas considerações sobre a produção das construções-*SE* em nosso *corpus* de estudo e os fatores que possivelmente têm influência nesse processo.

PALAVRAS-CHAVE: Construções-*SE*. ELE. Gramática Cognitiva. Estudos comparados. Linguística de *Corpus*.

ABSTRACT: This work aims to analyze the production of SE-constructions in Spanish by Brazilian learners of Spanish as a Foreign Language (SFL). In this analysis, which uses Cognitive Grammar as theoretical framework, reflexive, middle, impersonal and passive constructions will be specially examined. After a brief theoretical introduction, we present our corpus of study; then, we demonstrate and discuss the results of SE-constructions survey in this corpus. In the discussion, our data are compared with those observed in two native speakers' corpora, one for Iberian Spanish and other for Brazilian Portuguese (hereinafter PB). Finally, based on that comparison, we make some considerations on the production of SE-constructions in our corpus of study and the factors that possibly influence this process.

KEYWORDS: SE-constructions. SFL. Cognitive Grammar. Comparative studies. Corpus Linguistics.

1. O modelo de conceitualização adotado

Na classificação e interpretação dos nossos dados, adotaremos o modelo cognitivo do evento canônico (LANGACKER, 1991, p. 285). Este, por sua vez, é a combinação de dois outros: o modelo da bola de bilhar e o modelo do palco. De acordo com o primeiro (bola de bilhar), tendemos a conceitualizar os eventos do mundo como uma cadeia de ação, na qual um

*Doutor em Língua Espanhola pela FFLCH-USP e professor da Escola Superior de Propaganda e Marketing (ESPM-SP).

[1] Este trabalho é uma versão ampliada e modificada de outro apresentado no II Congresso Internacional de Professores de Línguas Oficiais do MERCOSUL (Buenos Aires, 2013), intitulado "As construções com *SE* na produção escrita de brasileiros aprendizes de espanhol como língua estrangeira".

elemento energético transfere energia e provoca efeitos em outro elemento tipicamente não energético. Quanto ao segundo (palco), relaciona-se à experiência perceptual e foi assim chamado porque nele o papel do conceitualizador é similar ao de um espectador assistindo a uma peça; ou seja, alguém que se encontra fora de cena, mas que constitui parte do evento global. No modelo do evento canônico (fig. 1), segundo a terminologia de Langacker (1991, p. 283), o iniciador da cadeia de ação (fonte de energia) chama-se trajetor e é também o sujeito/agente prototípico; o participante final (ralo por onde escoa a energia) chama-se marco e é também o objeto/paciente prototípico; o conceitualizador/observador está indicado na figura pela letra O. Conforme o modelo, no enunciado *Ernesto rompió la estatuilla* [PB: Ernesto quebrou a estatueta], *Ernesto* é o iniciador (fonte)/trajetor/sujeito/agente e *estatuilla* é o participante afetado (ralo)/marco/objeto/paciente.

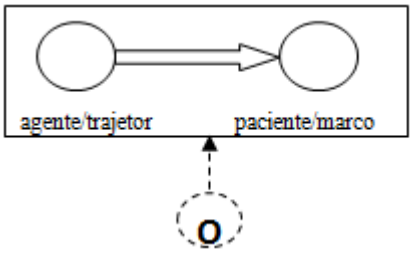


Fig. 1 – Modelo de evento canônico (LANGACKER, 1991, p. 285)

O modelo de evento canônico representa a observação usual de uma ação prototípica nas línguas do sistema nominativo-acusativo, como o PB e o espanhol, sendo, portanto, o modelo mais produtivo para nossa análise. Portanto, no modelo que adotamos, um evento canônico é transitivo por definição e nele temos dois participantes claramente distintos numa relação assimétrica, no qual o participante agente é animado e volitivo e afeta direta e totalmente o participante paciente por meio de uma força ou transferência de energia (GIVÓN, 1984, apud KEMMER, 1994).

2. As construções reflexivas e médias segundo Kemmer (1994)

Conforme já apontado por Hopper & Thompson (1980, p. 277), nos eventos em cuja codificação morfossintática aparece o pronome *SE*/SE[2], a assimetria entre os participantes é problemática, acarretando uma baixa na transitividade. Por exemplo, na sentença *Clara se vio*

[2] A notação *SE* (itálico) refere-se à ocorrência de construções com clítico em espanhol ou na produção não nativa nessa língua. A notação SE (normal) se aplica à incidência de construções com clítico no PB.

en el espejo (PB: *Clara se viu no espelho*), os papéis de trajetor/agente e marco/paciente são preenchidos pela mesma entidade referencial, de modo que a distinguibilidade entre esses dois participantes esmaece. Contudo, para Kemmer (1994, p. 207), é possível operar uma separação conceitual entre *Clara* como iniciadora da atividade (o participante que vê) e seu reflexo no espelho (o que é visto), de modo a poder distinguir dois participantes no evento. O mesmo ocorre com o enunciado *La presidenta se imaginó descansando en una playa desierta* [PB: *A presidenta se imaginou descansando numa praia deserta*]. Neste caso, o sujeito interage com uma representação de si mesmo situada em algum espaço mental distinto do discurso (MALDONADO, 2006, p. 270). A esse tipo de eventos, Kemmer (1994, p. 207) classifica como **reflexivos**. Ainda segundo a autora (1994), nos eventos reflexivos o iniciador (trajetor) atua sobre si mesmo como o faria sobre qualquer outra entidade.

Há ainda eventos nos quais a distinção conceitual entre os participantes é mínima. No exemplo *Felipe solo se afeita los domingos* [PB: *Felipe só se barbeia/faz a barba aos domingos*], o paciente em questão não é unicamente afetado: barbear o próprio rosto implica a atividade da parte diretamente envolvida (a cabeça), além de outras partes do corpo (o pescoço). Em eventos dessa natureza, ambos os participantes têm como referente uma única entidade cujos aspectos são conceitualmente indistinguíveis. Kemmer (1994) classifica esses eventos como **médios**, fundamentando-se no conceito de **voz média** (originalmente empregado para designar uma categoria inflexional das línguas clássicas indoeuropeias), cuja função, em termos semânticos, é "expressar eventos nos quais a ação ou estado afeta o sujeito do verbo ou seus interesses" (LYONS, 1970, p. 286). Após minucioso estudo translinguístico, Kemmer enumera dez situações que configuram um domínio semântico no qual a voz média apresenta algum tipo de marcação morfológica. Por motivos de concisão, daremos apenas exemplos do espanhol e do PB, nos quais o *SE*/SE funciona como marcador médio: (1) cuidados *corpora*is (*peinarse*; *frotarse* [PB: *pentear-se*; *esfregar-se*]); (2) movimento não translacional (*inclinarse* [PB: *inclinar-se*]); (3) mudança na postura *corpora*l (*levantarse*; *sentarse* [PB: *levantar-se*; *sentar-se*]); (4) movimento translacional (*irse* [PB: *ir-se*]); (5) eventos naturalmente recíprocos (*abrazarse*; *besarse* [PB: *abraçar-se*; *beijar-se*]); (6) média indireta (*preguntarse* [PB: *perguntar-se*], referente a uma atividade mental); (7) média de emoção (*enfadarse* [PB: *irritar-se*]); (8) discurso emotivo (*quejarse* [PB: *queixar-se*]); (9) média de cognição (*acordarse* [PB: *lembrar-se*)]; (10) eventos espontâneos (*originarse* [PB: *originar-se*]). No PB, em muitas dessas situações, coexistem as formas com realização do

clítico ou seu apagamento, como nos enunciados *Meu pai (se) levanta cedo* e *Os fieis (se) sentaram*.

Kemmer (1994) esquematiza os eventos reflexivos e médios conforme as figuras 2 e 3.

Fig. 2 – Esquema do evento reflexivo (KEMMER, 1994, p. 207)

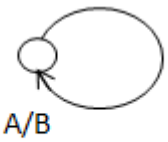


Fig. 3 – Esquema do evento médio (KEMMER, 1994, p. 207)

Em sua análise sobre transitividade, Hopper & Thompson (1980, p. 277) propõem as construções reflexivas como intermediárias entre as transitivas e as intransitivas. Kemmer (1994, p. 209) amplia essa proposta, incluindo no conjunto as construções médias e utilizando o grau de distinguibilidade dos participantes como parâmetro semântico para diferenciar uma construção da outra. Dessa forma, a autora constrói um diagrama no qual num dos extremos estão os eventos intransitivos (1 participante), e no outro os transitivos (2 participantes). Entre esses dois extremos, estão os eventos nos quais a distinção entre os participantes é baixa, caso dos reflexivos e médios.

Fig. 4 – Grau de distinguibilidade dos participantes (KEMMER, 1993, p. 73)

Segundo o diagrama, nas construções reflexivas/ recíprocas, pelo fato de conceitualmente podermos distinguir dois participantes (embora se refiram à mesma entidade), temos maior proximidade a um evento transitivo. Já nas construções médias, a baixa distinguibilidade entre os participantes aproxima-as das intransitivas. Kemmer (1994) relaciona a maior ou menor distinguibilidade dos participantes com uma propriedade semântica a qual denomina **elaboração relativa de eventos**.

A marcação morfológica das reflexivas/recíprocas e médias no PB e no espanhol, quando ocorre, normalmente se dá com o mesmo marcador SE/*SE*, conforme já mostrado em

exemplos anteriores. Nas referidas línguas, portanto, a distinção entre essas construções não é imediata e requer uma análise acurada do verbo em termos semânticos. No caso do PB, salientamos a existência de um fator que complica ainda mais essa classificação: a perda dos clíticos (inclusive o SE) que vem acontecendo nessa língua[3].

3. As construções passivas e impessoais segundo Maldonado (2006)

Para Maldonado (2006, p. 273), que se concentra fundamentalmente na análise dessa questão na língua espanhola, a presença do *SE* nas construções médias, passivas e impessoais se relaciona com a força indutiva envolvida no evento.

Com base nessa hipótese, o autor propõe a seguinte escala de classificação (figura 5): num dos extremos da escala estariam as construções transitivas, com máxima proeminência de transferência de energia; no outro extremo, estariam as construções absolutas, sem energia envolvida; na zona intermediária, as médias, passivas e impessoais[4].

Nessa classificação, além dos graus de manifestação da força indutiva, são levados em conta o requisito de agentividade humana do verbo, o aspecto (léxico e morfológico), a concordância e a ordem. Com base nesses critérios, Maldonado (2006) classifica como médias algumas construções consideradas passivas por muitos gramáticos e linguistas, como é o caso de D; para esse autor, tal passividade é aparente, uma vez que resulta mais da perda de proeminência da força indutora que da escolha do tema como figura do evento.

[3] Esse fenômeno foi analisado em diversos estudos. González (1994) leva em conta a perda dos clíticos no PB, ao falar das assimetrias entre o espanhol e o PB na descrição do preenchimento vs. não preenchimento das posições argumentais de sujeito e objeto nessas línguas. Não trataremos dos casos de perda, visto que o foco do nosso trabalho são as construções nas quais aparece o *SE*; porém, colocaremos o clítico entre parênteses toda vez que a construção puder aparecer sem ele.

[4] As impessoais intransitivas (*Se vive bien aquí* [PB: *Vive-se bem aqui*]) não apresentam ambiguidades quanto à classificação no E e no PB; portanto, neste trabalho, receberão mais atenção as impessoais com objeto direto (*Se alquila(n) casas* [PB: *Aluga(m)-se casas*]).

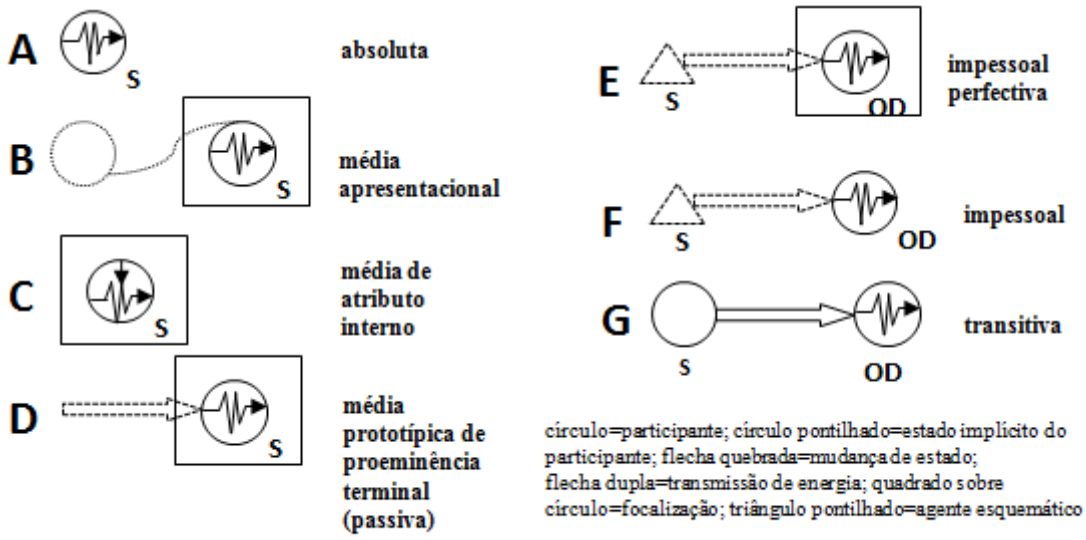


Fig. 5 – Níveis de proeminência da força indutiva (MALDONADO, 2006, p. 273)

A seguir, definimos sucintamente as construções da figura 5:

(a) **Absoluta:** sem energia em perfil. Ex.: *El ruído disminuyó* [PB: *O ruído diminuiu*].

(b) **Média apresentacional:** nível quase nulo de energia; simplesmente mostra um evento num domínio qualquer a partir do ponto de vista de um conceitualizador. Ex.: *Las Islas Fiji se encuentran en el Pacífico* [PB: *As Ilhas Fiji se encontram no Pacífico*].

(c) **Média de atributo interno:** a mudança de estado sofrida pelo tema depende mais de sua configuração interna que de uma força externa. Ex.: *Estos juegos no se venden bien* [PB: *Estes jogos não (se) vendem bem*].

(d) **Média prototípica de proeminência terminal (Passiva):** focaliza a porção terminal do evento (a mudança de estado devido a uma força externa esquemática) e não sua fase indutiva. Em nossa análise, utilizaremos a forma abreviada MPT para referir-nos a essa construção; igualmente, manteremos entre parênteses a denominação **passiva**, por ser a mais frequente na literatura. Ex.: *La pared se manchó* [PB: *A parede (se) manchou*].

(e) **Impessoal perfectiva:** a indução da força é tão importante quanto a mudança de estado, e tem que ser humana, embora seu agente não seja específico. Ex.: *Se pagó la deuda* [PB: *Pagou-se a dívida*/ *A dívida foi paga*[5]].

[5] Conforme mostrado em Araújo Júnior (2006), os enunciados com SE ocorrem no PB, porém a preferência dos falantes nativos dessa língua é pela passiva perifrástica.

(f) **Impessoal:** os marcadores imperfectivos reduzem a proeminência da mudança de estado e favorecem o que de homogêneo possa haver no evento; portanto, o que se perfila são as tendências naturais de mudança sofridas pelo tema. Ex.: *Estos productos se fabrican con materiales sintéticos* [PB: *Estes produtos se fabricam/ são fabricados com materiais sintéticos*].

(g) **Transitiva:** prototipicamente, são as construções nas quais um agente animado e volitivo transfere energia a um paciente, provocando neste uma mudança de estado. Ex.: *Miguel pintó la casa* [PB: *Miguel pintou a casa*].

4. Os *corpora*

Nosso *corpus* de estudo é o Dados EEC (doravante DEEC), que possui 178.066 palavras e está constituído por 1.172 produções escritas de alunos do curso *Español en el Campus*[6]. O DEEC foi construído para integrar um *corpus* maior, o COMET (*Corpus* Multilíngue para Ensino e Tradução), desenvolvido pelo CITRAT[7] com o objetivo de servir de suporte a pesquisas linguísticas, principalmente nas áreas de tradução, terminologia e ensino de línguas. O objetivo específico do DEEC era coletar e organizar dados longitudinais, que possibilitassem acompanhar um grupo de informantes do primeiro ao último estágio do EEC. A coleta de produções ocorreu entre os meses de agosto de 2003 e junho de 2006, e abarcou os níveis Básico, Intermediário e Avançado. Na tabela a seguir, mostramos uma síntese do perfil dos informantes do DEEC.

[6] Curso de ELE oferecido pela Área de Língua Espanhola e Literaturas Espanhola e Hispano-americana – DLM e mantido pelo Serviço de Cultura e Extensão da FFLCH/USP entre 1996 e 2010. A grade completa compreendia os níveis Básico I (B1), Básico II (B2), Intermediário I (I1), Intermediário II (I2), Avançado I (A1) e Avançado II (A2).

[7] Centro Interdepartamental de Tradução e Terminologia da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.

Tabela 1 – DEEC: Perfil dos informantes.

		QUANTIDADE	%
INFORMANTES POR SEXO	Masculino	50	27
	Feminino	136	73
	TOTAL	186	100
INFORMANTES POR FAIXA ETÁRIA	Até 20 anos	22	11,8
	De 21 a 34 anos	98	52,7
	De 35 a 49 anos	45	24,2
	De 50 a 69 anos	21	11,3
	TOTAL	186	100

Quanto à escolaridade, 94,6% do total de informantes chegou à universidade. No geral, esses colaboradores pertenciam à comunidade USP: eram estudantes (graduandos ou pós-graduandos), professores ou funcionários da instituição. Do total, apenas 4% afirmavam ter tido algum tipo de contato com a língua espanhola antes de ingressar no EEC.

As produções do *corpus* pertencem a diversos gêneros discursivos[8], tais como anúncios pessoais, cartas formais e informais, diários, resenhas, crônicas, artigos de opinião, notícias, sinopses, críticas de filmes, etc. Na tabela a seguir, disponibilizamos a quantidade de produções do DEEC segundo o nível dos informantes.

Tabela 2 – DEEC: Produções por nível.

NÍVEL	QUANTIDADE	%	Nº DE PALAVRAS
B1	545	46,5	62.376
B2	251	21,4	47.627
I1	147	12,5	22.514
I2	121	10,3	20.268
A1	60	5,1	14.420
A2	48	4,1	10.861
TOTAL	1.172	100	178.066

[8] Convém esclarecer que os gêneros presentes no *corpus* não são autênticos, mas sim projeções: não se trata, por exemplo, de jornalistas redatando crônicas, críticas ou notícias para um público leitor; o que se tem é o exercício de escrita em língua estrangeira (no caso, em espanhol) feito por alunos, mediante instrução formal e tendo como único destinatário o professor. Portanto, embora se tenha buscado a diversidade nas práticas textuais pedidas aos alunos, buscando aproximá-las dos gêneros correntes no cotidiano, essas produções são inevitavelmente atravessadas pela prática pedagógica.

Os resultados da análise das construções-*SE* no DEEC serão cotejados com aqueles observados em outros dois *corpora* de falantes nativos, um para o espanhol e o outro para o PB. Ambos são *corpora* já constituídos e na modalidade oral. No caso do espanhol, trata-se do *PRESEEA-ALCALÁ*, que por sua vez é parte integrante do *PRESEEA* (*Proyecto para el Estudio Sociolingüístico del Español de España y de América*)[9]. As 36 entrevistas que compõem esse *corpus* foram realizadas em 1998, na cidade de Alcalá de Henares[10]. Na tabela 3 aparecem informações detalhadas sobre o *corpus PRESEEA-ALCALÁ*, ao qual nos referiremos doravante como CPA.

Tabela 3 – *Corpus* PRESEEA-ALCALÁ: Resumo das principais características.

***Corpus* PRESEEA-ALCALÁ (CPA)**			
Faixa etária	**Sexo**	**Nível de escolaridade**	
		Superior (Nº de	Médio (Nº de
Faixa 1 20-34 anos	masculino	24.730	24.257
	feminino	24.442	26.471
	Total da faixa1	**49.172**	**50.728**
Faixa 2 35-54 anos	masculino	30.932	26.311
	feminino	22.882	25.443
	Total da faixa 2	**53.814**	**51.754**
Faixa 3 Acima de 55 anos	masculino	29.695	29.860
	feminino	33.940	30.398
	Total da faixa 3	**63.635**	**60.258**
Total por nível de escolaridade		**166.621**	**162.740**
Total geral		**329.361**	

Para o PB, utilizamos a Amostra SP2010 (Piloto), que é parte do *corpus* que está sendo desenvolvido pelo GESOL-USP (Grupo de Estudos e Pesquisa em Sociolinguística da USP) com o objetivo de subsidiar pesquisas na área de sociolinguística variacionista — em especial, na descrição e análise dos fenômenos variáveis e processos de mudança em curso na variedade paulistana do PB (MENDES, 2011). As 36 entrevistas que constituem esse *corpus*

[9] Projeto surgido em 1993 e que conta com equipes de pesquisa e documentação na Espanha, nos Estados Unidos e em alguns países americanos (Argentina, Colômbia, Cuba, Chile, Equador, Guatemala, México, Paraguai, Peru, Porto Rico, Uruguai e Venezuela). Fonte: http://linguas.net/portalpreseea/Inicio/tabid/441/language/es-ES/Default.aspx , consultado em 20/05/13.

[10] A informação completa acerca da metodologia de coleta desses materiais está em Moreno Fernández et alii (2002, 2004).

foram realizadas na cidade de São Paulo, entre 2008 e 2011. A tabela 4 mostra em detalhe as principais características da Amostra SP2010, à qual nos referiremos doravante como ASP[11].

Tabela 4 – Amostra SP2010 (Piloto) (GESOL-USP): Resumo das principais características.

Amostra SP2010 (Piloto)(GESOL-USP)			
Faixa etária	**Sexo**	**Nível de escolaridade**	
		Superior (Nº de	Médio (Nº de
Faixa 1 20-34 anos	masculino	28.145	30.414
	feminino	35.411	37.170
	Total da faixa1	**63.556**	**67.584**
Faixa 2 35-49 anos	masculino	36.330	33.977
	feminino	32.627	29.416
	Total da faixa 2	**68.957**	**63.393**
Faixa 3 50-69 anos	masculino	31.884	32.730
	feminino	27.737	33.357
	Total da faixa 3	**59.621**	**66.087**
Total por nível de escolaridade		**192.134**	**197.064**
Total geral		**389.198**	

Escolhemos o CPA e a ASP para este estudo, primeiramente, porque foram constituídos a partir dos mesmos critérios metodológicos. Em segundo lugar, ambos os *corpora* são recentes e, portanto, refletem o estado atual do espanhol e do PB com maior probabilidade de in*corpora*r fenômenos emergentes nessas línguas. Por último, os dois *corpora* apresentam pouca disparidade no quesito extensão: a ASP (389.198 palavras) supera o CPA (329.361 palavras) em apenas 18%; ambos são *corpora* médios, conforme a classificação proposta por Berber Sardinha (2004, p. 26)[12].

Embora as tabelas 3 e 4 ofereçam dados acerca da faixa etária, sexo e escolaridade dos informantes, tais variáveis não serão levadas em conta em nossa análise. O que aqui faremos é uma primeira aproximação a alguns aspectos referentes à produção de construções-*SE* em língua espanhola por aprendizes brasileiros de ELE; neste nível, portanto, nos limitaremos a analisar a incidência global e os percentuais dessas construções no *corpus* de estudo e compará-los com os resultados obtidos no CPA e na ASP.

[11] As tabelas 3 e 4 são uma versão simplificada daquelas presentes em Araújo Júnior (2013, p. 131-2).
[12] Segundo esse autor, estão nesse grupo os *corpora* que possuem entre 250.000 e 1 milhão de palavras.

5. Metodologia e análise do *corpus* de estudo

O levantamento das construções-*SE* no *corpus* DEEC foi feito em três etapas, conforme o procedimento que se descreve a seguir. A etapa inicial constou de três passos, sendo o primeiro dividir o *corpus* em seis *subcorpora* — B1, B2, I1, I2, A1, A2 —, considerando os níveis da grade (ver tabela 2). A seguir, processamos cada sub*corpus* no programa Kitconc[13], obtendo, assim, as concordâncias por meio da lematização **SE*. O Kitconc está escrito totalmente em português e reúne algumas funcionalidades usadas em Linguística de *Corpus*, tais como a listagem de palavras, frequências, concordâncias, colocados e a extração de palavras-chave; no caso, usamos o programa apenas como concordanciador. A lematização **SE* nos permitiu acessar todas as combinações do clítico no *corpus*, tanto as proclíticas (*SE levanta*, *SE hacía*, *SE ha comprado*, etc.) quanto as enclíticas (*marcharSE*, *verSE*, *realizándoSE*, etc.). O terceiro passo foi transferir as concordâncias para uma planilha Excel, a fim de facilitar sua leitura e triagem.

A segunda etapa, executada manualmente, consistiu em ler as concordâncias e eliminar aquelas sem interesse para o trabalho. Portanto, foram desprezadas as ocorrências do *SE* como variante de *LE*/*LES* (objeto direto em construções do tipo *SE la dio*, *SE lo entregó*, etc.) e outros casos (*faSE*, *eSE*, *informaSE*, etc.).

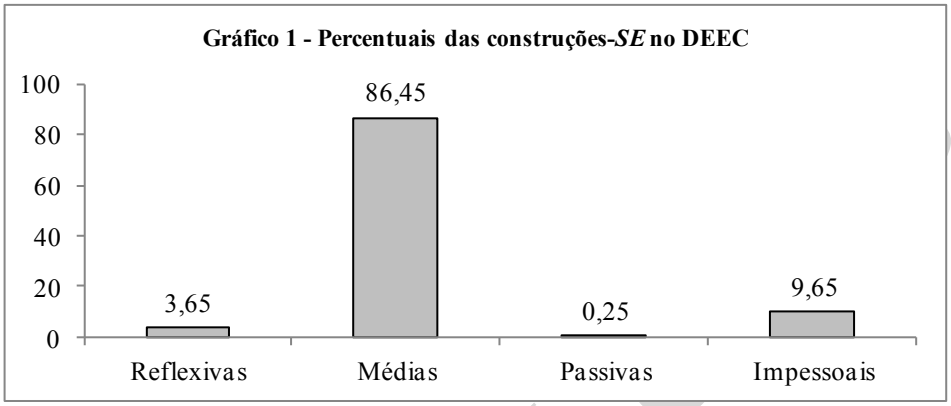
A etapa final consistiu em classificar as incidências do *SE* nas concordâncias remanescentes. Embasando-nos em Kemmer (1994) e Maldonado (2006), reunimos essas ocorrências em reflexivas, médias, passivas (MPT) e impessoais. O resultado final está na tabela a seguir:

Tab. 5 – DEEC: Totais por construção.

Sub*corpus*	Reflexivas	Médias	Passivas (MPT)	Impessoais	TOTAL
B1	6	644	0	21	671
B2	39	371	0	20	430
I1	1	158	0	21	180
I2	6	95	0	49	150
A1	5	68	1	26	100
A2	2	61	3	19	85
TOTAL	**59**	**1.397**	**5**	**155**	**1.616**
%	**3,65**	**86,45**	**0,31**	**9,59**	**100**

[13] Desenvolvido por José Lopes Moreira Filho e disponível como *freeware* na Internet. Neste trabalho, utilizamos a versão 3.0 do programa.

Observando os resultados de modo global, é notória a supremacia das médias (86,45%) com relação às demais construções-*SE* no *corpus*. Em segundo lugar vêm as impessoais (9,65%), seguidas das reflexivas (3,65%) e, finalmente as passivas (0,25%). No gráfico a seguir, esses percentuais podem ser melhor visualizados:



A alta incidência das médias pode ser explicada em função dos critérios de classificação adotados. Ou seja, delimitamos as construções reflexivas e médias a partir da proposta de Kemmer (1994), fundamentada na elaboração relativa de eventos e no grau de distinguibilidade de seus participantes (ou dos subeventos componentes); dessa forma, em nossa análise, classificamos como médias algumas construções que normalmente são consideradas reflexivas (sobretudo nas abordagens prescritivas) ou recebem outra denominação. A seguir, citamos exemplos do *corpus*[14] para algumas das situações mediais listadas por Kemmer (1994):

(1)*Hay que dejarlos se ducharen todos los días* (ED400494) [cuidados *corpora*is]
(2)*Entonces se levantó en un rato, paró un poco porque estaba fatigado* (EC300013) [mudança na postura *corpora*l]
(3)*A los 23 años, él se fué a vivir en Salvador* (EC300353) [movimento translacional]
(4)*En ese momento, ella se alarmó: "Pero, ¿estaré más tiempo sola, sin un novio?"* (EC300071) [média de emoção]

As reflexivas totalizaram pouco menos de 4% do total. Considerando-se apenas o conjunto dessas construções no *corpus*, houve incidência de reflexivas recíprocas (maioria,

[14] A informação entre parênteses, após cada exemplo, representa o código da produção escrita no DEEC.

com 83%), reflexivas diretas (15%) e reflexivas indiretas (2%), representadas, respectivamente, nos exemplos (5), (6) e (7).

(5)*Se quieren y se respetan* (ED101572)
(6)*A ella le gusta olharse en el espejo* (EA100384)
(7)*(...) se están preparando unos pasteles para comérselos* (EC300074)

No levantamento das impessoais e passivas, utilizamos a proposta de Maldonado (2006), embasada na força indutiva presente nos eventos. Embora fatores como a concordância e o aspecto sejam importantes na diferenciação dessas construções, o requisito de agentividade humana é preponderante: se a indução no evento é humana, embora esquemática, há maior probabilidade de termos uma construção impessoal (esquemas E e F na fig. 5); se a força indutiva tem menos proeminência no evento, a saliência passa a ser do tema e estamos no âmbito das médias e passivas (esquemas B, C e D na fig. 5).

As impessoais totalizaram cerca de 10% do total de incidências. Levando-se em conta apenas o conjunto dessas construções, obteve-se no *corpus* a seguinte distribuição: as impessoais intransitivas correspondem a 22% (exemplo 8); as impessoais com complemento direto, que foram maioria, totalizam 78%. Destas, 75% são imperfectivas (exemplo 9) e 3% são perfectivas (exemplo 10), segundo o aspecto verbal.

(8)*(...) se produce más y mejor en casa (...)* (ED400271)
(9)(...) *sólo se requiere buena capacidad de comunicación* (ED400171)
(10) *(...) es una cosa que nunca se ofreció* (EF600024)

As passivas constituem apenas 0,25% do total. Para essas construções, citamos os exemplos a seguir.

(11)*(...) no la compre [la mermelada] si ya se rompió el lacre* (EE500021)
(12)(...) *se extraviaron mis maletas desde hace una semana* (EF600021)

Nessas ocorrências, destacamos a pouca saliência da força indutiva, cuja fonte específica não é identificável: a ruptura do lacre ou o extravio das maletas podem não ter sido provocados por agentes humanos; os agentes — humanos ou não — são inespecíficos na codificação do evento, de modo que o elemento proeminente em cada construção é o tema (*el*

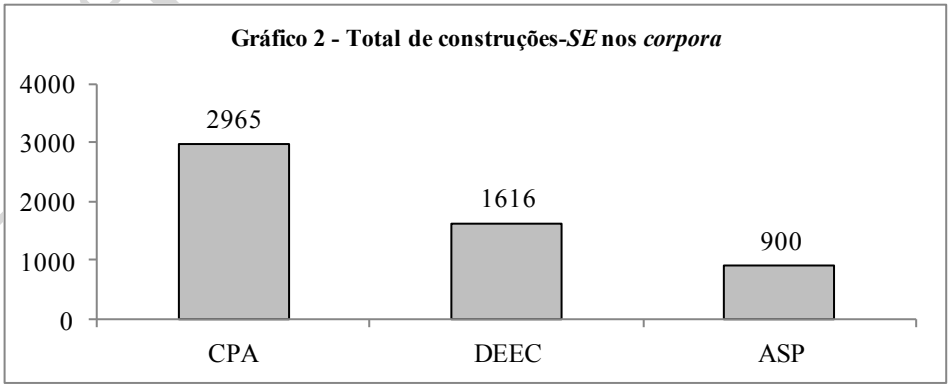
lacre, *las maletas*). Enfim, o que se focaliza nas passivas é a mudança de estado sofrida pelo tema e não o processo em si; daí o aspecto perfectivo presente nessas construções.

Cotejando os dados do DEEC com os do CPA (falantes nativos do espanhol peninsular) e da ASP (falantes nativos do PB), chegamos à seguinte tabela:

Tab. 6 – CPA/ DEEC/ ASP: Totais por construção.

CORPORA		**Reflexivas**	**Médias**	**Passivas (MPT)**	**Impessoais**	**TOTAL**
CPA	**total**	98	1.651	12	1.204	2.965
	%	3,31	55,68	0,4	40,61	100
DEEC	**total**	59	1.397	5	155	1.616
	%	3,65	86,45	0,31	9,59	100
ASP	**total**	96	626	2	176	900
	%	10,67	69,56	0,22	19,56	100

Comparando-se o total das construções-*SE* nos 3 *corpora*, o maior número de incidências ocorreu no CPA: o total nesse *corpus* supera em mais de 3 vezes o observado na ASP e, em quase o dobro, o total levantado no DEEC (gráfico 2). Tal resultado está dentro do esperado para a ASP, conforme os estudos comparativos de Araújo Júnior (2006) e González (1994), segundo os quais a perda ou apagamento dos clíticos (SE e os demais) no PB favoreceria a menor incidência de construções-SE nessa língua e na produção não nativa de lusoparlantes em espanhol.

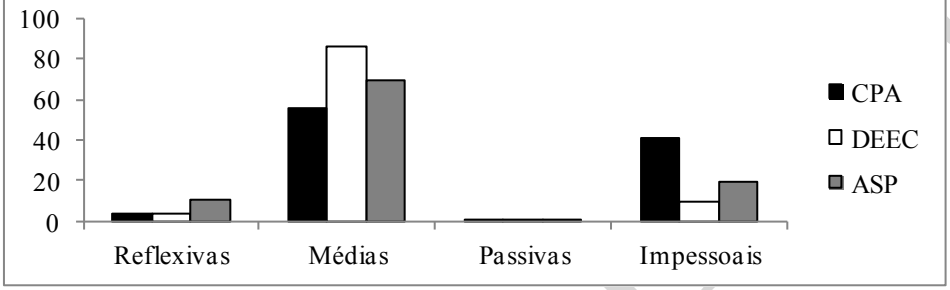


Os resultados no DEEC, tomados globalmente, estão a meio caminho entre o CPA e a ASP: se por um lado a instrução formal em LE influenciaria a presença significativa de construções-*SE* no *corpus* de aprendizes (quase o dobro da que aparece no *corpus* de falantes

nativos do PB), por outro lado, o apagamento do clítico implicaria uma frequência ainda aquém da observada no *corpus* do espanhol.

Ao comparar os percentuais por construção (gráfico 3), os resultados se mostram difusos — o CPA lidera na quantificação percentual das impessoais e passivas; a ASP tem a maior porcentagem de reflexivas e o DEEC está à frente no percentual das médias — e não sinalizam propriamente uma tendência.

Gráfico 3 - Percentuais de construções-*SE* nos *corpora* por tipo



Se a perda do clítico pode estar ocasionando os baixos percentuais das passivas no DEEC[15] (0,31%) e na ASP (0,22%), nas médias parece ocorrer o contrário: a alta incidência dessas construções no DEEC (86,45%) e na ASP (69,56%) sinaliza menor índice de apagamento do clítico na variedade em questão.

Sobre outros fenômenos observados no *corpus*, destacamos o baixo índice de construções sem concordância (0,43%). Citamos duas delas na sequência:

> *(13) Existe el sitio donde se fabrica las piezas* (ED400701)
> *(14) (...) havia formaciones minerales como aquellas que se ve en cavernas de la Tierra* (EB200302)

No caso, nossa expectativa era de que o enfraquecimento da concordância no PB (GALVES, 1993, apud GONZÁLEZ, 1994) resultasse na produção de mais construções discordantes no *corpus* de aprendizes, porém tal não ocorreu; não se pode atribuir o fato à influência da instrução formal em LE unicamente, uma vez que no próprio *corpus* do PB esse índice também foi baixo (0,8%). Outro fenômeno verificado no *corpus* de aprendizes,

[15] Em Araújo Júnior (2006), ao analisarmos o *corpus* DEEC, constatamos que a perda do clítico também favoreceria a predominância das passivas perifrásticas nesse *corpus*.

também apontado por González (1994), foi a supergeneralização dos clíticos (exemplos 15 e 16), que consiste na presença destes onde não deveriam aparecer. No caso específico do *SE*, computamos 2,1% dessas construções no *corpus*.

(15) Pero no fue eso que se sucedió (EF600744)
(16)Para el bebé, todo no se pasa de un juguete (EC300283)

Conforme González (1994, p. 411), tais formas são in*corpora*das à produção não nativa muito mais pelo efeito sonoro do que propriamente pela importância na construção do sentido e na referencialidade. Também observamos ocorrências no DEEC nas quais o *SE* não desempenha função argumental (exemplos 17 e 18), mas cuja presença pode ser creditada à influência da língua materna, uma vez que o fenômeno ocorre com relativa frequência no PB.

(17)Ésta es una pregunta dificil de se contestar (EF501114)
(18) Sólo había piedras y más piedras para se estudiar (EC201152)

Nesses casos — que totalizaram 0,8% no DEEC e 2,3% na ASP — o *SE*/SE acompanha infinitivos em construções impessoais e parece ter como finalidade reforçar o caráter humano do sujeito esquemático na estrutura.

5. Considerações finais

A análise do *corpus* de aprendizes que aqui procedemos, embora sucinta, permite-nos tecer algumas considerações.

De início, ressaltamos que a Gramática Cognitiva foi uma abordagem produtiva na análise do nosso *corpus* de estudo, uma vez que nos ofereceu critérios (especialmente semânticos) que nos auxiliaram a melhor delimitar e classificar as construções-*SE* levantadas. De acordo com os dados, a incidência global das construções-*SE* no *corpus* de aprendizes ainda está distante do observado no espanhol, possivelmente influenciada pela perda do clítico no PB. Entretanto, a análise das ocorrências por construção revelou baixo índice de passivas e alto percentual das médias no *corpus* de aprendizes; consideramos esses resultados um indício de que o apagamento do clítico no PB ocorreria de maneira diferenciada nas diferentes construções-*SE*, e que uma investigação específica desse fenômeno ofereceria mais respostas sobre a produção das referidas construções, em espanhol, por aprendizes brasileiros.

O contraponto ao apagamento do clítico seria a sua supergeneralização no *corpus* de aprendizes, cujo índice ficou em 2,1% do total e, conforme González (1994), seria produto da aprendizagem formal (por um efeito mimético) e não da aquisição espontânea.

Por fim, a presença no DEEC de construções impessoais com *SE* não argumental (não observada no *corpus* do espanhol) pode ser atribuída à influência da língua materna, nas quais são atestadas com relativa frequência. O clítico *SE* nessas construções, tal qual parece ocorrer no PB, enfatizaria o caráter humano do sujeito/agente selecionado pelo verbo, embora esse participante não esteja codificado na sentença.

Referências bibliográficas

ARAÚJO JÚNIOR, B. J. **As passivas na produção escrita de brasileiros aprendizes de Espanhol como língua estrangeira**. 2006. 111f. Dissertação (Mestrado em Língua Espanhola). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2006.

_____. **Limites precisos ou fronteiras que desaparecem?** As construções impessoais e passivas com o clítico SE/SE no português brasileiro e no espanhol. 2013. 203 f. Tese (Doutorado em Língua Espanhola). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2013.

BERBER SARDINHA, T. **Linguística de Corpus**. Barueri, Manole, 2004. 410 p.

GONZÁLEZ, N. T. M. **Cadê o pronome? — O gato comeu**. Os pronomes pessoais na aquisição/aprendizagem do espanhol por brasileiros adultos. 1994. 451f. Tese (Doutorado em Linguística). Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 1994.

HOPPER, P. J.; THOMPSON, S. Transitivity in grammar and discourse. **Language**, v. 56, n.2, p. 251-99, 1980. crossref http://dx.doi.org/10.1353/lan.1980.0017

KEMMER, S. Middle voice, Transitivity, and the Elaboration of Events. In: FOX, B.; HOPPER, P. J. (eds.) **Voice**: Form and Function. John Benjamins: Amsterdam/Philadelphia, 1994, p. 179-229. crossref http://dx.doi.org/10.1075/tsl.27.09kem

LANGACKER, R. W. **Foundations of Cognitive Grammar**: Descriptive Application. Stanford: Stanford University Press, 1991, v.2. 590 p.

LYONS, J. **Linguistique générale**. Paris: Larousse, 1970. 384 p.

MALDONADO, R. **A media voz**: Problemas conceptuales del clítico se. México: Universidad Nacional Autónoma de México, 2006. 482 p.

MENDES, R. SP2010 - Construção de uma amostra da fala paulistana. Projeto de Pesquisa apresentado à FAPESP (Proc. no. 2011/09278-6), 2011.

MORENO FERNÁNDEZ, F. et al. **La lengua hablada en Alcalá de Henares. *Corpus* PRESEEA-ALCALÁ. I. Hablantes de Instrucción Superior**. Alcalá de Henares: Universidad de Alcalá, 2002. CD-ROM.

_____. **La lengua hablada en Alcalá de Henares. *Corpus* PRESEEA-ALCALÁ. II. Hablantes de Instrucción Media**. Alcalá de Henares: Universidad de Alcalá, 2004. CD-ROM.

Artigo recebido em: 28.09.2014
Artigo aprovado em: 15.12.2014

Reflexões sobre anotação sintática e ferramentas de busca - Uso da linguagem XML para anotação sintática no *corpus* digital DOViC

Reflections on syntactic annotation and search tools - Using the XML for syntactic annotation in digital corpus DOViC

Cristiane Namiuti Temponi[*]
Aline Silva Costa[**]

RESUMO: *Penn TreeBank* para anotação sintática no *corpus* digital DOViC, uma vez que esta linguagem já é utilizada para a anotação de edições e de informações morfológicas neste *corpus*. Assim, uma única tecnologia pode ser usada para os diversos tipos de buscas automáticas. Para uma experimentação da anotação sintática com XML, implementamos um programa que faz a conversão do formato *Penn TreeBank* para a linguagem alvo, e foram realizadas algumas pesquisas sintáticas com a linguagem XPath, uma linguagem de consulta para a tecnologia XML. As buscas realizadas foram comparadas com as mesmas buscas feitas na ferramenta *corpus* Search, uma ferramenta específica para o formato *Penn TreeBank*. O uso de XML para todas as representações favorece a criação de recursos padronizados, que podem ser reutilizados, facilitando a extração de dados de *corpora*. A disponibilidade de anotação usando um padrão como XML também oferece independência tecnológica a outros grupos pesquisadores interessados no *corpus*.

PALAVRAS-CHAVE: *Corpus*. XML. XPath. *Penn TreeBank*.

ABSTRACT: This paper makes reflections on the use of XML as an alternative format for *Penn TreeBank* syntactic annotation in digital corpus DOViC, since this language is already used for the annotation of editions and morphological information in this corpus. Thus, a single technology can be used for various types of automatic searches. For a trial of syntactic annotation with XML, we implemented a program that does the conversion of the *Penn TreeBank* format for the target language, and some syntactic research with the XPath language, a query language for XML technology were performed. The queries were compared with the same search queries made in tool corpus Search, a tool for the specific format *TreeBank Penn*. The use of XML for any representations favors the creation of standard features, which can be re-used, facilitating the extraction of data from corpora. The availability of annotation using XML as a standard also offers technological independence to other researchers interested groups in the corpus.

KEYWORDS: Corpus. XML. XPath. *Penn TreeBank*.

1. Introdução

Nos últimos anos, *corpora* cada vez maiores de recursos linguísticos foram desenvolvidos e anotados pelos estudiosos da linguagem. Certos princípios de representação

[*] Doutora em Linguística pela Universidade Estadual de Campinas (UNICAMP) e professora do Departamento de Estudos Lingüísticos e Literários e do Programa de Pós-Graduação em Linguística da Universidade Estadual do Sudoeste da Bahia (UESB).
[**] Bacharel em Ciências da Computação pela Universidade Estadual do Sudoeste da Bahia (UESB) e Mestranda do Programa de Pós-Graduação em Linguística da UESB.

têm sido amplamente adotados, como o uso de anotação *stand-off*[1] ou o uso da linguagem XML, e foram feitas várias tentativas para proporcionar mecanismos e formatos de anotação genéricos. Apesar de tais esforços, os formatos de anotação variam consideravelmente para cada recurso linguístico, projeto ou *corpus*, muitas vezes para satisfazer as restrições impostas por determinado software de processamento. Tratando-se especificamente dos recursos sintáticos, existem diversos formatos para representar a estrutura de sentenças em *corpora* linguísticos digitais. Essa variedade de formatos, no entanto, dificulta o acesso aos dados sintáticos, uma vez que cada formato exige tecnologias de processamento específicas. Na análise de línguas naturais, os dados linguísticos podem ser reutilizados, servindo a diversas pesquisas. Mas pela variedade de representações existentes, as ferramentas e aplicações computacionais desenvolvidas são raramente reutilizadas. A comunidade de processamento de linguagem reconhece que a uniformização e interoperabilidade são cada vez mais prementes para permitir o compartilhamento, fusão e comparação de recursos linguísticos (IDE; ROMARY; DE CLERGERIE, 2004).

O *corpus* de Documentos Oitocentistas de Vitória da Conquista – DOViC – utiliza a linguagem XML para anotação de edições e representação da morfologia dos textos que o compõem. O esquema de anotação e ferramenta utilizados são os mesmos utilizados pelo *Corpus Histórico do Português Tycho Brahe*. No entanto, a representação da estrutura sintática no *Tycho Brahe* não é feita utilizando essa mesma linguagem, mas sim um outro formato, o *Penn TreeBank*. Dessa maneira, a extração dos dados sintáticos demanda o uso de uma tecnologia diversa, uma ferramenta que faça buscas em arquivos nesse formato específico. A ferramenta utilizada para este propósito no *Tycho Brahe* é o programa *Corpus Search*.

Este trabalho discute o uso da mesma tecnologia já utilizada na representação da morfologia, a linguagem XML, como uma alternativa ao formato *Penn TreeBank* para anotação sintática. XML é uma linguagem que permite descrever qualquer tipo de dado e é um padrão aberto para interoperabilidade e intercâmbio de informações. A existência de uma ampla variedade de tecnologias para esse padrão permite a criação de recursos padronizados, favorecendo a reutilização tecnológica e facilitando a extração de dados de *corpora*.

[1] Anotação *stand- off* é uma estratégia de anotação em que se mantém os dados anotados em documentos separados dos documentos com os dados originais (IDE; ROMARY; DE CLERGERIE, 2004).

Para experimentação de uma representação de estrutura sintática usand XML foi implementado um programa que realiza a conversão do formato *Penn TreeBank* para a linguagem alvo. Buscas automáticas nesse formato foram realizadas com a linguagem de consulta XPath. As buscas realizadas foram então comparadas com as mesmas buscas feitas na ferramenta *Corpus Search*. Finalmente, foi feita uma análise qualitativa de custo/benefício para uso da linguagem em questão no *corpus* DOViC.

Na seção dois, a linguagem XML será abordada brevemente. A seção três apresenta o *corpus* digital DOViC e sua metodologia de anotação. As seções quatro a sete tratam de padrões de anotação para *corpora*, do formato Penn TreeBank e do uso de XML em anotações sintáticas. A seção oito apresenta sucintamente uma linguagem para consultas em XML, a Xpath. As seções seguintes apresentam a proposta do trabalho, mostrando o resultado do programa implementado para a conversão de formatos e as buscas realizadas em XPath. Por fim, a seção onze faz uma análise qualitativa do custo/benefício para uso da proposta, seguida das considerações finais.

2. A linguagem XML

XML (*Extensible Markup Language*) é uma linguagem de editoração que oferece um formato universal para estruturação de documentos e dados na Web. Proposta pelo W3C[2] (*World Wide Web Consortium*) como uma nova alternativa à linguagem HTML (*Hiper Text Markup Language*), linguagem dominante na Web, a XML combina extensibilidade, poder e flexibilidade com a simplicidade exigida pela Web (SILVA FILHO, 2004; DEITEL; DEITEL; NIETO; LIN; SADU, 2005).

Documentos XML são documentos de texto que representam dados de maneira estruturada utilizando um conjunto de *tags*[3] ou elementos. Tal conjunto não é fixo nem limitado, podendo ser estendido. Assim, os autores dos documentos podem criar suas próprias *tags* para atender a necessidades específicas, o que torna a linguagem poderosa para representar qualquer tipo de dado conferindo-lhe a classificação como uma metalinguagem (SILVA FILHO, 2004; DEITEL; DEITEL; NIETO; LIN; SADU, 2005).

[2] O W3C é uma organização, fundada em 1994, destinada a desenvolver tecnologias interoperantes e de domínio público para a *World Wide Web* (DEITEL et al, 2005).
[3] Os termos *marca*, *elemento* ou *etiqueta* podem ser usados como sinônimo de *tag*.

Ainda que baseie-se em texto, "a XML não se limita a descrever somente dados textuais, mas também pode descrever imagens, gráficos vetoriais, animações ou qualquer outro tipo de dado para o qual seja estendida" (DEITEL; DEITEL; NIETO; LIN; SADU, 2005).

Dados representados por XML são estruturados de forma arbórea, e cada *tag* ou marca representa um nó ou elemento na árvore. "A sintaxe de XML requer um único elemento como nó raiz, uma marca de abertura e de finalização para cada elemento, marcas corretamente aninhadas e valores de atributos entre aspas." (DEITEL; DEITEL; NIETO; LIN; SADU, 2005).

O quadro 1 mostra um exemplo de um documento XML representando os dados de um livro, com as informações de autor, título e ISBN. O nó raiz é <livro> e este possui como filhos três nós <autor> e um nó <titulo>. A informação de ISBN foi representada como atributo do nó <livro> e seu valor no exemplo é "978-85-7244-800-0".

Quadro 1: Exemplo de um documento XML

```
<livro ISBN="978-85-7244-800-0">
  <autor> Carlos Mioto </autor>
  <autor> Ruth Lopes </autor>
<autor> Maria Cristina Figueiredo Silva </autor>
  <titulo> Novo Manual de Sintaxe</titulo>
  </livro>
```

Os documentos XML são legíveis para as pessoas e também manipuláveis por computadores. A ausência de instruções de formatação facilita a realização do processamento sintático de sua estrutura, o que a torna uma referência que pode ser usada para o intercâmbio de dados. Para obter funcionalidade e interoperabilidade na Web, desenvolvedores de software em todo o mundo estão integrando XML a seus aplicativos. Contudo, a XML não está limitada a aplicações Web (DEITEL; DEITEL; NIETO; LIN; SADU, 2005).

Atualmente, a linguagem XML é um dos formatos mais utilizados para compartilhamento de informação estruturada entre aplicativos, independente de plataforma. Como é um padrão aberto, existe uma grande quantidade de opções relacionadas às ferramentas para implementá-la, permitindo que o usuário escolha o que melhor se ajuste às suas necessidades (W3C, 2010; DEITEL; DEITEL; NIETO; LIN; SADU, 2005)

3. O *corpus* DOViC

O *corpus* DOViC (*corpus* de Documentos Oitocentistas de Vitória da Conquista) é um *corpus* digital de documentos manuscritos do século XIX, desenvolvido no âmbito do projeto "Memória conquistense: implementação de um *corpus* digital"[4] (NAMIUTI, 2013) em parceria com o projeto de pesquisa "Sintaxe diacrônica em *corpus* eletrônico: do português pré-clássico às variantes modernas" (NAMIUTI; SANTOS, 2010). Os documentos manuscritos que compõem o *corpus* estão guardados nos arquivos do Fórum de Vitória da Conquista-Bahia.

Os textos do corpus DOViC são transcritos, editados e anotados nos mesmos moldes do *Corpus Histórico do Português Tycho Brahe*, utilizando a mesma ferramenta e mesmo esquema de anotação. O *Corpus Tycho Brahe* é um *corpus* digital composto de textos em português escritos por autores nascidos entre 1380 e 1845, desenvolvido na Universidade Estadual de Campinas (UNICAMP). O desenvolvimento deste *corpus* se deu a partir de 1998, no âmbito do Projeto "Padrões Rítmicos, Fixação de Parâmetros e Mudança Linguística" (UNICAMP, 1998).

A transcrição e edição dos textos do *corpus* DOViC são feitos com o auxílio da ferramenta E-Dictor (KEPLER; PAIXÃO DE SOUSA; FARIA, 2010). O texto transcrito é salvo em um arquivo no formato texto simples. Edições como modernização, junção, segmentação e modernização de grafia são feitas por meio da interface gráfica da ferramenta, produzindo como resultado um arquivo anotado na linguagem XML. O software realiza anotação das informações morfológicas dos textos, também no formato XML e ambas as anotações são feitas num único arquivo. Esse esquema de anotação suportado pelo E-Dictor, utilizado tanto no *corpus Tycho Brahe* quanto no DOViC, foi concebido dentro do projeto "Memória dos Texto" (PAIXÃO DE SOUSA, 2006). Como esse processo é feito por meio da interface gráfica, o uso da linguagem XML é transparente para o usuário, ou seja, ele não lida diretamente com essa estrutura.

A anotação de edição é realizada identificando todos os itens acrescentados ao texto pelo editor como elementos <e>, e os itens originais correspondentes como elementos <o>. O tipo de edição é identificado através da propriedade "t" dos elementos <e> e os tipos possíveis são listados na tabela 1. A numeração dos elementos, identificados pela propriedade "id" são atribuídas automaticamente pela ferramenta (UNICAMP, 2007).

[4] Projeto financiado pelo CNPq (CNPq 485098/2013-0).

As anotações de informações morfológicas dos textos também são feitas em XML e mantidas no mesmo arquivo com as edições. A identificação de informação morfológica dá-se pela marcação do item lexical com o elemento <m>. A propriedade "v" marca o valor da categoria lexical. A figura 1 mostra um trecho da anotação gerada pelo E-Dictor para um texto do *corpus* DoVic.

Tabela 1: Tipos de edição possíveis para o *corpus Tycho Brahe* e representação na anotação XML.

Tipo de edição	**Atributo de <e>**	**Exemplo**
uniformização grafemática e de módulo	t="gra"	<e t="gra">serviço</e><o>feruiço</o>
separação ou junção de vocábulos	t="seg"	<e t="seg">que fe</e><o>quefe</o>
expansão de abreviatura	t="exp"	<e t="exp">Vossa Mercê</e><o>V.M.</o>
uniformização de pontuação	t="punc"	<e t="punc"> >> </e><or> " </or>
modernização de grafia	t="mod"	<e t="mod">inclita</e><o>inclita</o>
Correções	t="cor"	<e t="cor">depois</e><o>deqois</o>

Fonte: Unicamp (2007).

```
<text t="full" words="130" id="text_1">
  <sc id="sc_1">
    <p id="p_1">
      <s id="s_1">
        <w id="s_1#0">
          <o>eordeno</o>
          <e t="seg">e ordeno</e>
          <m v="CONJ"/>
          <m v="VB-P"/>
        </w>
        <w id="s_1#1">
          <o>atodos</o>
          <e t="seg">a todos</e>
          <m v="P"/>
          <m v="Q-P"/>
        </w>
        <w id="s_1#2">
          <o>osOfficiaes</o>
          <e t="mod">os oficiais</e>
          <e t="seg">os Officiaes</e>
          <m v="D-P"/>
          <m v="N-P"/>
        </w>
        <w id="s_1#3">
          <o>de</o>
          <m v="P"/>
        </w>
        <w id="s_1#4">
          <o>Justiça</o>
          <m v="NPR"/>
        </w>
        <w id="s_1#5">
          <o>desta</o>
          <m v="P+D-F"/>
        </w>
        <w id="s_1#6">
          <o>sobre</o>
```

Figura 1 - Arquivo XML gerado pelo E-Dictor para um documento do *corpus* DOViC.

A versão atual do programa E-Dictor (versão 1.0 beta 10) não realiza anotação da estrutura sintática. Tal informação é gerada separadamente utilizando um *parser* que recebe como entrada um arquivo anotado no formato POS (*Part of Speech*), e gera como saída um arquivo texto no formato *Penn TreeBank*, que será detalhado na seção 5. O treinamento do *parser* foi feito para o português clássico na Universidade da Pensilvânia. Para obtenção da

representação sintática nos textos do *corpus* DOViC, os textos deverão ainda passar pelo mesmo processo de etiquetagem.

4. Padrões para anotações de *corpora*

O aumento das pesquisas em Linguística de *Corpus* e o crescimento na disponibilidade de *corpora* eletrônico fizeram com que diversos formatos de codificação e anotação de textos surgissem. Cada projeto de compilação de *corpus* pode criar e/ou definir um formato, com o objetivo de atender requisitos das ferramentas de anotação e exploração de *corpus* específicas. A diversidade de formatos aumentou a importância e a necessidade de estabelecimento de padrões que facilitassem o compartilhamento, a combinação e o intercâmbio desses recursos. Entre os principais projetos e iniciativas com o propósito de definir um padrão de codificação e anotação de textos, podemos destacar: MuchMore, Tiger- XML[5], Text Encoding Initiative (TEI)[6], Corpus Encoding Standard (CES), Corpus Encoding Standard for XML (XCES) e padrão ISO TC37/SC4.

O XCES é a versão do padrão CES (Corpus *Encoding Standard*) baseado em XML. O CES é um padrão de codificação para *corpora* destinado a atender a necessidade do desenvolvimento de práticas de codificação padronizados para *corpora* linguísticos. O CES identifica um nível de codificação mínima que *corpora* devem alcançar para ser considerado padronizado em termos de representação descritiva (marcação de informação estrutural e linguística) (IDE, 1998; IDE; BONHOMME; ROMARY, 2000).

O Padrão ISO TC37/SC4 é um *framework* para anotação de informação linguística desenvolvido pela Organização Internacional de Padronização (*International Organization for Standardization)*. A ISO formou um subcomitê (SC4) no âmbito da Comissão Técnica 37 (TC37, *Terminology and Other Languages Resources*) com o objetivo de estabelecer padrões internacionais e recomendações para a modelagem de dados, anotação, intercâmbio de dados e avaliação de recursos linguísticos. Dentre os diversos grupos de trabalho do TC37/SC4, um grupo foi criado para prover um *framework* para anotação linguística. A intenção não é definir um esquema ou formato único e definitivo de anotação, mas fornecer uma arquitetura p que possa servir de referência para diferentes esquemas de anotação, permitindo a fusão ou comparação entre eles. A estrutura do *framework* tem como finalidade prover o máximo de

[5] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html
[6] http://www.tei-c.org/index.xml

flexibilidade para codificadores e anotadores, e ao mesmo tempo permitir e estimular o intercâmbio e reutilização de recursos linguísticos anotados (IDE; ROMARY; DE LA CLERGERIE, 2004).

O projeto MuchMore (*Multilingual Concept Hierarchies for Medical Information Organization and Retrieval*) propõe um formato de anotação linguística capaz de integrar múltiplos níveis de informação: anotação morfológica, sintática e semântica. O formato é baseado em XML e os níveis de informação podem ser organizados separadamente, sendo integrados através de referência a identificadores (BUITELAAR et al., 2003).

5. O formato *Penn TreeBank*

Assim como há vários formatos para representação e armazenamento de *corpora* linguísticos, há também um variado número de formatos para representação e anotação da estrutura sintática dos textos que os compõem, como Tipster, *Penn TreeBank*, Susanne e NeGra (LEZIUS; MENGEL, 2000).

O *Penn TreeBank Format* (Formato *Penn TreeBank*) é um esquema de anotação sintática de *corpora* desenvolvido pela Universidade da Pensilvânia. O esquema utiliza uma representação arbórea delimitada por parênteses etiquetados. Todos os parênteses abertos têm uma etiqueta associada, sendo uma etiqueta *phrase* (NP, ADJP, etc), associada a projeções máximas da teoria X-Barra, ou uma etiqueta *word* (N, ADJ , etc), associadas a núcleos da mesma teoria, representando os nós de uma árvore (SANTORINI, 2010; MARCUS; TAYLOR, 2002).

A cada palavra está associada uma etiqueta *word*, mas nem sempre uma etiqueta *phrase* será associada a cada nó correspondente em uma árvore da teoria sintática. As projeções intermediárias da teoria X-Barra (N', ADJ', etc) não são incuídas nessa representação. Outras categorias também são omitidas nesse esquema de anotação, como por exemplo, VP e DP (SANTORINI, 2010).

A representação parcial da estrutura sintática se dá por razões práticas, e por esse motivo não se mantém a mesma estrutura correspondente à árvore teórica. A categoria DP, por exemplo, é omitida porque o custo de incluí-la supera sua utilidade. Outra diferença para as árvores da teoria sintática é que nesse esquema de representação as árvores não são obrigatoriamente binárias, ou seja, cada nó pode ter mais de duas ramificações (SANTORINI, 2010).

Uma estrutura típica de análise sintática com anotação nesse formato é dada como exemplo no quadro 2. A figura 2 mostra a representação gráfica correspondente a esta mesma estrutura de análise.

Quadro 2 - Estrutura de análise de uma sentença na anotação *Penn TreeBank*

```
((IP-MAT (NP-SBJ (NPR Mary))
        (HVP has)
        (BEN been)
        (VAG meaning)
        (IP-INF (TO to)
              (VB go))
        (PP (P for)
           (NP (D a) (N week)))))
```

Fonte: Santorini (2010).

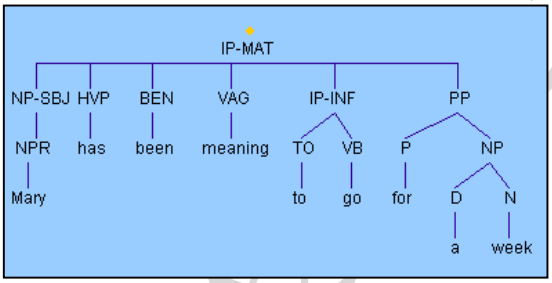


Figura 2 - Representação gráfica de estrutura de análise de uma sentença na anotação *Penn TreeBank*

6. A Ferramenta *Corpus Search*

Assim como há vários formatos para anotação sintática de *corpora*, há também várias ferramentas para extrair informação destes dados anotados, dentre as quais podemos citar: Tgrep2, TIGERsearch, Emu, Corpus Search, NiteQL, Lpath (IMS, 2013).

O *Corpus Search* é um programa que realiza pesquisas sintáticas em *corpora* anotados no formato *Penn TreeBank*. Asssim como o esquema de anotação, o software também foi desenvolvido na Universidade da Pensilvânia (CORPUS SEARCH, 2009).

Corpus Search é implementado na linguagem de programação Java, e portanto, é multiplataforma[7] e requer que o programa JRE[8] (*Java Runtime Environment*) esteja instalado no computador do usuário.

[7] Um programa multiplataforma pode ser executado em qualquer sistema operacional, desde que haja uma máquina virtual apropriada instalada.
[8] JRE (*Java Runtime Environment)* é um software desenvolvido para executar programas feitos na linguagem Java. O JRE possui como componente a máquina virtual Java (JVM- Java Virtual Machine) (ORACLE, 2010).

A execução do *Corpus Search* para realizar buscas sintáticas requer duas entradas: o arquivo do *corpus*, anotado no formato *Penn TreeBank*; e o arquivo com a especificação da consulta a ser realizada, também chamado de *command file*, em formato texto simples.

A especificação das buscas no arquivo de entrada deve estar de acordo com a sintaxe exigida pela linguagem de consulta do *Corpus Search*, que compreende chamadas a funções de busca e uso de operações lógicas. As funções de busca pesquisam relações existentes na estrutura sintática como dominância, c-comando, irmandade, entre outras.

Os resultados de uma busca realizada pelo *Corpus Search* podem ser vistos no arquivo de saída gerado pelo programa. O arquivo é produzido no formato texto simples e reúne informações sobre as sentenças contendo as restrições especificadas pela busca (CORPUS SEARCH, 2009).

7. Utilização da linguagem XML na anotação sintática

Este trabalho discute a utilização da linguagem XML como alternativa para anotação sintática de textos do *corpus* DOViC. Existem numerosos exemplos da implementação de XML em anotações de *corpora*. Entre eles estão os projetos Alpino Dependency *TreeBank*, Europarl Parallel Corpus, Wikipedia XML *corpora*, PDTB XML, e outros. Há ainda outros estudos que visam converter dados de *corpora* em XML ou desenvolver representações XML para unir os dados de *corpora* de múltiplas fontes. O *Expert Advisory Group on Language Engineering Standards* lançou uma codificação XML padrão para *corpus*, o XCES (*XML Corpus Encoding Standard*) (YAO; BORISOVA; ALAM, 2010).

Lezius e Mengel (2000) propõem um esquema de anotação sintática baseado em XML. Nessa abordagem, é proposta a utilização de basicamente quatro elementos XMLpara descrever a estrutura arbórea: elementos sentença <s> , elementos não-terminais <n>, elementos terminais ou palavras <w> e elementos de aresta <edge> , usado para nós de ligação. As categorias dos nós, como categoria sintática ou rótulo POS são representados como atributos das *tags* XML. Um exemplo da anotação proposta é mostrado na figura 3.

O padrão XCES (seção 4) descreve um padrão de codificação em XML para anotações linguísticas com informações morfossintáticas. Assim, informações sobre estrutura de sentenças e informações morfológicas são mantidas numa única estrutura. As informações morfossintáticas são anotadas utilizando-se dos elementos <tok>. A integração com os dados primários é feita através do atributo "xlink". Elementos <s> marcam sentenças e etiquetas

<par> marcam parágrafos. A figura 4 mostra um fragmento de um texto contendo anotações de informação morfossintática neste padrão.

```
<s id="s1" href="#id(n1_500)"/>

<n id="n1_500" cat="S">
  <edge id="edge1_1" href="#id(n1_501)"/>
  <edge id="edge1_2" href="#id(n1_502)"/>
</n>
<n id="n1_501" cat="NP">
  <edge id="edge1_3" href="#id(w1_0)"/>
  <edge id="edge1_4" href="#id(w1_1)"/>
</n>
<n id="n1_502" cat="VP">
  <edge id="edge1_5" href="#id(w1_2)"/>
  <edge id="edge1_6" href="#id(n1_503)"/>
</n>
<n id="n1_503" cat="NP">
  <edge id="edge1_7" href="#id(w1_3)"/>
  <edge id="edge1_8" href="#id(w1_4)"/>
</n>

<w id="w1_0" word="The"/>
<w id="w1_1" word="boy"/>
<w id="w1_2" word="likes"/>
<w id="w1_3" word="the"/>
<w id="w1_4" word="girl"/>
```

Figura 3 - Exemplo de anotação sintática usando XML proposta por Lezius e Mengel (2000).
Fonte: Lezius; Mengel (2000).

O projeto MuchMore (seção 4) propõe um formato de anotação baseado em XML onde diversos níveis de informação podem ser mantidos separadamente mas integrados através de identificadores. A figura 5 exemplifica um trecho de texto anotado no formato do MuchMore. O texto é representado pelo elemento <text>, que por sua vez, é composto de um ou mais elementos <token>, que identificam as palavras, os quais marcam através de atributos as informações morfossintáticas, além da forma canônica de cada palavra. As estruturas sintáticas são representadas pelos elementos <chunk> , cujos atributos "from" e "to" marcam onde começa e onde termina a estrutura. O atributo "type" classifica a estrutura como NP, PP, etc. Os atributos "id" dos elementos permitem a referência para a integração de múltiplos níveis de informação linguística.

Yao, Alam e Borisova (2010) apresentam o projeto PDTB XML, um projeto que converte os textos do *corpus Penn* Discourse *TreeBank* 2.0 para o formato XML. O PDTB é um grande *corpus* construído na Universidade da Pensilvânia, anotado com informações sintáticas e relações de discurso, argumentos, atribuições e sentido. O esquema de anotação utilizado possui o mesmo nome do *corpus*, *Penn* Discourse *TreeBank*.

```
<?xml version="1.0">
<chunk type="BODY" lang="en"
 xml:base=
"http://www.cs.vassar.edu/~ME/Oen.xcesDoc#">
 <par xlink:href="xptr(substring(//p[1]">
  <s xlink:href="xptr(substring(//p/s[1]">
   <tok type="WORD"
     xlink:href=
     "xptr(substring(//p/s[1]/text(),1,2">
      <orth>It</orth>
      <disamb>
        <base>it</base>
        <msd>Pp3ns</msd>
        <ctag>PPER3</ctag></lex>
      <lex>
        <base>it</base>
        <msd>Pp3ns</msd>
        <ctag>PPER3</ctag></lex></tok>
    <tok type="WORD"
      xlink:href=
      "xptr(substring(//p/s[1]/text(),4,2">

     <orth>was</orth>
     <disamb>
       <base>be</base>
       <msd>Vmis3s</msd>
       <ctag>PAST3</ctag></lex>
     <lex>
       <base>be</base>
       <msd>Vais1s</msd>
       <ctag>AUX1</ctag></lex>
     <lex>
   <base>be</base>
   <msd>Vais3s</msd>
   <ctag>AUX3</ctag></lex>
<lex>
   <base>be</base>
   <msd>Vmis1s</msd>
   <ctag>PAST1</ctag></lex>
<lex>
   <base>be</base>
   <msd>Vmis3s</msd>
   <ctag>PAST3</ctag></lex></tok>…
```

Figura 4 - Fragmento de texto com anotação no padrão XCES.
Fonte: Ide; Bonhomme; Romary (2000).

```
Balint syndrom is a combination of symptoms including simultanagnosia, a
disorder of spatial and object-based attention, disturbed spatial percep-
tion and representation, and optic ataxia resulting from bilateral pa-
rieto-occipital lesions.

<text>
  <token id="w1" pos="NN">Balint</token>
  <token id="w2" pos="NN">syndrom</token>
  <token id="w3" pos="VBZ" lemma="be">is</token>
  <token id="w4" pos="DT" lemma="a">a</token>
  <token id="w5" pos="NN" lemma="combination">combination</token>
  ...
  <token id="w20" pos="JJ" lemma="spatial">spatial</token>
  <token id="w21" pos="NN" lemma="perception">perception</token>
  <token id="w22" pos="CC" lemma="and">and</token>
  <token id="w23" pos="NN" lemma="representation">representation</token>
  ...
</text>

<chunks>
      <chunk id="c1" from="w1" to="w2" type="NP"/>
      <chunk id="c7" from="w20" to="w23" type="NP"/>
</chunks>
```

Figura 5 – Exemplo de anotações linguísticas no MUCHMORE.
Fonte: Buitelaar et al. (2003).

8. XPath

A linguagem XML descreve dados de forma flexível e eficiente através da marcação dos dados com *tags* descritivas. No entanto, ela não fornece uma maneira de localizar dados específicos dentro de um documento (DEITEL *et al*, 2005).

A linguagem XPath (*XML Path*), recomendada pelo W3C, fornece uma sintaxe para localizar dados específicos em um documento XML de forma efetiva e eficiente. XPath modela um documento XML como uma árvore de nós. É uma linguagem de expressões, baseada em *strings*, para localizar conteúdo dentro da árvore que representa o documento XML (W3C, 1999; DEITEL *et al*, 2005).

Exemplos de expressões XPath são dados nos quadros 3 a 5. Todos os exemplos podem ser aplicados ao documento XML dado como exemplo na figura 1. A expressão na figura 8 localiza todos os nós <titulo>, que sejam filhos de <livro>. A expressão na figura 9 localiza o nó <livro> que possua um atributo ISBN cujo valor seja "978-85-7244-800-0". E por fim, a expressão na figura 10 localiza o terceiro nó filho <autor> do nó <livro>.

Quadro 3 - Exemplo de expressão XPath para localizar nós <titulo> filhos de <livro>

```
/livro/titulo
```

Quadro 4 - Exemplo de expressão XPath para localizar nós <livro> com atributo ISBN com valor "978-85-7244-800-0"

```
/livro[@ISBN="978-85-7244-800-0"]
```

Quadro 5 - Exemplo de expressão XPath para localizar o terceiro nó <autor> filho de nós <titulo>

```
/livro/autor[3]
```

9. Conversor do formato *Penn TreeBank* para XML

Para a transformação do formato *Penn TreeBank* para XML, foi desenvolvido neste trabalho um programa na linguagem Java, que recebe como entrada um arquivo no primeiro formato e gera um arquivo de saída XML correspondente. O programa não implementa ainda a função de *parser*, e portanto, o arquivo de entrada deve ser um documento *Penn TreeBank* bem formado. Para uso futuro, o programa deve implementar a função de *parser* a fim de evitar entradas errôneas. O programa não possui interface gráfica, exibindo apenas uma janela de diálogo para fornecimento do arquivo de entrada pelo usuário.

Para o arquivo de saída, foram usados os mesmos nomes de rótulos para nomear as *tags*, com exceção do nó raiz e de rótulos com caracteres não aceitos pela linguagem XML.

Como o arquivo de entrada não possui um elemento raiz, foi inserido no arquivo de saída a *tag* <DOCUMENT> como raiz do documento. Nós com o sinal de pontuação "." no formato *Penn TreeBank* foram mapeados para *tags* <POINT>. Nós com o símbolo "," foram mapeados para *tags* <COMMA>. Houve a necessidade de substituir o caracter "$" pelo caracter "S". Assim, rótulos como "PRO$" foram mapeados para etiquetas "PROS". Os demais nomes das tags para o documento XML permaneceram os mesmos utilizados no formato *Penn* Tree Bank. Assim, cada nó do arquivo de entrada é mapeado numa tag XML com mesmo nome. Cada nó folha (nó sem filho) é gerado na saída como texto puro entre as tags.

O quadro 6 mostra um trecho de um arquivo do *corpus Tycho Brahe* com anotação sintática *Penn TreeBank* e o quadro 7 mostra o arquivo saída correspondente em XML gerado pelo programa.

Quadro 6 - Trecho de arquivo do *corpus Tycho Brahe* com anotação *Penn TreeBank*

```
( (IP-MAT (NP-SBJ *pro*)
          (VB-R Darei)
          (NP-ACC (N princípio))
          (PP (P a)
             (NP (D-F-P estas) (PRO$-F-P minhas) (N-P memórias)))
          (RRC (P por)
              (NP (D-F a) (PRO$-F minha) (N genealogia)))
          (. .)) (ID A_003_PSD,03.1))
```

Quadro 7 - Trecho de arquivo com anotação sintática em XML gerado pelo programa conversor

```
<IP-MAT>
<NP-SBJ>
*pro*
</NP-SBJ>
<VB-R>
Darei
</VB-R>
<NP-ACC>
<N>
princípio
</N>
</NP-ACC>
<PP>
<P>
a
```

</P>
<NP>
<D-F-P>
estas
</D-F-P>
<PROS-F-P>
minhas
</PROS-F-P>
<N-P>
memórias
</N-P>
</NP>
</PP>
<RRC>
<P>
por
</P>
<NP>
<D-F>
a
</D-F>
<PROS-F>
minha
</PROS-F>
<N>
genealogia
</N>
</NP>
</RRC>
<POINT>
.
</POINT>
</IP-MAT>
<ID>
A_003_PSD,03.1
</ID>

A hierarquia na estrutura gerada pode ser melhor visualizada usando qualquer ferramenta que represente as relações hierárquicas inserindo tabulações, como navegadores e outros. A figura 6 mostra a visualização do documento no navegador Firefox.

10. Buscas sintáticas utilizando anotação XML

As buscas nos arquivos de anotação sintática com XML podem ser feitas utilizando uma linguagem de consulta para esta linguagem. Neste trabalho, a linguagem XPath foi utilizada. Como exemplo, foram realizadas duas buscas em um arquivo com anotação sintática *Penn TreeBank* do *corpus Tycho Brahe*, envolvendo relações de dominância ou maternidade e irmandade. Depois de convertido o arquivo para a anotação XML proposta, buscas equivalentes foram realizadas dentro deste arquivo XML com a linguagem XPath.

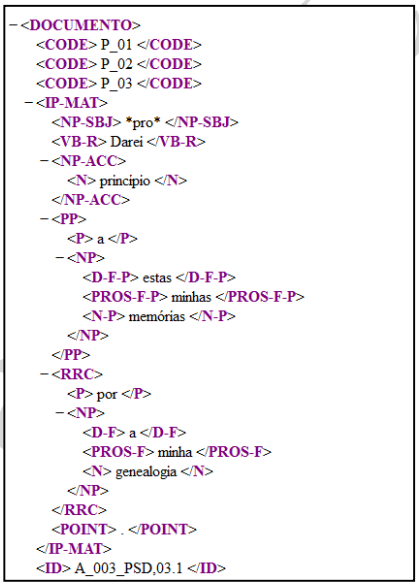


Figura 6 - Visualização da estrutura hierárquica de anotação com XML no navegador Firefox

Para as buscas com XPath, foi implementado um segundo programa na linguagem Java, utilizando a API[9] (*Application Programming Interface*) para XPath. Além da implementação e utilização deste programa, as mesmas buscas também foram feitas no navegador FireFox, através da instalação do *plugin XPath Checker*, disponível gratuitamente

[9] API (Application Programming Interface) é um conjunto de funções e padrões estabelecidos por um software para a utilização das suas funcionalidades por aplicativos que não pretendem conhecer detalhes da implementação do software, mas apenas em usar seus serviços.

na página de complementos para este navegador. Outros programas editores de XML com processadores XML integrados também estão disponíveis na Internet.

A busca para relação de dominância no *Corpus Search* utiliza a função *dominates.* Para realizar uma busca de nós NP que dominam nós PP utilizando essa ferramenta, a seguinte expressão foi utilizada no arquivo de consulta:

(NP dominates PP) (10.1)

Para realizar a busca equivalente em XPath, a expressão de consulta utilizada foi:

(//NP[PP]) (10.2)

Para pesquisas com relação de irmandade, o *Corpus Search* utiliza a função *hasSister*. Para realizar uma busca de nós P que possuem nós irmãos NP utilizando esta ferramenta, a seguinte expressão foi utilizada no arquivo de consulta:

(P hasSister NP) (10.3)

A busca equivalente na linguagem XPath foi feita da seguinte maneira:

//P/following-sibling::NP|NP/following-sibling::P (10.4)

11. Análise do uso de XML/XPath para anotação e buscas sintáticas

Um esquema de anotação sintática utilizando XML traz a vantagem de utilizar um padrão aberto para interoperabilidade e intercâmbio de dados. Utilizando formatos específicos para o esquema de anotação sintática, as tecnologias para recuperação da informação dificilmente são reutilizadas. Para cada tipo de anotação, são necessárias ferramentas de busca restritas àquela anotação em questão.

Para avaliar o uso do *Corpus Search* e da linguagem XPath como ferramentas de busca, a análise pode ser feita sob diversas perspectivas, tanto nos aspectos tecnológicos, quanto na utilização por usuários finais.

Considerando linguistas como usuários finais de tais ferramentas[10], o *Corpus Search* possui uma linguagem mais simples e mais fácil de aprender que XPath. Como o *Corpus Search* é específico para buscas sintáticas, os comandos foram projetados para este fim, trazendo assim mais simplicidade se comparado à XPath. Para pesquisar uma relação de C-comando, por exemplo, o *Corpus Search* possui a função *ccomands*. Na linguagem XPath seria necessária a combinação de várias expressões utilizando-se de operadores para realizar a busca equivalente. A linguagem XPath não é destinada ao uso por usuários leigos em programação de computadores. Ainda assim, o uso do *Corpus Search* também requer aprendizado de sua linguagem específica, além de requerer que o usuário final saiba trabalhar com linhas de comando, instalação e configuração da máquina virtual Java (JVM), não podendo ser um usuário totalmente leigo.

Se considerarmos a existência de uma aplicação intermediária que forneça uma interface para realização das buscas, a tecnologia utilizada para o usuário torna-se transparente, uma vez que não terá conhecimento do que está realmente sendo utilizado na busca, se *Corpus Search*, XPath ou qualquer outra tecnologia. A comparação neste caso, deverá ser feita apenas tratando aspectos tecnológicos sob o ponto de vista do desenvolvedor da aplicação, como facilidade de implementação e esforço exigido de programação.

Em se tratando dos aspectos tecnológicos, o uso do *Corpus Search* traz a vantagem de trazer no arquivo de saída os resultados das buscas, com os trechos encontrados e as estatísticas. Para buscas em XPath, os conjunto dos nós encontrados também é retornado, mas é preciso que o desenvolvedor da aplicação implemente um tratamento deste resultado para exibi-lo para o usuário final. De qualquer sorte, isto pode ser feito sem um exagerado esforço de programação uma vez que as APIs para XML e XPath da linguagem Java fornecem várias funções para isso. Se for desejável não mostrar para os usuários finais o arquivo de saída do *Corpus Search* tal como é exibido, também será exigido um esforço de programação para tratá-lo a fim de apresentar as informações de outra maneira. Se as buscas são feitas em XML, haverá maior flexibilidade ao desenvolvimento da aplicação para exibição dos resultados da consulta. No *Corpus Search*, os resultados são restritos ao que é trazido no arquivo de saída.

[10] Usuários finais usam a ferramenta diretamente, sem a existência de um outro programa que disponibilize uma interface para facilitar o uso

Com o uso de XML em *corpora* digitais, as buscas sintáticas tornam-se independentes de tecnologia específica, passando a utilizar tecnologias padrão. Em se tratando do *corpus* DOViC, a vantagem é a reaplicação da mesma tecnologia que será utilizada para as pesquisas morfológicas. Como a anotação morfológica e de edições dos textos do *corpus* já é feita em XML, as buscas nestes arquivos terão que ser feitas obrigatoriamente utilizando tecnologias para XML. Assim, a mesma tecnologia pode então ser reutilizada, dispensando o uso do *Corpus Search*.

A disponibilidade de uma vasta gama de implementações para XML torna o acesso a este formalismo mais fácil. Além de XPath, existem outras linguagens de busca, como XQuery, com mais poder e flexibilidade. Muitos SGBDs (Sistemas Gerenciadores de Bancos de Dados) já implementam o suporte a XML e fornecem um mecanismo de processamento das linguagens de consulta, como o banco de dados PostgreSQL.

As linguagens de consulta não são o único mecanismo de recuperação de dados num documento XML. As buscas podem ser realizadas utilizando-se apenas das APIs para XML, que estão disponíveis em diversas linguagens de programação, com variadas funções para navegação na estrutura arbórea do arquivo XML. Há também ferramentas de visualização que proporcionam uma visão geral da estrutura. O suporte a XML implementado pelos navegadores em combinação com folhas de estilo poderão ser utilizados para uma exibição customizada da estrutura sintática.

Outra vantagem importante de XML como formalismo de codificação para anotação sintática é que a marcação XML é altamente expansível. Isto significa que diferentes níveis de anotação podem ser combinados, como por exemplo o discurso e a sintaxe. O uso de XML na anotação sintática através da conversão *Penn TreeBank* para XML pode dispensar o uso do *Corpus Search*, mas ainda mantém a dependência do *parser* que gera o arquivo PTB. Como a XML já tem sido usada por diversos *corpora*, projetos futuros podem considerar o desenvolvimento de um *parser* que já produza a estrutura sintática em XML. De qualquer sorte, o uso de XML para todas as representações favorece a criação de recursos padronizados, que podem ser reutilizados, facilitando assim a extração de dados de *corpora*. A disponibilidade de anotação usando um padrão como XML se faz importante porque também oferece independência tecnológica a outros grupos pesquisadores interessados no *corpus* DOViC.

12. Conclusão

O uso de XML para anotação sintática evidenciou a vantagem de reutilizar a mesma tecnologia já utilizada para anotações morfológica e de edições no *corpus* DOViC. Como XML é um padrão, usá-lo para todas as representações nos textos do *corpus* favorece a criação de recursos padronizados, permitindo reuso de tecnologia, oferecendo mais flexibilidade para as buscas e exibição dos resultados, e independência tecnológica para grupos de pesquisa interessados em estudo neste *corpus*.

Trabalhos futuros poderão considerar o desenvolvimento ou emprego de um *parser* que faça anotação da estrutura sintática em XML, sem a necessidade de conversão. Este trabalho não considerou um novo esquema de anotação, como novos nomes de etiquetas, definição de atributos, etc. Assim, um esquema completo de anotação também pode ser desenvolvido, prevendo a sistematização das anotações de edições, morfologia, sintaxe e discurso, baseando-se em padrões existentes mas não deixando de atender às necessidades específicas do *corpus* em questão.

Referências

BUITELAAR, P.; DECLERCK, T.; SACALEANU, B.; VINTAR, S.; RAILEANU, D.; CRISPI, C. A multilayered, XML-based approach to the integration of linguistic and semantic annotations. In: **EACL 2003 Workshop on language technology and the semantic web (NLPXML'03)**, 2003, Budapeste. Proceedings of EACL 2003 Workshop on Language Technology and the Semantic Web (NLPXML'03). Cunningham: EACL, 2003. Disponivel em: <http://www.dfki.de/dfkibib/publications/docs/eacl03-xmlnlp.ps>. Acesso em: 23 set 2014.

CORPUS SEARCH**.** *Corpus Search* **Users Guide**. 2009. Disponível em: <http://corpussearch.sourceforge.net/CS-manual/Contents.html>. Acesso em: 25 jul 2013.

DEITEL, H.M.; DEITEL, P.J.; NIETO, T.M.; LIN, T.M.; SHADU, P.V. **XML**: Como programar. Porto Alegre: Bookman, 2005.

IDE, N. Encoding Linguistic Corpora. In **Proceedings of the Sixth Workshop on Very Large *Corpora***, 1998.

IDE, N.; BONHOMME, P.; ROMARY, L. XCES: An XML-based Encoding Standard for Linguistic Corpora. In: **International language resources and evaluation conference**, 2., 2000, Atenas. Proceedings of the Second International Language Resources and Evaluation Conference. Paris: European Language Resources Association, 2000.

IDE, N.; ROMARY, L.; DE LA CLERGERIE, E.. International standard for a linguistic annotation framework. **Journal of Natural Language Engineering**, Cambridge, v. 10 n. 3-4, pp. 307-326, Sept. 2004.

IMS (Institut für Maschinelle Sprachverarbeitung). **The TIGER-XML treebank encoding format.** Disponível em: <http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/TIGERSearch/doc/html/TigerXML.html>. [ca. 2002]. Acesso em: 08 out 2013.

LEZIUS, W.; MENGEL, A. **An XML-based representation format for syntactically annotated corpora**. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=14E13F7984717A2C1EB5E6CB039C4C92?doi=10.1.1.26.6389&rep=rep1&type=pdf>. 2000. Acesso em: 08 out 2013.

MARCUS, M.; TAYLOR, A.. **The Penn TreeBank Project**. Disponível em: <http://www.cis.u*Penn*.edu/~*TreeBank*/> 2002. Acesso em: 14 out 2013.

NAMIUTTI, C. (Coord.) **Novos meios para antigas fontes**: Sintaxe Diacrônica em corpus eletrônico do português. Projeto de Pesquisa. UESB, Vitória da Conquista, 2010.

NAMIUTI, C. (Coord.); SANTOS, J. V. (Co-coordenador) **Memória Conquistense:** implementação de um corpus digital. CNPq 485098/2013-0. UESB, Vitória da Conquista, 2013. (Projeto de Pesquisa).

ORACLE. **JDK 1.1 for Solaris Developer's Guide**. Java Programming Environment and the Java Runtime Environment (JRE). 2010. Disponível em: <http://docs.oracle.com/cd/E19455-01/806-3461/6jck06gqd/index.html>.Acesso em: 14 out. 2013.

PAIXÃO DE SOUSA, M. C.; KEPLER, F. N.; FARIA, P. P. E-Dictor: Novas perspectivas na codificação e edição de corpora de textos históricos. In: Tania Shepherd; Tony Berber Sardinha; Marcia Veirano Pinto. (Org.). **Caminhos da Linguística de Corpus**. Campinas: Mercado de Letras, 2010.

PAIXÃO DE SOUSA, M.C. Memórias do Texto. **Revista Texto Digital**, n.2., 2006. Disponível em: <http://www.textodigital.ufsc.br/num02/paixao.htm>. Acesso em: 01 out 2012.

SANTORINI, B. **Annotation manual for the Penn Historical *Corpora* and the PCEEC**. Disponível em: <http://www.ling.u*Penn*.edu/hist-*corpora*/annotation/index.html>. 2010. Acesso em: 08 out 2013.

SILVA FILHO, A. M. **Programando com XML**. Rio de Janeiro: Elsevier, 2004.

UNICAMP. **Padrões Rítmicos, Fixação de Parâmetros & Mudança Linguística**. 1998. Disponível em:<http://www.tycho.iel.unicamp.br/~tycho/prfpml/fase2/index.html> Acesso em: 31 jul. 2014.

_____. **Corpus do Português Histórico Tycho Brahe.** Manual de Preparação dos Textos. Sistema de Edições Eletrônicas do corpus *Tycho Brahe*. Campinas, 2007. Disponível em: < http://www.tycho.iel.unicamp.br/~tycho/corpus/manual/prep/manual_frameset.html>. Acesso em: 5 ago 2014.

W3C. **XML Technology**. 2010. Disponível em: < http://www.w3.org/standards/xml/> Acesso em 08 de outubro de 2013.

W3C. **XML Path Language (**XPath**)**. 1999. Disponível em: <http://www.w3.org/TR/XPath/>. Acesso em 08 de outubro de 2013.

YAO, X.; BORISOVA, I.; ALAM, M. **PDTB XML: the XMLization of the *Penn* Discourse *TreeBank*** 2.0. 2010. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2010/summaries/336.html>. Acesso em: 12 out 2013.

Artigo recebido em: 30.09.2014
Artigo aprovado em: 23.11.2014

Epistemic modality through the use of adverbs: a corpus-based study on learners' written discourse

Modalidade epistêmica por meio do uso de advérbios: um estudo baseado em *corpora* sobre o discurso escrito de aprendizes

Adriana Tenuta[*]
Ana Larissa A. M. Oliveira[**]
Bárbara Malveira Orfanó[***]

ABSTRACT: This paper discusses the grammatical category of modality and the variety of linguistic resources available for the expression of it, and presents a research that aimed at analyzing the expression of modality through the use of adverbs in academic writing. More specifically, the study presented investigated how Brazilian learners of English express modality through adverbs in their academic essays. Two corpora were used: a sub-corpus taken from the corpus of Brazilian Learners of English (CABrI) and another sub-corpus taken from the Louvain corpus of Native English Essays (LOCNESS). The prevalent adverbial items found in both corpora were identified and described, using corpus Linguistics tools. The analysis conducted revealed the rigidity of the expression of modality through adverbs in the learners' written discourse as opposed to a more varied way of this expression in the native speakers' data. This paper also discusses the way native speakers and learners differ in their written production and the possible pedagogical implications of these findings.

KEYWORDS: Corpora. Modality. Adverbs. Syntax. Learners' writing.

RESUMO: Este artigo discute a categoria gramatical da modalidade e a variedade de recursos linguísticos disponíveis para a expressão desta. O artigo apresenta uma pesquisa que teve como objetivo analisar a expressão da modalidade por meio do uso de advérbios na escrita de aprendizes. Mais especificamente, o estudo apresentado investigou como aprendizes brasileiros de Inglês expressam modalidade por meio de advérbios em redações produzidas em ambiente acadêmico. Foram utilizados dois *corpora*: um *sub-corpus*, retirado do *corpus* de Aprendizes Brasileiros de Inglês (Cabri) e outro *sub-corpus*, retirado do *corpus* Louvain de Redações em Inglês escritas por nativos (LOCNESS). Os itens adverbiais prevalentes em ambos os *corpora* foram identificados e descritos, usando ferramentas de Linguística de *corpus*. A análise realizada revelou a rigidez da expressão da modalidade por meio de advérbios na escrita dos aprendizes, bem como uma forma mais variada desta expressão nos dados de falantes nativos. Este artigo discute, ainda, a forma como os falantes nativos e os aprendizes diferem em sua produção escrita e as possíveis implicações pedagógicas desses resultados.

PALAVRAS-CHAVE: *Corpora*. Modalidade. Advérbios. Sintaxe. Escrita de aprendizes.

[*] Universidade Federal de Minas Gerais.
[**] Universidade Federal de Minas Gerais.
[***] Universidade Federal de São João del Rei.

1. Introduction

This paper reports on a study that investigated, using corpora, how Brazilian learners of English express modality through adverbs in the production of essays. Corpus-based studies on learners' production of written discourse have caught the attention of many researchers from different domains. Despite the difficulties in compiling and analyzing students' production, recent findings have contributed to the understanding of these students' interlanguage by identifying linguistic features that are prevalent in their discourse (BERBER-SARDINHA; SHEPERD, 2008; DUTRA, 2009).

Following the Hallidayan model (HALLIDAY, 2004), modality conveys stance and attitude of the sender of a message. In this study, then, we use a learners' corpus, aiming at identifying how Brazilian learners of English express stance and attitude by employing modality elements containing an adverb in their academic writing. We shall compare their production to that of native speakers of English in the same setting, that is, in the academic writing scenario. By identifying the most used adverbs in the expression of modality in learners' essays, we may have a better account of these speakers' expressions of stance and attitude. For that purpose, two corpora were investigated: our reference corpus, CABrI (Corpus of Brazilian English Learners, in construction – BERBER-SARDINHA, 2001; DUTRA, 2009), and LOCNESS (Louvain Corpus of Native English Essays - GRANGER; DAGNEAUX; MEUNIER; PAQUOT, 2009). We believe that such an approach to the study of modality in English can contribute to the emerging area of corpora as well as to the study of syntax.

The structure of this paper is the following: introduction, literature review and theoretical framework, analysis, and conclusion.

2. Theoretical framework

For a better understanding of how modality is conceptualized, we have structured the theoretical framework as follows: (1) outlining the main characteristics of modality, (2) describing the realization of modality in English and (3) discussing the interface between corpora and grammar in academic writing by reviewing previous research in the area.

2.1 The Expression of modal values in English

Mood and other modality resources are means for the expression of the speaker's attitude or commitment regarding the content of a proposition (PALMER, 1974). According to Palmer, mood is realized by verbal morphology, whereas modality is a feature related to a variety of linguistic phenomena, as described by Downing and Locke (2006), among which modal verbs play a central role. Modality is to be understood as a grammatical category that covers notions such as possibility, probability, necessity, volition, obligation and permission.

Modality, therefore, can be connected to basic logical meanings, categorized under a few types: (a) epistemic, (b) deontic and (c) dynamic (DOWNING; LOCKE, 2006), of which the first two (epistemic and deontic) are the central ones.

Epistemic modality is the expression of the various degrees of certainty/uncertainty about facts, events, situations, and, thus, it is related to limitations on the speaker's knowledge about these same facts, events, situations. Consequently, epistemic modality refers to meanings related to inference, prediction, expectation, and probability (BIBER: 1999; DOWNING; LOCKE, 2006). Epistemic modality expressed through different means is illustrated below:

1. *It might rain tomorrow.* (modality realized by a modal verb)
I expect that he be happy. (modality realized by *expect* + an embedded clause in the subjunctive)
It's very unlikely that they will accept our offer. (modality realized by an adverb)

Deontic modality, on the other hand, refers to meanings such as permission and obligation of various kinds, ranging from very strong to a milder obligation. Thus, deontic modality, differently from epistemic, is, associated with authority and judgment, rather than with knowledge or prediction. For this reason, deontic modality comprises language resources used to influence people to do (or not to do) things, whereas epistemic modality is used to express what speakers think is likely to happen.

In spite of the fact that epistemic and deontic meanings are different, the same modal verbs can be used in the expression of one or the other, depending on the context given.

2. *It must have been him*. (epistemic)
You must leave now. (deontic)

Additionally, on many occasions, it is necessary to consider the context of use for the precise interpretation of modality meanings. This is the case of the example below, in which *must* can express either the epistemic meaning of prediction or possibility (contextualized as: *I assume you are patient, given certain evidences*) or the deontic meaning of obligation or necessity (contextualized as: *there is a need for you to be very patient, according to my understanding of the situation*).

3. *You must be very patient*.

Although modality is centrally related to epistemic or deontic meanings, as we have stated, there are also other kinds of meanings associated with modality, all of them, however, play a more peripheral role in syntax and are grouped under the label *dynamic.*

These dynamic meanings are described as ability and courage (DOWNING; LOCKE, 2006) and ability, volition and courage (HUDDLESTON; PULLUM, 2005). They are often expressed by modal verbs, like *can* and *will,* and by semi-modals, like *dare*.

Some examples of dynamic modality are displayed below:

4. *I can speak Spanish*. (ability)
I daren' t say this. (courage)

In certain cases, we can interpret the same occurrence as dynamic and as epistemic, since both types of meanings can be identified in the occurrence:

5. You can 't be right. (probability and/or ability)
She can play the piano. (possibility and/or ability)
I can speak four languages. (possibility and/or ability)

Modality also conveys meanings related to the concept of remoteness, illustrated in the examples below by Huddleston and Pullum (2005).

6. *If she liked the place, she would have stayed*. (remote)

There are authors that group modality meanings differently. Biber (1999), in this corpus-based reference grammar, identifies three categories of modal verbs: (a) permission/possibility/ability - *can, could, may, might*; (b) obligation/necessity - *must, should, (had) better, have (got) to, need to, ought to, be supposed to;* and (c) volition/prediction - *will,*

would, shall, be going to. This categorization does not correspond exactly to the distinction deontic/epistemic adopted in this work.

From the perspective of Biber (1999), modals are divided into three groups, namely, modals, marginal auxiliary verbs and semi-modals. The first group encompasses *can, could, may, might, shall, should, will, would* and *must*. These modals have a number of specific features, such as (a) being invariant forms, (b) preceding the subject in yes-no questions and (c) being followed by a verb in the bare infinitive. Marginal auxiliary verbs correspond to *need (to), ought to, dare (to)* and *used to*. According to Biber (1999), these marginal auxiliary verbs are rare and occur almost only in British English. Fixed idiomatic phrases, such as *(had) better, have to, (have) got to, be supposed to* and *be going to,* are called semi-modals by Biber (1999). Semi-modals differ from central modals because they can be marked for both tense and person. Besides, they can also occur as non-finite forms.

Taking a multi-dimensional feature of modality into account, Carter and McCarthy (2006) state that the best candidates for modality meanings is the closed class of modal verbs, but the 'list' contains others, which are very high in frequency and carry related meanings. These include lexical modals, such as the verbs *look, seem* and *sound*; the adjectives *possible* and *certain,* and the adverbs *maybe*, *probably, definitely, apparently,* and *possibly*. The authors also suggest that the domain of modality needs to be expanded beyond the closed class of modal verbs, which is not a new idea, and they provide compelling evidence of the ubiquity of modality items in everyday spoken and written discourse.

The varied expression of modality is presented below, based on Downing and Locke (2006), considering two basic situations: modality expressed in the verb group and expressed elsewhere in the clause

When expressed in the VG, modality can be realized by:

- The modal verbs *may, might, should, must, can, would, will, ought to, shall, could, need*; the semi-modals (modals in certain uses): *need, dare, wish*
- The lexical auxiliaries (chain-like structures with primary verbs *be* and *have*): *be able to, be apt to, be due to, be going to, be liable to, be likely to, be certain to, be sure to, be to, be unlikely to, be supposed to, have to, have got to, had better, would rather, would sooner.*
- The phased structures composed of: *need, want, regret, try, manage, hesitate, happen, chance, tend, seem, appear, pretend* (in any tense) + a V in *-ing* or

infinitive; subjunctive forms in embedded clauses, introduced by verbs such as: *expect, suppose, recommend, require, request, suspect, intend, think, guess, assume.*

- The lexical verbs such as *allow, beg, command, forbid, guarantee, guess, promise, suggest, warn.*.
- The imperative forms.
- The past tense to indicate remoteness from reality, as in *I thought I'd go along with you, if you don't mind;* and conditional structures, as in *If you went, I would go too.*

Modality expressed elsewhere in the clause, may be found in adverbs and sentence modifiers: *maybe*, *supposedly*, *perhaps*, *possibly*; predicate adjectives: *possible, impossible, likely, conceivable, doubtful, certain, sure, positive* and nouns such as *possibility, probability, chance, likelihood*.

In contrast with modal verbs, adverbs, which are the focus of this study, are numerous in the area of certainty expression. According to Chafe (1986), all adverbs can be considered *evidential*[1] and can be classified along dimensions such as reliability and degree of expectation. *Certainly*, for example, indicates that the speaker expresses his or her assessment of the proposition. *Obviously* and *clearly* could be classified as markers of induction and of courses expresses that something is in line with expectations.

Modality can also be expressed in different points of the clause, concomitantly. Downing and Locke (2006) refer to this realization as modal harmony. According to them, modal harmony can be illustrated by the following example:

7. *I doubt she could possibly have said that.*

2.2 Corpora and grammar

The study of grammar is relatively recent in Corpus Linguistics, since the lexicon used to be the unit of investigation by excellence in early corpora studies. However, advances in

[1] In a broad definition of evidentiality, epistemic modality is a subcategory of evidentiality, which is neither marked for the mode of knowing nor for the source of knowledge and therefore is distinct from evidentiality in a narrow sense. In English, all evidential adverbs are modal.

automatic tagging and parsing, as well as the appearance of reasonably sized corpora containing detailed grammatical annotation have progressively enabled corpus linguists to shift their attention towards genuinely grammatical issues.

Fulfilling the main objective of this paper, we shall discuss relevant corpus based research in academic writing.

2.3 Empiric research on academic writing and modality from a Corpus Linguistics perspective

In the past few years, research on academic discourse has flourished promoting an interesting debate on the language of higher education. Many studies have concentrated on different aspects of how both native speakers and non-native speakers organize their essays.

Following a topic-oriented approach to academic writing, Hinkel (1995) showed how topics could influence the use and distribution of certain modal expressions in learners' production. In essays written in the discipline of Education, Chinese, Japanese and Vietnamese learners overused *must* and *should*. Native speakers, on the other hand, did not employ *must* when talking about political and educational issues. Hoye (1997) conducted an experiment with Spanish learners of English and native speakers of English on how modal values are expressed in English. The author designed a series of tests in which the participants were required to fill in gaps in a text with modals and/or adverbs. He observed that learners had trouble combining modal and adverbs to express attitude. Learners did not perceive the combinatory potential of modal expressions, which in turn hindered their performance in academic writing.

Biber (1999, 2001, and 2004) has contributed immensely to our understanding of academic discourse. Following a frequency-driven approach, Biber and his colleagues identified the most frequent bundles in academic discourse mainly describing their grammatical characteristics.

Hyland (2008) explores the structure and function of four word-bundles in a corpus of academic discourse. The data for his study consists of three corpora (research articles, doctoral thesis and master's dissertations) comprising 3.5 million words. The author suggests that the presence of certain bundles, for example, *as a result of* can help identify different text genres. The results indicated that students draw on different resources to develop their arguments. There were fewer lexical bundles in doctoral thesis, while in master's

dissertations, a greater number of bundles were found. This might be an indication that less proficient students rely more on formulaic expressions due to restrict vocabulary.

In another study, focusing on the field of second language academic writing, Hyland and Milton (2010) show how L2 writers differ significantly from native speakers in that the former group relies on a more limited range of items, offering stronger commitments, and exhibiting greater problems in conveying a precise degree of certainty. The authors also found that, in comparison to the L1 writers, the L2 writers rely on a more limited range of grammatical resources, including particular modal verbs and the expression *I think*. The authors posit that this distinction may be a consequence of non-native speakers´ limited language repertoire, which does not enable them to adjust different levels of stance when building up their writing.

In a similar direction and adopting a frequency-driven approach, Chen and Baker (2010) identify the most frequent lexical bundles in three corpora: a) a sub-corpus from FLOB (academic prose section), b) BAWE-CH (Chinese students of English), and c) BAWE- EN (English students). Their comparative study showed that there are differences and similarities between native speaker and learner academic writing. The use of lexical bundles in non-native and native student essays, for example, is very similar: from a structural point of view, they both have more verb phrase based bundles and discourse organizers than native expert writing whereas native professional writers exhibit a wider range of noun phrase based bundles and referential markers.

Following a pragmatic-functional approach, Simpson-Vlach and Ellis (2010) also looked at the most common lexical bundles in academic discourse, focusing on both oral and written corpora. They used the Michigan Corpus of Academic Spoken English (MICASE) and the oral academic part of the British National Corpus (BNC) also including in their research the Hyland Corpus (2004) and the written BNC files of various academic subjects. They extracted three and four word *n-grams* and had ESP instructors judge if the lexical bundles were chunks, if they had a function or if they were expressions that were worth teaching. As a result, they proposed the Academic Formulas List (AFL) with 435 lexical bundles distributed in 18 subcategories. However, research on the expression of modality in academic writing is scarce: the next paragraphs will outline previous studies in the area that prove to be significant to our analysis.

3. Data and Methodology

As we have already mentioned, this study uses two corpora: a sub-corpus taken from the corpus of Brazilian Learners of English (CABrI) and another sub-corpus taken from the Louvain Corpus of Native English Essays (LOCNESS). To CABrI is composed of academic essays written by advanced undergraduate students (B1 to C1 according to the Common European Framework) from the Language Course at Universidade Federal de Minas Gerais. Students from the Liberal Arts course are asked to write argumentative essays ranging from 300 to 500 hundred words. Students have to choose from 13 different titles such as *Crime does not pay* and *Feminists have done more harm to the cause of women than good* to write their essays. Learners' writings are converted to text files and stored, so corpus methodological tools can be employed for analysis. The texts chosen to compose the sub-corpus belong to the American argumentative section. In total, the LOCNESS sub-corpus used in this study contains 60,241 words. Now, CABrI contains around 36,187 words. The Louvain Corpus of Native English Essays (LOCNESS) presents essays written by American and British speakers, ranging from academic to literary texts.

For this analysis, first, word lists (frequency lists) were generated and adverbs with the potential to function as modal items were isolated. This procedure, according to O'Keeffe, McCarthy and Carter (2007), proves to be essential in identifying the core vocabulary of English, and it is considered a good resource for pedagogical purposes, which is one of the aims of this study. Comparing frequency lists is, then, an appropriate starting point; however, relying only on frequency lists would not be sufficient. For that reason, in order to get a better notion of the pragmatic function of the adverbs involved in the construction of modality in the essays under investigation, the next step was to compile lexical bundle lists containing an adverb with a modal function. After identifying the most common items in this list, concordance lines were analyzed so that the bundles containing modality adverb could be observed in their particular contexts through manual search of the data.

4. Analysis

We started analyzing the most frequent adverbs found in the learner corpus in order to verify their function in the data. These results are presented in the table below.

Table 1- Most frequent adverbs found in CABrI

Item	Raw Freq.	Freq. per million words
probably	13	472
certainly	9	326
maybe	9	326
likely	9	326
simply	7	254
unfortunately	5	181
actually	5	181
Total	44	1,594

As the main aim of this paper is to compare learners' data with native speakers' production, we shall analyze LOCNESS for corresponding results. In table 2, then, we present the most frequent adverbs found in LOCNESS.

Table 2- Most frequent adverbs found in LOCNESS

Items	Raw Freq.	Freq. per million words
likely	27	448
certainly	19	315
probably	17	282
perhaps	16	265
surely	13	215
possibly	12	199
maybe	11	182
unlikely	7	116
Total	122	1,952

Confronting tables 1 and 2, we observe that native speakers seem to use more adverbs than non-native speakers do. In addition, when analyzing the frequency of adverbs in each group, we can observe that the native speakers' use of adverbs is more evenly distributed than that of learners. This first finding is in line with Holmes (1998), is that it provides evidence of a more balanced use of linguistic resources by native speakers to express modality. Throughout this paper, we will make the case that learners tend to use a more fixed set of expressions to convey modality. In table 3, we contrast the adverbs found in both corpora, aiming to determine differences and/or similarities in them.

Table 3- Distribution of adverbs in the two corpora (raw results)

Items	LOCNESS	CABrI
Likely	27	9
Certainly	19	9
Probably	17	13
Perhaps	16	4
Surely	13	1
Possibly	12	0
Maybe	11	9
Unlikely	7	0

Complementing the quantitative analysis, we also submitted the data to statistic tests, believing that they would strengthen the argument that there is a striking difference between native speaker and learner corpora in the use of adverbs expressing modality. This difference is outlined in table 4.

Table 4 - Expected contingency table between both corpora

Adverbs	NS	NNS
Likely	26.3	9.70
Certainly	20.5	7.54
Probably	21.9	8.08
Perhaps	14.6	5.39
Surely	10.2	3.77
Possibly	8.77	3.23
Maybe	14.6	5.39
Unlikely	5.11	1.89

Chi-square = 18.1
degrees of freedom = 7
probability = 0.011

Based on the results of the statistical test, we can affirm that the difference in the use of adverbs in the corpora analyzed is significant. The fact that the *p value* is zero reinforces the claim that there is a striking difference between the corpora analyzed, as the results of the *chi square* (*chi square* = 18.1) have proved. From this preliminary analysis, one can speculate that the use of adverbs to express modality is underrepresented in learners´ academic essays, as stated in a research carried out by Tenuta, Oliveira, Orfanó (in press), which has shown that Brazilian learners rely on a rigid set of verbs to express modal values. However, the analysis proposed here intends to go beyond stating statistical differences. In fact, it aims at understanding the linguistic features that make up for these differences and their implications

for the learners' written discourse production. In order to do so, at this point of the analysis, we shall concentrate on the most common bundles containing an adverb in CABrI and LOCNESS. The following table shows this distribution in both corpora.

Table 5- Distribution of bundles containing an adverb in the data

Bundles	CABrI- raw freq	Freq. per m words	LOCNESS-raw frequency	Freq. per m words
likely to	9	33	16	265
Almost certainly	0	0	4	66
certainly not	0	0	4	66
will probably	9	32	0	0
would probably	0	0	6	99

Submitting the data to a *chi-square* test, we found that the difference between the corpora is significant, as illustrated below. Observing the *p value* from the results from table 6, one can verify that when looking, in particular, at bundles containing an adverb, the difference between the two corpora is significant. The items were analyzed and compared with results from LOCNESS as follows:

Table 6 - Expected contingency table

Bundles	CABrI	LOCNESS
Likely to	9.38	15.6
almost certainly	1.50	2.50
certainly not	1.50	2.50
will probably	3.38	5.62
Would probably	2.25	3.75

Chi-square = 23.4
degrees of freedom = 4
probability = 0.000

After identifying the most common bundles in both corpora, we focused on the analysis of each bundle independently. In CABrI, there are only two bundles being used, whereas in LOCNESS, we found four bundles. This fact reinforces the claim that native speakers express modality not only by using different adverbs, but also by combining them in different bundles. The only two bundles used by learners in the corpus analyzed were *likely to* and *will probably*.

4.1 *Likely to*

There are 33 occurrences of the bundle in CABrI and 265 in LOCNESS. This difference was expected since, as previously mentioned, there is lower frequency of adverbs expressing modality in the learners' productions. We considered important to speculate on this situation focusing on the learners' written discourse, bearing in mind that frequency differences between two datasets can indicate either overuse or underuse of linguistic features, which poses interesting pedagogical issues involving the teaching and learning of English.

> Extract 1 - Example of *likely to* from LOCNESS[2]
> *Now that rail privatisation has gone ahead, many people are likely to lose faith in trains, due to the percieved inefficiency of the operators (for example the time-table book full of errors or the recent survey in Which? magazine about overcharging). Fares are likely to increase, and many rural lines that used to be subsidised by the government face closure.*

Fig. 1 shows concordance lines for the bundle *likely to* in CABrI. In all examples, it is possible to see that the epistemic use is prevalent, which might be due to the text genre in focus. In texts of this genre, the writer, very frequently, has to commit him/herself, in different degrees, to the certainty of occurrence of a fact.

N			
1	ished severely and consequently they are more	**likely to**	commit crimes again.
2	country side towns around Brazil we are very	**likely to**	find a huge number of
3	by creating artificial dreams, graduates are	**likely to**	be shocked or unsure in
4	This kind of proficiency is more	**likely to**	be developed if one
5	This kind of proficiency is more	**likely to**	be developed if one
6	arget our limited resources for programs most	**likely to**	reduce recidivism and
7	Not	**likely to**	happen.
8	y and dreams be profitable, but they are also	**likely to**	be crucial ways to make
9	ot prepare graduate students to what they are	**likely to**	face in real life.
10	nt agencies show that such tragedies are more	**likely to**	occur to young adults

Fig. 1- Concordance lines for the bundle *likely to* in CABrI.

The function of this bundle is similar in both corpora; however, the frequency is significantly higher in the native speaker corpus. *Likely to* is the bundle containing an adverb preferred by native speakers to express possibility/probability, while English learners expressed possibility/probability by using the bundle *will probably*.

2 All extracts from the learner corpus have been preserved as written by students. Hence, corrections by any type are not included in this study.

The low frequency of modal adverbs in the learner´s academic writing was previously acknowledged by Tenuta, Orfanó, Oliveira (in press). The authors showed that learners seemed to rely on modal verbs to express epistemic modality. However, in the corpus investigated, learners used a very narrow range of modal verbs with epistemic meaning, for example, mainly *can* and *will*. In this regard, the findings revealed that the distribution of modals in the native corpus was more varied, since native speakers employed, for example, *should*, *could* and *would*.

4.2 Will probably

This bundle follows the pattern *will probably* + verb.[3] We tend to conclude that the bundle *will probably* might be more easily accessed by learners, becoming active in discourse through less mental effort (CHAFE: 1994), mainly because *will probably* is lexically and structurally closer to the Portuguese language. In this study, learners used *will probably* three times more than native speakers did.

> Extract 2- Example from CABrI
> *There will probably be many reasons for dreaming and three possible - and believable – ones could be its profitability (for the enterprises which provide entertainment, for example), its help in making us stand and try to change our stressful reality, and the health benefits it provides us. Oddly enough, imagining can make a big profit from generating – and selling – brilliant ideas.*

N	Concordance
1	is. All you have to do is read and read. In fact, you **will probably** get all
2	e she does not recognize. As for Dee's sister, she **will probably** feel
3	he references (authors, books, concepts, etc) that **will probably** help
4	e way she could free herself from her family. They **will probably** go along
5	no chance to see in loco how things really work, it **will probably** be very
6	e inhabit a modern and industrialised world. There **will probably** be many
7	ity degrees, such as philosophy and anthropology, **will probably** have
8	ity degrees, such as philosophy and anthropology, **will probably** have
9	er mother, Dee is not the same anymore and they **will probably** be apart

Fig. 2- Concordance lines for the bundle *will probably* in CABrI

In the next sections, we analyzed the bundles found in LOCNESS.

[3] We also observed that there was a high frequency in the use of *will probably* + a linking verb, since a third of the occurrences with the bundle *will probably* followed this pattern. This finding will be further investigated.

4.3 Would probably

The use of the bundle w*ould probably* in LOCNESS might reflect a feature more closely related to formal register, considered adequate to the essay genre. We might speculate that learners do not use it because they might not be very proficient in the use of linguistic features that would enable them to express a degree of both certainty and formality in written discourse, as seen in the concordance lines below. We could argue that the bundle has a twofold discursive function, which was employed by native speakers and ignored by learners in the corpora analyzed.

> Extract 3- Example of *would probably* from LOCNESS
> *If boxing was made illegal, then the sport would still take place, but it would not be under the control of doctors or referees and there would be no proper association to represent rules and regulations. It would probably become even more dangerous. Of course there are advantages banning the sport but they are very small compared to the disadvantages.*

1	By banning fox hunting the law	**would probably** drive the spor
2	ulations showed that charities	**would probably** lose out, coll
3	primary school and her mother	**would probably** die while the
4	It	**would probably** become even mo
5	everal thousands of pounds and	**would probably** have to sell u
6	were the highest bider as they	**would probably** hide it as it

Fig. 3- Concordance lines for the bundle *would probably* in LOCNESS.

5. Conclusion

In this study, we analyzed epistemic modality, which was a major feature in academic essays of both native speakers and learners of English. The choice for epistemic modality is justified by to the fact that this text genre favors this kind of modal expression, used to mitigate the level of commitment of the speaker to the certainty of the facts being stated.

Our findings indicate that the expression of epistemic modality in the native speakers' written production of academic essays, referring to, specifically, the use of adverbs, seems to be more varied than in the learner´s production. This can be evidenced, for example, by the non-use, from the part of learners, of expressions such as like *likely to*, which proved to be widely used by native speakers. On the same token, the data analyzed show that the learners' preferred expression, including adverb to express epistemic modality was *will probably*. These findings strengthen the argument that teachers and material developers should introduce students to a more varied set of vocabulary as well as structural choices, so that

their writing skills can be improved towards the production of texts which are better elaborated and more responsive to the grammatical and semantic, but also to the pragmatic demands of genre.

Another important remark about the learners' written production concerns the fact that, in this study, learners seemed to ignore the form *would probably*, which is highly used by native speakers as a way to express remoteness and little commitment to the certainty of facts. This finding also supports the claim that learners tend to rely on a very rigid set of structures to realize modality. It also poses the issue of the role of language instruction and material design to raise students' awareness of how they can structure their discourse in order to comply successfully with the requirements of academic writing.

We would also like to comment on the importance of analyzing empirical language data for a broader understanding of how native speakers and learners can differ in their production and, if that is the case, decide on the best teaching strategies to help learners write more fluently and effectively.

Also, from a pedagogical perspective, research studies like the one conducted by Holmes (1998) have shown that, in general, the use of modality in English has been underrepresented in teaching materials, since this grammatical category is often dealt with through the restricted use of modal verbs. The present study argues along the same lines, as it also shows that students seem to rely on modal verbs instead of using a variety of modal forms, which is present in the native speakers' production.

Finally, we consider it is important that instructors provide students with opportunities to engage in reflection on their own stance, using strategies while writing, as it has also been postulated by Silver (2013), from a critical perspective.

References

BERBER SARDINHA, A.P. O *corpus* de aprendizes Br-ICLE. **Intercâmbio**, v.10, 2001, p. 227-239.

BERBER SARDINHA, A. P. **Linguística de Corpus**. Barueri: Manole, 004.

BERBER SARDINHA, A. P.; SHEPERD, T. An online system for error identification in Brazilian learner English. Anais do **8th Teaching and Language Corpora Conference**. Lisboa: Associação de Estudos e de Investigação Científica do ISLA-Lisboa, 2008, p. 257-262.

BIBER, D. et al. **Longman Grammar of Spoken and Written English**, 1999.

BIBER, D. Stance in spoken and written university registers. **Journal of English for Academic Purposes**, 5, 2006, 97–116. crossref http://dx.doi.org/10.1016/j.jeap.2006.05.001

BIBER, D.; CONRAD, S.; LEECH, G. **Longman Student Grammar of Spoken and Written English**. Essex: Longman, 2002.

CHAFE, W. The nature of consciousness. In: **Discourse, consciousness and time**. Chicago & London: The University of Chicago Press, 1994, pp. 26-40.

DOWNING, A.; LOCKE, P. **English Grammar***:* a University Course. New York: Routledge, 2006, second edition

DUTRA, D. P. Conscientização linguística através de corpora online. **Caderno de resumos do 17 INPLA**. São Paulo: PUC-SP, 2009, p. 62.

DUTRA S.; DAGNEAUX, E.; MEUNIER, F.; PAQUOT, M. **International corpus of Learner English-Version2**. Louvain-La-Neuve: UCL Presses Universitaires de Louvain, 2009.

HALLIDAY, M. A. K.; HASAN, R. **Cohesion in English**. London: Longman, 1976.

HALLIDAY, M. A. K. **An Introduction To Functional Grammar**. London: Hodder Arnold Publication, Revised Edition, 2004.

HOLMES, R. Genre analysis, and the social sciences: An investigation of the structure of research article discussion sections in three disciplines. In **English for Specific Purposes**, 2007, 16: 321-3. crossref http://dx.doi.org/10.1016/S0889-4906(96)00038-5

HOYE, L. **Adverbs and Modality in English**. London & New York: Longman, 1997.

HYLAND, K.; MILTON, J. Qualifications and certainty in L1 and L2 students' writing. **Journal of Second Language Writing**, 6(2), 1997, pp. 183-205. crossref http://dx.doi.org/10.1016/S1060-3743(97)90033-3

O'KEEFFE, A.; MCCARTHY, M.; CARTER, R. **From corpus to classroom**: Language use and language teaching. Randi Reppen, Northern Arizona University, 2007. crossref http://dx.doi.org/10.1017/CBO9780511497650

PALMER, F. R. **The English verb**. London: Longman, 1974.

SILVER, M. The stance of stance: a critical look at ways stance is expressed and modeled in academic discourse. **Journal of English for Academic Purposes**, 2003, vol. 2, issue 4: 359-374. crossref http://dx.doi.org/10.1016/S1475-1585(03)00051-1

TENUTA, A. M. **Tempo, modo e aspecto verbal na estruturação do discurso narrativo**. Dissertação (Mestrado em Linguística). Belo Horizonte: UFMG, 1992.

TENUTA, A. M. A., OLIVEIRA, A. L. A. M., ORFANÒ B. M. How Brazilian learners express modality in their writing: a corpus-based study on lexical bundles. São Paulo: **Revista Intercâmbio**, no prelo.

Artigo recebido em: 07.10.2014
Artigo aprovado em: 11.12.2014

Do Português Clássico ao Português Europeu Moderno: o mapeamento do artigo

From Classic Portuguese to Modern European Portuguese: definite article description

Simone Floripi [*]

RESUMO: No decorrer dos séculos os padrões de aplicação do artigo na língua portuguesa foram sendo modificados, demonstrando ter havido uma mudança na gramática dessa língua. Tal mudança culmina na obrigatoriedade de utilização do artigo frente a sintagmas nominais possessivos, como observado atualmente no português europeu moderno. Diante deste panorama, buscamos investigar o uso do determinante em DPs possessivos desde o século 16 ao século 19 (Português Clássico) e demonstrar sua evolução no tempo. Foi realizado um estudo abrangente que considera fatores sintáticos e estruturais capazes de indicar o início da variação na colocação de determinantes e quando esta deixou de existir, momento no qual, o uso do artigo diante de possessivo passou a ser obrigatório no português europeu (cf. Floripi, 2008 e Castro, 2000). Apresentaremos os resultados do mapeamento geral dos contextos capazes de evidenciar a mudança sintática ocorrida e os fatores estruturais responsáveis pelo desencadeamento dessa mudança na gramática da língua portuguesa. Para tanto, foi realizada uma investigação diacrônica dos dados, por meio da quantificação destes, analisando-os sob uma abordagem minimalista (Chomsky (1995) e Kayne (1994)), tendo como pressupostos teóricos o Modelo de Princípios e Parâmetros.

PALAVRAS-CHAVE: Mudança linguística. Língua portuguesa. Artigo definido. Pronome possessivo. *Corpus* linguístico eletrônico.

ABSTRACT: Throughout the centuries, the use of definite article in Portuguese has been changed, indicating a linguistic change on its grammar. This grammar changed to a pattern that obligate the employ of the article in possessive nominal phrases as we can note in Modern European Portuguese. Considering this, we propose to investigate the employ of the article since century 16 to 19 in order to show it development during the centuries. We made a robust investigation which considers syntactic and structural factors that can determine the beginning of the process of variation and the final pattern with the requirement of definite article on the referred structure (cf. Floripi, 2008 e Castro, 2000). We will discuss the results, showing the contexts in which occurs the syntactic change. We made a diachronically view considering the Minimalist assumptions (Chomsky (1995) e Kayne (1994)) of Principles and Parameters Theory.

KEYWORDS: Linguistic change. Portuguese. Definite article. Possessive pronoun. Eletronic linguistic corpus.

[*] Fez mestrado e doutorado em Linguística pela Universidade Estadual de Campinas – UNICAMP e pós-doutorado pela Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP. Atualmente é professora em dedicação exclusiva da Universidade Federal de Uberlândia – UFU. Dentre seus interesses de pesquisa estão: sintaxe das línguas naturais, linguística histórica, estudos das tradições discursivas, estudos do texto e formação de professores de língua portuguesa. E-mail: simone.floripi@gmail.com

1. A estrutura do sintagma nominal possessivo

Em diversas línguas afirma-se que os pronomes possessivos estão localizados dentro do sintagma nominal (DP), em posição de núcleo de determinante (D) ou em especificador (Spec,DP) (cf. OLSEN, 1989; DEMSKE, 1995, entre outros). Tal posição comprova que em certas línguas como Inglês, Holandês e Francês o possessivo não co-ocorre com o determinante, conforme apresentado nos exemplos em (1).

(1)
a. *(*The) my book* Inglês
b. *(*Het) mijn boek* Holandês
c. *(*Le) mon livre* Francês
d. *O meu livro* Português
'o meu livro'

Contudo, como vemos em (1d) há línguas, como o português, em que é possível esta ocorrência. A explicação para tal fenômeno é que os possessivos não são equivalentes a determinantes definidos mesmo quando eles parecem ocupar a posição de um determinante; Fato que nos mostraria que existem posições específicas na estrutura do DP, uma para o determinante e outra diferente para o possessivo.

Outro caso, apresentado em (2) abaixo, trata da co-ocorrência de um artigo e um possessivo numa elipse de um DP. Como vemos em (2), em algumas línguas há diferentes formas para o possessivo, a depender de sua posição na estrutura, dada a necessidade da concordância da morfologia do possuidor.

(2)

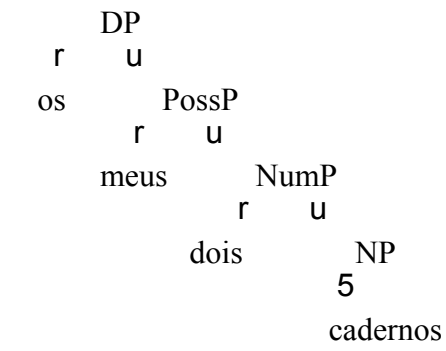
a. **my** book	a'. Your book, not **mine**	Inglês
b. **mon** livre	b'. Ton livre à toi, pas **le mien**	Francês
c. **mein** Buck	c´. Dein Buch, nicht **meins** / das **meinige**	Alemão
'meu livro'	'seu livro, não meu'	

Os exemplos acima também nos levariam a pensar que existem posições específicas para a realização de cada tipo de possessivo a depender de sua posição no DP. Dessa maneira, para indicarmos o posicionamento de cada um dos elementos tratados (determinante e possessivo) abordaremos a estrutura interna do DP que pode ser comparada à estrutura de uma sentença (CP), pois, do mesmo modo que ocorre nos CPs, o elemento possuidor carrega

certos traços que precisam ser checados fora do constituinte onde ele foi gerado como possuidor. Isso requer a existência de projeções funcionais dentro do DP para que os traços de Concordância e Caso possam ser checados.

Vejamos a seguir dados do português contemporâneo, utilizando a estrutura do DP (cf. Shoorlehmer (1998), Szabolcsi (1994), entre outros).

(3)
os meus dois cadernos

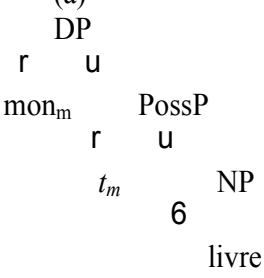


Os pronomes possessivos, por exemplo, podem ser inseridos em posição de núcleo, como Poss^0. Essa arquitetura captura para o português, a instanciação do pronome possessivo entre D e Num e é em seus domínios de checagem que um dado constituinte vai ser interpretado como o possuidor de N.

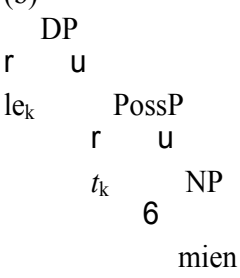
Uma vez apresentada a estrutura do DP, vale dizer que a linha na qual que pretendemos trabalhar enfoca as características do pronome possessivo para explicarmos a possibilidade da variação no uso do determinante nesse contexto. No intuito de introduzirmos a proposta de análise, imaginemos que no português clássico, assim como no francês atualmente, o pronome 'meu' exercia dois papéis semânticos, mas que no português clássico sua realização fonética era homófona, i.e., com apenas uma forma fonológica.

Assim, conforme assumido por Giorgi & Longobardi (1991) haveria uma distinção entre dois tipos de pronomes possessivos daquela época; sendo que estes poderiam se comportar como determinantes (por exemplo, como ocorre em *mon livre* do francês) ou como adjetivos, em que ele caracterizava-se como um predicado (por exemplo, como ocorre em *le mien* do francês). Apresentamos a ilustração destas duas estruturas, utilizando como exemplo o pronome 'meu' do francês, mas que na verdade este pronome representaria as duas formas semanticamente distintas do pronome 'meu' no português clássico.

(4) (a)



(b)



Segundo as estruturas acima, o pronome possessivo 'meu' no português clássico ocuparia uma posição específica a depender a interpretação semântica que ele fosse exercer. Em (4a) ele receberia um papel de determinante, ao ser alçado para a posição de D e em (4b) permaneceria no NP, exercendo a função de predicado da estrutura.

A hipótese de haver duas estruturas para o pronome 'meu' ainda pode ser fundamentada teoricamente pelo trabalho de Shoorlemmer (1998). A autora assume que os argumentos nominais são gerados na base dentro do NP e podem ser alçados ao Spec,PosP para seu licenciamento formal na busca de checar um traço de definitude, assim como ocorre para o licenciamento formal dos sujeitos, mas sem a necessidade de respeitar o Princípio de Projeção Extendida (EPP).

Shoorlemmer (1998) ainda argumenta que o alçamento leva em consideração um conjunto de diferenças entre línguas que apresentam a propriedade de checar o traço mais definido de construções possessivas em oposição a línguas que não operam desta forma, evidenciando o licenciamento de estruturas em que há a co-ocorrência de pronomes e artigos.

2. Apresentação do *corpus*

Este trabalho consiste no estudo do uso do artigo diante dos sintagmas nominais possessivo do português clássico, utilizando 23 textos de autores portugueses nascidos entre o século XVI e XIX. Baseamo-nos no trabalho de Silva (1982) e Magalhães (2002) sobre o uso de artigo frente a possessivos no português clássico com o intuito de determinar a evolução do uso do determinante no decorrer dos séculos.

Os respectivos autores de cada texto utilizado para a realização deste estudo estão abaixo elencados.

Século 16
Fernão Mendes Pinto (1510-1583) *Perigrinação* (52.555 palavras).
Francisco de Holanda (1517-1584) D*a Pintura Antiga (*52.538 palavras).
Diogo do Couto (1542 - 1606) *Décadas* (selecção, prefácio e notas de Antônio Baião) (47.448 palavras).
Luis de Sousa (1556 - 1632) *A Vida de Frei Bertolameu dos Mártires* (52.928 palavras).
F. Rodrigues Lobo (1579 - 1621) *Côrte na Aldeia e Noites de Inverno* (52.429 palavras).

Século 17
Manuel da Costa (1601 - 1667) *Arte de Furtar* (52.867 palavras).
António Vieira (1608 - 1697) *Sermões* (53.855 palavras).
António Vieira (1608 – 1697) *Cartas* (57.088 palavras).
F. Manuel de Melo (1608 - 1666) *Cartas* (58.070 palavras).
António das Chagas (1*631-1682) Cartas Espirituais* (54.445 palavras).
Manuel Bernardes (1644 - 1710) *Nova Floresta* (52.374 palavras).
J. Cunha Brochado (1651 - 1735) *Cartas* (35.058 palavras).
Maria do Céu (1658-1753) Rellacao da Vida e Morte da Serva de Deos a Veneravel Madre Elenna da Crus (27.410 palavras).
André de Barros (1675-1754) *A Vida do Padre António Vieira* (52.055 palavras).
Alexandre de Gusmão (1675-?) *Cartas* (32.433 palavras).

Século 18
Cavaleiro de Oliveira (1702 – 1783) *Cartas* (51.080 palavras).
Matias Aires (1705 - 1763) *Reflexão sobre a Vaidade dos Homens e Cartas sobre a Fortuna* (56.479 palavras).
Luís António Verney (1713-1792) *Verdadeiro Método de Estudar* (49.335 palavras).
Antonio da Costa (1714-?) Cartas do Abade Antônio da Costa (27.096 palavras).
Correia Garção (1724 - 1772) *Obras Completas* (24.924 palavras).
Marquesa D'Alorna (1750-1839) *Cartas e outros Escritos* (49.512 palavras).
Almeida Garrett (1799-1854) V*iagens na minha terra* (51.784 palavras).

Século 19
Ramalho Ortigão (1836 - 1915) *Cartas a Emília* (32.441 palavras).

Como ferramenta de estudo, para a busca dos contextos a serem investigados, contamos com a etiquetação morfológica dos 23 textos, estando estes alocados no *Corpus Anotado do Português Histórico Tycho Brahe*, disponibilizado na rede internacional de computadores no seguinte endereço www.ime.usp.br/~tycho.[1]

A busca dos contextos estudados resultou em dez tipos de contextos que dependiam:

[1] Esta é uma ferramenta que permite termos uma recuperação rápida e confiável dos dados, além de ser acessível a qualquer pesquisador interessado no estudo de textos do português médio.

(i) da presença ou ausência do artigo,
(ii) da sua posição na oração, e
(ii) da presença da preposição

Apresentamos a seguir os contextos observados:

i) Possessivo em posição inicial **(0 – poss)**;
Seu *corpo foi enterrado o mais solenemente que pôde ser, com grande dor, e sentimento de todos, de que era muito amado, como era razão o fosse um Rei*.
(Couto, 1542)

ii) Possessivo em posição inicial mais o Determinante **(0 – D – poss);**
O seu *pintar é trapos, maçonerias, verduras de campos, sombras de árvores, e rios e pontes, a que chamam paisagens, e muitas feguras para cá e muitas para acolá*. (Holanda, 1517)

iii) Possessivo em posição inicial mais a preposição **(0 – P – poss)**;
De sua *dificultosa conquista, a redução à Fé, empresa digna do grande coração de VIEIRA, e uma de suas maiores façanhas, demos já em separada obra completa relação*. (Barros, 1675)

iv) Possessivo em posição inicial antecedido pela preposição + Determinante **(0–P–D-poss)**
Contra o nosso *parecer, nunca achamos dúvida bastante, contra o dos outros sim.* (Aires, 1705)

v) Possessivo em posição inicial antecedido pela contração da preposição com o Determinante (0 – PD – poss);
No nosso *Evangelho diz o mesmo Senhor: Tunc videbunt: então verão: E aquelle então é agora: aquelle tunc é nunc: Tunc videbunt, et nunc est.*
(Sermões, Vieira, 1608)

vi) Possessivo precedido por um Determinante **(D – poss);**
São Paulo descrevendo este mundo, para nos desaffeiçoar de suas vaidades, diz que é como um theatro, em que as figuras cada uma entra a ***representar o seu*** *papel, e passa: Præterit enim figura hujus mundi*.
(Sermões, Vieira, 1608)

vii) Possessivo precedido por uma preposição **(P- poss);**
Nem cuideis, que vos conheço, quem quer que sois, nem que ponho o dedo ***em vossas*** *couzas em particular: o meu zelo bate só no commum.*
(Manuel Antônio da Costa, 1601)

viii) Possessivo precedido por uma preposição e um Determinante **(P – D -poss)**;
Porém, quanto a mi, o que da tenção dêstes autores convém mais ***com o nosso***

modo de fala, sal quer dizer graça, que é o contrário da frieza e sensaboria.
(Lobo, 1579)

ix) Possessivo precedido pela contração da preposição e Determinante **(PD – poss)** e;
No mês de Maio dos anos do Senhor de mil e quinhentos e catorze, reinando em Portugal el-Rei Dom Manuel, único deste nome, e presidindo na Igreja de Deus o Papa Leão X, pariu Maria Correa um filho, que ***bautizaram na sua*** *igreja e freguesia e chamaram Bertolameu*. (Souza, 1556)

x) Possessivo precedido por outro elemento qualquer que não um Determinante ou uma preposição **(X– poss).**
Agora encomendo eu muito a Vossa Mercê me sofra ***como seu*** *despertador e que se acorde do prometido a Deus*.
(Chagas, 1631)

Conforme podemos notar, o procedimento de busca dos contextos de sintagmas possessivos resultou em diversas possibilidades estruturais a depender das variáveis em observação. Para a investigação desta configuração, preocupamo-nos em observar as características intrínsecas dos sintagmas nominais possessivos com o intuito de compreender os mecanismos engatilhados na mudança sintática observada.

Os resultados obtidos por meio da busca, classificação e quantificação dos dados de sintagmas possessivos evidenciam o comportamento na aplicação do artigo no decorrer dos séculos, conforme discutido a seguir.

3. Delineando os contextos de mudança no português clássico

Ao lidar com os dados observados, trabalhamos com a presença e ausência do determinante em sintagmas possessivos considerando os contextos sintáticos em que foi realizado na sentença. Vale salientar que trabalhamos com um montante de cerca de 10.000 dados no total[2], considerando-se os 10 contextos estruturais apresentados anteriormente o que traz a este trabalho uma segurança na procura de identificar o comportamento da aplicação do determinante nos sintagmas possessivos sob uma perspectiva diacrônica.

Conforme apresentado na forma de gráfico, agrupando os dados em períodos de 50 anos, notamos que o uso do artigo em DPs possessivos revela um comportamento distinto de

[2] Os dados foram quantificados utilizando-se do pacote estatístico GoldVarb 2001 uma vez que lidávamos com um montante alto de ocorrências e com vários fatores estruturais envolvidos.

acordo com o contexto sintático em que é realizado. Vejamos os resultados em posição de sujeito, objeto direto, objeto indireto e adjuntos, apresentados nos gráficos abaixo elencados[3].

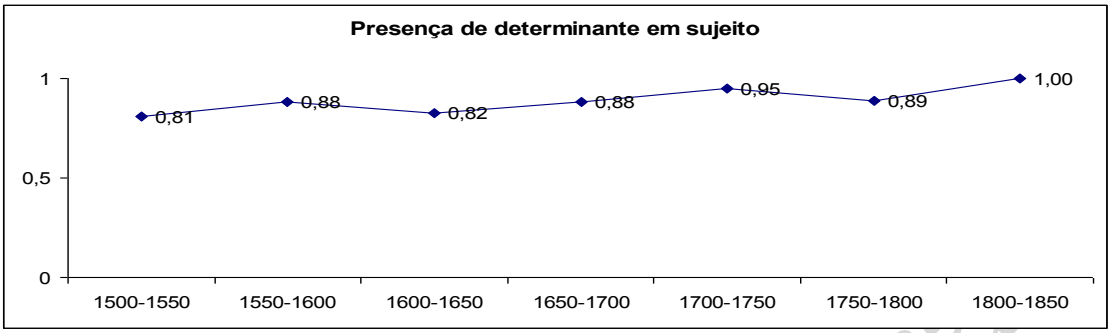


Gráfico 1 – Percentagem do uso de determinantes em DPs possessivos sujeitos

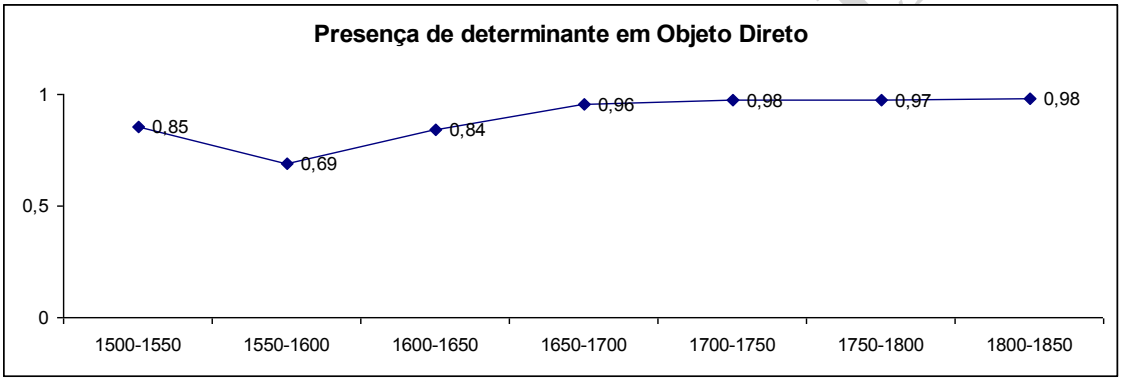


Gráfico 2 - Percentagem do uso de determinantes em DPs possessivos objetos diretos

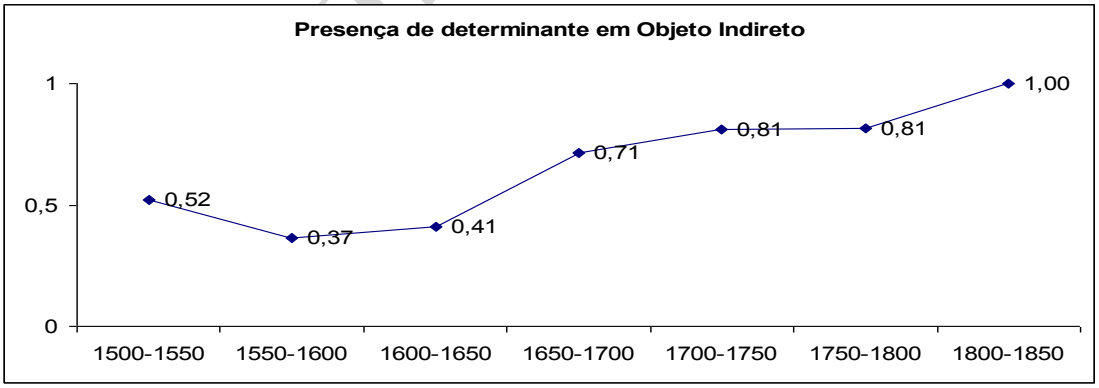


Gráfico 3 - Percentagem do uso de determinantes em DPs possessivos objetos indiretos

[3] Vale ressaltar que os valores de 0 a 1, apresentados em todos os gráficos, equivalem a uma escala de 0 a 100, configurando-se, portanto, como porcentagens e não como valores de pesos relativos, pois fizemos a opção de utilizar o GoldVarb apenas como uma ferramenta de auxílio para a contagem dos dados.

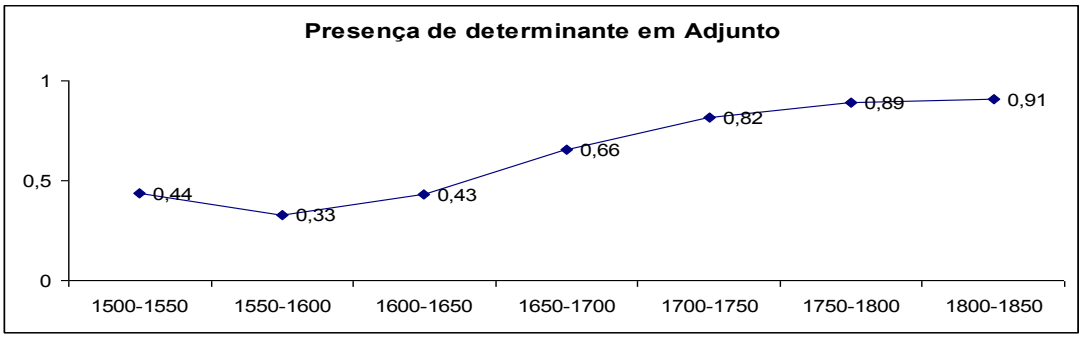
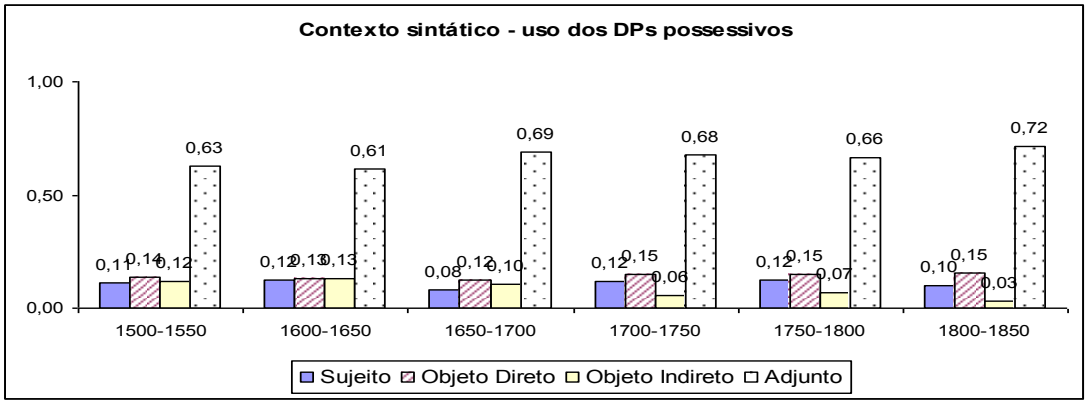


Gráfico 4 - Percentagem do uso de determinantes em DPs possessivos adjuntos

Conforme observamos por meio dos gráficos acima, é possível agrupar os casos de DPs em sujeitos e objetos diretos de um lado e os objetos indiretos e os adjuntos de outro. Com relação ao primeiro grupo, seu comportamento mostra casos em que o número de emprego de artigos era mais elevado desde o início do século 16 com uma pequena variação. Já para o segundo grupo, o número de ocorrências era mais baixo no início do século 16 havendo um posterior crescimento no decorrer dos séculos até uma quantidade elevada de uso do artigo no século 18. Essas diferenças não parecem ser aleatórias, pois justamente os contextos em que percebemos melhor a mudança na gramática do PC são aqueles em que se requer uma preposição, como nos objetos indiretos e nos adjuntos.

Ao fazermos as quantificações dos dados obtidos, averiguamos que as realizações dos DPs possessivos de acordo com o contexto sintático mostraram uma disparidade grande quanto ao emprego dos adjuntos em oposição aos demais contextos. Nota-se que a quantidade de aplicação de adjuntos sobressai-se dos demais contextos, pois em termos absolutos, o número de adjuntos é bastante superior aos outros, conforme visualizado no gráfico a seguir.



Gáfico 5 – Percentagem do uso do DP possessivo de acordo com o contexto sintático.

Como percebemos em números de realização de DPs possessivos, o contexto sintático mais utilizado sempre é o de adjunto, agrupado em dois tipos (adnominais e verbais). Em consequência do elevado número de ocorrências deste contexto é nos adjuntos que a língua visualizou a mudança sintática por meio dos nossos dados históricos. E é nos adjuntos que o licenciamento da preposição pode trazer influências acarretando em mudança. Isto quer dizer que a preposição nestes contextos desempenha um papel importante para a gramática da língua.

Quando agrupamos todos os contextos sintáticos, o número de ocorrências de adjunto é grande, mas o número elevado de adjuntos em relação aos demais contextos não tem significação para a análise, pois na verdade não é o adjunto em si que influencia na mudança, mas os contextos em que ocorrem uma preposição são aqueles que nos trazem uma melhor visualização da mudança. Portanto, não é a função sintática que está em jogo para evidenciar a mudança, mas sim o emprego de uma preposição. Se fosse a função sintática, esperaríamos que houvesse diferenças nos resultados de uso de artigo entre sujeito e objeto, por exemplo, mas como vemos, não há. Além disso, verificamos que os casos de objetos indiretos ocorrem em números bem menores que os de adjuntos, comparando-se àqueles de sujeito e objeto direto, mas mesmo assim o comportamento deste contexto assemelha-se ao de adjuntos. Isso quer dizer que o fato de haver menos casos de objetos indiretos não invalida o comportamento diferente deste contexto em relação aos demais, pois mesmo em menor número, funcionam nos mesmos moldes dos adjuntos devido a presença da preposição.

Compreende-se, dessa maneira, que a preposição desempenha um papel importante para a mudança. Passemos, então, aos resultados de DPs possessivos com relação ao uso ou não de uma preposição.

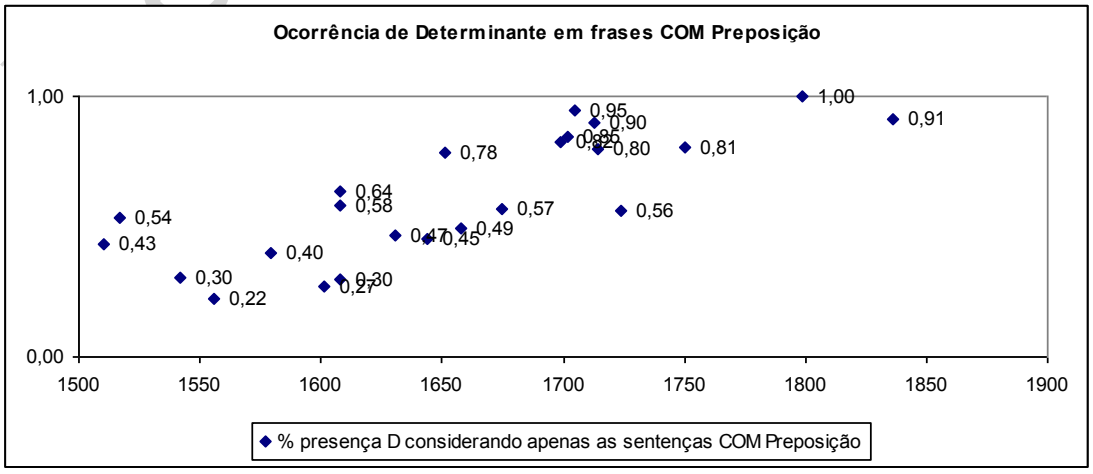


Gráfico 6 – Percentagem do uso de determinante em DPs possessivos preposicionados

De acordo com as realizações de artigo, verifica-se que até o século 17, nos casos de DPs possessivos acompanhados de uma preposição, o número de artigos era bastante reduzido, situando-se num patamar inferior aos 50% de ocorrência e que posteriormente sofreu um aumento devido a mudança na língua, chegando até a 100% de aplicação no DP.

E nos casos em que a preposição não era realizada, mesmo notando haver um período de variação, o número de artigos sempre foi maior, como representado no gráfico a seguir.

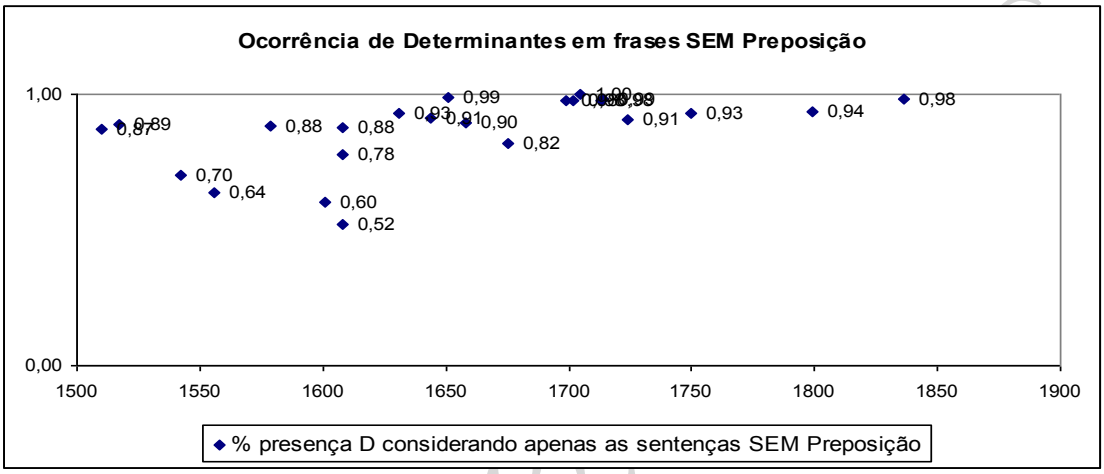


Gráfico 7 – Percentagem do uso de determinante em DPs possessivos não preposicionados

Mesmo sendo em números mais elevados com relação ao contexto sem preposição e com uma variação mais sutil, também verificamos uma mudança nos padrões de realização do artigo neste contexto a partir do século 17.

Verificamos que até os anos de 1650-1700, aproximadamente, há variação no uso do artigo, mas ainda não temos em questão um contexto de mudança estabelecido.

Ao atentarmos para os dois últimos gráficos, nota-se que até 1700 temos dois panoramas de mudança, o que remete a distintos sistemas possessivos sendo realizados. Vamos assumir que nessa época duas mudanças estavam em jogo.

Consideremos os resultados em um contexto que está mais livre de influência, como os casos sem a preposição com cerca de 80% de realização do determinante aproximadamente. Estes dados comprovam que no século 16 e 17 a média de 20% a 30% das realizações sem artigo correspondem a uma gramática como a do Francês em que o possessivo ao possuir os traços [+definido] e [+ possessivo] é alçado para Dº, inibindo o uso do artigo. Nesse pequeno número de ocorrências (cerca de 20%) serve como evidência clara de um pronome possessivo que não utiliza o determinante.

Ainda no mesmo contexto sem preposição, os 80 % restantes dos resultados em que o artigo é realizado correspondem a uma outra gramática com obrigatoriedade do uso do artigo. Mesmo assim, os casos que englobam estes 80% de ocorrências de artigo não são suficientemente evidentes para afirmar que correspondam todos a uma mesma gramática.

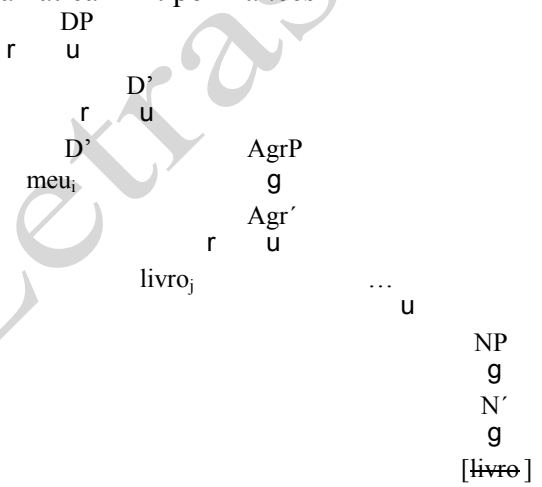
Apresentamos na próxima seção as características estruturais dos sistemas possessivos encontrados no período de 1500 até 1700 do português clássico.

4. Proposta de análise: Duas gramáticas para o sistema possessivo do português clássico

Primeiramente, consideremos os casos em que a preposição é realizada. No início do século 16, no português clássico, quando ocorria uma estrutura em que o possessivo detinha os traços [+posse] e [+definitude] este era capaz de ser alojado na posição de Dº e selecionado pela preposição.

Seria equivalente a uma estrutura semelhante a do Francês para o pronome *mon*, em uma configuração seguindo os mesmos moldes propostos por Brito (2001, 2003, 2007), mas que o artigo não era realizado como afirma a autora para o português europeu, conforme exemplificado em (5).

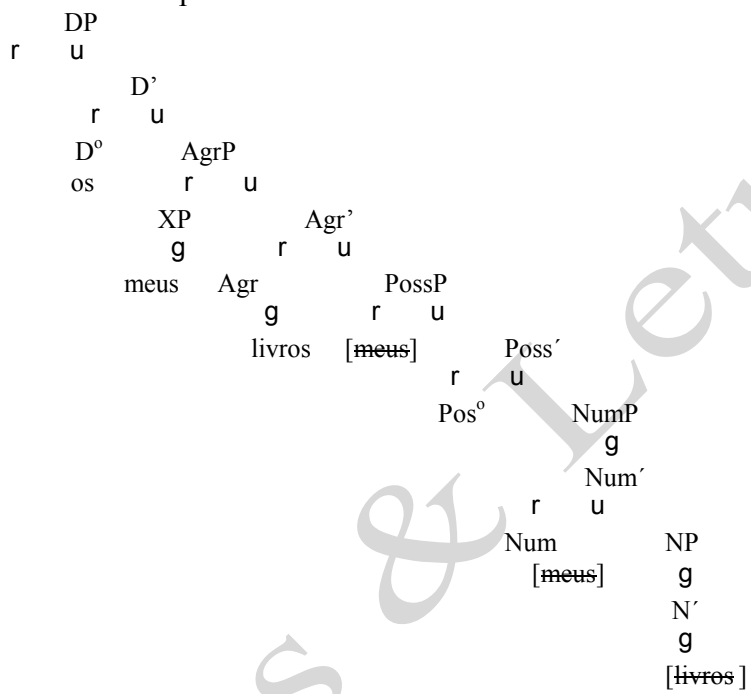
(5)
Gramática 1 – tipo *Francês*



No entanto, se esta fosse a única estrutura em vigor, o que dizer quando era licenciado um artigo como percebemos nos gráficos acima? Desse modo, podemos afirmar que havia um outro sistema possessivo.

Assumimos, portanto, uma outra estrutura possessiva baseada nos moldes de Brito (2007) semelhantemente à estrutura do Italiano em que o possessivo estaria em uma posição mais baixa dotado do traço [+posse] alojado em Spec,AgrP ou em PossP e o artigo dotado do traço [+definido] alojado em Dº.

(6)
Gramática 2 - tipo *Italiano*



Apresentamos abaixo, na forma de quadro, o esboço dos sistemas possessivos em co-ocorrência no português clássico que foram modificados na passagem para o português europeu, conforme nossa proposta de análise para a mudança em questão.

Quadro 1. Possessivos no português clássico.

	Português Clássico	Português Europeu
Gramática 1 *(Francês)*	[$_D$ meu] [+definido, +possessivo] Elemento Xº	Extinguiu-se
Gramática 2 *(Italiano)*	[$_D$ o] [$_{AgrP}$ meu] [+definido] [+possessivo] Elemento XP	Dialeto não padrão [$_D$ o] [$_{AgrP}$ meu] [+definido] [+possessivo] Elemento XP Reanálise da gramática tipo Italiano [$_D$ o meu]

		[+definido, +possessivo] Elemento Xº

Resumidamente, podemos dizer que no português clássico havia duas gramáticas em co-ocorrência: a gramática 1 em que o possessivo era dotado dos traços [+definido, +possessivo] realizado em núcleo de D, assemelhando-se à gramática do Francês, assim como uma outra, a gramática 2, em que o possessivo não era alçado para D, assemelhando-se à gramática do Italiano. Nessa última gramática os traços de definitude e posse eram checados por itens lexicais diferentes, alojados em seus respectivos núcleos funcionais.

A explicação para a mudança no sistema possessivo que ocorreu no PC deve-se a dois fatores: ao aumento no uso do determinante e a uma reanálise na categoria do pronome possessivo do Italiano. Não sabemos qual dos dois fatores ocorreu primeiro, mas estes podem ser recuperados nos nossos dados.

A gramática 1 (tipo do Francês) em que o possessivo localiza-se ora em PossP ou AgrP (como o *mien)* ora em D (como o *mon)* já era minoritária no período investigado. Como verificamos nos dados sem preposição, apenas 20% a 30% de ocorrências de possessivo pré-nominal sem determinante era realizado (estrutura que corresponderia a do pronome *mon*). Esta estrutura no decorrer dos séculos foi perdendo força até desaparecer totalmente no português europeu moderno. Com o aumento no uso do artigo, ocorreu uma reanálise da gramática 2 do possessivo (tipo do Italiano), acirrando ainda mais a competição com esta gramática 1 (tipo do Francês) em que o artigo e possessivo eram realizados em núcleo D.

Assumimos que a gramática 2 (do tipo Italiano), que correspondia a cerca de 70% a 80% de uso do artigo nos dados sem preposição, sofreu um processo de reanálise do seu pronome possessivo, sendo este anteriormente confirgurado como um elemento XP, passou a a ser licenciado pela gramática vencedora em posição de núcleo D juntamente com um artigo para o dialeto padrão[4].

[4] Em outros termos essa nova gramática é derivada do Italiano, mas que se assemelha de certa forma ao Francês pelo fato de o possessivo passar a ser realizado em D, em outras palavras, a configuração estrutural da gramática que corresponde ao sistema do Italiano em que o possessivo se localiza em uma posição mais abaixo de D foi reanalisada passando a ocorrer em núcleo D.

5. Conclusão

Nesta pesquisa, procuramos investigar a respeito do emprego dos artigos definidos em contextos de sintagmas nominais possessivos ao longo dos séculos XVI ao XIX. Constatamos ter havido uma mudança nos padrões de aplicação do artigo nesse período, sendo possível delinearmos por meio de gráficos o perfil dessa mudança sintática no Português Clássico.

Ao considerarmos alguns trabalhos anteriores que revelaram a importância de maiores estudos a respeito do artigo frente aos DPs possessivos, tais como Silva (1982), Magalhães (2002) e Castro (2006), entre outros, percebemos que a preposição evidencia um papel importante para a visualização da mudança. Além disso, procuramos estabelecer uma hipótese que considerasse os padrões estruturais da sintaxe da época, culminando no que sabemos existir no português europeu moderno, fase final da mudança. Portanto, nossa proposta central para a análise dos dados assume que o pronome [meu] do português clássico passou a ser reanalizado como [o meu] em um mesmo núcleo, passando a possuir os traços [+definido, + possessivo] e modificando sua categoria antes XP para Xº. Com tais considerações, buscamos trazer um panorama a respeito do contexto em estudo para que novas pesquisas a respeito possam ser embasadas com relação aos padrões de mudança ocorridos no português europeu.

Referências

ABNEY, S. P. **The English noun phrase in sentential aspects**, Ph.D. dissertation, Massachusetts: MIT, 1987.

BRITO, A. Presença / ausência de artigo antes de possessivo no Português do Brasil. In. **Actas do XVI Encontro da Associação Portuguesa de Linguística**. Faculdade de Letras do Porto. Centro e Linguística da Universidade do Porto, 551-575, 2001.

BRITO, A. Os possessivos em Português numa perspectiva de Sintaxe Comparada. Revista da Faculdade de Letras – **Línguas e Literaturas** XX:495-522, 2003.

BRITO, A. European Portuguese possessives and the structures of DP. **Cuadernos de Linguística del I.U.I Ortega y Gasset**, vol. 14, PP.27-50, 2007.

CASTRO, A. O Sistema dos Possessivos em Francês e em Português. In **Revista da Faculdade de Ciências Sociais e Humanas** (número especial em homenagem a Henriqueta Costa Campos), 1999.

CASTRO, A. Os Possessivos em Português Europeu e Português Brasileiro: Unidade e Diversidade. **Actas do XVI Encontro Nacional da Associação Portuguesa de Linguística de Coimbra,** Lisboa, APL, 599-613, 2000.

CASTRO, A**. On Possessives in Portuguese**. Ph.D. Diss, Universidade Nova de Lisboa - FCSH and Université Paris 8 –CLI, 2006.

CHOMSKY, N. **The minimalist Program**. Cambridge: MIT Press, 1995.

CORREIA, C. N. ***Estudos de Determinação****:* A Operação de quantificação-qualificação em sintagmas nominais. Lisboa: Fundação Calouste Gulbenkian, 2002.

COSTA, I. O uso do artigo definido diante de nome próprio de pessoa e de possessivo do século XIII ao século XVI. In Mattos e Silva & Machado Filho, A. V. L. (ed.) **O Português Quinhentista – Estudos Linguísticos**. EDUFBA/UEFS, 2002. p. 285-306,

FLORIPI, S. **O determinante em sintagmas nominais possessivos na história do português**. Universidade Estadual de Campinas. Tese de doutorado, 2008.

GARY-PRIEUR, M. N. **Gramaire du Nom Propre.** Paris: PUF, 1994.

LONGOBARDI, G. Reference and Proper Names: A Theory of N-Movement in Syntax and Logical Form. **Linguistic Inquiry** 25,4 : 609-665, 1994.

LOPES, O. **Gramática Simbólica do Português**. Lisboa: Fundação Calouste Gulbenkian, 1972.

MAGALHÃES, T. O Uso de Artigo Definido diante de Pronome Possessivo em Textos Portugueses do Século XVI a XIX. **DELTA**, 2004.

SCHOORLEMMER, M. Possessors, Articles and Definiteness In: **Possessors, Predicates and Movement in The Determiner Phrase** , Artemis Alexiadou & Chris Wilder (eds), 56-86, John Benjamins Publishing Company, 1998. crossref http://dx.doi.org/10.1075/la.22.04sch

MIGUEL SARMENTO, M. "O estatuto categorial dos possessivos: possessivos e adjetivos". In: **Actas do Encontro comemorativo dos 25 anos do CLUP**, 191-202. Porto, 2002a.

MIGUEL SARMENTO, M. Para uma tipologia dos possessivos. In: **Actas do XVII Encontro Nacional da Associação Portuguesa de Linguística**, Lisboa: Associação Portuguesa de Linguística. 287-299, 2002b.

MIGUEL SARMENTO, M. Possessive pronouns in European Portuguese and Old French. **Journal of Portuguese Linguistics** 1:215-240, 2002c.

MIGUEL SARMENTO, M. **O Sintagma Nominal em Português Europeu: posições de sujeito**. Tese de doutorado. Faculdade de Letras da Universidade de Lisboa, 2004.

SILVA, G. M. O. **Estudo da Regularidade na Variação dos Possessivos no Português do Rio de Janeiro**. UFRJ. Tese de Doutorado, 1982.

SILVA, G. M. O. Variação no sistema de Possessivo de Terceira Pessoa. **Tempo Brasileiro** (78/79): 54-72, 1994.

SILVA, G. M. de O.; CALLOU, D. O uso do artigo definido diante de possessivo. In: DUARTE, I. & LEIRIA, I. (orgs.) **Congresso Internacional sobre o Português**, Colibri/APL, Lisboa, Vol. III. p. 115-125, 1996.

SZABOLSCI, A. The Noun Phrase. In: KIEFER, N. F.; KISS, K. (eds) Syntax and Semantics 27 **The Syntactic Structure of Hungarian**. Academic Press, pp. 179-274, 1994.

Artigo recebido em: 10.10.2014
Artigo aprovado em: 04.12.2014

A função pragmática Tópico na legendagem brasileira de um filme argentino em um estudo de *corpus* paralelo

The pragmatic function Topic in Brazilian Portuguese subtitles of an Argentine movie: a parallel corpus study

Amanda Verdan Dib[*]
Paulo Pinheiro-Correa[**]

RESUMO: Neste estudo de *corpus* paralelo, analisamos a função pragmática Tópico na legendagem brasileira do filme argentino *O Segredo dos seus olhos/El secreto de sus ojos* (Espanha/Argentina, 2009), com o instrumental teórico da Gramática Discursivo-Funcional (GDF). Para levantar as ocorrências, utilizamos o programa *YouAlign* (Terminotix Inc.), de alinhamento de *corpora* paralelos. Foram analisados dois tipos de construções de tópicos: as topicalizações e os deslocamentos à esquerda. Os resultados indicaram um expressivo apagamento das construções de tópico dos diálogos originais em espanhol nas legendas brasileiras, o que pode dever-se à natureza do procedimento de legendagem, que tende à simplificação da sintaxe, mas também à resistência ao emprego dessas construções no português brasileiro escrito.

PALAVRAS-CHAVE: Linguística de *Corpus*. Espanhol. Tradução. Legendagem. Tópico.

ABSTRACT: In this parallel corpus study we have analyzed the pragmatic function Topic in Brazilian subtitles of the Argentine movie *The secret in their eyes/El secreto de sus ojos* (Spain/Argentine, 2009) within the theoretical framework of Functional Discourse Grammar (FDG). We used the software *YouAlign* (Terminotix Inc.) for alignment of parallel corpora. Two types of topic constructions were studied: topicalizations and left dislocations. Results indicate an erasing process of topic constructions from the original Spanish dialogues in Brazilian Portuguese subtitles. This could be due to the syntactic simplification typical of the subtitling process, but also to a trend of avoiding these constructions in written Brazilian Portuguese.

KEYWORDS: Corpus Linguistics. Spanish. Translation. Subtitles. Topic.

1. Introdução

Este artigo analisa as ocorrências da função informativa *tópico* no espanhol argentino (doravante EA) e português brasileiro (doravante PB), com o objetivo de descrever o uso dessa função pragmática em uma variedade e na outra por meio da análise de um *corpus* paralelo, tomando por base os pressupostos da Gramática Discursivo-Funcional (doravante, GDF), tal como propostos por Hengeveld e Mackenzie (2008).

Os dados analisados provêm de diálogos roteirizados em EA e a correspondente legendagem eletrônica em PB do filme *El secreto de sus ojos (2009)*.

[*] Mestranda em Estudos de Linguagem. Instituto de Letras, Universidade Federal Fluminense (UFF).
[**] Pós-Doutor em Linguística. Instituto de Letras, Universidade Federal Fluminense (UFF)/CAPES.

No restante desta seção, apresentamos o problema e os pressupostos teóricos, bem como os objetivos do trabalho. Na seção (2), apresentaremos a metodologia utilizada, o *software* empregado e a progressão dos passos metodológicos. Na seção (3) apresentamos os resultados e a discussão, à qual se segue a conclusão (seção [4]).

1.1. Definindo tópico

Dentro dos pressupostos da GDF (HENGEVELD; MACKENZIE, 2008), as funções pragmáticas, tais como tópico, foco ou contraste, têm uma função primordial e constituem primitivos gramaticais, de maneira que se considera que todo o enunciado é montado em sua ordenação lógica partindo-se das relações pragmáticas. Como observa Pezatti (2012, p.363):

> A pragmática, na GDF, refere-se ao modo como o falante modela as suas mensagens em relação às expectativas que tem do estado mental do ouvinte. Isso determina as partes de uma unidade linguística que serão apresentadas como particularmente salientes, as que serão escolhidas como ponto de partida do falante e as que serão consideradas compartilhadas pelo falante e pelo ouvinte.

Assim, o tópico, considerado tradicionalmente como um elemento do plano discursivo, neste modelo, é um componente estruturante do enunciado no que diz respeito à sua formalização lógica.

Dik (1997), dentro do modelo da Gramática Funcional que serve de base para a GDF caracteriza o tópico como: "a entidade sobre a qual a predicação predica alguma coisa em uma dada situação; em outras palavras, na predicação nós dizemos alguma coisa sobre o Tópico" (tradução de Pezatti [1998]). Assim, tal como em Lambrecht (1994), o tópico, para Dik (1997) é um elemento do plano pragmático claramente estabelecido, mas com um lugar que não fica claro dentro de um modelo de estruturação gramatical. No modelo da GDF, Hengeveld e Mackenzie (2008) estabelecem formalmente que a dicotomia tópico-comentário, bem como as demais funções pragmáticas, são primitivos gramaticais e os localizam no denominado nível interpessoal, na estruturação do enunciado, considerado o nível primordial da codificação gramatical. Nesse nível, mais primordial que o sintático e o semântico, é que são estruturados os participantes do acontecimento, a relação entre eles, o contexto e as funções informativas.

Desta maneira, dentro dos pressupostos da GDF, o tópico se define como uma função que pode vir a ser atribuída a certo elemento do enunciado e que vai assinalar a maneira como o conteúdo comunicado se relaciona ao registro que vai sendo construído gradualmente no componente contextual (HENGEVELD; MACKENZIE, 2008). Em outras palavras, é uma função operante na representação lógica do enunciado e da qual vai depender a maneira como o conteúdo comunicado vai ser apresentado ao interlocutor. Esta função pode fazer com que o elemento a ela associado venha a ser construído como ponto de partida do enunciado, o que vai determinar formalmente que este seja interpretado pelo interlocutor como tópico.

1.2. Tradução e funções informativas

No que tange aos Estudos da Tradução, com foco na sua modalidade audiovisual, especificamente, a legendagem, há questões teóricas e práticas que envolvem a passagem do texto oral (áudio original em EA) para o escrito em outra língua (legenda em PB). Em Lerma Sanchís (2012), há uma citação de Mayoral Ascencio (1999) sobre as mudanças que ocorrem nessas transposições: "este se refere à forma como a expressão linguística de significados potencialmente similares pode variar, dependendo das estratégias diferentes que dão lugar a segmentos textuais distintos". (MAYORAL, *apud* LERMA SANCHÍS, 2012, p.72).

Duro (2001), por sua vez, aponta três princípios da prática da tradução audiovisual, aos quais são atribuídas razões cognitivas cujo fundamento psicolinguístico não é explicitado no texto, mas que fazem parte da prática tradutória e que não podem deixar de ser considerados nesta pesquisa. São eles: a) sintetizar, b) sincronizar e c) tornar claro. O fator (a), a necessidade de sintetizar, pode acarretar perda de informação que poderia ser importante para a compreensão do sentido original, o que é um ponto crucial no que se refere ao *status* das funções pragmáticas. Em termos da GDF, nossa hipótese é a de que, ante a necessidade de sintetizar, passem a ser privilegiados elementos do chamado *nível representacional*, que é centrado no objeto da enunciação em si, em detrimento de elementos do nível interpessoal, relacionado ao contexto e aos participantes da enunciação. O fator (b) trata da sincronização entre a imagem (cena) e texto (legenda correspondente) uma vez que não havendo concordância na entrada e saída das legendas com base nos seus "pares" orais, a compreensão do sentido original pode ser afetada. Cada legenda deve surgir exatamente quando o personagem fala, de acordo com a dinâmica da cena. O fator (c), "deixar claro", por sua vez, implica facilitar o entendimento do espectador. Isto se concretiza através do emprego dos

sinais de pontuação considerados adequados, por exemplo, além da escolha de estruturas sintáticas porventura mais simplificadas que outras que poderiam eventualmente ter uma correspondência maior com o texto original. Acreditamos que pelo menos os fatores (a) e (c) podem interferir no texto da legenda e por isso devem ser considerados nos resultados.

1.3. O tópico no PB e no EA

Para a análise do tópico nos dados referentes ao Português do Brasil (PB), utilizamos a classificação clássica de Ross (1967), relida por Pontes (1987), que diferencia dois tipos de construções de tópico: topicalização (TOP) e deslocamento à esquerda (DE). Assim, as topicalizações são identificadas como aquelas que ocorrem sem retomada do componente inicial através de um termo anafórico, como no exemplo (1) abaixo, de Pontes (1987, p.66):

(1)TOP: Feijão, eu não gosto.

No exemplo acima (1), verificamos que o complemento verbal *feijão* encontra-se na posição inicial, cumprindo o papel da função pragmática tópico, pois resultou da inversão de ordem de palavras canônica (SVO) em favor do contexto situacional. Neste caso o SN tópico é incontável e não pode ser recuperado por um pronome elidido no comentário. A autora comenta que para o PB essa distinção não é categórica, uma vez que há casos em que a topicalização envolve objetos definidos que não são retomados por nenhum sintagma no comentário, mas que poderia ser um caso de deslocamento à esquerda com pronome elidido. Como em (2) e (2'), abaixo, ambos retirado de Pontes (idem):

(2) Meu cabelo desta vez eu não gostei nem um pouco.
(2')Meu cabelo desta vez eu não gostei nem um pouco dele.

Para a autora, no exemplo (2') o emprego do pronome *dele* ao final da sentença não promoveria nenhuma mudança semântica, razão pela qual em casos que envolvem um pronome elidido, ela postula ser difícil fazer a distinção entre TOPs e DEs.

Os DEs, apesar desta particularidade, se caracterizam fundamentalmente pela retomada do referente em posição inicial por um elemento anafórico, que pode ser o mesmo SN, outro SN ou um pronome, entre outros elementos, como no exemplo (3), abaixo, de Pontes (1987, p.12):

(3) DE: Os livros, eles estão em cima da mesa.

Diferentemente do que ocorre nos exemplos (1) e (2), em que o complemento é deslocado sem a retomada de um elemento anafórico correspondente, em (3) o pronome *eles* tem função sintática de sujeito, retomando o tópico *os livros,* e concordando com o mesmo em gênero e número.

Em ambos os casos (DE e TOP), para a GDF, a colocação do referente em primeira posição é determinada na estruturação lógica do enunciado no nível interpessoal e organiza pragmaticamente a sentença a partir deste ponto de partida, apresentando a informação na forma da dicotomia tópico-comentário, por razões discursivas.

Seguindo autores como Padilla (2005) e Sedano (2012) para o espanhol, de maneira geral e Kovacci (1992) para o EA, de maneira específica, observamos que a TOP envolve sujeitos ou outras funções sintáticas diferentes de objeto[1]. Já a DE envolve apenas objetos, segundo Padilla (2005) e também poderia incluir sujeitos, como propõe Sedano (2012), que, no entanto, não encontrou nenhum caso em seu estudo. Assim, se um objeto definido ocupar a posição de tópico, trata-se de um caso de DE, pelo fato de que o espanhol, por ser uma língua de sujeitos nulos, apresenta um inventário rico de clíticos de caso oblíquo, que prototipicamente estão presentes. Desta maneira, no que diz respeito à presença de objetos no tópico, a diferença entre a presença e a ausência do elemento anafórico no comentário em espanhol termina por distinguir os casos de tópico de outra estratégia informativa: o *contraste* (também chamado *foco contrastivo*, em versões anteriores do funcionalismo holandês). Nos exemplos (4) e (5) se estabelece a comparação entre tópico e contraste. O exemplo (4) é de um caso de tópico no espanhol argentino, retirado de Kovacci (1992, p.248):

(4) La escuela primaria la hice en casa.

Neste exemplo, o objeto *la escuela primaria*, por imposições do contexto é um tópico, (retomado por um clítico resumitivo) concordante em gênero e número com o objeto. Nos casos de contraste, o objeto definido aparece em posição anteposta e não é recuperado pelo clítico dentro da oração como em (5), também analisado por Kovacci (idem):

[1] No nosso entender, seguindo Kovacci (1992) que será discutida mais adiante, um objeto em posição de tópico sem pronome resumitivo no comentário é associado a uma estratégia de foco, e não, de topicalização.

(5) La escuela *primaria* hice en casa (no la secundaria).

Kovacci chama a atenção, neste exemplo, como viemos discutindo, de que não é toda anteposição que serve para topicalizar. Neste caso, a anteposição do objeto é "um rema contrastivo e vai acompanhada de reforço acentual" (KOVACCI, 1992, p.248). Desta maneira, a autora define claramente a construção de anteposição de objeto do exemplo como *foco contrastivo* (relido pela GDF como *contraste*) e a associa a uma proeminência acentual típica dessas construções[2].

Dessa maneira observa-se que EA e PB se comportam de maneira contrária quanto à retomada de referentes em função de tópico no comentário: a TOP, que envolve a não-retomada de referentes no comentário, em EA está associada aos sujeitos e os objetos estão diretamente associados ao DE. Em PB, a TOP envolve objetos e o DE envolve sujeitos que são recuperados no comentário.

1.4. Hipótese

Partimos da hipótese de que os resultados em PB apresentem indícios de uma tendência à manutenção da marcação de tópico na legendagem. Para isso, nos baseamos em Pontes (1987), que argumenta que o PB é uma língua de proeminência de sujeito e de tópico. Igualmente, seguimos Pezatti (2012) que, numa análise das diferentes construções do PB, conclui que esta é uma língua categorial tópico-orientada, o que significa que as construções categóricas da língua (em oposição às téticas e às apresentativas) se constroem orientadas para o tópico.

Assim, esperamos que a orientação para o tópico, característica que os estudos contemporâneos vêm revelando ser cada vez mais clara no PB, se manifeste na legendagem, apesar das especificidades da tradução audiovisual, comentadas na seção (1.2) acima.

1.5. Objetivo

O objetivo deste trabalho é, com base na análise de um *corpus* com os diálogos de um produto audiovisual, detectar construções marcadas no espanhol que apresentam a função

[2]O itálico corresponde ao sublinhado do original e faz referência à proeminência acentual.

informativa *tópico* e verificar, no seu processo de tradução ao português na elaboração das legendas as seguintes possibilidades: i. se os tópicos permanecem como função marcada; ii. se a construção permanece marcada na legendagem com outro recurso diferente do tópico; ou iii. se a construção deixa de ser marcada.

2. Metodologia

A fim de alcançar os objetivos acima descritos, escolhemos o filme *El secreto de sus ojos (2009)*, para análise. A escolha de tal filme se deve a estudos preliminares por nós conduzidos que o revelaram como uma boa fonte de dados para o estudo das funções *tópico* e *foco* na variedade argentina do espanhol em legendas de filmes. Quisemos dar continuidade aos estudos preliminares acerca do tema das funções informativas e, consequentemente, analisar o filme com mais profundidade, uma vez que o projeto relaciona duas grandes áreas: a Linguística Funcional e a Tradução Audiovisual. Para tanto, escolhemos trabalhar com o estudo comparado de *corpora*, pelo qual este trabalho se insere no campo da *Linguística de Corpus*, campo sobre o qual discorreremos brevemente nesta seção, com o objetivo de situar a pesquisa de que se ocupa este artigo no referido campo de pesquisa.

Baker (1995, p. 225) ao analisar o uso de *corpora* nos estudos de tradução, dá uma definição contemporânea de *corpus*, em oposição a definições que vigoravam anteriormente. Assim, segundo a autora:

> (i)Um *corpus* hoje em dia significa basicamente uma coletânea de textos legíveis em computador e capazes de serem analisados de forma automática ou semi-automática de várias maneiras; (ii) já não é restrito a 'escritos' mas inclui a linguagem falada, bem como textos escritos, e (iii) pode exibir grande quantidade de texto proveniente de uma variedade de fontes, de muitos autores e falantes e com diversidade de assuntos.[3]

Entre os tipos de *corpora* adequados aos estudos de tradução, ela menciona três tipos: *corpora* paralelos, multilíngues e comparáveis.

Nosso estudo utiliza um *corpus* paralelo, devido à especificidade dos dados. Um *corpus* paralelo pode ser definido como aquele que contém textos-fonte e suas traduções,

[3]Tradução dos autores, do original em inglês: (i) corpus now means primarily a collection of texts held in machine-readable form and capable of being analyzed automatically or semi-automatically in a variety of ways; (ii) a corpus is no longer restricted to 'writings' but includes spoken as well as written text, and (iii) a corpus may include a large number of text from a variety of sources, by many writers and speakers and on a multitude of topics.

podendo ser bilíngues ou multilíngues, uni-, bi- ou multidirecionais (MC ENERY; XIAO, 2007, entre outros). Neste sentido, o *corpus* que constituímos, contendo o texto original em EA e a legendagem desse texto em PB pode ser considerado bilíngue e unidirecional, por envolver apenas duas línguas e por não possuir a contraparte de textos originais em outra língua que não o EA.

Para o objetivo desta análise, era necessário analisar enunciado por enunciado em cada língua e a sua relação entre eles. Com a finalidade de obter um paralelismo semi-automático entre as falas originais e as legendas, utilizamos o *software* gratuito de alinhamento *on-line YouAlign* (Terminotix Inc., 2009-2014), o qual nos permitiu ordenar lado a lado os diálogos roteirizados em espanhol e suas legendas correspondentes em português, em um único arquivo.

Uma característica deste *software* é que o alinhamento é todo realizado *on-line*, ou seja, depois que o pesquisador se cadastra no site: *www.youalign.com*, deve alimentar o programa com os arquivos correspondentes. O procedimento passa a ser descrito a seguir:

(i.) deve-se manter apenas a legenda nos arquivos, no nosso caso, retirando a marcação de tempo e números de falas do arquivo de legendas original;

(ii.) deve-se ter um arquivo separado para os dados de cada língua (o programa admite arquivos de Word ou mesmo PDF*);*

(iii.) inserir os dois arquivos, separadamente, no *site*, para o alinhamento, na página *Alignment Settings;*

(iiii.) solicitar o alinhamento, que é realizado automaticamente, e o programa oferece duas opções de salvamento do arquivo alinhado, com as extensões HTML ou TMX. Optamos por salvá-lo em arquivo HTML. Ainda que quase todo o alinhamento seja realizado automaticamente, algumas correções devem ser feitas pelo pesquisador.

O programa revelou-se sensível inclusive aos casos em que não houve legendagem para algum segmento falado, deixando o espaço correspondente em branco e, para os casos em que a legendagem não corresponde ao enunciado original, o programa reconhece o espaço e o completa com a legenda correspondente, ainda que esta não apresente marcas léxicas ou morfológicas comuns com o enunciado original.

A etiquetagem das funções informativas se deu de maneira manual no arquivo já alinhado.

A obtenção do arquivo com as legendas em PB do filme analisado foi feita por adaptação de um arquivo baixado através do endereço eletrônico *http://www.opensubtitles.com*. O arquivo inicialmente obtido correspondia a uma legenda amadora do filme, do tipo elaborado majoritariamente por internautas ou aficionados à indústria cinematográfica, e não correspondia diretamente à legenda comercial, publicada no filme lançado em DVD. Como o projeto consiste em analisar a legenda comercial do filme, por uma questão de credibilidade, sobretudo, adaptamos cada legenda transmitida no filme em PB para o arquivo que já possuíamos. Por outro lado, o arquivo com a transcrição dos diálogos oralizados originais do filme foi obtido no *site Data base of movie dialogs (http://movie.subtitle.com)*. Estes também foram cotejados com os diálogos que efetivamente foram empregados no filme e revisados, quando necessário.

Uma última questão metodológica se refere ao tipo de análise do *corpus* desenvolvida. Esta partiu do levantamento das construções com tópicos marcados nos dados do EA para, a partir delas, se observar com se comportavam as legendas correspondentes em PB, observando-se as três possibilidades de resolução delineadas em (1.5), ou seja, se o tópico em PB permanecia marcado, se a construção permanecia marcada de outra maneira sem que o tópico fosse o marcado ou se a construção equivalente deixava de ser marcada. Dessa maneira, fica clara nossa decisão metodológica, de partir dos dados no idioma original. Ainda, a pequena quantidade de dados não justificou a realização de uma análise estatística e conduziu a uma análise de caráter mais qualitativo que quantitativo.

3. Resultados e discussão

3.1. Resultados

O resultado referente ao número total de construções de tópico no EA e PB foi discrepante entre uma língua e outra. Das 53 ocorrências de tópicos marcados, 42 foram casos de tópicos marcados no Espanhol Argentino (EA) e 11 casos de tópicos marcados no Português Brasileiro (PB), equivalente a 79% e 21% das ocorrências totais, respectivamente, como aponta a tabela 1:

Tabela 1: Tópicos marcados totais e por cada idioma (EA e PB).

Tópicos Marcados Totais	Tópicos Marcados no EA	Tópicos Marcados no PB
53	42	11
100%	79%	21%

Quanto aos processos de tradução ao português das construções em espanhol envolvendo a função tópico, feita a partir dos diálogos com base em roteiros no EA para a sua legendagem correspondente no PB, observamos 11 ocorrências (26%) em que a tópicos marcados no diálogo original correspondiam também tópicos marcados na legendagem (este resultado corresponde também aos números totais de construções de tópicos marcados). Houve 28 ocorrências (67% dos casos) de tópicos marcados no original em espanhol que corresponderam a construções não-marcadas na legendagem brasileira. Em 3 ocorrências (7%) os tópicos marcados no original em EA se resolveram com estruturas diferentes, como outra função informativa marcada ou uma legenda não correspondente nocionalmente ao enunciado original. A tabela (2) abaixo elenca estes resultados:

Tabela 2: Tipos de resolução das construções de tópico do EA na legendagem no PB.

Tópico Marcado EA> Tópico Marcado no PB	Tópico Marcado EA> Construção equivalente não-marcada no PB	Tópico Marcado EA> Outras construções no PB
11	28	3
26%	67%	7%

A primeira coluna aponta os resultados totais, a segunda, a quantidade de construções que foram traduzidas como marcadas e a terceira, aquelas que foram traduzidas como não-marcadas.

Por uma decisão metodológica, foi importante destacar, separadamente, as subclassificações das construções de tópico e suas manifestações em ambas as línguas analisadas. A tabela 3, abaixo, aponta as ocorrências de topicalização (TOP) marcadas no EA e as resoluções obtidas na legendagem em PB.

Tabela 3: Topicalizações marcadas no EA e suas resoluções ao PB.

TOPs Marcadas em EA (total)	TOPs Marcadas EA> Tópico Marcado no PB	TOPs Marcadas EA> Construção equivalente não-marcada no PB	TOPs Marcadas EA> Outras construções no PB
33 (100%)	8 (24%)	22 (67%)	3 (9%)

Na tabela, a primeira coluna indica os resultados totais de topicalizações no diálogo original, a segunda coluna, aquelas construções que permaneceram marcadas, a terceira, aquelas que foram traduzidas de maneira não-marcada e a quarta, as construções com um conteúdo nocionalmente diferente dos diálogos originais.

A tabela 4, abaixo, apresenta os resultados encontrados para os casos de deslocamento à esquerda marcados no EA e, logo, como foram resolvidos na versão ao PB. Verifica-se que há uma discrepância entre resultados comparados de TOP e DE entre as duas línguas.

Tabela 4: Deslocamentos à Esquerda marcados no EA e suas resoluções ao PB.

DEs Marcados EA(Total)	**DEs Marcados EA> Tópico Marcado no PB**	**DEs Marcados EA> Construção equivalente não-marcada no PB**	**DEs Marcados EA> Outras construções no PB**
9 (100%)	3 (33%)	6 (67%)	-

Como na tabela 3, a primeira coluna indica os resultados totais, desta vez, dos deslocamentos à esquerda dos diálogos originais, a segunda coluna, aquelas construções que permaneceram marcadas na legendagem, a terceira, aquelas que foram traduzidas de maneira não-marcada e a quarta, as construções que trariam um conteúdo nocionalmente diferente dos diálogos originais, para a qual não foi identificado nenhum caso.

Esta tabela demonstra – como será comentado com detalhe na próxima seção – que todos os casos de DE do *corpus* em espanhol que foram resolvidos como construção marcada no PB se resolveram como topicalizações.

3.2 Discussão

Como apontado na seção anterior, observamos que entre as variedades das línguas estudadas neste trabalho (EA e PB) não há somente semelhanças que devem ser analisadas, mas também suas diferenças e as estratégias linguísticas que foram empregadas nas soluções deste procedimento tradutório. O resultado obtido demonstra a disparidade entre as construções com tópico marcado no EA e no PB.

3.2.1. TOP

As ocorrências de TOP, uma das subclassificações das construções de tópicos, como vimos anteriormente, se manifestam de maneira similar em EA e PB quando desempenham construções marcadas, de acordo com o *corpus* analisado. Houve 8 casos de TOP marcada no

EA (24%), de um total de 33 casos de tópicos marcados no original, como nos exemplos (6) e (6'):

(6) (EA) No, para mí **la cárcel** toda la vida hubiera estado bien.
(6') (PB) "Não. Para mim, **a prisão perpétua** para ele estaria bem."

Por outro lado, dois terços das ocorrências (n=22, 67%) foram de casos de TOP silenciados na passagem ao PB, convertendo-se em construções não-marcadas. Nos exemplos a seguir, as construções de tópico, com sua ordenação invertida dos constituintes no original, foram resolvidas em PB com a manutenção da ordem SVO, considerada canônica e não-marcada. Entretanto, não é possível especular se há perda de informação semântica ou pragmática para o espectador, diante da especificidade da transmissão audiovisual, em que a informação chega ao espectador por meio de vários canais sensoriais. Seguem os exemplos (7) e (7') e (8) e (8'):

(7) (EA) Mire, ya **bastante problema** me trajo su quijotada con Romano.
(7') (PB) "Você já me causou **muitos problemas**. A sua briga com o coitado do Romano."

(8) (EA) Acá **el jefe** soy yo y **el subordinado** es usted.
(8') (PB) "Você tem que entender que eu sou **o chefe** e você, **meu subordinado**."

No exemplo (7), o objeto é pré-verbal em EA em uma configuração OVS e em (7'), do PB, este é pós-verbal, aparecendo em uma construção SVO. De maneira semelhante, os predicativos em EA (8) aparecem antes da cópula e em (8') aparecem sistematicamente depois da cópula e do sujeito, em PB.

3.2.2. DE

É importante destacar que todos os casos de DE encontrados em EA deram resultados em PB que não correspondiam ao recurso sintático de marcação do enunciado original. De acordo com os dados analisados, dos 9 casos de DE encontrados, todos os que foram traduzidos ao PB como construções marcadas (n=3, 33%) se converteram em TOP marcada, como em (9), (9'),(10) e (10'), abaixo:

(9) (EA) **Los presos** te **los** mando mañana.

(9') (PB) "**Os presos** eu mando amanhã."
(10) (EA) ***A Irene la*** *quiero matar.*
(10') (PB) "**A Irene**, tenho vontade de matar!"

Nos exemplos (11) e (11'), abaixo, houve o emprego do quantificador "todo" em lugar do pronome resumitivo, cumprindo estratégia sintático-semântica referente ao SN tópico *El del mes pasado*/"O do mês passado", (referente ao seu salário) que terminou por aproximar o conteúdo da legenda do conteúdo do diálogo em áudio original.

(11) (EA) **El del mes pasado** ya se **lo** chupó.
(11') (PB) "**O do mês passado**, ele já bebeu **todo**."

Nestes exemplos, o emprego do marcador de telicidade[4] 'se' em **se lo chupó** foi traduzido por 'todo', que mantém a semântica télica do evento.

Dois terços dos casos de DE obtidos no *corpus* em EA (n=6, 67%), no entanto, foram resolvidos em PB como construções não-marcadas, ou seja, aquelas formadas, principalmente, pela ordem SVO de constituintes, a chamada ordem canônica e que apresenta maior frequência. Os exemplos (12), (12'), (13) e (13') demonstram a perda da função pragmática em questão:

(12) (EA) (…) porque **este juzgado ¡**no **lo** pisas más en tu puta vida!
(12') (PB) "Você nunca mais vai voltar a pisar **neste tribunal**."

(13) (EA) Pero **este muchacho** no puede haber sido de ninguna manera.
(13') (PB) "Mas não pode ter sido **este garoto**, de jeito nenhum."

Comparando as ocorrências de TOP e de DE, as construções de TOP foram mais numerosas tanto em EA quanto em PB, mantidas as proporções, já que as construções marcadas em PB foram aproximadamente um terço das construções marcadas do EA. Curiosamente, dois terços das construções tanto de TOP quanto de DE originais deixaram de ser marcadas na legendagem brasileira, sendo construídas na ordem SVO.

[4] A telicidade é uma noção semântica que se refere à realização de um evento em sua completude. No exemplo em questão, dizer 'se lo chupó' atribui ao evento de beber um traço semântico indicador que este se deu inteiramente, que não sobrou nada da bebida consumida. Em sentido figurado, faz referência ao fato de o personagem, por ser alcoólatra, ter gasto integralmente o salário com bebida.

Com relação às ocorrências de DE, o *corpus* em EA apresentou uma maior quantidade de casos de DE sem comparação com o PB. Das 9 ocorrências encontradas em EA, apenas 1 caso em PB apontou proximidade com tal subclassificação nos níveis semântico-pragmático, através do uso de um quantificador que se relacionava com o SN tópico, que foram os exemplos (13) e (13').

Nossa hipótese inicial era a de encontrar uma grande quantidade de casos de tópicos marcados que refletisse as observações de Pontes (1987) e de Pezatti (2012). Os resultados evidenciam construções de tópicos-sujeitos, mas não necessariamente, marcados, o que não corroborou nossa hipótese. A pouca quantidade percentual de ocorrências de tópicos marcados nos dados das legendas em PB pode estar ligada de maneira óbvia à especificidade do processo de legendagem, tal como discutido na seção 1.2, mas convém aportar outros elementos à discussão.

Pontes (1987, pp.60-63), ao analisar a tradução da obra "Caminho da Perfeição", de Teresa de Ávila, compara as versões elaboradas pelas monjas carmelitas portuguesas e brasileiras e mostra uma discrepância entre as duas versões na tradução das construções de tópico. Enquanto na tradução portuguesa se conservam as numerosas construções de tópico marcado da edição espanhola na qual se basearam as tradutoras, a tradução brasileira exibe um sistemático apagamento dos tópicos e uma reformulação das construções na estrutura SVO. Esta observação de Pontes demonstra uma tendência ao apagamento de construções de tópicos marcados na língua escrita.

Assim, além das especificidades da legendagem, outro fator em jogo na determinação dos resultados encontrados pode ser a tendência ao apagamento dos tópicos marcados na língua escrita, uma vez que o procedimento tradutório analisado implica a passagem da modalidade oral para a escrita.

4. Conclusão

Com base no *corpus* analisado, que corresponde aos diálogos originais em espanhol do roteiro do filme O segredo dos seus olhos/*El secreto de sus ojos* (EA) e sua legendagem eletrônica em PB, pudemos analisar manifestações da função pragmática *tópico* e as sutilezas de sua tradução nos dois idiomas relacionados.

O resultado assinalou maior frequência de construções de tópicos marcados em EA que em PB, o que significa que a maioria das construções originalmente marcadas no EA

passaram a não-marcadas no PB, precisamente dois terços delas. Entre as construções marcadas, o EA apresentou três vezes mais topicalizações que o PB e os casos de deslocamento à esquerda em espanhol se converteram em topicalizações no PB, que não apresentou nenhuma ocorrência deste tipo.

Os resultados corroboraram apenas em parte a hipótese de partida, de que na legendagem transpareceria a propriedade sintática do PB de ser uma língua com características de orientação para o tópico, segundo Pontes (1987) e Pezatti (2012). Com isso, esperávamos uma grande quantidade de tópicos marcados, e o que obtivemos foi uma grande quantidade de tópicos-sujeitos, não marcados, consoantes com as observações de Pontes (1987) e Pezatti (2012), mas não necessariamente com nossa expectativa de tópicos marcados. Tal resultado poderia dever-se à especificidade do procedimento de legendagem, que determina que seja empregada uma ordem direta e construções simplificadas, mas também pode estar relacionado a uma tendência ao apagamento das construções de tópico marcado na língua escrita, demonstrada por Pontes (1987).

Outra hipótese, que esperávamos que fosse corroborada, se confirmou. Era a de que diante da necessidade de simplificação, a legendagem se concentrasse no nível conceitual, da expressão do conteúdo, segundo a GDF, e abandonasse a expressão de elementos do nível representacional – que codifica informações pragmáticas, tais como as funções informativas e ponto de vista, e é a escolha que aparentemente foi feita, já que em prol da simplificação estrutural, foram as funções informativas as que foram cortadas em dois terços dos dados da legendagem.

Por fim, o emprego da metodologia adequada, que foi a escolha da constituição de um *corpus* paralelo unidirecional – que pode vir a expandir-se com incorporações futuras – foi fundamental para a visualização dos recursos empregados nos casos analisados. Possibilitou, por um lado, a observação de comportamentos sistemáticos, e, por outro, diferentes recursos tradutórios empregados diante de situações contextuais aparentemente idênticas, que não seriam possíveis de visualizar sem a utilização deste recurso.

Referências

BAKER, M. Corpora in Translation Studies. An Overview and Suggestions for Future Research. **Target,** Amsterdam, 7 (2), p. 223-243, 1995. crossref http://dx.doi.org/10.1075/target.7.2.03bak

DIK, S. **The Theory of Functional Grammar.** Part 2: Complex and Derived Constructions. Berlim e Nova York: Mouton de Gruyter, 1997. 2nd. Edition.

DURO, M. (Coord.). **La traducción para el doblaje y la subtitulación.** Madrid: Ediciones Cátedra, 2001.

HENGEVELD, K.; J. L. MACKENZIE. **Functional Discourse Grammar**: A Typologically-Based Theory of Language Structure. Oxford: Oxford University Press, 2008. crossref http://dx.doi.org/10.1093/acprof:oso/9780199278107.001.0001

LAMBRECHT, K. 1994. **Information Structure and Sentence Form.** Topic, Focus and the Mental Representation of Discourse Referents. Cambridge, Cambridge University Press. crossrefhttp://dx.doi.org/10.1017/CBO9780511620607

LERMA SANCHÍS, M. D. Recepção do cinema espanhol em Portugal: uma experiência partilhada". In: DIAZ FOUCES, O. (Org.). **Olhares & Miradas**. Granada: Atrio, 2012.

KOVACCI, O. **El comentario gramatical. Teoría y práctica II**. Madrid: Arco Libros, 1992.

MC ENERY, A; XIAO Z. Parallel and comparable corpora: what are they up to?. In: G. M. ANDERMAN; M. ROGERS (Eds.). **Incorporating Corpora**: Translation and the Linguist.Translating Europe. Clevedon: MultilingualMatters, 2007.

PADILLA GARCÍA, X.A. **Pragmática del orden de palabras**. Alicante: Universidad de Alicante, 2005.

PEZATTI, E. G. Constituintes pragmáticos em posição inicial: distinção entre tema, tópico e foco. **Alfa**, São Paulo, 42, p.133-150, 1998.

_____. Ordenação de constituintes em construções categorial, tética e apresentativa. **D.E.L.T.A.**, São Paulo, 28 vol 2, p. 353-385, 2012.

PONTES, E. **O tópico no português do Brasil.** Campinas: Pontes, 1987.

ROSS, J. **Constraints on variables in Syntax.** Tese (Doutorado em Linguística), Massachusetts Institute of Technology, Cambridge, MA, EUA, 1967.

SEDANO, M. La dislocación a la izquierda en el discurso escrito. **Estudios de Linguística,** Alicante, 26, p. 319-341, 2012.

Programa utilizado
YouAlign. Desenvolvido por Terminotix Inc. 2009-2014. (Disponível em www.youalign.com. Acesso em setembro de 2014).

Filme analisado
O segredo dos seus olhos. Direção: Juan José Campanella. Tornasol Films, 2009.1 DVD (127 min), NTSC, color. Título original: El secreto de sus ojos. Distribuidora no Brasil: Europa Filmes.

Artigo recebido em: 13.10.2014
Artigo aprobado em: 25.11.2014

A segmentação linguística na legendagem para surdos e ensurdecidos (LSE) de 'Amor Eterno Amor': uma análise baseada em *corpus*[1]

Linguistic segmentation in the subtitling for the deaf and hard-of-hearing (SDH) of the Brazilian TV soap opera *Amor Eterno Amor*: a *corpus*-based analysis

Vera Lúcia Santiago Araújo[*]
Ítalo Alves Pinto de Assis[**]

RESUMO: A segmentação na legendagem relaciona-se à divisão dos diálogos de uma produção audiovisual em legendas, a qual pode ser realizada a partir de três critérios: a) visual – pelo corte; b) retórico – pelo fluxo da fala; c) linguístico – pela sintaxe. Os resultados de uma pesquisa realizada pelo grupo LEAD (Legendagem e Audiodescrição) da UECE (ARAÚJO; NASCIMENTO, 2011) sugeriram que uma segmentação linguística adequada pode garantir uma leitura confortável à comunidade surda mesmo quando a legenda possui velocidades de 160 e 180 palavras por minuto. O presente trabalho teve por objetivo analisar e descrever como acontecem os problemas de segmentação linguística na LSE de um capítulo da telenovela 'Amor Eterno Amor'. A metodologia envolveu uma análise descritiva baseada em *corpus*. Os problemas de segmentação foram identificados a partir de etiquetas criadas com base no trabalho de Chaves (CHAVES, 2012; CHAVES; ARAÚJO, 2014). Após o processo de anotação, o *corpus* foi analisado com o auxílio do programa *Wordsmith Tools* 5.0. Os resultados apontaram para uma grande quantidade de problemas de segmentação linguística, em cerca de 25,5% do total de legendas. Os problemas apareceram com mais frequência na ordem dos sintagmas verbal e nominal em legendas de 3 linhas e com velocidade alta.

ABSTRACT: Segmentation in subtitling is related to the division of the dialogues on films and TV programs into subtitles, which can be carried out according to three criteria: a) visual – based on the cuts; b) rhetorical – based on the flow of speech; c) linguistic – on syntax. The results of a piece of research (ARAÚJO; NASCIMENTO, 2011) conducted by the research group called LEAD ('Subtitling and Audiodescription' in Portuguese) at State University of Ceará (UECE) suggested that an appropriate linguistic segmentation can ensure comfortable reading to the deaf community even when the subtitle rate is 160 and 180 words per minute. This study aimed to analyze and describe how linguistic-related segmentation problems occur in the SDH of a Brazilian TV soap opera called *Amor Eterno Amor*. The methodology involved a *corpus*-based descriptive analysis. The segmentation problems were identified through a tag setting created by Chaves (CHAVES, 2012; CHAVES; ARAÚJO, 2014). After the annotation process, the corpus was analyzed by Wordsmith Tools 5.0. The results showed a large amount of linguistic-related segmentation problems, in about 25,5% of the subtitles. The problems were more frequent in noun and verb phrases in three-line and fast subtitles.

[1] Este trabalho apresenta resultados parciais do Projeto 'Segmentação na Legendagem para Surdos e Ensurdecidos (LSE): Uma análise baseada em corpus' (Projeto CORSEL), financiado pelo CNPq. Ele também traz os resultados parciais da dissertação em andamento do coautor, oriundos de uma pesquisa de TCC (trabalho de Conclusão de Curso) desenvolvida no ano de 2013 como requisito parcial para obtenção do Grau de Bacharel em Letras/Inglês da Universidade Estadual do Ceará (ASSIS, 2013).
[*] Universidade Estadual do Ceará (UECE)
[**] Mestrando Programa de Pós-Graduação em Linguística Aplicada da Universidade Estadual do Ceará (UECE).

PALAVRAS-CHAVE: Tradução Audiovisual. Legendagem para Surdos e Ensurdecidos (LSE). Linguística de *Corpus*. Segmentação Linguística.	**KEYWORDS**: Audiovisual Translation. Subtitling for the Deaf and Heard-of-Hearing (SDH). Corpus Linguistics. Linguistic Segmentation.

1. Introdução

A Tradução Audiovisual (doravante também TAV) tem se tornado um campo fecundo quanto à elaboração de recursos para o acesso a produtos audiovisuais por parte de surdos – ensurdecidos e pessoas com deficiência visual-PcDV**.** No caso da comunidade surda, a Legendagem para Surdos e Ensurdecidos (LSE) é um dos meios pelos quais seus membros podem desfrutar de novelas, peças de teatro, telejornais, programas de TV, filmes etc. Contudo, pesquisas de recepção realizadas pelo grupo de pesquisa LEAD (Legendagem e Audiodescrição) da Universidade Estadual do Ceará (UECE) (FRANCO; ARAÚJO, 2003; ARAÚJO, 2004, 2007, 2008) com surdos[2] de todo o Brasil têm demonstrado que as legendas do tipo *closed caption* oferecidas pelas emissoras de TV brasileiras precisam de ajustes, sugerindo que os parâmetros utilizados na confecção dessas legendas não atendem às necessidades da comunidade surda brasileira.

Os resultados obtidos nessas pesquisas (ARAÚJO; NASCIMENTO, 2011; ARAÚJO, 2012), conduzidas com 34 surdos de 4 regiões brasileiras, sugerem que legendas bem segmentadas promovem uma recepção eficaz do programa televisivo legendado mesmo que sejam rápidas (entre 160 e 180 palavras por minuto), o que nos surpreendeu, pois acreditávamos, com base em pesquisas anteriores (FRANCO; ARAÚJO, 2003; ARAÚJO, 2004, 2007, 2008), que o principal entrave para uma recepção eficiente de filmes e programas legendados por parte de surdos estaria na velocidade da legenda. Uma boa segmentação acontece, segundo a literatura da área (DIAZ-CINTAS; REMAEL, 2007; REID, 1990; GOTTLIEB, 1994; KARAMITROGLOU,1998; IVARSSON; CARROL, 1998), quando as falas de uma produção audiovisual são transformadas em legendas a partir de critérios que levam em conta o corte ou mudança de cena[3] (segmentação visual), o fluxo da fala (segmentação retórica) e a sintaxe, a semântica, a lexicogramática, a fonologia-fonética (segmentação linguística).

[2] Os ensurdecidos não foram contemplados na pesquisa, porque acreditamos que uma legenda adequada para surdos contemplaria bastante suas necessidades, já que os surdos teriam menos acesso à língua portuguesa do que os ensurdecidos, os quais, diferentemente dos surdos, tem o português como primeira língua.
[3] Para fins desse trabalho, os dois termos estão sendo usados como intercambiáveis.

Tendo em vista essas questões, temos como objetivo do trabalho aqui relatado identificar os problemas de segmentação linguística da LSE de um capítulo da telenovela 'Amor Eterno Amor' e analisá-los para descrevê-los, classificando-os. Os outros tipos de segmentação não foram analisados, dado que não se apresentaram como problemáticos em pesquisas anteriores (CHAVES, 2012; CHAVES; ARAÚJO, 2012) envolvendo filmes e telenovelas. Além da identificação e descrição-classificação, pretendemos sugerir possíveis 'ressegmentações' dos problemas encontrados no *corpus* a partir de parâmetros preconizados por pesquisadores da área. A metodologia utilizada foi, em grande parte, uma replicação daquela adotada na pesquisa de Chaves (CHAVES, 2012; CHAVES; ARAÚJO, 2014) sobre os problemas de segmentação na LSE do filme brasileiro 'Nosso Lar'. Assim como na referida pesquisa, a identificação, análise e descrição-classificação dos problemas na segmentação linguística de 'Amor Eterno Amor' foi realizada com o auxílio de anotação por etiquetas criadas pela autora. Após a etiquetagem, o *corpus* anotado foi submetido ao *Wordsmith Tools* 5.0.

Este artigo está dividido nas seguintes seções além desta introdução: na Seção 2, apresentamos uma breve revisão de trabalhos relacionados à legendagem e sua segmentação e à legendagem em interface com a Linguística de Corpus; na Seção 3, dedicada à metodologia, discorremos sobre os processos de extração do *corpus* e sua anotação, dentre outras etapas; na Seção 4, apresentamos e discutimos os resultados da análise do *corpus* e, por fim, na Seção 5, traçamos algumas considerações finais acerca deles.

2. Legendagem

2.1 A Legendagem *Closed Caption*

O sistema norte-americano de legendagem, mais conhecido como *closed caption,* é um tipo de LSE, no qual a linha de legenda, para ser visualizada pelo espectador, necessita ser transformada em códigos eletrônicos e inserida na linha 21 do intervalo vertical em branco da TV, que é uma barra horizontal posta entre as imagens (ARAÚJO, 2008, p. 62). O sistema brasileiro, na maioria das vezes, não edita a fala, constituindo-se quase numa transcrição e, portanto, não atendendo à maioria das especificações técnicas esperadas numa legendagem voltada tanto para surdos e ensurdecidos como para ouvintes (ARAÚJO, 2008; ARAÚJO; NASCIMENTO, 2011). Esta abordagem difere-se da europeia no que diz respeito à confecção das legendas para surdos e ensurdecidos. Os europeus adotam os mesmos parâmetros para

ambos os tipos de legendagem, com exceção da inclusão de informações adicionais que não podem ser recuperadas pela audição como a identificação de falantes e de efeitos sonoros.

Aqui no Brasil, os procedimentos são diferentes dependendo do tipo de legendagem. Para os ouvintes, nos quais são legendadas produções audiovisuais estrangeiras, a legenda segue os padrões europeus, ou seja, tem no máximo duas linhas e velocidades de 145, 160 e 180 palavras por minuto (DIAZ CINTAS; REMAEL, 2007:96). Quando a velocidade da fala é maior do que 180 palavras por minuto, é necessário uma edição dessa fala por meio da condensação, redução (uso de palavras mais curtas) ou omissão de palavras quando seu sentido pode ser apreendido pelas imagens. Para os surdos e ensurdecidos, produz-se uma tipo de legendagem semelhante ao norte-americano utilizado somente para traduções intralinguísticas, já que o sistema foi criado exclusivamente para atender à comunidade surda dos Estados Unidos. A LSE produzida para os canais de TV brasileiros são também intralinguísticas, visto que são produzidas a partir das versões dubladas de filmes e programas de TV em língua estrangeira. A pesquisa realizada na UECE não concorda com essa diferença de procedimentos, aproximando-se mais da concepção europeia de legendagem.

A LSE do tipo *Closed Caption* pode aparecer em dois formatos: como *pop-on* ou como *roll-up*. A legendagem *roll-up* é destinada a programas ao vivo, podendo ser desenvolvida por estenotipia ou por refalamento (*respeaking*). No primeiro caso, por estenotipia, é produzida por meio de um teclado de digitação rápida chamado de estenótipo. No segundo caso, por refalamento, é produzida a partir de um software de reconhecimento de voz, no qual o legendista repete o que está sendo dito na tela para que as legendas possam ser visualizadas. As legendas do tipo *roll-up* aparecem na tela palavra por palavra da esquerda para a direita. Quando a linha de legenda se completa ela desloca-se para cima e permanece na tela até que a linha de baixo esteja também completa. A legendagem *pop-on*, por sua vez, também feita por meio do estenótipo, aparece na tela em bloco e sai em bloco. É semelhante às legendas exibidas tanto na TV quanto no cinema, as quais já estamos acostumados. É reservada para programas pré-gravados, como novelas e filmes. Este foi o formato analisado no presente trabalho, já que, supostamente, por ser editável, possibilitaria ao legendista a elaboração de LSE de acordo com os preceitos da literatura da área, especialmente no caso do parâmetro da segmentação, que, como já dito, foi o foco deste trabalho.

2.2 A Segmentação em Legendagem

A segmentação em legendagem diz respeito à divisão da fala traduzida em porções de texto escrito na parte inferior da tela. A segmentação, tal como indicamos anteriormente, pode ocorrer de três formas: pautada pela linguística (sintaxe, semântica, lexicogramática e fonologia-fonética), pela retórica (fluxo da fala) e pelo visual (corte ou mudança de cena) (DIAZ-CINTAS; REMAEL, 2007; REID, 1990; GOTTLIEB, 1994; KARAMITROGLOU,1998; IVARSSON; CARROL, 1998). A segmentação linguística[4] pode acontecer de duas formas: 1) segmentação entre linhas, a qual ocorre na divisão das falas de uma produção audiovisual em diferentes legendas; ou 2) segmentação quebra de linha *(line break)*, a qual ocorre dentro da mesma legenda, envolvendo a divisão da informação entre as duas linhas dessa legenda (DIAZ CINTAS; REMAEL, 2007, p. 173). O foco aqui recairá sobre o segundo tipo de segmentação linguística.

Muitos teóricos da TAV enfatizam a importância de uma segmentação linguística dentro dos critérios supracitados para a recepção eficaz de um filme ou programa legendado. Diaz Cintas e Remael (ibidem, p. 172) alertam que "uma segmentação cuidadosa da informação pode ajudar a reforçar a coerência e a coesão na legendagem"[5]. Do ponto de vista da segmentação quebra de linha, o foco deste estudo, as legendas devem ser divididas de uma forma que sejam autossuficientes sintática e semanticamente, já que "quando segmentamos uma sentença, forçamos o cérebro a pausar o seu processamento linguístico por um momento até que os olhos captem a próxima informação"[6] (KARAMITROGLOU, 1998, p. 10), sendo, assim, necessário que essa segmentação siga determinadas diretrizes linguísticas. Karamitroglou (1998) considera que o processo deve se dar no mais alto nível sintático possível. Para tal, o mesmo faz uso de nódulos sintáticos oriundos de uma descrição sintática em árvore de base gerativista, que apresentamos na Figura 1.

[4] Reid (1990) utilizou o termo 'gramatical' ao invés de 'linguística'. Para além da questão teórica problematizada por Chaves (2012, p; 45), a qual sugeriu que a escolha por "linguística" deveu-se ao fato de que a palavra 'gramatical' está "arraigada de preconceito e por isso acaba conduzindo à discussão por um viés normativo da língua", preferimos o termo 'linguística' por o mesmo ser mais abrangente, porque a segmentação não envolve somente a gramática, ou seja o léxico e a sintaxe, mas também a semântica e fonologia-fonética.
[5] Tradução dos autores. Original: "A careful segmentation of the information can help reinforce coherence and cohesion in subtitling".
[6] Tradução dos autores. Original: "When we segment a sentence, we force the brain to pause its linguistic processing for a while, until the eyes trace the next piece of linguistic information."

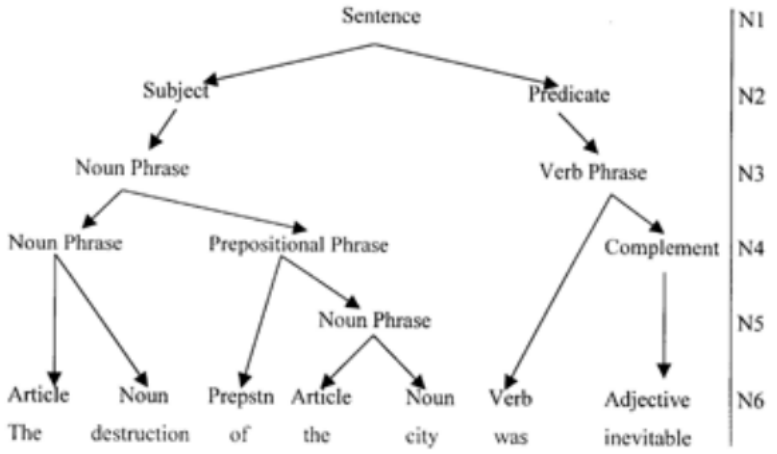


Figura 1 – Árvore sintática usada por Karamitroglou (1998) para explicar a segmentação linguística.
Fonte: Karamitroglou (1998, p. 9).

O autor divide a oração *the destruction of the city was inevitable* em seis nódulos: no primeiro, N1, está a oração em si, constituída de sujeito e predicado (N2), os quais são formados de sintagmas nominais e verbais respectivamente (N3). O sujeito *the destruction of the city* é constituído de dois sintagmas, um nominal (*the destruction*) e o outro preposicional (*of the city*), enquanto o predicado tem um complemento (N4). O sintagma preposicional também tem entre seus constituintes um sintagma nominal (*the city*, N5). Por fim, temos o léxico e a função dentro da estrutura gramatical (N6). Para o autor, se a oração em questão estivesse em uma legenda, ela deveria ser segmentada, idealmente, privilegiando as unidades *The destruction of the city* (sujeito) na linha de cima e *was inevitable* (predicado) na linha de baixo. Dessa maneira, a quebra aconteceria no mais alto nódulo sintático a partir do qual a segmentação de uma oração é possível, o N2, e, quando isso ocorre, "[...] maior é o agrupamento de carga semântica e mais completa é a informação a ser apresentada ao cérebro" (ibidem, p. 9)[7], facilitando, assim, a compreensão da legenda como um todo.

Um pequeno adendo a fazer é a respeito da distribuição das legendas na segmentação quebra de linha. Ao invés de utilizar os critérios de segmentação pelos critérios discutidos anteriormente, alguns legendistas preferem segmentar pelo número de caracteres, produzindo três formatos de legendagem. No primeiro, aparece quase o mesmo número de caracteres para as duas linhas, tendo um formato semelhante ao de um retângulo. No segundo e no

[7] Tradução dos autores. Original: "This occurs because the higher the node, the greater the grouping of the semantic load and the more complete the piece of information presented to the brain."

terceiro, aparecem mais caracteres na linha de cima ou na de baixo, respectivamente, lembrando um triângulo. O Quadro 1 mostra os três formatos:

Quadro 1: Formato das legendas na tela.

Formato	Legendas
Em forma de retângulo	O guardinha me parou por causa de uma bobagem da placa que caiu!
Em forma de triângulo com a linha de cima maior	Um tutuzinho de feijão, um lombinho.
Em forma de triângulo com a linha de cima menor	[Deolinda] já imaginava, por isso fiz o tutuzinho logo hoje.

Fonte: Os autores.

Karamitroglou (ibidem, p. 10) acredita que o ideal seria que a legenda de duas linhas, a qual pode ter até 39 caracteres em cada linha, tivesse um formato retangular, ao invés do triangular, por ser aquele ao qual os telespectadores estão mais acostumados. Contudo, quando as questões linguísticas não permitirem esse formato, os triangulares podem ser acessados. No que diz respeito ao capítulo de 'Amor Eterno Amor' em foco, quando foram propostas novas legendas conforme as diretrizes dos pesquisadores da área para a segmentação quebra de linha (DIAZ-CINTAS; REMAEL, 2007; REID, 1990; GOTTLIEB, 1994; KARAMITROGLOU,1998; IVARSSON; CARROL, 1998), o formato em forma de triângulo prevaleceu.

Por questões de espaço e, como foi dito anteriormente, dado que os outros tipos de segmentação não se apresentaram como problemáticos em pesquisas anteriores (CHAVES, 2012; CHAVES; ARAÚJO, 2014) envolvendo filmes e telenovelas, decidimos analisar neste trabalho apenas a segmentação quebra de linha, a qual será chamada aqui também de segmentação linguística.

2.3 Legendagem e Linguística de *Corpus*

O uso da Linguística de *Corpus* como metodologia nos Estudos da Tradução começou a ser preconizado por Baker (1993). Na Tradução Audiovisual, por sua vez, a análise baseada em *corpus* vem ganhando cada vez mais espaço entre os pesquisadores, sem ser ainda, contudo, grande o arcabouço teórico formado pela interface entre essas duas áreas disciplinares. A partir dos objetivos do presente trabalho, iremos detalhar, a seguir, as duas

pesquisas que embasaram nosso estudo. São, a saber: Perego (2008) e Chaves (2012; CHAVES; ARAÚJO, 2014).

Por meio da análise manual de um *corpus* de filmes, Perego (2008) examinou casos de má segmentação na legendagem interlinguística para ouvintes. Em seu trabalho, a autora investigou a segmentação linguística de um *corpus* heterogêneo de legendas de filmes para DVD e para cinema. A autora definiu categorias para a análise dessa má segmentação como quebra do sintagma nominal, quebra do sintagma preposicionado, quebra do sintagma verbal e quebra da oração complexa em seus constituintes, as orações coordenadas e subordinadas. Mesmo sem a utilização de uma análise baseada em *corpus*, como nos propusemos a fazer em nosso estudo, foi o trabalho de Perego que motivou e serviu como ponto de partida para a produção de etiquetas de Chaves (2012; CHAVES; ARAÚJO, 2014) para a análise de aspectos técnicos (número de linhas e caracteres, velocidade, tempo inicial e final da legenda) e, principalmente, de problemas de segmentação na LSE do filme brasileiro 'Nosso Lar'.

Na questão relativa à análise da segmentação linguística, Chaves (ibidem) utilizou os pressupostos de Karamitroglou (1998) sobre a importância de ela ocorrer no mais alto nível sintático possível. Para as definições de oração e dos vários tipos de sintagmas na sintaxe do português brasileiro, necessárias para a análise dos problemas de segmentação linguística, a autora fez uso da abordagem descritiva de orientação formalista de Perini (2010). Chaves (2012; CHAVES; ARAÚJO, 2014) formulou 3 etiquetas indicativas de problemas de segmentação (linguística, retórica e visual) e 8 para análise de parâmetros técnicos da LSE (número de linhas e caracteres, velocidade, tempo inicial e final da legenda). A partir dos conceitos, nomenclaturas e observações de Perini (2010) sobre a gramática do português brasileiro, a autora elencou um número de 19 subetiquetas referentes a categorias de problemas de segmentação linguística. Os resultados apontaram para uma maior quantidade de problemas de segmentação linguística no nível do sintagma verbal e nominal (41 e 26%, respectivamente), majoritariamente na quebra verbo+verbo, e em legendas com velocidade considerada alta (160 e 180 palavras por minuto). Abaixo, o Quadro 2 referente aos problemas de segmentação identificados por Chaves**.**

Quadro 2[8] – Quadro de etiquetas de problemas de segmentação de Chaves.

ETIQUETAS DE ANÁLISE DE PARÂMETROS TÉCNICOS DA LEGENDAGEM	
Número da legenda	<sub1>legenda 1</sub1>
Linhas por legenda	<1L>, <2L>
Tempos inicial e final de cada legenda	<t>início --> final</t>
Número de caracteres por linha (aplicada em legendas de 2 linhas)	<cpl>
Velocidade da legenda baixa (até 13cps) Velocidade ideal (14 a 15cps) Velocidade alta (a partir de 16cps)	<velocidade da legenda_baixa> <velocidade da legenda_ideal> <velocidade da legenda_alta>
ETIQUETA INDICATIVA DE PROBLEMA DE SEGMENTAÇÃO GRAMATICAL	
<PROSEGG>	
ETIQUETAS INDICATIVA DE PROBLEMA DE SEGMENTAÇÃO RETÓRICA	
<PROSEGR_antecipouinformação>	
<PROSEGR_atrasouinformação>	
ETIQUETA INDICATIVA DE PROBLEMA DE SEGMENTAÇÃO VISUAL	
<PROSEGV_vazou>	
ETIQUETAS DE ANÁLISE DE SINTAGMA NOMINAL (SN)	
<SN_pre-nucleares+subst>	
<SN_nominal+modif/modif+nominal>	
<SN_superlativo+adj>	
<SN_relativo+oração incompleta>	
<SN_nome próprio>	
<SN_título+nome próprio>	
<SN_colocações/idiom/conv>	
ETIQUETAS DE ANÁLISE DE SINTAGMA PREPOSICIONADO (SP)	
<SP_prep+subst>	
ETIQUETAS DE ANÁLISE DE SINTAGMA VERBAL (SV)	
<SV_verbo+verbo>	
<SV_verbo+adv>	
<SV_colocações>	
<SV_negação+verbo>	
<SV_(verbo)+oblíquo+verbo>	
ETIQUETAS DE ANÁLISE DE SINTAGMA ADVERBIAL (SAdv)	
<SAdv>	
ETIQUETAS DE ANÁLISE DE SINTAGMA ADJETIVO (SAdj)	
<SAdj_subst+adj>	
ETIQUETAS DE ANÁLISE DE ORAÇÃO COORDENADA (COORD)	
<COORD_coordenador+oração>	
<COORD_negativa>	
ETIQUETAS DE ANÁLISE DE ORAÇÃO SUBORDINADA (SUBORD)	
<SUBORD_conj+oração>	
<SUBORD_se>	

Fonte: Chaves (2012, p. 58)

Usando como ponto de partida as etiquetas da autora e valendo-se do pressuposto teórico de que "[...] mesmo a ação mais simples imaginável, a de contar palavras ou identificar a pontuação pressupõe uma teoria linguística" (GREFENSTETTE; TAPANAINEN, 1994; NUNBERG, 1990, *apud* SANTOS, p. 58) o mais coerente possível com o objetivo e o desenho metodológico no âmbito do Projeto CORSEL – este será mais detalhado na seção 3.1 – como um todo**,** passamos a embasar a análise da segmentação linguística a partir de um viés funcionalista, já que essa vertente dos estudos linguísticos possui uma maior aproximação epistemológica com a metodologia baseada em *corpus*. Para tal, a 'Nova Gramática do Português Brasileiro'(CASTILHO, 2012) – pela grande riqueza de

[8] Para mais detalhes sobre as diferenças entre as nossas etiquetas e as de Chaves, ver Assis (2013).

detalhes desta no que se refere à descrição dos sintagmas e das sentenças[9] complexas[10] da língua portuguesa brasileira, sendo menos sintética[11] que a descrição de Perini (2010) – foi adotada. Para Castilho (p. 249), "a estrutura sintática da sentença fundamenta-se nos arranjos lexicais de que ela é formada, os sintagmas, bem como nas funções que decorrem do relacionamento desses sintagmas", sendo estes um somatório de constituintes que apresentam um lugar previsível dentro na estrutura da sentença. Ainda segundo o autor, o sintagma tem uma estrutura composta por um NÚCLEO, uma margem esquerda, preenchida ou não pelos chamados ESPECIFICADORES do sintagma, e por uma margem direita, preenchida ou não pelos ditos COMPLEMENTADORES**.** No Quadro 3, mostramos os vários tipos de sintagma reconhecidos por Castilho e a estrutura geral de cada um bem como a estrutura geral dos dois tipos de sentença.

Quadro 3: Estrutura geral dos sintagmas e sentenças complexas (CASTILHO, 2012).

Sintagma Verbal (SV)	O SV é a construção que tem como núcleo o verbo. Como a sentença é um verbo que articula seus argumentos, a única diferença entre ele e uma sentença é que no SV não figura o sujeito.
Sintagma Nominal (SN)	O SN é uma construção cuja estrutura tem por núcleo um substantivo ou um pronome, tendo por Especificador o artigo e os pronomes e por Complementadores os SAdjs e os SPs.
Sintagma Adjetival (SAdj)	O SAdj tem por Núcleo o adjetivo, que é uma classe basicamente predicadora, funcionando como adjunto adnominal, enquanto constituinte do sintagma nominal, ou como predicativo, enquanto constituinte do sintagma verbal (CASTILHO, 2012, p. 516). O SAdj tem como Especificador advérbios predicativos qualificadores e como Complementador SPs e sentenças substantivas objetivas.

[9] Segundo o autor, a "designação da sentença não é pacífica na literatura" (p. 58), ao passo que podemos encontrar "[...] termos tais como oração, frase, período (conjunto de orações) etc." (ibidem). No Glossário de sua obra, Castilho, define o vocábulo 'sentença' da seguinte forma: "1. Sentença ou oração é a unidade da sintaxe estruturada por um verbo que seleciona o seu sujeito e seus complementos. Os adjuntos também integram uma sentença, mas não são selecionados pelo verbo [...]" (p. 691). Ao que nos parece, o autor considera os dois termos intercambiáveis, apesar de preferência pela utilização de 'sentença'. Esta pressuposição confirma-se pelo fato de que no vocábulo 'oração', há uma remissão ao vocábulo 'sentença' expressa por 'Veja **Sentença'** (p. 686). Em nossa análise preferimos utilizar o termo 'oração', pois, se 'sentença' e 'oração' são intercambiáveis, a utilização deste último se torna menos obscura. Inclusive, o termo 'sentença' nem mesmo consta 'Nomenclatura Gramatical Brasileira', ao contrário de 'oração'.

[10] No capítulo intitulado 'A Sentença Complexa e sua Tipologia', Castilho (2012) afirma que o termo 'sentença complexa' é preferido no lugar de 'período', dado que este não apresenta uma unidade sintática diferente da sentença simples, "ou seja, tudo o que ocorre numa sentença simples ocorre numa sentença complexa"(p. 336). Dado que não há uma diferença além de terminológica em relação a esses dois termos, iremos utilizar em nossa análise os termos 'oração coordenada' e 'oração subordinadas' (períodos compostos) pelos mesmo motivo supracitado.

[11] Por ser menos sintética, ela nos permitiu compreender de forma mais substancial os problemas de segmentação linguística.

Sintagma Preposicional (SP)	O SP tem por núcleo a preposição, por Especificador o advérbio, enquanto que o Complementador em um SP também pode ser: (i) um verbo, como em 'para comer', (ii) um pronome, como em 'para mim', 'para quem'; (iii) um quantificador definido, como em 'para dois'; (iv) um quantificador indefinido, como em 'para muitos', dentre outras possibilidades de realização.
Sintagma Adverbial (SAdv)	O SAdv tem como núcleo o advérbio e como Especificador e Complementador outros advérbios. Sintaticamente, Castilho (2012) diz que os advérbios mantém relação entre si, com e com os adjetivos verbos, adjetivos.
Sentenças Complexas	Enunciados que possuem mais de um verbo, ou seja, contém mais de uma sentença, podendo estabelecer relação de coordenação, subordinação e correlação entre si.

A partir dessa releitura, algumas etiquetas formuladas por Chaves (2012) permaneceram intactas, sendo outras mudadas em relação à nomenclatura, tendo acontecido, também, alguns casos de aglutinação de diferentes etiquetas em apenas uma. Outras surgiram ainda pelos casos inéditos encontrados no *corpus* do presente estudo. Nossa anotação fez uso, em relação aos tipos de problema de segmentação, de apenas u etiqueta, que foi para o problema do tipo linguístico – <PROSEGL>, pois não mais etiquetamos os problemas de segmentação dos tipos retórico e visual. Quanto aos problemas de segmentação linguística, fizemos uso de 12 subetiquetas, contra as 19 de Chaves. No que diz respeito aos parâmetros técnicos, optamos por replicar as etiquetas presentes no trabalho de Chaves, acrescentando a etiqueta <3L>, que não apareceu em seu *corpus*. Apresentaremos todas as etiquetas nas Subseções 3.2.1 e 3.2.2.

Na próxima seção, iremos tratar das etapas do percurso metodológico, embasado, principalmente, em ferramentas oriundas da Linguística de Corpus.

3. Metodologia

A pesquisa teve suporte teórico-metodológico, como já mencionado, nos Estudos da Tradução, nos estudos em TAV-LSE e na Linguística de Corpus. A metodologia envolveu uma dimensão descritiva pautada por análises quanti-qualitativas baseadas em *corpus*.

3.1 O *Corpus*

O *corpus* é do tipo especializado, composto da LSE do tipo *closed caption pop-on* de um episódio da telenovela 'Amor Eterno Amor', produzida e exibida pela Rede Globo entre 5 de março e 7 de setembro de 2012, em 161 capítulos, escrita por Elizabeth Jhin e com

direção geral de Pedro Vasconcelos. Essa LSE faz parte do Projeto CORSEL (Corpus, Segmentação e Legendagem), que tem como objetivo a identificação e descrição-classificação dos problemas de segmentação na LSE da TV brasileira e é vinculado ao grupo de pesquisa LEAD-UECE. Para o projeto CORSEL, foram gravadas duas semanas de programação dos canais de TV do Brasil que oferecessem LSE do tipo *closed caption pop-on*. A partir da nossa pesquisa sobre quais itens de programação da TV brasileira possuíam legenda com os requisitos acima descritos, chegamos a quatro principais tipos de programa: novelas; seriados e programas de humor. Optamos por utilizar como *corpus* para o nosso recorte a LSE de um episódio da novela 'Amor Eterno Amor'. Escolhemos apenas um episódio da novela em questão, em um universo de 5 que serão analisados no âmbito do projeto, por questões de espaço e de não haver tempo hábil de análise de todos os episódios para essa publicação.

Consideramos a LSE de um episódio da novela em questão, com um total de 294 legendas e 5.181 palavras corridas[12] (*tokens*), como sendo representativa para a análise feita, mesmo constituindo-se em um *corpus* pequeno[13]. Esse nosso posicionamento se justifica pelo fato de que, como um *corpus* é uma amostra de um fenômeno linguístico cuja extensão não se conhece, não se pode estabelecer o tamanho ideal representativo do fenômeno a ser descrito, não existindo critérios objetivos para a determinação dessa representatividade, a não ser aqueles estipuladas pelo próprio analista. Ou seja, como pontua Leech (1991, p. 27 *apud* BERBER SARDINHA, 2004, p. 25), "os usuários de um *corpus* atribuem a ele a função de ser representativo de certa variedade".

Para a extração das legendas do tipo *closed caption*, o software *CCExctractor 0.61*[14] foi o escolhido. Apresentamos sua interface na Figura 2.

[12] Sendo computados os números presentes no corpus relativos à inserção da legenda e aos tempos iniciais e finais, que, se não contabilizados, somariam 2.534 palavras corridas.
[13] Ao observar corpora realmente utilizados pela comunidade, Berber Sardinha (2002 *apud* BERBER SARDINHA, 2004, p. 26), descreve um corpus com menos de 80 mil palavras como pequeno, por exemplo. Mas é claro que, quão mais especializado o corpus, como é o nosso caso, menores são as proporções do que seria considerado como um corpus pequeno ou grande.
[14] Programa *freeware* que pode ser baixado no domínio http://ccextractor.sourceforge.net/

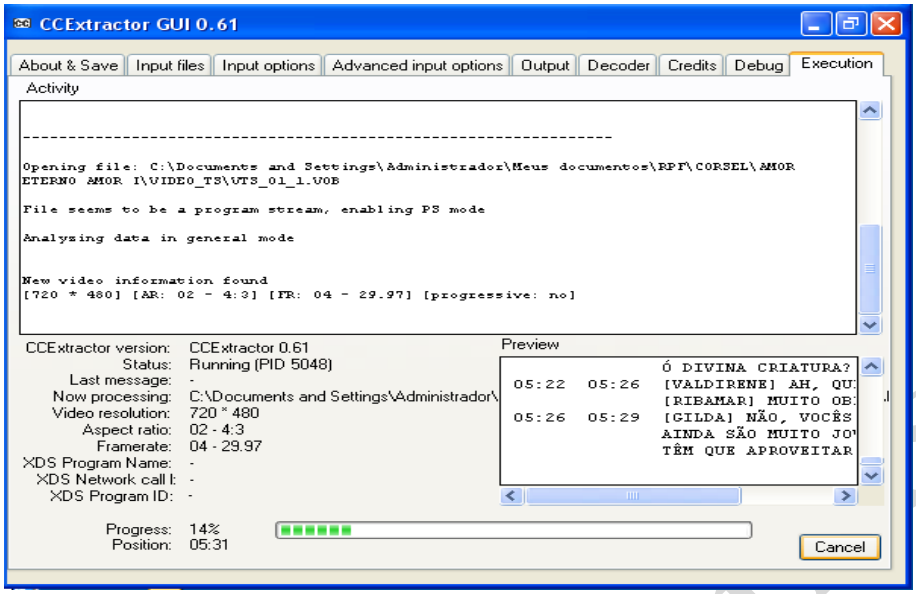


Figura 2: Interface *CCExtractor*.
Fonte: Os autores.

As vantagens do *CCExtractor* são inúmeras. Além de ter uma interface simples, sendo fácil de operar, o programa gera um arquivo de legendagem srt. com o número de inserção das legendas, assim como a marcação e sincronização das mesmas. Por vezes, contudo, o arquivo gerado apresenta pequenas falhas que precisam ser corrigidas manualmente, como no caso das legendas 1, 3, 6 e 7[15] na Figura 3.

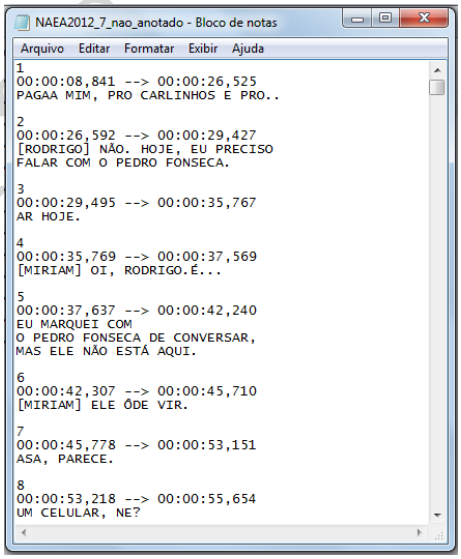


Figura 3: Arquivo de legenda em formato srt. aberto no Bloco de Notas.
Fonte: Os autores.

[15] A nossa experiência nos sugere que as legendas 1, 3, 6 e 7 apresentaram problemas, provavelmente, devido a instabilidade do sinal em que é veiculado a legendagem *closed caption*. Contudo, há a possibilidade de o erro ter ocorrido na própria confecção das legendas por parte da emissora, sendo difícil detectar com absoluta certeza em que instância desse processo o erro aconteceu.

Utilizando como exemplo a primeira inserção de legenda, podemos observar a primeira linha composta do número '1', indicando a posição daquela legenda na sequência de inserções. Logo abaixo, vemos a indicação do tempo de entrada daquela inserção '00:00:08, 841' e o tempo de saída '00:00:26,525', definindo assim a marcação e sincronização daquela legenda, ou seja, o tempo em que ela permanecerá na tela.

3.2 Etiquetagem

As etiquetas[16] foram feitas a partir do padrão *SGML* (*Standard Generalized Markup Language*), que fornece códigos escritos no formato <etiquetas de abertura>informação</etiqueta de fechamento/>, caracterizando-as e delimitando o conteúdo a ser analisado. Ainda com o intuito de replicar a metodologia formulada por Chaves (2012; CHAVES; ARAÚJO, 2014), utilizamos, na confecção das etiquetas, o sinal de (+) de modo a indicar a quebra indevida da estrutura sintagmática, caso em que ocorrem os problemas de segmentação linguística na maioria das vezes.

Após o processo de anotação do *corpus*, feita manualmente no arquivo txt aberto no programa Bloco de Notas[17], as etiquetas puderam ser tratadas de forma adequada pelo programa *Wordsmith Tools* 5.0. e analisadas com o auxílio ferramenta *Concord*[18], que, de acordo com Berber Sardinha (2009, p. 9), pode ser utilizada para realizar concordâncias no texto, listando palavras específicas (nódulos) juntamente com parte do texto ou cotexto em que a mesma ocorreu.

Agora, passamos a apresentar as etiquetas.

[16] Segundo a descrição de Tagnin (2010) sobre os tipos de etiquetagem, o tipo em que a nossa etiquetagem pode ser encaixado é no campo da "discursiva", processo pelo qual o conteúdo do corpus recebe etiquetas demarcando uma determinada parte do texto. Contudo, pela especificidade do projeto, podemos situar nossas etiquetas como 'etiquetas de segmentação'.
[17] Editor de texto disponível no sistema operacional *Windows*, da *Microsoft*.
[18] Em Assis (2013) foi utilizada também a *Wordlist*, que produz listas de palavras dos arquivos selecionados, contendo tanto suas frequências absolutas, quanto percentuais, de modo a fazer uma descrição geral das características quantitativas do corpus *per se*. Mas como esses resultados não apresentaram influência direta na análise da segmentação, optamos por deixá-los fora do estudo ora relatado.

3.2.1 Para os Parâmetros Técnicos

As etiquetas criadas para os parâmetros técnicos serviram, como já dito, para uma compreensão mais abrangente dos problemas de segmentação linguística. O Quadro 2 traz essas etiquetas.

Quadro 4: Etiquetas para parâmetros técnicos em legendagem.

ETIQUETAS DE ANÁLISE DE PARÂMETROS TÉCNICOS DA LEGENDAGEM	
Número da legenda	<sub1[19]>legenda1</sub1>
Linhas por legenda	<1L>, <2L> e <3L>
Tempos inicial e final de cada legenda	<t>início --> final</t>
Número de caracteres por linha (aplicada em legendas de 2 e 3 linhas)	<cpl[20]>
Velocidade da legenda baixa (145ppm[21]) Velocidade de legenda média (160ppm) Velocidade de legenda alta (180ppm)	<velocidade da legenda_baixa> <velocidade da legenda_média> <velocidade da legenda_alta>

Fonte: Os autores.

As etiquetas relativas ao número de linhas de cada inserção de legenda foram as seguintes: <1L>, <2L> e <3L>. Vale lembrar que a terceira não foi utilizada por Chaves (2012; CHAVES; ARAÚJO, 2014) por ausência de legendas de 3 linhas em seu *corpus*[22], característica esta comum na legendagem do tipo *closed caption pop-on* analisada aqui[23].

3.2.2 Para os Parâmetros Linguísticos

A base de sustentação para a formulação das etiquetas relacionadas à segmentação linguística, como já mencionado, são as considerações de Castilho (2012) acerca dos sintagmas e das orações coordenadas e subordinadas, as quais seguem a vertente linguística funcionalista-cognitivista do autor (ibidem, p. 32). As considerações do Castilho (ibidem), por sua vez, são respaldadas por exemplos a partir de excertos da norma culta da língua oral do

[19] Abreviatura de *subtitle*, 'legenda' em inglês.
[20] Acrônimo de 'caracteres por linha'.
[21] Acrônimo de 'palavras por minuto'.
[22] A LSE do filme 'Nosso Lar' seguiu os padrões da legendagem comercial, que permite no máximo duas linhas de legenda.
[23] A quantidade de 3 linhas por legenda não é recomendada por estudiosos de TAV por ser, segundo pesquisas (D'YDEWALLE et al., 1987), um fator que dificulta uma boa recepção.

português brasileiro, sendo, assim, ideais para o nosso propósito de identificar e analisar problemas de segmentação linguística na legendagem *closed caption pop-on* de uma novela da TV brasileira: são ideais porque, no Brasil, via de regra, as telenovelas mostram a norma culta oral e a LSE se constitui de tradução, no meio escrito, de textos produzidos o meio oral. O Quadro 5 traz as etiquetas.

Quadro 5[24]: Etiquetas para problemas de segmentação linguística em legendagem.

ETIQUETA INDICATIVA DE PROBLEMA DE SEGMENTAÇÃO LINGUÍSTICA
<PROSEGL>
SUBETIQUETAS PARA O SINTAGMA VERBAL (SV)
<SV_verbo+verbo> <SV_verbo+adv> → adv=advérbio <SV_(verbo)+oblíquo+verbo> <SV_negação+verbo>
SUBETIQUETA PARA O SINTAGMA NOMINAL (SN)
<SN_nominal_composto> <SN_especificador+subst> → subst=substantivo <SN_conj+subst> → conj=conjunção <SN_subst+adj> → adj=adjetivo
SUBETIQUETA PARA O SINTAGMA ADVERBIAL (SAdv)
<SAdv>
SUBETIQUETA PARA O SINTAGMA PREPOSICIONAL (SP)
<SP_prep+subst> → prep=preposição
SUBETIQUETA PARA A ORAÇÃO COORDENADA (COORD)
<COORD_conj+oração>
SUBETIQUETA PARA A ORAÇÃO SUBORDINADA (SUBORD)
<SUBORD_conj/pronome_rel+oração> → rel=pronome relativo

Fonte: Os autores.

Essas etiquetas foram elaboradas com o intuito de contemplarem aquela recomendação de Karamitroglou (1998) ao defender que, quando uma mesma legenda tiver que ser dividida em duas porções de texto em linhas distintas, essa quebra deve ocorrer no mais alto nível sintático. As etiquetas baseadas em Castilho (2012) viabilizam a operacionalização analítica da recomendação porque o que ela significa é que não devem ocorrer quebras entre os sintagmas constituintes de uma dada linha de legenda, como os verbais (SVs), nominais

[24] Não houve ocorrências de quebra do SAdj no corpus, sendo assim, a etiqueta correspondente a esse sintagma foi retirada do quadro de etiquetas.

(SNs), adjetivais (SAdj), adverbiais (SAdvs) e preposicionais (SPs), além das orações coordenadas e subordinadas, e que não seja quebrada a estrutura interna destes.

4. Problemas de Segmentação Linguística

4.1 Resultados da Etiquetagem

Nas subseções a seguir, serão apresentados os problemas de segmentação linguística identificados a partir das ocorrências de quebra de linha nas legendas em que ou as estruturas completas de sintagmas e/ou de orações foram indevidamente separadas ou a estrutura interna de sintagmas e/ou de orações foi desconsiderada. As porções de texto que deveriam aparecer preferivelmente na mesma linha de legenda foram sublinhadas. Serão apresentados, ainda, os resultados da análise descritivo-classificatória dos problemas identificados, o que será feito pela alocação dos problemas por tipo de sintagma e de oração em subseções distintas. Pela impossibilidade de se colocarem todas as ocorrências encontradas no *corpus* dada a restrição de espaço, optamos por trazer exemplos que mostrem apenas uma só categoria de problema por tipo de sintagma e oração. Para cada exemplo, mostramos também nossa proposta de 'ressegmentação'. Há ainda duas subseções dedicadas à análise da inter-relação dos parâmetros técnicos de velocidade e número de linhas da legenda com os problemas de segmentação.

4.1.1 Sintagma Verbal (SV)

Sobre as categorias de quebra problemática existentes no sintagma verbal, estas geralmente acontecem dentro de um sintagma verbal composto, normalmente composto de seu núcleo preenchido por verbo pleno numa forma nominal – infinitivo, particípio ou gerúndio – e especificado, ou seja, antecedido por um verbo auxiliar, como nos casos da quebra verbo+oblíquo+verbo, verbo+verbo etc. Outro caso considerado é o da quebra verbo+advérbio porque, apesar de não levar em conta constituintes que estejam relacionados dentro de uma estrutura verbal complexa, eles são tão inter-relacionados semanticamente (o segundo modifica o primeiro) que, quando os mesmos não ocorrem vizinhos um do outro em uma linha de legenda, o sentido imediato da linha pode ser comprometido. Na Figura 4, mostramos um fragmento da tela da ferramenta Concord com as ocorrências de problemas de segmentação no SV.

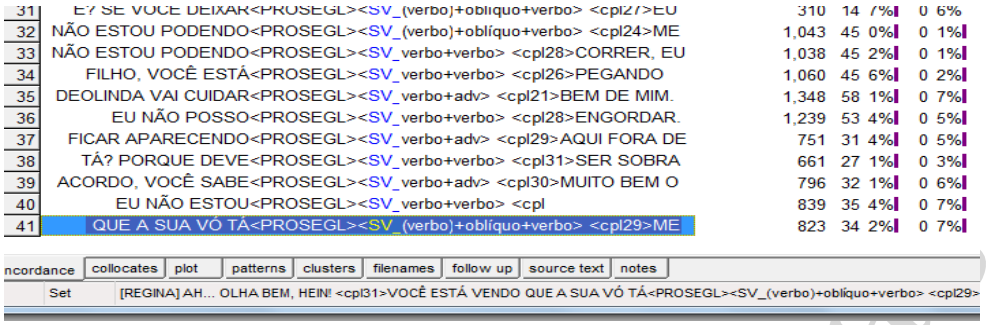


Figura 4: Fragmento de tela do *Concord* com ocorrências de problemas de segmentação linguística no SV.
Fonte: Os autores.

Ao todo, houve 41 problemas de segmentação na **ordem** do sintagma verbal, o que significa 50,6% dos problemas de segmentação na LSE de 'Amor Estranho Amor'. O Quadro 6 traz um exemplo da ocorrência da categoria representada pela etiqueta[25] <SV_verbo+verbo>, a mais presente em todo o *corpus* com 16 ocorrências ou 19,7%:

Quadro 6: Problema de segmentação linguística - quebra no sintagma verbal.

Nº da Legenda	**Número de Linhas**	**Velocidade da Legenda**	**Etiqueta**	**Legenda**
Sub247	2	Baixa	<SV_verbo+verbo>	[GRACINHA] JÁ VALE FICAR TRABALHANDO COM FEBRE, É?

Fonte: Os autores.

Para exemplo dos casos de quebra da estrutura verbo+verbo, observemos a legenda 247, que possui um sintagma verbal complexo – 'vale ficar trabalhando' – e um advérbio anteposto – 'já', os quais deveriam permanecer juntos para que o legendista tivesse seguido a recomendação de Karamitoglou (1998). Se considerarmos o parâmetro da sintaxe na segmentação linguística, esta legenda deveria ficar assim:

[GRACINHA] JÁ VALE FICAR TRABALHANDO
COM FEBRE, É?

[25] Nos casos em que a legenda contém mais de um problema de segmentação linguística, apenas o problema do sintagma discutido na subseção correspondente é ressaltado com o sublinhado.

É interessante notar que, nessa ressegmentação, a segmentação linguística prevalece, como defende Karamitroglou (1998), em relação ao formato considerado ideal pelo teórico: linha de cima e de baixo quase do mesmo tamanho, conforme os tipos de formato descritos no Quadro 1. Como dissemos anteriormente, esse não foi o padrão observado na legendagem de 'Amor Eterno Amor' como um todo.

4.1.2 Sintagma Nominal (SN)

Quanto às categorias de quebra problemática no âmbito do sintagma nominal, identificamos quebras entre o i) Especificador e o núcleo do SN, ii) o núcleo do SN e o adjetivo que o modifica, iii) conjunções e os substantivos a eles relacionados, iv) além dos casos que denominamos 'nominais compostos'[26]. As quebras problemáticas na ordem do sintagma nominal aconteceram 26 vezes, isto representando cerca de 32,1% do total de má segmentação linguística na LSE de 'Amor Estranho Amor'. O Quadro 7 apresenta um exemplo da categoria anotada pela etiqueta <SN_nominal_composto>, que, com 13 ocorrências ou 16%, ranqueou em segundo lugar:

Quadro 7: Problema de segmentação linguística - quebra no sintagma nominal.

Nº da Legenda	**Número de Linhas**	**Velocidade**	**Etiqueta**	**Legenda**
Sub121	2	Alta	<SN_nominal_composto>	NÃO VAI TER EQUIPAMENTO DE SEGURANÇA PRA TODO MUNDO.

Fonte: Os autores.

Nesse exemplo, relativo à legenda 121, a estrutura sublinhada que deveria estar unida em uma mesma linha corresponde ao sintagma nominal 'equipamento de segurança'. Defendemos a ideia de que o exemplo em questão se configura como uma quebra na ordem do SN se considerado o seguinte: no momento da leitura da legenda, o telespectador se depara com o substantivo 'equipamento' e se pergunta "que equipamento?". A resposta só surge na

[26] Essa etiqueta serviu de 'guarda-chuva', digamos, por ter aglutinado as seguintes etiquetas de Chaves: <SN_superlativo+adj> <SN_nominal+modif/modif+nominal >, <SN_nomepróprio>,<SN_título+nome próprio>, de modo que pudesse abarcar também a terminologia de Castilho (2012). Além disso, ela serviu para nomear os casos em que o SP, enquanto Complementador do sintagma nominal, que cumpre função adjetiva é separado do núcleo do SN. Contudo, em reuniões do nosso grupo de pesquisadores, essa etiqueta tem se demonstrado problemática, dado, que com um nome mais geral, tenta abarcar muitas situações específicas. É certo que em futuros trabalhos essa etiqueta será remodelada e diluída em outras.

linha seguinte, onde está o sintagma preposicional 'de segurança', que, na verdade, funciona como sintagma adjetival por conferir uma qualidade/característica ao substantivo. Uma possibilidade de quebra que conciliaria melhor a questão da sintaxe e a facilitação do processo de compreensão da porção de legenda em questão está logo abaixo:

NÃO VAI TER
EQUIPAMENTO DE SEGURANÇA PARA TODOS.

Neste caso, o formato não é o retangular preconizado por Karamitroglou como o que melhor representa uma boa segmentação. Porém, esse formato triangular permite uma boa visualização para uma legenda com a linha superior mais curta, conforme o Quadro 1.

4.1.3 Sintagma Preposicional (SP)

Utilizando a etiqueta <SP_prep+subst>, identificamos a categoria de quebra problemática do sintagma preposicional que diz respeito à separação entre o seu núcleo e o seu Complementador, que, no caso do *corpus* do estudo ora relatado, só se realizou por substantivo. Essa categoria ocorreu 6 vezes ou cerca de 7,4% do total de má segmentação linguística na LSE de 'Amor Estranho Amor'. A categoria em apreço está exemplificada no Quadro 8.

Quadro 8: Problema de segmentação linguística - quebra no sintagma preposicional.

Nº da Legenda	**Número de Linhas**	**Velocidade**	**Etiqueta**	**Legenda**
Sub4	3	Alta	<SP_prep+subst>	O GUARDINHA ME PAROU POR CAUSA DE UMA BOBAGEM DA PLACA QUE CAIU!

Fonte: Os autores.

Podemos observar na legenda 4 que há a separação, indevida a nosso ver, da preposição 'de' e do sintagma nominal 'uma bobagem'. A seguir, fazemos uma proposta de como a legenda poderia ficar se considerado o parâmetro linguístico:

O GUARDINHA ME PAROU POR CAUSA
DE UMA BOBAGEM DA PLACA QUE CAIU!

4.1.4 Sintagma Adverbial (SAdv)

Aqui, consideramos como quebra problemática relacionada ao sintagma adverbial a segmentação de estruturas adverbiais compostas, isto é, sempre que houve a separação entre dado advérbio e outro lhe acrescenta informação. Essa categoria foi anotada pela etiqueta <SAdv> e houve apenas 1 ocorrência, representando 1,2% dos problemas de segmentação linguística na LSE de 'Amor Estranho Amor'. A única ocorrência encontra-se no Quadro 9.

Quadro 9: Problema de segmentação linguística - quebra no sintagma adverbial.

Nº da Legenda	**Número de Linhas**	**Velocidade**	**Etiqueta**	**Legenda**
Sub131	3	Alta	<**SAdv**>	ENTROU COM UM PROCESSO LÁ NA EMPRESA EM QUE ELE TRABALHAVA, GANHOU UMA GRANA FERRADA.

Fonte: Os autores.

A legenda 131 contém um advérbio de lugar – 'lá' – e um sintagma preposicional com função de SAdv igualmente de lugar – 'na empresa' – que, juntos, formam uma só estrutura adverbial complexa. Como os dois segmentos estão em linhas distintas, houve a separação entre um advérbio e aquele que o modifica. A nosso ver, portanto, o SAdv complexo 'lá na empresa' deveria permanecer inseparável em uma única linha. Por se tratar de uma legenda com densidade lexical alta (compõe-se de 83 caracteres quando o máximo deveria ter 78 caracteres ou 39 caracteres por linha como informamos na Subseção 2.2), propor uma ressegmentação sem considerar o tempo disponível para a marcação é uma tarefa complicada. A legenda 131, provavelmente, teria que ou ser bem condensada ou ser ressegmentada em duas porções de legenda. Como nossas ressegmentações são apenas a título de exemplificação, podemos propor uma que seja condensada e respeite a segmentação linguística sem dividir a mesma em mais legendas. Contudo, para isso, achamos necessário retirar o advérbio 'lá', redundante no contexto:

ENTROU COM UM PROCESSO NA EMPRESA
EM QUE TRABALHAVA E GANHOU BOA GRANA.

4.1.5 Orações Subordinadas (SUBORD) e Coordenadas (COORD)

A identificação de quebras problemáticas em orações coordenadas e subordinadas foi um dos aspectos analíticos mais complicados, pois há uma instrução clara nos manuais de legendagem de, sempre que possível, deixar cada uma das orações que formam um período composto em uma linha da legenda, de forma a colocar o sentido completo de uma oração em cada linha (IVARSSON; CAROLL, 1998). Conciliar essa instrução com uma boa segmentação linguística foi um desafio.

Interpretando o trabalho de Chaves (2012; CHAVES; ARAÚJO, 2014) e de Perego (2008) sobre a segmentação em orações coordenadas e subordinadas e utilizando as perspectivas gramático-funcionais de Castilho (2012) sobre as mesmas, definimos como problemas de segmentação apenas os casos em que as conjunções coordenativas e subordinativas – incluindo nestas os pronomes relativos – aparecem desconectadas da oração coordenada ou subordinada que iniciam. Quebras problemáticas na ordem da oração complexa relativamente à oração subordinada ocorreram 6 vezes, representando 7,4% das ocorrências totais na LSE de 'Amor Estranho Amor'. O mesmo tipo de quebra, mas relacionada à oração coordenada, aconteceu apenas 1 vez ou cerca de 1,2% das ocorrências totais de quebras problemáticas no *corpus*. O Quadro 10 traz um exemplo da categoria representada pela etiqueta <SUBORD_conj/pronome_rel+oração>.

Quadro 10: Problema de segmentação linguística - quebra em oração subordinada.

Nº da Legenda	**Número de Linhas**	**Velocidade**	**Etiqueta**	**Legenda**
Sub21	3	Alta	<SUBORD_conj/pronome_ rel+oração>	NÃO TEM NADA NESSE MUNDO QUE SUBSTITUA CARINHO DE MÃE, NÉ?

Fonte: Os autores.

Nesse exemplo, ocorre separação entre o pronome relativo 'que' e o restante da oração subordinada por ele introduzida: 'que substitua carinho de mãe, né?'. Uma sugestão para a ressegmentação da legenda 21 seria:

NÃO TEM NADA NESSE MUNDO
QUE SUBSTITUA CARINHO DE MÃE, NÉ?

4.1.6[27] Quebras Problemáticas por Número de Linhas das Legendas

A análise mostrou que 72 inserções tiveram 1 linha de legenda ou 24,5% do total. A maior parte do *corpus* é composta de legendas de 2 linhas, num total de 137 ocorrências ou 46,6 % do total. As legendas de 3 linhas somaram 85 ocorrências ou 28,9% do total de inserções. A Tabela 1 mostra a relação entre os problemas de segmentação e o número de linhas das legendas.

Tabela 1: Ocorrências de problemas de segmentação por número de linhas das legendas.

Número de linhas das legendas	Percentual de ocorrência de quebras problemáticas
2L	33,3%
3L	66,7%

Fonte: Os autores.

Os problemas de segmentação linguística só apareceram em legendas de 2 e 3 linhas, já que a segmentação aqui analisada é a de quebra de linha, que só ocorre em legendas com mais de 1 linha. 54 ocorrências de problemas de segmentação de foram em legendas de 3 linhas, o que significa 66,7% do total. Os outros 27 problemas apareceram em legendas de 2 linhas, o que representa 33,3% das ocorrências.

4.1.7 Quebras Problemáticas por Nível de Velocidade das Legendas

Quanto à velocidade, em um universo de 294 legendas, a maior frequência de ocorrência no *corpus* foi de legendas com velocidade alta, em um total de 169 ou 57,5% do total. A segunda maior foi a de legendas com velocidade baixa, perfazendo 95 inserções ou 32,3% do total. Por último, as legendas com velocidade média foram em número de 30, ou 10,2% do total. Abaixo, apresentamos a Tabela 2 com os percentuais de problemas de segmentação por nível de velocidade.

[27] Para efeitos de análise, apenas alguns dos parâmetros técnicos foram levadas em conta, como caracteres por linha, número de linhas e velocidade da legenda, por apresentarem relações mais intrínsecas com o parâmetro da segmentação linguística. Os outros parâmetros etiquetados foram importantes para a localização da legenda no corpus, como a etiqueta que indica o número da legenda e a que delimita os tempos inicial e final da legenda. Apesar de ter sido importante para o cálculo da velocidade da legenda, o parâmetro técnico de 'caracteres por linha' (<cpl>), por não ter apresentado uma relação tão forte com a segmentação como os parâmetros de velocidade e número de linhas, decidimos, pelo espaço limitado não incluí-lo nas subseções de inter-relação entre parâmetros técnicos e problemas de segmentação.

Tabela 2: Ocorrências de problemas de segmentação por nível de velocidade.

Nível de velocidade das legendas	Percentual de ocorrência de quebras problemáticas
Alto	76,5%
Médio	12,4%
Baixo	11,1%

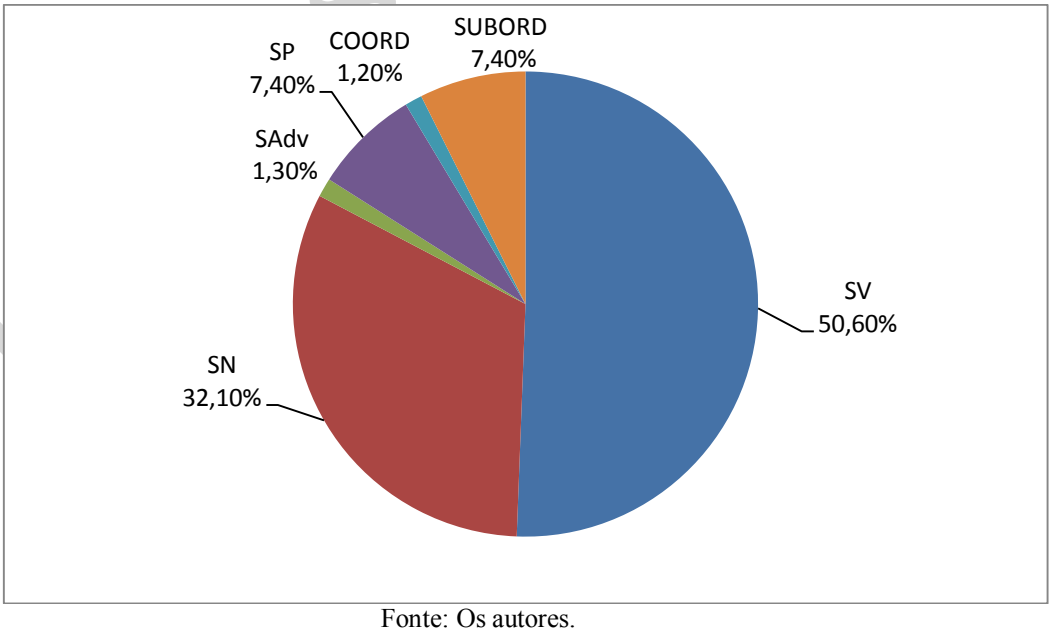
Fonte: Os autores.

A tabela mostra que, no que concerne à relação entre as velocidades das legendas e os problemas de segmentação encontrados, foi nas legendas com velocidade alta que os problemas predominaram: 62 ou 76,5% das ocorrências. O segundo lugar foi ocupado pelas legendas com velocidade média, com 10 problemas ou 12,4% das ocorrências. Em último lugar, ficaram a legendas com velocidade média, com 9 problemas ou 11,1% das ocorrências.

4.2 Sistematização e Discussão Geral dos Resultados

Sistematizando os resultados quantitativos decorrentes dos dados provenientes da análise baseada em *corpus* que viabilizou a identificação e a descrição classificatória dos problemas de segmentação linguística na LSE da novela 'Amor Eterno Amor', temos o Gráfico 1:

Gráfico 1: Sistematização dos resultados dos problemas de segmentação linguística na LSE da novela 'Amor Eterno Amor'.



Fonte: Os autores.

Como é possível observar no Gráfico 1, o maior número de problemas ocorre nas ordens do SV, com contribuição de 50,6%, e do SN, com participação de 32,1%, o que corrobora os resultados de Chaves (CHAVES, 2012; CHAVES; ARAÚJO, 2014). Além disso, assim como no trabalho de Chaves, a estrutura em que houve mais ocorrências de quebras problemáticas foi em verbo+verbo, assim como a maior parte dos problemas de segmentação linguística encontrados aconteceram em legendas com velocidade considerada alta (cf. Subseção 2.3).

Os resultados semelhantes de que obtivemos em relação à Chaves sugerem, mesmo em gêneros ficcionais audiovisuais diferentes (telenovela e filme em DVD), o que pode se apresentar como uma tendência para a segmentação linguística em gêneros ficcionais. Porém, ainda é cedo para apontar de forma substancial padrões característicos da segmentação linguística no Brasil, o que será possível com o aumento de *corpus* analisado em tamanho e variedade de gênero.

Contudo, apesar das similaridades entre os resultados de Chaves e o presente trabalho, é inegável, se examinados os números provenientes da análise da LSE no capítulo da novela 'Amor Eterno Amor', a grande relevância que os problemas de segmentação linguística têm no *corpus*. Ao todo, identificamos 81 problemas, todos anotados pela etiqueta <PROSEGL>. Estes, de uma forma geral, aparecem em 75 legendas em um universo de 294, representando 25,5% do *corpus*. Ou seja, a cada quatro legendas, um pouco mais de uma possui problema de segmentação, teoricamente dificultando bastante a compreensão do texto traduzido. Nas legendas de 2 linhas, cerca de 19,7% delas (27 casos em 27 legendas diferentes) possuem algum tipo de problema de segmentação, seja na ordem do sintagma ou da oração complexa. Nas legendas de 3 linhas, a porcentagem foi ainda maior, com 56,4% destas (54 casos em 48 legendas diferentes) apresentando uma ou mais ocorrências de algum tipo de problema de segmentação linguística, já que algumas legendas de 3 linhas apresentaram mais de um caso.

O grande número de legendas com uma velocidade considerada alta e de 3 linhas não foi algo de inteiro surpreendente, já que isso só reforça a noção equivocada da legendagem para surdos da TV brasileira, em desacordo com os parâmetros estabelecidos pelos pesquisadores da área. A alta velocidade das legendas é resultado previsível, se considerado que a legenda *closed caption* no Brasil é muito mais uma transcrição da fala do

que um processo tradutório que siga os preceitos dos estudiosos da Tradução Audiovisual, especificamente a LSE

Com todos esses resultados, pudemos identificar, de uma forma mais geral, o padrão de má segmentação linguística no *corpus*: na maioria das vezes, nos sintagmas verbal e nominal em legendas de 3 linhas e com velocidade alta. O fato de a segmentação ter sido problemática nas legendas mais longas – as de 3 linhas – e de velocidade alta se impõe como fator complicador para a recepção de produtos legendados, de acordo com o que estipula a literatura da área em relação à LSE. A partir dos resultados obtidos, podemos caracterizar a LSE no capítulo de 'Amor Eterno Amor' em questão como fora dos padrões de uma legendagem confortável de ser consumida pelos telespectadores surdos e ensurdecidos.

5. Considerações Finais

Nossas conclusões sobre a segmentação linguística estudada na telenovela 'Amor Eterno Amor' só foram possíveis graças à metodologia baseada em *corpus* proposta por Chaves (2012; CHAVES; ARAÚJO, 2014) para a análise do parâmetro linguístico em questão, mesmo com os poucos dados analisados aqui. Isso foi motivado, contudo, pelo tamanho e tempo restritos da própria pesquisa. A partir disso, nos foi mais interessante estudar uma pequena quantidade de legendas, que nos possibilitou empreender uma análise que desse mais vazão à interpretação qualitativa de nossos dados. Isso seria difícil de fazer com um grande *corpus* em tão pouco tempo, sendo, portanto, preferível à fazer um simples empilhamento de dados quantitativos.

A metodologia baseada em *corpus* também nos permitiu observar, de forma concomitante à análise da segmentação linguística em 'Amor Eterno Amor', uma amostra das características técnicas da LSE da TV Brasileira, contribuindo, assim, com nossos resultados, para a pesquisa em LSE realizada na UECE desde 2002, a qual visa estudar como esse tipo de Legendagem caracteriza-se e como ela é recebida pelo seu público alvo, os surdos e ensurdecidos. O Projeto CORSEL, ao qual os resultados desta pesquisa estão atrelados, está filiado nessa perspectiva. Este, ao seu fim, permitirá uma caracterização mais conclusiva a respeito das características técnicas e de como acontece a segmentação linguística na LSE do Brasil a partir da abrangência em quantidade e diversidade genérica de narrativas audiovisuais que o projeto pretende abarcar.

A análise dos parâmetros técnicos da legendagem *closed caption* no *corpus* foi bastante importante para a compreensão dos problemas de segmentação linguística encontrados, tendo embasado nossas considerações acerca destes e possibilitado uma correlação entre ambos. Para além disso, essa análise propiciou o que parece ser as características técnico-linguísticas da LSE na TV brasileira: legendas em sua maioria de velocidade alta, de 3 linhas e com uma grande quantidade de problemas de segmentação de ordem linguística. Do ponto de vista dos parâmetros linguísticos, os resultados apontaram o sintagma verbal e o nominal como os mais problemáticos em relação aos casos de má segmentação no *corpus*, tal como aconteceu em Chaves (2012; CHAVES; ARAÚJO, 2014).

Os dados resultantes do estudo que acabamos de relatar e dos demais no âmbito do CORSEL poderão, no futuro, possibilitar novas pesquisas dentro da interface Linguística de Corpus/LSE. Entre estas, por exemplo, está a elaboração e compilação de um *corpus* eletrônico de LSE acessível por computador, visando a consulta por legendistas e/ou seu treinamento através da plataforma do Projeto CoMET – Corpus Multilíngue para o Ensino e Tradução[28], um *corpus* monolíngue e paralelo voltado para pesquisadores e profissionais de legendagem, com o qual já mantemos *link* acadêmico. Outra possibilidade de utilização dos dados do Projeto CORSEL, assim como da presente pesquisa, é na investigação experimental com o auxílio de rastreamento ocular para o estudo da real influência do parâmetro da segmentação na recepção de legendas por pessoas surdas-ensurdecidas e ouvintes. Essa pesquisa se faz necessária, pois, apesar de ser um parâmetro considerado como essencial para a confecção de uma boa legendagem, o mesmo ainda carece de pesquisas experimentais que possam mensurar, de fato, a sua contribuição no processo de compreensão do produto legendado, sendo este um dos próximos passos da pesquisa em LSE pelo grupo LEAD na UECE, com auxílio da metodologia fornecida pela Linguística de Corpus.

Referências Bibliográficas

AMOR ETERNO AMOR. Direção: Pedro Vasconcelos. Brasil: TV Globo, 2012, 161 capítulos. sonor. color. Legenda CC disponível.

ARAÚJO, V. L. S. Closed subtitling in Brazil In: ORERO, P. (org.). **Topics in Audiovisual translation**. Amsterdã: John Benjamins Publishing Company, v. 1, p. 199 - 212, 2004. crossref http://dx.doi.org/10.1075/btl.56.20san

[28] http://www.fflch.usp.br/dlm/comet/

ARAÚJO, V.L.S. Subtitling for the deaf and hard- of - hearing in Brazil In: ORERO, P.; REMAEL, A. (orgs.). **Media for All**: Subtitling for the Deaf, Audio Description and Sign Language. Kenilworth: Nova Jersey, EUA: Rodopi, v. 30, p. 99- 107, 2007.

ARAÚJO, V. L. S. Por um modelo de legendagem para Surdos no Brasil. In VERAS, V. (org.). **Tradução e Comunicação**, Revista Brasileira de Tradutores, São Paulo: UNIBERO, n. 17, p. 59–76, 2008.

ARAÚJO, V. L. S.; NASCIMENTO, A. K. P. Investigando parâmetros de legendas para Surdos e Ensurdecidos no Brasil. In: FROTA, M. P.; MARTINS, M. A. P. (orgs.). **Tradução em Revista**, v. 2, p. 1- 18, 2011.

ARAÚJO, V. L. S. **Legendagem para surdos**: em busca de um modelo para o Brasil. Relatório Técnico n. 306948/2008-7. Fortaleza: CNPq. Fev/2012.

ASSIS, Í. A. P. **A segmentação na LSE de Amor eterno Amor**: uma análise baseada em *corpus*. (Bacharelado em Letras Inglês). Universidade Estadual do Ceará, Fortaleza-CE, 2013.

BAKER, M. Corpus Linguistics and Translation Studies: Implications and Applications. In: Baker, M.; Francis, G.; Tognini - Bonelli, E. (orgs.). **Text and technology:** In honour of John Sinclair. Philadelphia, Amsterdam: John Benjamins, p. 233- 250, 1993. crossref http://dx.doi.org/10.1075/z.64.15bak

BERBER SARDINHA, T. **Linguística de Corpus**. Barueri, São Paulo: Manole, 2004, 410 p.

BERBER SARDINHA, T. **Pesquisa em Linguística de *Corpus* com WordSmith Tools**. Campinas: Mercado de Letras, 2009, 299 p.

CASTILHO, A. de. **Nova Gramática do Português Brasileiro**. São Paulo: Contexto, 2012, 768p.

CHAVES, E. G. **Legendagem para Surdos e Ensurdecidos**: um Estudo Baseado em *corpus* da segmentação nas legendas de filmes brasileiros em DVD. 126f. Dissertação (Mestrado em Linguística Aplicada). Universidade Estadual do Ceará, Fortaleza- CE, 2012.

CHAVES, E. G.; ARAÚJO, V. L. S. Segmentation tags: a proposal for the analysis of subtitles. In: ALUÍSIO, S. M.; TAGNIN, S E. O. (orgs.) **New language, technologies and linguistic research**: a two way road. Newcastle upon Thyne: Cambridge Scholar's Publishing, 2014, p. 62-75.

DIAZ- CINTAS, J.; REMAEL, A. **Audiovisual Translation**: Subtitling. Manchester, UK, Kinderhook, N Y, UK : St. Jerome Publishing, 2007, 272 p.

DINIZ, N. S. L. **A Segmentação em Legendagem para Surdos e Ensurdecidos:** um Estudo Baseado em *Corpus*. 149f. Dissertação (Mestrado em Linguística). Universidade Federal de Minas Gerais, Belo Horizonte-MG, 2012.

FRANCO, E.; ARAUJO, V. L. S. Reading Television: Checking deaf people's Reactions to Closed Subtitling in Fortaleza, Brazil. In: GAMBIER, Y. (org.). **The Translator,** v. 9, n. 2, p.249- 267, 2003.

GOTTLIEB, H. Subttling: Diagonal Translation. In: **Perspective in Translatology,** v.2, n.1, p.101-121, 1994.

IVARSSON, J.; CARROLL, M.; **Subtitling**. Simrishamm, Suécia: TransEditHB, 1998,184 p.

KARAMITROGLOU, F. A Proposed Set of Subtitling Standards in Europe. In: **Translation Journal**, v. 2, n. 2, p. 1- 15, 1998. Disponível em: <http://translationjournal.net/journal//04stndrd.htm > Acesso em: 15 de Março de 2011.

PEREGO, E. What Would We Read Best? Hypotheses and Suggestions for the Location of Line Breaks in Film Subtitles. In: **The Sign Language Translator and Interpreter**. Manchester, UK : St. Jerome Publishing, p. 35- 63, 2008.

PERINI, M. A. **Gramática do português brasileiro**. São Paulo: Parábola Editorial, 2010, 366 p.

REID, H. Literature on the screen: subtitle translation for public broadcasting. In: BART, W.; D'HAEN, T. (Eds.). **Something understood**. Studies in Anglo- Dutch literary translation.Amsterdam: Rodopi, p. 97- 107, 1990.

SANTOS, D. Corporizando algumas questões. In: TAGNIN, S.E.O.; ARAÚJO, O. V. (orgs.). **Avanços da Linguística de *Corpus* no Brasil.** São Paulo: Humanitas, p. 41-55, 2008.

TAGNIN, S. E. O. Glossário de Linguística de *Corpus*. In: VIANA, V.; TAGNIN, S. E. O. (orgs.). ***Corpora* no ensino de línguas estrangeiras**. São Paulo: Hub Editorial, p. 357- 361, 2010.

WORDSMITH Tools 5.0: sítio do programa disponível em: <http://www.lexically.net/wordsmith/index.html>. Acesso em: 17 de março de 2013.

Artigo recebido em: 13.10.2014
Artigo aprovado em: 14.12.2014

Metáforas e domínios narrativos numa perspectiva da Linguística de *Corpus*[1]

Metaphors and narrative domains on an approach of Corpus Linguistics

Heberth Paulo de Souza*

RESUMO: Neste artigo, desenvolve-se uma abordagem da metáfora no escopo da cognição humana, utilizando os pressupostos teóricos da Linguística Cognitiva no âmbito das representações mentais e aplicando-os à descrição da articulação textual. Para alcançar esse intento, centramo-nos em alguns postulados do final do século XX que a consideram como um recurso de facilitação do raciocínio, através do qual conceitos mais complexos são elaborados na forma de conceitos mais simples. Considera-se também que a metáfora é um fenômeno presente em todos os níveis da comunicação, não se restringindo a algumas áreas e atividades do conhecimento humano. Baseando-se especialmente na Teoria dos Espaços Mentais, de Fauconnier (1994), e na Teoria da Mesclagem Conceitual, de Fauconnier e Turner (1994), a pesquisa desenvolveu-se sobre um corpus pequeno-médio, descrevendo o papel que a metáfora exerce na articulação textual. Com o suporte dos recursos eletrônicos do programa WordSmith Tools©, obteve-se uma sistematização de dados quantitativos para se proceder à pesquisa qualitativa, a partir de onde foi possível alcançar os resultados apresentados na tese de doutorado que deu origem a este artigo. Entre estes, destaca-se a constatação de que, subjacente à estruturação textual dos exemplares do nosso corpus, bem como em outros tipos textuais que também foram submetidos à análise, existe uma forma de organização de elementos típica do processo de narração, com a identificação de

ABSTRACT: In this paper an approach to metaphor, in the scope of human cognition, is developed, taking into account theoretical assumptions of Cognitive Linguistics within the ambit of mental representations, and having them applied to a description of textual articulation. In order to achieve this goal, we focused on a few late twentieth-century postulates. Metaphor is thus assumed to be a resource used to facilitate reasoning by means of which more complex concepts are elaborated in terms of more simple ones. Metaphor is also considered as a phenomenon present in all levels of communication, not being restricted to specific areas and activities of human knowledge. Based especially on both Mental Spaces Theory, by Fauconnier (1994), and on Conceptual Blending Theory, by Fauconnier and Turner (1994), the research was developed with the use of a small-medium-size corpus, describing the role that metaphor plays in textual articulation. With the aid of the electronic tool WordSmith Tools© it was possible to obtain a systematization of quantitative data in order to proceed to the qualitative research, from where the results in the original doctoral thesis were made available. It thus becomes clear that, underlying the textual structuring of the samples of our corpus, as well as in other textual types that were also submitted to analysis, there is a pattern of organization considered as typical of the narrative process that includes the identification of information relating to time, space and characters,

[1] Artigo-síntese de tese de doutorado defendida pelo POSLIN – Programa de Pós-graduação em Estudos Linguísticos da UFMG, em 27/08/2010, intitulada "A metáfora e a formação de esquemas narrativos em textos escritos de língua portuguesa".

* Doutor em Linguística pela UFMG – Universidade Federal de Minas Gerais. Professor de Língua Portuguesa e Metodologia Científica no IPTAN – Instituto de Ensino Superior Presidente Tancredo de Almeida Neves (São João del-Rei – MG).

informações relacionadas a tempo, espaço e personagens, considerando-se a inter-relação do nível metafórico e do não metafórico.

considering the interrelationship between metaphorical and nonmetaphorical levels.

PALAVRAS-CHAVE: Metáfora. Narrativa. Espaços Mentais. Mesclagem Conceitual. Corpus linguístico.

KEYWORDS: Metaphor. Narrative. Mental Spaces. Conceptual Blending. Linguistic corpus.

1. Introdução

Desde que foi evidenciada no campo dos estudos da Filosofia e da linguagem, a metáfora vem intrigando vários pesquisadores em virtude de algumas características que a vinculam a fenômenos diversos, tais como: a importância da metáfora na determinação de aspectos estilísticos do discurso, o papel da metáfora na oratória e na retórica, o papel das construções metafóricas na determinação do caráter literário ou não literário de um texto, a relação entre metáfora e o modo de pensar do ser humano, a inteligibilidade das ideias contidas num texto em função do seu nível metafórico, entre outros. Não foi por outro motivo que, a partir mesmo dos filósofos da Antiguidade clássica, a metáfora vem sendo abordada, ao longo de mais de vinte séculos, sob as mais variadas perspectivas, revelando-se sempre novas facetas desse fenômeno da linguagem, as quais se complementam, alternam-se e geram novas pesquisas sobre o tema.

Na segunda metade do século XX, os estudos da linguagem assistiram ao advento de uma área que finalmente se consolidou como distinta de outros ramos da Linguística e que tem como preceito básico o estudo do funcionamento da mente humana: a Linguística Cognitiva, que, embora também apresente diferentes modos de tratar os processos mentais que subjazem à comunicação humana, alcançou um desejável nível de solidez através da gramática cognitiva de Langacker (1987 e 1991). Nesses dois volumes, o autor descreve sistematicamente a relação existente entre a linguagem verbal e o modo como processamos as noções expressas pelo signo linguístico, proporcionando entendimentos e análises que até então ficavam circunscritas à nossa intuição como usuários e analistas da linguagem.

Paralelamente aos postulados da gramática cognitiva de Langacker, num outro contexto de estudos, Gilles Fauconnier desenvolvia importantes pesquisas sobre os espaços mentais, categorias de ordem cognitiva que também viriam a proporcionar entendimento mais conciso sobre a metáfora e outros fenômenos da linguagem, consolidando-se com a

publicação de 1994[2]. A identificação desses domínios cognitivos (espaços mentais), cujas características serão descritas neste trabalho, acrescentou à Linguística Cognitiva – mais especificamente à Semântica Cognitiva – um arcabouço teórico e um modelo de análise que adiantaram demasiadamente nossa compreensão da inter-relação entre linguagem e cognição humana.

Além da metáfora, vários outros fenômenos e elementos linguísticos puderam ser descritos com a fidedignidade de quem pretende conferir à linguagem verbal um *status* não fragmentado e não desvinculado de outras habilidades humanas. Nesse contexto, beneficiaram-se também sobremaneira da teoria dos espaços mentais as pesquisas que vinham sendo empreendidas em torno da narrativa – não em sua clássica abordagem dentro de gêneros textuais, mas também enquanto um fenômeno da cognição humana. Nesse sentido, Turner (1996) realizou um empreendimento de fundamental importância, demonstrando que a mente do ser humano é literária (entenda-se "narrativa") por excelência; em outras palavras, mesmo quando não estamos lidando diretamente com uma narrativa clássica, nosso modo de pensar é tipicamente narrativo.

Não é por outro motivo que Fauconnier e Turner estabeleceram uma parceria em suas pesquisas que vem rendendo, até recentemente, muitas publicações conjuntas[3]. A partir de Fauconnier e Turner (1994), a Semântica Cognitiva se viu enriquecida com o estabelecimento da teoria da mesclagem conceitual, um modelo de análise que discrimina a inter-relação de diferentes espaços mentais na produção do sentido, que se aplica especialmente no caso de sentidos emergentes, ou seja, ausentes nos espaços de origem, que se revelam numa sequência enunciativa em função principalmente de condições contextuais.

Até este ponto, vimos *en passant* o surgimento de duas importantes teorias que não foram criadas propriamente no bojo dos estudos de *corpora*, mas a nossa pesquisa tenciona justamente empreender uma análise no esteio da Semântica Cognitiva utilizando-se dos benefícios proporcionados por um segmento que nos capacita a lidar com um grande número de textos, cruzar dados quantitativos que proporcionem importantes questionamentos e hipóteses de cunho qualitativo e cujos resultados sejam atinentes a enunciados autênticos.

[2] Trata-se da publicação de Fauconnier (1994), obra que, na verdade, tinha sido publicada dez anos antes em língua francesa (FAUCONNIER, 1984), mas que atingiu patamares acadêmicos mais amplos a partir da publicação mais recente em inglês.
[3] Entre elas, Turner e Fauconnier (1995), Fauconnier e Turner (1996, 1998, 2000 e 2002).

Para atingir esse objetivo, será feita uma sucinta explanação sobre essas teorias semântico-cognitivas para, posteriormente, apresentarmos a abordagem realizada via análise de corpus.

2. A teoria dos espaços mentais e da mesclagem conceitual

Fauconnier (1994) caracteriza os espaços mentais como domínios cognitivos que são ativados por certas expressões linguísticas e por alguns mecanismos de reconhecimento de elementos em diferentes campos (psicológico, cultural, histórico, ficcional etc.). A dinâmica que envolve os espaços mentais se resume no seguinte: a referência a um determinado elemento "a" situa-o num domínio cognitivo específico, chamado domínio-fonte. Através de um conector, que pode ser uma expressão linguística ou um outro mecanismo construtor de espaço, as características desse elemento "a" são projetadas para um elemento "b" pertencente a outro domínio cognitivo, chamado domínio-alvo. Esquematicamente, temos o seguinte:

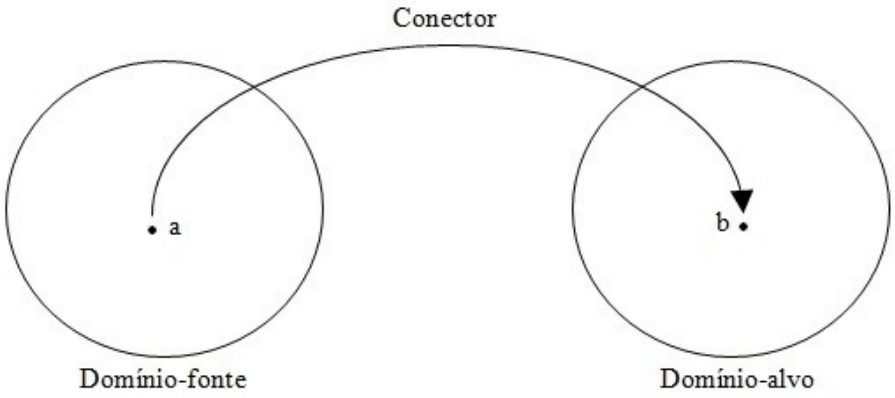


Figura 1 – Esquema de projeção de elementos entre espaços mentais diferentes

O modelo acima é o princípio de uma complexa rede de relações entre domínios cognitivos que se processa na linguagem. Durante uma prática comunicativa qualquer, ativamos vários espaços mentais e inter-relacionamos elementos de vários desses espaços, estabelecendo uma rede de projeções tal que a linguagem se configura como um intrincado emaranhado de elementos, domínios e projeções. Esse modelo nos permite entender que a linguagem humana é um jogo de projeções por excelência. Fazemos analogias o tempo todo, sendo tais o fundamento do nosso raciocínio em várias situações, desde a comunicação corriqueira mais elementar até as construções consideradas mais complexas.

Vejamos uma aplicação desse modelo de Fauconnier à sequência linguística que destacamos no pequeno texto abaixo:

> (1) Dois carregadores estão conversando e um diz: "*Se eu fosse Presidente da República*, eu só acordava lá pelo meio-dia, depois ia almoçar lá pelas três, quatro horas. Só então é que eu ia fazer o primeiro carreto."[4]

Nesse caso, o domínio-fonte engloba as informações referentes ao mundo do carregador (pobreza, necessidade de trabalhar, dificuldades de sobrevivência etc.), enquanto o domínio-alvo abarca os dados relativos à vida do Presidente da República (marcada pelo poder, regalias etc.). Para a compreensão do sentido do trecho, as informações do domínio do carregador são transpostas para o domínio do Presidente da República, e funciona como conector, nesse caso, a expressão introdutora da contrafactualidade, "se eu fosse". Nesse processo, toda a noção relativa aos comportamentos e estilo de vida do carregador é compreendida no âmbito de outro domínio, o do Presidente da República. Esquematicamente:

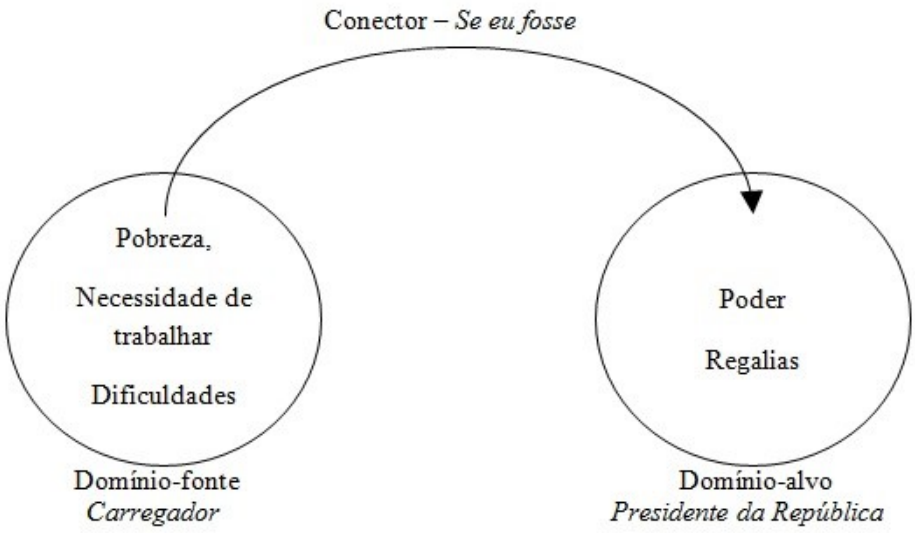


Figura 2 – Esquema de projeção de informações entre domínios diferentes

O modelo descritivo de Fauconnier é capaz de explicar como funciona a mente humana diante de situações em que operamos vários tipos de analogias, mas ele não é suficiente para explicar a seletividade que envolve o processo, ou seja, a imagem de um carregador que possui certas regalias de um Presidente da República, ou a imagem de um

[4] Transcrito de prova de Língua Portuguesa de Vestibular da Unicamp – SP. Grifo nosso.

Presidente da República que precisa fazer carreto. Isso vai concretizar-se mais tarde com a teoria da mesclagem conceitual, como veremos adiante. De toda forma, a teoria dos espaços mentais veio esclarecer como somos capazes de lidar com elementos de diferentes domínios cognitivos, projetando informações de um espaço para outro.

Esse modelo de projeção de informações de um domínio-fonte para um domínio-alvo atende a um princípio mais geral, o Princípio de Identificação, também chamado por Fauconnier (1997, p. 41) de Princípio de Acesso, segundo o qual se afirma o seguinte:

> Se dois objetos *a* e *b* se ligam por uma função pragmática F (*b* = F(*a*)), então uma descrição de *a* (d_a) pode ser usada para identificar sua contraparte *b*.

Por "função pragmática" entende-se o estabelecimento de "ligações entre objetos de natureza diferente por razões psicológicas, culturais ou localizadamente pragmáticas"[5], noção bem desenvolvida por Nunberg (1978). Em outras palavras, existem razões de natureza extralinguística que justificam o fenômeno da projeção, e esse é um ponto crucial para o nosso estudo sobre metáforas. No caso acima, não é por acaso que o carregador estabelece a analogia com o Presidente da República; existe uma série de características sobre esta entidade que motivam o processo de analogia.

O modelo de representação da mesclagem conceitual é uma evolução dos estudos realizados sobre os espaços mentais, tanto que o suporte daquela são os mesmos domínios cognitivos descritos anteriormente. A mesclagem conceitual surge como uma teoria que explica a dinâmica funcional dos espaços, com a vantagem de incluir outros domínios – indo além da simples relação entre domínios fonte e alvo –, o que enriquece sobremaneira a compreensão sobre o processamento do sentido.

Fauconnier e Turner (2002), fazendo uso de exemplos bem práticos, apresentam muitos detalhes sobre o processo de mesclagem, enfatizando especialmente os elementos que compõem essa rede de integração conceitual. E especialmente em Fauconnier e Turner (1996)[6] é apresentada a ideia de que os padrões gramaticais de uma língua refletem, em grande parte, as mesclagens conceituais e o processo de integração de eventos. Daí a noção de

[5] No original: "links between objects of a different nature for psychological, cultural, or locally pragmatic reasons" (FAUCONNIER, 1994, p. 3).

[6] Uma versão expandida desse trabalho se encontra em: <http://markturner.org/centralprocess.WWW/centralprocess.html>. Acesso em: 12 set. 2014.

que o estudo da linguagem verbal é a chave para se alcançar o entendimento dos processos da cognição humana.

Outra inovação no modelo da mesclagem conceitual é a identificação do espaço da mescla como uma estrutura emergente, sinalizado com um quadriculado. Nos estudos cognitivos em geral, a noção desse tipo de estrutura é de fundamental importância para a compreensão de vários fenômenos. Nesse aspecto, vale ressaltar a importância de trabalhos como o de Grady, Oakley e Coulson (1997), que mostram como uma sentença do tipo "aquele cirurgião é um açougueiro" apresenta uma série de significados emergentes, provando que o espaço da mescla não é um espaço de mera composicionalidade semântica.

Com essas modificações, o modelo utilizado para representar o processo de mesclagem é o que se mostra abaixo, no qual figuram o espaço de entrada 1 e o espaço de entrada 2 como domínios que apresentam elementos mapeados entre si, além do espaço genérico e o espaço emergente da mescla:

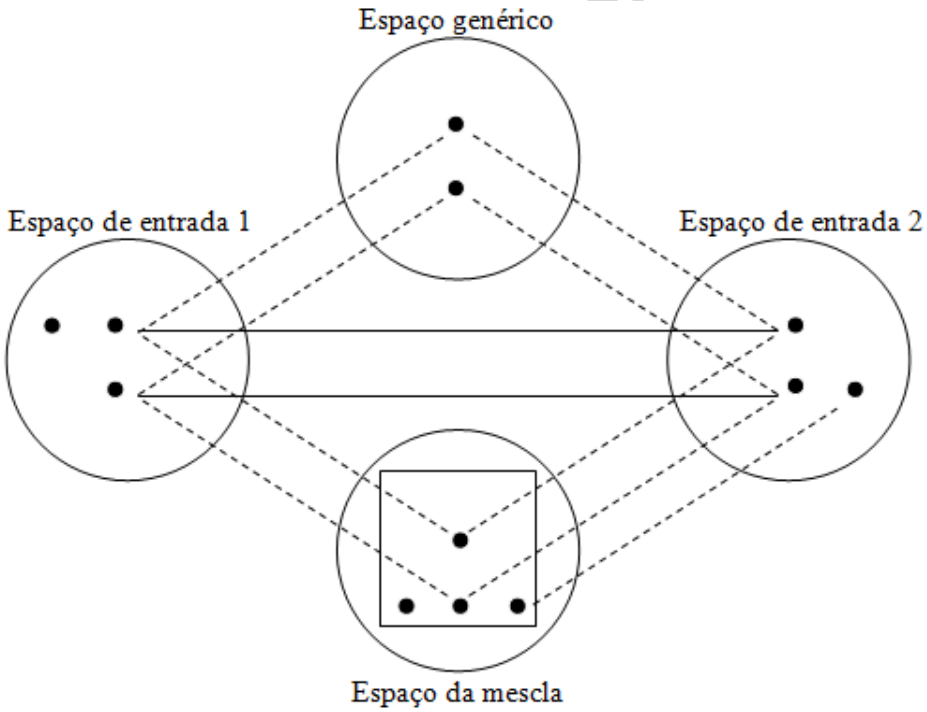


Figura 3 – Modelo de representação do processo de mesclagem conceitual

Observe-se que, pelo esquema apresentado acima, os espaços de entrada podem encerrar elementos que não são projetados para o espaço da mescla, bem como elementos projetados podem não apresentar uma contraparte no outro espaço de entrada. E, ainda, há

informações que emergem no espaço da mescla sem que tenham provindo de qualquer espaço de entrada (sentido emergente).

3. A metáfora no contexto da Linguística Cognitiva

Na segunda metade do século XX, vários estudos são empreendidos com foco no aspecto cognitivo. Essa característica se faz sentir em várias áreas do conhecimento humano, e não ocorre diferente em relação aos estudos da linguagem. Essa preocupação com o aspecto mentalista da linguagem vem desembocar nos estudos cognitivos, e com o advento da Gramática Cognitiva de Langacker esse espaço se consolida, desenvolvendo-se cada vez mais nos últimos decênios.

Em 1979, é amplamente difundida uma noção explicativa sobre o funcionamento da linguagem humana através de um clássico artigo de Michael Reddy. Segundo o autor, as palavras são concebidas como contêineres das ideias, e estas são transmitidas como que passando por um tubo de indivíduo para indivíduo. Dessa forma, as palavras podem ser entendidas como vazias de sentido ou plenas de significado, e o processo de transmissão de ideias pode ser entendido como susceptível a quaisquer vicissitudes típicas da passagem de objetos por um canal. É sobre essa noção que Reddy desenvolve o que ele denomina "metáfora do tubo" (*conduit metaphor*)[7].

Em 1980, foi publicada uma obra que revolucionou o pensamento acerca da metáfora, inclusive alargando a sua concepção e relacionando-a à experiência corporal, cultura, usos e costumes dos indivíduos. Lakoff e Johnson (1980) defendem a ideia de que as metáforas não são recursos especiais de linguagem, como era costume supor, específicos da linguagem literária ou retórica, mas fazem parte da linguagem corriqueira. E, mais do que isso, a metáfora também está presente no pensamento e nas ações humanas, não sendo tão somente um aspecto da linguagem verbal; nosso sistema conceitual é metafórico por natureza. O homem pensa, age e comunica através de metáforas.

Os autores apresentam metáforas fundamentais, a partir das quais muitos elementos comunicativos como expressões linguísticas, gestos e posturas são criados, como "para cima é bom; para baixo é ruim", "argumentar é lutar", "tempo é dinheiro", "ideias são objetos", "palavras são contêineres", "abstrato é concreto", "seres abstratos são entidades físicas",

[7] Cf. Reddy (1979).

"comunicar é enviar" etc. A título de exemplo, a primeira metáfora orientacional desta lista se manifesta através de uma série de expressões linguísticas (*A bolsa de valores fechou em alta*, *Fulano está no fundo do poço*, *Beltrano está em alto astral*, *Ela se encontra deprimida* (= em depressão), *Subir na vida*, *Chegar ao topo da carreira*, *Fazer parte do alto escalão*, *Hoje estou meio* down, *Os planos caíram por terra*), de gestos (polegar apontado para cima para indicar estado bom, polegar apontado para baixo para indicar estado ruim; referência ao céu para indicar o paraíso religioso, referência ao subterrâneo para indicar o inferno) e de posturas (ficar de cabeça erguida é bom, ficar cabisbaixo é ruim). Uma importante ideia defendida por Lakoff e Johnson é que não existe, a rigor, nenhum tipo de necessidade humana para se operarem tais conceitualizações; o que existe, e que justifica a concepção de uma ideia em termos de outra, é o apego à cultura da sociedade em que o indivíduo se encontra, além das suas experiências corporais. A metáfora orientacional que foi explicada acima, por exemplo, pode ser justificada pela própria experiência do ser humano, em seu primeiro ano de vida, ao tentar vencer a força gravitacional e manter-se de pé, em postura ereta.

É importante ressaltar que esses esquemas metafóricos não são propriamente universais semânticos, como poderia supor algum radical dentro dessa teoria. Trata-se, na verdade, de tendências de conceitualização manifestadas pelo ser humano de acordo com fatores ligados à sua vivência, cultura, constituição biológica. Portanto, apresentam um grau de uniformidade bastante considerável na espécie humana.

O advento da teoria sobre a metáfora conceitual impulsionou os estudos desse fenômeno sob a ótica da cognição humana e constituiu um grande impacto provocado sobre uma tradição de muitos séculos que encarava a metáfora como uma relação de simples-troca de expressões – com ressalva, obviamente, para importantes estudos empreendidos por filósofos desde alguns séculos passados que adiantam essa postura que veio consolidar-se ao final do século passado.

4. A metáfora analisada sob o prisma da Linguística de *Corpus*

Nas últimas décadas, têm crescido em larga escala os estudos linguísticos baseados em dados autênticos de linguagem, seja na modalidade oral ou escrita. Os avanços na área de Informática vêm proporcionando ganhos incomensuráveis nesse aspecto, fazendo com que a Linguística de *Corpus* enriqueça-se cada vez mais em termos de consistência técnica, teórica e metodológica, impulsionando o nível das pesquisas em todas as áreas da linguagem.

No Brasil, pesquisas nesse campo têm alcançado muito êxito, principalmente com a criação de programas específicos para análises linguísticas, como os etiquetadores, concordanciadores etc. Juntamente a isso, a montagem e o incremento de extensos bancos de textos disponíveis para análise – os *corpora* – têm proporcionado às nossas pesquisas enormes vantagens.

Berber Sardinha (2004) oferece uma boa visão desse tipo de pesquisa, reunindo os aspectos fundamentais para os estudos baseados em *corpora*, desde o histórico sobre essa área, a descrição de bancos de textos, até os detalhes de ordem técnica para utilização de ferramentas eletrônicas. Nas palavras do próprio autor:

> Há um debate na definição do status da área: a Linguística de Corpus é disciplina ou metodologia? Claramente, a Linguística de Corpus não é uma disciplina tal qual psicolinguística, sociolinguística ou semântica, pois seu objeto de pesquisa não é delimitado como em outras áreas. A Linguística de Corpus não se dedica a um assunto definido (...). Ao contrário, ocupa-se de vários fenômenos comumente enfocados em outras áreas (léxico, sintaxe, textura). É então uma metodologia da qual outras áreas podem se fazer valer? A princípio, sim. (...)
> Se a Linguística de Corpus é metodologia ou não, depende da definição de metodologia que está sendo usada. Entendendo metodologia como *instrumental*, então é possível aplicar o instrumental da Linguística de Corpus livremente e manter a orientação teórica da disciplina original. (BERBER SARDINHA, 2004, p. 35-36)

Uma clara contribuição dessa chamada "metodologia" para a Linguística é o fato de o pesquisador lidar com dados reais da linguagem, e não chegar a conclusões baseadas em exemplos construídos artificialmente, ainda que correspondendo à intuição dos falantes. E mais: com esse procedimento, o número de informações com que o linguista é capaz de lidar é inúmeras vezes maior, alcançando enorme fidedignidade entre as conclusões alcançadas em relação a um corpus e as conclusões que podem ser imputadas à língua como um todo. Enfim, quase todos os estudos quer da linha diacrônica, quer da sincrônica encontram na Linguística de *Corpus* um suporte jamais alcançado na história da pesquisa em linguagem.

Como os demais temas de pesquisa, os estudos sobre a metáfora também voltam os olhares para as técnicas e métodos proporcionados pela Linguística de *Corpus*, especialmente quando se pretende investigar as ocorrências dessa modalidade de linguagem no cotidiano dos usuários da língua. Berber Sardinha (2009, p. 1) destaca, na introdução de um texto ainda não publicado, que

> Tem existido um crescente interesse na utilização de corpora na pesquisa sobre metáfora nos últimos anos, e como resultado disso um certo número de ferramentas e técnicas tem sido proposto e utilizado para identificação de metáforas. No entanto, muito pouco se sabe a respeito de suas habilidades para recuperar todas e somente metáforas a partir dos corpora.[8]

Apesar dessa dificuldade, é inegável a contribuição que modernas tecnologias vêm dando à ciência no âmbito do estudo da metáfora. Questões jamais imaginadas até então passam a ser investigadas, como: qual a relação entre a metáfora e o processamento cognitivo humano? Quais são os limites entre o sentido literal e não literal na linguagem? Em que situações os falantes fazem uso de construções metafóricas em vez das correspondentes construções não metafóricas? Qual o grau de ocorrência de construções metafóricas numa dada língua?

Muita contribuição no sentido de possíveis respostas a esses questionamentos vem sendo dada por dois grandes estudiosos do assunto: Anatol Stefanowitsch e Stefan Gries. Stefanowitsch (2005), por exemplo, realiza um estudo de extrema relevância com vistas a explicar se o uso da linguagem metafórica é motivado por questões estilísticas ou por princípios cognitivos. O autor desenvolve essa questão analisando as ocorrências de algumas expressões metafóricas da língua inglesa, comparando as situações de uso das mesmas em contraposição à situação de uso das respectivas expressões não metafóricas.

Nesse artigo, o autor defende a hipótese cognitiva sobre a metáfora, segundo a qual ela é um elemento sistemático e pervasivo na linguagem cotidiana, um fenômeno conceitual/mental que nos possibilita a compreensão de uma ideia (mais abstrata) em termos de outra ideia (mais concreta), em oposição à hipótese estilística, cujos adeptos defendem que a metáfora é um recurso extraordinário de linguagem, uma figura de linguagem empregada para obter efeitos estéticos, largamente empregada na literatura, retórica e outros registros que utilizam a linguagem como "ornamento" das ideias (STEFANOWITSCH, 2005, p. 163). Como argumentos em favor da hipótese cognitiva, são apresentados os seguintes:

i) se a metáfora fosse um fenômeno estilístico simples, ela não apresentaria tão alto grau de sistematicidade e ocorrência;

[8] No original: "There has been growing interest in using corpora in metaphor research in recent years, and as a result a number of tools and techniques have been proposed and used for metaphor identification. However, very little is known about their ability to retrieve all and only metaphors from corpora."

ii) se a metáfora fosse um recurso ornamental da linguagem, existiria sempre uma expressão literal correspondente a cada expressão metafórica;
iii) nas metáforas, o mapeamento é sempre unidirecional, acontecendo do domínio mais concreto para o mais abstrato, e não vice-versa. Se a metáfora fosse um recurso puramente estilístico, a unidirecionalidade seria acidental, e não sistemática.

A ideia central sobre a linguagem metafórica na hipótese cognitiva é que o seu uso pode reduzir dificuldades de processamento do sentido. Assim, a metáfora pode ser descrita como um elemento que oferece "suporte conceitual" para a nossa apreensão de conceitos complexos. Daí o fato de concebermos os conceitos mais abstratos dentro de um domínio mais concreto.[9]

Estudos desse porte desmistificam a ideia de que o modo básico de utilização da linguagem humana é o uso do sentido literal e que o sentido metafórico é um mero correspondente opcional daquele. Tais estudos vêm demonstrando que a linguagem metafórica – e, por extensão, o raciocínio metafórico – é um elemento essencial da cognição humana. Gibbs Jr. (2002) já expusera em seu artigo que não faz sentido simplesmente contrapor o sentido literal ao sentido não literal, uma vez que não existe uma linha divisória entre essas duas formas de processamento do sentido, além de que não existe uma única forma de sentido literal nem tampouco uma única forma de sentido não literal. No bojo deste, existem, por exemplo, o sentido metafórico, o idiomático, o irônico, o metonímico etc. No processamento de uma sentença não literal, diferentes tipos de sentido são ativados em diferentes pontos da sentença.

Stefanowitsch utiliza, em várias de suas pesquisas, um procedimento bastante comum na Linguística de *Corpus*, que é a análise dos colocados, isto é, as palavras que ocorrem com frequência considerável na vizinhança de alguns nódulos (palavras e expressões) escolhidos para análise. Com esse procedimento, numa extensão da análise colocacional, Stefanowitsch e Gries desenvolveram um método através do qual é investigada a interação de lexemas e as estruturas gramaticais a eles associadas, com aplicação no estudo de expressões linguísticas

[9] Um bom exemplo disso é o fato de conceitualizarmos o tempo (abstrato) em termos de dinheiro (concreto), no emprego de várias expressões verbais: *gastar tempo, ganhar tempo, economizar tempo, perder tempo, ceder tempo, tomar tempo* etc. O contrário não ocorre, ou seja, não conceitualizamos dinheiro em termos de tempo, medindo-o em segundos, minutos, horas etc.

de vários níveis (palavras, expressões fixas, estruturas de argumento etc.). A esse procedimento os autores chamam de análise colostrucional (*collostructional analysis*)[10].

Stefanowitsch e Gries (2003, p. 210) afirmam que "recentemente (...) o foco dentro da linguística de corpus mudou para uma visão mais holística da língua"[11], chamando a atenção para o fato de que gramática e léxico não são elementos fundamentalmente diferentes, da maneira como essa antiga dicotomia tem sido vista nos estudos da linguagem, existindo muitas expressões ignoradas ao longo dos tempos que servem de importantes elos entre esses dois polos. Trata-se de um estudo que toma por base preceitos da chamada Gramática de Construções, aplicando-se de forma muito pertinente ao estudo de *collocations*, *chunks*[12] e outras expressões linguísticas. Não se trata especificamente de uma metodologia para estudo da metáfora, mas como a língua é plena de expressões metafóricas entrincheiradas[13], esse tipo de estudo também nos é de grande valia.

5. Descrição e análise linguística do corpus

5.1. Descrição do banco de textos

Procederemos à análise de textos escritos em língua portuguesa, extraídos do corpus organizado para esse fim, de forma que os resultados alcançados possam mostrar-se aplicáveis a uma ampla variedade de textos dentro da língua.

Para a composição do nosso corpus de análise, optou-se pelo gênero textual redação de vestibular. Trata-se de um tipo de produção textual muito difundido no meio escolar, cujo propósito é autorrecursivo, ou seja, o objetivo principal é treinar ou demonstrar habilidades de comunicação escrita dentro da norma padrão da língua. As redações de vestibular, sejam do estilo tradicional (realizado ao final do Ensino Médio) ou seriado (realizado ao longo dos anos que compõem o Ensino Médio), são produzidas num contexto específico de avaliação de desempenho de escrita e concatenação de ideias em torno de um tema. Elas não atendem a um propósito comunicativo externo à instituição de ensino e correspondem a um tipo de produção

[10] Cf. Stefanowitsch e Gries (2003).
[11] No original: "recently (...) the focus within corpus linguistics has shifted to a more holistic view of language".
[12] Mantivemos aqui os originais em inglês por não existirem adequadas traduções para esses termos em português.
[13] O entrincheiramento (do inglês *entrenchment*) corresponde ao fenômeno normalmente associado à cristalização com que certas palavras e expressões são utilizadas no sistema linguístico, como que se apresentando na forma de blocos imutáveis, diferentemente do uso de palavras livres nos enunciados sujeitas a modificações tanto no que diz respeito à forma quanto ao significado.

induzida, não espontânea. Essas características, no entanto, não invalidam estudos sobre esse tipo de produção. Nas palavras de Bezerra (2008, p. 138),

> Embora defendamos a utilização de situações efetivas de escrita em sala de aula, não estamos eliminando o fato de que o texto, ao chegar aí, perde parte da carga comunicativa que tem, já que se torna objeto de ensino/aprendizagem. Com isso, observamos que o trabalho com a redação (entendida como um texto inerte), com a produção textual (concebida como um texto produzido em uma situação comunicativa) e com o gênero textual (entendido como um enunciado produzido em uma situação comunicativa específica, de acordo com um tema, uma composição e um registro linguístico) tem um ponto comum, que é ser objeto de ensino. Por isso, não se deve desfazer-se dessa característica (...) sob o pretexto de que o importante é respeitar as práticas sociais da escrita e seus usos.

Em outras palavras, ressalvado o fato de que redações escolares – e aí incluímos as redações de vestibular – não atendem a um propósito comunicativo espontâneo, trata-se de um tipo textual muito difundido na prática escolar, capaz de revelar muitos fatos no âmbito do raciocínio com a linguagem.

O nosso corpus foi composto por um total de 500 (quinhentas) redações produzidas entre os anos de 2005 e 2007 em diferentes processos seletivos para ingresso no ensino superior da Universidade Presidente Antônio Carlos, instituição multicampi da rede particular cuja sede se localiza na cidade de Barbacena (MG) e que possui unidades de ensino em várias outras cidades, incluindo uma unidade no estado de Tocantins (Instituto Tocantinense Presidente Antônio Carlos).

A escolha desses textos foi aleatória. Esse procedimento faz parte do método estatístico da pesquisa científica, aplicando-se a seleção de amostragem casual simples, em que todos os conjuntos de textos disponíveis tinham igual probabilidade de serem escolhidos. Com isso, pretendemos detectar construções de uso metafórico no corpus e sobre elas realizar nossa análise qualitativa, de forma que tais construções tenham a probabilidade de serem representativas de todo o montante de textos à nossa disposição.

A proposta de analisar textos escritos autênticos justifica-se pelo objetivo de lidar com elementos da língua em uso real e efetivo (ainda que a produção dos textos seja induzida, conforme comentamos), e não criados para satisfazer a alguma hipótese de pesquisa. Na composição do corpus, mantivemos a escrita original dos textos, a fim de evitar qualquer tipo

de interferência que pudesse prejudicar os nossos resultados, ferindo a autenticidade dos mesmos.[14]

Para se ter uma noção da dimensão do nosso corpus, ele possui um total de 84.450 palavras, conforme se pode levantar através do listador de palavras do WordSmith Tools©, programa muito utilizado como suporte para vários tipos de análises linguísticas. Desse total de ocorrências (*tokens*), são identificadas 8.734 palavras diferentes, ou tipos (*types*). Vejam-se os dados na figura abaixo, em que está destacado o total de palavras ocorrentes no corpus:

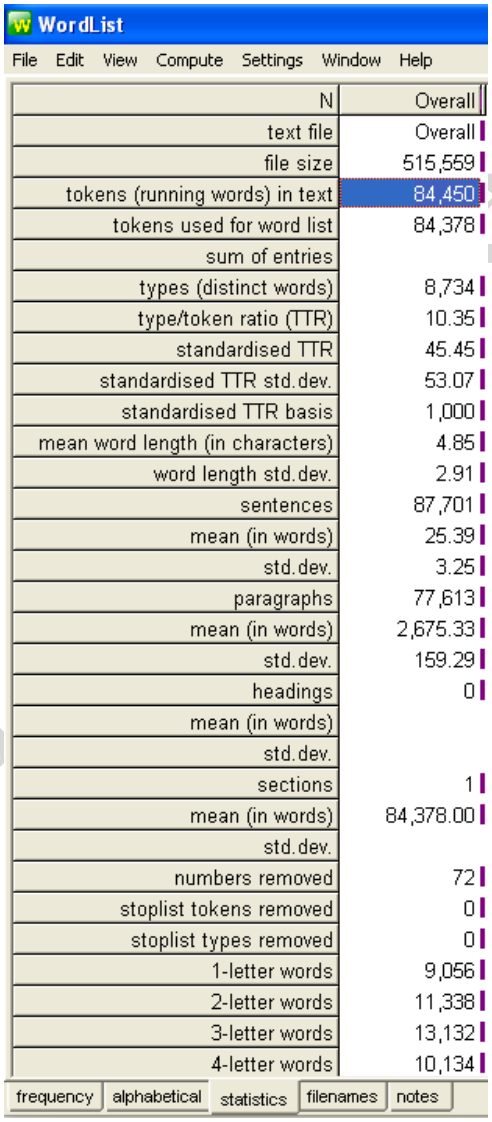


Figura 4 – Descrição geral do corpus pelo listador de palavras do WordSmith Tools©

[14] A escolha dos textos para a composição do corpus foi anterior ao trabalho de correção que a equipe do processo seletivo realiza para a classificação dos candidatos. Foram incluídas, assim, redações de níveis muito diferenciados, de candidatos que tanto foram aprovados quanto reprovados nos concursos. Portanto, a nota obtida pelos candidatos nas redações em nenhum momento influenciou a nossa escolha.

De acordo com Berber Sardinha (2004, p. 26), um corpus dessa natureza é classificado como pequeno-médio, em classificação baseada na observação dos *corpora* normalmente utilizados em pesquisas (um corpus pequeno-médio, segundo o autor, possui de 80.000 a 250.000 palavras).

Na montagem desse banco de textos, eles foram numerados de 001 a 500. A tabela a seguir apresenta uma descrição geral dos subconjuntos de redações que compõem o corpus, compreendendo a quantidade produzida em cada local, as cidades em que as mesmas foram produzidas e o tema que serviu de motivação para a produção de cada subconjunto:

Tabela 1 – Dados gerais dos textos do *corpus*, separados por grupos

REDAÇÕES	QUANTIDADE	LOCAL	TEMA (RESUMIDO)
001 a 181	181	Araguaína (TO)	Crimes virtuais
182 a 229	48	Araguaína (TO)	A pirataria no Brasil
230 a 246	17	Medina (MG)	A felicidade
247 a 256	10	Teófilo Otoni (MG)	Sonhos de simplicidade
257 a 466	210	Araguaína (TO)	A internet
467 a 500	34	Barbacena (MG)	A destruição da natureza

5.2. Procedimentos de tratamento dos textos

Como se trata da composição de um corpus de pesquisa que futuramente pode servir também a outros tipos de investigação, e com o intuito de não incorrermos em falhas metodológicas, seguimos os procedimentos gerais para tratamento do corpus, que normalmente integram esse tipo de abordagem, a saber:

i) Uma vez que as redações são manuscritas, após selecionadas elas foram transcritas ao computador utilizando-se o programa Microsoft Word for Windows©, em espaço simples, fonte Times New Roman tamanho 12, alinhamento de margem à esquerda. Seguimos um procedimento corriqueiro desse tipo de montagem de corpus: deixar um espaço em branco entre os parágrafos, apesar de não interessar diretamente para a nossa pesquisa a identificação de tais. Entre o título da redação – quando existente – e o primeiro parágrafo, deixaram-se dois espaços em branco para a identificação daquele.

ii) Após organizados em pastas no computador, os textos foram salvos também como "texto sem formatação" (com a extensão .txt), procedimento fundamental para que a aplicação de ferramentas eletrônicas como o WordSmith Tools© não seja prejudicada com a identificação de caracteres estranhos ao programa.

iii) A partir daí, levantamos as informações gerais sobre o corpus, a exemplo dos dados da Figura 4, para se ter uma noção geral do ambiente de pesquisa com que estamos lidando. O software utilizado serviu como ponto de partida para a identificação das características gerais do banco de textos e também para realizar buscas de palavras e expressões no corpus à medida que fomos realizando leituras e análises de cunho qualitativo.

As buscas de palavras e expressões no corpus com o apoio de recurso eletrônico são de fundamental importância num trabalho desse porte, uma vez que proporcionam levantamentos que seriam impossíveis de serem feitos somente através da chamada leitura manual. O grau de precisão dessas buscas é altíssimo, além da capacidade de obtenção de dados importantes para a análise em tempo imediato.

5.3. O uso da ferramenta eletrônica para análise

A ferramenta eletrônica escolhida para análise do nosso corpus, o WordSmith Tools©, possui três componentes básicos: o concordanciador (*Concord*), o listador de palavras (*WordList*) e o listador de palavras-chave (*KeyWords*).

O recurso do concordanciador permite que o analista visualize os "colocados", que são os itens lexicais que ocorrem com um nódulo de uma busca. Essa busca pode ser realizada com alcances diferenciados, chamados de "janelas" ou "horizontes", que consiste em quantidades de palavras à esquerda e à direita escolhidas pelo pesquisador. O concordanciador também fornece a "frequência", que é o número de ocorrências tanto do nódulo quanto de seus colocados.

Através do listador de palavras, outro recurso da ferramenta eletrônica, é possível obter informações do corpus analisado em três segmentos diferentes: um relativo às informações gerais do banco de textos (número de palavras, tipos e ocorrências; número de parágrafos; razão entre tipos e ocorrências; extensão média das palavras e dos parágrafos etc.); outro relativo à frequência de ocorrência de cada palavra do corpus, medida em porcentagem em relação às demais palavras do texto, da mais frequente até a menos frequente do corpus; e o último relativo à listagem de todas as palavras do corpus em ordem alfabética, acompanhadas da frequência em que ocorrem.

Por fim, o listador de palavras-chave estabelece uma comparação entre as palavras de um texto ou conjunto de textos selecionados em relação a um corpus que serve como referência. No caso de nossa pesquisa, não utilizamos esse recurso.

5.4. Análise do corpus

O grande desafio para os estudos sobre metáfora baseados em *corpora* é a identificação de nódulos que podem ser considerados metafóricos, já que os mapeamentos entre domínios cognitivos não estão ligados a formas linguísticas específicas. Na tentativa de dar um rumo à nossa análise tendo em vista essa dificuldade, buscamos suporte em Stefanowitsch (2006). O autor apresenta algumas estratégias para contornar esse problema, que se resumem no seguinte[15]:

i) Busca manual – Muitos estudos se baseiam na coleta manual das ocorrências de construções metafóricas, o que limita muito o trabalho do pesquisador, evidentemente, no caso de trabalhos baseados em *corpora* extensos.

ii) Busca por vocabulário de domínio-fonte – Algumas expressões metafóricas baseiam-se em itens lexicais específicos do domínio-fonte. Constitui, portanto, uma estratégia de pesquisa realizar a busca começando por elementos do léxico ou conjuntos de elementos que são potencialmente formadores de metáforas.

iii) Busca por vocabulário de domínio-alvo – Muitos estudos sobre a metáfora são realizados tendo-se em vista domínios-alvo específicos, bem como os mapeamentos conceituais que os estruturam; assim, esse tipo de busca pode ser bastante producente. Existem algumas restrições quanto a esse aspecto, especialmente o fato de que esse método se aplica muito bem quando se trata de um corpus muito representativo de textos que lidem com um domínio-alvo específico, além de funcionar bem, obviamente, quando se trata de construções cujo domínio-fonte apresente uma associação sistemática e previsível com o domínio-alvo em questão.

iv) Busca por sentenças que contenham itens lexicais tanto do domínio-fonte quanto do domínio-alvo – Os dois tipos de busca apresentados anteriormente podem combinar-se no mesmo processo. E, assim como os dois procedimentos anteriores não são completos, este também pode apresentar problemas. Trata-se de um processo que funciona muito bem em se tratando de expressões cujo mapeamento conceitual é conhecido de antemão – ou, para utilizar uma expressão do próprio Stefanowitsch, no caso de "padrões metafóricos" (*metaphorical patterns*).

[15] Cf. Stefanowitsch (2006, p. 2-6). O autor confere especial importância às três primeiras estratégias apresentadas.

v) Busca de metáforas baseada nos "marcadores de metáfora" (*markers of metaphors*) – Existe um certo número de expressões na língua, muitas vezes de natureza metalinguística, que sinalizam explicitamente a presença de metáforas, tais como "metaforicamente falando", "figurativamente falando", "literalmente" etc., bem como o recurso gráfico das aspas.

vi) Extração a partir de um corpus etiquetado por campos/domínios semânticos – A primeira estratégia descrita acima pode ser estendida da seguinte forma: pode-se especificar um domínio-fonte e operar a busca por todos os itens lexicais pertencentes a esse domínio, em vez de trabalhar com conjuntos de lexemas, que ficam sempre incompletos. Stefanowitsch qualifica esse método como bastante promissor.

vii) Extração a partir de um corpus etiquetado por mapeamentos conceituais – Esse tipo de busca seria grandemente valioso para os estudos da metáfora, mas o grande problema em relação a ele é justamente realizar as marcações que discriminem os mapeamentos conceituais.

A etiquetagem é muito produtiva em análises textuais, mas, no caso de pesquisas envolvendo metáforas, ainda se constitui um procedimento muito complexo. A identificação de metáforas realizada por recursos eletrônicos é feita, atualmente, em termos de probabilidade de emprego metafórico de uma determinada expressão com base na comparação com a co-ocorrência desse mesmo nódulo em outros *corpora* pré-analisados[16].

Levando-se em consideração todos os aspectos levantados, e tendo em vista o foco da nossa pesquisa voltado para uma análise qualitativa envolvendo metáforas e organização textual, diante dos recursos colocados à disposição para a pesquisa e que nos levem a um grau de total confiabilidade em relação aos resultados alcançados, estabelecemos os seguintes procedimentos metodológicos para análise dos textos:

i) Busca manual de metáforas mais relevantes em textos escolhidos aleatoriamente nos seis subgrupos de redações apresentados na Tabela 1, de maneira a contemplar uma análise preliminar em textos elaborados sob diferentes propostas de tema. Como os subgrupos de redações variam muito entre si, em relação ao número de textos que os compõem,

[16] São muito raros os programas de identificação de metáforas, sendo o único disponível na Internet à época desta pesquisa o do CEPRIL – Centro de Pesquisa, Recursos e Informação em Linguagem, da PUC-SP (disponível em: http://www.corpuslg.org/tools/), que realiza buscas em língua portuguesa e língua inglesa. Esse identificador funciona como um etiquetador, apresentando, para cada palavra do corpus a que o usuário pode submeter, uma informação correspondente à probabilidade de ela ser metafórica. Essa probabilidade varia de 0,01% (zero vírgula zero um por cento, ou seja, praticamente nenhuma probabilidade) a 100% (cem por cento, isto é, certeza de uso metafórico), que o programa oferece através da indicação "Avg.Prob".

estabelecemos a proporção de escolha de um texto para cada conjunto de no máximo setenta redações dentro de cada tema. Assim, foram selecionados: 3 textos no subgrupo I; 1 texto no subgrupo II; 1 texto no subgrupo III; 1 texto no subgrupo IV; 3 textos no subgrupo V; e 1 texto no subgrupo VI. Totalizam, assim 10 (dez) textos para esse procedimento inicial.

ii) Identificação de possíveis vocabulários de domínio-fonte e domínio-alvo a partir da busca manual nos dez textos mencionados acima. Embora esse procedimento não garanta o alcance de grande número de ocorrências metafóricas no corpus como um todo, pode constituir-se um ponto de partida para buscas mais minuciosas em etapas posteriores.

iii) Levantamento de construções metafóricas em outros textos do corpus, além dos dez textos iniciais, num processo que mescla a busca manual e a busca realizada através da ferramenta eletrônica com base nos possíveis vocabulários de domínio-fonte e domínio-alvo mencionados acima.

iv) Análise qualitativa de variados textos do corpus, de acordo com a relevância dos levantamentos feitos até então, com vistas ao comportamento da metáfora dentro desses textos.

5.4.1. Conclusões preliminares da busca manual

Procedendo-se ao levantamento inicial dos padrões metafóricos nas redações selecionadas em cada subgrupo do nosso corpus de análise, observa-se a recorrência de alguns esquemas metafóricos já salientados por vários estudiosos da Linguística Cognitiva como padrões de processamento mental existentes na espécie humana. O reconhecimento desses padrões se dá não só pelo grande número de ocorrências nos textos analisados, como também pela sistematicidade com que ocorrem. Isso nos leva a acreditar na existência de um modo de processamento de sentidos, no âmbito da cognição humana, que segue uma certa tendência na conceitualização de ideias.

A tendência identificada na nossa amostragem resume-se no seguinte:

i) elementos de natureza diversa, quer concretos, quer abstratos, são concebidos como lugares, espaços onde acontece alguma coisa;

ii) elementos abstratos – que, portanto, dizem respeito a ações e sentimentos – são concebidos como elementos concretos, que possuem uma corporeidade física;

iii) elementos inanimados, quer concretos, quer abstratos, são concebidos em termos de seres animados, que possuem vida própria, que praticam ações.

Outros esquemas metafóricos foram encontrados no corpus, mas esses três têm uma presença tão marcante nos textos a ponto de todo o conteúdo girar em torno deles. Não se trata de ocorrências isoladas; muitas vezes, são até interpenetrantes, ou seja, seres inanimados podem ser metaforizados como seres animados que ao mesmo tempo praticam ações em espaços também metafóricos.

A essas alturas, cremos já ter ficado bastante claro que a metáfora não é mero recurso estilístico, sendo um elemento intrínseco do modo de raciocinar humano. Dando continuidade a esse ponto de vista, através do levantamento que ora fizemos no nosso corpus, mais do que corroborar esse pensamento, fica visível que a metáfora participa sistematicamente da organização do texto como um todo, estabelecendo parâmetros de inserção de conteúdos no mesmo.

Porém, a metáfora não atua sozinha nessa função. Na verdade, informações de ordem metafórica e não metafórica se juntam no decorrer do texto e atuam concomitantemente na apresentação dos conteúdos. Ações são praticadas em espaços tanto metafóricos quanto não metafóricos, os elementos que possuem corporeidade física são apresentados no texto também de forma metafórica ou não metafórica, e assim por diante. Não se quer exaltar a importância do processo de metaforização em detrimento de outros recursos de organização textual, assaz importantes para a manutenção da coesão e da coerência textuais, mas fica claro que, sem as metáforas, esse quadro não seria instaurado – ou seria de forma incompleta, só no âmbito não metafórico.

Uma importante conclusão que vai nortear todo o rumo desta pesquisa é a seguinte: manifesta-se nítida a ideia de que a metáfora se situa num domínio cognitivo da organização textual capaz de nos fazer vislumbrar esses textos escritos à maneira de narrativas típicas, já que afloram, na tessitura do texto, os seguintes elementos:

i) espaço ou lugar, apresentado de maneira geral no texto, onde ocorrem todas as ações descritas, ou de forma localizada, existindo pequenos espaços para um grupo circunscrito de ações;

ii) personagens, elementos metaforizados ou não, que atuam ao longo do texto. A existência deles é vital para a compreensão dos textos como narrativas;

iii) tempo, informação nem sempre explícita nas redações, frequentemente não metafórica. Muitas vezes ele se manifesta na sequenciação das ações, sendo um importante recurso da coerência textual;

iv) ações, apresentadas explicitamente nos textos em relação a personagens metafóricos ou não, através de formas verbais que também podem ser metafóricas.

Dentro da tipologia textual clássica, raríssimos textos ou passagens de textos do nosso banco de redações poderiam ser categorizados como essencialmente narrativos, que normalmente são marcados com a existência dos elementos listados acima na própria superfície textual. Em outras palavras, para que um texto seja considerado narrativo, é necessário que apresente explicitamente os elementos espaço, personagem, tempo e ação, ou pelo menos a maioria deles, na nossa tradição de gêneros textuais. No caso do objetivo da produção de redações em processos seletivos, essa estrutura é até desaconselhável, uma vez que o comando para a elaboração dos textos direciona para a produção de ideias ou apresentação do ponto de vista dos candidatos sobre um determinado tema, ou seja, recai-se no esquema tradicional das dissertações, e não das narrações.

Na tentativa de esclarecer melhor essas questões e aprofundar um pouco mais o estudo da relação entre narração e metáfora, propomos, com base nos postulados da Gramática Cognitiva, a existência de um domínio no qual todas as informações narrativas são processadas, isto é, no qual convergem elementos como espaço, personagem, tempo e ação, cujo reflexo se manifesta no texto escrito. A esse domínio, que corresponde ao espaço da mescla no modelo de Fauconnier e Turner, estamos denominando Domínio Cognitivo da Narrativa, o qual descreveremos com mais detalhes a seguir.

5.4.2. O modelo dos Domínios Cognitivos da Narrativa

O Domínio Cognitivo da Narrativa (doravante DCN) é, pois, um espaço cognitivo no qual vislumbramos a conjugação de elementos metafóricos e não metafóricos na realização das ações e na apresentação do conteúdo narrativo. Os limites do DCN coincidem basicamente com os limites do texto, entendido este em sua acepção mais ampla, além da mera sequência de elementos da superfície (palavras, frases, parágrafos etc.), atingindo os fatores cognitivos envolvidos na sua organização. Esse domínio engloba, claro, informações de ordem pragmática, cultural, contextual etc., envolvidas no processo de composição textual.

O espaço da mescla é insuficiente para abarcar todas essas informações. Ele explica muito bem a ocorrência de metáforas e outros fenômenos, mas muitas informações contidas num texto estão fora da mescla, incluindo informações não metafóricas. O espaço da mescla é o espaço da compressão, e o texto não é só compressão. Em virtude disso, propomos situar o

DCN de forma a englobar o espaço-mescla e, ao mesmo tempo, abrigar as informações não metafóricas do texto e quaisquer outras informações que sejam pertinentes para a compreensão do mesmo em termos narrativos.

Assim, chegamos ao seguinte modelo de apresentação do DCN em relação ao modelo da mesclagem conceitual:

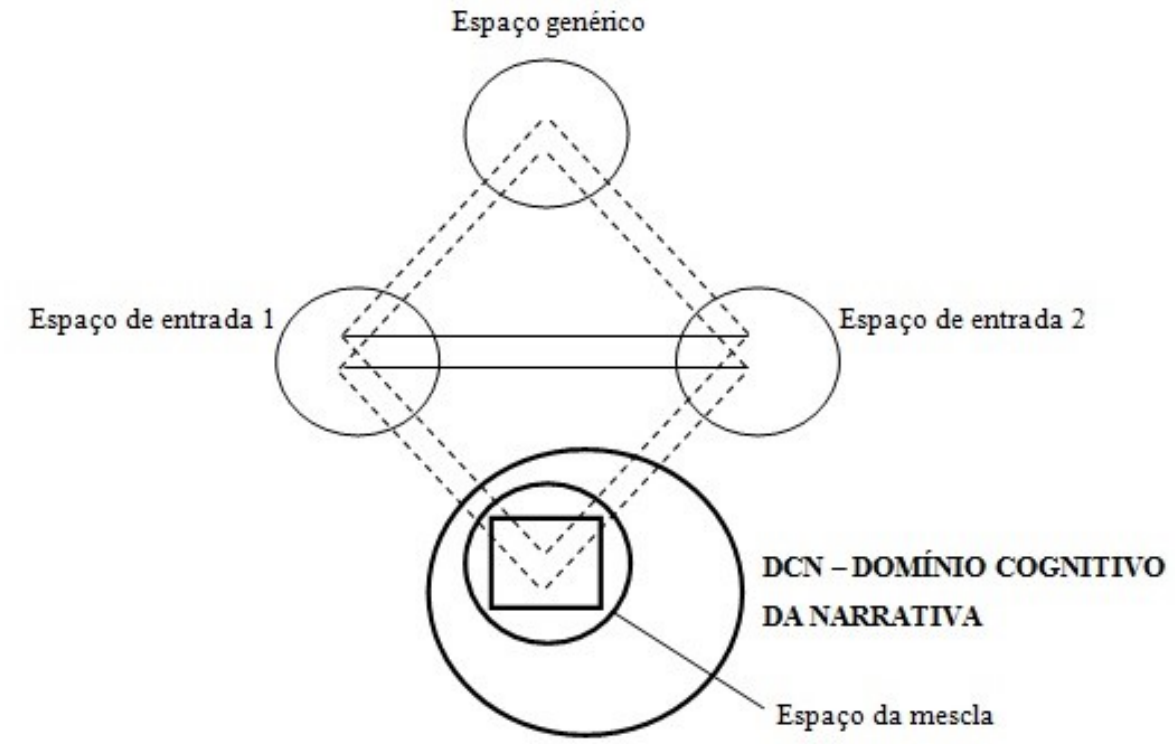


Figura 5 – Estrutura do DCN em relação ao modelo da mesclagem conceitual

Vejamos um exemplo de representação de um texto do nosso corpus de pesquisa nos parâmetros de representação do modelo aqui proposto, a saber, a redação de nº 187, que segue transcrita abaixo:

> (2) Pirataria, uma necessidade.
>
> Como se sobressair dos problemas que atingem hoje a maioria da sociedade brasileira? A miséria está presente em muitas famílias no nosso país, e isso faz com que elas procurem vários tipos de emprego, instantâneos, para que possam ir sobrevivendo. Devido à grande carência do povo brasileiro, a pirataria tornou-se um negócio tão lucrativo, e o único para alguns.
> É notável a situação em que vive a maioria do povo brasileiro, salários baixos, endividados devido ao grande consumo, e sem mão de obra qualificada. Tudo isso deixa a própria sociedade numa "saia justa", e para não entrarem no mundo da

> criminalidade, as pessoas vêem como solução trabalhar com produtos pirateados. A pirataria é ilegal sim, porém, para alguns é a única base de sobrevivência.
> No Brasil a pirataria tornou-se muito comum, e por ser um comércio que tanto cresceu beneficiando os mais pobres, e por ser também de baixo custo, causou abalo no mercado dos produtos originais, que por possuírem um alto custo para compra não são viáveis à comunidade mais pobre.
> É certo que em todo o mundo a pirataria é ilegal, um crime. Porém, dentro dos padrões de criminalidade do Brasil, esse é um crime suave, e por uma boa causa. Comparado com outros crimes atuais, a pirataria deve ser classificada como um bem, pois está dando comida e dignidade às famílias brasileiras que necessitam de apoio.

No texto acima, podemos apontar os seguintes elementos componentes da narrativa:

- Tempo (não metafórico): hoje.

- Personagens não metafóricos: a maioria da sociedade brasileira; o povo brasileiro; famílias brasileiras.

- Personagens metafóricos: miséria (ela está presente no espaço metafórico "muitas famílias"); famílias (elas procuram vários tipos de emprego para sua sobrevivência); a sociedade (fica numa "saia justa"); pirataria (está dando comida e dignidade às famílias brasileiras).

- Espaços não metafóricos: no nosso país; Brasil; todo o mundo.

- Espaços metafóricos: muitas famílias (local onde se encontra a "miséria"); mundo da criminalidade (onde as pessoas tentam não entrar); mercado dos produtos originais (sofreu abalo causado pela pirataria); saia justa (onde fica a sociedade).

Na redação de número 187 transcrita acima, da mesma maneira como acontece em outros textos, o tempo é marcado não metaforicamente, através do introdutor de espaço "hoje". Em relação a esse tempo, personagens e espaços são apresentados, no âmbito da metáfora e da não metáfora, conforme a listagem apresentada acima.

Em relação aos espaços metafóricos estabelecidos no texto, percebe-se que eles se ligam exclusivamente a alguns personagens, não funcionando como locais de ação de vários deles. Por isso, uma representação mais detalhada no modelo do DCN é capaz de representar melhor essa situação.

Os personagens e os espaços não metafóricos aparecem, em alguma proporção, repetidas vezes, através de expressões linguísticas bem similares, podendo ser resumidos no seguinte:

- Personagens não metafóricos: brasileiros.

- Espaços não metafóricos: Brasil; mundo.

Já em relação aos personagens e espaços metafóricos, nota-se uma variedade muito maior, não sendo possível resumi-los, como fizemos com os não metafóricos.

Diante desse quadro, podemos traçar a seguinte representação do DCN do texto transcrito em (2):

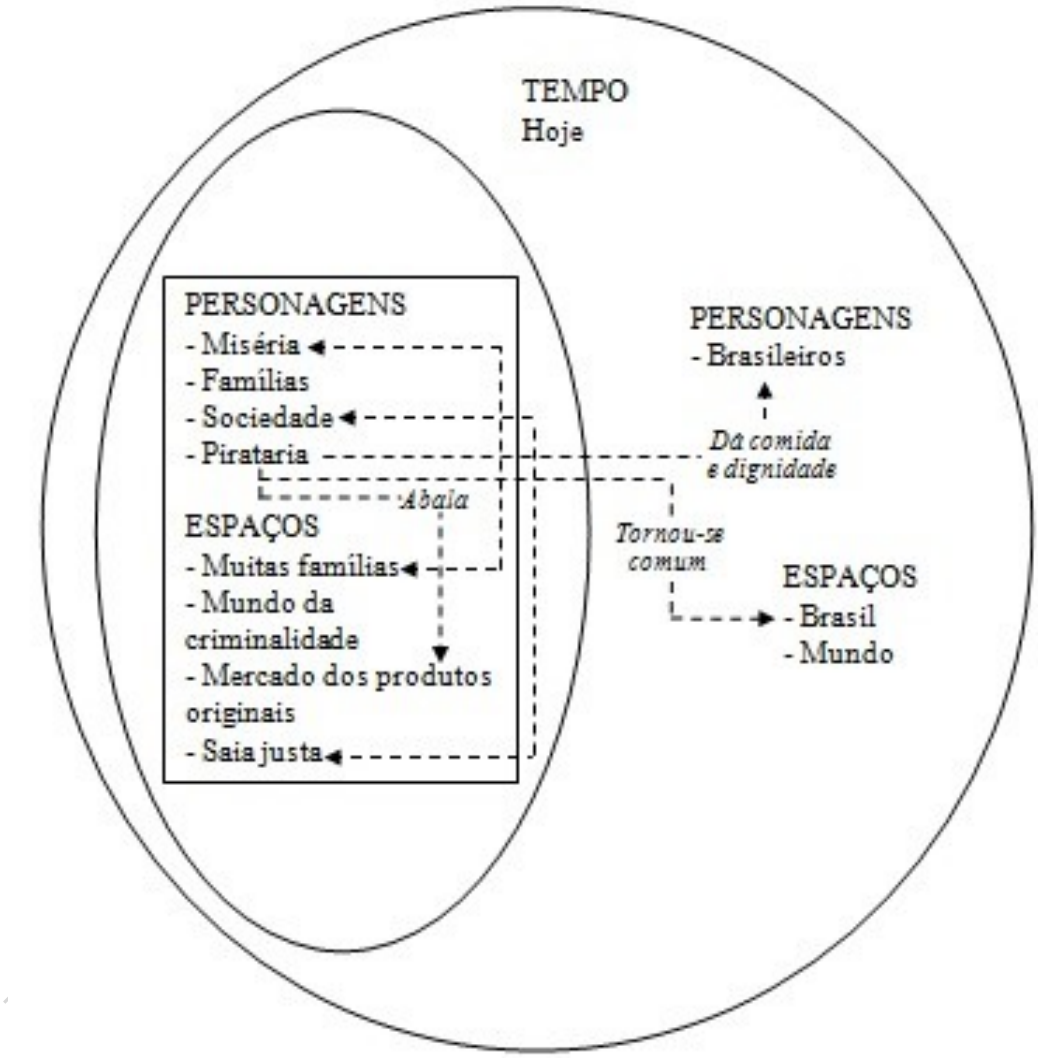


Figura 6 – Representação detalhada do DCN da redação nº 187

Nem todas as relações de sentido estão representadas na figura acima; portanto, a narrativa não se esgota com as relações que foram apresentadas. Essa é uma maneira de exemplificarmos como se dão tais relações, envolvendo diferentes elementos metafóricos e não metafóricos, para a constituição narrativa do texto no âmbito de sua representação semântico-cognitiva.

6. Considerações finais

Considerar que, no processo de produção de textos, a mente humana apresenta o funcionamento próprio da estruturação narrativa é, em outras palavras, apresentar a narração como o principal procedimento linguístico-textual, o princípio organizador das ideias, mesmo que, na estrutura superficial, o texto resulte numa não narrativa de acordo com a clássica tipologia textual.

O que se mostra como novidade no contexto da nossa pesquisa é a forma como esse processo se desenvolve, envolvendo metáfora (que se situa no espaço da mesclagem conceitual) e não metáfora.

A partir dessa constatação, o esquema dos DCNs pode ser incorporado pela Linguística Textual, que é a área por excelência que deu impulso às descobertas dos aspectos de coesão e coerência textuais aplicados a textos de natureza diversa; pode ser aproveitado também para os diversos tipos de estudo realizados no bojo da Semântica, área que trata dos processos de produção do sentido e sua correlação com aspectos que vão além do texto, interagindo com a Pragmática; o modelo se constitui também como um bom subsídio para os estudos empreendidos nas diversas vertentes da Análise do Discurso, uma vez que as informações sobre tempo, espaço e personagens são claramente relacionadas a fatores de ordem pragmática, histórica e linguística a que essa área comumente recorre; e, apesar de apresentar embasamentos teóricos e formas de abordagem diferentes das teorias linguísticas, o modelo também pode ser utilizado em estudos literários, uma vez que seja feita a necessária equalização dos conceitos de metáfora e narrativa. Mais do que uma contribuição teórica para a Literatura, pensa-se na possibilidade de aproveitamento do modelo de análise para esclarecer aspectos que não são exclusivos da teoria linguística.

Sendo mais específico em nossa abordagem, o trabalho apresenta também ampla abertura para a aplicabilidade de ferramentas eletrônicas da Linguística de *Corpus*, não só empreendendo pesquisas em direção ao grau de eficácia das mesmas quando o tema é a metáfora, mas contribuindo para mostrar também o nível de dificuldade e o alcance de procedimentos quando do seu uso efetivo. Na tese que deu origem a este artigo, não desenvolvemos um aparato dentro dessa vertente, mas lidamos com elementos de um corpus organizado, e cada trabalho que é feito com a utilização de *corpora* específicos constitui um ganho tanto no âmbito da análise linguística em si, quanto em relação à avaliação dos procedimentos técnicos capazes de serem empreendidos nessa análise.

Ademais, outras áreas do conhecimento humano podem ser beneficiadas de alguma maneira com a adoção do modelo proposto, desde que estejam interessadas na descrição das representações mentais envolvidas na articulação textual. Referimo-nos aqui superficialmente a algumas áreas mais ligadas à cognição humana, tais como a Psicologia, a Psicanálise, algumas vertentes da Pedagogia, a Ciência da Informação etc. Enfim, são muitas as possibilidades que se abrem a partir da adoção desse modelo, dentro e fora dos estudos linguísticos – na mesma proporção em que cada descoberta científica numa determinada área acarreta, no mínimo, muitas responsabilidades de investigação na própria área e nas suas correlatas. Não vamos nos enveredar aqui nessas possibilidades de aplicação do modelo em outras áreas, pois isso requereria conhecimentos específicos dentro das mesmas, mas lembramos que os termos "cognição" e, por extensão, "domínio cognitivo" aplicam-se muito bem a praticamente todo tipo de estudo que envolve processamento de sentido, raciocínio lógico, processos mentais, redes neurais, estados psicológicos etc. E, conforme mostramos, sendo a narração um processo inerente à espécie humana, a adoção de um modelo que considere a existência de um domínio cognitivo em que se processa a narração certamente é capaz de trazer muitos benefícios em termos de uma melhor compreensão de como funciona a mente humana.

Referências

BERBER SARDINHA, T. **An assessment of metaphor retrieval methods**. 2009. 25 p. Draft.

_____. **Linguística de corpus**. Barueri: Manole, 2004.

BEZERRA, M. A. Da redação ao gênero textual: a didatização da escrita na sala de aula. In: MOURA, D. (Org.). **Os desafios da língua**: pesquisas em língua falada e escrita. Maceió: EDUFAL, 2008. p. 135-138.

FAUCONNIER, G. **Espaces mentaux**: aspects de la construction du sens dans les langues naturelles. Paris: Minuit, 1984.

_____. **Mappings in thought and language**. Cambridge: Cambridge University Press, 1997. crossref http://dx.doi.org/10.1017/CBO9781139174220

_____. **Mental spaces**: aspects of meaning construction in natural language. Cambridge: Cambridge University Press, 1994. crossref http://dx.doi.org/10.1017/CBO9780511624582

FAUCONNIER, G.; TURNER, M. Blending as a central process of grammar. In: GOLDBERG, A. (Ed.). **Conceptual structure, discourse, and language**. Stanford: Center for the Study of Language and Information (CSLI) / Cambridge University Press, 1996. p. 113-129.

_____. Conceptual integration networks. In: **Cognitive science**, vol. 22 (2), p. 133-187, 1998. crossref http://dx.doi.org/10.1207/s15516709cog2202_1

_____. **Conceptual projection and middle spaces** (1994). Report 9401. University of California, San Diego. Disponível em <http://www.cogsci.ucsd.edu/research/files/technical/9401.pdf>. Acesso em: 08 fev. 2008.

_____. Compression and global insight. In: **Cognitive Linguistics** 11 – 3/4, p. 283-304, 2000.

_____. **The way we think**: conceptual blending and the mind's hidden complexities. New York: Basic Books, 2002.

GIBBS JR., R. W. A new look at literal meaning in understanding what is said and implicated. **Journal of pragmatics**, 34, p. 457-486, 2002. crossref http://dx.doi.org/10.1016/S0378-2166(01)00046-7

GRADY, J. E.; OAKLEY, T.; COULSON, S. Blending and metaphor. In: GIBBS JR., R. W.; STEEN, G. J. (Eds.). **Metaphor in cognitive linguistics**; selected papers from the Fifth International Cognitive Linguistics Conference. Amsterdam / Philadelphia: John Benjamins Publishing Company, 1997. p. 101-124.

LAKOFF, G.; JOHNSON, M. **Metaphors we live by**. Chicago/London: The University of Chicago Press, 1980.

LANGACKER, R. W. **Foundations of cognitive grammar**, volume I – Theoretical Prerequisites. Stanford: Stanford University Press, 1987.

_____. **Foundations of cognitive grammar**, volume II – Descriptive Application. Stanford: Stanford University Press, 1991.

NUNBERG, G. **The pragmatics of reference**. Bloomington: Indiana University Linguistics Club, 1978.

REDDY, M. J. The conduit metaphor: a case of frame conflict in our language about language. In: ORTONY, A. (Org.). **Metaphor and thought**. Cambridge: Cambridge University Press, 1979. p. 284-324.

STEFANOWITSCH, A. Corpus-based approaches to metaphor and metonymy. In: STEFANOWITSCH, A.; GRIES, S. T. (Eds.). **Corpus-based approaches to metaphor and metonymy**. Berlin/New York: Mouton de Gruyter, 2006. p. 1-16 (Trends in Linguistics 171). crossref http://dx.doi.org/10.1515/9783110199895

_____. The function of metaphor. **International Journal of Corpus Linguistics**, 10:2, p. 161-198, 2005. crossref http://dx.doi.org/10.1075/ijcl.10.2.03ste

_____; GRIES, S. T. Collostructions: investigating the interaction of words and constructions. **International Journal of Corpus Linguistics**, 8:2, p. 209-243, 2003. crossref http://dx.doi.org/10.1075/ijcl.8.2.03ste

TURNER, M. **The literary mind**: the origins of thought and language. New York / Oxford: Oxford University Press, 1996.

TURNER, M.; FAUCONNIER, G. Conceptual integration and formal expression. In: JOHNSON, M. (Ed.). **Journal of metaphor and symbolic activity**, vol. 10, n. 3, p. 183-203, 1995.

Artigo recebido em: 14.10.2014
Artigo aprovado em: 21.11.2014

A terminologia do futebol: um estudo direcionado pelo *corpus*

Football terminology: a *corpus*-driven study

Sabrina Matuda[*]
Stella E. O. Tagnin[**]

RESUMO: Este artigo tem como objetivo estudar a terminologia do futebol em inglês e português. Para tanto, a fundamentação teórica embasa-se na Linguística de *Corpus*, na Terminologia Textual, na tradução técnica como condicionante cultural e no conceito forma-representação. O *corpus* de estudo possui aproximadamente um milhão de palavras em cada língua. Para etiquetar o *corpus,* utilizamos o etiquetador *Tree Tagger,* desenvolvido por Helmut Schmid, e, para explorá-lo, o *WordSmith Tools,* de Mike Scott.

PALAVRAS-CHAVE: Tradução. Linguistíca de *Corpus*. Terminologia. Futebol e cultura.

ABSTRACT: This article aims at investigating football terminology in English and Portuguese. To that aim, the study is based on the notions of *Corpus* Linguistics and Textual Terminology. To explain cultural differences, technical translation is viewed as a communicative act subject to cultural restraints and the concept of 'form-representation' is called upon to elucidate such differences. Our corpus consists of approximately two million words. To tag the corpus we used Helmut Schmid's Tree-Tagger and to explore the corpus we used Mike Scott's WordSmith Tools.

KEYWORDS: Translation. Corpus Linguistics. Terminology. Football and Culture.

1. Por que estudar a terminologia do futebol

O futebol é o esporte mais praticado no Brasil e no mundo. É reconhecido mundialmente como competição, manifestação cultural e até mesmo como um mercado na ordem econômica. A FIFA (Federação Internacional de Futebol Associado) congrega mais de 140 milhões de jogadores de 300 mil clubes, em 207 países afiliados.

O Brasil, conhecido mundialmente como o "País do Futebol", tem aproximadamente 40 milhões de praticantes, entre atletas profissionais e amadores, cerca de 11 mil jogadores federados, 800 clubes, mais de dois mil atletas atuando em outros países e 580 estádios. Além de, pelo menos, 20 mil "campinhos" de pelada, nos bairros de classe mais baixa , escolinhas de futebol e milhares de torcedores.

[*] Doutoranda da Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH) da Universidade de São Paulo (USP).
[**] Livre Docente pela Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH) da Universidade de São Paulo (USP).

O esporte é praticado em todo o país por camadas sociais distintas e em diferentes espaços: campos de várzea, quadras, praias, ruas, escolas, clubes etc. Enfim, é parte do cotidiano de muitos brasileiros. DaMatta (1982) destaca que o futebol praticado, vivido e discutido no Brasil é um dos modos pelo qual a sociedade brasileira fala, se apresenta e se deixa descobrir. Nessa mesma linha, é também notório o reconhecimento do futebol, no Brasil, como objeto das ciências humanas ao longo dos últimos anos. Sobretudo a partir da década de 1990, ampliaram-se as pesquisas acadêmicas e publicações editoriais acerca do futebol (GIGLIO; SPAGGIARI, 2010).

Essa popularidade, tanto no Brasil quanto em outros países, tem aumentado a participação da indústria do futebol na economia mundial, movimentando cerca de 250 bilhões de dólares anuais (LEONCINI; SILVA, 2005).

É fato que as relações futebolísticas entre o Brasil e os países da Europa crescem cada vez mais (CRUZ, 2005), seja pelo intercâmbio de contratos de jogadores e técnicos, pelos direitos de transmissão de campeonatos, pelo patrocínio de jogadores por grandes marcas ou por qualquer outra negociação que envolva um produto relacionado ao futebol.

Para que todas essas relações se materializem, estabelece-se uma comunicação que, na grande maioria das vezes, se dá em língua inglesa. No entanto, cada nação tem a sua maneira de jogar, torcer e narrar, maneira, essa, expressa por meio de sua língua materna. O problema surge quando se quer expressar essas particularidades em uma língua estrangeira.

Embora seja, muitas vezes, relacionado somente ao lazer, o futebol não deixa de ser uma área de especialidade tanto no Brasil quanto em outros países. Sendo assim, possui uma linguagem própria utilizada para descrever o universo a ele relacionado. Essa linguagem é padronizada e, justamente por esse motivo, não pode ser utilizada de qualquer forma. Ao falar em padronização, não pretendemos, de forma alguma, tentar normatizar o léxico do futebol para favorecer a eficácia das comunicações especializadas em torno dessa temática. Ao contrário, pretendemos favorecer as peculiaridades de cada texto dentro de seu discurso (KRIEGER; FINATTO, 2004) levando em conta os aspectos históricos e socioculturais presentes em cada cultura.

A fim de melhor entender como essas particularidades históricas e socioculturais se manifestam na língua, utilizamos o conceito forma-representação proposto por Toledo (2002). Segundo o antropólogo, o conjunto de regras que define a atividade como esporte não delimita as maneiras de jogar. Na verdade, é a apropriação e a interpretação cultural que cada

região faz das regras que determinam as "formas de jogo". Justapostas, as regras e as "formas de jogo" dão origem às "representações", ou seja, o ajustamento da observação empírica das "formas de jogo" em um plano simbólico que, por sua vez, consolida as conhecidas "escolas", "jeitos" e "estilos" de se jogar.

2. Objetivos

O principal objetivo deste artigo, recorte de nossa pesquisa de doutorado que está sendo desenvolvida no programa de Estudos da Tradução da Universidade de São Paulo, é demonstrar como a extração terminológica e a identificação de equivalentes tradutórios em *corpora* comparáveis contribui para uma investigação terminológica que transpõe a esfera linguística e que considera fatores culturais e históricos no processo de entendimento do léxico especializado. Para tanto, estabelecemos dois princípios, a saber:

1) a pesquisa foi feita do português para o inglês;
2) como se trata de um estudo exploratório, ilustramos os passos seguidos para a extração terminológica e a identificação de equivalentes tradutórios por meio de um estudo de caso com 'gol', segundo termo mais frequente no *corpus*, e a unidade fraseológica[1] mais frequente que comporta o termo: 'fazer um/o gol'.

3. Tradução técnica como condicionante cultural

Entendemos tradução técnica como um ramo da tradução que se ocupa da tradução de textos de línguas de especialidade. Os tradutores, assim como os intérpretes, os redatores técnicos, os jornalistas e os documentalistas são usuários indiretos das terminologias, pois a eles interessa o uso adequado dos termos, das fraseologias e de expressões idiomáticas para que o texto esteja de acordo com as normas de convencionalidade que regem a produção do tipo e do gênero textual em questão na cultura em que é produzido.

Muito comum é a ideia de que a tradução técnica, diferente da literária, constitui um universo à parte, pois, para alguns, sua terminologia não se deixa contaminar por relações contextuais e pragmáticas, possuindo certa estabilidade que favorece e facilita o processo tradutório. Com base nesses preceitos, a tradução técnica é, muitas vezes, definida como uma

[1] Em nosso trabalho, consideramos unidades fraseológicas associações de, no mínimo, duas palavras, sendo uma delas obrigatoriamente uma palavra-chave no corpus e que apresentem frequência maior ou igual a três.

operação de transcodificação em que os conceitos "constituem uma [sic] amálgama indissolúvel e imune aos efeitos do tempo e do espaço, a fim de poderem resistir a uma série de condicionantes a que estão expostos." (AZENHA, 1999, p. 10).

Adotamos a proposta de Azenha (1999, p. 60), a de uma tradução técnica que vai além dos limites do texto, uma tradução que define o ato tradutório tendo como ponto de partida o ato de comunicação. Em virtude disso, o texto técnico, assim como qualquer outra forma de comunicação, está atrelado a uma realidade sócio-histórico-cultural.

Para o autor, não podemos deixar de lado as relações que o texto técnico trava com o contexto em que é produzido. Ao contrário, os termos, as fraseologias, as definições, os equivalentes e as expressões devem ser empregados de maneira convencional[2] no texto de chegada, respeitando as variáveis ligadas ao emissor, receptor, situação e objetivo de comunicação.

Neste estudo, consideramos "equivalente" uma unidade fraseológica que funcione no texto de chegada como funciona no texto de partida (TAGNIN, 2007, p. 1). O conceito, bastante amplo, nos permite identificar equivalentes não só no nível da palavra, do texto ou da frase, mas também equivalentes pragmáticos, ou seja, equivalentes que, embora não reflitam uma tradução direta para a língua de chegada, são utilizados no mesmo contexto e com o mesmo objetivo comunicativo.

Ao considerarmos que os termos são empregados em dado cenário histórico-cultural e, portanto, condicionados por normas sociais e linguísticas sempre sujeitas a alterações (AZENHA, 1999, p. 22), não vemos outra possibilidade senão o uso de *corpora* comparáveis para a identificação dos equivalentes tradutórios.

4. Terminologia Textual

Neste estudo, adotamos a Terminologia Textual como aporte teórico para nossa pesquisa.

Seria inviável falar de Terminologia Textual sem discutir, ainda que brevemente, seu objeto de estudo: o texto especializado que, segundo Hoffmann (1998 *apud* KRIEGER; FINATTO, 2004, p. 113) , possui dois eixos básicos: um horizontal, relacionado ao critério temático, englobando diferentes disciplinas e suas eventuais subdivisões, e um vertical,

[2] Adotamos a definição de convencionalidade proposta por Tagnin (2005): "forma peculiar de expressão de uma dada língua ou comunidade lingüística" p.14.

relacionado ao grau de especialização do texto, ou seja, à densidade terminológica. A classificação de Hoffmann ressalta o fator pragmático no âmbito das comunicações especializadas e enfatiza o papel que as unidades lexicais assumem em diferentes contextos especializados de comunicação.

Alguns tipos de texto possuem maior densidade terminológica do que outros: em nosso caso, as notícias de resultados de partidas possuem densidade terminológica bem menor do que os relatos de partida minuto a minuto. Por outro lado, seria ingenuidade de nossa parte analisar a densidade terminológica de nosso *corpus* somente pela divisão de gêneros textuais. Por esse motivo, optamos por adotar um caráter gradual para observar a densidade terminológica dos textos, pois, por exemplo, um texto sobre resultado de partida publicado por um tabloide inglês pode apresentar uma terminologia bem distinta daquela expressa em um texto do mesmo gênero publicado por um jornal mais tradicional. Essa variação terminológica não se refere apenas à densidade terminológica de um texto, mas também às diferentes terminologias encontradas em periódicos distintos. Por esse motivo, mais do que o tema, o grau de densidade informativa, a forma de se comunicar, a situação em que se comunica e para quem se comunica constituem fatores determinantes do grau de especialização de um texto.

Krieger e Finatto (2004) propõem que a variação tipológica no âmbito da comunicação especializada se reflete, por exemplo, na distinção entre uma tese, um artigo de periódico altamente especializado em determinada área do conhecimento e um texto de jornal ou de revista informativa redigido com a finalidade de divulgar ao grande público um acontecimento científico. Essa distinção não pode ser aplicada ao nosso caso por vários motivos. Primeiramente, devido à grande repercussão que o futebol tem na sociedade moderna, a delimitação de "grande público" e de especialistas se torna um pouco controversa. Em um primeiro momento, tenderíamos a caracterizar os jornalistas esportivos, jogadores, técnicos e membros da comissão técnica como especialistas e os leitores e fãs de futebol em geral como "grande público". No entanto, o que acontece é que, em se tratando de futebol, todo torcedor pode ser considerado um especialista na área.

Devido a essa grande dificuldade de delimitação de "grande público" e de especialistas, adotamos, aqui, a classificação de atores sociais do futebol proposta por Toledo (2002, p. 15):

> Parto de três [atores sociais] dentro do campo esportivo: os profissionais (jogadores, técnicos, dirigentes, juízes, preparadores, médicos, etc.), os especialistas (a crônica esportiva) e o conjunto genérico de torcedores, "comuns" ou nomeados e reunidos em certas coletividades específicas.

Essa classificação é utilizada por Toledo para uma análise das lógicas do futebol, com o propósito de caracterizar os grupos sociais que se expressam por meio do futebolês. Dentro dessa ordem, os profissionais são todos aqueles que interferem diretamente no jogo, quer dentro do campo, como os jogadores, quer fora do campo, como os dirigentes e instituições como federações. Os especialistas são representados pela crônica esportiva e pelo jornalismo esportivo e, segundo Toledo, ocupam um lugar simbólico equidistante entre os profissionais e os torcedores, pois não participam efetivamente da partida, mas também não se comprometem em um nível de emoção partidária, pelo menos em teoria. Por fim, o "grande público" é caracterizado pelos torcedores.

É a classe dos especialistas que mais nos interessa neste trabalho, pois é através do produto do fazer jornalístico que observamos o futebol fora de seu *locus* de ritualização máxima, que é a partida. Ademais, é também por meio dos textos jornalísticos que o futebol alcança o "grande público". Contudo, não podemos deixar de enfatizar a importância que o terceiro grupo, o conjunto genérico de torcedores, desempenha em nossa pesquisa, pois frequentemente recorremos a torcedores e amantes do esporte para melhor entender uma jogada ou drible.

Deve-se somar, ainda, à classificação de Toledo, o grau de densidade informativa da notícia, a forma de se comunicar, os traços de impessoalidade do jornalista, o contexto em que a notícia foi produzida[3], o propósito da comunicação, que pode ser mais informativo, mais descritivo, mais técnico, para que, assim, seja possível caracterizar o texto técnico propriamente.

Nessa concepção de texto especializado, a presença de termos deixa de ser o elemento primordial que configura o caráter de uma comunicação especializada (KRIEGER; FINATTO, 2004). Ao contrário, os mecanismos linguísticos, textuais e pragmáticos dos quais um texto especializado faz uso também constituem elementos caracterizadores de dada língua

[3] Uma notícia sobre o resultado de partida de um jogo do campeonato brasileiro provavelmente é produzida em um contexto mais específico e, portanto, voltada para um público mais específico do que uma notícia sobre uma partida de Copa do Mundo, pois durante esse acontecimento mais pessoas torcem, inclusive torcedores não habituais.

de especialidade. O uso de todos esses elementos em conjunto nos permite enxergar a complexidade do texto especializado.

Uma vez caracterizado seu principal objeto de estudo, passemos, agora, a uma breve discussão sobre as características da Terminologia Textual.

Para Krieger e Finatto (2004,), essa Terminologia, na qual o objeto 'termo' dá lugar ao objeto 'texto', possui duas características principais. A primeira refere-se ao reconhecimento do papel do cenário comunicativo e, consequentemente, do texto especializado para a descrição de uma terminologia. A segunda está relacionada ao estudo e caracterização do texto especializado.

Ao reconhecer o papel do cenário comunicativo, a Terminologia textual parte do pressuposto de que os termos são utilizados em situação de comunicação e que, portanto, não devem, ou melhor, não podem ser estudados à parte do contexto sociocultural em que ocorrem.

Os termos passaram a ser analisados em uso, ou seja, em textos especializados, permitindo, assim, a verificação de fenômenos até então ignorados, ou deixados em segundo plano pelos terminólogos. O contexto discursivo, antes considerado insignificante, passa a representar uma das principais características da Terminologia, contribuindo para um novo tratamento das Unidades Terminológicas (UTs), deixando de lado a ideia de que os termos constituem construtos teóricos idealizados em um sistema linguístico independente.

5. A Linguística de *Corpus* neste estudo

São inúmeras as definições para Linguística de *Corpus* (LC) e não nos cabe aqui, por questões de tempo e próposito de pesquisa, apresentar um panorama com todas as definições, que diferem consideralvelmente umas das outras.

Os teóricos da LC frequentemente discordam quanto à sua caracterização. Autores como Berber Sardinha (2000), consideram a LC uma abordagem, outros, como é o caso de Rocha (2001), uma metodologia. Existem ainda os estudiosos que preferem ser mais neutros em suas definições e não tomam nenhum partido como, por exemplo, Aijmer e Altenberg (1991, p. 2), que definem a LC como "o estudo da língua por meio de *corpora*". Há ainda os que adotam as duas definições como Bowker e Pearson (2002, p. 20): "uma abordagem ou metodologia para o estudo da língua em uso".

Berber Sardinha chama atenção para o fato de que a LC não constitui somente um

metodologia, um instrumental, do qual outras áreas podem se valer para o estudo da linguagem. Para o autor, a LC apresenta também uma nova perspectiva de se chegar à linguagem possibilitando aos seus seguidores produzir conhecimento novo que muitas vezes coloca teorias linguísticas tradicionais em questão.

Neste trabalho, adotamos duas definições de LC que, até o presente momento, nos pareceram as mais abrangentes. Primeiramente, assim como Berber Sardinha (2004 p.32), acreditamos que a LC constitui uma abordagem empirista que toma a língua como sistema probabilístico, refletindo uma nova maneira de enxergar a linguagem que, consequente e futuramente, poderá dar origem a uma nova teoria linguística. Adotamos, também, a visão de McEnery e Hardie (2012:1), que definem LC como "[...] área que foca em um conjunto de procedimentos, ou métodos, para o estudo da língua [...][4].

As duas definições acima se enquadrariam perfeitamente nos propósitos de nossa pesquisa se utilizássemos a LC unicamente para explorar um fenômeno linguístico. No entanto, o escopo de nossa pesquisa vai além da esfera linguística; estabelecemos, sempre que possível, um paralelo entre língua, cultura e história. Ademais, acreditamos que a LC pode e deve ser utilizada como metodologia, e, sim, unicamente como metodologia para pesquisas de outras áreas como, por exemplo, história, antropologia e jornalismo. Por esse motivo, acreditamos que as visões de Berber Sardinha e de McEnery e Hardie se mostram limitadas às áreas de estudos da linguagem. É fato que existe uma tendência mais atual de utilizar a LC em áreas afins; um bom exemplo dessa tendência foram os três cursos de verão oferecidos pela Lancaster University em julho de 2013: *UCREL Summer School in Corpus linguistics, ESRC Summer School in GIS in Corpus Approaches for Social Sciences* e *ERC Summer School in GIS for the Digital Humanities*. Como podemos observar, somente um dos cursos foi direcionado a linguistas.

Enfim, não nos cabe, aqui, criar uma nova definição de LC, apenas atentamos para o fato de que, embora utilizemos essas duas definições, acreditamos que a LC pode ser utilizada para outros fins, que não se restrinjam a pesquisas linguísticas.

Um *corpus* pode ser utilizado de diferentes maneiras para validar, exemplificar, contestar ou formular teorias linguísticas. Tognini-Bonelli (2001) distingue duas abordagens principais de pesquisa realizadas em *corpora*: abordagem baseada em *corpus* (*corpus-based*) e abordagem direcionada pelo *corpus* (*corpus-driven*).

[4] "[…] area which focuses upon a set of procedures, or methods, for studying language […]

Na abordagem baseada em *corpus*, o linguista utiliza o *corpus* para explicitar, testar e exemplificar teorias e hipóteses pré-existentes e, principalmente, para extrair exemplos.

A vantagem dessa abordagem é que a extração de exemplos autênticos, seja para fins lexicográficos ou para a validação de hipóteses, confere mais autoridade à pesquisa. Por outro lado, utilizar o *corpus* somente para verificar dados limita a visão do linguista, que ignora novos fenômenos deixando de fazer novas descobertas e de desafiar teorias já existentes.

Na abordagem direcionada pelo *corpus*, o linguista analisa o *corpus* sem hipóteses pré-concebidas. O *corpus* mostra-lhe o caminho a ser percorrido. As descrições são feitas sempre com base nas evidências do *corpus*, possibilitando, assim, novas descobertas. Por isso, dizemos que nessa abordagem o linguista não busca evidências para classificá-las dentro de categorias pré-definidas. Ao contrário, se no decorrer da pesquisa não forem encontrados padrões linguísticos ou se os padrões encontrados não puderem ser classificados em alguma categoria, os achados constituirão argumentos de extrema relevância para a descrição da linguagem ou para a descoberta de novos fenômenos.

Nessa abordagem, o caminho metodológico percorrido pelo linguista é claro: a observação dos dados conduz à formulação de hipóteses que, consequentemente, leva à generalização dos resultados possibilitando, assim, a formulação de novas teorias (TOGNINI-BONELLI, 2001).

Apesar de as duas abordagens apresentarem características bem distintas, acreditamos que podem ser utilizadas em conjunto. Neste trabalho, utilizamos a abordagem direcionada pelo *corpus* para extrair os termos a serem estudados por meio das palavras-chave e de seus agrupamentos (*clusters*). Por outro lado, lançamos mão da abordagem baseada em *corpus* quando partimos de uma tradução *prima facie* para a busca dos equivalentes nas linhas de concordância.

6. O *corpus* de estudo

O *design* e a qualidade do *corpus* de estudo constituem o pilar de qualquer pesquisa em *corpus*. O quadro que segue mostra o *design* do ***corpus*** utilizado neste estudo[5]:

[5] Chamamos atenção para o fato de que o *corpus* de estudo utilizado na presente pesquisa foi compilado para atender os objetivos de nossa pesquisa de doutorado, a saber: 1) verificar como os diferentes jeitos de jogar, a história do futebol em cada cultura, a apropriação cultural das regras na Inglaterra e no Brasil e outros fatores de ordem histórico-social contribuíram para a criação do léxico do futebol em português e inglês; 2) criar um glossário de futebol composto por verbetes que evidenciem diferenças culturais entre o Brasil e a Inglaterra.

Quadro 1: composição do *corpus* de estudo.

<table>
<tr><td>Conteúdo</td><td colspan="4">especializado</td></tr>
<tr><td>Assunto</td><td colspan="4">futebol</td></tr>
<tr><td>Autoria</td><td colspan="4">de língua nativa</td></tr>
<tr><td>Língua</td><td colspan="4">português (BR) e inglês (ING); comparável</td></tr>
<tr><td>Finalidade</td><td colspan="4">de estudo</td></tr>
<tr><td>Meio</td><td colspan="4">eletrônico</td></tr>
<tr><td>Modo</td><td colspan="4">escrito</td></tr>
<tr><td>tipo de texto</td><td colspan="4">resultados de partidas, narrações minuto a minuto e narrações minuto a minuto com comentários de internautas.</td></tr>
<tr><td>Período</td><td colspan="4">2013 – 2014</td></tr>
<tr><td>Seleção</td><td colspan="4">de amostragem; balanceado</td></tr>
<tr><td rowspan="3">Tamanho</td><td colspan="4">aproximadamente 500 mil palavras em cada língua</td></tr>
<tr><td colspan="2">Inglês</td><td colspan="2">Português</td></tr>
<tr><td>No. palavras</td><td>No. textos</td><td>No. Palavras</td><td>No. textos</td></tr>
<tr><td></td><td>600.079</td><td>612</td><td>469.765</td><td>864</td></tr>
</table>

O *corpus* utilizado na pesquisa é composto por textos de três periódicos ingleses sobre resultados de partidas da primeira divisão do campeonato inglês de 2013/2014[6] e por textos de três periódicos brasileiros sobre resultados de partidas da primeira divisão do campeonato brasileiro de 2013[7].

O *corpus* é caracterizado como *corpus* especializado, uma vez que é composto por textos de uma única área de especialidade: futebol. Coletamos somente textos escritos originalmente em português brasileiro e em inglês britânico.

No que se refere ao modo, nosso *corpus* é escrito, pois não trabalhos com textos orais. Optamos por coletar três tipos de textos: resultados de partidas, narrações minuto a minuto, que são escritas durante a partida por um narrador *on-line*; e narrações minuto a minuto com comentários de internautas, também escritas durante a partida por um narrador e com contribuições de internautas que compartilham suas opiniões sobre os lances do jogo.

7. O *corpus* de referência

O *corpus* de referência é utilizado para contrastar com o *corpus* de estudo a fim de evidenciar as formas mais frequentes nesse último, filtrando os elementos mais genéricos. Em

[6] Os periódicos selecionados foram: o jornal *The Guardian*, o tabloide *Daily Mail* e o site sobre futebol *Football.com*.
[7] Os periódicos selecionados foram: o jornal *O Estado de São Paulo*, o jornal esportivo *Gazeta Esportiva* e a revista sobre futebol *Placar*.

geral, deve-se incluir vários gêneros textuais em um *corpus* de referência, de modo que proporcione uma escolha não-marcada das palavras-chave, pois suas características influenciam de forma direta os tipos de palavra que podem se tornar chave (BERBER SARDINHA, 2004).

O tamanho do *corpus* de referência pode influenciar o número de palavras-chave obtidas. Berber Sardinha (2004, p. 102) recomenda que o *corpus* de referência seja entre três e cinco vezes maior que o *corpus* de estudo.

Compilar um *corpus* muito maior do que o recomendado não retornará, necessariamente, maior número de palavras-chave (BERBER SARDINHA, 2004). No entanto, não existem restrições que limitem o tamanho do *corpus* de referência.

Em nossa pesquisa utilizamos dois *corpora* de referência: o BNC (British National *Corpus*) [8] e o Banco de Português[9]. O BNC, *corpus* fechado (1990-1994), possui 100 milhões de palavras e foi desenvolvido com o objetivo de ser representativo das variantes escrita e falada do inglês britânico. O Banco de Português é um *corpus* monitor do português do Brasil, ou seja, está aberto e é constantemente atualizado. Conta com aproximadamente um bilhão de palavras.

8. Metodologia

Para nossa pesquisa, utilizamos o *software WordSmith Tools* versão 5.0, desenvolvido por Mike Scott e publicado pela *Oxford University Press*. O programa possui três ferramentas principais *WordList, KeyWords* e *Concord* e uma série de aplicativos extremamente úteis para a análise linguística. Ressaltamos que o *WordSmith* Tools 5 é um *software* riquíssimo para a análise linguística, sendo que cada uma de suas ferramentas possui vários instrumentos de análise. Contudo, descreveremos aqui somente as ferramentas e aplicativos que estão sendo utilizados em nossa pesquisa.

Para extrair os candidatos a termo, geramos uma lista de palavras-chave para o *corpus* de português, já que a extração terminológica foi realizada na direção português-inglês. Para tanto, comparamos a *wordlist* do nosso *corpus* com a *wordlist* do *corpus* de referência para obter as *keywords*.

[8] Disponível em: < http://corpus.byu.edu/bnc/ > Acesso em: 15 ago 2014.
[9] Disponível em: http://www2.lael.pucsp.br/corpora/bp/conc/ Acesso em: 15 ago 2014.

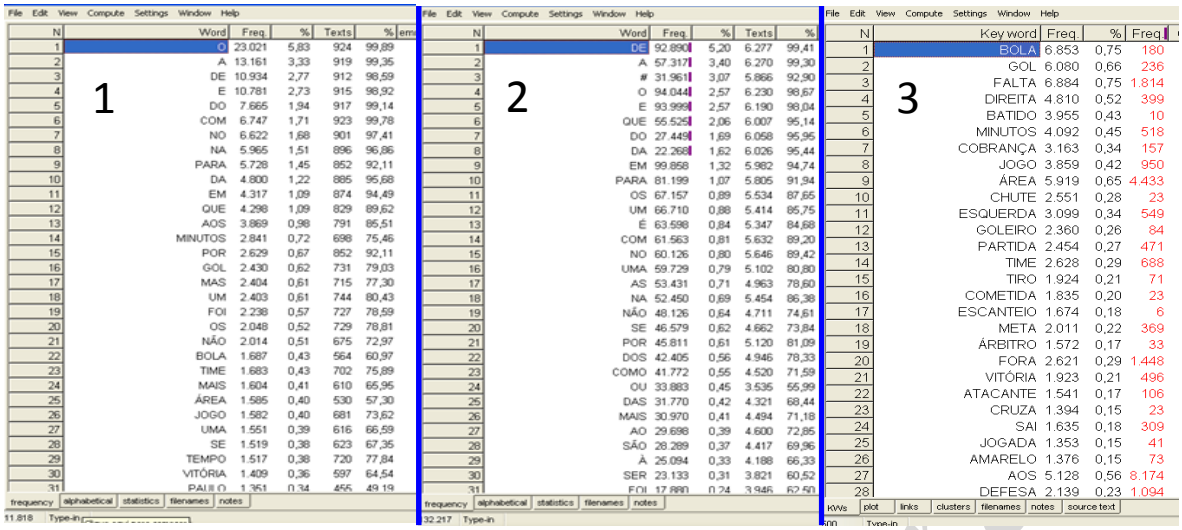
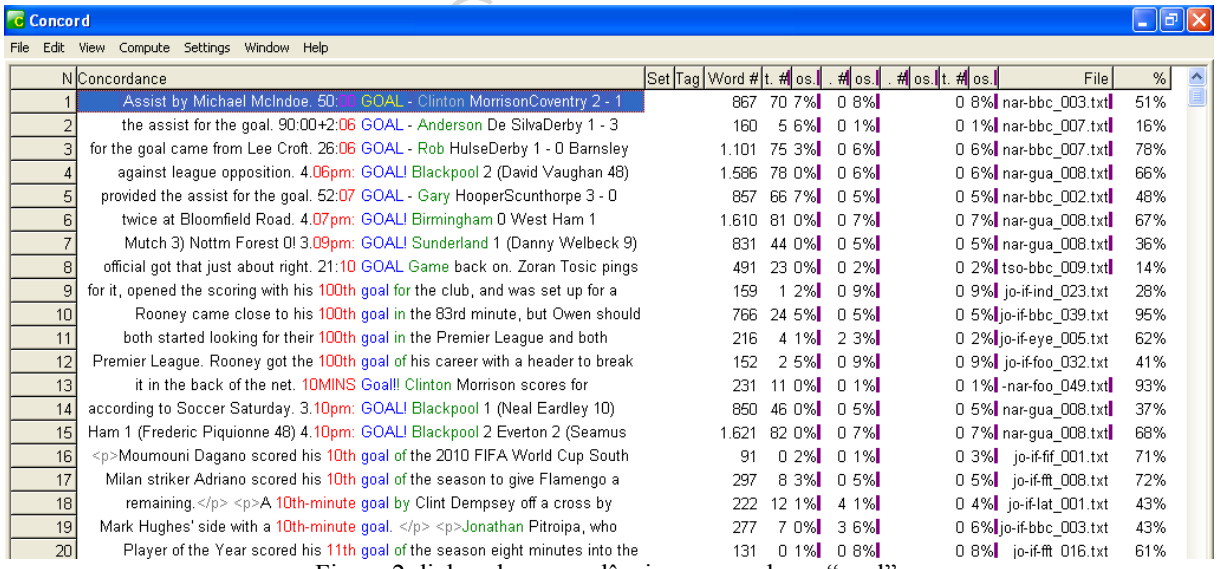


Figura 1: 1: lista de palavras do *corpus* de estudo; 2: lista de palavras do *corpus* de referência; 3: lista de palavras-chave.

Após obter as palavras-chave, geramos linhas de concordância para as palavras-chave e examinamos seus agrupamentos (*clusters*) a fim de encontrar colocações e unidades fraseológicas. Para tanto, utilizamos a ferramenta *Concord,* que gera listas das ocorrências de um item específico, chamado de 'palavra de busca' ou 'nódulo'. Esse item pode ser formado por uma ou mais palavras e é apresentado com o contexto ao seu redor (Figura 2).



N	Concordance	Set	Tag	Word #	t. #	os.	. #	os.	. #	os.	t. #	os.	File	%
1	Assist by Michael McIndoe. 50:00 GOAL - Clinton MorrisonCoventry 2 - 1			867	70	7%	0	8%			0	8%	nar-bbc_003.txt	51%
2	the assist for the goal. 90:00+2:06 GOAL - Anderson De SilvaDerby 1 - 3			160	5	6%	0	1%			0	1%	nar-bbc_007.txt	16%
3	for the goal came from Lee Croft. 26:06 GOAL - Rob HulseDerby 1 - 0 Barnsley			1.101	75	3%	0	6%			0	6%	nar-bbc_007.txt	78%
4	against league opposition. 4.06pm: GOAL! Blackpool 2 (David Vaughan 48)			1.586	78	0%	0	6%			0	6%	nar-gua_008.txt	66%
5	provided the assist for the goal. 52:07 GOAL - Gary HooperScunthorpe 3 - 0			857	66	7%	0	5%			0	5%	nar-bbc_002.txt	48%
6	twice at Bloomfield Road. 4.07pm: GOAL! Birmingham 0 West Ham 1			1.610	81	0%	0	7%			0	7%	nar-gua_008.txt	67%
7	Mutch 3) Nottm Forest 0! 3.09pm: GOAL! Sunderland 1 (Danny Welbeck 9)			831	44	0%	0	5%			0	5%	nar-gua_008.txt	36%
8	official got that just about right. 21:10 GOAL Game back on. Zoran Tosic pings			491	23	0%	0	2%			0	2%	tso-bbc_009.txt	14%
9	for it, opened the scoring with his 100th goal for the club, and was set up for a			159	1	2%	0	9%			0	9%	jo-if-ind_023.txt	28%
10	Rooney came close to his 100th goal in the 83rd minute, but Owen should			766	24	5%	0	5%			0	5%	jo-if-bbc_039.txt	95%
11	both started looking for their 100th goal in the Premier League and both			216	4	1%	2	3%			0	2%	jo-if-eye_005.txt	62%
12	Premier League. Rooney got the 100th goal of his career with a header to break			152	2	5%	0	9%			0	9%	jo-if-foo_032.txt	41%
13	it in the back of the net. 10MINS Goal!! Clinton Morrison scores for			231	11	0%	0	1%			0	1%	-nar-foo_049.txt	93%
14	according to Soccer Saturday. 3.10pm: GOAL! Blackpool 1 (Neal Eardley 10)			850	46	0%	0	5%			0	5%	nar-gua_008.txt	37%
15	Ham 1 (Frederic Piquionne 48) 4.10pm: GOAL! Blackpool 2 Everton 2 (Seamus			1.621	82	0%	0	7%			0	7%	nar-gua_008.txt	68%
16	Moumouni Dagano scored his 10th goal of the 2010 FIFA World Cup South			91	0	2%	0	1%			0	3%	jo-if-fif_001.txt	71%
17	Milan striker Adriano scored his 10th goal of the season to give Flamengo a			297	8	3%	0	5%			0	5%	jo-if-ft_008.txt	72%
18	remaining. A 10th-minute goal by Clint Dempsey off a cross by			222	12	1%	4	1%			0	4%	jo-if-lat_001.txt	43%
19	Mark Hughes' side with a 10th-minute goal. Jonathan Pitroipa, who			277	7	0%	3	6%			0	6%	jo-if-bbc_003.txt	43%
20	Player of the Year scored his 11th goal of the season eight minutes into the			131	0	1%	0	8%			0	8%	jo-if-ft_016.txt	61%

Figura 2: linhas de concordância para a palavra "goal".

A figura 2 mostra as linhas de concordância de 'goal' ordenadas alfabeticamente pela primeira palavra à esquerda.

O próximo passo foi examinar as linhas de concordância e, quando necessário, gerar

clusters para essas linhas. A figura 3 mostra os *clusters* de 'gol':

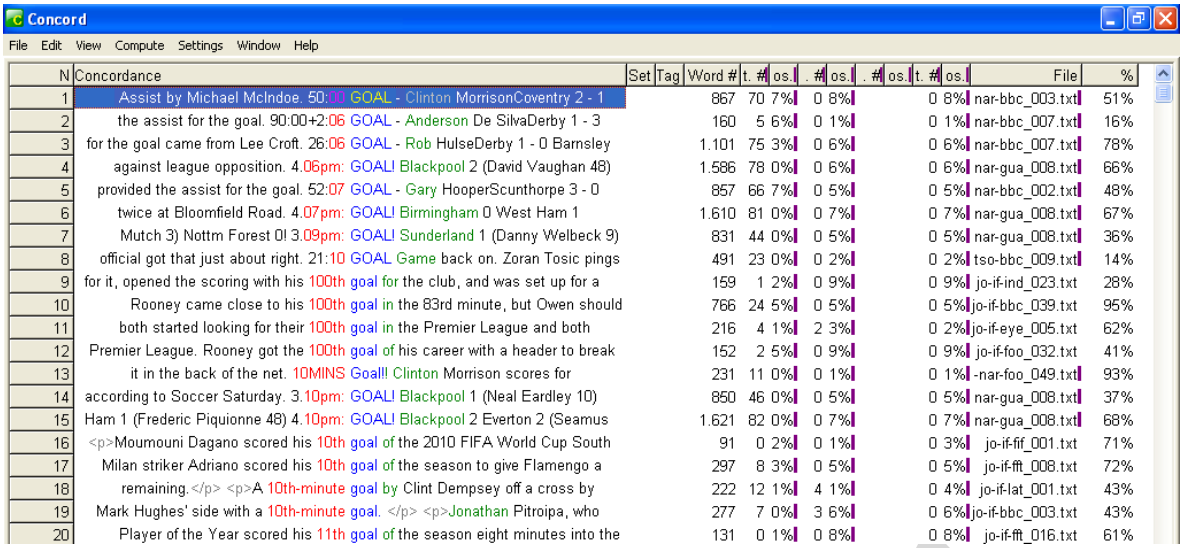
N	CLUSTER	FREQ.	N	CLUSTER	FREQ.
1	GOL FEITO POR	561	11	ESQUERDA NO GOL	169
2	NO GOL FEITO	561	12	DE ESQUERDA NO	169
3	DIREITA NO GOL	391	13	DO GOL E	165
4	DE DIREITA NO	391	14	O SEGUNDO GOL	149
5	PARA DO GOL	318	15	DEPOIS DO GOL	146
6	POR CIMA DO	230	16	O GOL DA	126
7	CIMA DO GOL	221	17	SAÍDA DEPOIS DO	123
8	DO GOL DE	196	18	COM GOL DE	109
9	O PRIMEIRO GOL	173	19	LONGE DO GOL	108
10	O GOL DE	173	20	SAI DO GOL	106

Figura 3: primeiros 20 *clusters* de "gol"

No ajuste realizado acima para 'gol', o programa foi preparado para encontrar *clusters* de três palavras, com frequência mínima de três e em uma janela de cinco palavras à direita e cinco à esquerda.

O último passo para extrair os candidatos a termo foi expandir as linhas de concordância das palavras-chave e acessar o texto integral da ocorrência, sempre que necessário, para melhor entender seu funcionamento e validá-las como termos.

Utilizamos procedimentos diferentes para o estabelecimento de equivalentes tradutórios de cada termo. A não padronização de uma metodologia deu-se pelo fato de que alguns termos apresentam um equivalente *prima-facie* em inglês, sempre mais simples de ser encontrado, ao passo que outros não apresentam esse tipo de equivalente e alguns sequer possuem equivalente. Em suma, geramos linhas de concordâncias para os termos e unidades fraseológicas especializadas (UFEs) em português, para melhor entender seu funcionamento, ou seja, o tipo de texto, a situação em campo e o contexto em que ocorrem. Após entender o funcionamento dos termos e UFEs em português, geramos linhas de concordância para suas traduções *prima facie* em inglês. Por exemplo, para chegar ao equivalente de "gol", geramos linhas de concordância para *goal*. A figura 4 mostra as linhas de concordância do termo *goal*:



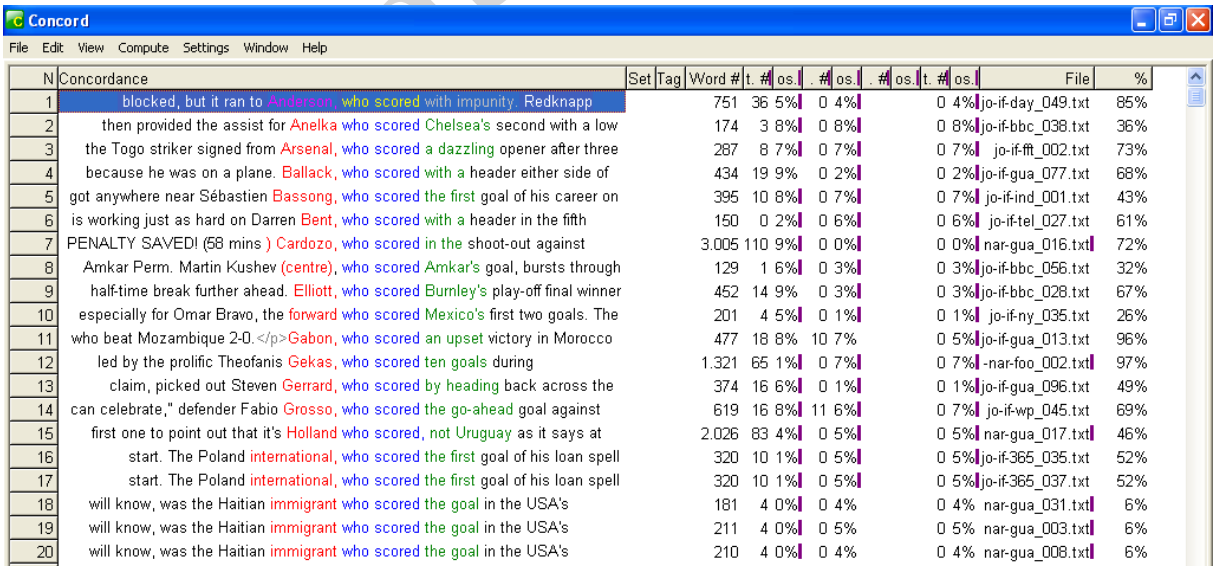
Concord

File Edit View Compute Settings Window Help

N	Concordance	Set	Tag	Word #	t. #	os.	. #	os.	. #	os.	t. #	os.	File	%
1	Assist by Michael McIndoe. 50:00 GOAL - Clinton MorrisonCoventry 2 - 1			867	70	7%	0	8%			0	8%	nar-bbc_003.txt	51%
2	the assist for the goal. 90:00+2:06 GOAL - Anderson De SilvaDerby 1 - 3			160	5	6%	0	1%			0	1%	nar-bbc_007.txt	16%
3	for the goal came from Lee Croft. 26:06 GOAL - Rob HulseDerby 1 - 0 Barnsley			1.101	75	3%	0	6%			0	6%	nar-bbc_007.txt	78%
4	against league opposition. 4.06pm: GOAL! Blackpool 2 (David Vaughan 48)			1.586	78	0%	0	6%			0	6%	nar-gua_008.txt	66%
5	provided the assist for the goal. 52:07 GOAL - Gary HooperScunthorpe 3 - 0			857	66	7%	0	5%			0	5%	nar-bbc_002.txt	48%
6	twice at Bloomfield Road. 4.07pm: GOAL! Birmingham 0 West Ham 1			1.610	81	0%	0	7%			0	7%	nar-gua_008.txt	67%
7	Mutch 3) Nottm Forest 0! 3.09pm: GOAL! Sunderland 1 (Danny Welbeck 9)			831	44	0%	0	5%			0	5%	nar-gua_008.txt	36%
8	official got that just about right. 21:10 GOAL Game back on. Zoran Tosic pings			491	23	0%	0	2%			0	2%	tso-bbc_009.txt	14%
9	for it, opened the scoring with his 100th goal for the club, and was set up for a			159	1	2%	0	9%			0	9%	jo-if-ind_023.txt	28%
10	Rooney came close to his 100th goal in the 83rd minute, but Owen should			766	24	5%	0	5%			0	5%	jo-if-bbc_039.txt	95%
11	both started looking for their 100th goal in the Premier League and both			216	4	1%	2	3%			0	2%	jo-if-eye_005.txt	62%
12	Premier League. Rooney got the 100th goal of his career with a header to break			152	2	5%	0	9%			0	9%	jo-if-foo_032.txt	41%
13	it in the back of the net. 10MINS Goal!! Clinton Morrison scores for			231	11	0%	0	1%			0	1%	-nar-foo_049.txt	93%
14	according to Soccer Saturday. 3.10pm: GOAL! Blackpool 1 (Neal Eardley 10)			850	46	0%	0	5%			0	5%	nar-gua_008.txt	37%
15	Ham 1 (Frederic Piquionne 48) 4.10pm: GOAL! Blackpool 2 Everton 2 (Seamus			1.621	82	0%	0	7%			0	7%	nar-gua_008.txt	68%
16	Moumouni Dagano scored his 10th goal of the 2010 FIFA World Cup South			91	0	2%	0	1%			0	3%	jo-if-fif_001.txt	71%
17	Milan striker Adriano scored his 10th goal of the season to give Flamengo a			297	8	3%	0	5%			0	5%	jo-if-fft_008.txt	72%
18	remaining. A 10th-minute goal by Clint Dempsey off a cross by			222	12	1%	4	1%			0	4%	jo-if-lat_001.txt	43%
19	Mark Hughes' side with a 10th-minute goal. Jonathan Pitroipa, who			277	7	0%	3	6%			0	6%	jo-if-bbc_003.txt	43%
20	Player of the Year scored his 11th goal of the season eight minutes into the			131	0	1%	0	8%			0	8%	jo-if-fft_016.txt	61%

Figura 4: linhas de concordância de *goal*

Esse tipo de pesquisa nos permitiu validar *goal* como equivalente de "gol" uma vez que observamos que o contexto em que as duas palavras ocorrem é o mesmo. No entanto, não se mostrou eficiente para a validação de alguns equivalentes como, por exemplo, o equivalente da unidade fraseológica "fazer um gol", já que não encontramos um verbo que co-ocorre com *goal* com frequência relativamente alta. Por esse motivo, resolvemos gerar linhas de concordância para "scored", tradução *prima facie* de 'marcar') a fim de verificar se o termo co-ocorre com 'gol'. Observemos a figura 5:



Concord

File Edit View Compute Settings Window Help

N	Concordance	Set	Tag	Word #	t. #	os.	. #	os.	. #	os.	t. #	os.	File	%
1	blocked, but it ran to Anderson, who scored with impunity. Redknapp			751	36	5%	0	4%			0	4%	jo-if-day_049.txt	85%
2	then provided the assist for Anelka who scored Chelsea's second with a low			174	3	8%	0	8%			0	8%	jo-if-bbc_038.txt	36%
3	the Togo striker signed from Arsenal, who scored a dazzling opener after three			287	8	7%	0	7%			0	7%	jo-if-fft_002.txt	73%
4	because he was on a plane. Ballack, who scored with a header either side of			434	19	9%	0	2%			0	2%	jo-if-gua_077.txt	68%
5	got anywhere near Sébastien Bassong, who scored the first goal of his career on			395	10	8%	0	7%			0	7%	jo-if-ind_001.txt	43%
6	is working just as hard on Darren Bent, who scored with a header in the fifth			150	0	2%	0	6%			0	6%	jo-if-tel_027.txt	61%
7	PENALTY SAVED! (58 mins) Cardozo, who scored in the shoot-out against			3.005	110	9%	0	0%			0	0%	nar-gua_016.txt	72%
8	Amkar Perm. Martin Kushev (centre), who scored Amkar's goal, bursts through			129	1	6%	0	3%			0	3%	jo-if-bbc_056.txt	32%
9	half-time break further ahead. Elliott, who scored Burnley's play-off final winner			452	14	9%	0	3%			0	3%	jo-if-bbc_028.txt	67%
10	especially for Omar Bravo, the forward who scored Mexico's first two goals. The			201	4	5%	0	1%			0	1%	jo-if-ny_035.txt	26%
11	who beat Mozambique 2-0. Gabon, who scored an upset victory in Morocco			477	18	8%	10	7%			0	5%	jo-if-gua_013.txt	96%
12	led by the prolific Theofanis Gekas, who scored ten goals during			1.321	65	1%	0	7%			0	7%	-nar-foo_002.txt	97%
13	claim, picked out Steven Gerrard, who scored by heading back across the			374	16	6%	0	1%			0	1%	jo-if-gua_096.txt	49%
14	can celebrate," defender Fabio Grosso, who scored the go-ahead goal against			619	16	8%	11	6%			0	7%	jo-if-wp_045.txt	69%
15	first one to point out that it's Holland who scored, not Uruguay as it says at			2.026	83	4%	0	5%			0	5%	nar-gua_017.txt	46%
16	start. The Poland international, who scored the first goal of his loan spell			320	10	1%	0	5%			0	5%	jo-if-365_035.txt	52%
17	start. The Poland international, who scored the first goal of his loan spell			320	10	1%	0	5%			0	5%	jo-if-365_037.txt	52%
18	will know, was the Haitian immigrant who scored the goal in the USA's			181	4	0%	0	4%			0	4%	nar-gua_031.txt	6%
19	will know, was the Haitian immigrant who scored the goal in the USA's			211	4	0%	0	5%			0	5%	nar-gua_003.txt	6%
20	will know, was the Haitian immigrant who scored the goal in the USA's			210	4	0%	0	4%			0	4%	nar-gua_008.txt	6%

Figura 5: parte das linhas de concordância de *scored* ordenadas pela duas primeiras palavras à esquerda.

Em um primeiro momento, fomos levados a validar *score* ou *score a goal* como

equivalente de "fazer/marcar um gol", o que não está errado. Entretanto, a diferença entre a frequência da UFE em português "marcar/fazer um gol" e da UFE em inglês "score a goal" era muito grande. Decidimos, então, conduzir buscas para todos os verbos do *corpus* utilizando as seguintes etiquetas morfossintáticas: VV (verbo no infinitivo), VVD (verbo no passado simples), VVG (verbo no gerúndio), VVN (verbo no particípio passado) e VVZ (verbo no presente – 3a pessoa do singular) a fim de encontrar outros verbos que descrevessem o ato de fazer um gol.

A etiquetagem morfossintática, ou *part of speech tagging* (POS), consiste em colocar em cada palavra do *corpus* uma etiqueta que indique sua classe gramatical. A etiquetagem foi realizada com o etiquetador *Tree-Tagger*, disponibilizado pelo professor Tony Berber Sardinha, da Pontífica Universidade Católica de São Paulo (PUC), no site do LAEL (Programa de Linguística Aplicada e Estudos da Linguagem)[10].

Os resultados dessas buscas serão detalhados mais adiante no estudo de caso. Ressaltamos que priorizamos os *corpora* para a extração e validação dos termos e UFEs, bem como para a busca de equivalentes. No entanto, sempre que necessário, recorremos à internet quando não encontramos o equivalente no *corpus* ou quando precisamos esclarecer a origem ou o significado de algum termo.

9. Estudo de caso: FAZER[11] um gol

Ao analisar as linhas de concordância de "gol" em nosso *corpus* de português, encontramos, ao total, 920 ocorrências de **FAZER|MARCAR [o|um] gol / MARCAR o tento / FINALIZAR / EMPATAR**, termos e unidades fraseológicas utilizados para descrever um gol.

Ao buscar pelos possíveis equivalentes em inglês, encontramos 259 ocorrências de:

- SCORE({a | the} goal)
- SNATCH [a goal]
- DELIVER a goal
- ADD {a|the} ORDINAL (goal)
- BLAST in ([a goal])

A primeira pergunta que nos ocorreu foi: por que temos mais ocorrências de "gol" em

[10] disponível em: <http://corpuslg.org/tools/etiquetagem/>
[11] utilizamos letras maiúsculas para indicar a forma lematizada dos verbos

português do que de *goal* em inglês, já que o número de gols da Série A do Campeonato Brasileiro é quase o mesmo da *Premier League*? A figura 6 mostra o número de gols do Campeonato Brasileiro de 2006 a 2013 e o número de gols das temporada 2011/2012, 2012/2013 e 2013/2014 da *Premier League*:

Figura 6: número de gols do Campeonato Brasileiro de 2006 a 2013[12] e número de gols da *Premier League* das temporadas de 2011/2012, 2012/2013 e 2013/2014[13]

Como podemos observar, o Campeonato Brasileiro de 2013 teve 936 gols e a *Premier League* de 2013/2014 teve 1052 gols. Nosso segundo questionamento foi: se *goal* não descreve os gols marcados, como esses gols estão sendo narrados? Nesse momento, nos demos conta de que o verbo *score*, presente na unidade fraseológica *SCORE [{a | the} goal]*, muitas vezes ocorria sozinho, observemos os exemplos que seguem:

a) The United States got out to an early lead as Clint Dempsey **scored** just 32 seconds into the game for the fastest goal in World Cup history for the U.S.

b) Striker Asamoah Gyan also **scored** right before the first half ended.

Mais adiante, ao ler os textos que coletávamos para compilar o *corpus*, nos demos conta de que a língua inglesa utiliza outros verbos que, muito frequentemente, não co-ocorrem com o termo "gol", ou seja, são usados como sinônimos de *SCORE [{a | the} goal]*. Sendo assim, geramos linhas de concordância para as etiquetas de verbo do *corpus* em inglês

[12] Disponível em: < http://www.srgoool.com.br/Noticia/Brasileirao-2013-teve-o-menor-numero-de-gols-e-a-pior-media-de-tentos-dos-pontos-corridos> Acesso: 16 jun 2014.
[13] Disponível em: <http://www.worldfootball.net/stats/eng-premier-league/1/ > Acesso: 16 jun 2014.

e analisamos todas as ocorrências à procura de verbos que descrevessem um gol. A tabela 1 mostra o número de linhas de concordância para cada etiqueta do *corpus*:

Tabela 1: Número de linhas de concordância para cada etiqueta gramatical

etiqueta	**número de ocorrências**
VB - Verb, base form	25.283
VBD - Verb, past tense	28.778
VBG - Verb, gerund or present participle	19.406
VBN – verb, past particple	20.457
VBZ - Verb, 3rd person singular present	13.853

Aplicamos o dispositivo *Re-sort*, que permite ordenar as linhas de concordância pelo nódulo de busca, e ajustamos as configurações para que as linhas fossem ordenadas pela etiqueta gramatical e pela primeira palavra à direita (forma canônica do verbo) e à esquerda.

Observamos todas as linhas de concordância e identificamos 43 verbos que poderiam fazer parte de UFEs utilizadas para relatar um gol. Depois, geramos concordâncias no *corpus* sem etiquetas gramaticais para esses verbos a fim verificar como ocorrem nos textos, ou seja, os padrões em que ocorrem.

O passo seguinte foi validar essas unidades fraseológicas por meio da análise das concordâncias e do contexto expandido. Ainda no caso de *fire*, ampliamos os contextos das linhas em que *fire home* era um possível equivalente de "fazer um gol". Observemos, por exemplo, a ocorrência número 153:

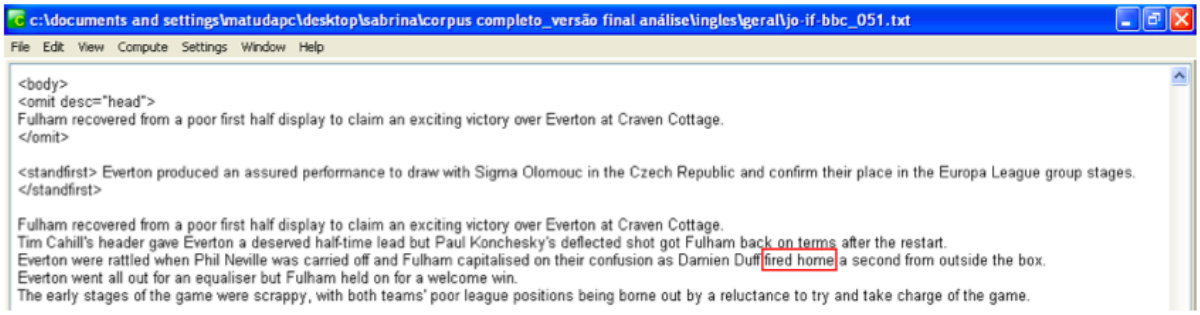


Figura 7: Parte do contexto expandido da linha de concordância 153 de *fire**

Após a validação das unidades fraseológicas, agrupamos as UFEs por categoria semântica tendo como base o *Roget`s International Thesaurus*. Optamos por esse tipo de

classificação porque, durante a análise, notamos que o significado de algumas UFEs é bastante similar e que o sentido que carregam está relacionado ao sentido do verbo na linguagem geral..

Apresentaremos, nos parágrafos que seguem, os equivalentes por categoria semântica. Embora intercambiáveis, todos apresentam características particulares que denotam categoria semântica, prosódia semântica e *lexical priming*[14] distintos, causando, dessa forma, reações diferentes no leitor.

9.1 Pontuação

As unidades fraseológicas ***SCORE ({a|the} goal)*** e ***NOTCH {his ORDINAL|CARDINAL}(goal)*** compõem a categoria semântica de pontuação. Ambas são utilizadas para pontuar os gols e seus significados recaem sobre o ato de adicionar gols ao placar de uma partida.

Os exemplos abaixo ilustram o uso de *notch* e *score* seguido do número de gols, enfatizando a ideia de soma de pontos:

> Rooney **notched his fifth goal** of the season and showed that United are serious contenders to retain their Premier League crown despite the loss of Cristiano Ronaldo during the summer.
>
> But Eduardo still had time **to score the sixth**, netting from close in after Andrey Arshavin's shot had come back off a post with two minutes to go.

9.2 Ação violenta

Durante o agrupamento semântico, identificamos 13 UFEs que transmitem a ideia de uma ação violenta. São elas:

a) FIRE {home | in (a goal) | (the ball) past [goalkeeper]| into the net} [15]
b) HAMMER {in | home}

[14] Teoria linguística desenvolvida por Michael Hoey (2005) que parte do princípio de que à medida que temos contato com uma nova unidade lexical, seja verbo, substantivo, adjetivo etc., essa nova unidade é adquirida juntamente com os contextos linguísticos, sociais, culturais em que frequentemente ocorre. Dessa forma, ao reproduzir a unidade lexical tendemos a utilizá-la nos mesmos contextos em que a encontramos. Toda a informação adquirida por meio da unidade lexical em questão é considerada o *lexical priming* da palavra.
[15] {} elenco de opções; | combinações possíveis; () uso opcional; [] hiperônimo.

c) SLAM {(the ball) past [goalkeeper] | in | into the roof of the net}
d) BLAST IN (a goal)
e) LASH home {(a goal) | into an empty net}
f) SMASH home
g) STAB the ball {home | past [goalkeeper]}
h) POKE{theballhome|homepast[goalkeeper]}
i) KNOCK {the ball {into the net | in | home}} / NUMBER goals
j) THUMP {home | past [goalkeeper]}
k) CLIP the ball into the net
l) SNATCH [a goal]
m) THUNDER a goal home

A alta ocorrência de verbos que transmitem a ideia de ação violenta em inglês nos levou à seguinte pergunta: será que o uso de verbos dessa categoria semântica também é comum em português?

Para responder a pergunta, geramos linhas de concordância para a etiqueta gramatical V[16], que indica um verbo, no *corpus* de português e observamos as ocorrências.

As concordâncias nos mostram que em português não utilizamos verbos de ação violenta com a mesma frequência que em inglês. Encontramos 222 ocorrências do verbo "empurrar", sendo 87 sinônimas de "fazer um gol". No entanto, em grande parte das ocorrências, o verbo "empurrar" estava relacionado à facilidade com que o gol foi feito, e não à intensidade do chute dado. Observemos o exemplo abaixo:

> Em desvantagem o Corinthians tentou pressionar, enquanto o Santos tentava concluir os bons contra-ataques que estava conseguindo. Mas aos 34 minutos Bill aproveitou bola rebatida na trave e **empurrou** para o gol, empatando o jogo no Pacaembu.

Também encontramos 6.000 ocorrências do verbo "bater", das quais apenas 28 utilizadas para descrever um gol. Ao analisar as linhas de concordância, constatamos que das 28 UFEs com o verbo "bater", somente 16 narram um gol feito. As outras representam tentativas que não deram certo:

> Jorge Wagner domina e **bate** para o gol. Mas a bola vai fraca e sem direção. MUCHA SALVA!! Robben faz bela jogada pela esquerda e cruza rasteiro. Mathijsen **bate** para o gol e bola explode no rosto de Mucha.

[16] As etiquetas gramaticais em português não fazem distinção entre os tempos verbais.

Após analisar cuidadosamente as linhas de concordância, expandindo o contexto e visitando o texto integral, concluímos que a diferença na frequência dos verbos de ação violenta é resultado da própria forma de jogar futebol no Brasil e na Inglaterra. Em outras palavras, se os *corpora* são realmente representativos, eles devem refletir, por meio do léxico, a forma como o futebol é jogado em diferentes culturas.

Se perguntarmos para qualquer pessoa que goste, acompanhe ou entenda, pelo menos um pouco, de futebol, sobre as diferenças entre o futebol brasileiro e o futebol inglês, as respostas serão muito parecidas: o futebol inglês é mais rápido e mais "pegado", mais "forte". Observemos o depoimentos de Ramirez, volante do Chelsea, em entrevista ao site da FIFA em outubro de 2010 sobre o futebol inglês:

> "O futebol inglês é muito físico e rápido. Aqui você não tem muito tempo para ficar com a bola no pé e tem que definir logo a jogada. Além de procurar as jogadas quando tenho a bola, eu sempre marquei forte. Por isso, não estou encontrando muitas dificuldades. Já estou me acostumando ao ritmo das partidas aqui.[17]"

Se tomarmos o depoimento acima, podemos caracterizar o futebol inglês como um futebol que tem a força como um de seus elementos centrais. Por esse motivo, tratamos a discrepância da frequência de verbos de ação violenta em inglês e português por meio do conceito forma-representação de Toledo (2002).

Retomando Toledo (2002), o conjunto de regras não determina nem influência as maneiras de jogar. Na verdade, a interpretação e apropriação cultural que cada região faz das regras é que revelam as "formas de jogo". Justapostas, as regras e as "formas de jogo" dão origem às representações, ou seja, o ajustamento da observação empírica das "formas de jogo" em um plano simbólico que, por sua vez, consolidam as "escolas", "jeitos" ou "estilos" próprios que se traduzem no código linguístico.

Investigar a apropriação cultural das regras no Brasil e na Inglaterra foge ao escopo deste artigo estudo. Contudo, reconhecer que essa interpretação influencia o "estilo" de jogo e, por conseguinte, a terminologia utilizada nas duas culturas é essencial para entendermos algumas discrepâncias linguísticas. Assim, o fato de o futebol inglês ser mais "pegado" transparece na grande quantidade de verbos de ação violenta.

---

[17] Disponível em: <http://www.guiame.com.br/v4/70368-1663-Destaque-no-site-da-Fifa-Ramires-diz- estar-adaptado-ao-futebol-ingl-s.html>. Acesso em: 12 abr. 2011.

Para comprovar nossa hipótese, compilamos um *mini-corpus* de, aproximadamente, 1000 palavras em cada língua, composto por 15 textos sobre o estilo do futebol brasileiro e 15 textos sobre o estilo do futebol inglês. Utilizamos *football, Brazil, style* e *school* como palavras de busca no *Google* para encontrar textos sobre o estilo do futebol brasileiro e *football, England, UK, school* e *style* para encontrar textos sore o estilo do futebol inglês. Esse *mini-corpus* é monolíngue, coletamos textos somente em inglês uma vez que o *corpus* foi compilado para comprovar a hipótese que seria apresentada em um congresso na Inglaterra.

A análise foi feita de forma bastante simples. Primeiramente, geramos listas de palavras dos *mini-corpora*. Em seguida, comparamos essas listas com listas de palavras de *corpora* de referência a fim de obter as palavras-chave. Como *corpus* de referência, utilizamos um *corpus* jornalístico, com aproximadamente de 100.000 palavras. Em seguida, analisamos as palavras-chave e selecionamos aquelas utilizadas para descrever as características dos dois futebóis. O quadro 2 mostra as *keywords* associadas às duas escolas:

Quadro 2: palavras-chave associadas ao estilo brasileiro e ao estilo inglês

Palavras-chave associadas ao estilo brasileiro	**Palavras-chave associadas ao estilo inglês**
- [good] dribblers - creative - spontaneous - improvisational - free-flowing - possession - samba [beat] - control [of the ball] - [highly] skilled - [flowing] passing [game] - beautiful - flair - invention	- physical [prowess/strength] - fast - athleticism - rough - [long] shoots - quick - [direct] passes - [no] risk - courageous - hard - [break-neck] speed

Ao fim da análise, nos questionamos se a frequência de verbos que denotam ação violenta tende a aumentar em português, uma vez que a prática e o aprendizado do futebol mudou bastante no Brasil. Antigamente as crianças jogavam futebol na rua, no pátio da escola e em "campinhos". Hoje em dia, as crianças frequentam, desde muito cedo, escolinhas de futebol, ambientes em que meninos e meninas são treinados, por meio de acompanhamento

tático, prática de jogadas e, muitas vezes, condicionamento físico. É comum ouvir de comentaristas e jornalistas esportivos que o "estilo" do futebol brasileiro, sempre caracterizado por jogadas de habilidade, está passando por um processo de transformação: a troca do "futebol arte" pelo "futebol força" ou, até mesmo, a junção desses dois futebóis. De qualquer forma, cremos que essas mudanças, futuramente, terão reflexo na constituição do léxico do futebol em português, incorporando neologismos. No entanto, embora as mudanças venham acontecendo há um tempo, seu reflexo no sistema linguístico, muito provavelmente, ocorrerá de forma mais lenta e, até o momento, o "futebol força" não é a "forma de jogo" que temos em nosso plano simbólico ao pensar no futebol brasileiro e, por esse motivo, não é traduzido em nosso código linguístico.

9.3 Facilidade

Outro campo semântico bastante comum em inglês, e que não foi encontrado em português, pelo menos não com frequência significativa, é o de "facilidade". Observemos as seguintes unidades fraseológicas

a) SLIDE {home his ORDINAL (goal) | the ball home};
b) SWEEP {home | the ball into the net};
c) DINK in [a goal];
d) TUCK home (a goal);
e) STROKE a shot past [goalkeeper].

Nelas, o sentido principal recai sobre a facilidade com que o gol foi feito. Na realidade, a facilidade não caracteriza o gol, mas sim a trajetória da bola, a posição do jogador ou o toque que o jogador dá na bola.

9.4 Velocidade

As UFEs ***TAP {in|home}*** e **FIZZ {the ball past goalkeeper|a sidefooter at goal}** compõem a categoria semântica de velocidade. Ambas enfatizam a velocidade do gol, não somente a velocidade com que o jogador recebe e chuta ao gol, mas também a rapidez dos toques até o momento do gol:

Scott Carson, oh Scott Carson. Drogba lines up the shot, **fizzes** a sidefooter at

> goal, but it's straight at the keeper. Who proceeds to let the ball squirm from his grasp, drop to Mikel who dinks it square for Malouda to **tap in**.

Novamente, se tomarmos os depoimentos dos jogadores, jornalistas e técnicos sobre o futebol inglês, veremos que a alta ocorrência desses verbos de velocidade nos periódicos ingleses reflete o "estilo" do futebol inglês. Retomemos o depoimento de Ramirez: "O futebol inglês é muito físico e rápido. Aqui você não tem muito tempo para ficar com a bola no pé e tem que definir logo a jogada...".

Sabemos que o português, assim como o inglês, utiliza palavras de outras categorias gramaticais para indicar a rapidez e velocidade de uma jogada como, por exemplo, os termos "rápido" e "de primeira". No entanto, chamamos atenção para o fato de que, pelo menos em nosso *corpus*, o jogo rápido se mostra tão característico no futebol inglês que existe uma necessidade de expressá-lo por meio de verbos específicos.

Algumas UFEs, não descrevem nenhuma particularidade do estilo do futebol inglês, mas refletem, de certa forma, uma característica da cultura e, consequentemente, da língua inglesa: um alto grau de detalhamento nas informações.

Para entender essa particularidade cultural, recorremos aos conceitos de *high-context culture* e *low-context culture* do antropólogo Edward Hall (1976). A teoria de Hall parte do pressuposto de que a quantidade de informação linguística e contextual necessária para transmitir o significado varia de acordo com a cultura e, para descrever as culturas, o autor cria dois grupos: *high-context culture*, formado por países como Brasil, Itália, Grécia e países da África, Asia e América Latina; e *low-context culture*, formado por países como Estados Unidos, Alemanha, Suíça e outros países da Europa ocidental e do Reino Unido.

As culturas que pertencem ao primeiro grupo, as *high-context cultures*, são caracterizadas por recorrerem a uma grande quantidade de elementos contextuais, os quais auxiliam seus integrantes a entenderem as regras, ou seja, o modo como a cultura funciona. Por esse motivo, parte da informação aparece de forma implícita, é o dito pelo não dito, informações que estão subentendidas na fala ou, em nosso caso, em um texto, e podem contribuir muito mais para a transmissão do significado do que aparentam. Essas culturas possuem um forte senso de comunidade, fator que implica em diferentes estilos comportamentais (MANCA, 2010, p. 373), que são dados como subentendidos e, raramente, aparecem de forma explícita.

As culturas que se enquadram no segundo grupo, *low-context*, são caracterizadas por

explicitar os elementos contextuais. Quase nada é subentendido, os assuntos são debatidos exaustivamente de modo que não restem dúvidas ou margem para uma interpretação equivocada. A comunicação é feita da forma mais explícita possível e é realizada por meio da transmissão de fatos, sem a expressão de sentimentos.

A teoria de Hall é bastante complexa e detalhada. Contudo, pode ser utilizada para entender o processo de produção de textos sobre futebol nas duas culturas.

9.5 Detalhamento

As UFEs que apresentaremos nos próximos parágrafos transmitem um detalhamento que não encontramos nos verbos e UFEs utilizados para descrever um gol em português. As UFEs do primeiro grupo expressam o movimento que a bola faz para dentro do gol:

> a) **SLOT {home([agoal)]|(theball)into{an|the}emptynet|past[goalkeeper]| in}**
> Robinho, who was the game's outstanding player, then provided one of the assists of the championship so far as he threaded a delightful ball to Elano in the 72nd minute before the midfielder **slotted home** a second.
>
> b) **DRILL home**
> The defender, 25 yesterday, settled a tight game and embarked on a celebration that would have looked more at home on Strictly Come Dancing than Match of the Day. He **drilled home** Swansea's 69th-minute winner after Leon Britton had rattled the home side's bar with a stunning overhead kick.
>
> c) **BURY the ball past [goalkeeper]**
> For good measure, Fabregas popped in a peach for his second and Arsenal's fifth as he ran up field, picked his spot and **buried the ball past** Howard.

No caso desses verbos, só entendemos o funcionamento das UFEs, que utilizam *drill, bury* e *slot* de forma figurativa, pelos seus significados na língua geral.

9.6 Realização

As UFEs ***FINISH (off)*** e ***DELIVER a goal*** expressam o sentido de 'completar', 'realizar uma tarefa', assim como o verbo "finalizar" em português:

> **a)** And 43 minutes into his full debut, Stanislas **finished off** a great end-to-end move by tapping in Luis Boa Morte's cross.

b) It's been 25 years since a match between these two teams **delivered a goal** and it doesn't look like we're going to get one tonight.

c) O goleiro Felipe nem se mexeu. Novamente impaciente com o atacante Souza, a torcida do Corinthians passou a pedir a entrada de Dentinho. Só parou quando Elias, aos 42, **finalizou** de fora da área no ângulo: 2 a 1.

9.7 Movimento da bola

Também detectamos a presença de duas UFEs que descrevem o movimento da bola ou seu controle por um jogador:

a) **HOOK in ([a goal])**
When defender Marius Zaliukas **hooked in** the second shortly after the break, belief they were capable of a sensational result coursed through the home ranks.

b) **STEER the ball past [goalkeeper]**
His pass picks out the run of Osman who **steers the ball past** Schwarzer.

Analisamos as ocorrências de *HOOK in ([a goal])* no *corpus* e chegamos a duas possibilidades do uso. Primeiramente, pensamos que a UFE poderia descrever o tipo de jogada realizada para fazer o gol, já que *hook in* descreve uma jogada feita com o pé no ar, a meia altura, conhecida como "voleio" em português, fato que foi confirmado após a observação de outras ocorrências no Google.uk. Nossa segunda hipótese era que a UFE seria utilizada como o termo *hook* de uso corrente no *cricket* e no golfe, cujo significado indica um modo de bater na bola que faz com que ela desenhe uma curva em vez de ir reto. Confirmar essa segunda hipótese foi um pouco mais complicado do que a primeira, pois se o verbo *hook* designa, em seu significado, o movimento da bola, o texto, muito provavelmente, não usa nenhum outro recurso linguístico que possamos analisar para validar nossa hipótese. O detalhamento das notícias em inglês foi de grande valia para confirmar nossa ideia. Acessamos o site Football.co.uk, do qual um dos textos que contém a UFE foi coletado, vimos as fotos dos gols do jogo e clicamos em um link do YouTube que direcionava para um vídeo que mostrava os gols da partida.

A última UFE, ***STEER the ball past [goalkeeper]***, não descreve o movimento da bola, como a anterior. Seu significado recai sobre o controle da direção que a bola toma, como no exemplo em que o jogador direciona a bola e chuta para o gol:

His pass picks out the run of Osman who steers the ball past Schwarzer.

9.8 Gols de cabeça

Identificamos três UFEs utilizadas para narrar gols de cabeça:

a) **NOD {home (into an empty goal) | in | into the net};**

b) **HEAD (the ball) {home|past (goalkeeper)|into the (roof of the) net};**

c) **{HAMMER | POUND} a header into the net.**

A alta frequência dos verbos que narram gols de cabeça em inglês nos chamou a atenção quando comparada à frequência de "cabecear", "bater de cabeça" e "tentar de cabeça", 1037 e 387 ocorrências em cada língua, respectivamente.

Recorremos novamente ao conceito forma-representação de Toledo para entender essa discrepância. Tomemos o depoimento de Robinho, em sua apresentação ao Milan, sobre o futebol inglês:

> "Não tive nenhum problema com o Roberto Mancini. O que acontece é que o futebol inglês não é muito bom para jogador brasileiro, muita bola alta e a gente gosta de jogar com ela no chão[18]."

Jogadas aéreas, principalmente de bola na área, definem um estilo bem inglês, uma vez que exigem a força, a impulsão e a altura características dos jogadores ingleses. Desse estilo, e também das condições climáticas da Inglaterra -- devido à chuva constante as equipes eram forçadas a mandar a bola para o alto, já que era quase impossível fazê-la correr na grama encharcada -- surgiu a expressão "chuveirinho", para designar a grande quantidade de jogadas aéreas na área, que caracterizou, por muitos anos, a escola inglesa.

O cabeceio, que definia o "estilo" inglês, era considerado como a melhor forma de atacar e de concluir uma jogada, levando a seleção da Inglaterra à conquista da Copa do Mundo de 1966, mas perdendo sua força nos anos seguintes.

Não nos cabe, aqui, discutir a origem e a eficácia do "estilo" inglês, mas sim reconhecer que o papel das jogadas aéreas no futebol inglês é fundamental para entender a

[18] Disponível em: http://opiodopovo.wordpress.com.tab/chuveirinho/ Acesso em: 21 mai. 2011.

divergência entre a frequência de UFEs que narram gols de cabeça em inglês e em português.

9.9 Adição

A estrutura de ***ADD {a|the} ORDINAL (goal)*** pode ser enquadrada na categoria semântica de adição, pois a referência ao número do gol que foi marcado é parte integrante da UFE.

> a) Rumblings of discontent were just beginning to emerge as Arsenal took the lead through Abou Diaby in the 18th minute and the midfielder **added a second** three minutes later, prompting Wenger to compare his midfielder with a Highbury great.

9.10 Empate

Identificamos os verbos ***EQUALISE, DRAW*** e ***TIE*** como equivalentes de EMPATAR. Entretanto, cada um apresenta suas particularidades. O verbo *EQUALISE* é o mais comum para narrar um gol de empate. Os trechos que seguem exemplificam seus principais usos:

> Martinez`s lads were 13 minutes from victory until Clint Dempsey **equalised** with a diving header and Zamora had the last word with his bonce on 81 minutes in this gripping fifth-round replay.

Os verbos ***TIE*** e ***DRAW***, outros possíveis equivalentes do verbo "empatar" em português, também ocorrem no *corpus*. Entretanto, o primeiro ocorre majoritariamente em jornais e revistas americanos e o segundo ocorre para narrar um jogo que empatou, e não um gol de empate.

> The Cottagers beat Arsenal 1-0 in August, **drew** 2-2 with Chelsea over Christmas and stunned United 2-0 last month.

9.11 Posicionamento

A UFE ***NESTLE {in | into} the (back of the) net*** nos dá a idéia de acomodar a bola no gol:

> Not a brilliant penalty-struck low and only a few feet to the left of the centre of the goal, but Julio Cesar went the other way and the ball **nestled in the back of the net**.

Como podemos observar, o exemplo ilustra a posição da bola após o gol.

9.12 Realização

Identificamos três UFEs que pertencem à categoria de realização, entendida aqui como consecução, ou seja, o ato de alcançar um objetivo. Observemos os exemplos:

> a) **NET [a goal]**
> Liverpool's season is firmly back on track after Yossi Benayoun **netted** a hat-trick in Saturday's 4-0 Premier League home win over Burnley. Substitute Stefan Maierhofer **netted** a late consolation for Wolves.
>
> b) **BAG [a goal]**
> Dave Kitson is finally beginning to find his feet at Stoke as he bagged the winner in a 1-0 victory over Sunderland at the Britannia Stadium. Drogba was a real menace throughout, heading just over from a Jose Bosingwa cross before **bagging** the all-important goal.
>
> c) **NICK [a goal]**
> Birmingham City had the early Premier League leaders begging for the final whistle before Aaron Lennon **nicked** a winner for Tottenham deep into stoppage time. Lions fans were already leaving the Den when Paul Shaw **nicked** a consolation goal on 79 minutes but the roar that went up when Tim Cahill (right) made it 3-2 within seconds brought many back.

Independentemente do significado de seus verbos na linguagem geral, as três UFEs descrevem um ato de realização. Quando utilizadas, enfatizam a conquista de um gol que é necessário ou importante. Se tomarmos como exemplos os colocados da UFE nos exemplos acima, *hat-trick* (três gols marcados por um jogador na mesma partida), *consolation* (gol de honra), *the winner* (gol da vitória) e *the-all-important goal* (gol importante, que pode ser decisivo para a vitória ou para a permanência de um time em um campeonato), veremos que além de descrever um gol feito, a UFE realça a importância do gol, tanto pelos seus colocados como pelo seu uso em detrimento de algum outro equivalente.

10. Considerações Finais

Ao final de nosso estudo de caso, verificamos que a LC é fundamental para expandir o escopo de pesquisas terminológicas e contribui, de forma significativa, para a identificação de aspectos culturais de uma área de especialidade.

A existência de UFEs que expressam o sentido de violência, explosão, rapidez, facilidade e gols de cabeça em inglês e a ausência do mesmo tipo de UFE em português pode ser explicada pelo conceito forma-representação, do mesmo modo que a maior variação de UFEs que descrevem a trajetória da bola, tanto em um gol quanto em uma tentativa, pode ser explicada pelo conceito de *high* e *low-context culture*. O torcedor brasileiro *(high-context)* espera um relato mais breve da partida e ênfase no produto final, o gol. Já o torcedor inglês *(low-context)* espera que os lances sejam narrados de forma precisa, com ênfase nos meios, ou seja, nas jogadas, que levaram ao gol.

Dessa forma, acreditamos que a Terminologia não é uma atividade prescritiva, na qual os termos devem ser normatizados a fim de garantir a eficácia de uma comunicação especializada. Ao contrário, os fatores culturais, o contexto e a situação e a finalidade de uso influenciam de forma direta o funcionamento das terminologias; por esse motivo, o fazer terminológico, principalmente o bilíngue, deve considerar todos esses elementos na compilação de obras terminográficas.

Referências bibliográficas

AIJMER, K.; BENGT, A. **English Corpus Linguistics**. New York: Longman, 1991.

AZENHA JR., J. **Tradução técnica e condicionantes culturais**: primeiros passos para um estudo integrado. São Paulo: Humanitas/FFLCH-USP, 1999.

BERBER SARDINHA, A. Lingüística de Corpus: Histórico e problemática. **Revista D.E.L.T.A.**, São Paulo, v. 16, n. 2, p. 323-367, 2000.

_____. **Lingüística de Corpus**. Barueri: Manole, 2004.

BOWKER, L.; PEARSON, J. **Working with Specialized Language.** A Practical Guide to Using Corpora. London/New York: Routledge, 2002. crossref http://dx.doi.org/10.4324/9780203469255

CRUZ, A. H. O. **A Nova Economia do Futebol**: Uma análise do processo de modernização de alguns estádios brasileiros. Dissertação (Mestrado em Antropologia Social) Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2005.

DAMATTA, R. (Org.). Esporte na sociedade: Um ensaio sobre o futebol brasileiro. IN: DAMATTA, R. **Universo do Futebol**. Rio de Janeiro: Pinakotheke, 1982.

GIGLIO, S. S.; SPAGGIARI, E. A produção das ciências humanas sobre futebol no Brasil: um panorama (19980 – 2009). **Revista de História** (USP), v. 163, p. 293-350, 2010. crossref http://dx.doi.org/10.11606/issn.2316-9141.v0i163p293-350

GIL, G. A copa da cultura no futebol. **Jornal O Globo**, ed. 5 set. 2004. <http://www.suapesquisa.com/futebol/> Acesso em 16 nov. 2008.

KRIEGER, M. da G.; FINATTO, M. J. B. **Introdução à Terminologia**. Teoria & prática. São Paulo: Contexto, 2004.

LEONINCE, M. P. ; SILVA, M. T. Entendendo o futebol como negócio: um estudo exploratório. **Gestão e Produção**, v.12, n.1, p.11-23, jan./abr. 2005. crossref http://dx.doi.org/10.1590/S0104-530X2005000100003

MANCA**,** E. From phraseology to culture: Qualifying adjectives in the language of tourism. In: RÖMER, U.; RAINER, S. **Patterns, Meaningful Units and Specialized Discourses**, 105–122, 2010.

MCENERY, T and HARDIE, A. **Corpus Linguistics**: Method, Theory and Practice. Cambridge: Cambridge University Press, 2012.

ROCHA, M. A. E. O uso de corpora computadorizados no ensino de língua portuguesa: metodologia e avaliação. In. GRIMM CABRAL, L. et all (orgs). **Lingüística e ensino**: novas tecnologias, Blumenau: Nova Letra, p.137-55, 2001.

TAGNIN, S. E. O. A identificação de equivalentes tradutórios em corpora comparáveis. In: **I Congresso Internacional da ABRAPUI**, Minas Gerais, 2007.

TOGNINI-BONELLI, E. **Corpus Linguistics at Work.** Amsterdam/Philadelphia: John Benjamins, 2001. crossref http://dx.doi.org/10.1075/scl.6

TOLEDO, L. H. de. **Lógicas no Futebol**. São Paulo: Hucitec/Fapesp, 2002.

Artigo recebido em: 15.10.2014
Artigo aprovado em: 30.11.2014

A música e os ruídos na legendagem francesa para surdos e ensurdecidos

Songs and noises in the French subtitling for the deaf and hard of hearing

Ana Katarinna Pessoa do Nascimento*
Stella E. O. Tagnin**

RESUMO: Além das imagens, o áudio exerce um papel fundamental na criação do significado do enredo de um filme. O universo sonoro de um filme é composto por três elementos básicos: a fala, a música e os ruídos. Sem o auxílio de legendas que traduzam também os efeitos sonoros para além da fala, o espectador surdo ou ensurdecido não tem acesso a esses aspectos das produções audiovisuais. Por isso, a legendagem para surdos e ensurdecidos (LSE) precisa indicar o falante e os efeitos sonoros. A acessibilidade audiovisual tem sido discutida na França desde 1986 através da lei sobre a liberdade de comunicação. Tendo em vista a tradição francesa na LSE, esta pesquisa buscou analisar a tradução dos efeitos sonoros de três filmes franceses comercializados em DVD: *Nos jours heureux* (2006), *Les femmes du sixième étage* (2010) e *L'écume des jours* (2013). Para isso, as músicas e ruídos desses filmes foram anotados com etiquetas discursivas nas seguintes categorias: música de fosso, música de tela, música qualificada, música não qualificada, sons causados pelo homem, sons causados por objetos, sons da natureza, sons de animais, sons ficcionais e silêncio. Os arquivos anotados foram analisados pela ferramenta *Concord* do *WordSmith Tools 5.0*. Os dados revelaram que os efeitos sonoros na LSE francesa nos filmes estudados foram traduzidos levando em conta a função de cada som dentro do filme, o que produz uma tradução de maior qualidade. A observação dessas funções pode ajudar os legendistas aprendizes na produção de legendas que sejam de fácil compreensão pelo público

ABSTRACT: In addition to images, audio plays a key role in creating meaning in a movie plot. The sound universe of a movie consists of three basic elements: speech, music, and noise. Without the aid of subtitles that also translate sound effects besides speech, deaf or hard-of-hearing audiences do not have access to those features on audiovisual productions. Therefore, the subtitling for the deaf and hard of hearing (SDH) must inform the speaker and the sound effects. Audiovisual accessibility has been discussed in France since 1986 through the 'Freedom of Communication' Bill. Given the French tradition in SDH, this study sought to examine the translation of sound effects in three French films on DVD: *Nos jours heureux* (2006), *Les femmes du sixième étage* (2010), and *L'écume des jours* (2013). For the analysis, their sound effects were annotated with discursive tags for the following categories: gap music, screen music, qualified music, non-qualified music, sounds made by men, sounds made by objects, nature sounds, sounds made by animals, and fictional sounds and silence. The annotated files were analyzed with the Concord tool on WordSmith Tools 5.0. The data revealed that the sound effects in the French SDH on the studied films were translated taking into account the function of each sound within the movie, which means a higher quality translation. Not leaving these functions aside may help subtitles-in-training to produce easily understandable subtitles for the deaf and hard-of-hearing audience.

* Doutoranda do Programa de Pós-Graduação em Estudos da Tradução da Universidade de São Paulo.
** Professora Associada da Universidade de São Paulo.

surdo e ensurdecido.

PALAVRAS-CHAVE: Tradução Audiovisual. Legendagem para surdos e ensurdecidos. Música e ruídos. Linguística de *Corpus*.

KEYWORDS: Audiovisual Translation. Subtitling for the deaf and hard of hearing. Songs and noises. Corpus Linguistics.

1. Introdução

As normas que tratam da acessibilidade audiovisual no Brasil ainda são muito incipientes. Em 2006, o Ministério das Comunicações publicou a Portaria 310, que torna oficial a Norma Complementar nº 1/2006. Essa Norma regula a implantação da grade de TV aberta para pessoas com surdos e ensurdecidos, tendo estabelecido os requisitos técnicos para torná-la acessível a esse público através da chamada legendagem para surdos e ensurdecidos ou LSE, que, nesta situação, é sempre intralinguística. Em cumprimento à Norma Complementar, as emissoras de televisão utilizam, para a LSE, o sistema de legenda fechada norte-americano, o *closed caption,* que é acessado apenas por quem possui um televisor capaz de acionar o sistema. Quanto ao cinema, os surdos e ensurdecidos brasileiros só podem assistir às produções estrangeiras por meio de uma legendagem que não lhes é a mais adequada (a legendagem para ouvintes, que não possui identificador dos efeitos sonoros para além das falas; nesse caso, o que eles necessitam é de uma LSE interlinguística). No tocante às produções nacionais, essas lhes ficam quase sempre alheias, pela ausência ainda muito frequente de LSE intralinguística. A luta para mudar essa realidade é bastante antiga, mas apenas em 2012, a partir de uma ação civil pública da Procuradoria da República do Estado de São Paulo, passou a haver obrigatoriedade do uso de legendas em língua portuguesa nos filmes brasileiros produzidos com recursos e financiamentos públicos. No entanto, essas regulamentações têm em vista as legendas confeccionadas para televisão e cinema, ficando os DVDs de filmes nacionais à mercê das produtoras e distribuidoras.

Na França, porém, a realidade é diferente. Já em 1986, a lei sobre a liberdade de comunicação, previa a LSE na grade de programação das emissoras de TV como meio de acessibilidade surdos e ensurdecidos. Posteriormente, em 2005, a lei pela igualdade de direitos e oportunidades, participação e cidadania das pessoas com deficiência, previa a ampliação de programas de televisão acessível a pessoas surdas ou ensurdecidas, com exceção de publicidade, *trailers*, vídeos sob demanda (VOD), entre outros. Com a lei de 2005, as emissoras precisam apresentar uma porção considerável da programação de forma acessível.

Essa porção depende da audiência da emissora: quanto maior a audiência, maior o percentual de programação legendada. Em relação aos DVDs, estima-se que, em 2007, 7% dos DVDs de filmes franceses apresentavam LSE.

Tendo em vista a já longa preocupação francesa com a acessibilidade em narrativas audiovisuais (programas televisivos e filmes), este trabalho tem como objetivo descrever a tradução de músicas e ruídos na LSE de filmes franceses em DVD (*Nos jours heureux*, 2006; *Les femmes du sixième étage*, 2009; *L'écume des jours*, 2013) para examinar se leva em conta a função de cada som na trama fílmica. Para atingir o objetivo, utilizou-se a metodologia aplicada em Nascimento (2013). Nele, as músicas e os ruídos dos filmes foram anotados com etiquetas discursivas denotativas das seguintes categorias: música de fosso, música de tela, música qualificada, música não qualificada, sons causados pelo homem, sons causados por objetos, sons da natureza, sons de animais, sons ficcionais e silêncio. Os arquivos de legendas etiquetadas foram, em seguida, rodados no *WordSmith Tools 5.0* e as legendas por ele analisadas. Os resultados obtidos mostraram que a legendagem da trilha sonora somente pode contribuir para a reconstrução do significado da trama dos filmes quando o legendista leva em consideração a função de cada som legendado (NASCIMENTO, 2013). A maioria das músicas e dos ruídos encontrados na LSE nos três filmes brasileiros estudados (‘Irmãos de fé’, 2005; ‘O Signo da Cidade’, 2008; ‘Nosso Lar’, 2010) foram aparentemente traduzidos sem preocupação de explicitar a relação entre o som e sua significação no enredo, pois a tradução não foi feita levando em consideração o papel da música e dos ruídos em todo o filme, mas apenas em cenas isoladas, o que pode ocasionar perda na compreensão da trama do filme como um todo (NASCIMENTO, 2013).

Este artigo, para além desta seção introdutória, tem ainda tantas outras seções. Na Seção 2, são apresentadas as principais características da LSE e um breve apanhado sobre o som no cinema. A seção 3, trata dos procedimentos metodológicos para a realização da pesquisa, descrevendo o *corpus* e os procedimentos metodológicos. A quarta seção é dedicada às análises das traduções das músicas e ruídos do *corpus*. Por fim, na quinta parte, são apresentadas as considerações finais com pontos conclusivos pertinentes ao tema da pesquisa e considerações acerca de futuras pesquisas.

2. Pressupostos teóricos

2.1 Legendagem para surdos e ensurdecidos (LSE)

A partir do advento de falas no cinema, surgiu a necessidade de legendas para aqueles que não ouvem, pois, até então, com o cinema mudo, surdos, ensurdecidos e ouvintes tinham o mesmo acesso às produções fílmicas. Com a introdução da fala – normalmente em forma de diálogos – e demais efeitos sonoros, o público surdo ficava impossibilitado de assistir a filmes independentemente (DE LINDE; KAY, 1999). Alguns produtores buscaram inserir intertítulos[1] durante toda a extensão do filme em busca de torná-los acessíveis, mas o procedimento, além de prolongar os filmes, era bastante caro. Outro problema era a dificuldade do público alvo em identificar o falante e em compreender os *inputs* do filme proporcionados pelas músicas e ruídos (CHAUME, 2004). Essa situação começou a mudar com o advento da televisão e da técnica da legenda fechada (*closed caption*), que passou a permitir o acesso à LSE, através do controle remoto, por parte de quem dela necessitasse (FRANCO; ARAÚJO, 2003).

Quanto à forma de aparição em tela, a legendagem pode ser de dois tipos: *roll-up* e *pop-on*. Na legendagem *roll-up*, as palavras são digitadas da esquerda para a direita e deslizam de baixo para cima na parte inferior da tela da televisão; na legendagem *pop-on*, cada legenda aparece e desaparece da tela em sincronismo com fala e imagem. Enquanto nas televisões brasileiras as duas formas são utilizadas, sendo que a *roll-up* é mais comum em programas ao vivo e a *pop-on* é normalmente utilizada em programas pré-gravados (FRANCO; ARAÚJO, 2003), na França, utiliza-se predominantemente a *pop-on,* sendo a *roll-up* mais popular em países anglófonos e no Canadá (DIU, 2009).

A condensação, a marcação e a segmentação são os principais fatores para uma legendagem de qualidade (DÍAZ CINTAS; REMAEL, 2007). A condensação ocorre quando é preciso reduzir o texto oral, de partida, para que as legendas não fiquem rápidas demais, impedindo o espectador de lê-las confortavelmente, ~~e~~ nem sejam demasiadamente demoradas, fazendo com que o espectador leia uma mesma legenda repetidamente (AUBERT; MARTI, [s.d]). A marcação determina os tempos de entrada e saída da legenda na tela, e a segmentação fragmenta o texto oral em duas ou mais legendas e/ou uma legenda em mais de uma linha. Essa segmentação pode ser feita seguindo três critérios: o visual, o retórico e o

[1] Textos em cartazes utilizados nos filmes mudos para comentar a ação ou inserir diálogos. As imagens dos cartazes intercalavam-se com as cenas do filme.

linguístico (REID, 1996). O visual diz respeito a mudanças de tomada no filme ou programa: a cada nova cena, uma nova legenda deverá ser inserida. O critério retórico é aquele pelo qual o fluxo da fala deve ser seguido, ou seja, quando houver pausas no discurso, deve-se inserir uma nova legenda. E por fim, o critério linguístico é aquele determinado pelas regras da sintaxe para que seja facilitada a leitura da legenda por parte do espectador, devendo o legendista evitar separar determinante e determinado, sujeito e predicado etc. bem como quebrar sintagmas de qualquer tipo.

Os parâmetros descritos acima são característicos de qualquer tipo de legendagem. A LSE, no entanto, exige a presença de dois parâmetros exclusivos: a identificação do falante e os efeitos sonoros. A identificação dos falantes é importante, pois muitas vezes os surdos não conseguem inferir a troca de turno de fala somente pela imagem. Além disso, a presença de dois ou mais personagens ao mesmo tempo em cena pode dificultar a identificação de quem está falando. Nesses casos, a compreensão da narrativa audiovisual pelos surdos e ensurdecidos pode ficar comprometida (ARAÚJO, 2008; NASCIMENTO; ARAÚJO, 2011).

A identificação de falantes no Brasil e na França se dá de maneiras diferentes. Enquanto no Brasil o falante é identificado através do nome do personagem entre colchetes, na França, além de a identificação ser posicionada sob o falante, a legenda muda de cor para diferentes personagens e situações.

A atenção à tradução dos efeitos sonoros de uma narrativa audiovisual faz-se necessária na medida em que os componentes acústicos não verbais colaboram para a construção de sentido, pois, sem eles, a construção da narrativa audiovisual perde um dos seus elementos significadores (NEVES, 2005). Não há regras explícitas para esse tipo de tradução em nenhum dos dois países; na França, porém, segundo Diu (2009, p. 18), "os tradutores buscam a maior precisão ao traduzir ruídos e música, especialmente porque esses participam de fato da trama". Contudo, a autora alerta, ainda, que a legendagem de músicas muitas vezes se dá através da indicação do estilo musical. No Brasil, segundo Nascimento (2013), não há nenhuma convenção de como os efeitos sonoros devem ser legendados; e por isso, as traduções são feitas sem preocupação com a trama, o que pode acarretar dificuldade de compreensão por parte do espectador surdo e ensurdecido. Também não há consenso no formato na apresentação das legendas de ruídos e de música, mas no Brasil a maioria delas são postas entre colchetes, enquanto na França, não apresenta nenhuma característica específica.

2.2 O som no cinema

O som no cinema causa efeitos que, muitas vezes, passam despercebidos. Mas é ele o responsável por alterar e direcionar o modo de se receber as imagens (CHION, 2008). Som e imagem devem interagir com o público, sem que esse possa diferenciar os dois elementos (HUNTER, 2008). Para uma cena ser considerada verossímil, o espectador deve ouvir ruídos provenientes dos inúmeros objetos percebidos dentro daquela cena (JULLIER, 2006), já que a ausência total de efeitos sonoros pode acarretar expectativa acerca do silêncio, que deve sempre ser significativo nas cenas em que ocorre (JULLIER, 2006).

Chion (2008) chama a atenção para o conceito de acusmática, que é definido como "[ruído] que ouvimos sem ver a causa". Segundo o autor, no decorrer de um filme, o [ruído] pode aparecer de duas formas: 1) quando há a visualização de um objeto e, só em seguida, seu respectivo ruído é ouvido; 2) quando o ruído produzido por dado objeto é ouvido inicialmente e ele é apresentado posteriormente. Essa última forma é muito utilizada em filmes de mistério para causar suspense sobre algum aspecto de determinado objeto/fenômeno/personagem, criando uma expectativa em torno dele.

Outro conceito importante é o de "ruídos de ambientação" (JULLIER, 2006). Eles são utilizados para tornar os ambientes mais reais, como já referido acima, mas podem, também, informar aos espectadores onde exatamente determinada cena está se passando (HUNTER, 2008). Portanto, os ruídos no cinema não passam despercebidos pelo espectador, pois acrescentam informações.

O termo 'trilha sonora' representa todo o conjunto sonoro de um filme, ou seja, as falas, geralmente em forma de diálogos, músicas e ruídos (BERCHAMPS, 2006). Quando uma composição é criada para um filme, é chamada de música original do filme. A música original pode criar temas específicos e próprios à personalidade dos personagens e à mensagem do filme, muitas vezes tornando-se algo tão conhecido quanto o próprio enredo, ~~e~~ caracterizando-o.

A música de um filme pode ser dividida em duas categorias: música de fosso e música de tela. A música de fosso[2] é afastada do local, do tempo e da ação em cena e apenas os espectadores podem escutá-la; já a música de tela é aquela que provém do local da ação, e pode ser escutada também pelos personagens (CHION, 2008). Essa distinção é bastante

[2] Música de fosso é um termo inspirado nas orquestras que ficavam localizadas em um fosso sob o palco nas óperas e cinemas antigos.

fluida, podendo a música de fosso passar a ser de tela e vice-versa (CHION, 2008). Neste trabalho, a música foi classificada como qualificada quando apresentou adjetivos ou outros elementos que indicassem sua função na trama, como, por exemplo, *Musique Flamenco*, e não qualificada quando não definiu sua função dentro do enredo (*Musique à la radio)*.

As *cues* são inserções de música no filme e são utilizadas para salientar passagens importantes, interligar cenas, e rotular acontecimentos. Bordwell (2008) entende que a escolha e a combinação dos sons podem criar padrões dentro da obra fílmica. O ritmo, a melodia e harmonia das músicas afetam a reação emocional do espectador. Dessa forma, uma melodia ou trecho de música pode ser associado com determinado personagem, situação ou ideia, criando *motifs* musicais. O uso da música pode ajudar a montar uma história cujo enredo acompanhe inúmeros personagens e localizações, permitindo criar uma unidade para o filme e garantir a fluidez da narrativa.

3. Metodologia

3.1 *Corpus*

O *corpus* utilizado no presente trabalho é do tipo especializado, pois a legenda no âmbito da LSE de filmes em DVD é o único gênero textual que o compõe. Além disso, é de língua nativa, com textos escritos em francês europeu. Em termos de extensão, é composto por tantas legendas que contêm tantas palavras.

A escolha dos três filmes franceses de onde a LSE foi extraída –*Nos jours heureux* (*NJH*) (NAKACHE, 2006), *Les femmes du 6ème étage* (*F6E*) (LE GUAY, 2010) e *L'écume des jours* (*LDJ*) (GONDRY, 2013), como já mencionados – se deu por terem sido sucesso de bilheteria na França (estando entre os filmes franceses mais assistidos dos últimos 50 anos segundo o site L'internaute[3]), e pertencerem a gêneros similares. Além disso, englobam uma passagem de tempo abrangente, que poderia revelar modificações temporais nas escolhas de tradução. Os dois primeiros são classificados como comédia e o terceiro como comédia dramática pelo *site* IMDb[4].

[3] http://www.linternaute.com/cinema/business/films-francais-les-plus-vus-des-50-dernieres-annees/
[4] http://www.imdb.com/

3.2 Procedimentos

O processo de análise de traduções de efeitos sonoros em LSE por meio da Linguística de *Corpus* pressupõe que os textos compilados estejam em formato eletrônico. As legendas dos filmes a serem analisados foram retiradas dos seus respectivos DVDs através do *software SubRip* 1.50. Este software reconhece apenas arquivos em extensão VOB, ou seja, no formato de DVD. Após escolhido o arquivo, o programa dá início à extração das legendas. Algumas vezes alguns caracteres não são reconhecidos pelo programa e devem ser inseridos manualmente pelo operador do *software*, daí em diante o programa passa a sempre reconhecer determinado caractere pelo que foi inserido manualmente. O processo todo de extração de legendas dura apenas alguns minutos.

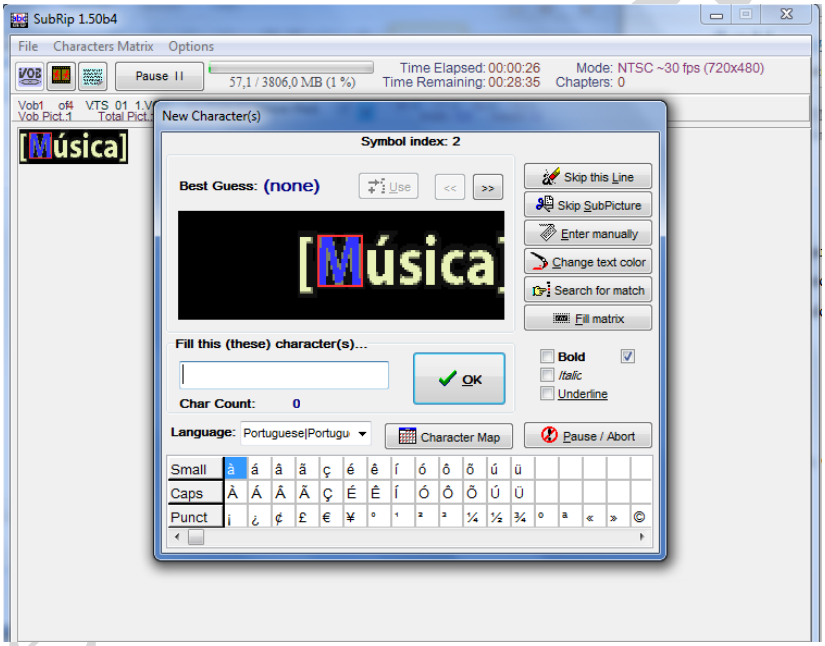


Figura 1: Interface do extrator de legendas *SubRip*.
Fonte: as autoras.

Os arquivos de legenda extraídos possuem extensão .srt; porém, o *software WordSmith Tools* (*WST*) lê apenas arquivos cuja extensão é .txt. Para transformá-los, bastou abrir os arquivos e salvá-los com a nova extensão no *software* Bloco de Notas. As legendas foram salvas em três arquivos diferentes: NJH, F6E e LDJ. Posteriormente a esse procedimento, deu-se início à anotação do *corpus*, o que foi feito com as categorias propostas por Nascimento (2013) e representadas pelas etiquetas também propostas pela mesma autora. As categorias e respectivas etiquetas são apresentadas no Quadro 1.

Quadro 1 – Categorias e respectivas etiquetas.

Categoria	**Etiqueta**
música de fosso	<mus_fosso>
música de tela	<mus_tela>
música qualificada	<mus_qualif>
música não qualificada	<mus_nqualif>
sons causados pelo homem	<som_hom>
sons causados por objetos	<som_obj>
sons da natureza	<som_nat>
sons de animais	<som_anim>
sons ficcionais	<som_ficc>
Silêncio	<sil>

Fonte: as autoras.

É importante ressaltar algumas considerações sobre as etiquetas: 1) as legendas relativas às músicas recebiam dupla etiquetação: Fosso ou tela se combinava com qualificada ou não qualificada. Por exemplo, a legenda *Musique douce* recebeu as etiquetas <mus_fosso> e <mus_qualif>; 2) As etiquetas para som de animais e silêncio, apesar de previstas, não foram utilizadas, já que não houve legendas que correspondessem a elas no *corpus* de estudo; 3) As etiquetas para sons ficcionais são aplicadas aos sons que não poderiam ser produzidos no mundo real. Os filmes foram assistidos em computador e, a cada legenda de música ou ruído, o filme era parado, a legenda era classificada segundo a categoria pertinente e era inserida a respectiva etiqueta no arquivo .txt. Esse processo de anotação foi manual. Um trecho do *corpus* etiquetado, pronto para análise pelo *WST*, é apresentado na Figura 2:

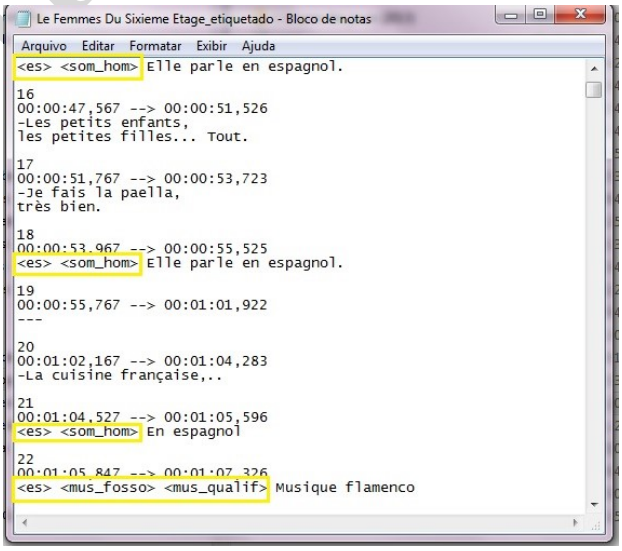


Figura 2 – Trecho do *corpus* etiquetado.
Fonte: as autoras.

Na Figura 2, há três etiquetas de sons produzidos pelo homem <som_hom>, como o próprio nome sugere, essa etiqueta é utilizada quando o ruído produzido pelo homem é não verbal, ou não foi transcrito para a legenda. No exemplo acima, os personagens falam em espanhol; isso foi indicado pela legenda, mas não transcrito. Já as etiquetas música de fosso <mus_fosso> e música qualificada <mus_qualif>, foram utilizadas na legenda Musique Flamenco, pois é música de fosso, ou seja, que apenas os espectadores podem ouvir e foi qualificada, uma vez que há o adjetivo Flamenco que a caracteriza na trama.

Para auxiliar a busca das etiquetas por arquivo/filme pela ferramenta *Concord* do *WST*, as linhas de concordância foram reorganizadas por ordem alfabética da primeira palavra à direita das etiquetas, que é o nódulo de busca (Figura 3). Dessa forma, as legendas iguais apareceram em sequência, tendo sido possível, assim, analisar qualitativamente a maneira pela qual os efeitos sonoros foram traduzidos na LSE dos filmes, se não considerando sua função nas tramas tal como foi feito na LSE dos filmes 'Irmãos de fé', 'O signo da cidade' e 'Nosso lar' (NASCIMENTO, 2013).

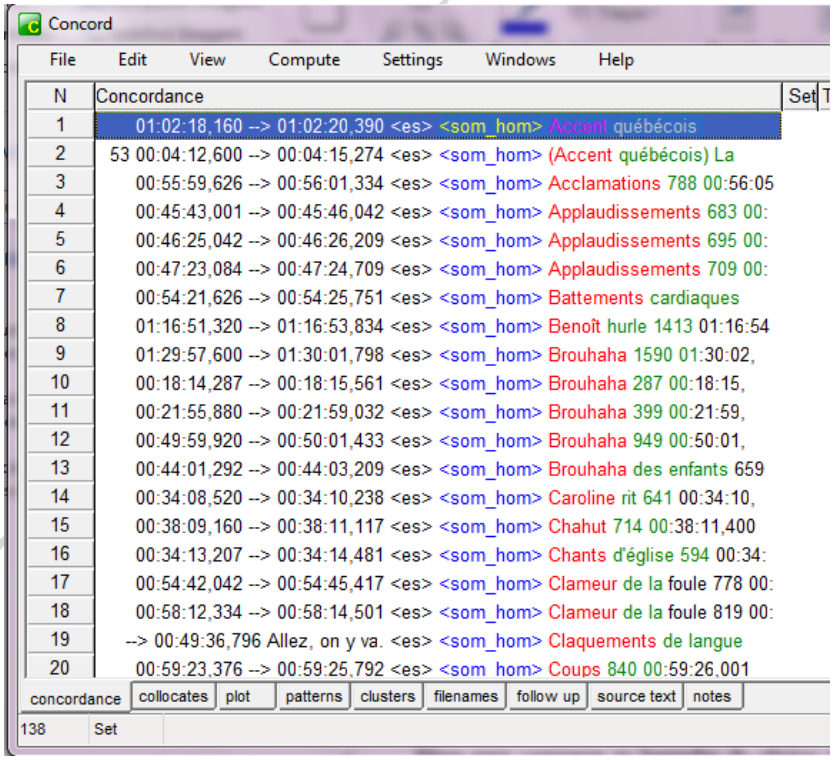


Figura 3 – Fragmento de tela de concordância com nódulos em ordem alfabética.
Fonte: as autoras.

Além da análise das linhas de concordância produzidas pelo *WST*, recorreu-se ao filme para comparar as legendas de efeitos sonoros com os sons do filme. Somando-se as noções

sobre o som no cinema revisadas na Subseção 2.2 a esses procedimentos analíticos, foi possível atingir o objetivo do trabalho.

Com a ferramenta *Concord*, foi feita também uma análise quantitativa, tendo-se contabilizado as etiquetas em números absolutos (o número de ocorrências de dada etiqueta é mostrado no canto esquerdo inferior da tela, como pode ser visto na Figura 3). Os dados quantitativos foram incluídos na discussão da análise qualitativa, permitindo a tomada de conclusões e obtenção dos resultados. Na seção a seguir, apresentamos os resultados da análise dos dados extraídos do *WordSmith Tools*.

4. Resultados da análise dos dados e discussão

Quanto à análise quantitativa geral, ela resultou nos seguintes dados: foram encontradas 283 inserções de tradução de música e ruídos na LSE de *NJH*, *F6E* e *LDJ*, sendo que dessas, 97 correspondem à tradução de música, 138 de sons causados pelo homem, 44 de sons causados por objetos, 3 de sons ficcionais e apenas 1 de som da natureza. Não foram encontradas legendas com tradução de sons de animais ou silêncio Estas etiquetas, apesar de terem sido criada.

O filme com mais efeitos sonoros legendados é *LDJ*, com um total de 126, seguido por *F6E* com 95 e, por fim, *NJH* com apenas 62. A presença de um grande número de traduções de efeitos sonoros em *LDJ* parece decorrer do fato de seu enredo dar bastante destaque à sua trilha sonora, repleta de músicas de artistas famosos, como Duke Ellington, e interpretadas pelo baixo de Paul McCartney. O mesmo ocorre com *F6E*, que possui uma música original composta por Jorge Arriagada.

Em quase todas as inserções de tradução de músicas nos três filmes, de um total de 97, as músicas foram qualificadas nas legendas, o que pode ser explicado por uma possível intenção de permitir que o espectador surdo ou ensurdecido possa fazer inferência acerca da função delas no enredo (*Musique douce, Musique entraînante)*. Apenas uma música foi traduzida sem qualificação em todo o *corpus* (*Musique à la radio)* e, ainda assim, supriu a necessidade em questão no enredo: fazer saber ao espectador que o rádio estava funcionando, não importando o gênero musical.

Uma tendência na legendagem francesa do efeito sonoro ‘música’ notada no *corpus* é a inclusão do nome da música, do cantor e/ou gênero musical ("*Never even thought", slow par Murray Head*); isso pode favorecer, principalmente os ensurdecidos, que ficam capazes de

depreender mais facilmente qual teor a música imprime à cena. É preciso salientar, porém, que esse tipo de tradução não é a preferida por surdos brasileiros. Isso acontece porque eles não conseguem depreender a relação entre música e filme a partir do cantor e/ou nome da música e/ou seu gênero, pois eles raramente sabem de quem se trata ou o que o gênero significa (NASCIMENTO, ARAÚJO, 2011).

Outra questão acerca da tradução de música encontrada no *corpus* foi a diferença do emprego da palavra *Musique* e *Chanson.* Quando a música foi traduzida como *Chanson*, ela apresentava letra; porém, quando foi traduzida como *Musique*, tratava-se apenas de música instrumental.

Quanto aos ruídos apenas de ambientação, ou seja, aqueles que apenas ajudam o espectador a identificar o espaço da ação (como é o caso de sirenes em hospitais, ou ruídos de animais em fazendas), é possível afirmar que não foram traduzidos. Apesar de estarem presentes nos filmes, a legenda não os contemplou.

Isso pode indicar uma preocupação maior na legendagem francesa em traduzir apenas sons que participem ativamente na trama, ou seja, que causem reação nos personagens. O excesso de tradução de ruídos pode prejudicar a compreensão do espectador, criando expectativas que não se concretizam no enredo. Além disso, os dados do projeto MOLES[5] (NASCIMENTO; ARAÚJO, 2011) mostraram que o excesso de "legendas entre colchetes" cansou o espectador, que passou a ignorar as informações nelas contidas.

O ruído mais traduzido no *corpus* foi o som causado pelo homem (<som_hom>). Esse tipo de som apareceu com frequência na LSE dos três filmes do *corpus*: 49 vezes em *F6E* e em *LDJ* e 40 vezes em *NJH*. Os principais sons traduzidos nessa categoria foram risos (*rires, elle rit)*, gritos (*cris, cris de protestation)*, além de burburinho (*brouhaha*). Isso deve ter ocorrido porque frequentemente esses sons expressam reações emotivas dos personagens e, se não fossem legendados, os surdos e ensurdecidos perderiam pelos menos parte da emoção da cena. Uma ocorrência frequente relativa aos sons causados pelo homem diz respeito a quando os personagens falavam em língua estrangeira. Esse ponto é importante, já que nos filmes em questão, nem mesmo o espectador ouvinte tem acesso à tradução durante o filme; então, os legendistas apenas apontaram que os personagens falavam uma língua estrangeira (*Elle parle*

[5] O título da pesquisa é Tradução para surdos: em busca de um modelo de legendagem fechada para o Brasil (NASCIMENTO; ARAÚJO, 2011)

en espagnol; Elle parle américain), o que bastou para que os surdos e ensurdecidos compreendessem a obra de acordo com a pretensão do diretor.

No tocante à tradução dos ruídos relativos aos sons causados por objetos (<som_obj>), tem-se apenas 4 em *NJH*, 14 em *F6E* e 26 em *LDJ*. Os números baixos podem demonstrar a preocupação em legendar apenas os ruídos que fazem parte efetiva do enredo fílmico; logo os sons que não contribuem para a significação fílmica foram pouco ou não legendados. O som mais traduzido na categoria foi campainha (*sonnette, sonnerie),* pois ele causa reação dos personagens, que frequentemente viram-se em direção à porta ou a abrem após ouvir o ruído. Sem a legenda o espectador surdo não saberia o porquê da reação dos personagens. Um outro frequentemente legendado foi o ruído de quando alguém bate em uma porta ou a abre (*On frappe, On tape à la porte),* pois também suscita resposta dos personagens.

A única tradução de som da natureza no *corpus* (*Coup de tonnerre*) ocorre em *LDJ*, em razão de um trovão que anuncia a chuva que irá ocasionar a doença da personagem principal. Apesar de ser um ruído causado pela natureza, não se configura como um simples som de ambientação, já que este serve apenas para tornar verossímil a cena (JULLIER, 2006), mas configura um marco importante na história do filme. A legenda, portanto, cumpre o papel do som ao salientar, em conjunto com a imagem, o acontecimento relevante. Uma possível causa da reduzida quantidade de traduções nesta categoria é a relevância ao enredo. Como, em todo o *corpus*, esse foi o único ruído da natureza relevante para a trama, ele foi o único contemplado com a legenda.

Há apenas três ocorrências de sons ficcionais legendados, todas no filme *LDJ*. '*Grognements*' (Figura 4) é ficcional nesse filme, pois se trata de um sapato soltando grunhidos como de cachorro. A tradução se faz indispensável, visto que, sem a legenda, o espectador surdo-ensurdecido precisaria apoiar-se apenas na imagem para compreender que o sapato se comporta como um cachorro. A baixa ocorrência desse som já era esperada, pois os filmes do *corpus* são do gênero comédia e/ou comédia dramática, que não costumam apresentar sons ficcionais. LDJ, porém, passa-se num universo absurdo, no qual é possível passear de teleférico sem fios de sustentação, cultivar armas de fogo e sapatos com comportamento canino. Mais uma vez, a baixa ocorrência dessa etiqueta pode apontar para uma legendagem mais voltada aos sons que efetivamente suscitam reações dos personagens.

Figura 4 – Legenda de som ficcional
Fonte: as autoras.

Diferentemente do que ocorreu em Nascimento (2013), no qual as músicas e ruídos nos filmes brasileiros estudados não foram legendados levando em conta as suas funções na trama, os dados obtidos no presente trabalho indicaram que a legendagem de música e ruídos em NJH, F6E e LDJ procurou traduzir apenas aquilo que é relevante ao enredo. Isso pôde ser percebido tanto pela quase totalidade de músicas qualificadas, quanto pelos baixa incidência de ruídos traduzidos que não suscitavam reação dos personagens. Analisando o corpus, especula-se a existência de dois parâmetros para a legenda de músicas e ruídos: 1) a legenda de música deve apresentar uma qualificação que indique sua função na trama, sempre que essa função for relevante; 2) deve-se buscar traduzir principalmente os ruídos que participam efetivamente da trama.

Nesta seção, foi observado como as diferentes categorias de efeitos sonoros foram traduzidas nos três filmes que compõem o *corpus* da presente pesquisa. A seção seguinte apresentará as considerações finais.

5. Considerações finais

Este trabalho se propôs a analisar a tradução de efeitos sonoros na LSE de três filmes franceses comercializados em DVD (*Nos jours heureux*, *Les femmes du sixième étage, L´écume des jours),* para examinar se os efeitos sonoros são traduzidos levando-se em conta a função de cada som na trama fílmica. Para responder essa questão, as legendas de efeitos sonoros dos filmes que compõem o *corpus* foram etiquetadas a partir de uma categorização de sons. Esse procedimento possibilitou a análise eletrônica das legendas pelo *software WordSmith Tools* com a ferramenta *Concord.* Aliados a essa metodologia, além dos

conhecimentos sobre o som no cinema, sempre que necessário se recorreu aos filmes (que são os textos originais) para sanar eventuais dúvidas acerca dos sons traduzidos.

A tradução de efeitos sonoros na LSE francesa, no âmbito do *corpus* analisado no estudo aqui relatado, procura incluir sons que suscitam reações dos personagens em tela, ou seja, que são relevantes para a compreensão da trama, sendo que não foram encontradas legendas de ruídos que servissem apenas para ambientar cenas no *corpus* de estudo. No que concerne às músicas presentes nos filmes, a maioria delas foi traduzida apresentando um elemento qualificador que ajuda o espectador a depreender sua função no enredo. Isso é importante, já que, sem esse elemento qualificador, a informação seria irrelevante ao espectador surdo-ensurdecido, podendo, inclusive, prejudicar sua compreensão da trama (NASCIMENTO, 2013).

Levando em consideração os resultados obtidos, é possível afirmar que os efeitos sonoros presentes no *corpus* de estudo foram traduzidos procurando respeitar suas funções dentro da trama fílmica. Dessa forma, conclui-se que a tradução de música e ruídos no *corpus* estudado possui boa qualidade, e pode servir como base para o treinamento e desenvolvimento de legendistas brasileiros iniciantes, além de contribuir para o aperfeiçoamento dos parâmetros de LSE investigados pelo projeto CORSEL[6], da Universidade Estadual do Ceará.

A metodologia baseada em *corpus* usada a partir dos pressupostos da Linguística de *Corpus* foi bastante eficaz para este trabalho, pois viabilizou a consecução do objetivo. É legítimo dizer isso porque houve a possibilidade de analisar cada categoria de som separadamente, tornando-a mais dinâmica e permitindo a observação de cada legenda em contexto. No entanto, além da observação do arquivo de texto da legenda, foi necessário sempre recorrer ao filme, por se tratar de um texto multimodal, e comparar cada som com sua tradução, observando, assim, a sua relevância na trama e como ocorreu a respectiva tradução. Dessa forma, cada legenda foi comparada não só com as outras legendas da mesma categoria, mas com seu respectivo som. Assim, foi possível verificar se cada legenda traduzia ou não a significação do filme.

Não se pretende encerrar o assunto com este trabalho, mas, ao contrário, incentivar mais pesquisas na área para aprimorar o conhecimento sobre a LSE e, assim, desenvolver uma tradução de qualidade no Brasil que atenda satisfatoriamente o público alvo de surdos e

[6] Corpus de segmentação em LSE

ensurdecidos, proporcionando-lhe acessibilidade cultural cada vez mais efetiva. Sugere-se, portanto, uma pesquisa de recepção com espectadores surdos-ensurdecidos para testar os resultados aqui encontrados.

É preciso ressaltar que as questões técnicas da LSE, tais como tempo de permanência em tela e quantidade de caracteres por minuto, não foram abordadas no presente trabalho por questões de restrição de espaço, mas podem ser contempladas em pesquisas futuras de caráter teórico-descritivo, promovendo a interface com as questões de legendagem dos efeitos sonoros via os pressupostos metodológicos propiciados pela Linguística de *Corpus*.

Referências

ARAÚJO, V. L. S.; NASCIMENTO, A. K. P. Investigando parâmetros de legendas para surdos e Ensurdecidos no Brasil. In: FROTA, M. P.; MARTINS, M. A. P. (orgs.). **Tradução em Revista**, v. 2, p. 1-18, 2011.

ARAÚJO, V. L. S. Por um modelo de legendagem para Surdos no Brasil. In VERAS, V. (org.). **Tradução e Comunicação**, Revista Brasileira de Tradutores, São Paulo: UNBERO, n. 17, p. 59–76, 2008.

AUBERT, J. P.; MARTI, M. **Quelques conseils pour le sous-titrage.** Disponível em : < http://lingalog.net/dokuwiki/_media/cours/sg/trad/methodest.pdf>. Acesso em: 14 de outubro de 2013.

BERCHAMPS, T. **A música do filme:** tudo o que você gostaria de saber sobre a música do cinema. São Paulo: Escrituras Editora, 2006.

BORDWELL, D.; THOMPSON, K. **Film art:** an introduction. New York: McGraw Hills, 2008.

CHAUME, F. **Cine y traducción.** Madri: Cátedra, 2004.

CHION, M. **A audiovisão:** som e imagem no cinema. Lisboa: Edições texto & grafia, 2008.

De LINDE, Z.; KAY, N. **The semiotics of subtitling**. Manchester: St. Jerome Publishing, 1999.

DÍAZ CINTAS, J.; REMAEL, A. **Audiovisual Translation**: Subtitling. Manchester: St. Jerome Publishing, 2007.

DIU, N. **Un métier à découvir :** adaptateur de programmes télévisés pour les sourds et malentendants. Disponível em : <http://www.semainedusoustitrage.org/IMG/pdf/Un_metiers_a_decouvrir.pdf>. Acessado em: 14 de outubro de 2013.

FRANCO, E; ARAÚJO, V. L. S. Reading Television: Checking Deaf People's Reactions to Closed Subtitling in Fortaleza, Brazil. In: GAMBIER, Y. (org.). **The Translator,** v. 09, n. 2, pp. 249-267, 2003.

HUNTER, C. **The use of sound effects and stylised ambiences in filmmaking**, 2008. Disponível em <http://freedownload.is/pdf/cinema-sound-effects>. Acesso em: 15 de março de 2012.

IRMÃOS de fé. Direção: Moacyr Góes. Brasil: Columbia, 2004. 1DVD (105 min), região 4, NTSC, color., legendas (para surdos em português), janela de LIBRAS, audiodescrição e audionavegação.

JULLIER, L. **Le son au cinéma.** Paris : Cahiers cinéma, SCEREN (CNPD), 2006.

NASCIMENTO, A. K. P. **Linguística de *corpus* e legendagem para surdos e ensurdecidos (LSE)**: uma análise baseada em *corpus* da tradução de efeitos sonoros na legendagem de filmes brasileiros em DVD. 2013. 109p. Dissertação (Mestrado em Linguística Aplicada). Pós Graduação em Linguística Aplicada, Universidade Estadual do Ceará, Fortaleza-CE. 2013.

NEVES, J. **Audiovisual translation:** subtitling for the deaf and hard of hearing. Tese (Doutorado). Universidade de Surrey Roehampton, Inglaterra, 2005. Disponível em: <http:// rrp.roehampton.ac.uk/artstheses/1>. Acesso em 15 de janeiro de 2012.

NOSSO Lar. Direção: Wagner de Assis. Brasil: Fox do Brasil, 2010. 1DVD (102 min), região 4, color., legendas (para surdos em português) e audiodescrição.

O SIGNO da cidade. Direção: Carlos Alberto Riccelli. Europa Filmes, 2008. 1DVD (95 min), região 4, color., legendas (para surdos em português) e audiodescrição.

REID, H. Literature on the screen: subtitle translation for public broadcasting. In: BART, W.; D'HAEN, T. (Eds.). **Something understood.** Studies in Anglo-Dutch literary translation. Amsterdam: Rodopi, p. 97-107, 1990.

Artigo recebido em: 15.10.2014
Artigo aprovado em: 11.12.2014

Contextos de ocorrência das perífrases de gerúndio e particípio no português do Brasil e na variedade do espanhol do México e sua significação aspectual

Contexts of occurrence of the gerund and participle periphrases in Brazilian portuguese and Mexican Spanish and its aspectual meaning

Anne Katheryne Estebe Maggessy*
Maria Mercedes Riveiro Quintans Sebold**

RESUMO: Neste artigo, centramos nossa análise na noção aspectual para contrastar duas línguas ricas morfologicamente. Trataremos das perífrases de gerúndio e particípio e seus contextos de ocorrência no Português do Brasil (doravante PB) e na variedade do espanhol do México (doravante EM) a partir da metodologia da Linguística de *Corpus*. Na primeira seção deste artigo, trataremos da noção de aspecto. Na segunda seção, trataremos da noção de perífrase e dos traços que caracterizam as perífrases de gerúndio e de particípio. Na terceira seção, trataremos dos contextos de ocorrência das perífrases de gerúndio e particípio no PB e no EM e, finalmente, na quarta seção, trataremos das noções aspectuais veiculadas por tais perífrases. Compartilhamos com Wachowicz (2006) que nas perífrases pode haver uma diferença de acordo com a forma nominal que compõe a mesma. Nessa perspectiva, a terminação da forma nominal –ndo, do gerúndio se combinaria com eventos atélicos, ao passo que a terminação –do, do particípio, se combinaria com eventos télicos. E compartilhamos com Bertinetto (2001) que as línguas combinam diferentemente telicidade/atelicidade e perfectividade/imperfectividade, podendo haver uma convergência entre atelicidade e imperfectivo e telicidade e perfectivo. O objetivo é verificar como essas relações entre formas nominais, telicidade e aspecto se dão nessas duas línguas.

PALAVRAS-CHAVE: Aspecto. Perífrase. Formas nominais. Linguística de *Corpus*.

ABSTRACT: In this paper, we focus our analysis on the aspectual notion to contrast two morphologically rich languages. We will tackle the periphrases of gerund and participle, and their contexts of occurrence in Brazilian Portuguese (PB) and variety of Mexican Spanish (MS) using the methodology of Corpus Linguistics. In the first section of this paper will deal with the notion of aspect. In the second section will address the notion of periphrasis and the traces that characterize the periphrases of gerund and participle. The third section will deal with the contexts of occurrence of the gerund and participle periphrases in PB and MS and finally, in the fourth section, we discuss the aspectual notions conveyed by such circumlocutions. Sharing with Wachowicz (2006) that the circumlocutions can be a difference according to the nominal form that composes it. In this perspective, the termination of the nominal form -ndo, the gerund would combine with atelic events, while -do termination of the participle, would combine with telic events. And we share with Bertinetto (2001) that the languages combine differently telicity/atelicity and perfective/imperfective, and there may be a convergence between atelicity and imperfective and telicity and perfective. The goal is to see how these relationships between nominal forms, telicity and appearance are given in both languages.

KEYWORDS: Aspect. Periphrase. Nominal forms. Corpus Linguistics.

* Mestre em Língua Espanhola pela UFRJ.
** Professora Doutora do Departamento de Letras Neolatinas e do Programa de Pós Graduação em Letras Neolatinas da Faculdade de Letras da UFRJ.

1. Introdução

Estudos sobre línguas tipologicamente próximas como o Português do Brasil (doravante PB) e o espanhol têm mostrado um maior ou menor distanciamento com relação a determinados fenômenos.

O objetivo deste artigo é o de contrastar duas línguas tipologicamente próximas a partir da informação aspectual. Trataremos das perífrases de gerúndio e particípio e sua produtividade no Português do Brasil (doravante PB) e na variedade do espanhol do México (doravante EM). Na primeira seção, trataremos da noção de aspecto. Na segunda seção, trataremos da noção de perífrase e dos traços que caracterizam as perífrases de gerúndio e de particípio. Na terceira seção, trataremos dos contextos referentes ao tipo de verbo e ao tipo de evento em que se encontram essas perífrases e finalmente da sua leitura aspectual veiculada composicionalmente.

Nossa motivação para tal estudo tem origem em Wachowicz (2006) que propõe que a forma nominal impõe restrições aos demais elementos da sentença. Nessa perspectiva, a terminação da forma nominal –ndo, do gerúndio se combinaria com eventos atélicos, ao passo que a terminação –do, do particípio, se combinaria com eventos télicos.

Central a essa ideia é a noção de telicidade. A origem do termo "telicidade" está na palavra grega "telos", entendida como "fim". Sendo assim, um evento télico é aquele que tem um fim ou uma meta previsível a ser atingida e que, portanto, pode ser considerado terminado quando este fim ou meta é alcançado, como no exemplo "correr uma maratona". Um evento atélico não terá um fim ou meta previsível e, por isso, pode continuar indefinidamente, como no exemplo " correr maratonas", no qual não é possível visualizar o fim do evento..

Neste artigo, vamos trabalhar com a noção de telicidade relacionada também aos tipos de verbo a partir da classificação proposta por Vendler (1957) – estado, atividade, *accomplishments* (processo culminado) e *achievements* (culminação); e da divisão em traços aspectuais proposta por Smith (1991): estado (atélicos, durativos e pontual), atividades (atélicos, durativos e não-pontual), processo culminado (télicos, durativos e não-pontual) e culminação (télicos, não-durativos e não-pontual).

Com relação às categorias de telicidade/atelicidade e perfectividade/imperfectividade, Bertinetto (2001) afirma que não co-variam, mas se comportam geralmente de forma independente, justamente por pertencerem a noções diferentes, sendo a primeira relativa à

Acionalidade (*Aktionsart* ou aspecto lexical) e a segunda relativa ao aspecto (ou aspecto gramatical).

2. Sobre o aspecto lexical e o aspecto gramatical

Comrie (1976) define tempo como uma categoria dêitica, isto é, que faz referência a outro tempo, o momento da enunciação. Em contrapartida, o aspecto é definido por este mesmo autor como uma categoria não dêitica porque se refere à situação em si.

O termo aspecto pode se referir a traços de naturezas bastante diversas. As propriedades inerentes às raízes verbais bem como a outros itens lexicais dizem respeito ao aspecto lexical. O aspecto lexical também está relacionado às classes aspectuais, ou seja, os verbos podem sugerir uma forma particular de conceber a noção de tempo (time). Seguimos a divisão proposta por Vendler (1957) acrescida dos traços aspectuais propostos por Smith (1991). Tal classificação está centrada em uma divisão em quatro classes: estados - **saber**, atividades - **correr**, processo culminado - **comeu dois doces** e culminações - **perdeu o livro**. A proposta de Smith (1991) se centra em dois eixos. No primeiro eixo, o foco está no ponto final natural do evento (ponto télico). O segundo eixo mostra se o evento é dinâmico ou se apresenta em estágios.

Nessa perspectiva, eventos com verbos do tipo "estado", como, por exemplo, "ver", ou "atividade", como, por exemplo, "correr" são [- télicos] (menos télicos) ou atélicos. Por outro lado, os verbos do tipo processo culminado, como, por exemplo, "construir", e culminação, como, por exemplo, "cair", têm um ponto final natural e, por isso, são considerados [+ télicos] (mais télicos) ou télicos.

No que diz respeito ao aspecto gramatical, as marcações são feitas na morfologia (seja por auxiliares e/ou morfemas flexionais e derivacionais), como, por exemplo, "estou comendo" indicando um evento contínuo e "comi", indicando um evento fechado.

Em seu estudo sobre o aspecto, Comrie (1976) propõe que as línguas humanas apresentam diferentes tipos de oposições aspectuais e busca comprovar sua hipótese através da análise dos tempos verbais de passado em línguas como inglês e português, por exemplo. Sua categorização postula a existência de dois aspectos básicos: o perfectivo e o imperfectivo. O autor, entretanto, propõe um diagrama em que somente o aspecto imperfectivo se subdividiria em diferentes categorias, tais como: habitual e contínuo (progressivo e não progressivo).

Entendemos que a bipartição proposta por Comrie (1976) está pensada para o aspecto gramatical e diz respeito a uma visão global do evento centrada na pontualidade ou não do evento.

Na sentença vista sob uma perspectiva composicional, temos, portanto, informação antes da flexão, no domínio lexical, sobre a pontualidade ou não pontualidade do evento, como, por exemplo, morrer [+ pontual] e morar [- pontual]. E na flexão, no domínio gramatical, temos a informação sobre a completude ou não completude do evento, como, por exemplo, morou [+ completude] e morava [- completude].

Línguas de morfologia flexional mais rica como o PB e o espanhol dispõem para expressão da imperfectividade da terminação *–ia/-va; ía/aba* e da terminação *–ndo / iendo* de gerúndio.

Para a expressão da perfectividade, o PB dispõe do tempo pretérito perfeito simples cujas desinências dão ao evento uma leitura télica, e a morfologia de particípio *ado/-ido*. O espanhol também dispõe do *pretérito perfecto simple* ou *pretérito indefinido* e também do particípio *ado/-ido*.

Uma língua menos rica morfologicamente como o inglês, por exemplo, não distingue morfologicamente o perfectivo e o imperfectivo e isso somente pode ser feito com o acréscimo de termos adjacentes na estrutura.

Bertinetto (2011) ao explorar a relação entre o léxico e a flexão ou entre as leituras [±télico] e [±perfectivo] propõe uma provável convergência em muitas línguas naturais entre atelicidade e imperfectivo e telicidade e perfectivo. Entretanto, o autor propõe que as línguas combinam diferentemente telicidade/atelicidade e perfectividade/imperfectividade e esse pode ser um elemento que as aproxima ou as distancia.

É interessante notar, que para Bertinetto (idem), a imperfectividade, em contextos progressivos, com a perífrase estar + gerúndio, suspende o valor télico do verbo como os do tipo processo culminado e culminação. O que não significa dizer, segundo o autor, que a imperfectividade e a atelicidade necessariamente convirjam. Quando essas duas categorias interagem, o produto dessa interação é passível de um conjunto bastante restrito de possibilidades. Com relação à telicidade e à imperfectividade, o autor aponta para a manutenção do caráter télico do evento em sentenças que expressam o imperfectivo habitual.

Além disso, Bertinetto (idem), também acha necessário diferenciar perfectividade de telicidade, por serem categorias comumente confundidas devido ao fato de ambas indicarem o

ponto final do evento. Com isso, o autor relaciona o aspecto à terminatividade e a telicidade à fronteira ou ao limite (*boundness*).

O autor propõe que a telicidade implica em perfectividade, mas perfectividade não implica em telicidade. Ou seja, todo evento télico seria perfectivo, mas nem toda a perfectividade seria expressa por eventos télicos. Reproduzimos a seguir a representação proposta por Bertinetto (idem):

Quadro 1. Perfectividade e Telicidade.

	+ perfectivo	- perfectivo
+ télico	[a] SIM	[b] (NÃO)
- télico	[c] SIM	[d] SIM

Essa relação é bastante compreensível, segundo o autor, pois se a imperfectividade é vista como envolvendo intervalos abertos, é esperado que essa propriedade entre em contradição com a definição do traço de telicidade, que justamente envolve um evento fechado. Como a noção de habitualidade é oposta à noção de progressividade dentro do aspecto imperfectivo, a telicidade pode estar presente na expressão da habitualidade.

3. Sobre a perífrase

Embora essa definição não seja consensual, Ilari (1997) e Castilho (2002) são autores que defendem que nas perífrases verbais há um "todo indivisível", com papéis bem definidos, tanto para o verbo auxiliar quanto para verbo principal. O verbo auxiliar marca o tempo e os traços de flexão, como pessoa e número e o segundo verbo particulariza o evento ou a ação expressa.

Longo e Campos (2002) também definem as perífrases como um complexo unitário que reúne um verbo e uma forma de infinitivo, gerúndio ou particípio numa só predicação (LONGO E CAMPOS, 2002, p. 447). As autoras definem os seguintes critérios para classificar um verbo como auxiliar:

1. não possibilidade de desdobramento da oração intimamente relacionada à existência de um sujeito único.

2. perda sofrida pelo auxiliar de atribuir papéis temáticos aos elementos nominais com os quais se combinam.

Neste artigo, também partimos da ideia de que a perífrase verbal é a combinação de dois elementos verbais: um verbo auxiliar flexionado e uma forma não flexionada (particípio e gerúndio). São perífrases aqueles complexos que funcionam como uma única unidade verbal.

Considerando-se que numa forma verbal simples como "canto" podemos encontrar informações sobre tempo e aspecto, o que podemos dizer de uma perífrase verbal? Essas informações de tempo e aspecto estão separadas no auxiliar e na forma não flexionada?

Seguindo uma determinada linha de pensamento, na sentença "Alice **está telefonando** para seu chefe.", o tempo estaria concentrado na forma em presente do auxiliar "está" e a noção aspectual de duração estaria concentrada na forma não flexionada de gerúndio "telefonando".

Longo & Campos (2002) dividem as perífrases em aquelas que só indicariam o tempo em que o evento ocorre em relação ao momento de fala, chamadas de temporais: ex.: **havíamos programado**; e aquelas que indicariam além do tempo, o aspecto, ou seja, como o evento se desenrola no tempo, chamadas aspectuais: ex.: **estão entendendo**. Para as autoras, na perífrase temporal o verbo auxiliar tem localização temporal separada do verbo principal — marca somente o tempo do evento.

Entretanto, neste artigo, atribuímos as diferenças temporais e aspectuais a fatores, tais como: o tempo verbal do auxiliar e a forma nominal que compõem a perífrase (se se trata de particípio ou se se trata de gerúndio).

Defendemos, portanto, a possibilidade de que haja oscilações entre as leituras aspectual e temporal nas diferentes perífrases, bem como, a possibilidade de sobreposição de diferentes leituras.

Compartilhamos com Wachowicz (2006) que nas perífrases pode haver uma diferença de acordo com a forma nominal que compõe a mesma. Nessa perspectiva, a terminação da forma nominal –ndo, do gerúndio, favorece a leitura de eventos iniciados mas que não têm necessariamente um fim e, portanto, seriam atélicos, ao passo que a terminação –do, do particípio, favorece a leitura de eventos que parecem ter um fim determinado e, portanto, esta terminação geraria eventos télicos.

Com relação à perífrase estar + particípio, que analisaremos neste artigo, Castilho (1968) afirma que esta é a principal forma de expressão do aspecto perfectivo resultativo, pois representa um completamento total da ação que implica em um resultado que decorre desse completamento. (CASTILHO, 1968, p. 78) Já Travaglia (2006) afirma que essa perífrase poderá expressar tanto o aspecto perfectivo quanto o aspecto imperfectivo como em "O assaltante está preso". (TRAVAGLIA, 2006, p. 175)

Para o espanhol, o "*Diccionario de perífrasis verbales*" de 2006 afirma que as perífrases de particípio estão estreitamente relacionadas com a voz passiva, pois ou são passivas ou foram na sua origem. Além disso, o autor Gómez Torrego (1988) afirma que o particípio fará sempre referência a um fato anterior ao tempo designado pela frase verbal principal. Para ele, isso se deve ao fato de que existem nas perífrases de particípio valores aspectuais que remetem sempre à ideia de perfectividade (ação acabada) do verbo. No entanto, o autor chama a atenção para a perífrase estar + particípio que pode adquirir um aspecto durativo de valor estativo, que o aproxima à significação do gerúndio. Como no exemplo dado pelo próprio autor "La casa de gobierno está vigilada por la policía" (= la policía la está vigilando). (GÓMEZ TORREGO, 1988, p. 195)

Já no caso das perífrases com gerúndio, tanto no português quanto no espanhol, é unânime a ideia de que estas só denotam imperfectividade.

4. Metodologia e *Corpus*

Neste artigo, utilizamos a Linguística de *Corpus* porque acreditamos que a utilização de uma grande quantidade de dados relativos ao efetivo uso da linguagem e a ajuda de programas de computador nos oferecerá subsídios para o estudo da gramática particular dos indivíduos. Utilizamos na nossa investigação do fenômeno linguístico proposto *corpora* orais de fala espontânea compilados nos últimos anos da década de 90. Analisamos três entrevistas sociolinguísticas das amostras do espanhol falado na Cidade do México do Projeto PRESEEA – México, e, para análise da variedade carioca do português do Brasil (doravante PB), utilizamos oito entrevistas sociolinguísticas do Projeto NURC-RJ.[1] Os informantes, cujas falas serão analisadas, possuem nível de escolaridade superior e compreendem a faixa etária

[1] Os temas das entrevistas selecionadas do espanhol são amigos, trabalho, estudos, situação familiar, filhos, casamento, vida acadêmica, história, costumes e problemas do bairro Tepito, E os temas das entrevistas do português são instituições, ensino, Igreja, cidade, comércio, vida social, diversões, família, ciclo de vida e saúde.

jovem (20 a 35 anos). O número desproporcional de entrevistas é justificado pela grande diferença de números de palavras em cada entrevista. Nas três entrevistas referentes à variedade da Cidade do México, encontramos 34.648 palavras e nas oito entrevistas da variedade carioca encontramos 32.801 palavras.

Com relação à análise dos dados, realizamos um levantamento quantitativo das ocorrências de sentenças com 'estar' + gerúndio e 'estar' + particípio tendo o auxiliar no presente do indicativo e verbos transitivos nas perífrases de particípio, tanto no PB quanto no espanhol do México (doravante EM), e de alguns fatores sintáticos que circundam cada sentença. Em seguida, realizamos uma análise qualitativa dos contextos identificados para chegar a um resultado que, de certa forma, mostre os fatores linguísticos que influenciam na interpretação aspectual perfectiva e imperfectiva dessas perífrases.

Para tal análise, o programa de computador utilizado foi o *WordSmith Tools* (versão 5). A ferramenta computacional *WordSmith Tools* (versão 5.0), desenvolvida por Mike Scott e *Oxford University Press*, consiste em uma metodologia atual, organizada e totalmente confiável que facilita a análise quantitativa, pois nos mostra todas as ocorrências existentes do que estamos pesquisando. Dentre as ferramentas disponibilizadas pelo software, *WordList*, *Keywords* e *Concord* entre outros utilitários, a ferramenta *Concord* foi a de maior relevância para a análise realizada na pesquisa, pois busca as palavras selecionadas dentro do seu contexto.

5. Análise dos dados

Inicialmente optamos por fazer uma análise quantitativa dessas perífrases na variedade carioca do português do Brasil (doravante PB) e na variedade do espanhol da Cidade do México (doravante EM). Embora não tenhamos conseguido o tempo de interação de cada entrevista, resolvemos equilibrar a análise a partir do número de palavras transcritas em cada uma, o que nos proporcionou analisar três entrevistas da variedade da Cidade do México com um total de 34.648 palavras e oito entrevistas da variedade carioca com um total de 32.801 palavras.

Tabela 1. Análise Quantitativa.

ANÁLISE QUANTITATIVA		
	PB (32.801 palavras)	EM (34.648 palavras)
Estar + particípio	11 perífrases	10 perífrases
Estar + gerúndio	62 perífrases	50 perífrases

Com isso, encontramos no PB 11 sentenças com a perífrase estar + particípio e 62 sentenças com estar + gerúndio. E no EM encontramos 10 sentenças com estar + particípio e 50 sentenças com estar + gerúndio. O que nos leva a considerar a partir dos dados levantados a possibilidade de uma maior seleção da perífrase de gerúndio em ambas as línguas analisadas, sendo que no PB parece haver ainda uma maior ocorrência, já que em uma contagem menor de palavras apresentou mais sentenças com essa perífrase.

Considerando a proposta de Wachowicz (2006) segundo a qual a forma –ndo marca a atelicidade e mais acentuadamente a habitualidade, essa seleção da forma nominal de gerúndio nas duas línguas pode ter repercussão na seleção dos demais elementos da sentença. Dessas 62 perífrases de gerúndio do PB, desconsideramos da análise 11 sentenças, e das 50 perífrases de gerúndio do EM, desconsideramos 8 sentenças. Isso porque as consideramos como expressões lexicalizadas que poderiam comprometer a qualidade da análise, como por exemplo “está dando aula”, “ está fugindo do assunto”, “está sendo beneficiada” no PB e “está pasando en la tele”, “estoy yendo bien a la chava”, “lo estoy echando como una carga” no EM.

Tabela 1. Sentenças analisadas.

SENTENÇAS ANALISADAS		
	Português do Brasil	Espanhol do México
Estar + particípio	11	10
Estar + gerúndio	51	42

Para a análise qualitativa das sentenças com as perífrases de particípio e gerúndio, consideramos como variáveis dependentes os aspectos perfectivo e imperfectivo e observamos a composicionalidade aspectual a partir das seguintes variáveis independentes linguísticas: transitividade verbal, o tipo de verbo segundo as categorias de Vendler (1957) e a telicidade do evento - se era télico como em “tá vendo exatamente o dia-dia”, atélico como em “a gente tá vendo coisas” ou sem telicidade delimitada como em “agora não tá estudando,

não sei o que", que é uma sentença sem um complemento determinado próximo do verbo ou de fácil recuperação.

Tabela 2. Estar + particípio, português do Brasil.

Português do Brasil Estar + particípio		PERFECTIVO			IMPERFECTIVO		
		[+T]	[-T]	X	[+T]	[-T]	x
TRAN	Estado	---	---	---	---	---	---
	Atividade	---	---	---	----	---	---
	Proc. Culminado	---	---	2	---	---	---
	Culminação	3	---	---	---	---	1
INTR	Estado	---	---	1	---	---	---
	Atividade	---	---	1	---	---	---
	Proc. Culminado	---	---	---	---	---	---
	Culminação	---	---	3	---	---	---

Das 11 sentenças encontradas com a perífrase estar + particípio no PB, 10 sentenças apresentaram o aspecto perfectivo, por representarem um evento já acabado, sendo 3 com verbos transitivos do tipo culminação com evento télico, 3 verbos intransitivos do tipo culminação e sem telicidade determinada, 2 sentenças com verbos transitivos do tipo processo culminado e sem telicidade determinada, 1 verbo intransitivo do tipo estado sem telicidade determinada e 1 verbo intransitivo do tipo atividade e também sem telicidade determinada. Exemplos:

(1) "Você já **tá garantido** no profissional" (perfectivo, culminação, transitivo e télico)
(2) "Ele **tá casado** há dois anos" (perfectivo, culminação, intransitivo e sem telicidade determinada)
(3) "Aqui no Rio, não **tá acostumado**, né? (perfectivo, processo culminado, transitivo, sem telicidade determinada)
(4) "e você não tiver um, ou uma família rica ou um convênio, **tá morto**!" (perfectivo, estado, intransitivo e sem telicidade determinada)
(5) "Agora a gente **tá brigado**" (perfectivo, atividade, intransitivo e sem telicidade determinada)

Embora tenhamos obtido 2 sentenças com verbos do tipo [- télico], ainda assim os dados parecem sugerir uma relação entre telicidade e perfectivo, assim como entre telicidade e particípio como afirmado por Bertinetto e Wachowicz, respectivamente.

Já com o aspecto imperfectivo, obtivemos apenas 1 sentença que dá a ideia de evento não acabado com repetição devido à expressão adverbial "o tempo todo", apresentando verbo transitivo e sem evento determinado. Exemplo:

> (6) "... telefone **tá ocupado** o tempo todo..." (Culminação, transitivo, sem telicidade determinada e imperfectivo/iterativo)

Essa ocorrência nos leva a supor o importante papel da expressão adverbial na composicionalidade aspectual de uma sentença.

Tabela 3. Estar + particípio, espanhol do México.

Espanhol do México Estar + particípio		PERFECTIVO			IMPERFECTIVO		
		[+T]	[-T]	X	[+T]	[-T]	x
TRAN	Estado	---	---	---	---	---	---
	Atividade	---	---	---	----	---	---
	Proc. Culminado	4	---	---	---	---	---
	Culminação	3	---	2	---	---	---
INTR	Estado	---	---	---	---	---	---
	Atividade	---	---	---	---	---	---
	Proc. Culminado	---	---	---	---	---	---
	Culminação	---	---	1	---	---	---

Diferentemente do PB, absolutamente todas as sentenças do EM apresentaram uma significação perfectiva, por representarem um evento já acabado. Com relação ao tipo de verbo, 6 eram do tipo culminação e 4 do tipo processo culminado, todos do tipo [+ télico]. Das sentenças com verbos do tipo culminação, 3 sentenças eram com verbo transitivo e evento télico, 2 eram com verbo transitivo e não apresentaram telicidade determinada e 1 era com verbo intransitivo e também sem telicidade determinada. Enquanto que das sentenças com verbos do tipo processo culminado, todas eram com verbo transitivo e com eventos télicos. Exemplos:

> (7) "porque todo **está modificado**" (perfectivo, culminação, transitivo, télico)
> (8) "siento que si se puede porque **está comprobado**" (perfectivo, culminação, transitivo e sem telicidade determinada)
> (9) "Ya acuando llegas ya estoy dormido" (perfectivo, culminação, intransitivo e sem telicidade determinada)
> (10) "y ahorita pues **está planeado** construir todo um cuarto arriba" (perfectivo, processo culminado, transitivo e télico)

A partir desses dados pudemos verificar uma semelhança entre as duas línguas na expressão do aspecto perfectivo em sentenças com perífrases de particípio. Ambas apresentaram sentenças em sua maioria com eventos télicos bem marcados, tanto no tipo de verbo quanto no tipo de evento. Assim a relação da perfectividade com a telicidade e de particípio com telicidade também pôde ser verificada no espanhol.

Para a análise das sentenças com a perífrase estar + gerúndio, também consideramos os aspectos perfectivo e imperfectivo, e também observamos a composicionalidade aspectual a partir do tipo de verbo segundo as categorias de Vendler (1957) e os traços de Smith (1991), a telicidade do evento e a transitividade do verbo.

Com relação ao aspecto, diferentemente das perífrases de particípio que alternaram, de alguma forma, entre os aspectos considerados, as perífrases de gerúndio apresentaram apenas o aspecto imperfectivo, tanto no PB quanto no EM. Pois, a perífrase estar + gerúndio parece realmente suscitar, exclusivamente, uma leitura de eventos iniciados, mas que não têm necessariamente um fim.

Analisando o tipo de verbo nas sentenças do PB, observamos 3 sentenças com o verbo do tipo estado, 21 com o verbo do tipo atividade, 23 com o do tipo processo culminado e 4 com o do tipo culminação.

Tabela 4. Estar + gerúndio. Português do Brasil.

Português do Brasil Estar + gerúndio		PERFECTIVO			IMPERFECTIVO		
		[+T]	[-T]	x	[+T]	[-T]	x
TRAN	Estado	---	---	---	---	2	1
	Atividade	---	---	---	----	14	3
	Proc. Culminado	---	---	---	11	---	6
	Culminação	---	---	---	---	---	2
INTR	Estado	---	---	---	---	---	---
	Atividade	---	---	---	---	3	---
	Proc. Culminado	---	---	---	6	---	---
	Culminação	---	---	---	1	1	1

Das 3 sentenças com o verbo do tipo estado, todos são verbos transitivos, sendo 2 sentenças com eventos atélicos e 1 sentença sem telicidade determinada. Exemplos:

> (11) “uma pessoa que **tá querendo** fazer Eletrônica” (imperfectivo, estado, transitivo e atélico)
> (12) “se você **tá sendo** seguido por algum carro” (imperfectivo, estado, transitivo e sem telicidade determinada)

Das 20 sentenças com o verbo do tipo atividade, 17 apresentaram verbos transitivos, sendo 14 com eventos atélicos e 3 sem telicidade determinada, e 3 com verbos intransitivos, todas com eventos atélicos. Exemplos:

(13) "isso nunca aconteceu no Brasil, não **tô entendendo** isso" (imperfectivo, atividade, transitivo e atélico)
(14) "essa loja **tá vendendo** mais do que a gente" (imperfectivo, atividade, transitivo e sem telicidade determinada)
(15) "porque eu **tô namorando** há quatro anos e meio" (imperfectivo, atividade, intransitivo e atélico)

Das 23 sentenças com o verbo do tipo processo culminado, 17 apresentaram verbos transitivos, sendo 11 com eventos télicos e 6 sem telicidade determinada. As outras 6 sentenças, do total de 23 sentenças, apresentaram verbos intransitivos, sendo todos eventos télicos. Exemplos:

(16) "quando eu estou em casa eu **estou fazendo** alguma coisa" (imperfectivo, processo culminado, transitivo e télico)
(17) "ele **tá tentando** né, ele é um cara corajoso" (imperfectivo, processo culminado, transitivo e sem telicidade determinada)
(18) "tem que tomar muito cuidado, ele **tá crescendo** – bairro da Barra" (imperfectivo, processo culminado, intransitivo e télico)
Das 5 sentenças com verbos do tipo culminação, 2 são verbos transitivos com eventos sem telicidade determinada e os outros 3 são verbos intransitivos, sendo 1 com evento télico, outro com evento atélico e outro sem telicidade determinada. Exemplos:
(19) "e tá morando em Bonsucesso" (imperfectivo, culminação, transitivo e sem telicidade determinada")
(20) "ah, **tá mudando** um pouco" (imperfectivo, culminação, intransitivo e télico)
(21) "não tô nem aí, mas **tá mudando**" (imperfectivo, culminação, intransitivo e sem telicidade determinada)

Considerando os dados analisados, pudemos verificar que diferentemente do perfectivo com sentenças télicas ou sem telicidade delimitada, o imperfectivo apresentou sentenças atélicas como o esperado, mas também sentenças télicas, tanto no tipo de verbo quanto no tipo de evento, assim como sentenças sem o complemento expresso. O que nos leva a supor que a relação entre imperfectivo e atelicidade não é tão estreita quanto a relação entre perfectivo e telicidade no PB.

Analisando o tipo de verbo das sentenças com a perífrase estar + gerúndio no EM, observamos que das 42 sentenças, 3 são do tipo estado, 17 são do tipo atividade, 21 são do tipo processo culminado e 1 do tipo culminação.

Tabela 5. Estar + gerúndio, espanhol do México.

Espanhol do México Estar + gerúndio		PERFECTIVO			IMPERFECTIVO		
		[+T]	[-T]	x	[+T]	[-T]	x
TRAN	Estado	---	---	---	---	1	1
	Atividade	---	---	---	----	8	6
	Proc. Culminado	---	---	---	19	---	---
	Culminação	---	---	---	1	---	---
INTR	Estado	---	---	---	---	---	1
	Atividade	---	---	---	---	---	3
	Proc. Culminado	---	---	---	2	---	---
	Culminação	---	---	---	---	---	---

Das 3 sentenças do tipo estado, 2 são com verbos transitivos sendo 1 com evento atélico e outra sem telicidade determinada, e 1 sentença com verbo intransitivo com evento sem telicidade determinada. Exemplos:

(18) “es algo que **se está viendo** ultimamente” (imperfectivo, estado, transitivo e atélico)
(19) “aparte **me está gustando**” (imperfectivo, estado, transitivo e sem telicidade de terminada)
(20) “donde **estoy vivendo** ahorita” (imperfectivo, estado, intransitivo e sem telicidade determinada)

Das 17 sentenças do tipo atividade, 14 são com verbos transitivos, sendo 8 com eventos atélicos e 6 com eventos sem telicidade determinada. As outras 3 sentenças restantes contém verbos intransitivos com eventos sem telicidade determinada. Exemplos:

(21) “**se está metiendo** lo que es este artículos de computación” (imperfectivo, atividade, transitivo e atélico)
(22) “ahorita **se está usando** así” (imperfectivo, atividade, transitivo e sem telicidade determinada)
(23) “**estoy radicando** ... en el Pueblo de San Mateo” (imperfectivo, atividade, intransitivo e sem telicidade determinada)

Das 21 sentenças com verbos do tipo processo culminado, 19 contém verbos transitivos com eventos télicos e 2 são verbos intransitivos também com eventos télicos. Exemplos:

> (24) "**estoy haciendo** mis trámites" (imperfectivo, processo culminado, transitivo e télico)
> (25) "lo que **estoy vivendo** y experimentando" (imperfectivo, processo culminado, intransitivo e télico)

E a única sentença com verbo do tipo culminação, tem o verbo transitivo e o evento télico:

> (26) "lo que hace sentir bien/ es que cosas malas/ a lo mejor no les **estoy dejando**".

De igual modo, no EM, a expressão do imperfectivo em sentenças com as perífrases estar + gerúndio também não apresenta uma relação tão estreita entre o imperfectivo e a atelicidade. Nos dados analisados do EM, também encontramos sentenças expressando o imperfectivo com a telicidade marcada no tipo de verbo e no tipo de evento, assim como sentenças sem a telicidade delimitada no complemento.

Tais resultados nos levam a supor um comportamento semelhante entre essas duas línguas na expressão do aspecto perfectivo em sentenças com perífrases de particípio e do aspecto imperfectivo em sentenças com perífrases de gerúndio.

6. Considerações finais

Os dados levantados mostraram-se relevantes posto que por se tratar de línguas de morfologia rica esperávamos a confirmação das tendências descritas pelos estudos de aspecto apresentados aqui. Ou seja, esperávamos que todas as sentenças com perífrases de particípio apresentassem o traço [+télico] e que todas as sentenças com gerúndio apresentassem o traço [- télico] conforme afirmado por Wachowicz (2006), o que não aconteceu.

Inicialmente, observamos uma maior ocorrência da perífrase estar + gerúndio em ambas as línguas analisadas, sendo que no PB parece haver ainda uma maior seleção dessa perífrase, já que em uma contagem menor de palavras observamos um maior número de sentenças com estar + gerúndio.

No EM, todas as perífrases de estar + particípio expressaram o aspecto perfectivo. No PB, apenas uma expressou o aspecto imperfectivo. No caso dessa sentença, a leitura de imperfectividade vem dada por um marcador de duratividade.

Nos dados analisados não encontramos nenhuma sentença no EM com o aspecto imperfectivo, embora Gómez Torrego (1988) afirme que a perífrase estar + particípio pode adquirir um aspecto durativo de valor estativo.

Sobre a relação da perfectividade com a telicidade, no PB não encontramos nenhum evento notadamente atélico com verbos transitivos, pois ou os eventos eram télicos ou não possuíam telicidade determinada, ou apresentavam verbos do tipo [+ télico]. No entanto, com verbos intransitivos encontramos duas sentenças com verbos do tipo [- télico] expressando a perfectividade.

Ainda assim, os dados sugerem que há uma relação entre telicidade e perfectividade, corroborando a proposta de Bertinetto (2001), assim como uma relação entre o particípio e a telicidade, corroborando também a proposta de Wachowicz (2006). A ocorrência de um único evento télico com leitura imperfectiva e de duas sentenças com verbos intransitivos do tipo [- télico] com leitura perfectiva, mostrou a necessidade de investigar com maior detalhe essas possibilidades nas duas línguas.

Sobre as perífrases de gerúndio, confirmamos realmente a sua leitura imperfectiva favorecida pela terminação –ndo do gerúndio que nos oferece um evento aberto, sem um fim determinado. No entanto, não podemos afirmar que a leitura imperfectiva do progressivo modifica a telicidade do evento, como afirmado por Bertinetto (2001), justamente por termos encontrado sentenças em imperfectivo com um complemento télico que marcava o fim inerente do evento, como em "**estoy escribiendo** la tesis". Além disso, tanto no PB quanto no EM encontramos mais verbos do tipo [+ télico] como processo culminado e culminação em sentenças imperfectivas com a perífrase de gerúndio, do que verbos do tipo [- télico] como os verbos do tipo estado e atividade. O que nos leva a considerar que a relação entre perfectivo e télico parece ser um pouco mais estreita que a relação entre imperfectivo e atélico.

Ao relacionarmos os aspectos lexical e gramatical, observamos tanto no PB quanto no EM uma semelhança na composicionalidade das sentenças com estar + gerúndio na expressão do aspecto imperfectivo e das sentenças com estar + particípio na expressão do aspecto perfectivo. Embora tenhamos encontrado uma sentença com estar + particípio no PB

expressando o aspecto imperfectivo, em ambas as línguas os eventos e tipos de verbo eram majoritariamente [+ télico].

As sentenças analisadas com estar + gerúndio, tanto no PB quanto no EM, expressaram exclusivamente o aspecto imperfectivo e não confirmaram a relação entre atelicidade e imperfectivo, pois a maioria dos verbos eram do tipo [+ télicos] e a maioria dos eventos também eram [+ télicos].

Com relação à transitividade verbal, pudemos verificar através dos dados a necessidade de um estudo mais aprofundado quanto à sua relevância tanto na formação de perífrases de particípio quanto nas de gerúndio. Visto que em ambas as línguas houve uma ocorrência bastante inferior de sentenças com verbos intransitivos.

Dessa forma, a partir dos dados levantados, neste estudo, podemos propor que o falante de PB e EM leva mais em conta a forma nominal selecionada para a expressão do aspecto, dando preferência para o gerúndio em uma leitura imperfectiva e para o particípio em uma leitura perfectiva. O reduzido número de eventos com telicidade determinada com particípio e gerúndio parece sugerir que essa marcação pode não ser tão relevante para o falante dessas duas línguas quando o seu interesse é marcar um evento mais aberto ou mais fechado.

Os resultados deste estudo são parciais e sugerem a necessidade de verificar tais tendências em *corpus* de registro escrito, tendo em vista a menor necessidade de recuperação do referente e, consequentemente, favorecendo uma telicidade indeterminada.

Finalmente, no que diz respeito à proposta de Bertinetto (2001) segundo a qual as línguas combinam diferentemente telicidade/atelicidade e perfectividade/imperfectividade e esse pode ser um elemento que as aproxima ou as distancia, no caso do PB e da variedade do espanhol analisada, as duas línguas demonstraram um comportamento semelhante na combinação de tais fatores sem distanciamentos significativos.

Referências bibliográficas

BERTINETTO, P. M. On a frequent misunderstanding in the temporal-aspectual domain: the ‘perfective-telic confusion’**.** In: CECHETTO, C. et alii. **Semantic Interfaces**: reference, anaphora and aspect. Stanford: CSLI Publications, 2001.

CASTILHO, A. T. de. **Introdução ao estudo do aspecto verbal na língua portuguêsa**. Marília: FFCL, coleção teses, 1968.

COMRIE, B. **Aspect**: An Introduction to the Study of Verbal Aspect and Related Problems. Cambridge: Cambridge:Cambridge University Press, 1976.

GARCÍA FERNÁNDEZ, L. (dir.), Ángeles Carrasco Gutiérrez, Bruno Camus Bergareche, María Martínez- Atienza & María Ángeles García García-Serrano: **Diccionario de perífrasis verbales**. Madrid: Gredos, 2006.

GÓMEZ TORREGO, L. **Perífrasis verbales. Sintaxis, se- mántica y estilística**. Arco/Libros, Madrid, 1988.

ILARI, R. **A expressão do tempo em português**. São Paulo: Contexto,1997.

LONGO, B. de O.; CAMPOS, O. de S. **A auxiliaridade**: perífrases de tempo e aspecto no português falado. In: Gramática do português falado: Volume VIII - Novos estudos descritivos. Campinas/ SP: Ed da Unicamp, 2002.

SCOTT, M. **Wordsmith Tools**. V 3.0. Oxford University Press, 1999.

SMITH, C. S. **The Parameter of Aspect**. Dordrecht, Kluwer Academic Press, 1991. crossref http://dx.doi.org/10.1007/978-94-015-7911-7

TRAVAGLIA, L. C. **O aspecto verbal no português**: a categoria e sua expressão. 4.ed. Uberlândia: EDUFU, 2006.

VENDLER, Z. **Linguistics in philosophy.** Ithaca: Cornell University Press, 1957.

Bibliografia

BASSO, Renato Miguel. Telicidade e Detelicização.In: **Revista Letras**, Curitiba, n.72, p.215-232, MAIO/AGO. 2007.

CASTILHO, A. **Introdução ao estudo do aspecto verbal na língua portuguesa**. Marília: Alfa, Faculdade de Filosofia, Ciências e Letras de Marília, ed. 12, 2002.

CASTILHO, A. T. de; MORAES de CASTILHO, C. M. O aspecto verbal no português falado. In: **VIII Seminário do Projeto de Gramática do Português Falado**. Campos do Jordão. 1994. (Mimeo)

CAVALLI, S. Perífrases verbais (vir+gerúndio e ter+particípio) e iteratividade. Comunicação apresentada em **VI FORUM DE LETRAS da PUCPR**, 2005 Curitiba. (CD-ROM).

CAVALLI, S. & WACHOWICZ, T.C. Verbos auxiliares vs aspectualizadores. Simpósio apresentado em **54.º seminário do GEL** – UNESP- Araraquara-SP, 2006. (HANDOUT)

CEGALLA, D. P. **Novíssima Gramática da Língua Portuguesa**. São Paulo: Companhia Editora Nacional, 46. ed., 2005.

CHIERCHIA, G. **Semântica**. Campinas: Editora da Unicamp; Londrina: Eduel, 2003.

COSTA, S. B. B. **O aspecto em português**. 3. ed. São Paulo: Contexto, 2002.

GILI GAYA, S. **Curso superior de sintaxis española**. 1948. Barcelona: VOX. 15º edição. Reimpressão: outubro de 2000.

ILARI, R. Notas para uma semântica do passado composto em português. In: **4º CELSUL**. Curitiba/UFPR, 2000. (mimeo)

_____ et al. Considerações sobre a posição dos advérbios. In CASTILHO, A . T. de (Org.). **Gramática do português falado**: a ordem. Campinas: Ed. da Unicamp, v. 1, p. 63-141, 1990.

_____ ; MANTOANELLI, I. As formas progressivas do português. In: **Caderno de Estudos Lingüísticos**. Campinas: IEL, Unicamp, n. 5, p. 27-60, 1983.

LENCI, A. & BERTINETTO, P. M. Aspect, adverbs and events – habituality vs. Perfectivity. In: HIGGINBOTHAM, J.; PIANESI, F.; VARZI, A. (eds.). **Speaking of events**. New York, Oxford: Oxford University Press, p.245-287, 2000.

MAGGESSY, A. K. E. **A significação aspectual iterativa da perífrase "estar" + gerúndio no português do Brasil e no espanhol do México.** Dissertação de mestrado. Rio de Janeiro, UFRJ, 2013.

MASACHS, L. G. **La adquisición del tempo verbal en el aprendizaje del espanhol como lengua extranjera.** Tese de doutorado. Departamento de Filologia Espanhola. Universidade Autónoma de Barcelona. 1998.

MENDES, R. B. Ter + particípio ou estar + gerúndio? Aspecto verbal e variação em PB. In: **Estudos Linguísticos** XXXIII, p. 1280-1285, 2004.

SANTOS, S. do R. C. dos. **Perífrases Durativas do Português Brasileiro**. Dissertação apresentada ao Curso de Pós-Graduação em Letras, Área de Concentração em Estudos Linguísticos, Setor de Ciências Humanas, Letras e Artes da Universidade Federal do Paraná. Curitiba, 2008.

TREVIÑO, E. La iteración de eventos. **Acta Poética**, 25-2. 145-183, 2004.

VARGAS, X. G. **Las perífrasis verbales resultativas en español de Chile**. Formas compuestas y perífrasis verbales en el español de Chile. Exploraciones sobre el desarrollo y el uso del aspecto perfecto. Tesis para optar al grado de licenciado o licenciada en Lengua y Literatura Hispánica con mención en Lingüística. Universidad de Chile. Facultad de Filosofía y Humanidades. Departamento de Lingüística. Santiago, Chile, 2009.

WACHOWICZ, T. C. As leituras aspectuais da forma de progressivo do PB. In: **Revista Letras**, Curitiba, n. 58, p. 397-406, jul./dez. 2002.

_____. Marcas linguísticas de iteratividade em PB. In: **Anais do 6° Encontro Celsul**-Círculo de Estudos Linguísticos do Sul, 2006.

_____ ; FOLTRAN, M. J. Sobre a noção de aspecto. In: **Cad. Est. Ling**., Campinas, 48 (2):211-232, 2006.

YLLERA, A. Las perífrasis verbales de gerundio y participio. **Gramática descriptiva de la lengua española**. Ignacio Bosque y Violeta Demonte, eds. Vol. 2. p. 3391-3441. Madrid: Espasa, 1999.

Anexos

ESTAR + PARTICÍPIO / PRESEEA-MÉXICO / 3 ENTREVISTAS PRESENTE: 10		
1	(E9) mejor en todos los aspectos/ ya sea como familia/ estando como pareja/ o independientemente// pero siento eso/ siento que sí se puede/ porque a-/ **está [comprobado]** 39 E: [mh] 40 I: ¿no?/ está comprobado// pero / no sé como que todo eso ya lo tengo así/// visualizado y pensado// digo si desafort	Culminação --- Transitivo PERFECTIVO
2	(E9) como familia/ estando como pareja/ o independientemente// pero siento eso/ siento que sí se puede/ porque a-/ está [comprobado] 39 E: [mh] 40 I: ¿no?/ **está comprobado**// pero / no sé como que todo eso ya lo tengo así/// visualizado y pensado// digo si desafortunadamente tuviera alguna complicación y	Culminação --- Transitivo PERFECTIVO
3	(E8) iba/ se va/ a poner no sé/ el lavadero o el cuarto de// donde metas la lavadora [y todo] 73 E: [¿para] quitar / de allá abajo? 74 I: ajá/ sí/ te digo/ **está planeado**// de esa manera 75 E: no pero a futuro claro que queda/ [poco a poco] 76 I: [pero va/ va a durar] mucho/ [o sea sí/ sí]	Processo Culminado Télico Transitivo PERFECTIVO
4	(E8) E: mh 86 I: sí/ porque/ hablaba de de/ él me/ me comentó algo// que/ a ellos les llama mucho la atención cuando vienen a esta zona// porque todo **está modificado**/ todo todo [todo] 87 E: [mh] 88 I: aquí/ se supone de/ que en un principio// todo tenía un/	Culminação Télico Transitivo PERFECTIVO
5	(E8) I: entonces/ arriba se va a hacer otro cuarto más grande 69 E: ah/ pues va a quedar muy bien/ [¿no?] 70 I: [como] un departamento/ yo creo/ pero eso **está planeado** para mis hermanos	Proc. Culminado Télico Transitivo PERFECTIVO
6	(E8) I: ahorita/ porque te digo/ que hay muchas cosas/ pero/ pues estaba ella y eran sólo/ sólo sus cosas// y <u>ahorita</u> pues **está planeado**// construir este/ todo un cuarto arriba/ [pues] 65 E: [mh] 66 I: ya ves que aquí está este// ampliado/ [ya se amplió]	Proc. Culminado Télico Transitivo PERFECTIVO
7	(E8) según ella/ sus planes son// quitar el/ el/ el campo/ ¿sí sabes que hay un campo de futbol? 909 E: mm 910 I: ¿no?/ bueno/ lo que pasa es que en el día **está/ tapado** 911 E: mh 912 I: los puestos y las lonas lo tapan// justo aquí a/ una/ dos calles	Culminação Télico Transitivo PERFECTIVO
8	(E8) cosa curiosa/ en/ en ningún este/ libro o/ o/ o en/ así/ oficialmente/ no está/ [no ha-/ ajá] 583 E: [no ha-/ no hay una delimitación] 584 I: o no/ no **está consignado** como/ barrio de	Proc. Culminado Télico

	Tepito o sea/ es la colonia Morelos	Transitivo PERFECTIVO
9	(E8) sabes qué es lo que venden 566 I: sí 567 E: este/ ¿cómo es esa? 568 I: ¿esa parte del barrio? 569 E: ajá 570 I: mira// eh/ el barrio// eh/ digamos que **está delimitado**/ bueno/ las calles que lo delimitan es/ el eje/ el eje uno norte/ ¿no?	Culminação Télico Transitivo PERFECTIVO
10	(E9) decir/ "oye mamá"/ este/ no por miedo al reproche de "ay mamá pues es que tú nunca me…" 74 I: o "tú/ llegas tarde 75 E: ajá 76 I: ya cuando llegas ya estoy dormido"/ no sé	Culminação --- Intransitivo PERFECTIVO

ESTAR + PARTICÍPIO NURC – RJ – 8 ENTREVISTAS		
1	1 h::... as opções... de/de diversão... entendeu?... então nós... a gente tem muita opção não sei se é porque... eu/eu moro aqui há vinte e seis anos eu **tô acostumada** mas eu também eu morei fora então eu vejo o que eu sentia muita falta também...	Proc. Culminado Transitivo --- PERFECTIVO
2	23 e tem assim, tipo um parque, maravilhoso lá, quer dizer, principalmente pra gente que a gente vê neve fica babando né (risos) aqui no Rio, <u>não</u> **tá acostumado**, né?	Proc. Culminado Transitivo --- PERFECTIVO
3	33 fila de metrô o pessoal tá numa fila realmente, não tá naquele amontoado, que nem aqui no Rio, e o cara fica parado e o Metrô pára exatamente onde você **tá parado**.	Estado Intransitivo --- PERFECTIVO
4	46 cê quer. Então cursos procurados como Eletrônica, Produção, tem gente que fica de fora, por causa do CR, entendeu, e isso é um outro problema, porque você, se você fizer no Básico não, entrou, você <u>já</u> **tá garantido** no profissional.	Culminação Transitivo Télico PERFECTIVO
5	47 se realmente, se você, se acontece um acidente de carro, que você, coisas que podem acontecer e você não tiver um, ou uma família rica ou um convênio, **tá morto**! Só de operação, de anestesia, você vai, morrer numa dessas,	Estado Intransitivo --- PERFECTIVO
6	78 to pra lavar, e que ela não vai fazer, aí ele vai fazer. DOC. - E ... bem, você falou que o, você tem um irmão casado. Ele tem filhos? LOC. - Não. Ele **tá casado** há dois anos, é mais novo do que eu até, tem vinte ... vinte e cinco, é, tem vinte e cinco.	Culminação Intransitivo --- PERFECTIVO
7	81 Agora a gente **tá brigado**, tem duas semanas que a gente nem se fala, aí, mas aí, até passar a raiva, pensar se, vale a pena voltar, não vale a pena voltar e tal	Atividade Intransitivo --- PERFECTIVO
8	85 Aí, uma maior bagunça, né a casa tá sempre cheia tem os amigos dos meus irmãos que tão sempre aqui, telefone **tá ocupado** <u>o tempo todo</u>, quando não sou eu é meu irmão, aí ... é assim.	Culminação Transitivo ---- IMPERFECTIVO
9	92 um acaso, elas não têm nenhum irmão, homem, sabe, é só irmã, então, acho que, eu acho que a parte negativa é essa, entendeu, hoje em dia, tem uma que **tá casada** e ela atura coisas	Culminação Intransitivo ---

	do marido dela que eu não aturaria de um inimigo meu, sabe, de não poder fazer isso, não quero você conversando	PERFECTIVO
10	117 eligível) LOC. - Então né, como é que a gente vai querer ser, poder criticar alguma coisa, se você não tá fazendo nada, nem por você nem pelo social, **tá parado** no teu lugar, né, até que ponto você não tá sendo instrumento, reprodutor desse sistema	Culminação Transitivo Télico PERFECTIVO
11	136 Produção, tem gente que fica de fora, por causa do CR, entendeu, e isso é um outro problema, porque você, se você fizer no Básico não, entrou, você já **tá garantido** no profissional. DOC. - E tem como o cara, vamos supor, ele terminou o básico e ...	Culminação Transitivo Télico PERFECTIVO

OCO.		**ASPECTO**
1.	(P1M3) Então uma pessoa que **tá querendo** fazer Eletrônica, vai ter Cálculo, vai ter parte de Desenho,	Estado Transitivo Atélico IMPERFECTIVO
2.	(P1M4) tá vendo, último ano, você tinha que tá estudando, agora não **tá estudando**, não sei o que!	Atividade Transitivo ---- IMPERFECTIVO
3.	(P2M4) É, nossa, é, e a gente **tá vendo** coisas que, tava há séculos né, parado	Atividade Transitivo Atélico IMPERFECTIVO
4.	(P2M5) Brizola tá fazendo coisa pra caramba, a gente tá vendo que ele **tá fazendo**.	Proc. Culminado Transitivo Télico IMPERFECTIVO
5.	(P2M6) se você não **tá fazendo** nada, nem por você nem pelo social, tá parado no teu lugar, né,	Atividade Transitivo Atélico IMPERFECTIVO
6.	(P2M7) Então é perceber o que **tá acontecendo** e tentar interferir dentro disso de alguma forma,	Atividade Transitivo Atélico IMPERFECTIVO
7.	(P2M9) Brizola tá fazendo coisa pra caramba, a gente **tá vendo** que ele tá fazendo.	Atividade Transitivo Atélico IMPERFECTIVO
8.	(P2M11) A sociedade **tá aumentando**, as pessoas tão aumentando em número, o espaço físico é o mesmo,	Processo Culm. Transitivo Télico IMPERFECTIVO

9.	(P2M13) Brizola **tá fazendo** coisa pra caramba, a gente tá vendo que ele tá fazendo.	Atividade Transitivo Atélico IMPERFECTIVO
10.	(P2M15) cê vê, **estão asfaltando**, "n" ruas do Rio, principalmente as de acesso à Barra da Tijuca, né.	Atividade Intransitivo Atélico IMPERFECTIVO
11.	(P2M16) O Grajaú. Eles **tão, asfaltando**, melhorando também.	Processo Culm. Intransitivo Télico IMPERFECTIVO
12.	(P2M17) É, pois é, eles **tão fazendo**, ali no, no Borel, o Ciep de lá,	Processo Culm. Transitivo Télico IMPERFECTIVO
13.	(P2M18) A Avenida das Américas, eles **tão recapeando** ela toda, né.	Processo Culm. Intransitivo Télico IMPERFECTIVO
14.	(P2M19) Tem outro, a própria, aquela rua, Teodoro da Silva né, eles **tão recapeando** também.	Processo Culm. Intransitivo Télico IMPERFECTIVO
15.	(P2M20) É, mas eles **tão mexendo** sabe, melhor do que ficar só parado. A Linha Vermelha, vai melhorar né, eu acho que...	Atividade Transitivo Atélico IMPERFECTIVO
16.	(P2M21) **26** pessoas tão aumentando em número, o espaço físico é o mesmo, mas isso em vez de, socializar mais as pessoas pelo contrário, **tão deixando** elas mais, agressivas.	Atividade Transitivo Atélico IMPERFECTIVO
17.	(P2M22) A sociedade tá aumentando, as pessoas **tão aumentando** em número, o espaço físico é o mesmo,	Atividade Transitivo Atélico IMPERFECTIVO
18.	(P2M23) esses comerciantes, eles tem uma um olho clínico aí pra ver o que que as pessoas tão precisando, a hora que elas **tão precisando**, vem diretamente proporcional,	Estado Transitivo Atélico IMPERFECTIVO
19.	(P2M24) eles tem uma um olho clínico aí pra ver o que que as pessoas **tão precisando**, a hora que elas tão precisando,	Atividade Transitivo Atélico

		IMPERFECTIVO
20.	(P13M1) porque quando eu estou em casa eu **estou fazendo** alguma coisa	Processo Culm. Transitivo Télico IMPERFECTIVO
21.	(P13M2) o tempo que eu tiver podendo ouvir também **estou ouvindo**...não tenho preferência por tipo de música...	Atividade Transitivo Atélico IMPERFECTIVO
22.	(P23M1) tenho a minha avaliação dele... mais até do que quando eu **estou jogando**...	Atividade Transitivo Atélico IMPERFECTIVO
23.	(P23M2) ou pelo menos tenho sempre uma opinião a dar quando eu **estou vendo** o jogo... tenho a minha avaliação dele...	Processo Culm. Transitivo Télico IMPERFECTIVO
24.	(P3F1) essas duas que **eu tô falando**, por um acaso, elas não têm nenhum irmão, homem, sabe, é só irmã,	Proc. Culminado Intransitivo Télico IMPERFECTIVO
25.	(P3F2) Depende, entendeu, porque eu **tô namorando** há quatro anos e meio, e aí fica meio dependente de namorado,	Atividade Intransitivo Atélico IMPERFECTIVO
26.	(P3F3) Eu acho que, eu acho que existe, cobrança, por exemplo, cobram, se você **tá namorando** há muito tempo, cobram, que você tem que casar,	Atividade Intransitivo Atélico IMPERFECTIVO
27.	(P3F4) as minhas amigas que, achavam ridículo igreja, véu e grinalda, hoje em dia, elas **tão casando** de véu e grinalda,	Culminação Intransitivo Atélico IMPERFECTIVO
28.	(P12F1) Que isso, isso nunca aconteceu no Brasil, não **tô entendendo** isso.	Atividade Transitivo Atélico IMPERFECTIVO
29.	(P12F2) é um problema, muito sério e você, hoje em dia, dirigindo um carro, você, eu, pelo menos, sou assim, paro no sinal, você não **tá, passeando** pela cidade, você não consegue passear.	Atividade Transitivo Atélico IMPERFECTIVO
30.	(P12F3) aqui no Rio a maioria das pessoas mora em prédios né, São Paulo, geralmente o pessoal mora em casa, ah, **tá mudando** um pouco, mas, é, São Paulo é uma cidade mais, esparramada,	Culminação Intransitivo

		Télico IMPERFECTIVO
31.	(P12F4) Pô, tá velho, o cara não **tá cuidando**! (risos) quer dizer, os caras pensam em tudo,	Atividade Transitivo Atélico IMPERFECTIVO
32.	(P12F5) Cê pára no sinal, você presta atenção se tem pivete, se tem alguém olhando meio suspeito, se você **tá sendo** seguido por algum carro, entendeu, então...	Estado Transitivo --- IMPERFECTIVO
33.	(P12F6) É, é meio acomodado. Pô tá tudo bem, não tô nem aí, mas, **tá mudando**!	Culminação Intransitivo ---- IMPERFECTIVO
34.	(P12F7) Ah essa loja tá vendendo mais do que a gente, essa outra **tá vendendo** mais,	Atividade Transitivo --- IMPERFECTIVO
35.	(P12F8) Ah essa loja **tá vendendo** mais do que a gente, essa outra tá vendendo mais,	Atividade Transitivo --- IMPERFECTIVO
36.	(P12F9) Porque eu não tava sabendo, você não **tá convivendo** o dia-a-dia do país, então, você olhava aquela notícia, você fica um pouco chocado.	Processo Culm. Transitivo Télico IMPERFECTIVO
37.	(P12F10) quando você vai num esquema desse de competição, você não **tá vendo** exatamente o dia-a-dia da cidade,	Processo Culm. Transitivo Télico IMPERFECTIVO
38.	(P15F1) a Denise a respeito da injustiça das mudanças que **estão havendo** nas leis né... e o Paulo Goulart falou sobre família...	Processo Culm. Transitivo Télico IMPERFECTIVO
39.	(P25F1) gente que... poderia muito bem morar na Avenida Atlântica e **tá morando** em Bonsucesso... entendeu?	Culminação Transitivo --- IMPERFECTIVO
40.	(P25F2) é contra o que eu digo mas eu acho que ::ele é um cara que **tá tentando** entendeu?...	Processo Culm. Transitivo Télico IMPERFECTIVO
41.	(P25F3) Barra é um/ é um::... é um bairro que tem que tomar muito cuidado ele **tá crescendo**... né?	Processo Culm.

		Intransitivo Télico IMPERFECTIVO
42.	(P25F4) Barra é um/ é um::... é um bairro que tem que tomar muito cuidado ele **tá crescendo**... né?	Processo Culm. Intransitivo Télico IMPERFECTIVO
43.	(P25F5) poderia muito bem morar na Avenida Atlântica e **tá morando** em Bonsucesso... entendeu?	Culminação Transitivo --- IMPERFECTIVO
44.	(P25F6) ele não tá tentando destruir a cidade ele tá tentando melhorar como nenhum outro fez né? ele **tá tentando**... né ele é um cara corajoso...	Processo Culm. Transitivo --- IMPERFECTIVO
45.	(P25F7) ele tá tentando:: modificar e::... ele **tá tentando**::... éh::... a/éh/agradar mesmo o povo carioca...	Processo Culm. Transitivo --- IMPERFECTIVO
46.	(P25F8) ele **tá tentando**:: modificar e::... ele tá tentando::... éh::... a/éh/agradar mesmo o povo carioca...	Processo Culm. Transitivo --- IMPERFECTIVO
47.	(P25F9) ele faz tudo meio atrapalhado mas eu... eu acho que ele... ele **tá tentando** né...	Processo Culm. Transitivo --- IMPERFECTIVO
48.	(P25F10) ele não tá tentando destruir a cidade ele **tá tentando** melhorar como nenhum outro fez né?	Processo Culm. Transitivo --- IMPERFECTIVO
49.	(P25F11) ele não **tá tentando** destruir a cidade ele tá tentando melhorar como nenhum outro fez né?	Processo Culm. Transitivo --- IMPERFECTIVO
50.	(P25F12) já te/te... **tão construindo** muitos prédios no Recreio então quer dizer...	Processo Culm. Transitivo Télico IMPERFECTIVO
51.	(P25F13) já te/te... **tão construindo** muitos prédios no Recreio então... quer dizer... éh::...	Processo Culm. Transitivo Télico IMPERFECTIVO

1.	(E3M1) me creerás que ultim-/ ahorita estoy es-/ lo **estoy estudiando**// y / a mí se me hace padre	Atividade Transitivo --- IMPERFECTIVO
2.	(E3M2tú háblame güey// aunque sepas que **estoy trabajando**/ tú háblame el fin de semana//	Atividade Intransitivo --- IMPERFECTIVO
3.	(E3M3) o sea la gente que ya no quiero que/ tener cerca de mí/ o sea / yo no **estoy pensando** en qué estará haciendo esa persona/	Atividade Intransitivo --- IMPERFECTIVO
4.	(E3M5)¡sí!/ pero si te digo que lo **estoy estudiando** desde/ que tengo/ cinco años	Atividade Transitivo --- IMPERFECTIVO
5.	(E3M6) o sea **estoy haciendo** el curso/ para hacer el Toefl // pero/ pero/ ¿cuándo empiezo?/ mañana/ mañana [empiezo]	Processo Culm. Transitivo Télico IMPERFECTIVO
6.	(E3M7)¡no!/ pues es que/ digo/ he estudiado toda la vida// y ahorita más bien **estoy haciendo** como el// curso para el Toefl	Processo Culm. Transitivo Télico IMPERFECTIVO
7.	(E3M8) [entre] ellos/ también te lo comenté hace rato/ **estoy aprendiendo** inglés	Processo Culm. Transitivo Télico IMPERFECTIVO
8.	(E3M9) ya terminé mi carrera// pero eh/ yo terminé t-/ t-/ obtuve mi certificado// ya te **estoy contando** toda la historia	Processo Culm. Transitivo Télico IMPERFECTIVO
9.	(E3M10) **estoy radicando** que es este/ te digo que en el pueblo de San Mateo	Atividade Intransitivo --- IMPERFECTIVO
10.	(E3M11) eh/ yo después/ este// del ochenta y cinco me vine a vivir ya a lo que/ donde **estoy viviendo** ahorita/ [donde]	Estado Intransitivo --- IMPERFECTIVO
11.	(E3M12) apenas voy a sacar mi título/ **estoy haciendo** mis trámites para el título/ ya tengo todos mis papeles/	Processo Culm. Transitivo Télico

		IMPERFECTIVO
12.	(E3M14) pues sí/ pero bueno/ nunca es tarde/ ¿no? I: no/ o sea/ lo **estoy haciendo**/ por eso te digo o sea// me arrepiento de una cosa me/	Processo Culm. Transitivo Télico IMPERFECTIVO
13.	(E3M15) sí/ es lo que **estoy haciendo** ahorita// ya tengo todo// a ver/ vamos a ver si no falta nada/ nada lo entrego//	Processo Culm. Transitivo Télico IMPERFECTIVO
14.	(E3M16) o hacer algo y/ pues más o menos te enteras de/ qué está diciendo// y y más/ si ya **estás estudiando** inglés/	Processo Culm. Transitivo Télico IMPERFECTIVO
15.	(E3M17)¡sí/ a fuerzas! y aprovéchenlo/ tú// ¿no **estás estudiando** entonces inglés?	Processo Culm. Transitivo Télico IMPERFECTIVO
16.	(E3M18) sea escuela/ trabajo/ tu barrio// donde caigas/ o sea/ conoces gente/ si no te gusta la gente que **estás tratando**/ pues simplemente te aparta	Processo Culm. Transitivo Télico IMPERFECTIVO
17.	(E3M19) "ay"/ no/ o sea/ tengo interés// por aprenderlo y/ aparte me **está gustando**/ o sea ya/ ya le agarré gusto 687	Estado Transitivo --- IMPERFECTIVO
18.	(E3M20) vas viendo la película/ y más o menos como que/ cuando ves a alguien platicar// o/ o hacer algo y/ pues más o menos te enteras de/ qué **está diciendo**// y y más/ si ya estás estudiando inglés/	Atividade Transitivo --- IMPERFECTIVO
19.	(E3M21) carpetas/ o sea todo lo que es artículo/ de hecho/ ahorita es- ya se **está metiendo** lo que es este/ artículos de/ computación//	Atividade Transitivo Atélico IMPERFECTIVO
20.	(E3M23) entonces/ ella se dedica a lo que es fiscal// y / pero igual o sea el/ la conocí porque / **estamos estudiando**/ inglés	Processo Culm. Transitivo Télico IMPERFECTIVO
21.	(E3M25) me lo **están diciendo** con el afán de E: sí/ no/ [no es por] I: [de ayudarme] E: por molestarte	Atividade Transitivo Atélico IMPERFECTIVO
22.	(E3M26) todo ese tipo de cosas ya/ se **están introduciendo**/ est- esta empresa maneja/ varias marcas	Processo Culm. Transitivo

		Télico IMPERFECTIVO
23.	(E8F1) yo creo que es algo así/ es un fenómeno o algo así/ que se **está viendo** <u>últimamente</u> porque/ antes sí// sí siempre ha sido peligroso el barrio/	Estado Transitivo Atélico IMPERFECTIVO
24.	(E8F3) pues <u>ahorita</u> sí/ se **está/ usando** así/ porque hay un montón de triques ahí	Atividade Transitivo --- IMPERFECTIVO
25.	(E8F4) después pues ya que/ te vuelves más consciente/ te das cuenta de es lo que **está sucediendo**// pero de/ a últimas fechas/ bueno de unos años para acá/	Processo Culm. Transitivo Télico IMPERFECTIVO
26.	(E8F5) no/ no/ es es lo que te digo// m-/ los/ o sea los/ los medios de comunicación/ no te van a decir la verdad// en primer lugar/ además algo **están pensando**/ porque bueno/ cómo te explicas	Atividade Transitivo Atélico IMPERFECTIVO
27.	(E8F6) pues n-/ mm/ tan truculento que no me imagino/ pero hay algo así muy// muy oscuro/ muy negro ¿no?/ que/ que **están pensando** en hacer aquí// ¿qué qué qué podría ser pues?	Atividade Transitivo Atélico IMPERFECTIVO
28.	(E8F7) sí/ me platicaba un amigo/ un ex vecino/ que ya no vive aquí// que una vez/ su tío/ venía a visitarlos/ ¿no?/ el tío era joven/ **estamos hablando** yo/ qué te diré/ del cincuenta/ sesenta	Atividade Transitivo Atélico IMPERFECTIVO
29.	(E9F1) porque sí hay momentos que llegas a ese grado// pero digo "no/ o sea yo decidí estudiar esto/ me gusta/ lo **estoy disfrutando**/ y creo que es lo que va a hacer que/ que valore todavía más"/ ¿ya se acabó?	Processo Culm. Transitivo Télico IMPERFECTIVO
30.	(E9F2) ve que me subo al autobús// me voy// aparte de del aspecto aca-/ aspecto académico que he aprendido <u>muchísimas cosas</u>/ que **estoy aprendiendo**/// este// trato de aprovecharlo// porque// aprendo/ de// de las vivencias de/ de mis compañeros/ de mis compañeras///	Processo Culm. Transitivo Télico IMPERFECTIVO
31.	(E9F3) le digo/ "bueno"/ como quiera/ lo que me hace sentir bien/ es que cosas malas/ a lo mejor <u>no</u> les **estoy dejando** E: claro	Culminação Transitivo Télico IMPERFECTIVO
32.	(E9F4) como que todo eso me ha servido/ y hace que/ que yo no desista de mis objetivos y hace de que yo no/ hace/ que yo no desista de / de lo que **estoy viviendo** y experimentando como pareja/ entonces/ eso hace/ eso hace también que me motive y no desista/ del aspecto escolar//	Processo Culm. Intransitivo Télico IMPERFECTIVO
33.	(E9F5) si es/ si es lo que/ no quieres"/ un ejemplo no/ [<u>no</u> **estoy**	Atividade

	diciendo que/ que así sea mi pareja	Transitivo Atélico IMPERFECTIVO
34.	(E9F6) en mi cabeza nunca pasó la idea de tener novio/// ni mucho menos vivir con alguien// sucedió/ tuve novio// **estoy viviendo** con una persona// ahorita estoy tranquila/	Processo Culm. Intransitivo Télico IMPERFECTIVO
35.	(E9F7) ciertas notas para su trabajo// yo me siento y estoy estudiando// **estoy leyendo**// o luego le digo "¿cómo ves?/	Atividade Transitivo --- IMPERFECTIVO
36.	(E9F8) elaborar// ciertas notas para su trabajo// yo me siento y **estoy estudiando**// estoy leyendo// o luego le digo	Atividade Transitivo --- IMPERFECTIVO
37.	(E9F9) **145** hijas o mis hijos" 157 E: mh I: creo que no/ que es lo más tonto e inmaduro/ ¿no?/ que puedo hacer/ porque ni estás resolviendo el problema/ y les **estás provocando** conflictos emocionales a tus hijos/ si no es que ya están dañados E: claro I: entonces este// digo/ "no/	Atividade Transitivo Atélico IMPERFECTIVO
38.	(E9F10) como adultos/ aparte de que te comparten conocimiento/ de la carrera que **estás estudiando**// te comparten conocimientos de sus otras áreas que hay	Processo Culm. Transitivo Télico IMPERFECTIVO
39.	(E9F11) como profesionista/ porque no estás este/ limitando esa parte// y aparte **estás confiando** en mí	Processo Culm. Transitivo Télico IMPERFECTIVO
40.	(E9F12) creo que no/ que es lo más tonto e inmaduro/ ¿no?/ que puedo hacer/ porque ni **estás resolviendo** el problema/ y les estás provocando conflictos emocionales a tus hijos	Processo Culm. Transitivo Télico IMPERFECTIVO
41.	(E9F13) que te toca ver/ escuchar y dices// "¿cómo es posible?"// se **está demostrando** ¿no?/ que son personas egoístas/ porque ni a él como padre/	Atividade Transitivo Atélico IMPERFECTIVO
42.	(E9F14) te puede hacer bromas/ y que tú le dices "¿sabes qué?/ que **estamos organizando** una comida/ vamos a salir/ no sé/ ¿vas?"/	Processo Culm. Transitivo Télico IMPERFECTIVO

Artigo recebido em: 15.10.2014
Artigo aprovado em: 14.11.2014

O *corpus* combinado e a pesquisa nos Estudos da tradução baseados em *corpora*

The use of bidirectional parallel *corpus* within *Corpus* based translation studies

Silvana Maria de Jesus[1]

RESUMO: Este artigo aborda as relações de tradução de SAY/DIZER em textos ficcionais no par linguístico inglês-português. Adotando uma perspectiva empírica de observação de dados em *corpus* combinado, este trabalho aborda os tipos de *corpora* e suas características, detalhando o *corpus* combinado e os procedimentos metodológicos ao se utilizar de dois programas computacionais – *WordSmith Tools* e SPSS. O *corpus* combinado apresentado é composto de três romances originais em inglês e suas traduções para o português e três romances originais em português e suas traduções para o inglês, e faz parte do CORDIALL (*Corpus* Discursivo para Análises Linguísticas e Literárias) desenvolvido pelos pesquisadores do LETRA (Laboratório Experimental de Tradução) da Faculdade de Letras da UFMG. Os programas *WordSmith Tools* e SPSS (*Statistical Package for the Social Sciences*) são utilizados para a extração de dados quantitativos para a construção de um banco de dados. Os resultados apresentam os equivalentes possíveis de SAY/DIZER, segundo as ocorrências do *corpus*, detectando-se padrões distintos nas relações de tradução, conforme a direção, seja do inglês para o português ou do português para o inglês.

PALAVRAS-CHAVE: Estudos da tradução baseados em *corpora*. *Corpus* combinado. Relações de tradução. Linguística de *Corpus*.

ABSTRACT: This work reports on a study developed at LETRA - Laboratory for Experimentation in Translation, Faculdade de Letras, UFMG. Building on corpus based translation studies, it examines equivalence relations between two verbs in the pair English-Portuguese, SAY/DIZER. Equivalence is looked at from an empirical perspective drawing on data gathered from translated and non-translated fiction in both languages and in both translation directions. Equivalence relations are studied applying Corpus Linguistics methodology for quantitative data analysis, using two software – *WordSmith Tools* and SPSS (*Statistical Package for the Social Sciences)* – and a bidirectional parallel corpus (comparable and parallel) compiled to that end. It investigates the originals and their translations to search and account for possible equivalents. Results point to different patterns of these verbs, relating to their functions and to equivalence relations.

KEYWORDS: Corpus based translation studies. Bidirectional parallel corpus. Equivalence relations. Corpus Linguistics.

[1] Professora Doutora do Curso de Tradução do Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia (UFU).

1.Introdução

O uso de *corpora* de textos traduzidos em pesquisas de tradução tem como base os trabalhos de Baker (1993, 1995, 1996), que por sua vez, têm como base os trabalhos de Toury (1995) com sua proposta de Estudos descritivos da tradução. Baker introduziu uma terminologia para os tipos de *corpus*, que se tornou bastante difundida, embora tenha recebido críticas e reformulações. Os trabalhos de Baker também são importantes pela abordagem que propõem – a busca de universais de tradução – que se constituiu em importante marco inicial para a investigação dos textos traduzidos.

Este artigo não pretende fazer um histórico sobre a Linguística de *Corpus*, tampouco sobre os Estudos da Tradução baseados em *corpora*, assuntos sobre os quais o leitor encontrará ampla discussão em Berber-Sardinha (2004) e no periódico *Cadernos de Tradução IX* (TAGNIN, 2002), um número especial sobre Tradução e *Corpora*.

O objetivo deste trabalho é trazer uma discussão sobre a terminologia utilizada para os tipos de *corpora* e introduzir um tipo de *corpus* pouco utilizado, o *corpus* combinado; pretende-se ainda mostrar como os dados brutos (*raw data*) do *corpus* oferecem material de análise ao pesquisador, a partir dos quais análises mais refinadas podem ser desenvolvidas, bem como descrever os procedimentos metodológicos para a obtenção de dados, destacando-se os recursos informáticos utilizados, especialmente os software *WordSmith* e SPSS.

Inicia-se, portanto, este artigo, com a discussão sobre os tipos de *corpora* e a definição de *corpus* combinado. Em seguida, descrevem-se os procedimentos metodológicos para a obtenção de dados do *corpus* através de programas computacionais, observando-se como é possível formular perguntas de pesquisa a partir destes dados brutos e como estas perguntas iniciais podem nortear o refinamento da análise. Por fim, para ilustrar a metodologia, desenvolve-se uma pequena análise das relações de tradução entre SAY e DIZER, a partir dos dados de um *corpus* combinado.

2. Os diferentes tipos de *corpora*

Desde os trabalhos seminais de Baker (1993, 1995, 1996), que marcaram o início da interface entre os Estudos da Tradução e a Linguística de *Corpus*, os tipos de *corpora* utilizados nas pesquisas têm se diversificado de forma a contemplar diversos aspectos da descrição e comparação de línguas.

Inicialmente, foram definidos por Baker (1995:230) três tipos: i) *corpora* paralelos – textos originais em língua A e suas traduções em língua B; ii) *corpora* multilíngues – textos não-traduzidos compilados a partir de critérios comuns, como por exemplo, o gênero do discurso, nas línguas A e B; iii) *corpora* comparáveis – textos originalmente produzidos em uma língua A e textos traduzidos para essa mesma língua A, ou seja, um *corpus* monolíngue, formado de textos não-traduzidos e textos traduzidos.

O termo *comparável*, utilizado por Baker para se definir *corpora* de textos traduzidos e não-traduzidos, pode ser usado como um termo geral, no sentido comum de se comparar duas ou mais características, que podem ser textos traduzidos x não-traduzidos, língua A x língua B, ou ainda gêneros diferentes, como textos acadêmicos x textos midiáticos, o que pode gerar certa confusão terminológica. Nas pesquisas linguísticas, fora do âmbito da tradução, o termo é usado para se nomear um *corpus* bilíngue ou multilíngue de textos não-traduzidos em que aspectos de diferentes línguas estão sendo comparados.

Nos Estudos da tradução, tem-se optado pelos termos *corpus comparável bilíngue* (ou multilíngue) e *corpus comparável monolíngue* para se distinguir entre a comparação de textos não-traduzidos em línguas diferentes e textos traduzidos e não-traduzidos em uma mesma língua, respectivamente (Olohan, 2004:34). Hansen e Teich (2001), em uma análise contrastiva do inglês e do alemão, utilizam os termos *textos comparáveis monolíngues*, para o *corpus* de textos traduzidos e não-traduzidos em alemão, *textos paralelos*, para o *corpus* de textos originais em inglês e textos traduzidos em alemão, e *textos comparáveis multilíngues*, para o *corpus* de textos não-traduzidos em inglês e em alemão.

Observam-se, ainda, os tipos de *corpora* citados por Ana Frankenberg-Garcia (2006), considerados pela autora como relevantes para os estudos da tradução: *corpora* comparáveis bilíngues, *corpora* comparáveis monolíngues, *corpora* paralelos unidirecionais e *corpora* paralelos bidirecionais. A nomenclatura dos tipos de *corpora* por si só não define as características e finalidades dos mesmos, sendo então explicitados pela autora.

Corpora comparáveis bilíngues são formados de um sub*corpus* de textos na língua A e um sub*corpus* de textos na língua B, não-traduzidos, não-paralelos, semelhantes em gênero e função e podem ser usados para extração de terminologia e no dia-a-dia de tradutores. *Corpora comparáveis monolíngues* são formados de um sub*corpus* de textos não-traduzidos e um sub*corpus* de textos traduzidos, ambos em língua A, não-paralelos, semelhantes em gênero e função e podem ser usados para se observarem as diferenças entre a linguagem

traduzida e a não-traduzida e para estudos teóricos sobre a tradução. *Corpora paralelos unidirecionais* são formados de um sub*corpus* de textos originais em língua A e um sub*corpus* com os textos traduzidos para a língua B, são paralelos e semelhantes em gênero e função e podem ser usados para a pesquisa de equivalentes tradutórios, como suporte para a criação de dicionários bilíngues e para a tradução automática. *Corpora paralelos bidirecionais* são *corpora* bastante complexos e merecem uma explicação mais detalhada.

Os *corpora paralelos bidirecionais*, segundo Frankenberg-Garcia (2006), são formados de quatro sub*corpora* que, combinados entre si, podem formar outros *corpora*. Os quatro sub*corpora* se compõem em: 1) textos originais na língua A e 2) as suas traduções para a língua B, 3) textos originais na língua B e 4) as suas traduções para a língua A.

A autora explica que estes *corpora* formam dois *corpora* comparáveis monolíngues (1 e 4, traduzidos e não traduzidos na língua A, e 2 e 3, traduzidos e não traduzidos na língua B) e dois *corpora* comparáveis bilíngues (1 e 3, não-traduzidos na língua A e não-traduzidos na língua B, e 2 e 4, traduzidos na língua A e traduzidos na língua B).

Este tipo de *corpus* é denominado por Vela e Hansen-Schirra (2006) de *corpus* combinado paralelo-comparável (*combined parallel-comparable corpus*), ou simplesmente, *corpus combinado* e apesar de sua complexidade tem-se mostrado mais produtivo para as pesquisas em tradução. Hansen (2002:20) utiliza o termo *corpus* comparável e paralelo (*comparable and parallel corpus*) e descreve dois tipos de *corpus* combinado: o bilíngue e o multilíngue (mais de duas línguas).

Um dos estudos pioneiros na utilização de um *corpus* combinado, embora o autor não tenha utilizado esta denominação, é o de Johansson (1998). O autor utiliza um *corpus* formado de textos originais em inglês e suas traduções para o norueguês e textos originais em norueguês e suas traduções para o inglês. Johansson sugere a utilização dos seguintes termos para a denominação dos diferentes tipos de *corpora*: i) *corpora* comparáveis bilíngues/multilingues ou *corpora* de textos originais comparáveis, ii) *corpora* de tradução ou *corpora* de textos originais e suas traduções, podendo ser bilíngue (original na língua A e tradução na língua B) ou multilingue (original na língua A e traduções nas línguas B, C, D), e iii) *corpora* comparáveis monolíngues ou *corpora* de textos originais e traduzidos na mesma língua.

Observando-se as classificações apresentadas, nota-se que os *corpora*, que geralmente são formados por distintos sub*corpora*, são classificados segundo três aspectos: i) a(s) língua(s) envolvida(s), ii) o status do texto, e iii) a direcionalidade.

Em relação à língua, os *corpora* podem ser monolíngues (somente textos em língua A), bilíngues (textos em língua A e B), ou multilíngues (textos em língua, A, B, C... n).

O status classifica o texto na sua relação com outros textos, podendo ser considerado como original, tradução e texto não-traduzido. Destaca-se aqui que este status não se refere ao texto em si, mas à sua posição em relação a outro(s) texto(s) do *corpus*. Um texto será considerado como original se ele estiver em relação de tradução com outro(s) texto(s), formando assim um *corpus* paralelo bilíngue (ou multilíngue) com original e tradução (ou traduções). Este mesmo texto pode ser classificado como texto não-traduzido se ele estiver sendo analisado em relação a um texto na mesma língua que é um texto traduzido, formando assim, um *corpus* comparável monolíngue com texto(s) não-traduzido(s) em língua A e texto(s) traduzido(s) em língua A.

O termo *original* é visto por alguns teóricos, principalmente aqueles voltados para a interface entre os Estudos da tradução e os Estudos culturais, como inadequado, por pressupor aspectos relativos às questões de hierarquia. Entretanto, opta-se por utilizar os termos *original/tradução* para se destacar que são textos que estão em relação de tradução, sendo os termos *texto traduzido*/*texto não-traduzido* utilizados para se referir aos textos que não estão em relação de tradução entre si.

Aspectos discursivos também podem ser apontados no status, como tipo textual, registro ou gênero textual. O *corpus* pode ser compilado de forma a contrastar as características de um gênero textual em duas línguas ou de gêneros textuais diferentes em uma mesma língua.

O aspecto direcionalidade indica se a análise é feita apenas na direção da língua A para a língua B ou se ela considera também a direção inversa, ou seja, da língua B para a língua A. Pode, portanto, apresentar-se como unidirecional, quando, por exemplo, compara-se uma dada característica em textos originais na língua A e em suas traduções na língua B; ou bidirecional, quando esta análise é ampliada com a comparação desta mesma característica em textos originais na língua B e em textos traduzidos para a língua A, ou seja, quando se utiliza um *corpus* combinado.

Estes tipos básicos podem ser combinados conforme as necessidades de pesquisa e, como não há ainda uma terminologia consagrada para cada aspecto, cabe ao pesquisador explicitar as diversas características do *corpus* utilizado.

Outros aspectos relacionados à tipologia de *corpus*, como modo (falado ou escrito), tempo (sincrônico ou diacrônico) e outros, são abordados por Berber Sardinha (2004:20). A título de ilustração de diferentes tipos de *corpora*, podem-se observar alguns *corpora* em desenvolvimento nos Estudos da tradução, como por exemplo, o TEC, o CORDIALL e o CroCo.[2] Remete-se ainda o leitor ao trabalho de Dayrell (2005:48) para um levantamento de *corpora* existentes em português.

3. O *corpus* combinado

O *corpus* combinado apresentado nesse artigo foi usado por Jesus (2008), sendo parte do Projeto CORDIALL (*Corpus* discursivo para análises linguísticas e literárias, desenvolvido pelos pesquisadores do LETRA – Laboratório Experimental de Tradução da Faculdade de Letras da UFMG)[3]. É composto por doze romances, somando um total de 1.237.970 palavras. Caracteriza-se como um *corpus* combinado que, como foi dito, é composto de vários *corpora* básicos, cujas combinações permitem a composição de vários *corpora* de análise. Os quatro *corpora* básicos desta pesquisa são[4]:

1- *Corpus* PO - para português original ou não-traduzido
2- *Corpus* IT - para inglês traduzido
3- *Corpus* IO - para inglês original ou não-traduzido
4- *Corpus* PT - para português traduzido

Neste *corpus* os sub*corpora* 1-2 e 3-4 são paralelos, ou seja, no sub*corpus* 2 temos as traduções dos textos que compõem o sub*corpus* 1 e no sub*corpus* 4 as traduções dos textos que compõem o sub*corpus* 3. Em outra combinação, os sub*corpora* 1-4 e 2-3 são comparáveis monolíngues, ou seja, são formados de textos traduzidos e não-traduzidos em português e em

[2] TEC (http://www.monabaker.com/tsresources/TranslationalEnglishCorpus.htm), CORDIALL (http://letra.letras.ufmg.br/letra/index.xml) e CroCo (http://fr46.uni-saarland.de/croco/index_en.html).
[3] Site do LETRA - http://letra.letras.ufmg.br/letra/
[4] As siglas IO, IT, PO, PT foram utilizadas por Tagnin (2002a:203) na construção do corpus para o Projeto terminológico para tradutores. A opção de utilizá-las na presente pesquisa deve-se ao fato de que as siglas facilitam a visualização do tipo de corpus, destacando-se ainda que os textos não-traduzidos utilizados no corpus desta pesquisa são também os textos originais.

inglês, respectivamente. Há ainda um *corpus* comparável bilíngue formado por 1-3, com textos não-traduzidos em português e inglês.

Os textos foram utilizados anteriormente por pesquisadores do LETRA, oferecendo, portanto, a vantagem de já se encontrarem digitalizados. A compilação do *corpus* obedeceu aos seguintes critérios: romances disponíveis no CORDIALL, no par linguístico português-inglês, publicados no século XX. Foram selecionados três romances para cada um dos sub*corpora*. O Quadro 1 apresenta os dados bibliográficos de cada romance e os *corpora* que eles compõem.[5]

Quadro 1 – Dados bibliográficos dos romances do *corpus* combinado

Corpora	Título	Sigla	Autor/Tradutor	1ª edição	Edição utilizada
IO	*Point counter point*	PCP	Aldous Huxley	1928	1994
	Interview with the vampire	IWV	Anne Rice	1976	1997
	Beloved	BEL	Toni Morrison	1987	1998
IT	*Macunaíma*	MAC	E A Goodland	1984	1984
	Gabriela, clove and cinammon	GAB	James L Taylor e William Grossman	1962	1962
	The hour of the star	THS	Giovanni Pontiero	1986	1992
PO	*Macunaíma: o herói sem caráter*	MHS	Mário de Andrade	1928	1980
	Gabriela, cravo e canela	GCC	Jorge Amado	1958	1958
	A hora da estrela	AHE	Clarice Lispector	1977	1998
PT	*Contraponto*	CPO	Érico Veríssimo	1934	1971
	Entrevista com o vampiro	ECV	Clarice Lispector	1977	1996
	Amada	AMA	Evelyn Kay Massaro	1987	1987

[5] Para maiores detalhes sobre os corpora, remete-se o leitor aos trabalhos de Assis (2004) para os romances *Beloved* e *Amada*; Jesus (2004) para os romances *Point counter point* e *Contraponto*; Bueno (2005) para os romances *Macunaíma, o herói sem caráter* e *Macunaíma*; Cançado (2005) para os romances *Interview with the vampire* e *Entrevista com o vampiro*; Morinaka (2005) para os romances *Gabriela, cravo e canela* e *Gabriela, clove and cinammon*; e Rodrigues (2005) para os romances *A hora da estrela* e *The hour of the star*.

Estes sub*corpora* podem ser combinados, segundo os objetivos de análise, formando outros *corpora* para a investigação de aspectos específicos:

i) *corpus paralelo IO-PT* – formado de textos originais em inglês (IO) e suas traduções para o português (PT)
ii) *corpus paralelo PO-IT* – formado de textos originais em português (PO) e suas traduções para o inglês (IT)
iii) *corpus comparável monolíngue IO-IT*– formado de textos não-traduzidos em inglês (IO) e textos traduzidos em inglês (IT)
iv) *corpus comparável monolíngue PO-PT* – formado de textos não-traduzidos em português (PO) e textos traduzidos em português (PT)
v) *corpus comparável bilíngue IO-PO* – formado de textos não-traduzidos em inglês (IO) e textos não-traduzidos em português (PO)

Em relação aos três aspectos que são observados em um *corpus*, quais sejam, língua(s), status e direcionalidade, este *corpus* caracteriza-se como bilíngue, composto por textos do par linguístico português-inglês; combinado, ou seja, contendo textos em ambas as relações – traduzidos/não-traduzidos e traduzidos/originais; e bidirecional, visto que a análise pode ser feita tanto na direção do português para o inglês quanto do inglês para o português. A próxima seção apresenta os procedimentos metodológicos para obtenção dos dados com o uso de programas computacionais.

4. Procedimentos metodológicos

O trabalho com *corpus* exige do pesquisador o uso de ferramentas computacionais para a obtenção de dados quantitativos. Esta seção introduz dois programas que podem ser usados para análises linguísticas, e os procedimentos metodológicos utilizados, ilustrando como as perguntas de pesquisa vão sendo desenvolvidas concomitantemente com a análise do *corpus*.

4.1 O programa *WordSmith Tools* e os dados brutos

Após a compilação do *corpus* da pesquisa, o tratamento do *corpus* é feito com a combinação de procedimentos manuais por parte do pesquisador e automáticos com a utilização de programas computacionais.

O software *WordSmith Tools*[6] tem sido bastante utilizado em pesquisas com *corpus*. Para maiores detalhes sobre o programa remeto o leitor a Jesus (2004), Berber Sardinha (2004), Alves e Morinaka (2004) e ao site do programa[7].

Com a ferramenta WordList é possível obter os seguintes dados do *corpus*: i) o número total de vocábulos (*types*) e o número total de ocorrências (*tokens*) do *corpus* e também de cada romance e/ou sub*corpus*; e ii) o número de ocorrências das formas de SAY/DIZER no *corpus*. A partir das formas encontradas na lista de palavras fornecida pela ferramenta Wordlist, buscam-se todas as ocorrências de SAY/DIZER com a ferramenta Concord.

A ferramenta Concord permite a busca de todas as formas de SAY/DIZER colocando-se como nódulo, que é a palavra de busca, a forma say*/said para SAY e disse*/diz*/diria*/diga*/digo/dito/ditas/dir-se-ia para DIZER. A definição das formas a serem utilizadas como nódulo foi feita através da observação das listas de palavras obtidas com a ferramenta WordList. O asterisco (*) indica ao programa que se deve buscar todas as palavras que iniciem pelas letras antes do asterisco, assim, no caso de say* o programa busca também as formas *saying* e *says*. No caso de DIZER, quando se usa disse* buscam também as formas *disseram*, *dissemos* e assim por diante. Neste processo, o programa busca também palavras que não são formas de SAY/DIZER, como por exemplo, *dizimados* para a forma diz*, sendo necessário checar as linhas de concordância e eliminar estes casos. Para se eliminar as linhas, utiliza-se o comando deletar e o botão ZAP.

A Tabela 1 apresenta os dados de cada sub*corpora* do *corpus* combinado.

[6] Uma outra opção é o programa AntConc, que não possui as mesmas funcionalidades, mas é de uso gratuito http://www.laurenceanthony.net/antconc_index.html .
[7] Mike Scott's Web http://www.lexically.net/wordsmith/ .

Tabela 1 – Dados dos quatro *corpora* básicos do *corpus* combinado

	IO	**IT**	**PO**	**PT**
Types	29.392	22.594	26.336	40.434
Tokens	402.525	234.599	203.504	397.342
Total de ocorrências de SAY/DIZER	2.723	776	707	2.385
% de SAY/DIZER x tokens	0.67	0.33	0.34	0.60

O tamanho do *corpus* é um aspecto bastante discutido na sua compilação (Olohan, 2004:45), com uma tendência, nos Estudos da tradução baseados em *corpora*, para o uso de *corpora* de mesmo tamanho (em relação ao número de textos ou de *tokens*). Entretanto, o *corpus* combinado aqui apresentado é formado de romances inteiros, originais e traduções, o que dificulta a padronização do tamanho. Além disso, serão analisadas 150 ocorrências de cada sub*corpus*, não sendo, portanto, relevante a questão do tamanho do *corpus*.

Os dados brutos já permitem algumas considerações para análise:

i) *corpus* comparável bilíngue IO-PO

Há 2.723 ocorrências de SAY em IO e 707 ocorrências de DIZER em PO.

Logo, a ocorrência de SAY em textos não-traduzidos (IO) é significativamente *maior* do que a ocorrência de DIZER em textos não-traduzidos (PO); SAY representa 0.67% de tokens, ao passo que DIZER representa 0.34%;

ii) *corpus* paralelo IO-PT

Há 2723 ocorrências de SAY em IO e 2.385 ocorrências de DIZER em PT.

Logo, a ocorrência de SAY nos textos originais é *maior* do que a de DIZER nos textos traduzidos. SAY equivale a 0.67% de *tokens* dos textos originais, enquanto que nas traduções a ocorrência de DIZER é de 0.60%.

iii) *corpus* paralelo PO-IT

Há 707 ocorrências de DIZER em PO e 776 ocorrências de DIZER em IT.

Logo, a ocorrência de DIZER nos textos originais é *semelhante* à ocorrência de SAY nos textos traduzidos. DIZER equivale a 0.34% de *tokens* dos textos originais enquanto que nas traduções a ocorrência de SAY é de 0.33%. Nota-se que a diferença é mínima em termos percentuais, ao passo que, em termos de ocorrências, o texto traduzido tem mais ocorrências

de SAY do que os originais têm de DIZER. Nota-se que a ocorrência de SAY é maior do que a ocorrência de DIZER tanto no *corpus* comparável bilíngue (IO-PO) quanto nos *corpora* paralelos (PO-IT e IO-PT);

iv) *corpus* comparável monolíngue (IO-IT)

Há 2.723 ocorrências de SAY nos textos não-traduzidos (IO) e 776 ocorrências de SAY nos textos traduzidos (IT). SAY equivale a 0.67% de *tokens* dos textos não-traduzidos, enquanto que nas traduções a ocorrência é de 0.33%. Nota-se, portanto, que a ocorrência de SAY nas traduções é significativamente *menor* do que a ocorrência em textos não-traduzidos.

v) *corpus* comparável monolíngue (PO-PT)

Há 707 ocorrências de DIZER nos textos não-traduzidos (PO) e 2.385 ocorrências de DIZER nos textos traduzidos (PT). DIZER equivale a 0.34% de *tokens* dos textos não-traduzidos, enquanto que nas traduções a ocorrência é de 0.60%. Nota-se, portanto, que a ocorrência de DIZER nas traduções é significativamente *maior* do que a ocorrência em textos não-traduzidos.

Estas considerações iniciais, obtidas através da observação dos dados brutos do *corpus*, permitem a elaboração de perguntas de pesquisas que orientarão a análise qualitativa do *corpus*. Cada um dos sub*corpora* do *corpus* combinado permite a análise de distintos aspectos das relações entre os textos que, somadas, contribuem para uma compreensão mais abrangente das relações de tradução. Estas perguntas baseiam-se no modelo de análise proposto por Teich (2003):

1- Qual a relação entre as orações verbais realizadas por SAY em textos ficcionais não-traduzidos em inglês e orações verbais realizadas por DIZER em textos ficcionais não-traduzidos em português?
2- As orações verbais realizadas por SAY são frequentemente traduzidas por orações verbais realizadas por DIZER e vice-versa? Qual a probabilidade de equivalência entre SAY/DIZER? Quais os outros itens lexicais em relação tradutória com SAY/DIZER? É possível estabelecer padrões que condicionem as relações de tradução?

3- Qual a relação entre as orações verbais realizadas por SAY/DIZER em textos não-traduzidos e as orações verbais realizadas por SAY/DIZER em textos traduzidos? Os textos traduzidos apresentam os mesmos padrões dos textos não-traduzidos ou estes padrões são condicionados pela relação dos textos traduzidos com os textos originais?

Para a investigação de cada pergunta, focaliza-se uma combinação diferente do *corpus*: 1- relações entre textos não-traduzidos em português e inglês (IO-PO); 2- relações entre textos traduzidos e textos não-traduzidos em inglês (IO-IT); 3- relações entre textos traduzidos e textos não-traduzidos em português (PO-PT); 4- relações entre textos originais em inglês e suas traduções para o português (IO-PT); e 5- relações entre textos originais em português e suas traduções para o inglês (PO-IT).

4.2 O programa SPSS e o banco de dados

A partir dos dados brutos e das primeiras considerações que eles permitem, é possível uma análise mais refinada dos dados. No total, há cerca de 6500 ocorrências de SAY/DIZER no *corpus*, que podem ser localizadas utilizando-se conjuntamente as ferramentas *WordList* e *Concord* do software *WordSmith Tools*, como foi explicado. Optou-se por selecionar cerca de 10% das ocorrências e, pensando-se em uma amostragem com o mesmo número de ocorrências de cada romance, optou-se por selecionar as primeiras 50 ocorrências de cada romance, totalizando 599 ocorrências, visto que o número total de ocorrências em um dos romances é 49.

Dado o foco do estudo, os processos verbais, foram eliminadas 30 ocorrências, 04 de SAY e 26 de DIZER, de orações em que SAY/DIZER realizam processos simbólicos e processos relacionais, fugindo, portanto, do escopo da análise.

Cada uma das 599 ocorrências foi analisada na linha de concordância fornecida pela ferramenta *Concord* e selecionaram-se apenas as ocorrências em que SAY/DIZER realizam processos verbais. Após este recorte, restaram 569 linhas de ocorrências, sendo 296 ocorrências de SAY (149 em textos não-traduzidos (IO) e 147 em textos traduzidos (IT)) e 273 ocorrências de DIZER (132 em textos não-traduzidos (PO) e 141 em textos traduzidos (PT)).

Estes dados já apontam o fato de que o papel de SAY na realização de orações verbais parece ser mais frequente do que o de DIZER, visto que foram eliminados apenas 04 casos de SAY e 26 casos de DIZER.

O próximo passo foi analisar cada ocorrência em relação às categorias focalizadas. O início da análise foi feito através de anotação[8] manual do *corpus*, que consiste em o pesquisador estabelecer códigos para as categorias analisadas e inserir estes códigos no *corpus* dentro de etiquetas (denominam-se etiquetas os parênteses angulares < e >, dentro dos quais se insere o código no *corpus*). Posteriormente, as etiquetas são contabilizadas automaticamente com o programa *WordSmith Tools* para a obtenção de dados, possibilitando ao pesquisador desenvolver sua análise qualitativa a partir dos dados quantitativos. Entretanto, este processo é bastante limitado para análise de grande quantidade de dados, principalmente quando se utilizam várias categorias. Optou-se, portanto, pelo uso de um programa estatístico, o SPSS (*Statistical package for the social sciences*).

O programa SPSS trabalha com casos ou ocorrências, variáveis e categorias, formando um banco de dados em uma planilha semelhante às utilizadas pelo programa Excel. O SPSS apresenta muitas facilidades: a construção do banco de dados é simples; é possível classificar separadamente cada aspecto observado e depois cruzar os dados; as ferramentas de geração de tabelas e gráficos são simples e com muitos recursos de produção e edição; estão disponíveis diferentes formas de visualização dos dados. Dentre suas limitações pode-se apontar: custo muito alto do programa; dificuldade para se trabalhar com uma variável que possua mais de cinco categorias; não apresenta nenhuma interface com o texto, visto que o programa não foi desenvolvido especificamente para linguistas; necessidade de conhecimentos de estatística para análises mais elaboradas.

O programa destaca-se pela facilidade de construção de um banco de dados que possibilita o cruzamento de variáveis, a extração de dados de frequência e a geração de tabelas e gráficos, sem aprofundamento em testes estatísticos. Olohan (2004:86) aponta que vários linguistas consideram que os dados de frequência podem ser suficientes para a análise e que os testes de significância não são sempre necessários. A autora cita uma observação de Halliday sobre esse aspecto, em que o autor diz que "dados brutos de frequências geralmente são suficientes para se aceitar a afirmativa do pesquisador de que alguma característica é

[8] Confira Feitosa (2005) para um estudo mais detalhado sobre anotação de corpora.

relevante no texto e que se deve verificá-la"[9]. Desta forma, o programa foi utilizado para o processamento automático de frequência das variáveis e categorias analisadas com a geração de tabelas e gráficos[10].

Inicialmente, o pesquisador estabelece as variáveis e categorias de análise que irão fazer parte do banco de dados do SPSS. Recomenda-se a construção de um único banco de dados, portanto, as variáveis e categorias devem englobar todas as ocorrências do *corpus*. Algumas variáveis podem existir apenas para facilitar a organização do banco de dados, como, por exemplo, a variável OCORRÊNCIAS, que não possui categorias e em que cada número corresponde a uma das 50 linhas de concordância selecionada para investigação em cada romance. Outras variáveis servem para facilitar a seleção de parte dos dados dentro do banco de dados, por exemplo, a variável STATUS cujas categorias são inglês original (IO), português original (PO), inglês traduzido (IT), e português traduzido (PT). O programa permite a seleção de parte dos dados, segundo as diferentes categorias, para a geração de dados, o que permite ao pesquisador, no exemplo da variável STATUS, selecionar os dados em relação às ocorrências de português em geral, ou somente de português traduzido ou somente de português original, segundo os objetivos da pesquisa.

Após vários testes, foram estabelecidas nove variáveis para a análise com suas respectivas categorias, descritas a seguir.

1) Variável: OCORRÊNCIAS. Sem categorias

Foram selecionadas 300 ocorrências de SAY e 299 ocorrências de DIZER que se constituíram das 50 primeiras ocorrências de cada romance, sendo que um deles possui apenas 49 ocorrências de DIZER. O SPSS conta cada ocorrência como um caso, ocupando uma linha do banco de dados e numerando-as de 1 a 599. Depois de eliminados os 30 casos em que SAY/DIZER não realizam processo verbal, restaram 569 casos. Na variável OCORRENCIA cada uma foi numerada segundo as linhas de concordância selecionadas com o programa *WordSmith*, ou seja, de 1 a 50 pela ordem de ocorrência.

[9] Minha tradução de: "a rough indication of frequencies is often just what is needed: enough to suggest why we should accept the analyst's assertation that some feature is prominent in the text, and allow us to check his statements"(Halliday apud Olohan, 2004:86).
[10] Para exemplos de pesquisas linguísticas com o uso do SPSS, veja a tese de Rothe-Neves (2002) na área de tradução e a dissertação de Oliveira (2006) na área de fonologia da língua portuguesa.

2) Variável: LEXICO. Categorias: SAY e DIZER

Cada ocorrência foi classificada como SAY ou DIZER, conforme faça parte do *corpus* em português ou do *corpus* em inglês.

3) Variável: CORPUS. Categorias: os doze romances do *corpus* combinado

Foi assinalada a origem de cada ocorrência, segundo os doze romances que compõem o *corpus* combinado. Essa classificação permite que se façam análises separadas de cada romance, procedimento necessário, por exemplo, para a localização de linhas de concordância a serem utilizadas como exemplos. Atribuiu-se uma sigla para cada um, conforme descrito na Tabela 1.

> PCP - Point counter point; IWV - Interview with the vampire; BEL – Beloved;
> MAC – Macunaíma; GAB – Gabriela, clove and cinammon; THS - The hour of the star;
> MHS – Macunaíma: o herói sem caráter; GCC – Gabriela, cravo e canela;
> AHE - A hora da estrela; CPO – Contraponto; ECV - Entrevista com vampiro;
> AMA - Amada

4- Variável: FORMA. Categorias: as formas de SAY/DIZER

Foi analisada a frequência das diferentes formas de SAY/DIZER. Enquanto SAY possui apenas quatro formas, *say, says, said, saying*, DIZER possui várias, sendo as mais frequentes, *disse, dizer, dizia*.

5- Variável: MODOPROJ. Categorias: relato, citação, verbiagem

Nesta variável, cada oração verbal foi analisada segundo o modo de projeção: relato ou citação. Os casos classificados como verbiagem indicam que não ocorreu a projeção.

6- Variável: MODOEXP. Categorias: congruente e metafórico

Cada oração verbal foi classificada segundo o modo de expressão: congruente ou metafórico. As orações de verbiagem não foram analisadas em relação aos modos de expressão.

7- Variável: TRADUÇÃO. Categorias: traduções de SAY/DIZER no *corpus* paralelo

Para a classificação desta variável foi necessário realizar antes o alinhamento do *corpus* paralelo. Alinhar significa colocar um trecho do texto original seguido do trecho do texto traduzido correspondente (Lawson, 2001). Este trabalho foi feito com o auxílio de bolsistas do LETRA e utilizaram-se dois recursos diferentes: a ferramenta *Viewer and Aligner* do programa *WordSmith* e o processador de texto *Word*. O resultado não é muito distinto. O *Viewer* faz o alinhamento automático, por sentenças ou parágrafos, mas apresenta problemas que precisam ser corrigidos manualmente pelo pesquisador; o texto alinhado pode ser exportado para ser manuseado no *Word*. No *Word*, o alinhamento é feito manualmente dividindo-se as sentenças ou parágrafos em duas colunas de uma tabela, uma para o texto original e outra para o texto traduzido.

Então, cada equivalente encontrado passa a ser uma categoria, sendo que esta variável tem tantas categorias quantas forem os possíveis equivalentes, agrupados sob a forma do infinitivo, que funciona como lema. Por exemplo, os itens lexicais mais frequentes correspondentes a SAY no *corpus* são DIZER, FALAR e RESPONDER. E os mais frequentes correspondentes a DIZER são SAY, TELL e EXPLAIN. Os dados observados nos *corpora* alinhados foram lançados diretamente no SPSS.

8- Variável: RELAÇÕES. Categorias: prototípico, típico, atípico, omissão e não-verbal

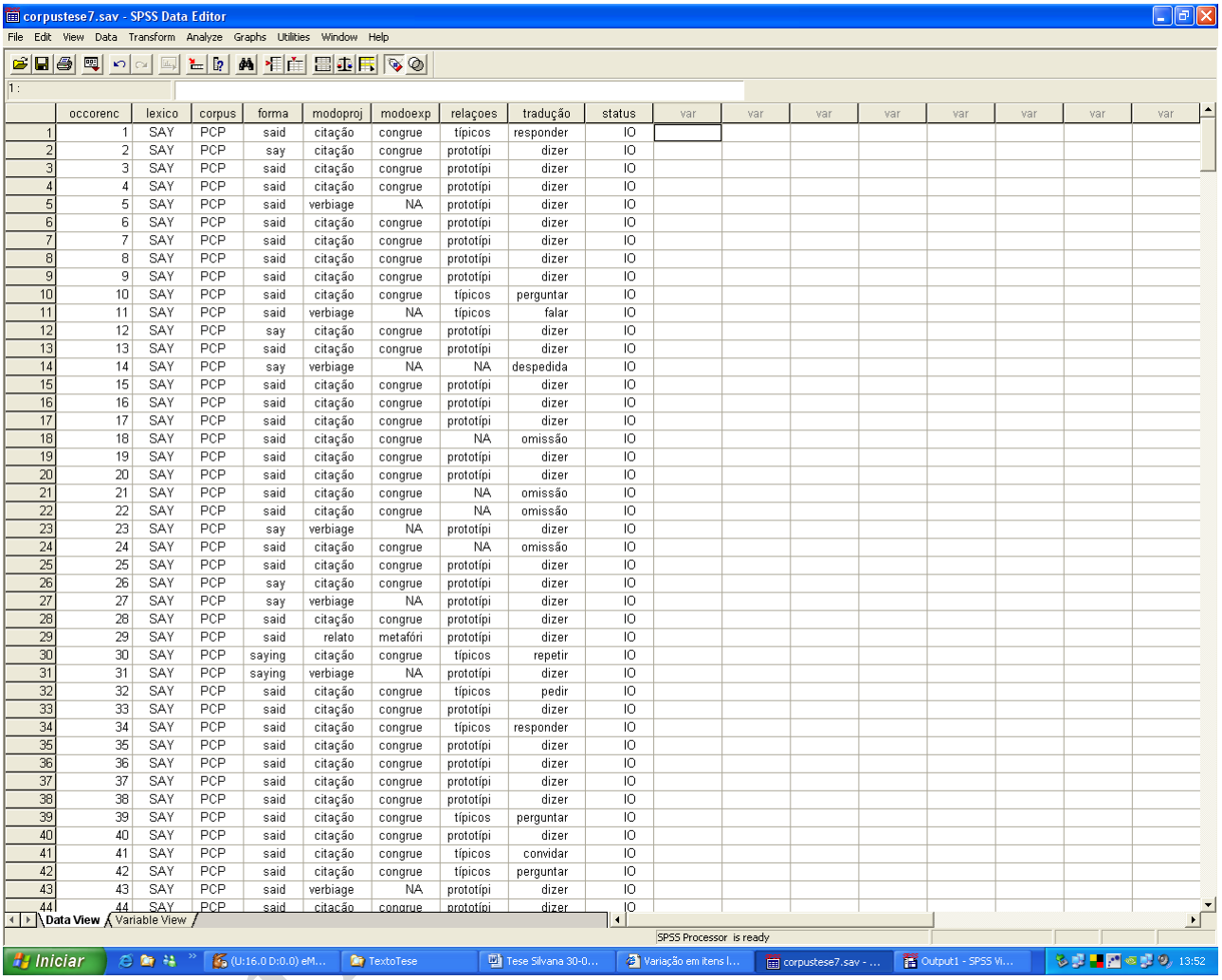
Os itens lexicais que aparecem no *corpus* em relação de tradução com SAY/DIZER, foram agrupam nestas categorias, segundo o tipo de verbo dentro dos verbos que podem realizar processo verbal.

9- Variável: STATUS. Categorias: inglês/português e original/traduzido

Cada ocorrência foi classificada em relação aos quatro *corpora* básicos: inglês original (IO), português original (PO), inglês traduzido (IT) e português traduzido (PT).

Todas as 569 ocorrências foram classificadas em relação a estas nove variáveis, formando assim o banco de dados da pesquisa. O SPSS permite que se realize uma análise parcial, excluindo-se uma ou mais variáveis e/ou categorias, facilitando a exploração do banco de dados para diferentes tipos de análise. O programa gera tabelas e gráficos de dados, com extensas possibilidades de edição e apresentação.

A Figura 1 mostra uma janela do SPSS com as nove variáveis de análise e uma amostra parcial das ocorrências que compõem o banco de dados.



	occorenc	lexico	corpus	forma	modoproj	modoexp	relações	tradução	status	var	var	var	var	var	var	var	var
1	1	SAY	PCP	said	citação	congrue	típicos	responder	IO								
2	2	SAY	PCP	say	citação	congrue	prototípi	dizer	IO								
3	3	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
4	4	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
5	5	SAY	PCP	said	verbiage	NA	prototípi	dizer	IO								
6	6	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
7	7	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
8	8	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
9	9	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
10	10	SAY	PCP	said	citação	congrue	típicos	perguntar	IO								
11	11	SAY	PCP	said	verbiage	NA	típicos	falar	IO								
12	12	SAY	PCP	say	citação	congrue	prototípi	dizer	IO								
13	13	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
14	14	SAY	PCP	say	verbiage	NA	NA	despedida	IO								
15	15	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
16	16	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
17	17	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
18	18	SAY	PCP	said	citação	congrue	NA	omissão	IO								
19	19	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
20	20	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
21	21	SAY	PCP	said	citação	congrue	NA	omissão	IO								
22	22	SAY	PCP	said	citação	congrue	NA	omissão	IO								
23	23	SAY	PCP	say	verbiage	NA	prototípi	dizer	IO								
24	24	SAY	PCP	said	citação	congrue	NA	omissão	IO								
25	25	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
26	26	SAY	PCP	say	citação	congrue	prototípi	dizer	IO								
27	27	SAY	PCP	say	verbiage	NA	prototípi	dizer	IO								
28	28	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
29	29	SAY	PCP	said	relato	metafóri	prototípi	dizer	IO								
30	30	SAY	PCP	saying	citação	congrue	típicos	repetir	IO								
31	31	SAY	PCP	saying	verbiage	NA	prototípi	dizer	IO								
32	32	SAY	PCP	said	citação	congrue	típicos	pedir	IO								
33	33	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
34	34	SAY	PCP	said	citação	congrue	típicos	responder	IO								
35	35	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
36	36	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
37	37	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
38	38	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
39	39	SAY	PCP	said	citação	congrue	típicos	perguntar	IO								
40	40	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								
41	41	SAY	PCP	said	citação	congrue	típicos	convidar	IO								
42	42	SAY	PCP	said	citação	congrue	típicos	perguntar	IO								
43	43	SAY	PCP	said	verbiage	NA	prototípi	dizer	IO								
44	44	SAY	PCP	said	citação	congrue	prototípi	dizer	IO								

Figura 1 – Janela do programa SPSS com visão parcial do banco de dados

A Tabela 2 é um exemplo de tabela produzida pelo SPSS e mostra as ocorrências de SAY/DIZER nos romances do *corpus*, abrangendo os quatro *corpora* básicos do *corpus* combinado. O *corpus* de textos originais em inglês (IO) com 149 ocorrências, o *corpus* de textos originais em português (PO) com 132 ocorrências, o *corpus* de textos traduzidos em inglês (IT) com 147 ocorrências e o *corpus* de textos traduzidos em português (PT) com 141 ocorrências, totalizando 569 casos ou ocorrências de SAY/DIZER. Nota-se que as tabelas produzidas pelo SPSS não seguem os padrões da ABNT.

Tabela 2 – Ocorrências de SAY/DIZER no *corpus* combinado (exemplo de tabela produzida pelo SPSS)

Romances do corpus * Texto original ou tradução Crosstabulation

Count

		Texto original ou tradução				
		ING nao-trad	POR nao-trad	ING trad	POR trad	Total
Romances do corpus	PCP	50				50
	IWV	49				49
	BEL	50				50
	MAC			47		47
	GAB			50		50
	THS			50		50
	MHS		40			40
	GCC		47			47
	AHE		45			45
	CPO				49	49
	ECV				45	45
	AMA				47	47
Total		149	132	147	141	569

A Tabela 2 está no formato produzido pelo SPSS, mas o programa tem amplos recursos para edição de tabelas. Pode-se, por exemplo, mudar o título, excluir colunas, alterar a fonte, inserir notas, alterar o tamanho e outras edições e formatações, conforme exemplificado na Tabela 3.

Tabela 3 – Exemplo de tabela editada com os recursos do SPSS

Ocorrências de SAY/DIZER nos romances do corpus combinado[a]

Count

		Texto original ou tradução			
		ING nao-trad	**POR nao-trad**	**ING trad**	**POR trad**
Romances do corpus	PCP	50			
	IWV	49			
	BEL	50			
	MAC			47	
	GAB			50	
	THS			50	
	MHS		40		
	GCC		47		
	AHE		45		
	CPO				49
	ECV				45
	AMA				47
Total		149	132	147	141

a. Corpus paralelo e comparável bidirecional no par linguístico inglês-português

Além do manual do programa (em inglês), é possível encontrar manuais em português com o passo-a-passo para a utilização dos recursos do programa.

Todas estas variáveis foram detalhadamente analisadas em Jesus (2008). A próxima seção apresenta, parcialmente, a análise de uma das variáveis, as relações de tradução, para ilustrar a metodologia utilizada.

5. As relações de tradução entre SAY-DIZER

O conceito leigo de equivalência geralmente está associado à ideia de que para um determinado item em dada língua A há um item equivalente na língua B. Entretanto, os estudos da tradução desde muito apontam que, para um determinado item em uma língua A, existem *vários* itens equivalentes na língua B ou, nos termos de Halliday (1992:16), "um item X na língua do original tem um grupo de itens equivalentes em potencial - A, B, C, D, E, F - na língua da tradução"[11]. Portanto, pode-se dizer que existem diferentes *relações de tradução* entre um item do texto original e seus vários possíveis equivalentes no texto traduzido.

A análise das relações de tradução de SAY/DIZER tem o objetivo de responder as seguintes perguntas: i) Dado que, em dicionários bilíngues, DIZER é o equivalente prototípico de SAY, com qual frequência SAY é traduzido por DIZER? ii) Dado que há ocorrências de SAY que *não* são traduzidas por DIZER, pois não há apenas um equivalente, quais seriam os possíveis equivalentes de SAY além de DIZER? E qual seria a frequência de ocorrência destes equivalentes? As mesmas perguntas são colocadas em relação em relação a DIZER, pensando-se na direção do português para o inglês.

Considerando-se SAY/DIZER como verbos que realizam Processo verbal, é natural que SAY/DIZER estejam em relações de tradução com outros verbos que também realizam processo verbal. A Gramática sistêmico-funcional (GSF, doravante) (HALLIDAY e MATTHIESSEN 2004) apresenta dois grupos de verbos que realizam processo verbal, segundo o modo de citação e o modo de relato.

Quadro 2 - Verbos que realizam Processo em orações verbais no modo de citação

	Proposição	Proposta
(1) Membro geral	SAY	SAY
(2)Verbos específicos funções de fala: a) oferta	aa) Declarações: TELL receptor), REMARK, POINT OUT, ANNOUNCE.	(+a) Ofertas: SUGGEST, OFFER; OBSERVE,THREATHEN, VOW, REPORT,PROMISE, AGREE.
b) demanda	b) Perguntas: ASK,	DEMAND,b) Comandos: CALL, ORDER,

[11] Minha tradução de: You are aware that an item X in the source language has a potential equivalence of items A, B, C, D, E, F in the target language (Halliday, 1992:16).

	Proposição	Proposta
	INQUIRE, QUERY.	REQUEST, TELL, PROPOSE, DECIDE; URGE, PLEAD, WARN.
(3)Verbos com características adicionais a) circunstanciais	REPLY, EXPLAIN, PROTEST, CONTINUE, ADD, INTERRUPT, WARN.	[consultar (2) acima]
b)especificadores de modo ou conotação	INSIST, COMPLAIN, CRY, SHOUT, BOAST, MURMUR, STAMMER.	[em grande parte como as proposições] BLARE, THUNDER, MOAN, YELL, FUSS.

Fonte: Halliday e Matthiessen (2004:448). Tradução de Alves (2006:29)

A GSF aponta SAY como o membro geral da classe de verbos que realizam processos verbais em citações (Halliday e Matthiessen, 2004:448), conforme apresentado no Quadro 2. Destaca-se seu papel como o membro não-marcado deste grupo de verbos (p. 252). Thompson (1994:34-35) considera SAY como o verbo de elocução básico, utilizado quando não se quer indicar nenhuma finalidade ou modo do dizer, portanto, um verbo neutro. Considera ainda, como neutros, os verbos*: tell, ask, write, speak, talk.*

Halliday e Matthiessen (2004:448) consideram ainda que outros verbos "que não são verbos de dizer", podem, especialmente na narrativa ficcional, realizar Processos verbais. Os autores destacam os verbos que tipicamente realizam Processos comportamentais (como *sob, snort, breathe*) e que quando realizam processos verbais expressam atitudes, emoções e gestos que acompanham o ato de fala.

A classificação dos diferentes tipos de verbos que realizam processos verbais é relevante para a análise das relações de tradução de SAY/DIZER, visto que aqueles são os mais prováveis candidatos ao quadro de equivalentes possíveis destes verbos.

Para esta análise, utilizaram-se os *corpora* paralelos do *corpus* combinado, compostos de textos originais em inglês (IO) e suas traduções para o português (PT) e de textos originais em português (PO) e suas traduções para o inglês (IT). A análise é bidirecional, ou seja, observa as relações de tradução tanto na direção do português para o inglês, quanto na direção do inglês para o português.

Foram selecionadas as primeiras 50 ocorrências de SAY/DIZER de cada romance. Estas ocorrências foram analisadas, excluindo-se as orações em que SAY/DIZER não realizam processo verbal, como mencionado. Portanto, analisam-se 149 orações verbais realizadas por SAY e 132 orações verbais realizada por DIZER.

Inicialmente, foram observadas as relações de tradução de SAY/DIZER, considerando-se o estrato léxico-gramatical, ou seja, as relações de equivalência de SAY (em IO) com DIZER e outros itens lexicais do português (em PT) e, por sua vez, as relações de equivalência de DIZER (em PO) com SAY e outros itens lexicais do inglês (em IT).

Ao invés de se considerar isoladamente cada item em relação de tradução com SAY/DIZER, decidiu-se por agrupá-los em categorias semânticas, na perspectiva experiencial de realização de Processos verbais, segundo a classificação da GSF para os diferentes tipos de verbos: i) membro geral, ii) verbos específicos e iii) verbos que, tipicamente, não são verbos de dizer, mas podem realizar esse significado.

Observa-se que, embora não existam descrições do português baseadas na LSF, Neves (2000:48) apresenta uma classificação dos verbos *discendi*, que se aproxima da proposta da GSF. A autora classifica os verbos de elocução como neutros (*dizer* e *falar*), verbos que indicam algum aspecto do dizer, como modo ou cronologia (*repetir*, *gritar*, *responder*) e verbos que podem funcionar como introdutores do discurso (*consolar*, *sorrir*). Menciona-se ainda a classificação de Garcia (1986:131), que considera como verbos de elocução os verbos *discendi* (*dizer, perguntar, responder, exclamar*) e os verbos *sentiendi* (*gemer, suspirar, lamentar-se, queixar-se*), que introduzem o discurso e realizam emoções, atitudes, comportamentos. Utilizou-se, ainda, para a análise dos verbos em português, o *Dicionário Houaiss da língua portuguesa*[12], observando-se as significações possíveis para os verbos e suas relações com o universo do dizer. Portanto, utiliza-se uma classificação das relações de tradução de SAY/DIZER em três tipos: i) prototípicos, ii) típicos e iii) atípicos.

Considera-se como prototípico o verbo que é o membro geral da classe, ou seja, SAY/DIZER. E como verbos típicos todos os outros verbos que realizam orações verbais nos modos de citação e relato. Como verbos atípicos consideram-se os verbos que tipicamente realizam *outro* tipo de Processo (mentais, materiais e outros), mas que nos textos ficcionais geralmente realizam Processos verbais ou que eventualmente podem realizar este tipo de Processo.

Assim, consideram-se três tipos de relações de tradução de SAY e DIZER: i) SAY é traduzido por DIZER, e vice-versa, em que a relação se dá com o verbo prototípico; ii) SAY/DIZER são traduzidos por outros verbos típicos de processos verbais; iii) SAY/DIZER

[12] Dicionário online, disponível para assinantes em http://educacao.uol.com.br/dicionarios/ .

são traduzidos por verbos que não são verbos de dizer, mas que podem realizar processo verbal, em que a relação se dá com o verbo atípico.

Ocorrem ainda, no *corpus*, casos em que SAY/DIZER são traduzidos por um substantivo, ou seja, a relação de tradução não se dá entre verbos. Ou ainda, casos em que a oração verbal realizada por SAY/DIZER é omitida.

Portanto, as relações de tradução de SAY/DIZER encontradas no *corpus* analisado são classificadas em cinco tipos: prototípico, típico, atípico, não-verbal e omissão. Esta classificação é feita a partir dos *corpora* paralelos: i) SAY/DIZER aparecem como equivalentes, ou seja, SAY é traduzido por DIZER e DIZER corresponde a SAY, casos classificados como prototípicos; ii) SAY/DIZER são traduzidos por outros verbos que tipicamente realizam Processo verbal como *falar, perguntar, responder, pedir, tell, speak, remark, point out*, casos classificados como típicos; iii) SAY/DIZER são traduzidos por verbos que tipicamente realizam outros tipos de processos, como *lembrar, fazer, mean, think*, que são classificados como atípicos; iv) casos em que SAY/DIZER são traduzidos por um substantivo, e v) casos de omissão.

Inicialmente, foram observados os itens lexicais em relação de tradução com SAY no *corpus*, apresentados na Tabela 4.

Tabela 4 – Itens lexicais em relação de tradução com SAY

		Frequency	Valid Percent
Valid	dizer[a]	98	65,8
	omissão	21	14,1
	falar	12	8,1
	responder	5	3,4
	perguntar	3	2,0
	não-verbal	2	1,3
	comentar	2	1,3
	pedir	1	,7
	convidar	1	,7
	repetir	1	,7
	concordar	1	,7
	fazer	1	,7
	retrucar	1	,7
	Total	149	100,0

a. Corpus paralelo inglês-português (IO-PT)

O *corpus* confirma a expectativa de que DIZER seja o equivalente mais frequente de SAY; a relação com o verbo prototípico ocorre em 65.8% dos casos. Nota-se que a segunda

relação mais frequente é a de omissão, que ocorre em 14.1%. O terceiro item mais frequente foi *falar,* um dos verbos típicos, que ocorre em 8.1% dos casos. Há dois casos (1.3%) de relação não-verbal. Agrupando-se estes itens lexicais, segundo as categorias de análise, tem-se outra perspectiva das relações de tradução, apresentadas na Tabela 5.

Tabela 5 – Relações de tradução de SAY segundo as categorias

		Frequency	Valid Percent
Valid	prototípico[a]	97	65,1
	típicos	28	18,8
	atípicos	1	,7
	nao-verbal	2	1,3
	omissão	21	14,1
	Total	149	100,0

a. Corpus paralelo inglês-português (IO-PT)

O equivalente prototípico ocorre em 65.1% dos casos, enquanto que verbos típicos de processos verbais ocorrem em 18.8%. Como foi dito, a omissão ocorre em 14.1% dos casos, há duas ocorrências de relação não-verbal (1.3%) e uma ocorrência de verbo atípico (0.7%). Exemplos do *corpus* são apresentados no Quadro 3.

Quadro 3 – Exemplos do *corpus* das relações de tradução de SAY em IO-PT

Prototípico
04 "You forgetting how little it is," **said** her mother. (BEL) -- Você está se esquecendo de que ela é muito pequenina, **disse** a mãe. (AMA)
Típico
45 “But I mean we want to get married.” “You just said so. And I **said** all right.” (BEL) --O que estou querendo dizer é que vamos nos casar. -- Já me disse isso. E eu **falei** que está tudo bem. (AMA)
Atípico
46 'Were you?' **said** Lady Edward smiling and looking from one to the other. (PCP) - Estavam? - **fez** Lady Edward, sorridente, olhando ora para uma ora para outra. (CPO)
Omissão
17 "We and Denver," she **said**. "That all right by you? (BEL) -- Eu e Denver. -- E está tudo bem com você? (AMA)
Não-verbal
14 When the time came to **say** good-bye, he had shaken the skeleton hand. (PCP) **Chegara o momento da despedida**: ele apertou na sua a mão esquelética do doente, a mão que jazia inerte sobre a coberta. (CPO)

08 She had not thought to ask him and it bothered her still that it might have been possible-that for twenty minutes, a half hour, **say**, she could have had the whole thing, every word she heard the preacher say at the funeral (and all there was to say, surely) engraved on her baby's headstone: Dearly Beloved. (BEL)
Não pensara em perguntar-lhe e ainda se perturbava com a probabilidade de ter sido possível - que por vinte minutos, **talvez** meia hora, teria obtido a coisa inteira, cada palavra que ouvira o pastor dizer no funeral (e certamente tudo o que havia para ser dito) gravada na lápide de sua filhinha: Caríssima Amada. (AMA)

Exemplos do *corpus* paralelo: inglês original (IO) e português traduzido (PT)

Em seguida, observaram-se os itens lexicais em relação de equivalência com DIZER, apresentados na Tabela 6.

Tabela 6 – Itens lexicais em relação de tradução com DIZER

		Frequency	Valid Percent
Valid	say[a]	61	46,2
	tell	14	10,6
	omissão	8	6,1
	mention	4	3,0
	explain	4	3,0
	não-verbal	3	2,3
	speak	3	2,3
	think	3	2,3
	mutter	2	1,5
	ask	2	1,5
	laugh	1	,8
	insist	1	,8
	believe	1	,8
	conclude	1	,8
	pronounce	1	,8
	divulge	1	,8
	inform	1	,8
	snap	1	,8
	retort	1	,8
	observe	1	,8
	call	1	,8
	remark	1	,8
	confide	1	,8
	object	1	,8
	comment	1	,8
	express	1	,8
	beg	1	,8
	rejoin	1	,8
	reply	1	,8
	hear	1	,8
	add	1	,8
	go	1	,8
	resume	1	,8
	accuse	1	,8
	find	1	,8
	concede	1	,8
	cover	1	,8
	imagine	1	,8
	Total	132	100,0

a. Corpus paralelo português-inglês (PO-IT)

DIZER se destaca como o equivalente prototípico de SAY, embora não chegue a 50% dos casos (46.2%). O segundo item mais frequente é *tell*, com 10.6%. A porcentagem de omissão de DIZER é de 6.1% e há três ocorrências de relação não-verbal (2.3%). A Tabela 7 mostra as relações de tradução de DIZER, quando se agrupam estes itens segundo as categorias de análise.

Tabela 7 – Relações de tradução de DIZER segundo as categorias

		Frequency	Valid Percent
Valid	prototípico[a]	61	46,2
	típicos	49	37,1
	atípicos	11	8,3
	nao-verbal	3	2,3
	omissão	8	6,1
	Total	132	100,0

a. Corpus paralelo português-inglês (PO-IT)

O equivalente prototípico ocorre em 46.2% dos casos, enquanto que outros verbos típicos de processos verbais ocorrem em 37.1% e verbos atípicos em 8.3%. A omissão ocorre em 6.1% e a relação não-verbal em 2.3%. Exemplos no Quadro 4.

Quadro 4 – Exemplos do *corpus* das relações de tradução de DIZER em PO-IT

Prototípico
17 -- Isso é revoltante - **dizia** o Doutor enquanto o grupo caminhava pela rua sem calçamento, contornando o morro. (GCC) "This is revolting," **said** the Doctor as the group walked along the unpaved street skirting the hill.
Típico
37 O ticotiquinho ficava azaranzado porque estava padecendo fome e aquele nhenhenhém-nhenhenhém azucrinando ele atrás, **diz**-que "Telo decumê!... telo decumê!..." (MHS) The cowbird swallowed everything and **resumed** its habitual refrain "Boo-hoo! Mama, I'm famished, I'm famished!" (MAC)
Atípico
32 E todos com muito medo foram correndo pra dentro. Então Chuvisco desapeou e **disse** pra Macunaíma: Está vendo? (MHS) Shower came down and **laughed** at Macunaíma: "Did you see that?" (MAC)
Omissão
12 Em Itamaracá Macunaíma passou um pouco folgado e teve tempo de comer uma dúzia de manga-jasmim que nasceu do corpo de dona Sancha, **dizem**. (MHS) They ran to the island of Itamaracá, where the hero ate some mangoes, a dozen jasmin mangoes that had sprung from the grave of the late Dona Sancha; (MAC)
Não-verbal

03 O seu rico andor bordado de ouro, levavam-no sobre os ombros orgulhosos os cidadãos mais notáveis, os maiores fazendeiros, vestidos com a bata vermelha da confraria, **e não é pouco dizer**, pois os coronéis do cacau não primavam pela religiosidade... (GCC)
The gold-embroidered litter bearing the image of the saint was carried on the shoulders of the town's most important citizens, the owners of the largest plantations, dressed in the red gowns of the lay brotherhood. **This was significant**, for the cacao colonels ordinarily avoided religious functions. (GAB)

21 Pelópidas de Assunção d'Ávila descendia de uns Ávilas, fidalgos portugueses estabelecidos nas bandas de Ilhéus ainda no tempo das capitanias. Pelo menos assim o afirmava o Doutor, dizendo-se baseado em documentos de família. Opinião ponderável, de historiador. Descendente desses celebrados Ávilas, cujo solar elevara-se entre Ilhéus e Olivença, hoje negras ruínas ante o mar, cercadas de coqueiros, mas também de uns Assunções plebeus e comerciantes - **diga**-se em sua homenagem, ele cultuava a memória de uns e de outros com o mesmo fervor exaltado. (GCC)
Pelópidas de Assunção d'Ávila (the Doctor) maintained that he was descended from the Portuguese noblemen named Ávila who had settled near what was now Ilhéus during the period of royal land grants. The line of descent could be clearly traced, he found, in family documents. The solid opinion of a historian. He was descended also from certain plebeian, shopkeeping Assunções, and, **to his great credit,** he cherished the memory of these ancestors and of the Ávilas with the same exalted fervor. (GAB)

01 Si o incitavam a falar exclamava:
- Ai! que preguiça!...
e não **dizia** mais nada. (MHS)
If anyone tried to make him speak he would exclaim, "Aw! What a fucking life!" but **nothing more**.

Exemplos do *corpus* paralelo: português original (PO) e inglês traduzido (IT)

Vários aspectos se destacam nas relações de tradução de SAY e DIZER a partir destes dados do *corpus*. Poder-se-ia pensar que SAY é a tradução de DIZER na mesma proporção que DIZER é a tradução de SAY, mas os dados apontam diferenças nas relações de tradução, conforme a direção seja do inglês para o português ou do português para o inglês. Estas diferenças são discutidas em cada uma das relações de tradução observadas para SAY e DIZER.

Relação de tradução I - equivalente prototípico

A ocorrência do equivalente prototípico é maior na direção do inglês para o português (65.1%) do que do português para o inglês (46.2%).

Relação de tradução II – equivalente típico

SAY apresenta relações de tradução com 09 verbos típicos de Processos verbais no português além de DIZER: *falar, responder, perguntar, comentar, pedir, convidar, repetir, concordar* e *retrucar*. DIZER, por sua vez, aparece em relação de tradução com 26 verbos típicos de Processos verbais além de SAY: *tell, mention, explain, speak, mutter, ask, insist, pronounce, divulge, inform, snap, retort, observe, call, remark, confide, object, comment,*

express, beg, rejoin, reply, add, resume, accuse e *concede*. Estatisticamente, 18.8% das ocorrências de SAY estão em relação de tradução com verbos típicos e 37.1% das ocorrências de DIZER.

Relação de tradução III – equivalente atípico

SAY apresenta relação de tradução com 01 verbo atípico (*fazer*), ao passo que DIZER apresenta relações de tradução com 09 verbos atípicos: *think, laugh, conclude, believe, hear, go, find, cover* e *imagine*. Este tipo de relação representa 0.7% das ocorrências com SAY e 8.3% das ocorrências com DIZER.

Relação de tradução IV – equivalente não-verbal

Há 02 casos com SAY (1.3%) e 03 com DIZER (2.3%).

Relação de tradução V – omissão

A omissão ocorre 14.1% com SAY e 6.1% com DIZER.

6.Apontamentos finais

Este artigo analisou as relações de tradução de SAY/DIZER a partir do uso de um *corpus* combinado. Foram observados os equivalentes possíveis destes itens, a partir dos correspondentes encontrados no *corpus*. Apresentou-se, ainda, a metodologia da pesquisa e os primeiros passos da análise, visto que a metodologia não se desenvolve de forma linear, mas vai sendo construída a partir das primeiras etapas da pesquisa em análises piloto. Discutiu-se a terminologia utilizada para classificação dos diferentes tipos de *corpus*, observando-se que o *corpus* combinado oferece uma maior abrangência de possibilidades de pesquisa, permitindo ao pesquisador ampliar as perspectivas de análise. Foram descritas as características de um *corpus* combinado, bem como os procedimentos metodológicos para extração de dados. Foram introduzidos os programas computacionais *WordSmith Tools* e SPSS e as variáveis e categorias que podem ser usadas para pesquisa com estes programas. Mostrou-se como iniciar a pesquisa a partir dos dados brutos, que possibilitam ao pesquisador as primeiras observações a partir das quais a análise pode desenvolver-se. Desta forma, o trabalho ilustra a produtividade do uso de *corpus* combinado nos Estudos da tradução baseados em *corpora*.

Referências bibliográficas

ALVES, D. A. S. **Aspectos da representação do discurso em textos traduzidos**: os verbos de elocução neutros. 2006. Dissertação (Mestrado em Estudos Linguíticos). Faculdade de Letras, Universidade Federal de Minas Gerais.

ALVES, D. A. S.; MORINAKA, E. M. **Compilação de procedimentos metodológicos adotados por pesquisadores(as) em Estudos da Tradução e interfaces com as Linguísticas Sistêmico-Funcional e de *Corpus***. Disponível em < www.geocities.com/xalaskero/UFMG/Metodologia >. Acesso em < 15 junho 2007 >.

ASSIS, R. C. **A transitividade na representação de Sethe no *corpus* paralelo "Beloved-Amada".** 2004. Dissertação (Mestrado em Estudos Linguísticos). Faculdade de Letras, Universidade Federal de Minas Gerais.

BAKER, M. Corpus linguistics and translation studies: implications and applications. In: BAKER et al. (Ed.). **Text and technology**: in honour of John Sinclair. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1993. Cap., p. 233-250.

BAKER, M. Corpora in translation studies: an overview and some suggestions for future research. **Target**, Amsterdam, v. 7, n. 2, p. 223-243, 1995. crossref http://dx.doi.org/10.1075/target.7.2.03bak

BAKER, M. Corpus-based translation studies: the challenges that lie ahead. In: SOMERS, H. (Ed.). **Terminology, LSP and translation**: studies in language engineering in honour of Juan C. Sager. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1996. Cap., p. 177-186.

BERBER SARDINHA, T. **Linguística de corpus.** Barueri, SP: Manole, 2004.

BUENO, L. T. **Transitividade, coesão e criatividade lexical em "Macunaíma", de Andrade, e "Macunaíma", de Goodland.** 2005. Dissertação (Mestrado em Estudos Linguísticos). Faculdade de Letras, Universidade Federal de Minas Gerais.

CANÇADO, T. **Transitividade e representação do discurso no *corpus* paralelo "Interview with the Vampire/Entrevista com o Vampiro".** 2005. Dissertação (Mestrado em Estudos Linguísticos). Faculdade de Letras, Universidade Federal de Minas Gerais.

DAYRELL, C. **Investigating lexical patterning in translated Brazilian Portuguese:** a corpus-based study. 2005. 316 p. Tese (Doutorado em Estudos da Tradução). Centre for Translation and Intercultural Studies , The University of Manchester.

FEITOSA, M. P. **Uma proposta de anotação de *corpora* paralelos com base na linguística sistêmico-funcional.** 2005. Dissertação (Mestrado em Estudos Linguísticos). Faculdade de Letras, Universidade Federal de Minas Gerais.

FRANKENBERG-GARCIA, A. ***Corpora* e Tradução**. Disponível em < http://www.linguateca.pt/escolaverao2006/ >. Acesso em < 15 junho 2007 >.

GARCIA, O. M. **Comunicação em prosa moderna.** 13. ed. Rio de Janeiro: Fundação Getúlio Vargas, 1986.

HALLIDAY, M.A.K. Language theory and translation practice. **Rivista internazionale di tecnica della traduzione,** Triesta, n. 1 (pilot issue), p. 15-25, 1992.

HALLIDAY, M.A.K.; MATTHIESSEN, C. M. I. M. **An introduction to functional grammar.** 3rd edition, rev. ampl. London: Arnold, 2004. 689 p.

HANSEN, S.; TEICH, E. Multi-layer analysis of translation *corpora*: methodological issues and practical implications. In: **EUROLAN 2001**: Workshop on Multi-layer *Corpus*-based Analysis, 2001, **Proceedings…** Disponível em <http://www.coli.uni-saarland.de/publications/show.php?year=2001>, Acesso em < janeiro de 2008 >.

HANSEN, S. **The nature of translated text**: an interdisciplinary methodology for the investigation of the specific properties of translations. 2002. 245 p. Tese (Doutorado em Estudos da Tradução). Department of Applied Linguistics, Saarland University. (Saarbrucken dissertations in computational linguistics and language technology, v. 13).

JESUS, S. M.. **Representação do discurso e tradução**: padrões de textualização em *corpora* paralelo e comparável. 2004. 128 p. Dissertação (Mestrado em Estudos Linguísticos). Faculdade de Letras, Universidade Federal de Minas Gerais.

JESUS, S. M. **Relações de tradução**: SAY e DIZER em *corpora* de textos ficcionais. 2008. Tese (Doutorado em Linguística Aplicada). Faculdade de Letras, Universidade Federal de Minas Gerais.

JOHANSSON, S. On the role of corpora in cross-linguistic research. In: JOHANSSON, S; OKSEFJELL, S. (Ed). **Corpora and cross-linguistic research**: theory, method, and case studies. Amsterdam: Rodopi, 1998. Cap. 1, p. 3-24.

JOHANSSON, S.; OKSEFJELL, S. (Ed). **Corpora and cross-linguistic research**: theory, method, and case studies. Amsterdam: Rodopi, 1998. 376 p. (Language and computers: studies in practical linguistics, n. 24)

LAWSON, A. Collecting, aligning and analyzing parallel corpora. In: GHADESSY, M. et al. (Ed.). **Small corpus studies and ELT**. Theory and practice. Amsterdam: John Benjamins, 2001. Cap., p. 47-67.

MORINAKA, E. M. **"Gabriela, cravo e canela" and its (re)textualization in English**: representation through lexical relations. 2005. Dissertação (Mestrado em Estudos Linguísticos). Faculdade de Letras, Universidade Federal de Minas Gerais.

NEVES, M. H. de M. **A gramática funcional.** São Paulo: Martins Fontes, 1997. (Coleção Texto e Linguagem).

NEVES, M. H. de M. **Gramática de usos do português.** São Paulo: UNESP, 2000: 31-53.

OLIVEIRA, A. J. **Variação em itens lexicais terminados em /l/+vogal na região de Itaúna/MG.** 2006. Dissertação (Mestrado em Estudos Linguísticos). Faculdade de Letras, Universidade Federal de Minas Gerais.

OLOHAN, M. **Introducing corpora in Translation Studies**. London: Routledge, 2004. 220 p.

RODRIGUES, R. R. **A organização temática em "A hora da estrela" e "The hour of the star"**: um estudo de caso. 2005. Dissertação (Mestrado em Estudos Linguísticos). Faculdade de Letras, Universidade Federal de Minas Gerais.

ROTHE-NEVES**,** R**. Características cognitivas e desempenho em tradução: investigação em tempo real.** 2002. Tese (Doutorado em Estudos Linguísticos). Faculdade de Letras, Universidade Federal de Minas Gerais.

SCOTT, M. **WordSmith Tools.** 1999. Disponível em < http://www.lexically.net/wordsmith/ >Acesso em: 12 março 2003.

TAGNIN, S. E. O. (Org). **Cadernos de Tradução** IX (Número especial sobre Tradução e Corpora). Florianópolis: UFSC/NUT, 2002, n. 1, 278 p.

TAGNIN, S. E. O. Os *corpora*: instrumentos de auto-ajuda para o tradutor. **Cadernos de Tradução** IX, Tradução e Corpora. Florianópolis, 2002a, n. 1, p. 191-219.

TEICH, E. Towards a model for the description of cross-linguistic divergence and commonality in translation. In: STEINER, E.; YALLOP, C. (Ed.). **Exploring translation and multilingual text production**: beyond content. Berlin, New York: Mouton de Gruyter, 2001. Cap, p. 191-227.

TEICH, E. **Cross-linguistic variation in system and text**: a methodology for the investigation of translations and comparable texts. Berlin: Mouton de Gruyter, 2003. 276p. (Text, translation, computational processing, 5). crossref http://dx.doi.org/10.1515/9783110896541

THOMPSON, G. **Collins Cobuild English Guides 5**: Reporting. London: HarperCollins Publishers, 1994.

TOURY, G. **Descriptive translation studies and beyond.** Amsterdam: John Benjamins, 1995. 311 p. (Benjamins Translation Library).

VELA, M; HANSEN-SCHIRRA, S. The use of multi-level annotation and alignment for the translator. In: ASLIB Translating and the computer 28 conference, 2006, Londres. **Proceedings….** Disponível em < http://fr46.uni-saarland.de/croco/publication_en.html > Acesso em < 14 junho 2007 >

***Corpus* analisado**

AMADO, J. **Gabriela, cravo e canela.** Rio de Janeiro: Record, 1958.

AMADO, J. **Gabriela, clove and cinnamon.** Trad. Taylor, J. e Grossman . New York: Avon Books, 1962. (Tradução de *Gabriela, cravo e canela*).

ANDRADE, M. **Macunaíma**: o herói sem nenhum caráter. São Paulo e Belo Horizonte: Martins e Itatiaia, 1980.

ANDRADE, M. **Macunaíma**. Tradução de E. A. Goodland. London: Quartet Books, 1984. (Tradução de: *Macunaíma: o herói sem caráter*).

HUXLEY, A. **Point counter point**. London: Flamingo, 1994. (Coleção Modern Classic).

HUXLEY, A. **Contraponto**. Trad. Érico Veríssimo. Porto Alegre: Editora Globo, 1971. (Tradução de: *Point counter point* – Coleção Imortais da Literatura).

LISPECTOR, C. **A hora da estrela.** Rio de Janeiro: Rocco, 1998.

LISPECTOR, C. **The hour of the star**. Trad. Giovanni Pontiero. New York: New Directions Books, 1992. (Tradução de *A hora da estrela*).

MORRISON, T. **Beloved** . New York: Alfred A. Knopf, 1998. 275p.

MORRISON, T. **Amada**. Tradução de Evelyn Kay Massaro. São Paulo: Ed. Best Seller, 1987. 321p. (Tradução de: *Beloved*).

RICE, A. **Interview with the vampire.** New York: Ed. Ballantine Books, 1997.

RICE, A. **Entrevista com o vampiro**. Trad. Clarice Lispector. Rio de Janeiro: Rocco, 1996. (Tradução de: *Interview with the vampire*).

Artigo recebido em: 15.10.2014
Artigo aprovado em: 25.11.2014

A equivalência tradutória de Partículas Modais: um estudo baseado em *corpus*

Translation equivalence of Modal Particles: a *corpus*-based study

Adriana Silvina Pagano[*]
Arthur de Melo Sá[**]
Kícila Ferreguetti[***]

RESUMO: Este artigo apresenta resultados de um estudo sobre Partículas Modais realizado pelo grupo de pesquisa "*Modelagem sistêmico-funcional da tradução e da produção textual multilíngue*", do Laboratório Experimental de Tradução da UFMG. Segundo Figueredo (2011), as Partículas Modais são recursos gramaticais utilizados por falante e ouvinte com a função de negociar os seus papéis durante a interação. Por serem uma particularidade do sistema linguístico do português brasileiro e seu uso estar mais circunscrito à linguagem falada, as Partículas Modais são uma categoria escassamente representada em *corpora* de grandes dimensões e pouco estudada, constituindo um rico campo de pesquisa, sobretudo nos estudos multilíngues uma vez que podem representar um problema de tradução. Para investigar essa categoria, foi compilado um *corpus* paralelo bilíngue português brasileiro – inglês, formado por histórias seriadas da Turma da Mônica e suas respectivas traduções para o inglês. O objetivo da pesquisa foi identificar quais Partículas eram utilizadas com mais frequência no *corpus*, como elas foram traduzidas para o inglês, e se era possível verificar um padrão para as opções tradutórias. A metodologia foi dividida em três etapas principais: 1) compilação, preparação e alinhamento do *corpus*; 2) busca e extração das linhas de concordância contendo Partículas Modais em português brasileiro e suas traduções para o inglês e 3) anotação das ocorrências segundo a descrição proposta por Figueredo (2011) e com base em cada opção tradutória verificada. Os

ABSTRACT: This paper reports the main results of a study on Modal Particles carried out by the research group *Systemic-functional modeling of translation and multilingual text production* at LETRA/FALE/UFMG. As defined by Figueredo (2011), Modal Particles are grammatical resources used by speaker and listener to negotiate their roles during their interaction. Due to their unique status as a grammatical resource in Brazilian Portuguese and the fact that their use is mostly confined to spoken language, Modal Particles are a category that is underrepresented in large *corpora* and scarcely studied, thus constituting a rich field of research in multilingual studies, particularly in view of their being a potential source for translation problems. In order to investigate this category, a bilingual parallel *corpus* was compiled with comic strips from *Turma da Mônica* in Brazilian Portuguese and their translation into English. The aim was threefold: 1) to identify which Modal Particles were most frequent in the *corpus*; 2) to verify how they were translated into English and 3) to observe whether translation patterns can be found. The analysis comprised three stages. First, the *corpus* was compiled, prepared and aligned; then it was queried for occurrences of Modal Particles that were extracted along with their translations. After that, they were annotated according to the description on Figueredo (2011) and based on each translation choice. Results revealed significant frequencies of occurrence for some of the Particles as well as patterns in their translation equivalents in English.

[*] Universidade Federal de Minas Gerais. apagano@ufmg.br
[**] Universidade Federal de Minas Gerais. arthurdemelosa@gmail.com
[***] Universidade Federal de Minas Gerais.kicilaferreguetti@yahoo.com.br

resultados evidenciaram frequências de ocorrência significativas para algumas das Partículas encontradas bem como padrões nos equivalentes tradutórios na língua inglesa.

PALAVRAS-CHAVE: Estudos da Tradução baseados em *corpus*. *Corpus* Paralelo Bilíngue. Equivalência tradutória. Partículas Modais. Linguística Sistêmico-Funcional.

KEYWORDS: Corpus-based Translation Studies. Parallel Bilingual Corpus. Translational Equivalence. Modal Particles. Systemic Functional Linguistics.

1. Introdução

O presente artigo apresenta uma análise de uma categoria gramatical pouco estudada no português brasileiro[1], tendo como base dados obtidos de um *corpus* de textos nessa língua e seus equivalentes tradutórios numa segunda língua, neste caso, o inglês. Trata-se das Partículas Modais utilizadas de maneira geral na interação oral entre falante e ouvinte com a função de validar o argumento do falante durante essa interação. As Partículas Modais representam um desafio para seu estudo, uma vez que demandam a compilação de um *corpus* oral, devidamente transcrito de forma a preservar essas construções, o que poderia ser encontrado em alguns estudos do português falado.

Todavia, se desejamos examinar equivalentes tradutórios dessas Partículas numa outra língua, passíveis de ser obtidos de um *corpus* paralelo bilíngue, o desafio torna-se ainda maior, haja vista que são escassos os textos orais que são traduzidos para o inglês e compilados na forma de *corpus*. A forma encontrada para superar esta limitação foi a compilação de um *corpus* de textos escritos do português brasileiro, que possuem tradução publicada em inglês e que pelas características do discurso – histórias em quadrinhos que visam reproduzir alguns aspectos da oralidade – incluem ocorrências de Partículas Modais. O *corpus* compilado é uma seleção de histórias em quadrinhos da Turma da Mônica e suas traduções para o inglês, nas quais temos a encenação de diálogos e narrativas majoritariamente entre crianças, as quais se valem de Partículas para as diversas funções interativas.

[1] As descrições da língua portuguesa em textos de gramática publicados no Brasil não contemplam as Partículas Modais. Apenas Castilho (2010) dedica um capítulo de sua gramática à linguagem da conversação e apresenta uma relação de "marcadores discursivos" que incluem algumas das realizações do que se entende neste estudo por das Partículas Modais. O autor recolhe resultados de estudos anteriores sobre "marcadores discursivos", notadamente em Neves (1999). Cumpre destacar que "marcadores discursivos" abrangem diversas classes de palavra e funções assim classificados sob uma perspectiva funcionalista da linguagem, a qual difere da perspectiva sistêmico-funcional aqui adotada.

As Partículas Modais pesquisadas neste *corpus* são aquelas descritas por Figueredo (2011) em sua *Introdução ao perfil metafuncional do português brasileiro* e sua descrição está baseada na teoria sistêmico-funcional (HALLIDAY; MATTHIESSEN, 2004; CAFFAREL; MARTIN; MATTHIESSEN, 2004). O objetivo foi verificar quais Partículas eram utilizadas com mais frequência no *corpus* compilado com histórias seriadas da Turma da Mônica e como elas foram traduzidas para o inglês, com vistas a examinar como o significado de Partículas Modais do português brasileiro é realizado nas traduções em inglês e o quê o mesmo revela sobre o seu funcionamento.

O estudo das Partículas Modais em português brasileiro oferece a oportunidade de se preencher uma lacuna nos estudos descritivos baseados em *corpus*, que como foi dito não contam com *corpora* com um número significativo de ocorrências dessa categoria gramatical. Complementando as observações no *corpus* monolíngue, os equivalentes tradutórios obtidos de um *corpus* paralelo bilíngue oferecem indícios de como os significados realizados pelas Partículas Modais podem ser construídos em outras línguas. Em uma perspectiva aplicada, as Partículas possuem um grande potencial de pesquisa no campo dos Estudos da Tradução, uma vez que, por se tratarem de uma particularidade do sistema linguístico do português brasileiro, podem configurar-se como um problema de tradução.

O estudo foi realizado no escopo dos projetos do grupo de pesquisa "*Modelagem sistêmico-funcional da tradução e da produção textual multilíngue*", do Laboratório Experimental de Tradução (LETRA) da UFMG.

2. Revisão da literatura

De acordo com Caffarel, Martin e Matthiessen (2004), as línguas possuem um potencial de significação[2] que incorpora três tipos de significado: o ideacional, o interpessoal e o textual. Esses significados recebem o nome de metafunções. Assim, a teoria sistêmico-funcional considera que as metafunções ideacional, interpessoal e textual são as responsáveis por construir de forma simultânea a mensagem em qualquer texto.

O foco de estudo desta pesquisa é a metafunção interpessoal, que como Matthiessen e Halliday (2009) afirmam, é a responsável pela *interação* entre falante e ouvinte(s). Ainda de acordo com os autores, a metafunção interpessoal possui os recursos necessários para, em

[2] *Potencial de significação* é a tradução adotada por Figueredo (2011) para *meaning potential*.

uma interação dialógica, realizarem a encenação de papéis sociais (de maneira geral) e de papéis discursivos (de maneira específica) (MATTHIESSEN e HALLIDAY, 2009).

O sistema de MODO é uma das principais categorias gramaticais da metafunção interpessoal, e dentro dele, encontra-se um componente específico: o sistema de VALIDAÇÃO, que é realizado pelas Partículas Modais. Ambos serão detalhados a seguir.

2.1 A metafunção interpessoal

Como definido por Martin e White (2005, p. 7), "os recursos interpessoais lidam com a negociação de relações sociais: como as pessoas interagem, incluindo os sentimentos que tentam compartilhar"[3]. Partindo-se desta ideia, a língua pode ser analisada e descrita com base nas relações entre falante e ouvinte, e esta é a proposta desta pesquisa: analisar a relação entre falante e ouvinte por meio da observação das Partículas Modais. Na metafunção interpessoal encontram-se os sistemas de MODO e VALIDAÇÃO, assim como os conceitos de função discursiva, proposição e proposta, discutidos a seguir.

De acordo com Matthiessen e Halliday (2009):

> ao interagir com o outro, o ser humano se insere em uma variedade de relacionamentos interpessoais, fazendo escolhas entre diferentes estratégias semânticas, como bajular, persuadir, incitar, solicitar, ordenar, sugerir, asseverar, insistir, duvidar, etc. A gramática fornece os recursos básicos para a expressão dessas funções discursivas na forma de um conjunto de *sistemas* da oração altamente generalizados conhecidos como MODO[4].

Dessa maneira, o MODO é um sistema para a expressão da mensagem que o falante quer passar. Este faz determinadas escolhas neste sistema e, assim, produz seu discurso por meio da fala ou da escrita.

Esta pesquisa se detém à análise no estrato da lexicogramática. O sistema de MODO realizado neste nível é semelhante para o português brasileiro e para o inglês, sendo realizado na forma de duas opções principais: indicativo e imperativo. O indicativo, por sua vez, é ainda

[3] Nossa tradução para: "Interpersonal resources are concerned with negotiating social relations: how people are interacting, including the feelings they try to share".
[4] Nossa tradução para: "In interacting with one another, we enter into a range of interpersonal relationships, choosing among semantic strategies such as cajoling, persuading, enticing, requesting, ordering, suggesting, asserting, insisting, doubting, and so on. The grammar provides us with the basic resource for expressing these speech functions, in the form of a highly generalized set of clause **systems** referred to as MOOD" (grifo do autor).

realizado por duas outras opções: declarativo e interrogativo. Sendo assim, esta pesquisa baseia sua análise nessas três opções iniciais do sistema de MODO: imperativo, declarativo e interrogativo. A Figura 1, adaptada de Matthiessen e Halliday (2009), representa esse sistema.

Dentro da concepção da metafunção interpessoal também é definido o conceito de *função discursiva*. De acordo com Figueredo (2011, p. 173), são "funções semânticas responsáveis pela troca de informação ou de bens-&-serviços de um falante que assume determinado papel na troca". Entende-se, então, que o falante pode assumir diferentes papéis em uma interação (troca). Nessa troca, o produto pode ser informação ou bens-&-serviços. Quando o produto da troca é uma informação, esta troca recebe o nome de *proposição*, e quando são bens-&-serviços, recebe o nome de *proposta* (cf. CAFFAREL, MARTIN e MATTHIESSEN, 2004).

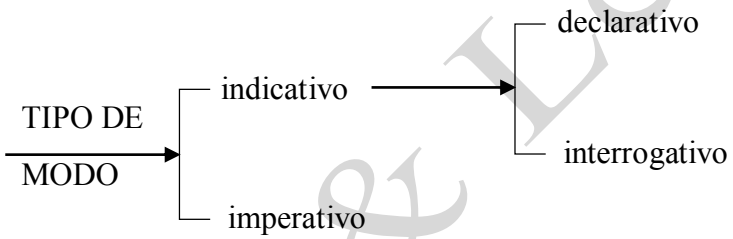


Figura 2 – Rede do sistema de MODO do português brasileiro e do inglês.
Fonte: Traduzida e adaptada de Matthiessen e Halliday (2009).

Figueredo (2011, p. 174) ainda explica que "as proposições se caracterizam pela troca de declarações e perguntas, enquanto as propostas pela troca de comandos e ofertas". Tem-se então uma relação entre esses quatro conceitos: MODO, função discursiva, proposição e proposta. Sendo assim, as diferentes combinações de mercadorias e papéis formam as quatro funções discursivas iniciais do português brasileiro e do inglês: declaração, pergunta, comando e oferta; e essas funções discursivas são realizadas prototipicamente por determinadas opções do sistema de MODO. A relação apresentada aqui pode ser melhor visualizada no Quadro 1, adaptado de Figueredo (2011).

Quadro 2 - Relação entre MODO, função discursiva, proposta e proposição.

	Informação	*bens-&-serviços*
Fornecer	[DECLARAÇÃO] proposição MODO declarativo	[OFERTA] proposta sem MODO específico
Demandar	[PERGUNTA] proposição MODO interrogativo	[COMANDO] proposta MODO imperativo

Fonte: Adaptado de Figueredo (2011, p. 175)

É possível observar no quadro as quatro funções discursivas iniciais. A primeira é a opção de fornecer informação: uma proposição denominada declaração e realizada por orações no modo declarativo. Já a segunda é a opção de fornecer bens-&-serviços: uma proposta denominada oferta e não realizada por um modo específico. A terceira, por sua vez, é a opção de demandar informação: uma proposição denominada pergunta e realizada por orações no modo interrogativo. A quarta, e última, é a opção de demandar bens-&-serviços: uma proposta denominada comando e realizada por orações no modo imperativo.

Esses conceitos e a relação entre eles são relevantes para esta pesquisa, pois muitas Partículas Modais somente são realizadas em um determinado ambiente do MODO. Assim, as Partículas são realizadas na oração de forma contínua ao sistema de MODO e realizam funções discursivas juntamente com este sistema (cf. FIGUEREDO, 2011, p. 217), como veremos a seguir.

2.2 O sistema de VALIDAÇÃO do português brasileiro

Em português brasileiro, a metafunção interpessoal também apresenta o sistema de VALIDAÇÃO ilustrado na Figura 2, abaixo. De acordo com Figueredo (2011, p. 218), este sistema é "realizado por Partículas Modais e tem a função de validar as proposições ou propostas enquanto um argumento da troca". O autor também afirma que a principal função das Partículas Modais é "negociar entre os interlocutores o papel do falante em uma proposição ou proposta a fim de que esta se torne um significado compartilhado na interação" (FIGUEREDO, 2011, p. 220).

Além disso, o autor argumenta que os resultados da análise das Partículas Modais em sua pesquisa demonstram que, em geral, "os falantes empregam os recursos deste sistema quando necessitam que seus interlocutores endossem seu papel de 'falante' na troca" (FIGUEREDO, 2011, p. 221). Dessa forma, o que se percebe é que o sistema de VALIDAÇÃO

tem a função de negociar o papel do falante com relação ao ouvinte, e os papéis de ambos com relação à proposição ou proposta.

Ainda segundo o autor:

> Na ordem da oração, as Partículas do sistema de VALIDAÇÃO encerram duas funções interpessoais complementares: indicam a forma pela qual a oração deve ser validada em termos de concordar, assentir, exortar, etc.; e são retomadas pelo ouvinte como forma de dar continuidade a troca. Em muitos casos [...] elas respondem sozinhas por um argumento. (FIGUEREDO, 2011, p. 229)

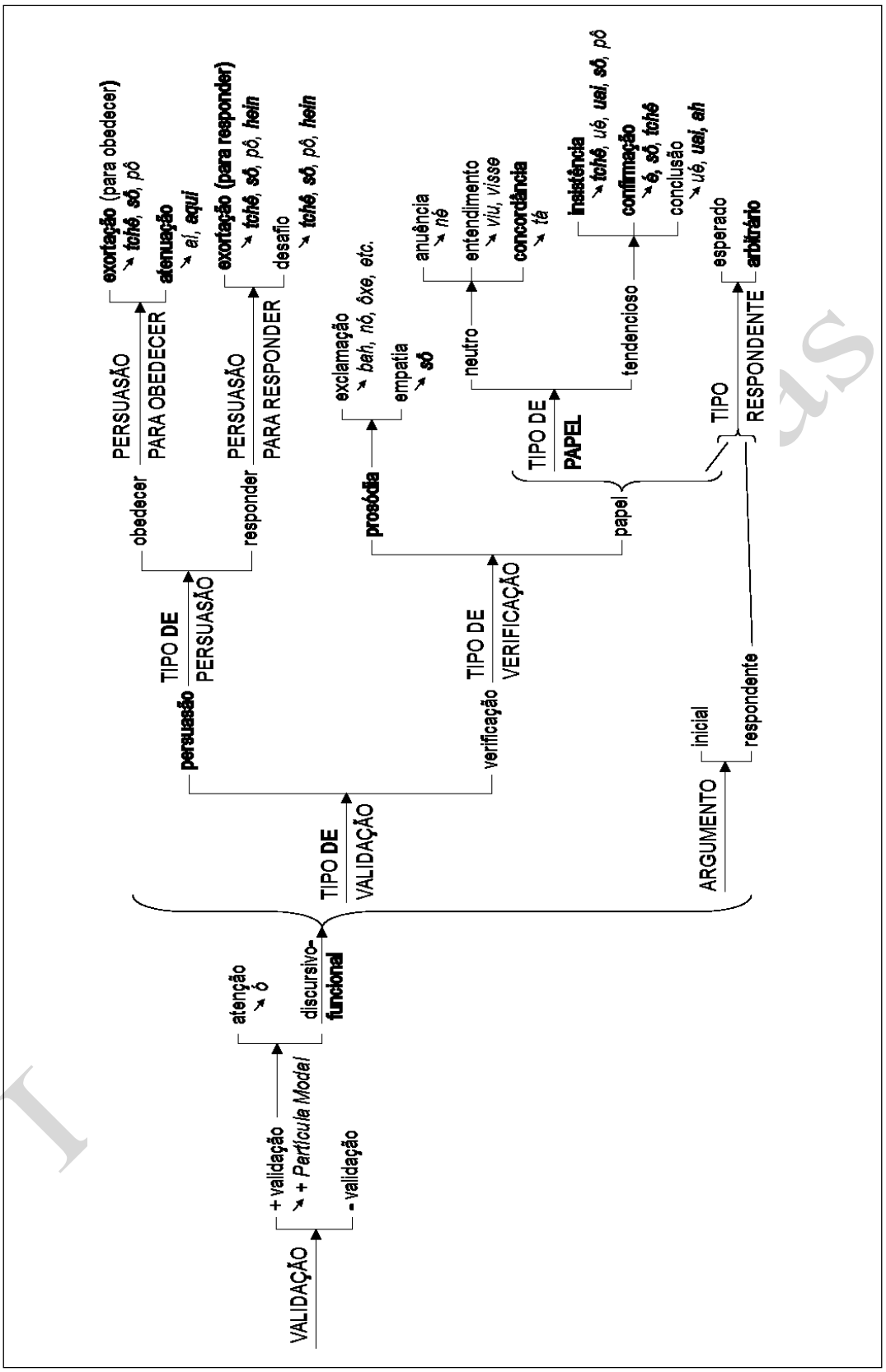


Figura 3 – Sistema de VALIDAÇÃO
Fonte: Adaptada de Figueredo (2011, p. 234)

Portanto, o sistema de VALIDAÇÃO tem a função não só de avaliar os papéis do falante e do ouvinte, mas também de, por meio dessa validação, dar continuidade ao discurso, pois somente mediante a validação do ouvinte é que o falante dá prosseguimento ao discurso. De outra forma, caso o ouvinte não valide algum (ou todos) dos papéis em questão, há uma quebra no discurso, dando início a uma nova troca por parte do próprio ouvinte.

Sabendo que o sistema de VALIDAÇÃO atua na construção da mensagem de forma contínua ao MODO (FIGUEREDO, 2011, p. 2017), este lhe acrescenta mais um nível de delicadeza. Assim, quando o falante constrói uma proposta ou proposição e esta apresenta uma Partícula Modal, há uma seleção primária pelo TIPO DE MODO, seguida de uma seleção secundária de acordo com a categoria de avaliação. Dessa forma, o sistema de VALIDAÇÃO não só é influenciado por escolhas feitas no sistema de MODO, mas também participa dessas escolhas.

Por fim, o estudo de Figueredo (2011) ainda define opções mais delicadas do sistema de VALIDAÇÃO ilustradas na Figura 2 e exemplificadas no Quadro 2 (a seguir) juntamente com as realizações de todas as Partículas. Para fins deste artigo, serão enfocadas as Partículas Modais dos tipos *Anuência* e *Confirmação*. A escolha dessas Partículas foi embasada por testes estatísticos que revelaram que esses dois tipos foram os únicos que apresentaram relevância estatística na sua tradução; no caso desta pesquisa, apresentaram um valor-*p* maior que 0,05 – portanto, em conformidade com H_0 e dentro da área de não-rejeição da distribuição chi-quadrado[5].

A Partícula de *Anuência* é realizada em proposições, que, por sua vez, são realizadas por orações declarativas. Figueredo (2011) afirma que:

> esta Partícula possui a função de pedir ao ouvinte que dê seu assentimento ao falante para que a proposição possa se tornar parte do "conhecimento compartilhado" entre eles. Neste caso, o ouvinte não precisa, necessariamente, concordar com a opinião do falante, mas apenas dar seu aval para que o falante permaneça no lugar de "avaliador da proposição". (p. 225)

O falante, portanto, pede o aval do ouvinte para que sua proposição seja compartilhada entre eles, consolidando assim o seu papel na troca. Isso permite que ele prossiga com o argumento e que o ouvinte permaneça nessa posição. Como exemplo, pode-se ver no

[5] Os testes estatísticos foram realizados no ambiente de programação R (R CORE TEAM, 2014) e serão detalhados na seção de metodologia.

Exemplo 1[6] que o falante (*CBe*) pede ao ouvinte (*Zdr*) que dê o seu aval para que a proposição seja compartilhada. O ouvinte valida o papel do falante e ainda constrói uma nova proposição complementar à do falante, demonstrando que validou o papel de *CBe*. O contexto neste exemplo é: os dois personagens estão pegando içás (um tipo de formiga) para fazer um prato culinário. Alguns içás escapam, enquanto outros, os machos (*bitus*), vão atrás destes.

Exemplo 1 – Partículas de Anuência:
CBe Tem uns qui inda iscapa, né, Zé?
Zdr É! Com os noivos bitus voando atrás.

No que diz respeito à Partícula de *Confirmação*, Figueredo (2011) argumenta que o uso dessa Partícula ocorre da seguinte maneira: "o ouvinte constrói uma proposição – ou realiza uma ação – sobre a qual o falante constrói uma proposição imaginando que o ouvinte a compreendeu e irá negociá-la da maneira esperada" (FIGUEREDO, 2011, p. 227).

O Exemplo 2 demonstra o uso dessa Partícula em uma situação na qual o ouvinte construiu uma proposição, que é levada em consideração pelo falante da maneira descrita acima. O contexto é a chegada do personagem *Fra* na base secreta de *Nim* e *DCm*. Estes últimos questionam como o primeiro encontrou a base, que revela a fonte dessa informação (*Joaquim da padaria*). O personagem *Nim* infere quem havia revelado a *Joaquim* o local da base: *Mag*; e então produz uma proposição direcionada a *Mag*, mas que é respondida por *Fra*, que confirma a inferência de *Nim*. Dessa forma, a proposição de *Nim* é negociada da maneira esperada.

Exemplo 2 – Partículas de Confirmação:
Nim Ei! Como você entrou aqui?
Fra Pela porta.
DCm Não! Como achou meu esconderijo?
Fra O Joaquim da padaria me contou.
Nim Padaria, hein?
Mag Glup!
Fra É! Ela deu o endereço pra entrega de quinhentos pãezinhos!

O Exemplo 3 demonstra o uso dessa Partícula numa situação na qual o ouvinte realizou uma ação. Dessa forma, o falante constrói sua proposição em relação a esse comportamento, e

[6] Todos os exemplos apresentados ao longo do texto foram extraídos do *corpus* da pesquisa.

espera que o ouvinte a compreenda e negocie da forma esperada. No entanto, o ouvinte não negocia da forma esperada, empregando uma Partícula (*não*) na negociação. O contexto aqui é o personagem *Pir* fugindo do personagem *Cas* em uma luta com espadas.

Exemplo 3 – Partículas de Confirmação:
Cas Fugindo da minha espada, hein?
Pir Não! Fugindo do seu cheiro!

É importante ainda elucidar que o sistema de VALIDAÇÃO possui dois tipos de argumento: inicial e respondente. O argumento inicial é aquele no qual o falante requisita por parte do ouvinte a validação de sua proposição ou proposta. O argumento respondente é aquele no qual o ouvinte faz a sua validação do que foi apresentado pelo falante. O movimento respondente ainda é realizado por duas opções mais delicadas: esperado ou arbitrário; pois o ouvinte pode responder conforme o esperado pelo falante, ou pode responder de forma arbitrária, invalidando a proposição ou proposta do falante. No caso de ofertas, demandas e perguntas elementais, o argumento respondente pode não apresentar Partículas Modais. Mesmo assim, nesses casos o ouvinte valida a mensagem do falante ao reagir da maneira esperada por ele (aceite, obediência ou resposta), ou então o ouvinte invalida a mensagem, ao reagir de maneira arbitrária (rejeição, recusa ou renúncia).

3. Metodologia

Para a realização desta pesquisa, foi compilado um *corpus* paralelo bilíngue unidirecional (português brasileiro – inglês) formado por onze histórias seriadas da Turma da Mônica, criadas por Maurício de Souza, e suas respectivas traduções para o inglês[7]. O *corpus* é constituído de aproximadamente 38.205 *tokens*, distribuídos conforme a Tabela 1, abaixo.

Para a compilação do *corpus*, as histórias seriadas da Turma da Mônica foram transcritas, seguindo alguns critérios previamente definidos, tais como: 1) os personagens são indicados por siglas referentes aos seus nomes em português brasileiro. A mesma sigla foi utilizada nos textos em inglês, visando facilitar o alinhamento e as buscas no *corpus*; 2) as transcrições não apresentam itálicos, negritos ou outro tipo de formatação adicional de estilo

[7] Disponível em: <http://www.monica.com.br/comics/seriadas.htm>. Acesso em 14 de outubro de 2014.
O *corpus* foi compilado com anuência dos editores da revista; seu acesso é restrito aos pesquisadores do LETRA/UFMG.

do texto, pois os textos foram compilados em arquivo eletrônico do tipo *.txt*, que não permite esse tipo de formatação; 3) aspas, parênteses, vírgulas e reticências foram removidos, pois podem causar conflitos quando o texto é submetido a algum software para a análise do *corpus*; e 4) foram mantidas as letras maiúsculas de nomes próprios após ponto final, ponto de interrogação, ponto de exclamação, e no início de cada sentença. Ao final do processo de transcrição e obedecida a padronização exemplificada acima, as compilações em ambas as línguas foram armazenadas em arquivos eletrônicos de texto com a extensão *.txt*.

Tabela 4
Composição do *corpus* em número de tokens

	Batmenino? Eternamente	Coelhada nas estrelas	Comandante gancho	Horacic Park	Mônica e os bárbaros	O unicórnio	Os doze trabalhos da Mônica	Ou nós acabamos com as formigas...	Que furada de reportagem!	Romeu e Julieta	Superparque	Total
português	1730	908	570	1096	2803	1446	1536	1707	3112	3275	734	18917
Inglês	1857	947	597	1057	2987	1507	1540	1673	3181	3181	761	19288
Total	3587	1855	1167	2153	5790	2953	3076	3380	6293	6456	1495	38205

Após a transcrição, as compilações nas duas línguas foram alinhadas no nível da sentença, utilizando o *software* ParaConc (Barlow, 2012). A Figura 3 ilustra esta etapa.

A busca por Partículas Modais foi feita com o auxílio da ferramenta de pesquisa do *ParaConc,* tomando o português brasileiro como ponto de partida e baseada no Quadro 2 abaixo, proposto por Figueredo (2011), onde o autor lista as Partículas Modais observadas em sua pesquisa.

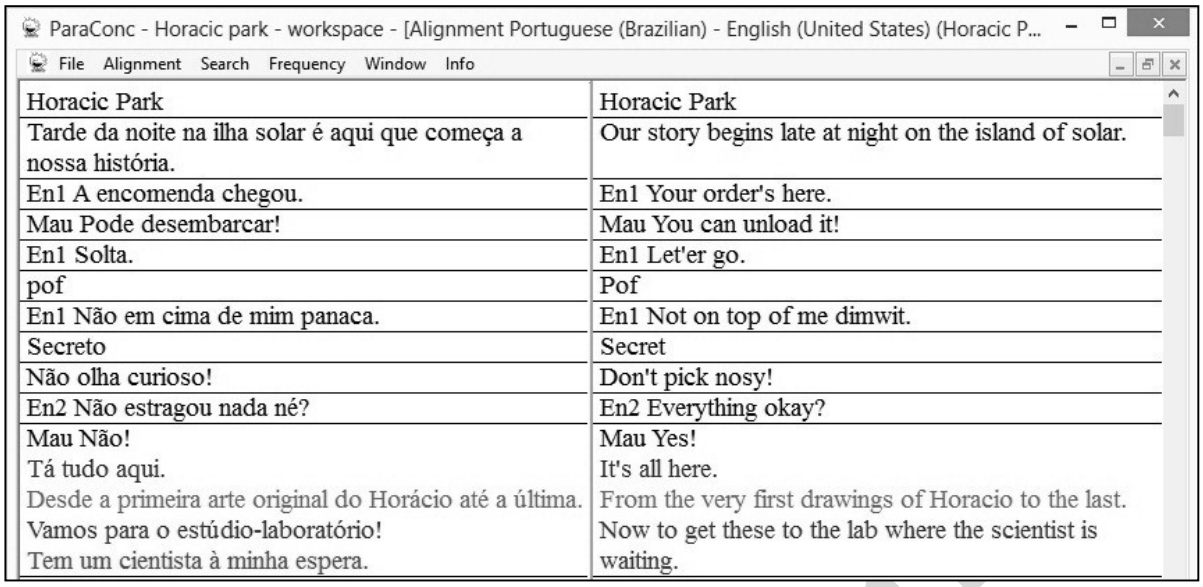


Figura 4 – Captura de tela do *ParaConc*: Alinhamento.

Quadro 3 - Partículas Modais de acordo com Figueredo (2011).

Ambiente para as partículas		**Partículas Modais**			
orientação e mercadoria	**ambiente do MODO**	**Função**	***Inicial***	***Respondente esperada***	***arbitrária***
		ATENÇÃO	*ó*	*tá*	*não*
fornecer informação	Declaração: indicativo declarativo	**ANUÊNCIA**	*né*	*é*	*né não*
		CONCORDÂNCIA	*tá*	*tá*	*não*
		INSISTÊNCIA	*tchê, ué, sô, pô*	*ah, é; ah, tá; tá*	*não sô; não ué; não tchê*
		CONCLUSÃO	*ué, uai, ah*	*é*	*ah*
		ENTENDIMENTO	*viu, visse*	*viu*	*não*
		CONFIRMAÇÃO	*hein, é*	*tá (curto), é*	*tá (longo)*
		EMPATIA	*sô*	---	---
		EXCLAMAÇÃO	*bah, nó, oxe, pô, tchê, etc.*	---	---
demandar informação	Pergunta Polar: indicativo interrogativo	**CONFIRMAÇÃO**	*é, sô, tchê*	*é*	*não*
	Pergunta Elemental: indicativo interrogativo	**EXORTAÇÃO p/ responder** **DESAFIO**	*tchê, sô, pô, hein*	resposta	renúncia
demandar bens-e-	Comando: imperativo	**EXORTAÇÃO p/ obedecer**	*sô, tchê, pô*	obediência	recusa

serviços		**ATENUAÇÃO p/ obedecer**	*aí, aqui*	obediência	recusa
Fornecer bens-e-serviços	Oferta: imperativo ou indicativo interrogativo	**ATENUAÇÃO p/ aceitar**	*tchê, sô, aí, aqui*	aceite	rejeição

Fonte: Figueredo (2011, p. 222).

Essa busca no *corpus* gerou listas de concordância por meio das quais foi possível classificar os itens lexicais como Partículas Modais ou não. Os itens identificados com função de Partícula foram ainda classificados segundo os níveis mais delicados do sistema de VALIDAÇÃO, enquanto os itens identificados com outras funções não foram analisados mais profundamente, sendo assim descartados.

É importante mencionar que, de acordo com Figueredo (2011), a diferença entre algumas Partículas, como é o caso de *tá (longo)* e *tá (curto),* é prosódica, sendo passível de ser observada pela extensão da sílaba tônica. Essa diferença pode ser percebida em textos orais transcritos com essa anotação; porém, a menos que haja uma diferenciação gráfica para essa extensão da sílaba tônica, ela é impossível de ser observada em textos escritos, como é o caso do presente corpus, composto por textos escritos que não apresentam essa diferenciação gráfica. No presente estudo, optou-se pela busca de ocorrências de *tá* em geral e uma vez localizadas, proceder-se a sua análise numa tentativa de categorização delas como Partícula de Confirmação em um argumento respondente, isto é, como *tá (curto)* e *tá (longo)*.

Esse tipo de análise foi possível graças ao princípio da agnação, que, segundo Figueredo (2011, p. 85), "significa trabalhar tanto com as realizações, quanto com outras possíveis realizações para um mesmo elemento". Ainda de acordo com o autor,

> constitui-se como um princípio analítico a utilização de acréscimos, substituições e subtração de itens e funções; a inversão da estrutura que as realiza; e a possibilidade de expansão ou retração dos termos do sistema para determinada função, observando sempre as mudanças no registro. Tais "exercícios" de possibilidade são adotados como critério para se entender quais os itens que compõem um determinado sistema e os que compõem sistemas diferentes. (FIGUEREDO, 2011, p. 85)

Como ilustração desse princípio, observa-se o item *né* no Exemplo 4 a seguir.

Exemplo 4 – Partícula de Anuência:
Mon Hum, você não me trouxe aqui só pra não apanhar, né?
Ceb E-eu? Clalo que não.

É possível classificar o *né* como uma Partícula de Anuência comparando-o com outros tipos agnatos de Partícula, por exemplo: de Entendimento e de Concordância, que também realizam função discursivo-funcional: de verificação: de papel: neutra – como pode ser observado na Figura 2. Isso porque se pode fazer a substituição do *né* pelas realizações dos outros tipos de Partícula, por exemplo: 1) pela de Entendimento *viu*: "*você não me trouxe aqui só pra não apanhar, viu*"; e 2) pela de Concordância *tá*: "*você não me trouxe aqui só pra não apanhar, tá*".

Com essa substituição, verifica-se que é mantido o contraste na delicadeza entre esses três tipos de Partícula, pois cada uma realiza uma função diferente na oração. Além disso, também observa-se que o contraste é mantido em relação às Partículas de Insistência (e.g. *pô*), de Confirmação (e.g. *sô*) e de Conclusão (e.g. *uai*), pois estas realizam função discursivo-funcional: de verificação: de papel: tendenciosa na oração (cf. Figura 2). Dessa forma, ao substituir o *né* por outros itens com outras funções, é possível classificá-lo como uma Partícula de Anuência, pois os contrastes com as outras opções neutras e com as opções tendenciosas dentro do sistema de TIPO DE PAPEL são mantidos.

Após a obtenção e extração das ocorrências das Partículas Modais e suas traduções, estas foram arquivadas no formato .txt e passou-se à sua análise propriamente. Para isso, foram estabelecidos alguns parâmetros e utilizadas algumas ferramentas, ambos apresentados a seguir.

Primeiramente, foi feita a análise quantitativa utilizando o ambiente de programação R (R CORE TEAM, 2014), com base nos dados extraídos pelo *ParaConc*, que possibilitou observar o número total de ocorrências de cada item lexical pesquisado nos textos em português brasileiro. Esses itens foram então analisados individualmente, para verificar se eram ou não uma Partícula. Foi então feita uma comparação entre o número de ocorrências totais do item lexical e o número de vezes em que ele é usado como Partícula. Com isso, foi feito o cálculo da frequência relativa entre as ocorrências totais dos itens e as ocorrências de Partículas. Como exemplo, o item lexical *não* ocorreu 432 vezes; desse total, somente duas ocorrências configuram Partículas Modais; isso representa uma frequência relativa de 0,46%, na qual, dentre todas as ocorrências de *não*, somente 0,46% são Partículas Modais.

Além disso, cada item lexical foi analisado qualitativamente com o objetivo de definir quais tipos de Partícula Modal eram realizados por eles. Esta análise teve como base o

Sistema de VALIDAÇÃO do português brasileiro proposto por Figueredo (2011), ilustrado pela Figura 2 anteriormente. Caso os itens lexicais se tratassem de Partículas Modais, eles eram anotados, de acordo com as opções mais delicadas do sistema, como *Anuência* ou *Confirmação*, por exemplo. Para a realização desta etapa de anotação semiautomática, foi usado o *software* UAM Corpus Tool (O'Donnell, 2014).

No que diz respeito às traduções das Partículas Modais para o português, cada ocorrência foi anotada manualmente segundo a opção tradutória, visando observar como foi realizada a tradução de cada Partícula. Cada opção tradutória foi quantificada, possibilitando a investigação de padrões nas realizações dos itens em inglês.

Por fim, os dados foram analisados estatisticamente (cf. GRIES, 2013) no ambiente de programação *R* (R Core Team, 2014). Os testes utilizados foram o teste exato de Fisher para a comparação de proporções e o teste chi-quadrado para a verificação da distribuição dos dados e da igualdade de proporções.

Para o teste exato de Fisher, a hipótese nula é de que as proporções são iguais entre os itens com observações positivas e negativas da variável, e a hipótese alternativa é de que essas proporções são diferentes. Por exemplo: a hipótese nula sugere que as proporções são iguais entre 1) o item *né* com função de Partícula e com outra função e 2) o item *ó* com função de Partícula e com outra função, enquanto a alternativa sugere que essas proporções são diferentes.

Para o teste de aderência chi-quadrado, a hipótese nula é de que a amostra (neste caso as ocorrências de Partículas Modais nos textos selecionados) se encontra numa distribuição chi-quadrado, e que, portanto, a população (isto é, as ocorrências de Partículas no português brasileiro) possui o mesmo comportamento, enquanto a hipótese alternativa é de que a amostra não se encontra nessa distribuição, e consequentemente a população se comporta de maneira diferente. Por exemplo: segundo a hipótese nula, a população se comporta da mesma forma que a amostra de Partículas Modais encontradas no *corpus*, mas, segundo a hipótese alternativa, a população se comporta de maneira diferente.

Além disso, o teste chi-quadrado pode ser aplicado somente aos itens da pesquisa de forma independente com o objetivo de testar a igualdade entre as proporções das observações das variáveis. A hipótese nula assume que as proporções entre as observações são iguais, enquanto a alternativa, que as proporções são diferentes. Por exemplo: as proporções de

ocorrências de *né* com função de Partícula Modal e com outras funções podem ser iguais (H_0) ou diferentes (H_A).

Primeiramente, foram aplicados os testes às proporções entre os itens lexicais em português brasileiro e os que podiam ser considerados Partículas Modais. E em uma segunda etapa, foram aplicados testes às proporções entre as Partículas e seus equivalentes tradutórios.

Os resultados da pesquisa são apresentados a seguir.

3. Resultados

3.1 Análise das ocorrências dos itens lexicais em português brasileiro

A Tabela 2 mostra: na primeira coluna, os itens lexicais encontrados no *corpus* (e.g. *né*, *ué*, *bah*, etc.); na segunda, o total de ocorrências desses itens lexicais; na terceira, o número de ocorrências do item com função de Partícula Modal; e na última coluna, a frequência relativa entre o total de ocorrências do item lexical no *corpus* e as ocorrências que realizam função de Partícula.

TABELA 5.
Total de ocorrências, total de Partículas e frequência relativa.

Item lexical	Total de ocorrências	Função de Partícula	Frequência Relativa
né	**19**	**19**	**100%**
ué	**8**	**8**	**100%**
bah	**5**	**5**	**100%**
uai	3	3	100%
ah, é	6	3	50%
ó	**10**	**4**	**40%**
hein	30	9	30%
ah	120	22	18,3%
tá	56	7	12,5%
aí	49	6	12,2%
é	488	5	1,02%
não	432	2	0,46%
aqui	68	0	0%
viu	11	0	0%
nó	3	0	0%
sô	2	0	0%
Totais	**1310**	**93**	**7,10%**

Ao todo, foram encontradas 1310 ocorrências dos itens lexicais pesquisados no *corpus*. Porém, somente 93 dessas ocorrências realizam a função de Partículas Modais, e os itens que realizam essa função são: *né*, *ué*, *bah*, *uai*, *ah, é*, *ó*, *hein*, *ah*, *tá*, *aí*, *é*, *não*, *aqui*, *viu*, *nó* e *sô*. Além disso, os tipos de Partícula Modal verificados no *corpus* são: Anuência, Confirmação, Exclamação, Conclusão, Atenuação, Atenção, Concordância e Exortação, sendo as suas respectivas ocorrências no *corpus*: 21, 18, 38, 3, 5, 5, 3 e 1.

É importante mencionar que neste *corpus* não foram encontradas Partículas Modais com função de Insistência, Entendimento, Empatia, Desafio, Exortação (para obedecer) e Atenuação (para aceitar).

O teste utilizado para a verificação da distribuição dos dados e validação das probabilidades (teste chi-quadrado) evidenciou que os itens lexicais *uai*, *ah, é*, *hein*, *ah*, *tá*, *aí*, *é*, *não*, *aqui*, *viu*, *nó* e *sô* apresentaram valor-*p* inferior a 0,05, indicando que todos encontram-se na área de rejeição de H_0. Dessa forma, é maior a probabilidade de esses itens terem uma proporção diferente da população (i.e. as ocorrências de Partículas no português brasileiro), e, por isso, provavelmente eles se comportam de maneira diferente.

Já para os itens *né, ué, bah* e *ó*, o teste de aderência chi-quadrado não rejeitou a hipótese nula de que esses dados (as ocorrências dos itens lexicais *né, ué, bah* e *ó* no *corpus*) estavam dentro da distribuição chi-quadrado, e, portanto, a população (todas as ocorrências dos itens lexicais *né, ué, bah* e *ó* no português brasileiro) teria um comportamento semelhante.

É possível observar na Tabela 2 que todas as ocorrências dos itens lexicais *né, ué* e *bah* neste *corpus* têm a função de Partícula Modal. Para estes três itens o resultado do valor-p no teste estatístico é igual a 1, o que demonstra uma probabilidade alta de a população também ter o mesmo comportamento: realizar somente a função de Partícula Modal. Os Exemplos 5, 6 e 7 ilustram os usos dessas Partículas.

Exemplo 5 – Partícula de Anuência:
Vam Legal quando o cliente fica contente, né?
Pen E ainda nos elogia!

Exemplo 6 – Partícula de Exclamação:
Mag Eu sou contra qualquer casamento secreto!
Mon Ué! Por que?

Exemplo 7 – Partícula de Exclamação:
Cas Desgraça pouca é bobagem.
Ceb Bah! Detesto ditados.

Observa-se ainda na Tabela 2 que o item lexical *ó* ocorre dez vezes neste *corpus*. No entanto, somente em quatro ocorrências ele possui a função de Partícula. Em todas as outras, o item tem a função de um Adjunto que acompanha o Vocativo, como em "Fiz exatamente como pediu, ó Zeus". O Exemplo 8 ilustra o uso do *ó* como Partícula Modal:

Exemplo 8 – Partícula de Atenção:
CBe Si bem qui o corno dele mais parece um sorvete na testa.
PZe Bobagem! Ele é um unicórnio igualzinho ao do livro, ó!

No caso desse item, o valor-*p* resultado do teste estatístico é igual a 0,1088. Considerando que o valor de α utilizado foi 0,05, o teste revela que também neste caso o item se encontra dentro da área de não-rejeição da distribuição chi-quadrado. Assim, embora a probabilidade de a população se comportar da mesma forma seja menor, ainda é válido afirmar que o item *ó* possui uma probabilidade de 40% de realizar a função de Partícula Modal no português brasileiro.

Cabe ainda uma importante observação sobre o item *ah, é*. Como é possível observar na Tabela 2, ele ocorre seis vezes no *corpus*, sendo três dessas ocorrências de Partículas; mais especificamente, os três casos são de Partículas Modais de Exclamação. O Exemplo 9 ilustra o uso de *ah, é* com função de Partícula.

Exemplo 9 – Partícula de Exclamação:
Her Não sou sua avó!
Mon Ah, é! Desculpe, vodrasta!

Nos outros três casos de *ah, é* os itens provavelmente tratam-se de Partículas Modais, porém, a análise de Figueredo (2011) não contempla essas construções, as quais aparentemente seriam uma outra opção do sistema, diferente das descritas por ele. A análise desses itens, apresentados nos Exemplos 10, 11 e 12, revela que os três são encontrados antes de propostas – nos Exemplos 10 e 11 precedem uma demanda de informação, e no Exemplo 12, uma demanda de bens-&-serviços. Observa-se também que o falante constrói sua proposta baseado na fala do ouvinte, utilizando a Partícula *ah, é* para se opor ao ouvinte. Assim, o falante aparentemente espera que o ouvinte valide seu papel como contrário. Dessa forma, é possível afirmar que, nestes casos, o item *ah, é* se trata de uma Partícula Modal com a função

de Oposição. Ainda assim, seria necessária uma pesquisa mais aprofundada sobre esse item para confirmar essa hipótese, pois somente foram encontradas três ocorrências de *ah, é* com essa provável função.

Os três exemplos apresentados a seguir demonstram todos os usos do item *ah, é* com a possível função de Partícula de Oposição. Para o Exemplo 10, o contexto é um diálogo entre os personagens *Mon* e *Dar*, no qual o primeiro é interrogado pelo segundo sobre a localização de uma base rebelde.

> Exemplo 10 – Partícula de Oposição:
> *Mon Devolve meu coelhinho, lorde Feio!*
> *Dar Primeiro conta onde fica a base rebelde!*
> *Mon Eu não sei!*
> *Dar Ah, é? Não vai falar, gorducha?*
> *Mon Como é que é?*
> *Dar Igor vai obrigá-la a falar.*

Para o Exemplo 11, o contexto é uma luta entre os personagens *Mon* e *Had*. Após derrotar os guardas de *Had*, *Mon* tenta atacá-lo:

> Exemplo 11 – Partícula de Oposição:
> *Mon E agora você, seu feioso!*
> *Had Ah, é? Pensa que pode comigo?*

E para o Exemplo 12, o contexto é um diálogo entre o personagem *Tit* e *Ceb*, sendo que este último participa de um jogo de bolas-de-gude e o primeiro aparece em cena, interrompe o jogo e faz um pronunciamento:

> Exemplo 12 – Partícula de Oposição:
> *Tit Por ordem de sua alteza, o príncipe Xaveco, ficam todos avisados que estão proibidas as brigas e duelos entre inimigos. E quem perturbar a paz será severamente castigado!*
> *Ceb Ah, é? Então sai daí, que você está atlapalhando o jogo!*

Na seção seguinte, serão examinados os equivalentes tradutórios em inglês para as Partículas encontradas no *corpus*.

3.2 Análise da tradução das partículas modais

Nesta etapa da análise, foi observado como cada Partícula foi traduzida para o inglês, e quantificada cada uma das opções tradutórias. Também foi investigada a existência de padrões na tradução de algumas Partículas, sendo que apenas as de Anuência e de Confirmação (em declarações) apresentaram padrões estatisticamente relevantes para as suas traduções.

Esta seção apresenta os resultados da análise da tradução das Partículas Modais encontradas neste *corpus*, começando pela tradução das Partículas de Anuência.

TABELA 6.
Tradução das Partículas de Anuência.

Partícula de Anuência	Item em inglês	Total de ocorrências do item em inglês traduzindo essa Partícula	Total de ocorrências do item em inglês no corpus
né	**(*tag question*)**	**7**	**9**
	Huh	4	21
	Right	4	46
	Eh	**2**	**8**
	(não realizado)	2	NA
é	*Yep*	1	6
não	*No*	1	133
Totais		**21**	**223**

Começando pela Partícula *né*, é possível observar na Tabela 3 que ela é traduzida como uma *tag question* sete vezes, como *huh* e *right* quatro vezes, como *eh* duas vezes e houve a *não-realização* em outras duas ocasiões. As Partículas *é* e *não* foram traduzidas como *yep* e *no*, respectivamente e somente uma vez cada.

A Partícula de Anuência *né* foi traduzida sete vezes como uma *tag question*. Isso pode ser considerado um padrão na tradução deste item, o que é reforçado pelo valor-*p* (0,6171) do teste chi-quadrado para este item, que é maior que o valor de α (0,05) e, por isso, encontra-se na área de não rejeição da distribuição. Ainda, o princípio da agnação permitiu notar que na tradução para o inglês, o texto alvo apresenta uma pequena mudança na prosódia. Enquanto a entonação prototípica de uma *tag question* é semelhante à de uma pergunta polar, a sentença que contém essa tradução apresenta uma entonação semelhante à de declarações. Isto é, as *tag questions* das traduções de *né* não são um estímulo à resposta, mas um convite do falante para

que o ouvinte valide sua proposição. O Exemplo 13 exemplifica o uso dessa Partícula e sua tradução para o inglês como *tag question*.

> Exemplo 13 – Tradução da Partícula de Anuência *né*:
> *Zeu Hera espero que tenha aprendido a sua lição! Você não ganha nada sendo tão violenta! E você Mônica também aprendeu uma lição, né?*
> *Mon Claro! De hoje em diante vou ser mais cuidadosa!*
>
> _____
> *Zeu I hope you learned your lesson Hera! You gain nothing by being violent! And you learned a lesson too, didn't you, Monica?*
> *Mon Yes! From now on I'll be more careful!*

Na Tabela 3 também é possível observar que a Partícula *né* foi traduzida como *huh* e *right*, cada uma em quatro ocasiões. No entanto, o teste estatístico revela que estes itens se encontram na área de rejeição, pois o valor-*p* é menor que o de α: 0,0006739 e 2,855e-09, respectivamente. Dessa forma, não é possível afirmar que a tradução de *né* como *huh* e *right* se configura como um padrão, dado que é provável que a população se comporte de maneira diferente.

A Partícula *né* ainda foi traduzida como *eh* duas vezes. As duas ocorrências desse item permitem dizer que este seja outro padrão na tradução de *né*. O teste chi-quadrado para este item revela que o valor-*p* (0,05778) é pouco maior que o de α (0,05). Embora seja um resultado limítrofe, ainda é possível afirmar que a população se comporta da mesma forma, porém espera-se que a frequência seja menor em relação à tradução de *né* como *tag question*. O Exemplo 14 apresenta a tradução de *né* como *eh*.

> Exemplo 14 – Tradução da Partícula de Anuência *né*:
> *Mon Pronto pessoal! Caso resolvido! É só alimentá-lo uma vez ao dia!*
> *Ps2 Tudo graças a você, Mônica!*
> *Ceb O Com a minha assistência, é clalo!*
> *Mon Em vez de só falar bem que você podia ajudar mesmo, né, Cebolinha?*
>
> _____
> *Mon There people! Case solved! Just feed him once a day!*
> *Ps2 It's all thanks to you, Monica!*
> *Ceb With my help, of couwse!*
> *Mon Instead of just yakking you could actually help me, eh, Jimmy?*

Houve, ainda, dois casos de *não realização*. Em um dos casos, a sentença no texto fonte se tratava de uma declaração, com a Partícula realizada no fim da oração. No texto alvo, a

estrutura da sentença muda, passando a ser uma pergunta polar. Esse caso é demonstrado pelo Exemplo 15. O contexto é o desembarque de uma caixa grande por dois entregadores, sendo que um deles, *En2*, ao desembarcar a caixa, deixa-a cair em cima do outro entregador; a caixa é uma encomenda para *Mau*.

Exemplo 15 – Tradução da Partícula de Anuência *né*:
En2 Não estragou nada, né?
Mau Não! Tá tudo aqui. Desde a primeira arte original do Horácio até a última.

_____
En2 Everything okay?
Mau Yes! It's all here. From the very first drawings of Horacio to the last.

É possível observar que no texto fonte o personagem *En2* faz uma declaração "não estragou nada" e pede que o ouvinte valide essa declaração "né?". No texto alvo, o falante faz uma pergunta para o ouvinte, e exige dele uma resposta. A agnação mostra ainda que a pergunta em inglês possui uma entonação semelhante às traduções de *né* para *tag questions*. Dessa forma, a tradução para o inglês mantém o convite à validação do falante, porém, na forma prosódica somente. Assim, quando o ouvinte responde "Yes" e completa "it's all here", ele está validando o papel de En2 e sua proposição.

O Exemplo 16 mostra o outro caso em que houve uma *não-realização* na tradução de *né*. O contexto aqui é a personagem *Mon* procurando pelo personagem *Sebolak*. Ao encontrar o personagem *Cebolinha* preso em uma masmorra, *Mon* fica surpresa.

Exemplo 16 – Tradução da Partícula de Anuência *né*:
Mon Oh! Cebolinha! Então foi aqui que você se escondeu, né, moleque?

_____
Mon Oh! Jimmy Five! So this is where you came to hide, you scamp!

Neste caso, a agnação permitiu verificar que o significado de *né* é realizado na forma de prosódia no texto alvo, com um prolongamento das vogais, de forma que o falante negocia não somente a mensagem, mas também seu posicionamento como falante. Assim, o ouvinte deve validar esse posicionamento.

Como o número de ocorrências da tradução como *não realizado* é pequeno, não é possível afirmar que este seja o padrão neste *corpus*. Além disso, como a *não-realização* na tradução para o inglês não envolve um item lexical, não é possível fazer a contagem das ocorrências nos textos em inglês. Dessa forma, cada ocorrência de *não-realização* deve ser

considerada como única, não sendo possível, no escopo desta pesquisa, fazer uma análise mais aprofundada desse item ou uma modelagem estatística.

Por fim, tem-se as traduções das Partículas *é* e *não*. Estas Partículas estão presentes no *corpus* em movimentos respondentes, mais especificamente, como um movimento respondente esperado e um arbitrário, para *é* e *não*, respectivamente. Por isso, a tradução também segue esses movimentos, sendo traduzida como *yep* para o movimento esperado, e como *no* para o arbitrário. Como essas Partículas e suas traduções somente ocorrem uma vez no *corpus*, não é possível fazer afirmações sobre o padrão nas traduções desses itens.

O próximo tipo de Partícula Modal abordado é o de Confirmação. De acordo com Figueredo (2011), este tipo de Partícula pode ser encontrado em declarações ou em perguntas polares. Neste artigo são apresentados somente os dados da análise de declarações.

A tradução das Partículas Modais de Confirmação em sentenças declarativas pode ser vista na Tabela 4. Nota-se que houve quatro *não-realizações* na tradução da Partícula *hein*, duas traduções para *eh*, e duas para *huh*. A Partícula *tá (longo)* foi traduzida duas vezes como *okay* e duas como *yeh*. Além disso, *é* foi traduzido como *huh* e *yes* uma vez cada, e *não* também foi traduzido somente uma vez como *no*.

TABELA 7.
Tradução das Partículas de Confirmação (em declarações).

Partícula de Confirmação (em declarações)	Item em inglês	Total de ocorrências do item traduzindo essa Partícula	Total de ocorrências do item em inglês
Hein	(não realizado)	4	NA
	eh	**2**	**8**
	huh	2	21
tá (longo)	*okay*	2	45
	yeh	2	3
É	*huh*	1	21
	yes	1	44
Não	*no*	1	133
Totais		**15**	**275**

Na Tabela acima, observa-se a ocorrência de quatro *não-realizações* na tradução de *hein*. Novamente, como no caso das traduções da Partícula de Anuência *né*, os casos de *não-realização* do item em inglês devem ser tratados como ocorrências únicas, não sendo possível uma análise mais aprofundada ou um tratamento estatístico.

A Partícula de Confirmação *hein* foi também traduzida como *eh* em duas ocasiões. O teste chi-quadrado revela que o valor-*p* para esses dados (0,05778) é maior que o valor de α (0,05). Ainda que seja um valor limítrofe, é possível afirmar que a população se comporta de forma semelhante. Sendo assim, pode-se considerar que o padrão da tradução de *hein* neste *corpus* é *eh*. O Exemplo 17 ilustra esse caso.

Exemplo 17 – Tradução da Partícula de Confirmação *hein*:
Ceb Mas você é bem pesadinha, hein?

Ceb But you're pretty heavy, eh?

As demais Partículas de Confirmação não apresentam um padrão em sua tradução, sendo isso revelado pelo teste chi-quadrado que revela valores-*p* menores que o de α (0,05) para esses itens ou uma ocorrência muito baixa para a realização do teste (como é o caso da Partícula *tá (longo)* e sua tradução como *yeh*).

4. Discussão dos resultados e conclusão

Foram encontradas ao todo 1310 ocorrências dos itens lexicais pesquisados no *corpus*, mas somente 93 dessas ocorrências realizam a função de Partículas Modais. Os itens que realizam função de Partícula Modal são: *né*, *ué*, *bah*, *uai*, *ah, é*, *ó*, *hein*, *ah*, *tá*, *aí*, *é*, *não*, *aqui*, *viu*, *nó* e *sô*. Os tipos de Partícula Modal verificados no corpus são: Anuência, Confirmação, Exclamação, Conclusão, Atenuação, Atenção, Concordância e Exortação.

A análise dos textos na língua fonte – português brasileiro – revelam um padrão na realização das Partículas de Anuência *né*, de Exclamação *ué* e *bah*, e de Atenção *ó*. A análise estatística desses itens revelou que esses itens se encontram na área de não-rejeição da hipótese nula da distribuição chi-quadrado e, por isso, é provável que, na população (no português brasileiro), as Partículas do mesmo tipo sejam realizadas pelos mesmos itens lexicais.

Além disso, a caracterização das Partículas de Oposição configura-se como uma importante contribuição desta pesquisa. No entanto, como se trata de somente três ocorrências, um estudo com um *corpus* maior poderia revelar se a população também se comporta da mesma maneira, havendo assim um novo tipo de Partícula Modal em português brasileiro.

Em relação aos outros tipos de Partículas Modais encontrados no *corpus* (Confirmação, Conclusão, Atenuação, Concordância e Exortação) o tratamento estatístico dos dados revelou que, para os itens que as realizam, a hipótese nula deve ser rejeitada. Dessa forma, deve-se assumir a hipótese alternativa de que a população se comporta de maneira diferente. Também é possível afirmar que a quantidade de ocorrências para essas Partículas Modais foi pequena, e que os dados observados são justamente os que se encontram na área de rejeição. Sendo assim, uma pesquisa que envolva um *corpus* maior poderia revelar com maior precisão a relação entre os itens lexicais encontrados nesta pesquisa e os tipos de Partícula Modal que realizam: se a população de fato se comporta de maneira diferente ou não.

Ainda, a análise dos textos na língua alvo revelou alguns padrões na tradução para o inglês de Partículas Modais de Anuência e Confirmação (em declarações). O padrão na tradução das Partículas de Anuência revelou-se ser para o item *né*: 1) as *tag questions*, acompanhadas de queda no movimento tônico, e 2) o item lexical *eh*. A Partícula de Confirmação (em declarações) apresentou um padrão na tradução de *hein* como sendo o item lexical *eh*.

Esses padrões revelam que os significados realizados por algumas Partículas Modais em português brasileiro são realizados por elementos específicos em inglês. Isso revela especificidades do sistema de VALIDAÇÃO em inglês, o qual, diferentemente do português brasileiro, seria realizado não somente por itens lexicais (e.g. *eh*), mas também por recursos gramaticais (e.g. *tag questions*) e prosódicos (e.g. nivelamento do movimento tônico).

Como aplicação dos resultados desta pesquisa, profissionais na área de tradução podem se basear neles para aprimorarem suas traduções. Os profissionais que podem melhor desfrutar desta pesquisa são os que trabalham com traduções na modalidade oral e com legendagem de filmes, tendo em vista que as Partículas Modais desempenham um papel importante na construção da mensagem e que são encontradas com maior frequência em textos orais. Portanto, os padrões das traduções podem ser também seguidos por esses profissionais a fim de que os significados das Partículas Modais sejam realizados também nos textos em inglês.

Referências Bibliográficas

BARLOW, M. **ParaConc**. Houston: Athelstan, 2012.

CAFFAREL, A.; MARTIN, J. R.; MATTHIESSEN, C. M. I. M. (Eds.). **Language typology:** a functional perspective. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2004. crossref http://dx.doi.org/10.1075/cilt.253

CASTILHO, A. T. **Gramática do português brasileiro**. São Paulo: Contexto, 2010.

FIGUEREDO, G. **Introdução ao perfil metafuncional do português brasileiro**: contribuições para os estudos multilíngues. Belo Horizonte: Faculdade de Letras da UFMG / PosLin, 2011. (Tese, Doutorado em Linguística Aplicada).

GRIES, S. T. **Statistics for Linguistics with R**. 2. ed. Berlin/Boston: Walter de Gruyter GmbH, 2013. crossref http://dx.doi.org/10.1515/9783110307474

HALLIDAY, M. A. K.; MATTHIESSEN, C. M. I. M. **An introduction to functional grammar**. 3rd. ed. London: Edward Arnold, 2004.

MARTIN, J. R.; WHITE, P. R. R. **The Language of Evaluation:** Appraisal in English. London: Palgrave, 2005.

MATTHIESSEN, C. M. I. M.; HALLIDAY, M. A. K. **Systemic Functional Grammar:** A First Step Into The Theory. Beijing: Higher Education Press, 2009.

MAURICIO DE SOUSA PRODUÇÕES LTDA. Quadrinhos - História Seriada. **Portal da Turma da Mônica**, 1996. Disponivel em: <http://www.monica.com.br/comics/seriadas.htm>. Acesso em: Abril a agosto 2012.

NEVES, M. H. M. de. (Org.). **Gramática do português falado**. v.7. Novos estudos. São Paulo: Humanitas/FFLCH/USP; Campinas: Editora da Unicamp, 1999.

O'DONNELL, M. **UAM Corpus Tool**. Madri: Universidad Autónoma de Madrid, 2014. Disponível em: http://www.wagsoft.com/CorpusTool/.

R CORE TEAM. **R:** A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2014. Disponivel em: <http://www.R-project.org/>.

Artigo recebido em: 15.10.2014
Artigo aprovado em: 04.12.2014

Linguística de *Corpus* e ensino: a compilação de um *corpus* de especialidade para preparação e implementação de um curso preparatório rápido para exame de proficiência

Corpus Linguistics and teaching: the compilation of a specialized corpus for the preparation and implementation of a crash prep course for a proficiency examination

Stella E. O. Tagnin[*]
Danilo S. Murakami[**]

RESUMO: Este artigo apresentará o processo de compilação de um *corpus* de especialidade na área de Relações Exteriores e seu uso para definir o conteúdo programático e a preparação de material didático para candidatos a um exame de proficiência em inglês para preenchimento de um cargo público no âmbito do governo federal.

ABSTRACT: This article will present the compilation of a specialized corpus in Foreign Affairs and how it was used to define the syllabus and support the preparation of teaching materials for candidates that would take a proficiency exam as one of the requirements to fill a federal government position.

PALAVRAS-CHAVE: Linguística de *Corpus*. Ensino de inglês. Conteúdo programático. Material didático.

KEYWORDS: Corpus Linguistics. English teaching. Syllabus. Teaching materials.

1. Introdução

Neste artigo apresentaremos o processo de compilação de um *corpus* de especialidade para informar um curso preparatório rápido para candidatos a um cargo público federal na área de Relações Exteriores. A investigação do *corpus* norteou o conteúdo programático assim como a preparação do material para o curso.

O curso destinava-se a candidatos que deveriam se submeter a um exame de proficiência em língua inglesa como um dos requisitos para o preenchimento das vagas. A duração do curso era de 12 horas, divididas em quatro dias, ou seja, aulas de três horas por dia. Os candidatos já haviam passado por uma prova que envolvia uma redação em inglês. Essas redações foram corrigidas manualmente e avaliadas sob três quesitos: conteúdo, estrutura e expressão. Na parte do 'conteúdo', que valia 20 pontos, era julgada a "perspectiva adotada no tratamento do tema; [a] capacidade de análise e senso crítico em relação ao tema proposto [e a] consistência dos argumentos, clareza e coerência no seu encadeamento". No

[*] Livre docente, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.
[**] Mestrando, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.

quesito 'estrutura', que valia 30 pontos, foram avaliados "o respeito ao gênero solicitado; [a] progressão textual e encadeamento de ideias; [e a] articulação de frases e parágrafos (coesão textual)". Finalmente, no quesito 'expressão', responsável pelos 50 pontos restantes da prova, foram levados em conta o domínio, a proficiência e o emprego correto das estruturas próprias da língua inglesa escrita. Como a parte relativa à 'expressão' era responsável por metade do valor da prova, a entidade organizadora da prova de proficiência optou por privilegiar esse aspecto num curso rápido preparatório a ser oferecido aos candidatos qualificados, isto é, aos que haviam atingido ou ultrapassado a nota mínima exigida na redação em língua inglesa. Colocava-se o problema de 'o quê' abordar nesse curso, uma vez que o cargo a que os candidatos almejavam previa a redação de textos em língua inglesa. O primeiro passo foi analisar os erros mais recorrentes detectados durante a correção da redação. A fim de abordar esses problemas num contexto real de uso, compilou-se um *corpus* com textos relacionados à área de Relações Exteriores. Na seção seguinte (seção 2) discutiremos os objetivos que pretendíamos alcançar com o curso. Em seguida (seção 3), apresentaremos os princípios da abordagem denominada *Data-Driven Learning* (DDL). O *corpus* compilado e as ferramentas utilizadas para definir o conteúdo a ser abordado no curso e para preparar os respectivos exercícios são discutidos na seção 4. Na seção 5 analisaremos os resultados alcançados para concluir, na seção 6, com algumas considerações sobre as possibilidades de apropriação da metodologia por parte dos candidatos.

2. Um curso de revisão para exame de proficiência em língua inglesa – o que abordar?

Dado o curto tempo disponível para a pretendida revisão gramatical, optamos por, inicialmente, conscientizar os candidatos sobre os aspectos convencionais da língua (TAGNIN, 2013), ou seja, embora gramaticalmente seja por vezes possível expressar-se de certa forma, o convencional, ou seja, o usual, não é empregar essa forma. Por exemplo, apesar de 'gatos e cães' ser uma forma gramaticalmente correta, não é a forma consagrada de nos referirmos a 'cães e gatos'. Da mesma forma, 'severamente ferido', apesar de gramatical, não é a forma usual de denominarmos alguém que foi 'gravemente ferido'. Em outras palavras, a língua é um sistema probabilístico (HALLIDAY, 1961), em que, embora certas formas sejam *possíveis*, elas não são *prováveis* de ocorrer numa produção fluente da língua. Nesse sentido, dominar as estruturas convencionais contribui de forma significativa para se alcançar a fluência nessa língua. A Linguística de *Corpus* tem provado ser extremamente eficaz em

identificar essas formas convencionais, uma vez que são padrões que se repetem. Familiarizar os candidatos com os princípios dessa abordagem foi outro objetivo, pois que lhes daria autonomia para fazerem suas próprias investigações. Observe-se, no entanto, que eram candidatos com várias formações (Administração de Empresas, Direito, Economia etc.), mas praticamente sem qualquer formação linguística. Por essa razão, pretendíamos apresentar-lhes *corpora on-line* a que pudessem recorrer sempre que tivessem alguma dúvida de redação. Para tanto, foram preparados vários exercícios baseados em *corpora*, numa abordagem denominada *data-driven learning* (DDL) (JOHNS, 1991), que preconiza o aprendizado pela observação de várias instâncias de um fenômeno linguístico.

3. Como "ensinar" o conteúdo programático? A abordagem *data-driven learning* (DDL)

Diferentemente de métodos de ensino mais consagrados em que o professor transmite ao aluno o que sabe acerca da língua, na perspectiva DDL, o aprendiz participa ativamente do processo de construção de conhecimento. Conforme aponta Johns (2002), o instrutor fornece ao aprendiz ocorrências autênticas de uso da língua na forma de linhas de concordância. O aprendiz, por sua vez, examina os dados reais da língua estudada a fim de encontrar padrões linguísticos recorrentes. Tal processo leva o aprendiz a levantar hipóteses sobre a organização linguística das informações que observa nesses dados. A partir de suas observações, o aprendiz é capaz de testar as hipóteses que levantou com o objetivo de fazer generalizações a respeito do funcionamento da língua estudada. Nessa abordagem, portanto, o aprendiz atua como um pesquisador linguístico (JOHNS, 2002, p. 108), uma vez que tem de aprender por meio da análise de dados de que dispõe.

Tendo em vista que cabe ao aluno observar-levantar hipóteses – fazer generalizações para aprender uma língua, pode-se ter a errônea impressão de que o instrutor que adota essa abordagem, não teria o trabalho de "ensinar". No entanto, Johns (1988, p. 10) chama a atenção para o fato de que o papel do professor está no trabalho prévio de preparação e apresentação das informações que serão examinadas pelo aprendiz. Em vez de preparar um texto, que poderia ser concebido artificialmente, ou seja, escrito pelo próprio professor para a lição, o instrutor deve preparar o material de modo que contemple especificamente o tópico a ser abordado e seja adequado ao nível de conhecimento de seus aprendizes.

Entretanto, a concepção dos estudantes de que, para compreender qualquer elemento da língua, eles devem entender todos os elementos que o cercam, pode, no início, configurar

uma desvantagem da DDL (JOHNS, 1988, p. 10). Quando consideramos a exposição do aprendiz a linhas de concordância, pode-se levantar pelo menos dois pontos que podem causar-lhe certa estranheza. Primeiramente, uma concordância tem o formato peculiar em que o tópico em estudo é apresentado centralizado de maneira que há a mesma quantidade de caracteres em ambos os lados, esquerdo e direito. Tal formatação limita o número de caracteres dispostos a partir do centro, independentemente do limiar entre sentenças ou palavras. Dessa forma, uma linha de concordância pode ou não coincidir com o início de uma sentença ou com uma sentença completa, da mesma forma que a palavra no início e no fim da linha pode ser lida em sua plenitude (*palavra*) ou vermos somente algumas letras que a compõem (*[pala]vra)*. Em segundo lugar, dado que as linhas de concordância provêm de usos reais da língua, o aprendiz pode encontrar palavras que desconhece, as quais podem desviar sua atenção do foco de aprendizado (Quadro 1)

Quadro 1: linhas de concordância para *Brazil*.

N	Concordance
1	nded by the Trade Ministers of Brazil, Egypt, Argentina, Ind
2	oviding inputs to negotiators. Brazil 's WTO Ambassador, Clo
3	o shift the blame to India and Brazil for not offering enoug
4	tral to Doha Round," India and Brazil state positions Inglês
5	tral to Doha Round," India and Brazil state positions Hong K

Entre as vantagens de adotar a DDL no ensino, podemos citar a independência da metalinguagem gramatical. Muitas vezes o estudante não domina, especialmente em língua estrangeira, os termos usados pelo professor em sua explicação, porém pode alcançar sucesso nas tarefas práticas de análise da língua que não necessariamente dependem da nomeação de fenômenos da língua. Além disso, o aprendiz pode descobrir algo que o professor não pensou em incluir em sua exposição tradicional, ou algum fato que até mesmo dicionários ou gramáticas não demonstram com clareza.

4. Por trás dos bastidores

Como dissemos, o primeiro passo foi compilar um *corpus* com textos na língua inglesa relacionados às Relações Exteriores. Esses foram coletados da internet a partir de jornais, revistas ou *sites* especializados em relações exteriores, segundo os critérios descritos em Berber Sardinha (2004).

Para a coleta dos textos que compuseram o *corpus*, utilizamos a ferramenta BootCat (BARONI e BERNARDINI, 2004). Apesar de ser comum a prática de criar um *corpus* a partir da coleta manual de textos disponíveis na Internet, ou seja, acessando a página em que cada texto é encontrado para obter seu conteúdo, a ferramenta adotada realiza o processo de forma automática. A BootCat funciona a partir de palavras iniciais que atuam como "sementes" (do inglês, *seeds*). A ideia é que tais palavras, selecionadas no início do processo, identifiquem o domínio que se deseja investigar, no nosso caso, Relações Exteriores. Uma vez que temos esse número inicial, inserimos essas sementes na ferramenta para buscar páginas da web que as contenham. Em nosso trabalho, usamos como sementes unidades lexicais como *Brazil, USA, foreign policy, global warming, climate change, Doha Round*, entre outras (Figura 2).

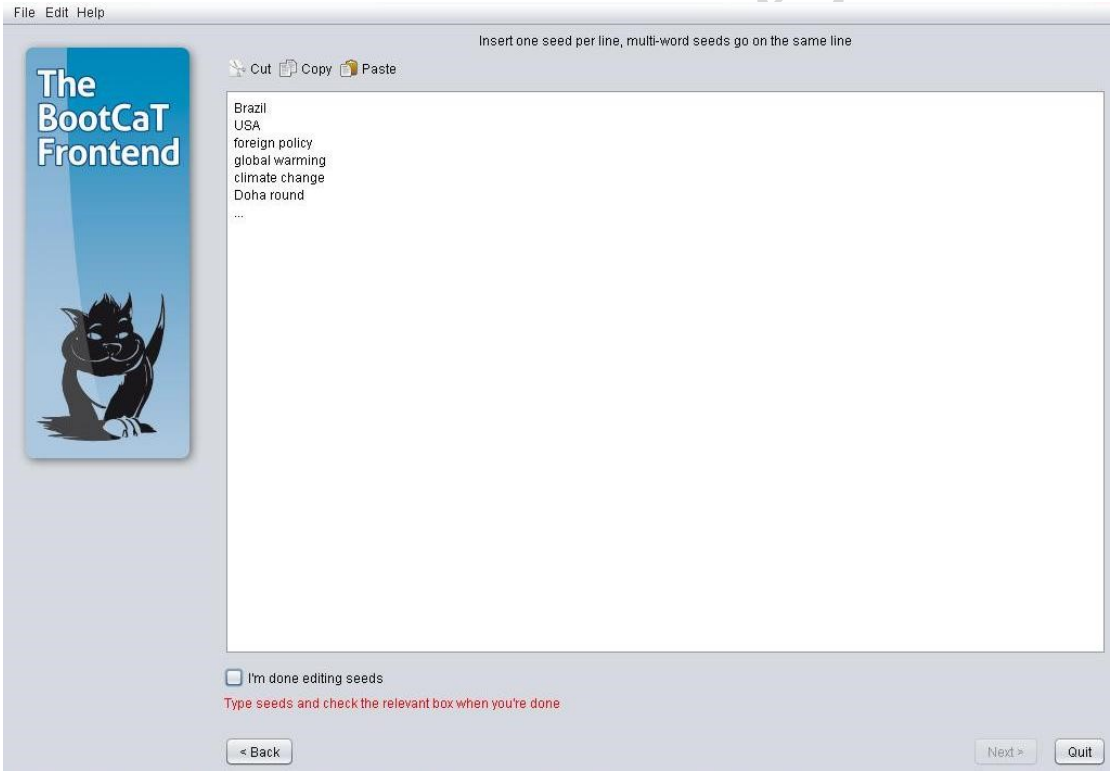


Figura 1: sementes (*seeds*) inseridas na BootCat.

A ferramenta permite que tais palavras sejam organizadas em conjuntos que combinam aleatoriamente uma quantidade específica de "sementes", conjuntos que o programa chama de *tuples*. Nessa etapa, a BootCat procura páginas que contenham necessariamente todas as palavras que fazem parte de um conjunto. Em outras palavras, na criação de nosso *corpus*, selecionamos dez conjuntos que continham combinações de três

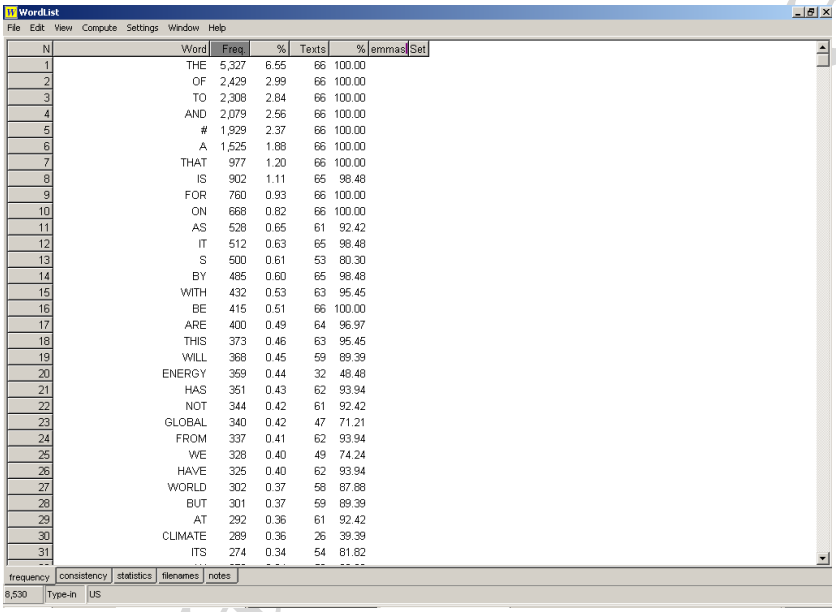
palavras (e.g. *Brazil, USA, foreign policy; global warming*, *USA, Doha round*; etc). É também possível excluir os domínios de rede que não queremos pesquisar, para evitar que certos conteúdos provenientes de páginas que julguemos inapropriadas apareçam entre os resultados da busca. Uma vez que pretendíamos averiguar com mais cuidado o conteúdo das páginas encontradas, não recorremos ao recurso de limitar domínios. O resultado que a ferramenta nos fornece é uma lista de endereços URL que contém os conjuntos de *tuples* determinados anteriormente. É importante ressaltar que um dos procedimentos que tornam a ferramenta eficiente é a possibilidade de gerar novas sementes a partir das páginas encontradas nos resultados. Consequentemente, pode-se reiniciar o processo a fim de coletar mais dados. No entanto, uma vez que havíamos adotado um número que julgávamos suficiente de sementes iniciais, consideramos que não foi necessário repetir o procedimento.

Como mencionamos anteriormente, uma forma de coletar textos que comporiam o *corpus* seria vasculhar páginas da Internet e baixar o seu conteúdo. Optamos por adotar a ferramenta BootCat devido a algumas de suas vantagens. Primeiramente, não conhecíamos em profundidade a área de Relações Exteriores. Não poderíamos, portanto, acessar diretamente páginas que contemplassem tal domínio de conhecimento, uma vez que não estávamos familiarizados com seus textos. Nesse aspecto, valemo-nos de palavras ou termos encontrados nas redações dos alunos para usar como sementes na busca de textos. Outra vantagem está relacionada ao propósito da coleta do *corpus*. Como se tratava de um *corpus* "descartável" (VARANTOLA, 2002), ou seja, construído para uma tarefa específica, não foi necessário seguir os critérios rígidos que se estabelecem ao construir um *corpus* de especialidade para fins lexicográficos ou terminográficos. Em outras palavras, foi uma tarefa bem menos trabalhosa. Por fim, em consequência do limite de tempo, a ferramenta foi de grande valia, pois o prazo para apresentação do material para o curso não permitia uma coleta demorada. Com a automatização que a ferramenta proporciona, obtivemos de maneira rápida e objetiva o *corpus* que serviu como recurso para a etapa seguinte, a criação do material didático.

À época de nosso trabalho utilizamos a interface da BootCat disponível no *site* da ferramenta. Hoje ela está apenas disponível para *download*[1]. Desenvolvimentos mais recentes estão relatados em Bernardini & Ferraresi (2013).

[1] BootCat: http://bootcat.sslmit.unibo.it

O produto resultante da utilização da BootCat, o *corpus* de Relações Exteriores é composto por sessenta e nove artigos integrais escritos em inglês, provenientes de jornais, revistas e *sites* especializados, num total de 58.104 palavras. Os textos compreendem o período entre 1999 e 2009. O passo seguinte foi determinar o vocabulário mais recorrente nesse *corpus*. Para isso foi feita uma Lista de Palavras (*WordList*) (Figura 2) com o software WordSmith Tools, versão 5 (SCOTT, 2007) que foi, em seguida, comparada a uma lista de palavras de um *corpus* jornalístico de 2 milhões de palavras a fim de evidenciar as palavras mais típicas (palavras-chave ou *KeyWords*) do *corpus* de Relações Exteriores (Figura 3).


WordList

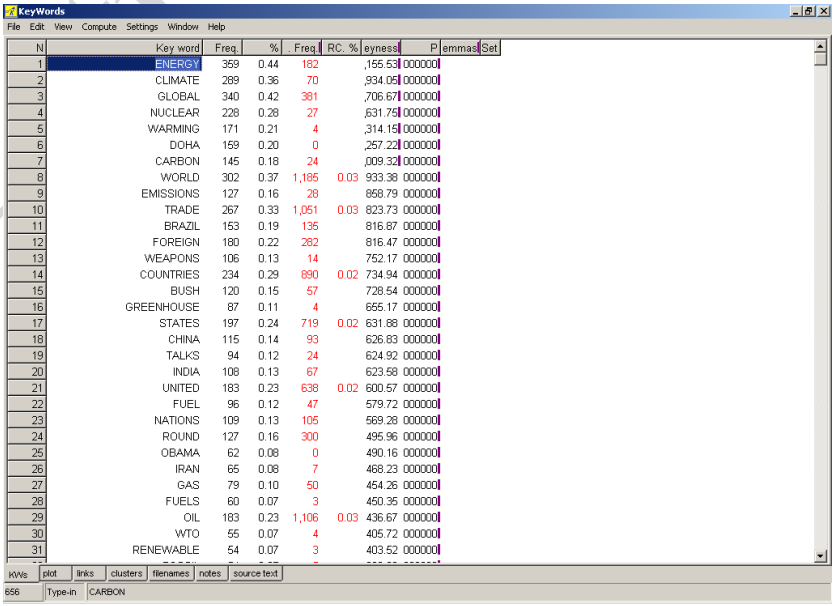
File Edit View Compute Settings Window Help

N	Word	Freq.	%	Texts	%	emmas	Set
1	THE	5,327	6.55	66	100.00		
2	OF	2,429	2.99	66	100.00		
3	TO	2,308	2.84	66	100.00		
4	AND	2,079	2.56	66	100.00		
5	#	1,929	2.37	66	100.00		
6	A	1,525	1.88	66	100.00		
7	THAT	977	1.20	66	100.00		
8	IS	902	1.11	65	98.48		
9	FOR	760	0.93	66	100.00		
10	ON	668	0.82	66	100.00		
11	AS	528	0.65	61	92.42		
12	IT	512	0.63	65	98.48		
13	S	500	0.61	53	80.30		
14	BY	485	0.60	65	98.48		
15	WITH	432	0.53	63	95.45		
16	BE	415	0.51	66	100.00		
17	ARE	400	0.49	64	96.97		
18	THIS	373	0.46	63	95.45		
19	WILL	368	0.45	59	89.39		
20	ENERGY	359	0.44	32	48.48		
21	HAS	351	0.43	62	93.94		
22	NOT	344	0.42	61	92.42		
23	GLOBAL	340	0.42	47	71.21		
24	FROM	337	0.41	62	93.94		
25	WE	328	0.40	49	74.24		
26	HAVE	325	0.40	62	93.94		
27	WORLD	302	0.37	58	87.88		
28	BUT	301	0.37	59	89.39		
29	AT	292	0.36	61	92.42		
30	CLIMATE	289	0.36	26	39.39		
31	ITS	274	0.34	54	81.82		

frequency consistency statistics filenames notes

8,530 Type-in US

Figura 2: Lista de Palavras do *corpus* de Relações Exteriores.


KeyWords

File Edit View Compute Settings Window Help

N	Key word	Freq.	%	. Freq.	RC. %	eyness	P	emmas	Set
1	ENERGY	359	0.44	182		,155.53	000000		
2	CLIMATE	289	0.36	70		,934.05	000000		
3	GLOBAL	340	0.42	381		,706.67	000000		
4	NUCLEAR	228	0.28	27		,631.75	000000		
5	WARMING	171	0.21	4		,314.15	000000		
6	DOHA	159	0.20	0		,257.22	000000		
7	CARBON	145	0.18	24		,009.32	000000		
8	WORLD	302	0.37	1,185	0.03	933.38	000000		
9	EMISSIONS	127	0.16	28		858.79	000000		
10	TRADE	267	0.33	1,051	0.03	823.73	000000		
11	BRAZIL	153	0.19	135		816.87	000000		
12	FOREIGN	180	0.22	282		816.47	000000		
13	WEAPONS	106	0.13	14		752.17	000000		
14	COUNTRIES	234	0.29	890	0.02	734.94	000000		
15	BUSH	120	0.15	57		728.54	000000		
16	GREENHOUSE	87	0.11	4		655.17	000000		
17	STATES	197	0.24	719	0.02	631.88	000000		
18	CHINA	115	0.14	93		626.83	000000		
19	TALKS	94	0.12	24		624.92	000000		
20	INDIA	108	0.13	67		623.58	000000		
21	UNITED	183	0.23	638	0.02	600.57	000000		
22	FUEL	96	0.12	47		579.72	000000		
23	NATIONS	109	0.13	105		569.28	000000		
24	ROUND	127	0.16	300		495.96	000000		
25	OBAMA	62	0.08	0		490.16	000000		
26	IRAN	65	0.08	7		468.23	000000		
27	GAS	79	0.10	50		454.26	000000		
28	FUELS	60	0.07	3		450.35	000000		
29	OIL	183	0.23	1,106	0.03	436.67	000000		
30	WTO	55	0.07	4		405.72	000000		
31	RENEWABLE	54	0.07	3		403.52	000000		

kWs plot links clusters filenames notes source text

656 Type-in CARBON

Figura 3: Palavras-chave do *corpus* de Relações Exteriores.

Em seguida foram geradas linhas de concordância com as palavras ou padrões que se pretendia discutir no curso (quadro 2).

Quadro 2: Linhas de concordância para *global*.

There are many, many other costs, ranging from **global** warming to the toxic pollution of the
human beings as occupying a place within the **global** community which allowed them the
Design Association links designers in a **global** network to provide information,
the farming of sea vegetation as a potential **global** disaster. [p] His triumphant
the rain forests, protect the whales, reduce **global** warming, take a metaphorical axe to
findings. 76 think that people are causing **global** warming, but only 41 think there is
IS DESIGNED TO HELP THE MALDIVES LIVE WITH **GLOBAL** WARMING, NOT TO HELP PREVENT IT. WHEN
GLOBAL WARMING SPREAD TROPICAL DISEASES? If **global** warming does take off and climate
Gulf having concluded that the potential for **global** effects is very small. For global
see Table 4.2). The first version is **global** in scope and historically dilated; the
of Bangladesh now, they will be tenfold under **global** warming. One scenario suggests that 1

No quadro acima já é possível identificar a colocação *global warming*, que se repete seis vezes em 11 linhas.

Os exercícios foram preparados a partir das dificuldades detectadas durante a correção das redações, assim como do vocabulário mais frequente revelado pelas palavras-chave. Preparado o material, ele foi compilado em formato de apostila e distribuído aos candidatos.

5. Na sala de aula

O curso foi dividido em quatro sessões de 3 horas cada. As sessões foram realizadas em quatro dias subsequentes.

5.1 Primeiro dia: noções de Convencionalidade e Linguística de *Corpus*

Como dissemos, os candidatos foram primeiramente apresentados aos princípios da Linguística de *Corpus*, que parte da observação de grandes quantidades de textos para descrever padrões que se repetem. Ao contrário de abordagens anteriores que privilegiavam a intuição do falante ou pesquisador, ou seja, bastava a forma ser gramaticalmente correta para ser considerada aceitável, isto é, possível. Já a Linguística de *Corpus*, observando a linguagem autêntica em uso, pretende descrever aquilo que é mais provável de ocorrer, justamente por se configurar como um padrão recorrente. Esses padrões podem ser classificados em diversas categorias convencionais que variam de colocações como *foreign policy* e *climate changes*, ou binômios como *Republicans and Democrats,* até unidades fraseológicas mais extensas como *on the other hand* (TAGNIN, 2013; BARNBROOK, MASON e KRISHNAMURTHY, 2013).

Em seguida, os candidatos foram apresentados a linhas de concordância e instruídos na forma de como ler esse material a fim de identificar padrões recorrentes e sentidos diversos de uma mesma palavra (TRIBBLE). Num primeiro momento, os alunos estranham esse formato por não apresentar períodos completos. Mas o estranhamento desaparece quando lhes é explicado que uma leitura vertical identifica padrões recorrentes, como vimos no quadro 2, e uma leitura horizontal das linhas salienta sentidos diversos, como se observa no quadro 3, que apresenta três sentidos diferentes de *issue:* exemplar de uma revista (linhas 2, 4 e 5), 'assunto', 'questão' (linhas 1, 3, 6 e 8) e, finalmente, como o verbo 'anunciar' ou 'emitir' (linha 7). Uma leitura mais cuidadosa ainda irá revelar, na linha 9, a colocação verbal *take issue with*, que significa 'discordar'.

Quadro 3: Linhas de concordância para *issue.*

1.	eld convictions and stirs ancient taboos. The **issue** needs to be tackled with
2.	only commitment is to choose one item from each **issue** of the magazine minimum
3.	a parting of the ways over a money or moral **issue**. Let common sense prevail-
4.	Wedgwood Designs For Brides', before the next **issue** of Brides is published,
5.	fixtures. The first league table will appear in **ISSUE** 176 - ie: in two weeks'
6.	the disaster aid programme is a humanitarian **issue** which transcends politics.
7.	be dragged into it. I think if the Americans **issue** a unilateral ultimatum to
8.	the biggest obstacle to negotiations on that **issue**. It's a view echoed by Mr.
9.	incompetent. If a hospital or the state took **issue** with Ken's directives about

Como não se esperava que os candidatos construíssem, de imediato, um *corpus* especializado que atendesse suas necessidades, apresentamos alguns *corpora* que podem ser consultados *on-line*, dentre eles três *corpora* que fazem parte da plataforma de Mark Davies, da Brigham Young University, quais sejam: o *Corpus of Contemporary English*[2] (COCA), o *British National Corpus*[3] (BYU-BNC) e o *Corpus do Português*[4]. Outros *corpora* foram o Lácio-Ref[5], um *corpus* de português com 9 milhões de palavras, o COMPARA[6], um *corpus* bilíngue inglês-português com textos literários e o projeto COMET[7], que é constituído de um *corpus* técnico, o CorTec[8], com textos em inglês e português em mais de 20 áreas técnicas, e um *corpus* de traduções, o CorTrad[9], com originais em inglês e português e respectivas traduções. Como tínhamos acesso à internet, demonstramos como fazer buscas em cada um

[2] www.american*corpus*.org
[3] *corpus*.byu.edu/bnc/
[4] www.*corpus*doportugues.org
[5] www.nilc.icmc.usp.br/lacioweb
[6] www.linguateca.pt/COMPARA
[7] www.fflch.usp.br/dlm/comet
[8] www.fflch.usp.br/dlm/comet/consulta_cortec.html
[9] www.fflch.usp.br/dlm/comet/consulta_cortrad.html

desses *corpora*. A maioria dos candidatos mostrou-se bastante entusiasmada com esses recursos.

5.2 Segundo dia: trabalho com concordâncias

A partir dos problemas identificados durante a correção, foram elaborados os exercícios que compuseram a apostila recebida pelos candidatos. Um dos aspectos mais recorrentes foi o uso indevido de *what* em lugar de *which,* como no trecho abaixo:

> Another consequence to Brazilian diplomacy – and to the government as a whole – is the increase in international pressure to enhance its forests protection, what would give rise to the enlargement in Itamaraty's scope of action.

A fim de conscientizar os candidatos para a diferença entre *what, which* e *that* geramos 60 linhas de concordância para cada um desses elementos, das quais transcrevemos dez de cada abaixo (quadro 4). As instruções pedem que os candidatos comparem os três conjuntos para descobrir a forma como cada um dos elementos é usado:

Quadro 4: Linhas de concordância para *what, which* e *that.*

I. Compare the three sets of concordance lines below and see if you can tell the different ways in which the search words are used.

1	sp of procedure, someone who knows	**what** an official class is. If I don
2	Do you know	**what** anti-lock brakes are for? They
3	thing. We stop to look at	**what** appears to be a roadside memor
4	e Quality Street But	**what** are the local schools like?
5	g dismissed. "**What**	can I say?" he said. "I was dr
6	will do to computers	**what** computers did to slide rules?
7	both Max and I by the arm . . . "	**What**? Did the copy ed
8	to kids like Lionel, I don't know	**what** does," Lewis said. "It
9	e way its comedy steadily darkens.	**What** else could one exp
10	o libraries and bookstores to ask "**what**	else do you have?"
1	ary Road'' -- a feature with	**which** I myself have no complaint. '
2	VW,	**which** struggled in the early 1990s,
3	Image] The Senate vote,	**which** may come Friday, will
4	words similar to the old version,	**which** refers to "an
5	planted at the right time,	**which** is October and November.
6	lebrity, particularly that	**which** inhabits the House of Windsor
7	an Football Champion-ship, specify	**which** prize you're af
8	he most valued in society,	**which** might be fine if the exercise
9	alker--the aptly named Sea Shanty,	**which** is smaller than the P
10	known as Liverpool rummy,	**which** is a favorite of Bill Clinton
1	ependent BBC has accepted,	**that** there should be no tactless me
2	or in chief of Cruise Week, agrees	**that** cruise ships
3	imity and calmly authoritative air	**that** the reader can almost
4	n't built in a day and all	**that**, I still believe Mr I can turn
5	ay presented a Democratic analysis	**that**, she said, showed that the Rep
6	ter the Labor Department announced	**that** consumer prices edged up just
7	week, announcing	**that** gold-gold collisions had produ
8	eople's Lottery. Camelot can argue	**that** it was carrying
9	Gore himself has argued	**that** his dual origins gave him a pe
10	rats who have been arguing	**that** budget surpluses won't be enou

Após alguns minutos, pedimos que apresentassem suas conclusões, que íamos transcrevendo no quadro. Ao final da discussão, projetamos um resumo dos diferentes usos em *slides* do PowerPoint, que foram disponibilizados ao público ao término do curso. O quadro 5 mostra parte de um dos slides salientando duas das funções de *what.*

Quadro 5: Parte do *slide* apresentando algumas funções de *what.*

➢	**Pronome**
5	g dismissed. "**What** can I say**?**" he said. "I was dr
23	"**What** is the most suffering? **What** is less suffering**?**" As
34	sed, "Scientific Dilemmas: **What**'s Lurking Out There in Space**?**"
➢	**Introdução de uma oração que complementa verbo, preposição ou conectivo**
6	will **do** to computers **what** computers did to slide rules?
13	The CHC put together a case **about what** had happened to the family.
15	Everybody in this room knows **what** happened there. That Ja

Outro problema recorrente nas redações foi o uso inadequado de preposições, em especial *on*. Uma concordância com uma seleção de ocorrências com essa preposição foi apresentada aos candidatos solicitando que identificassem os padrões em que ocorre. O quadro 6 traz um recorte dessa linhas:

Quadro 6. Seleção de linhas de concordância para a preposição *on.*

45	dinner invitations. "**On** the other hand," Bloodworth-Thom
48	ect of spending last night **on** the streets. 'There were
49	troom. "TV violence is not **on** trial. Professional wrestling is
50	all nine of their number **on** the Judiciary Committee vote aga
51	plan to detail their objections **on** the Senate floor before the fina
52	Hughes, Mr Mehmet had told police **on** at least ten occasions that he k
53	will invite us to reflect **on** the beastliness of the Germans o
55	inter. Alexander said **on** NBC's "Meet the Press" that ther
56	out **on** the streets,' Ms Casey said **on** Today . 'I reach a stage
57	g or deliver the Socratic seminars **on** litter laws and good cit
58	eased from prison and sent **on** community service programmes ins
59	t... different. Policies are short **on** ideas and entail more
60	0 witnesses, including specialists **on** voting rights and voting
61	o monitor and control big spending **on** television and print ads be used
63	he ANC, with its eccentric stances **on** Aids and press freedom, proved t
64	gest Bush administration statement **on** the subject to date, Fle
67	y. "My personal views **on** abortion are well known," Ashcro
68	d order issues. However, his views **on** certain issues —

Além de identificarem o conector *On the other hand*, também observaram o uso da preposição com o verbo *spend* (*spending last night on the streets, big spending on television and print ads*), com o adjetivo *short* (*short on ideas*), com o substantivo *views* (*views on abortion, views on certain issues*), entre outros. No entanto, o que lhes chamou a atenção foi *including specialists on voting rights* (linha 60), pois acreditavam que a preposição correta

seria *in.* De fato, é o que os dicionários apresentam. Por exemplo, o *Macmillan English Dictionary for advanced learners of American English* (2006) define *specialist* como

> Someone whose training, education, or experience makes them an expert in a particular subject: *a web design specialist* **a**. a doctor who is an expert in a specific area of medical work: *Both her husband and brother are specialists in hand surgery.* (p. 1350)

Nota-se que, tanto na definição, quando emprega o sinônimo *expert* (*expert in a particular subject, expert in a specific área of medical work*), quanto no exemplo na acepção **a**, aparece a preposição *in*: *specialists in hand surgery*.

Já o *Oxford Collocations dictionary for students of English* (2002), na entrada para *specialist* em que elenca as preposições com que o substantivo co-ocorre, dá exemplos com ambas as preposições:

> **~in** She is a specialist in eighteenth-century English painting. **~on** a specialist on the history of this city. (p. 733)

No entanto, não oferece qualquer explicação para diferenciar o uso de uma ou outra. Numa tentativa de identificar alguma diferença, buscamos outros exemplos, que apresentamos em *slides* (Quadros 7 e 8):

Quadro 7: *slide* com exemplos para *specialist in.*

➢ Specialist **in**
• Charles K Hole, specialist **in restoration of antique furniture**
• Dental Association, Specialist **in Community Medicine**
• another specialist **in cancer medicine** is
• Garratt & Co. is a specialist **in assisting firms** to make
• Maria Balinska, a specialist **in North American affairs**,
• David Willis, a specialist **in US affairs**, reports.
• officer, a specialist **in Oriental languages**
• becoming a specialist **in his chosen area**.

Quadro 8: *slide* com exemplos para *specialist on.*

➢ Specialist **on**
• Jonathan Fryer, a specialist **on southern African affairs**, explains.
• Jacques Bekaert, a specialist **on Indochina**,
• the well-known Daily Express specialist **on defence and intelligence**,
• the cla's top specialist **on the Soviet Union**.
• specialist **on animal behaviour**.
• A specialist **on defense issues**, he covered US
• Gary Milhollin, a specialist **on nuclear proliferation**
• Dr. Klaus Peter Treiter, a specialist **on European issues**,

Embora não se possa estabelecer uma diferença categórica, pois, por exemplo, as duas preposições ocorrem com *affairs*, é possível postular que há uma tendência para o uso de *in* quando se refere a uma área (*câncer medicine, Community medicine, oriental languages*) ou a serviços a serem prestados (*restoration of antique furniture, assisting firms*), enquanto *on* ocorre quando se trata de assuntos (*issues*) específicos (*Indochina, Soviet Union, animal behaviour, defense issues, nuclear proliferation, European issues*).

Com essa questão, os candidatos se familiarizaram com uma forma de esclarecer dúvidas, ou seja, recorrer a um *corpus* para buscar exemplos e compará-los.

Uma palavra que apresentou vários erros colocacionais foi *opportunity*. Um dos problemas foi a preposição que a segue: *to* ou *for*. Para essa questão, o exercício foi comparar concordâncias com uma e outra preposição. Mas também ocorreram erros com os verbos que co-ocorrem com *opportunity*. Além disso, notamos que os candidatos pouco variavam os adjetivos com os quais a palavra co-ocorria, privilegiando, na maioria dos casos, *good* ou *great*. Dessa forma, foi-lhes apresentado um exercício em que deveriam identificar os adjetivos que qualificavam *opportunity* e os verbos que co-ocorriam com esse substantivo. O quadro 9 apresenta uma seleção das linhas de concordância do exercício:

Quadro 9: Seleção de linhas de concordância para *opportunity/opportunities*.

20	Patrick Byrne, however, the biggest **opportunity** is in picking through
21	decades. "The NHS now has the best **opportunity** it has ever had to bri
22	aid. "We think he gives us the best **opportunity** right now to win a bas
23	eir slowness in grasping commercial **opportunities** offered by the inter
24	d was later revived, and gave early **opportunities** to young directors i
25	s neighbourhood, makes people enjoy **opportunities** and ideas which were
26	the time and they will seize every **opportunity** to get into the game."
27	vels. Amidst the risks are exciting **opportunities**, too. A clear signal
28	ng year, and a range of fascinating **opportunities** have recently arisen
29	e chance, they will seize the first **opportunity** to leap out of the tra
30	No Beat of Drum (1966), which gave **opportunity** for her to have her pr
31	y scoring chances. "We were getting **opportunities** and were not putting
32	rement years could provide a golden **opportunity** for Chile. It could he
33	d to take advantage of every golden **opportunity**. The Boilermakers did
34	ng, they simply spurned more golden **opportunities** than any Premiership
35	ove to manage," he said. "If a good **opportunity** comes up, I would jump
36	p, I would jump at it. I had a good **opportunity** here, and it didn't wo
37	place where you can find some good **opportunities**." The tight labor ma
38	igned this offseason. "It's a great **opportunity** for me to come in ther
39	attitude. "The states have a great **opportunity** to change public healt
40	se free hours are no longer a great **opportunity** but become filled with
41	teachers, and in some cases greater **opportunities** to get involved in y
42	small school, you will have greater **opportunities** for knowing your tea
43	training courses, may offer greater **opportunities** for customizable pub
44	s him the chance, he will seize his **opportunity** with both hands. "Thin
45	ut taking advantage of the historic **opportunity** that will present itse
46	can take, but also what interesting **opportunities** you might have to se
47	cow market for promising investment **opportunities**. Logoshin ascribes C
48	e says, is about family and "missed **opportunities**." So often, she says
49	But it is also a tale of missed **opportunities** - how an impressive t
50	try's recent expansion present more **opportunities** for career advanceme

As colocações identificadas foram resumidas no seguinte *slide:*

Quadro 10: *Slide* com o resumo das colocações para *opportunity/opportunities*.

- **Adjectival collocations:** ample, biggest, best, early, exciting, fascinating, golden good, great, greater, historic, interesting, new, rare, perfect
- **Verbal collocations:** offer, give, have, miss, provide, pursue, grasp, seize, take advantage of, spurn, open up
- **Intensifiers:** more, a lot of, plenty of, a wealth of

Alguns dos outros tópicos abordados foram as diferenças entre

- the /a(n) / zero article
- concerning / regarding / regard / regards
- despite / in spite of
- economy / economic / economical
- such / like
- big / large /great
- reduce / decrease
- increase / enlarge / enhance

5.3 Terceiro dia: avaliação em sala de aula

O terceiro dia foi dedicado a exercícios de averiguação do aprendizado, sem qualquer tipo de avaliação formal. Um dos exercícios requeria o preenchimento de lacunas em trechos de textos jornalísticos relacionados a assuntos que foram identificados a partir das palavras-chave do *corpus*. Dessa forma o candidato estaria trabalhando com textos pertinentes ao universo de sua atuação futura (Figura 12).

Quadro 11: Exercício de preenchimento de lacuna.

_1__ United States and _2__ allies have long recognized _3__ power of _4__ jihadi narrative, but _5__ attempts to overcome it have been ill-conceived. _6__ Al Qaeda's story justifies _7__ violence by claiming _8__ mantle of _9__ victimhood, but _10__ Bush administration responded like _11__ boxer, hitting back with _12__ "war on terror" and _13__ "battle of ideas." _15__ war metaphor only played into _16__ terrorists' hands, and _17__ Obama administration is rightly moving away from it.

No exercício acima o candidato deveria preencher as lacunas com o artigo definido *the* ou indefinido *a(n)*, quando necessário. O texto corretamente preenchido foi apresentado em *slide* (quadro 12):

Quadro 12: Exercício de preenchimento de lacuna completado.

The United States and **its** allies have long recognized **the** power of **the** jihadi narrative, but **Ø** attempts to overcome it have been ill-conceived. **Ø** Al Qaeda's story justifies **Ø** violence by claiming **the** mantle of **Ø** victimhood, but **the** Bush administration responded like **a** boxer, hitting back with **a** "war on terror" and **a** "battle of ideas." **The** war metaphor only played into **the** terrorists' hands, and **the** Obama administration is rightly moving away from it.

Outro tipo de exercício consistia de concordâncias com apagamento da palavra de busca. No exercício abaixo o candidato deveria descobrir qual forma de *concern** ou *regard** seria correta.

Quadro 13 : Exercício de preenchimento da palavra de busca[10].

In the following exercise, try to figure out what the missing word is.

a. Concern*/regard*

1	cans support the key recommendations of the Iraq Study Group	the withdrawal of U.S. troops and talking with I
2	elming majority of both U.S. and international media reports	"global warming" as a foregone conclusion is tha
3	irely upon climate models. We all know the frailty of models	the air-surface system." — Atmospheric scientist
4	SCO in place, the future for export may require concessions	the nuclear issue with Iran. • The U.S. is also
5	e South. It would allow developing economies to grow without	to atmospheric limits—and without the budgetary
6	China, Japan, Russia, and India, seek to emulate it in this	, the already voracious military component of glo
7	d countries with diminishing reserves, and security concerns	Iranian profits from inflated oil prices. The an

Como, em alguns casos, *concerning* e *regarding* podem ser usados como sinônimos, foram dadas as duas opções quando pertinente (quadro 14):

Quadro 14: Exercício de preenchimento da palavra de busca completado.

1	cans support the key recommendations of the Iraq Study Group **concerning / regarding**	the withdrawal of U.S. troops and talking with I
2	elming majority of both U.S. and international media reports **concerning / regarding**	"global warming" as a foregone conclusion is tha
3	irely upon climate models. We all know the frailty of models **concerning / regarding**	the air-surface system." — Atmospheric scientist
4	SCO in place, the future for export may require concessions **concerning / regarding**	the nuclear issue with Iran. • The U.S. is also
5	e South. It would allow developing economies to grow **without regard to**	atmospheric limits—and without the budgetary
6	China, Japan, Russia, and India, seek to emulate it **in this regard** , the	already voracious military component of glo
7	d countries with diminishing reserves, and security concerns **regarding**	Iranian profits from inflated oil prices. The an

Por fim, os candidatos receberam nove textos para serem traduzidos para o português, uma vez que, em seu futuro cargo, deveriam estar aptos a compreender e eventualmente traduzir textos em inglês. A correção, a partir de uma tradução proposta, suscitou várias sugestões por parte dos candidatos, em especial quando se tratava de terminologia específica da área. Assim, enquanto a tradução sugerida – e apresentada em slide – apresentava as opções:

> "integrantes/membros do Escritório do Representante de Comércio/Comercial dos Estados Unidos / os representantes do comércio dos Estados Unidos"

[10] A fonte dessas concordâncias foi reduzida para se adequar ao espaço da revista. Na apostila, elas aparecem em formato paisagem.

os candidatos sugeriram "integrantes/membros do Escritório de Representação Comercial dos Estados Unidos", opção não apresentada na tradução proposta.

Em outro texto, a lista de sugestões para traduzir o adjetivo *overwhelming* em

The solution to the climate crisis is as simple as it is overwhelming.

foi bem extensa: 'atordoante', 'arrebatadora', 'gigantesca', 'hercúlea', 'homérica', 'complexa', enquanto a tradução proposta foi 'impressionante'.

Nos exercícios de tradução os candidatos pareciam estar mais à vontade e mais motivados a contribuir com sugestões diretamente relacionadas a sua futura área de atuação.

5.4 Exercícios de versão

O último dia foi dedicado a exercícios de versão para o inglês, que os candidatos deveriam ter feito em casa. Entretanto, poucos o fizeram, certamente por falta de tempo, uma vez que no outro período do dia assistiam a um curso preparatório de português. Porém, como já havíamos preparado uma versão para ser apresentada e, principalmente, para ser discutida com os candidatos, o exercício foi bastante proveitoso, principalmente porque muitas sugestões ou dúvidas puderam ser esclarecidas por meio de consulta ao COCA, uma vez que tínhamos acesso à internet e a tela de busca podia ser projetada para visualização da classe.

6. Resultados

O método de revisão utilizando a Linguística de *Corpus* foi aplicado com resultados satisfatórios, conforme se pode observar no decorrer das atividades. O contato com uma metodologia ainda pouco conhecida no ensino de línguas estrangeiras foi bastante motivante, em especial por se tratar de uma abordagem computacional, universo com que os candidatos estão familiarizados. Em avaliação informal a maioria dos alunos expressou satisfação em relação ao método, que correspondeu às suas expectativas de revisão do conteúdo face ao curto tempo disponível, embora alguns tenham criticado o fato de "sempre se recorrer ao *corpus*" para sanar alguma dúvida. Enquanto para os primeiros a Linguística de *Corpus* representou uma porta para um novo universo, que poderiam explorar por si, para os últimos representava uma falha no conhecimento da ministrante do curso, que, em caso de dúvidas, "tinha de recorrer ao *corpus*". A esses interessavam "respostas corretas" que deveriam ser

fornecidas pela professora. No geral, no entanto, afirmaram que ficaram estimulados a passar a usar a metodologia em suas vidas profissionais.

7. Considerações finais

Como era de se esperar, a Linguística de *Corpus* mostrou-se uma abordagem pedagógica eficaz para o tipo de curso descrito – uma rápida revisão gramatical da língua inglesa, visando também a tradução e a versão de textos da área de Relações Exteriores. O resultado alcançado comprovou que a abordagem pode ser replicada para qualquer área profissional, desde que haja *corpora* disponíveis. Quando não os há, construir um *corpus* para esse fim é tarefa bastante fácil e rápida, com vimos acima. Quanto aos alunos, uma vez familiarizados com a metodologia, esses podem tornar-se pesquisadores capazes de buscar respostas para questões que eventualmente surjam durante seu aprendizado, imergindo-se em um processo autodidata de aprendizagem. Ressalte-se, entretanto, que sempre haverá os que preferem depender do conhecimento do professor a desenvolver sua própria autonomia de aprendizado.

Referências bibliográficas

BARNBROOK, G.; MASON, O.; KRISHNAMURTHY, R. **Collocations:** Applications and Implications. Basingstoke: Palgrave Macmillan, 2013. crossref http://dx.doi.org/10.1057/9781137297242

BARONI, M.; BERNARDINI, S. **BootCat - Bootstrapping corpora and terms from the web**. Proceedings of LREC 2004 Conference. Lisboa: [s.n.]. 2004. p. 1313-1316.

BERBER SARDINHA, T. **Linguística de *Corpus***. Barueri: Manole, 2004.

BERNARDINI, S.; FERRARESI, A. Old needs, new solutions - comparable *corpora* for language professionals. In: SHAROFF, S., et al. **Building and Using Comparable Corpora**. Berlin - Heidelberg: Springer Verlag, 2013. p. 303-319. crossref http://dx.doi.org/10.1007/978-3-642-20128-8_16

HALLIDAY, M. A. K. Categories of the theory of grammar. **Word**, Vol. 17, No. 3 1961. 241-292.

JOHNS, T. Whence and whither classroom concordancing? In: BONGAERTS, T., et al. **Computer applicatons in language learning**. Dordrecht: Foris Publications, 1988. p. 9-27.

JOHNS, T. Should you be persuaded: two samples of data driven learning. **ELR**, vo. 4 1991. 27-46.

JOHNS, T. **Data-driven Learning:** The Perpetual Challenge. Proceedings of the Fourth International Conference on Teaching and Language Corpora - Graz, 19-24 July 2000. Amsterdam / New York: Rodopi. 2002. p. 107-117.

MACMILLAN EDUCATION. **Macmillan English Dictionary for advanced learners of American English**. Oxford: Macmillan Publishers Limited, 2006.

OXFORD UNIVERSITY PRESS. **Oxford Collocations dictionary for students of English**. Oxford: Oxford University Press, 2002.

SCOTT, M. **WordSmith Tools**. Oxford: Oxford University Press, 2007.

TAGNIN, S. E. O. **O Jeito que a Gente Diz**. São Paulo: Disal, 2013.

TRIBBLE, C. **Concordances in the classroom**. [S.l.]: [s.n.].

VARANTOLA, K. Disposable corpora as intelligent tools in translation. **Cadernos de Tradução**, 2002. 171-189.

Artigo recebido em: 15.10.2014
Artigo aprovado em: 02.12.2014

A (in)existência de neutralidade: um estudo de caso baseado em *corpus* com roteiros de audiodescrições francesas de filmes via Teoria da Avaliatividade

(Non)existence of neutrality: a case study based on a corpus composed of French audiodescription scripts of movies via Appraisal Theory

Cristiene Ferreira da Silva[*]
Pedro Henrique Lima Praxedes Filho[*]

RESUMO: Este artigo relata um estudo descritivo que dedicou-se à Audiodescrição (AD), modalidade da Tradução Audiovisual Acessível (TAV-Ac) responsável pela acessibilidade sociocultural de deficientes visuais (DVs), e tratou do registro 'roteiro de AD de longas', buscando investigar, via Linguística de Corpus (LC), a presença-ausência de neutralidade segundo a Teoria da Avaliatividade (TA)-Linguística Sistêmico-Funcional (LSF). O corpus foi dividido em dois subcorpora, cada um contendo um roteiro de AD escrito pela mesma audiodescritora francesa: Intouchables (C1)-Minuit à Paris (C2). Foram elaboradas sete etiquetas para cada subcorpus via termos/escolhas dos sistemas da rede de avaliatividade da TA até o segundo nível de delicadeza, as etiquetas foram inseridas e os subcorpora compilados foram submetidos às ferramentas WordList e Concord do Word Smith Tools. Os resultados demonstraram a inexistência de neutralidade por serem os roteiros avaliativos/interpretativos por 'atitude' e seus três tipos, por 'engajamento' e seus dois tipos e por 'gradação' e seus dois tipos, tendo evidenciado uma tendência de padrão avaliativo caracterizado pela predominância de 'gradação'-'força' e 'atitude'-'apreciação'. A LC foi imprescindível para o êxito do estudo por ter viabilizado a inserção de 14 etiquetas derivadas de 10 termos/escolhas avaliativos em 10.387 palavras (C1+C2) e uma análise com 2.422 etiquetagens. A partir desses

ABSTRACT: This article reports on a descriptive study that dealt with Audiodescription, a branch of Accessible Audiovisual Translation involved in the socio-cultural accessibility of the blind and visually impaired, and the register 'AD script of feature films', aimed at investigating, via Corpus Linguistics, the presence-absence of neutrality through Appraisal Theory-Systemic-Functional Linguistics. The corpus was divided into two subcorpora, each of which containing an AD script written by the same French audiodescriber: *Intouchables* (C1)-*Minuit à Paris* (C2). Seven tags were made up for each subcorpus via the terms/choices within the systems of the appraisal network up to the second delicacy level, the tags were inserted, and the compiled subcorpora were treated by the WordList and Concord tools on Word Smith Tools. The results evidenced the **non**existence of neutrality as the scripts are evaluative/interpretative from the perspectives of 'attitude'+its three types, 'engagement'+its two types, and 'graduation'+its two types, having indicated a tendency of an evaluative pattern characterized by the predominance of 'graduation'-'force' and 'attitude'-'appreciation'. CL was vital for the study's success as it made feasible the insertion of 14 tags derived from 10 evaluative terms/choices in 10,387 words (C1+C2), and an analysis with 2,422 taggings. Based on these results, we suggest other studies: (i)

[*] Mestre em Linguística Aplicada pelo Programa de Pós-Graduação em Linguística Aplicada (PosLA) do Centro de Humanidades (CH) da Universidade Estadual do Ceará (UECE).
[*] Doutor em Letras e professor do Curso de Letras e do Programa de Pós-Graduação em Linguística Aplicada (PosLA) do Centro de Humanidades (CH) da Universidade Estadual do Ceará (UECE).

resultados, sugerimos outros estudos: (i) descrição da assinatura avaliativa da audiodescritora via corpus com mais roteiros de AD de longas por ela escritos e (ii) descrição do estilo avaliativo do registro 'roteiro de AD de longas' via corpus com roteiros de AD de um mesmo longa francês, escritos por diferentes audiodescritores.

PALAVRAS-CHAVE: Tradução Audiovisual Acessível. Audiodescrição francesa fílmica. Descrição de neutralidade. Linguística de *Corpus*. Teoria da Avaliatividade-Linguística Sistêmico-Funcional.

description of the audiodescriber's evaluative signature via a corpus with more AD scripts of feature films written by her, and (ii) description of the evaluative style of the register 'AD script of feature films' via a corpus with AD scripts of the same French feature film, written by different audiodescribers.

KEYWORDS: Accessible Audiovisual Translation. French film Audiodescription. Neutrality description. Corpus Linguistics. Appraisal Theory-Systemic-Functional Linguistics.

1. Introdução

Este estudo trata da descrição de roteiros de audiodescrições francesas de filmes quanto à existência, ou não, de neutralidade. Em outras palavras, a temática diz respeito à ausência ou presença de interpretação através de marcas de posicionamento avaliativo por parte do tradutor/audiodescritor.

Epistemologicamente, o estudo se insere na área maior dos Estudos Descritivos da Tradução (EDT); nela, localiza-se na subárea dos Estudos da Tradução Baseados em *Corpus* (ETBC), na qual, por sua vez, restringe-se ao recorte da Tradução Audiovisual Acessível (TAV-Ac) quanto à modalidade Audiodescrição (AD) para deficientes visuais (DVs). Institucionalmente, encontra-se inserido no projeto 'A neutralidade em audiodescrições de produtos audio(visuais) e/ou o estilo do roteiro de AD e/ou a assinatura do audiodescritor: um estudo via Teoria da Avaliatividade (TA)', conduzido pelo coautor, Dr. Pedro Henrique Lima Praxedes Filho, e equipe no âmbito do projeto maior 'A locução na audiodescrição para pessoas com deficiência visual: uma proposta para a formação de audiodescritores' (LOAD), desenvolvido, por seu turno, sob a coordenação geral da Dra. Vera Lúcia Santiago Araújo no Laboratório de Tradução Audiovisual (LATAV) do Programa de Pós-Graduação em Linguística Aplicada (PosLA) do Centro de Humanidades (CH) da Universidade Estadual do Ceará (UECE). O primeiro projeto tem por objetivo, dentre outros, fornecer evidência empírica ao segundo quanto à (in)existência de neutralidade relativamente ao parâmetro prescritivo de neutralidade em roteiros de AD.

No que se refere ao parâmetro de neutralidade, nos Estados Unidos, país onde a AD começou (FRANCO; SILVA, 2010) como profissão no universo extra-academia, a última

edição, ainda em vigor, do documento *Standards for Audio Description and Code of Professional Conduct for Describers* (2009) da organização *Audio Description Coalition* mantém as seguintes orientações prescritivas:

> Esta é a primeira regra de descrição [descreva o que você vê]: *o que você vê* é o que você descreve. Você vê aparências e ações físicas; você *não vê* motivações ou intenções. Nunca descreva o que você *acha* que vê. (...)
> Permita que os ouvintes [DVs] formem suas próprias opiniões e cheguem às suas próprias conclusões. Não edite, interprete, explique, analise (...)
> Se a conclusão é que um personagem está com raiva, descreva o que lhe levou a essa conclusão – os gestos/as expressões faciais do personagem. Os humores, as razões ou o raciocínio de um personagem não são visíveis e, portanto, não devem ser descritos. (...)
> Use somente aqueles adjetivos e advérbios que não oferecem julgamentos de valor e que não são (...) sujeitos à interpretação. (...)
> Ao invés de dizer que uma pessoa, uma roupa, um objeto etc. é bonito/a, descreva o que você viu e que lhe levou a essa conclusão, de tal forma que os ouvintes [DVs] possam chegar às suas próprias conclusões. (...)
> É mais interessante listar os itens que estão em um amontoado de coisas, se o tempo permitir, do que dizer: "O sótão está amontoado". (...)
> Não acrescente 'cerca de' ou 'aproximadamente' para qualificar (...) dimensões estimadas (...).[1] (itálico no texto fonte) (p. 1-3)

Foram os EUA que exportaram a AD e o parâmetro de neutralidade obrigatória para o mundo. Na Alemanha, por exemplo, Benecke (2004), ao estabelecer etapas para a AD alemã, afirmou que uma "[b]oa audiodescrição deve ser discreta [invisibilidade da voz do audiodescritor] e neutra (...)"[2] (p. 80). Também no Brasil, trabalhos recentes manifestam-se favoráveis à não interpretação ou à neutralidade do roteiro de AD. Para Vilaronga (2009), "a busca da fidelidade ao filme deve ser perseguida pelo audiodescritor(a), evitando antecipar, *julgar ou interpretar* o filme" (itálico nosso) (p.1060). Silva *et al.* (2010) avalizam diretrizes para a invisibilidade do audiodescritor, igualmente

[1] As traduções, salvo indicação contrária, são de nossa autoria. "This is the first rule of description: what you see is what you describe. One sees physical appearances and actions; one does not see motivations or intentions. Never describe what you think you see. … Allow listeners to form their own opinions and draw their own conclusions. Don't editorialize, interpret, explain, analyze … If the conclusion is that a character is angry, describe what led to that conclusion – the gestures/facial expressions of the character. Character's moods, motives or reasoning are not visible, thus, not subject to description. … Use only those adjectives and adverbs that do not offer value judgments and that are not … subject to interpretation. … Instead of saying the person, clothing, object, etc. is beautiful, describe the things observed that caused your conclusion – so listeners may draw their own conclusion. … It is more interesting to name the items in the clutter if time permits than to say, 'The attic is cluttered'. ... Don't add 'about' or 'approximately' to qualify … estimated dimensions (…)."

[2] "Good audio-description should be unobtrusive and neutral (…)." Na verdade, Rai, Greening e Petré (2010, p. 8) informam que "...os parâmetros alemães estabelecem que todas as palavras escolhidas devem ser o mais possível imparciais a fim de que o espectador [DV] possa ter a oportunidade de tomar sua própria decisão..." (texto fonte: "…the Germam guidelines state that all the words chosen should be as impartial as possible – so that the viewer has a chance to make up his own decision...").

prescrevendo uma atitude de neutralidade. Tratando sobre a inclusão social de DVs em relação à publicidade brasileira via AD, Navarro (2012) endossa 'regras de ouro' para o audiodescritor como: "descrever o que está lá, *não dar uma versão pessoal do que está lá*" (itálico nosso) (p. 18).

Em relação à França – um outro país engajado na promoção de AD, cujo marco inicial data de 1989 (FRANCO; SILVA, 2010) –, há o documento *La Charte de Qualité de l'Audiodescription*[3] (FRANÇA, 2008), que é contraditório quanto à questão da neutralidade. A *Charte* (carta) expressa os princípios e as orientações para a AD francesa de produtos (áudio)visuais, estabelecendo padrões mínimos de referência para os profissionais da área. Dentre as informações relevantes, ela, inicialmente, estabelece

> [u]m quadro ético dos princípios básicos:
> *O trabalho de audiodescrição é um trabalho de autor. É um trabalho completo de criação: trata-se de escrever um texto inédito a partir de um suporte visual. Descrever uma obra é compreendê-la, analisá-la, restituir o sentido, para transmitir sua mensagem e provocar emoção pela verbalização.*[4] (itálico no texto fonte) (FRANÇA, 2008, p. 5)

Em caminho oposto e incompatível a essas recomendações sobre a AD enquanto trabalho de autor e de criação, que deve provocar emoção pela verbalização, o mesmo documento também determina, paradoxalmente, que

> [o]s seguintes princípios devem ser seguidos: (...) O audiodescritor *não deve interpretar as imagens, mas descrevê-las* (...);
> Modo operacional: (...) A voz deve (...) *manter uma certa neutralidade*.[5] (itálicos nossos) (FRANÇA, 2008, p. 5-8)

Ainda na França, Bernengo (2012) defende que "[o] audiodescritor deve se manter fiel ao que é factual, objetivo e *não deve emitir julgamento* (...)"[6] (p. 27) (itálico nosso). Esse posicionamento reforça o paradoxo presente na *Charte*.

---

[3] A tradução deste documento para a língua portuguesa, na variante brasileira, feita pela autora Cristiene Ferreira da Silva, foi objeto de estudo do Trabalho de Conclusão de Curso, nível *lato sensu* (UECE), intitulado 'Aspectos relevantes na tradução da *charte* francesa para a audiodescrição' (FERREIRA, 2013).

[4] "Un cadre éthique, des principes fondamentaux:
Le travail d'audiodescription est un travail d'auteur. C'est un travail de création à part entière: il s'agit d'écrire un texte inédit à partir d'un support visuel.
Décrire une oeuvre, c'est la comprendre, l'analyser, la décrypter pour transmettre son message et provoquer émotion par la verbalisation."

[5] "Les principes suivants doivent être suivis: (...) L'audiodescripteur ne doit pas interpréter les images mais les décrire (...); Mode opératoire: (...) La voix doit (...) doit néanmoins garder une certaine neutralité."

[6] "Le descripteur doit rester dans le factuel, objectif et ne pas émettre de jugement (...)."

Portanto, podemos depreender que o parâmetro prescritivo de neutralidade, surgido inicialmente como parte do *modus faciendi* da AD e posteriormente incorporado pela academia, diz respeito à ausência da voz autoral do audiodescritor no roteiro de AD no sentido da necessidade, pelo 'bem' das pessoas DVs, de essa voz ser imparcial. Entendemos por necessidade de imparcialidade a interdição de que o audiodescritor avalie ou interprete, fazendo juízos de valor, o texto visual que descreve verbalmente.

Apesar de a tendência prescritiva ser favorável à não avaliação/interpretação, as contradições no documento francês incitam questionamentos relativos à neutralidade e surgem as dúvidas: um roteiro de AD escrito sob a égide da regra da neutralidade é realmente neutro? Essa neutralidade é possível?

Essa falta de unanimidade sobre a questão da neutralidade nos levou a querer estudar o assunto nos roteiros de AD, em língua francesa, dos filmes *Intouchables* (de Eric Toledano e Olivier Nakache) e *Minuit à Paris* (de Woody Allen, dublado em francês), audiodescritos na França, em 2011, pela mesma profissional. As dúvidas surgidas a partir da problematização identificada definiram as metas do estudo. No nível geral, o objetivo foi estudar a (in)existência de neutralidade, concernente ao parâmetro prescritivo de neutralidade para ADs, em roteiros de AD francesa de filmes, tendo sido o objetivo específico assim estabelecido:

- Investigar a presença ou ausência de neutralidade nos roteiros de AD em francês dos filmes *Intouchables* e *Minuit à Paris*, quanto a possíveis marcas de posicionamento avaliativo/interpretativo de 'atitude', 'engajamento' e/ou 'gradação'[7] por parte da audiodescritora.

Esse objetivo suscitou as perguntas abaixo elencadas:

- Como se caracterizam os roteiros de AD em francês dos filmes *Intouchables* e *Minuit à Paris* acerca da presença ou ausência de neutralidade operacionalizada pela ausência ou presença, respectivamente, de marcas de posicionamento avaliativo/interpretativo...
  I. ... de 'atitude' quanto aos TIPOS DE ATITUDE: 'afeto', 'julgamento' e/ou 'apreciação'?

[7] A 'atitude', o 'engajamento' e a 'gradação' são as três grandes áreas de significados avaliativas previstos pela Teoria da Avaliatividade (MARTIN; WHITE, 2005) e serão apresentadas na Seção 2.4.

II. ... de 'engajamento' quanto aos TIPOS DE ENGAJAMENTO: 'monoglossia' ou 'heteroglossia'?
III. ... de 'gradação' quanto aos TIPOS DE GRADAÇÃO: 'força' e/ou 'foco'? e
IV. Os roteiros de AD em francês dos filmes *Intouchables* e *Minuit à Paris* são neutros ou avaliativos/interpretativos e, se avaliativos/interpretativos, assim são de modo semelhante ou diferente?

Quando a AD deixa de ser da exclusividade do mundo da prática e passa a ser do interesse também do mundo teórico-acadêmico, com sua inserção na subárea da TAV-Ac dos EDT, o parâmetro de neutralidade perde definitivamente seu status de unanimidade pelo simples fato de os teóricos da tradução – por terem, em geral, formação também em Linguística e/ou Linguística Aplicada –, saberem que nenhum texto pode ser neutro[8]. Contudo, para que aqueles que passaram a teorizar sobre AD possam convencer os praticantes a respeito desse fato, tem sido necessário desenvolver pesquisas com o fim de demonstrá-lo empiricamente.

Esse foi o caso de PRAXEDES FILHO e MAGALHÃES (2013a,b), que analisaram roteiros de AD de pinturas via TA, Holland (2009), que discorreu – de modo impressionístico –, sobre a impossibilidade de neutralidade em roteiros de AD nas artes visuais, e Jiménez Hurtado (2007), que estudou roteiros de AD de filmes. Apesar de os estudos empíricos, o primeiro e o terceiro, terem demonstrado a ausência de neutralidade, o estudo ora relatado se justificou pelo fato de ter sido o primeiro a tratar de roteiros de AD em língua francesa para o cinema sob o olhar da TA, que trata da avaliatividade de maneira holística. Não obstante Jiménez Hurtado (2007) ter também estudado roteiros de AD fílmica, eram todos em língua espanhola e a neutralidade foi abordada sob outra perspectiva teórica, que consideramos reducionista pois contempla somente uma das nuances avaliativas descritas pela TA.

Passamos, agora, para a revisão da literatura.

[8] Para textos em geral, originais e traduzidos, Martin e White (2005, p. 94) asseveram que até asserções categóricas são permeadas por *intersubjetividade*. Para os textos traduzidos especificamente, Jakobson (1995/1959) fala em *interpretação* de signos por outros e a Linguística Sistêmico-Funcional os entende como retextualizações de textos fonte por novos autores, os tradutores, cujas vozes também se impõem.

2. Referencial Teórico

2.1 Estudos Descritivos da Tradução

Inicialmente, ressaltamos, pelas palavras de Hermans, o delineamento do escopo dos EDT:

> Em essência, os descritivistas consideram, como seu objeto de estudo, o que os tradutores, professores e críticos da tradução fazem e dizem. Desta forma, não só as traduções, como também declarações sobre elas, incluídos enunciados prescritivos e avaliativos, são matérias-primas para os estudos descritivos.[9] (HERMANS, 1999, p. 35)

Nesse sentido, entendemos, então, que o parâmetro da neutralidade orientado para o trabalho do audiodescritor é 'matéria-prima' do campo disciplinar dos EDT.

Ainda para Hermans (1999), os EDT marcam oposição ao prescritivismo e objetivam descrever os textos traduzidos (TTs) para poder entender e explicar sua natureza e compreendê-la no sistema de chegada. Para este estudo, elegemos os seguintes pressupostos hermansianos para além daquele já citado acima:

> Ao rejeitar uma abordagem de tradução prescritiva ou normativa, os descritivistas querem realizar pesquisas que se justifiquem por si mesmas, e não pesquisas que determinem conselhos práticos ou diretrizes para uma boa tradução ou regras de ouro para serem seguidas pelos tradutores ou ainda critérios com os quais críticos e comentaristas possam avaliar a qualidade de uma tradução. Desse modo, o termo 'descritivo' assinala um movimento deliberado da pesquisa 'aplicada' para a pesquisa 'pura', em um contexto histórico no qual a tendência 'aplicada' dominava há muito tempo (...).
>
> (...) o descritivismo redefine os objetivos de se estudar tradução ao defender a legitimidade de investigações que têm caráter de 'esclarecimento' em contraposição àquelas que têm caráter de 'uso', seguindo os termos de Holmes. Deseja estudar as traduções como elas são, dando conta de suas ocorrências e natureza. Estes objetivos podem gerar ideias que acabam sendo de relevância prática tanto para tradutores, quanto para professores e críticos de tradução (...).[10] (p. 35)

[9] "In essence, descriptivists regard what translators do and say, and what translation teachers and critics do and say, as their object of study. In this way not only translations but also statements about translation, including prescriptive and evaluative pronouncements, are grist to the descriptive mill."

[10] "In rejecting a prescriptive, or normative, approach to translation, the descriptivists want to conduct research for its own sake and not in order to distil from it practical advice or guidelines for good translating, or rules of thumb which translators should follow when they translate, or criteria with which critics and reviewers can assess the quality of translation. 'Descriptive' thus signals a deliberate shift away from 'applied' to 'pure' research, in a historical context in which the 'applied' tendency had long been dominant (…).
(…) descriptivism redefines the aims of studying translation by claiming legitimacy for research which is 'of light' rather than 'of use', to speak in Holmes's terms. It wants to study translations as they are, and to account for their occurrence and nature. These endeavors may yield insights that turn out to be of practical use to translators and to translation teachers and critics (…)."

Continuando a discussão a respeito da abordagem descritivista adotada neste estudo, vale salientar o que Pagano e Vasconcelos (2003) dizem, por reconhecerem seu importante papel acadêmico-científico, a respeito da proposta de Holmes (1988/1972) quanto aos subcampos dos Estudos da Tradução, à qual Hermans (1999) explicitamente se refere:

> Acredita-se que ele [Holmes] consegue capturar as mais tradicionais vertentes da pesquisa na área; além disso, a distinção por ele proposta entre *estudos aplicados* (voltados para a prática) e *estudos puros* (ou seja, estudos teóricos e descritivos [orientados para o produto, o processo ou a função] feitos sem preocupação com uma aplicação prática e direta) e suas subseqüentes divisões servem de norteamento para a pesquisa de tradução. (itálicos das autoras) (PAGANO; VASCONCELOS, 2003, p. 14)

Nessa perspectiva e consoante à sistematização da proposta holmesiana feita por Toury (1995), o que ficou conhecido como 'Mapa de Holmes', a pesquisa aqui relatada localiza-se no subcampo dos estudos 'puros', 'descritivos' e 'orientados para o produto'. Logo, descrevemos, sem intenção aplicada apriorística, produtos finais do fazer tradutório de uma dada tradutora/audiodescritora, isto é, textos traduzidos/roteiros de AD por ela produzidos.

2.2 Estudos da Tradução Baseados em Corpus e Linguística de *Corpus*

Nos Estudos da Tradução, a mudança para o enfoque descritivista possibilitou a expansão de novas perspectivas. Aliado ao descritivismo, os avanços na tecnologia computacional contribuíram, no decorrer dos anos 1990, conforme Bassnett (2003), para o surgimento de uma linha de pesquisa interessada em investigação com *corpus*. Nesse sentido, os trabalhos de Mona Baker são pioneiros no que tange à proposta de metodologias baseadas em *corpora* eletrônicos, usando ferramentas da Linguística de *Corpus* (LC), para estudar características do texto traduzido.

Baker (1996) esclarece que a pesquisa baseada em *corpora* propiciou contribuições significativas para os Estudos da Tradução, tendo dado origem aos Estudos da Tradução Baseados em *Corpus* (ETBC). No âmbito dos estudos descritivos, os ETBC fornecem suporte a várias linhas de pesquisa, possibilitando a busca por regularidades e padrões de TTs a partir de bases de dados eletrônicos ou *corpora* eletrônicos, passíveis de análise graças à LC. Portanto, os ETBC constituem-se numa abordagem interdisciplinar abrangendo os EDT e a LC, a fim de poder examinar dados autênticos e evidências empíricas da atividade tradutória. Nessa perspectiva, esta pesquisa apresenta sua primeira interface interdisciplinar, ao propor investigação envolvendo os EDT e a LC via ETBC.

Acerca do tema, Sinclair (1991), um grande expoente da LC, posicionando-se a favor da metodologia de investigação com *corpora*, evidencia as contribuições dos recursos tecnológicos para estudos na área da linguagem verbal, dentre as quais: maior precisão, abrangência, sistematização e confiabilidade na coleta e na análise de dados sob a forma de *corpora* eletrônicos. Conforme Berber Sardinha (2004), a definição de *corpus* proposta por Sanchez é a mais completa por levar em conta pontos importantes, tais como

> (a) a origem: os dados devem ser autênticos; (b) o propósito: o corpus deve ter a finalidade de ser um objeto de estudo linguístico; (c) a composição: o conteúdo do corpus deve ser criteriosamente escolhido; (d) a formatação: os dados do corpus devem ser legíveis por computador; (e) a representatividade: o corpus deve ser representativo de uma língua ou variedade; e (f) extensão: o corpus deve ser vasto para ser representativo. (BERBER SARDINHA, 2004, p. 18-19)

Nesse sentido, Saldanha (2009) afirma que a representatividade depende do propósito para o qual o *corpus* é usado, bem como dos traços linguísticos a serem estudados. No que diz respeito à extensão, Berber Sardinha (2004) classifica como pequenos os *corpora* com menos de 80 mil palavras.

Os trabalhos que envolvem *corpora* eletrônicos precisam contar com o auxílio de *softwares* que possibilitem análises e descrições detalhadas dos dados linguísticos, viabilizando maior eficiência, confiabilidade e a possibilidade de detectarmos fenômenos a partir dos dados. Ressaltamos que, para Berber Sardinha (2004), o programa *WordSmith Tools* de Mike Scott é eficiente como auxílio analítico. O referido *software* dispõe de três ferramentas básicas – que funcionam sob os princípios da ocorrência, da recorrência e da coocorrência –, quais sejam: *WordList, KeyWord* e *Concord*, respectivamente.

2.3 Tradução Audiovisual e Audiodescrição

No âmbito da TAV, Díaz Cintas (2004) destaca a importância de validar diversos postulados articulados nos EDT. Acrescenta que esses postulados representam um ponto de partida valioso para a TAV. O autor declara que,

> [a]o transcender a dimensão puramente linguística, os postulados apresentados pelos EDT têm a vantagem de colocar os pesquisadores em uma posição que lhes permite concentrar seus esforços no objeto de estudo a partir de uma perspectiva plural e interdisciplinar. *A tradução é considerada um ato de comunicação intercultural, ao invés de simplesmente interlinguística* (...). *As abordagens linguística e cultural não devem ser*

> *vistas como paradigmas antagônicos, mas sim complementares*.[11] (itálicos nossos) (DÍAZ CINTAS, 2004, p. 31)

Foi a ultrapassagem do exclusivamente linguístico que pavimentou o caminho que levou às pesquisas em TAV. Assim sendo, Díaz Cintas (2007) classifica a TAV como subárea dos Estudos da Tradução e defende que ela "(...) engloba as diferentes práticas tradutórias implementadas nos meios audiovisuais (...)[, nas quais] há uma interação semiótica entre o som e as imagens"[12] (p. 13). Ainda nesse sentido, Franco e Araújo (2011), discutindo questões terminológicas e conceituais no âmbito da TAV, nos fazem observar que, para esse autor, "(...) o meio audiovisual inclui todos os espaços onde há um sinal acústico e um sinal visual, independentemente de ser transmitido através de uma tela (...) ou de um palco (...)" (p. 3). Para Aderaldo (2014), com base em Jiménez Hurtado e Seibel (2007), deve ser considerado que a TAV tem uma subárea por ela denominada de TAV acessível (TAV-Ac), "relacionada às boas práticas de inclusão social" (p. 18) através da Legendagem para Surdos e Ensurdecidos (LSE) e da Audiodescrição (AD) para DVs, que é lócus onde se encontra, de fato, o estudo aqui relatado.

Vale ressaltar, agora, que foi Díaz Cintas que – em 2005, em artigo publicado no periódico *Translating Today* –, se manifestou pela inclusão da AD no escopo da TAV. De modo geral, a AD traduz o visual para o verbal, podendo ser, portanto, compreendida como a tradução de imagens em palavras. Quanto a sua realização, a AD pode ser pré-gravada, ao vivo ou simultânea. Em filmes, ela é pré-gravada e requer um roteiro com rubricas detalhadas para que, em estúdio, um texto oral resultante da leitura do roteiro seja inserido entre os elementos sonoros e gravado. A AD torna acessível – às pessoas totalmente sem visão ou com baixa visão (DVs), como já dito –, expressões artísticas, como: pinturas, esculturas, exposições, espetáculos de teatro e dança, filmes, programas de TV etc. Portanto, essa modalidade de TAV-Ac possibilita, às pessoas DVs, acessibilidade sociocultural, isto é, sua inclusão social e cultural, empoderando-as como cidadãs com direitos plenos.

Nos centros de pesquisa que desenvolvem estudos em TAV-Ac, o caráter tradutório da AD é justificado pela tipologia de Jakobson (1995/1959), que estabeleceu três possibilidades

[11] "By transcending the purely linguistic dimension, the postulates put forward by DTS have the advantage of placing translation researchers on a starting grid that allows them to channel their efforts into the object of study from a plural and interdisciplinary perspective. Translation is viewed as an act of intercultural communication, rather than simply interlinguistic (…). The linguistic and cultural approaches should not be viewed as antagonistic paradigms but, rather, complementary."

[12] "(...) encapsula las diferentes prácticas traductoras que se implementan en los medios audiovisuales (...) en un formato en el que hay una interacción semiótica entre el sonido y las imágenes."

de tradução: intralinguística, que consiste na 'interpretação' de signos verbais por meio de outros signos da mesma língua; interlinguística, que consiste na 'interpretação' de signos verbais por meio de signos de outra língua; e intersemiótica, que consiste na 'interpretação' de signos verbais por meio de signos de sistemas não-verbais. Nessa perspectiva, a AD é TAV-Ac do tipo intersemiótica. A esse respeito, Mascarenhas (2012) esclarece que

> (...) a partir de uma revisão da taxonomia proposta por Jakobson (...) para o conceito de tradução, incluindo a ela dimensões visuais e acústicas, verificamos que tanto a legendagem para surdos e ensurdecidos (LSE), quanto a audiodescrição (AD) podem ser consideradas práticas tradutórias. A primeira por sua natureza intralingual (...) – uma interpretação de códigos verbais orais por meio de códigos verbais escritos na mesma língua – ao passo que a segunda por sua essência intersemiótica – no caso, uma interpretação de códigos visuais por meio de códigos verbais orais. (p. 23)

Franco e Silva (2010) oferecem um panorama histórico da AD no Brasil e no exterior. Esclarecem que se trata de uma prática com pouco mais de trinta anos de existência, tendo nascido "em meados da década de 70 nos Estados Unidos, a partir das ideias desenvolvidas por Gregory Frazier" (p. 24). Acrescentam, ainda, que na década seguinte, a AD expande-se para fora do território americano, chegando na Europa pela Inglaterra, Espanha, França e Alemanha e ganhando atenção no Brasil em 2003, durante o festival temático *Assim Vivemos: Festival Internacional de Filmes sobre Deficiência*. Na atualidade, informam as autoras, Estados Unidos, Inglaterra, *França*, Espanha, Alemanha, Bélgica, Canadá, Austrália e Argentina são os países que mais investem em AD, tanto na televisão como no cinema e no teatro.

Observando ainda o contexto histórico mundial, verificamos que, há mais de vinte anos, a AD[13] foi introduzida na França. Conforme informações difundidas nesse país pelo *Conseil Supérieur d'Audiovisuel*[14] (*CSA*), pela *Association Valentin Haüy*[15] (*AVH*), pela *Association Française d'Audiodescription*[16] (*AFA*) e pela associação *En Aparté*[17], entre outras organizações, em 2008, instâncias governamentais, profissionais da área audiovisual e entidades engajadas na questão da acessibilidade para DVs reuniram-se para assinar *La Charte de Qualité de l'Audiodescription*. No total, 14 assinaturas atestaram a aprovação desse documento de referência, redigido, após dois anos de

[13] Na França, a AD é denominada por dois termos intercambiáveis: '*audiodescription*' e '*audiovision*' (Fonte: http://www.avh.asso.fr/rubriques/audiovision/tout_savoit_audiovision.php).
[14] Disponível em: <http://www.csa.fr/Espace-Presse/Communiques-de-presse/Public-non-voyant-ou-mal voyant-le-CSA-signe-la-Charte-de-l-audiodescription>. Acesso em: 29 set. 2012.
[15] Disponível em: <http://www.avh.asso.fr/rubriques/audiovision/tout_savoit_audiovision.php>. Acesso em: 29 set. 2012.
[16] Disponível em: <http://audiodescriptionfrance.wordpress.com/acteurs/> Acesso em 29 set. 2012.
[17] Disponível em: <http://www.enaparte.org/audiodescription/La_Charte_files/Historique-charte.html>. Acesso em 30 jul.2012.

pesquisa, por Frédéric Gonant e Laure Morisset (cofundadores da associação *En Aparté*), que contaram com a colaboração de: Maryvonne Simoneau – discípula de Auguste Coppola e pioneira, em 1989, com seus pares Marie-Luce Plumauzille e Jean-Yves Simoneau, da AD na França; Patrick Gohet, da *Délégation Interministériel aux Personnes Handicapées* (*DIPH*)*;* Gilbert Montagné da *Société Civile des Auteurs Multimédia* (*SCAM*)*;* e Michel Boyon do *Conseil Supérieur d'Audiovisuel* (*CSA*)*,* entre outros. Vale relembrar que o texto francês concebe a AD como um trabalho de autor e de criação, que consiste na concepção de um texto inédito a partir de um suporte visual; contudo, essa concepção é contradita quando o mesmo documento recomenda que o audiodescritor não deve interpretar o que vê, significando que a *Charte* é ambígua quanto ao parâmetro de neutralidade.

Concernente a essa questão, algumas pesquisas representam o estado da arte. Jiménez Hurtado (2007) estudou roteiros de AD de filmes redigidos em espanhol quanto, dentre outros aspectos, ao parâmetro da neutralidade quanto a somente a avaliações em torno dos sentimentos emotivos (examinou apenas orações com verbos de ligação e predicativo do sujeito), tendo chegado ao resultado de que não são neutros desse ponto de vista. Em seu ensaio, Holland (2009), com base exclusivamente em suas impressões a partir de sua experiência como audiodescritor profissional na Inglaterra, argumentou a favor da impossibilidade de neutralidade em roteiros de AD para o teatro e as artes visuais em geral; além disso, relatou o resultado de uma pesquisa de recepção de pequena escala em que comparou a preferência de DVs entre dois roteiros de AD de uma pintura escritos em inglês: um com muito pouca e o outro com muita avaliação. O preferido foi o segundo!

Praxedes Filho e Magalhães (2013a,b)[18], na pesquisa 'A audiodescrição de pinturas é neutra? um estudo descritivo via teoria da avaliatividade', descreveram roteiros de AD de pinturas quanto a presença ou ausência de neutralidade e o fizeram sob a perspectiva pragmático-funcionalista da Teoria da Avaliatividade (TA) no escopo da Linguística Sistêmico-Funcional (LSF). Nesse estudo pioneiro, os pesquisadores demonstraram que seis roteiros de AD de pinturas em inglês e seis em português, mesmo tendo sido escritos por audiodescritores americanos e brasileiros treinados segundo a prescrição da neutralidade obrigatória, são interpretativos dos pontos de vista das três grandes áreas de significados avaliativos da TA ('atitude', 'engajamento' e 'gradação'). Os resultados quantitativos evidenciaram, então, a ausência de neutralidade nos roteiros em ambas as línguas e, face aos

[18] Praxedes Filho e Magalhães (2013a) apresentam os resultados parciais em relação aos resultados globais relatados em Praxedes Filho e Magalhães (2013b), tendo, portanto, ambos trabalhos derivado do mesmo estudo.

mesmos, os autores concluem que "parece haver um padrão avaliativo caracterizado pela predominância de avaliações/interpretações atitudinais em termos de apreciações estéticas bem como de avaliações/interpretações de gradação em termos da força com a qual as apreciações são expressas" (PRAXEDES FILHO; MAGALHÃES, 2013b, p. 59), dado que os termos/escolhas no segundo nível de delicadeza assim ranquearam para os *corpora* em português e inglês, respectivamente: **'força'** > **'apreciação'** > 'foco' > 'heteroglossia' > 'monoglossia' > 'afeto' > 'julgamento' e **'força'** > **'apreciação'** > 'heteroglossia' > 'foco' > 'afeto' > 'julgamento' > 'monoglossia'.

Araújo e Aderaldo (2013) reuniram estudos desenvolvidos nas Universidades Federal da Bahia (UFBA), de Minas Gerais (UFMG) e na Universidade Estadual do Ceará (UECE), que demonstram a situação acadêmico-científica atual da modalidade AD da TAV-Ac brasileira. Os trabalhos tratam da elaboração de roteiros, da locução e da recepção de ADs realizadas para o teatro, obras de arte, cinema e televisão. Contudo, nessas instituições de ensino, nenhum estudo, exceto o de Praxedes Filho e Magalhães (2013a,b), foi realizado para fornecer informações empíricas acerca da questão do parâmetro prescritivo de neutralidade sob a luz da TA-LSF. É desta interface que passamos a tratar.

2.4 Teoria da Avaliatividade-LSF

Pesquisas embasadas na Linguística Sistêmico-Funcional (LSF) ganham, conforme Vian Jr. *et al* (2010), significativa expansão no Brasil com a inclusão de "diversos tipos de texto e contextos, estabelecendo diálogos com diferentes disciplinas e ampliando muitos aspectos da teoria em diferentes campos" (p. 11), expandindo as possibilidades de pesquisa em Linguística e Linguística Aplicada através da proposta teórico-metodológica de Michael Halliday. Os autores explicam, ainda, que

> um desses aspectos é o Sistema de Avaliatividade, um conjunto de significados interpessoais que se debruça sobre os mecanismos de avaliação veiculados pela linguagem [verbal], configurados em um sistema que oferece aos usuários possibilidades de utilizar itens avaliativos em suas interações cotidianas. (VIAN JR. *ET AL*, 2010, p. 11)

A respeito da abrangência da Teoria hallidayana, Praxedes Filho e Magalhães (2013b) observam que

> a LSF – dado seu viés funcionalista em contraposição ao viés formalista –, não se limita a estudar a língua apenas do ponto de vista intralinguístico do

> significado (semântica), da forma (lexicogramática) e da expressão (fonologia e fonética-grafologia e grafética). Antes de chegar aos estratos intralinguísticos, a teoria hallidayana parte do estrato ainda extralinguístico dos contextos de cultura e de situação (social) (HALLIDAY; MATTHIESSEN, 2004). (p. 15)

A LSF (HALLIDAY, 1985, 1994; HALLIDAY; MATTHIESSEN, 2004) tem por base a concepção de língua enquanto fenômeno principalmente social, levando em conta, portanto, a relação de interdependência entre os contextos de cultura e de situação (imediata social) e a língua. Para a LSF, as variáveis (i) 'campo' (atividade social, objetivo comunicativo e assunto), (ii) 'relações' (participantes, seus papeis e as relações entre eles quanto ao poder, ao afeto e à duração) e (iii) 'modo' (papel da língua, canal, meio e modo retórico) do segundo contexto, o de situação, ativam (ou são realizadas), no estrato da semântica, respectivamente, (i) os (pelos) significados ideacionais (experienciais-representação subjetiva das experiências cotidianas e lógicos-sequenciamento das experiências), (ii) os (pelos) significados interpessoais (negociação e avaliação) e (iii) os (pelos) significados textuais (construção de textos coesos e coerentes); essas áreas de significados constituem-se nas funções universais da linguagem verbal ou metafunções. Cada tipo de significado, por sua vez, constrói (realiza) cada variável do contexto de situação, as quais definem o registro (tipo de texto) correspondente ao tipo de situação social do qual dado contexto de situação é uma instância, sendo o contexto de situação, por seu turno, o lócus onde um dado texto oral, escrito ou sinalizado é produzido como instância do registro[19]. Esses significados, por sua vez, ativam (ou são realizados), no estrato da lexicogramática, respectivamente, as (pelas) áreas formais (i) de transitividade + relações táticas e lógico-semânticas, (ii) de modo + recursos lexicogramaticais de avaliatividade (incluem as modalidades de modalização e modulação) e (iii) de tema e informação; dialeticamente, cada área formal constrói (realiza) cada tipo de significado. Halliday (1985) explica que as metafunções se interligam na construção de textos, sejam orais, escritos ou sinalizados.

Ancorada na LSF (HALLIDAY, 1985, 1994; HALLIDAY; MATTHIESSEN, 2004), a Teoria da Avaliatividade (TA) (MARTIN; WHITE, 2005) trata dos significados avaliativos na língua. Esses significados existem como recursos à nossa disposição para que façamos escolhas, as quais são realizadas lexicogramaticalmente. Se são recursos a serem escolhidos,

[19] Os textos que compõem o *corpus* deste trabalho instanciam o registro 'roteiro de AD de filmes de longa-metragem'.

são sistematizados em uma rede de sistemas ou paradigmas de significados avaliativos (rede de sistemas de avaliatividade). Se são de significados que a rede de sistemas de avaliatividade se constitui, ela situa-se no estrato da semântica e, nele, insere-se na área da língua responsável pelo estabelecimento das relações interpessoais, ou seja, insere-se no âmbito dos significados interpessoais ou da metafunção interpessoal porque, quando avaliamos / interpretamos / posicionamo-nos, construímos nossas identidades, o que só pode ser feito na relação com o outro.

Martin e White (2005) demonstram que a rede de sistemas de avaliatividade se organiza, inicialmente, em torno do sistema TIPOS DE AVALIATIVIDADE, cujos termos/escolhas, os mais gerais ou menos delicados[20], são 'atitude', 'engajamento' e/ou 'gradação'. A partir dessas primeiras escolhas, a rede de sistemas de significados avaliativos se expande através de outros sistemas cujos termos/escolhas vão se refinando em até seis níveis de delicadeza ou especificidade. Relativo a esse entendimento, Praxedes Filho e Magalhães (2013b) esclarecem que

> [u]ma rede de sistemas é um conjunto de sistemas inter-relacionados, cuja organização relacional se dá através dos níveis de delicadeza da escala de delicadeza ou refinamento/detalhamento. Um sistema, por sua vez, é um conjunto de termos mutuamente excludentes/não-simultâneos ou simultâneos dentre os quais o falante/escritor faz escolhas. Cada rede de sistemas tem uma condição de entrada inicial que estabelece seu ambiente/escopo e enseja que sejam feitas as escolhas dentre os termos dos sistemas no primeiro nível de delicadeza. (...) Cada termo escolhido em um dado sistema pertencente a um dado nível de delicadeza passa a ser condição de entrada a outro sistema à direita, pertencente ao nível de delicadeza subsequente. Foi convencionado que, enquanto os nomes de sistemas devem ser grafados em letras maiúsculas, os nomes dos termos de um sistema devem ser grafados em letras minúsculas e, quando aparecem em textos verbais, devem ser acrescentadas aspas simples [(MATTHIESSEN, 1995, p. 749-754)]. Foi também convencionado que termos ou sistemas que podem ser escolhidos simultaneamente devem ser envolvidos por chaves, enquanto termos ou sistemas que são necessariamente mutuamente excludentes devem ser envolvidos por colchetes. (p. 17)

Mais adiante, os mesmos autores continuam, informando que

> [a] condição de entrada [inicial] 'avaliatividade' possibilita a entrada no sistema de primeiro nível de delicadeza, chamado TIPOS DE AVALIATIVIDADE, cujos termos são 'atitude' e/ou 'engajamento' e/ou 'gradação'. Os termos 'atitude', 'engajamento' e 'gradação', quando

[20] Seguindo Figueredo (2011), essa é nossa tradução do termo '*delicate*' da LSF. Seguindo o mesmo autor, traduzimos '*delicacy*' por 'delicadeza'.

> escolhidos, passam a ser novas condições de entrada a sistemas mais refinados à direita ou sistemas de segundo nível de delicadeza: TIPOS DE ATITUDE, TIPOS DE ENGAJAMENTO e TIPOS DE GRADAÇÃO, respectivamente. Os termos do sistema TIPOS DE ATITUDE são 'afeto' e/ou 'julgamento' e/ou 'apreciação'. Quanto ao sistema TIPOS DE ENGAJAMENTO, seus termos são 'monoglossia' ou 'heteroglossia'. Para o sistema TIPOS DE GRADAÇÃO, seus termos são 'força' e/ou 'foco'. (PRAXEDES FILHO; MAGALHÃES, 2013a, p. 76)

Os autores – em conformidade com Martin e White (2005), Martin e Rose (2007), Navarro (2012), Macken-Horarik (2004) e Bednarek (2008; 2010) –, elaboraram a representação gráfica da rede de sistemas de avaliatividade até o segundo nível de delicadeza apresentada na Figura 1.

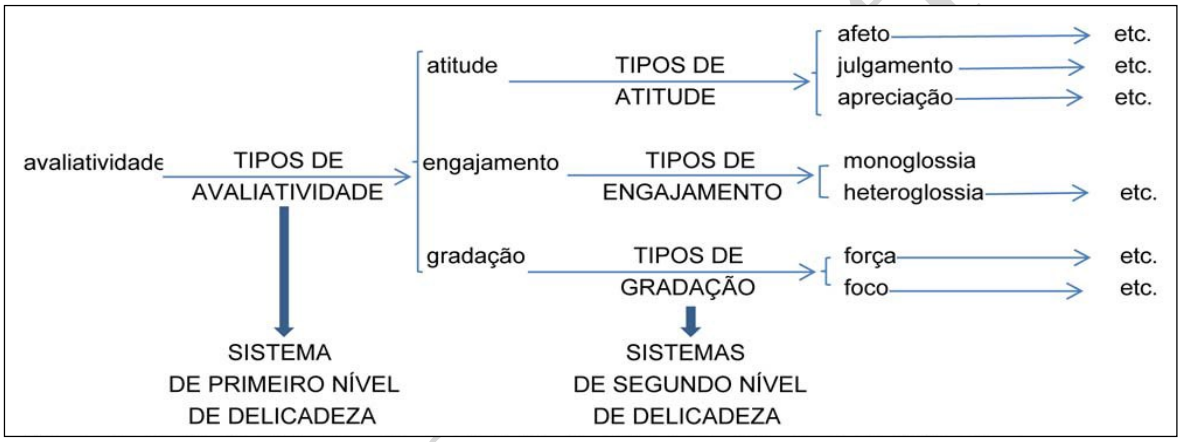


Figura 1 – Rede de sistemas de avaliatividade até o segundo nível de delicadeza.
Fonte: Praxedes Filho e Magalhães (2013b, p. 26)21

A respeito dos termos dos sistemas que abrangem os dois primeiros níveis de delicadeza, tal como ilustra a Figura 1, passamos a discorrer nas subseções que seguem.

2.4.1 'atitude'

É o termo/escolha de significado interpessoal avaliativo no âmbito do sistema TIPOS DE AVALIATIVIDADE, de primeiro nível de delicadeza, ligado aos sentimentos emotivos, éticos e estéticos do falante, do escritor ou de terceiros. O sistema TIPOS DE ATITUDE, em segundo

[21] Para alertar os leitores não iniciados em LSF, informamos que, na Figura 1, não constituem a rede de sistemas os seguintes elementos: 1) as setas que levam aos grupos nominais 'SISTEMA DE PRIMEIRO NÍVEL DE DELICADEZA' e 'SISTEMAS DE SEGUNDO NÍVEL DE DELICADEZA bem como os grupos nominais propriamente ditos e 2) as seis ocorrências de 'etc.'. Mantivemos os acréscimos de Praxedes Filho e Magalhães (2013b, p. 26) tendo em vista tornar a representação gráfica da rede mais clara. Cada 'etc.' significa que, se escolhidos os termos 'afeto', 'julgamento', 'apreciação', 'heteroglossia', 'força' e/ou 'foco', outros termos estão disponíveis para escolha até seis níveis de delicadeza.

nível de delicadeza, desdobra-se, por sua vez, nos termos/escolhas 'afeto' e/ou 'julgamento' e/ou 'apreciação', os quais são simultâneos, o que é representado, na Figura 1, pela chave. De acordo com Praxedes Filho e Magalhães (2013a), o tipo de atitude 'afeto' evoca a "área emotiva dos sentimentos; diz respeito a avaliações sobre as emoções das pessoas (...)", o tipo de atitude 'julgamento' evoca a "área ética dos sentimentos; tem a ver com avaliações sobre o comportamento das pessoas (...)" e o tipo de atitude 'apreciação' evoca a "área estética dos sentimentos; contempla avaliações sobre o aspecto estético das coisas e dos fenômenos, tanto os semióticos quanto os naturais (...)" (p. 77).

O 'afeto' envolve sentimentos domésticos do dia-a-dia ou do senso comum e contempla o registro de estados emocionais que são experienciados. Sentimos 'felicidade' ou tristeza (emoções intimistas, ligadas aos assuntos do coração), 'segurança' ou ansiedade (emoções ligadas ao bem estar ecossocial), 'satisfação' ou frustração (emoções ligadas ao *telos*/a consecução de objetivos).

O 'julgamento' abrange sentimentos institucionalizados (que saem do senso comum) relativos aos "valores comunitários compartilhados" (MARTIN; WHITE, 2005, p. 45) e tem a ver com avaliações que fazemos a respeito do caráter e do comportamento das pessoas: atitudes que 'estimamos' socialmente ou criticamos, elogiamos ou reprovamos/condenamos; valores que 'sancionam' socialmente o indivíduo perante a lei ou a religião.

A 'apreciação' engloba também sentimentos institucionalizados e diz respeito às avaliações que envolvem a estética de pessoas, objetos, coisas e fenômenos em geral (semióticos ou naturais): revelam posicionamentos sobre como tais entidades e fenômenos são valorizados ou não. Ainda nesse sentido, vale salientar que os aspectos estéticos apreciados podem ser manifestados pelas avaliações quanto a 'reação' que provocam, a 'composição' que apresentam e o 'valor social' que possuem.

Em relação de simultaneidade ao sistema TIPOS DE ATITUDE, há os sistemas de POLARIDADE e TIPOS DE REALIZAÇÃO DE ATITUDE. Pelo sistema de POLARIDADE, cada escolha avaliativa de 'atitude' pode ser 'positiva' ou 'negativa' ou, seguindo Bednarek (2008; 2010), 'ambígua' (não são explicitamente positivas ou negativas). Quanto aos TIPOS DE REALIZAÇÃO DE ATITUDE, cada escolha avaliativa de 'atitude' pode ser 'inscrita' ou 'evocada'. A primeira diz respeito à avaliação atitudinal explícita por meio de itens lexicais ou estruturas inscritas no texto. A segunda tem a ver com avaliação atitudinal implícita no texto, podendo ser: 'provocada' por metáforas lexicais ou 'convidada/sinalizada' por avaliações de gradação e outros meios ou 'convidada/propiciada' pelo conteúdo ideacional-experiencial dos enunciados.

2.4.2 'engajamento'

É o termo/escolha de significado interpessoal avaliativo no âmbito do sistema TIPOS DE AVALIATIVIDADE, de primeiro nível de delicadeza, relacionado à forma pela qual falantes e escritores assumem alguma postura em relação ao que dizem em seus textos e em relação ao que outros dizem sobre os mesmos assuntos. O sistema TIPOS DE ENGAJAMENTO, em segundo nível de delicadeza, desdobra-se, por sua vez, nos termos/escolhas 'monoglossia' ou 'heteroglossia', os quais são mutuamente excludentes, o que é representado, na Figura 1, pelo colchete. Segundo Praxedes Filho e Magalhães (2013a), a 'monoglossia' "tem a ver com asserções categóricas que não permitem o questionamento ou que não dão margem à dialogia (...)" e a 'heteroglossia' "tem a ver com o reconhecimento, por parte do falante/escritor, de que existem outras vozes ou pontos de vista acerca do assunto que está tratando (...)" (p. 78). Os autores (2013b) consideram que "é através da heteroglossia que o falante/escritor traz não só os seus próprios juízos de valor – mas também os de outros, alinhando-os ou desalinhando-os com os deles e negociando com o ouvinte/leitor uma relação de solidariedade ou não (...)" (p. 45).

A respeito do termo/escolha 'monoglossia', Praxedes Filho e Magalhães (2013b), levando em conta a especificidade do registro geral 'roteiro de AD' a partir do que evidenciou a análise dos dados, passaram a defender que o engajamento monoglóssico ocorre por desvios ou inferências descritivas categóricas. Tratando dessas duas situações, os pesquisadores esclarecem:

> 1) descrição não-modalizada de dado aspecto de uma pintura em desacordo com o referido aspecto tal como aparece na pintura (desvio descritivo categórico) e 2) descrição não-modalizada de dado aspecto de uma pintura por extrapolação da caracterização do referido aspecto tal como o pintor a construiu (inferência descritiva categórica). (p. 42)

No presente estudo, incorporamos as definições do termo/escolha 'monoglossia' tal como propostas por Praxedes Filho e Magalhães (2013b).

2.4.3 'gradação'

É o termo/escolha de significado interpessoal avaliativo no âmbito do sistema TIPOS DE AVALIATIVIDADE, de primeiro nível de delicadeza, relativo à regulação, para mais ou menos, do grau das avaliações de atitude e engajamento. O sistema TIPOS DE GRADAÇÃO, em segundo nível de delicadeza, desdobra-se, por sua vez, nos termos/escolhas 'força' e/ou 'foco', os quais são

simultâneos, o que é representado, na Figura 1, pela chave. Para Praxedes Filho e Magalhães (2013a), pelo tipo de gradação 'força', "o falante/escritor ajusta as avaliações quanto à sua 'quantidade' (...) ou 'intensidade' (...)" e, pelo tipo de gradação 'foco', "o falante/escritor ajusta as avaliações quanto à sua 'prototipicalidade' (...) ou a precisão pela qual as fronteiras de uma categoria são definidas (...)" (p. 78). A direção da gradação de 'força' e 'foco' pode ser 'aumentando' ou 'diminuindo'.

Quando a escolha é pelo termo 'força', a avaliação pode envolver (i) 'intensidade', podendo dizer respeito a 'qualidades' (grupos adjetivais, grupos adverbiais e modalidades) e 'processos' (grupos verbais), e (ii) 'quantidade', podendo dizer respeito a valores numéricos imprecisos de entidades, valores imprecisos da massa ou presença de entidades e a valores imprecisos do espalhamento de entidades no tempo e no espaço. Ainda sobre as escolhas relativas à intensidade e quantidade, para Martin e White (2005), elas são "(...) categorias que envolvem avaliações inerentemente escalares, como por exemplo as avaliações atitudinais ... [graduáveis ao longo de um contínuo positividade-negatividade], mas também as avaliações de tamanho, vigor, extensão, proximidade (...)"[22] (p. 137).

A realização do tipo de gradação 'força' pode ser: (i) 'isolada' – em que a gradação "(...) é realizada por um item isolado, individual (...)"[23] (MARTIN; WHITE, 2005, p. 141), como os intensificadores (*muito* feliz) –, ou (ii) 'fusionada', em que a realização da gradação encontra-se fundida em uma dada palavra e só se explicita na comparação paradigmática com outras palavras relacionadas semanticamente (gostar, amar, adorar).

Conforme Martin e White (2005), a escolha pelo termo 'foco' permite ao falante/escritor fazer uma avaliação que "(...) atua na medida em que os fenômenos são graduados em relação a até que ponto eles se enquadram no centro de uma categoria semântica ou se assemelham a uma instância exemplar dessa categoria"[24] (p. 137).

As subredes da rede de sistemas de avaliatividade, descritas nas Subseções 2.4.1, 2.4.2 e 2.4.3, forneceram categorias analíticas somente até o segundo nível de delicadeza. Essa delimitação de abrangência atende à necessidade que se impõe para a consecução dos objetivos desta pesquisa. Destarte, a TA permite verificar a (in)existência de neutralidade por parte do audiodescritor dos

[22]"(…) categories which involve inherently scalar assessments – for example the attitudinal assessments … [gradable along clines of positivity/negativity] but also assessments of size, vigour, extent, proximity, (…)".
[23] "(...) is realised by an isolated, individual item (...)."
[24] "(…) operates as phenomena are scaled by reference to the degree to which they match some supposed core or exemplary instance of a semantic category."

pontos de vista de suas avaliações atitudinais, de seu engajamento com sua própria voz e com outras vozes avaliativas e de como ele gradua suas atitudes e seus posicionamentos de engajamento.

Passamos a descrever o desenho metodológica adotada no estudo.

3. Metodologia

Considerando os objetivos e o percurso teórico, esta pesquisa apresenta-se como exploratória, descritiva e quanti-qualitativa, visando analisar, descrever e discutir a neutralidade em roteiros autênticos de ADs francesas de filmes, sem manipulá-los. Trata-se também de um estudo de caso por contemplar somente uma profissional da AD francesa e apenas dois roteiros de sua autoria. Outra característica importante diz respeito ao caráter interdisciplinar posto em relevo pelas interfaces da pesquisa e que segue a seguinte trajetória: Estudos da Tradução ↔ Estudos Descritivos da Tradução ↔ Estudos da Tradução Baseados em *Corpus* ↔ Linguística de *Corpus* ↔ Tradução Audiovisual ↔ Tradução Audiovisual Acessível ↔ Audiodescrição ↔ Linguística Sistêmico-Funcional ↔ Teoria da Avaliatividade.

O *corpus* foi constituído por textos que instanciam o registro 'roteiro de AD de filmes de longa-metragem'. Os filmes, do gênero fílmico comédia, são, tal como já mencionado, *Intouchables* (de Eric Toledano e Olivier Nakache) e *Minuit à Paris* (de Woody Allen), audiodescritos em 2011 e disponibilizados no mercado francês. A seleção dos mesmos ocorreu por consultas na internet e, no propósito de agilizar o estudo, solicitamos à audiodescritora os roteiros autênticos, os quais nos foram prontamente disponibilizados pelo envio de dois arquivos no formato *.pdf*. No sentido de facilitar a construção das etiquetas, decidimos alocar os filmes em dois *subcorpora*: *Subcorpus* 1 ou C1 com o roteiro de AD de *Intouchables* e *Subcorpus* 2 ou C2 com o roteiro de AD de *Minuit à Paris*.

Ressaltamos que, na interface com a LC, foram adotados procedimentos de tratamento de *corpus* e de análise de dados via o *software Wordsmith Tools 5.0* de Mike Scott. Assim, no que se refere à descrição geral de cada *subcorpus*, obtida por meio das ferramentas desse programa, destacamos aqui, que C1 apresenta 6.948 palavras corridas (*tokens*), 1.204 palavras distintas (*types*) e 33,08% de variedade lexical padronizada (*standardised*) e C2 possui 3.439 palavras corridas, 813 palavras distintas e 36,10% de variedade lexical padronizada.

Fica claro, pelo exposto até aqui sobre os *subcorpora*, que, com base nas características apontadas por Berber Sardinha (2004) e Saldanha (2009), eles caracterizam-se por sua origem autêntica e por sua extensão pequena mas representativa porque atendem ao

propósito deste estudo, que trata roteiros de AD linguisticamente para descrevê-los apenas quanto à (in)existência de neutralidade e não quanto a seus padrões avaliativos. Apesar de ter cerca da metade da extensão de C1, C2 parece ser tão representativo quanto C1 pelo fato de ambos apresentarem percentuais próximos de variedade lexical padronizada.

Os procedimentos metodológicos foram os seguintes:

- Concepção das etiquetas de análise, visando a identificação de cada *subcorpus* e das categorias (termos/escolhas) da rede de sistemas de avaliatividade até o 2º nível de delicadeza. Nesta fase, foram concebidas sete etiquetas que apresentam as iniciais C1, relativas ao *Subcorpus* 1 (filme *Intouchables*), como mostra o Quadro 1. Outras sete foram confeccionadas, distintas das demais somente por apresentarem as iniciais C2, indicativas do *Subcorpus* 2 (filme *Minuit à Paris*). As etiquetas foram acompanhadas de parênteses angulares (< >) e apenas uma foi definida para marcar o término de cada ocorrência: </t>.

Quadro 1 – Etiquetas do *Subcorpus* 1 (C1) – Identificação das categorias da rede de sistemas de avaliatividade até o segundo nível de delicadeza.

1. <C1_ATIT_AFETO>	identifica ATITUDE afeto em C1
2. <C1_ATIT_JULG>	identifica ATITUDE julgamento em C1
3. <C1_ATIT_APREC>	identifica ATITUDE apreciação em C1
4. <C1_ENG_MONOGL>	identifica ENGAJAMENTO monoglossia em C1
5. <C1_ENG_ HGL>	identifica ENGAJAMENTO heteroglossia em C1
6. <C1_GRAD_FOCO>	identifica GRADAÇÃO foco em C1
7. <C1_GRAD_FORÇA>	identifica GRADAÇÃO força em C1

Fonte: os autores.

- Compilação dos *subcorpora* para leitura pelo *software Wordsmith Tools 5.0*. Nesta etapa, os conteúdos dos arquivos recebidos em *.pdf* foram inicialmente selecionados, copiados e colados no programa *Word* e, em seguida, salvos em *.doc*. Escritos para serem lidos em voz alta, os roteiros das ADs possuem marcações (como o tempo da entrada e saída da AD, dentre outras) que foram identificadas e postas entre parênteses angulares (< >) com o propósito de não serem interpretadas pelo *software*. Em seguida, foram criados outros dois arquivos no formato *.txt* do 'Bloco de Notas', a partir dos arquivos *.doc*, a fim de serem lidos pelo *Wordsmith Tools 5.0*.

- Análise quanto à identificação das categorias e a inserção das etiquetas nos *subcorpora*. Para a realização desta fase, foram criados, a partir dos dois arquivos no formato *.txt*, já

compilados, 14 arquivos, correspondentes às etiquetas criadas para os *subcorpora*, sendo especificamente: sete arquivos *.txt* para C1 e sete arquivos *.txt* para C2. Essa foi uma medida preventiva a fim de evitar que, dentro de um mesmo arquivo *.txt*, várias etiquetas fossem, em um mesmo trecho de texto, inseridas lado a lado. Essa medida estratégica foi motivada pelo fato de que a abordagem teórico-analítica adotada é fundamentada em termos/escolhas de sistemas em uma rede de sistemas que podem ou não ocorrer simultaneamente e, portanto, incidirem ou não em um mesmo trecho de texto. Na sequência, após a identificação das categorias avaliativas até o 'segundo nível de delicadeza', as etiquetas foram inseridas no início do trecho de ocorrência de dada categoria avaliativa e a etiqueta </t> marcou o fechamento da ocorrência.

- Revisão da análise e extração de dados quantitativos via *software*. Esses procedimentos dizem respeito à exploração de C1 e C2 por meio das ferramentas integradas ao *Wordsmith Tools 5.0*, a *WordList* e a *Concord*. Nesse propósito, os *subcorpora* no formato *.txt* foram inicialmente lidos pela *WordList*, que disponibilizou a descrição de características específicas dos mesmos, fornecendo dados como: o número de palavras corridas, de palavras distintas e o percentual da variedade lexical padronizada. Em seguida, a revisão da análise/etiquetagem foi realizada via *Concord*, que permitiu a exibição e consulta de todas as ocorrências das etiquetas, acompanhadas pelo cotexto (em ordem alfabética) no qual elas foram inseridas e, ainda, a visualização das mesmas no interior de cada *subcorpus*. Esses passos possibilitaram a revisão da análise pela confirmação (ou não) das categorias etiquetadas e pela verificação de suas localizações. O concordanciador executou a contagem de todas as ocorrências das etiquetas em C1 e C2, forneceu listas das mesmas, e dados em números absolutos, os quais receberam tratamento estatístico básico.
- Tratamento estatístico básico dos dados em números absolutos visando a comparabilidade dos resultados entre *subcorpora*. Considerando que cada *subcorpus* apresenta dimensão diferente, ou seja, possui um número total desigual de palavras corridas e considerando que quanto mais extenso um *corpus*, maior a probabilidade de ocorrência de dada categoria linguística, o Índice de Frequência Simples (IFS) foi recurso estatístico adotado para o controle dessa variável. Um IFS corresponde ao número de ocorrências de uma categoria por cada 1.000 palavras de texto. Para se chegar a um IFS, como exemplifica o Quadro 2, o número ou valor absoluto (VA) de ocorrências de uma etiqueta é dividido pelo VA total de palavras corridas do *subcorpus* e o resultado é multiplicado por 1.000. Esse procedimento foi executado utilizando os dados fornecidos pela *WordList* e o *Concord*. Em seguida, os IFS(s) foram transformados em percentuais.

Quadro 2 – Exemplo de cálculo dos IFS(s) dos dados quantitativos extraídos via *WordList* e *Concord* do *Wordsmith Tools 5.0*.

IFS = VA da Etiqueta 3 ÷ VA total de palavras corridas de C1 X 1000
ou
IFS = 360 ÷ 6.948 X 1000 = 51,8

Fonte: os autores.

Pelo exemplo, o C1 apresenta 51,8 ocorrências da Etiqueta 3 (<C1_ATIT_APREC>) por cada 1.000 palavras corridas de texto. Se o mesmo VA ocorresse em C2, ele representaria valor em torno do dobro pelo fato de C2 ser cerca de metade de C1. Portanto, o IFS para um VA de 360 em C2 seria 104,6.

A seguir, apresentamos os resultados e os discutimos.

4. Resultados e Discussão

A Tabela 1 mostra os VA(s) e os resultados quantitativos finais em IFS(s) e percentuais relativos às 2.422 ocorrências de posicionamentos avaliativos/interpretativos encontradas em ambos os *subcorpora*.

Tabela 1 - Dados em VA, IFS e percentual (%) – resultados finais das ocorrências avaliativas nos dois primeiros níveis de delicadeza.

1º NÍVEL DE DELICADEZA – TIPOS DE AVALIATIVIDADE

SUBCORPUS	'atitude'			'engajamento'			'gradação'		
	VA	IFS	%	VA	IFS	%	VA	IFS	%
C1	739	106,3	46,7	145	20,8	9,1	699	100,6	44,1
C2	370	107,6	44,1	64	18,6	7,6	405	117,7	48,2

2º NÍVEL DE DELICADEZA

SUBCORPUS	TIPOS DE ATITUDE									TIPOS DE ENGAJAMENTO						TIPOS DE GRADAÇÃO					
	'afeto'			'julgamento'			'apreciação'			'monoglossia'			'heteroglossia'			'força'			'foco'		
	VA	IFS	%	VA	IFS	%	VA	IFS	%	VA	IFS	%	VA	IFS	%	VA	IFS	%	VA	IFS	%
C1	242	34,8	15,3	137	19,7	8,7	360	51,8	22,7	110	15,8	6,9	35	5,0	2,2	624	89,8	39,4	75	10,8	4,7
C2	109	31,7	13,0	78	22,7	9,3	183	53,2	21,8	48	13,9	5,7	16	4,7	1,9	366	106,4	43,6	39	11,3	4,6

Fonte: os autores

A Tabela revela que os percentuais de ocorrência em C1 e C2, em todos os tipos de avaliatividade, não diferem muito. Destacamos que a semelhança entre *subcorpora* de tamanhos díspares é evento analítico coerente ao considerarmos a proximidade dos percentuais da variedade lexical padronizada, dado que a diferença, entre eles, é de apenas 3,02 pontos percentuais. Em

conformidade com os dados quantitativos finais encontrados com a abordagem teórico-analítica adotada, passamos a detalhar os resultados por pergunta de pesquisa.

4.1 Primeira Pergunta

Na primeira pergunta de pesquisa, questionamos como se caracterizam os roteiros das ADs de C1 e C2 quanto à presença ou ausência de neutralidade operacionalizada pela ausência ou presença, respectivamente, de marcas de posicionamento avaliativo/interpretativo de 'atitude', quanto aos TIPOS DE ATITUDE 'afeto', 'julgamento' e/ou 'apreciação'.

De acordo com a Tabela 1, C1 e C2 caracterizam-se pela presença de 'atitude' (C1- 46,7%; C2- 44,1%), com ocorrências de seus três tipos. As ocorrências do tipo 'apreciação', referente aos sentimentos estéticos, ranquearam em 1ª posição, tanto em C1 (22,7%) quanto em C2 (21,8%), com uma diferença pró-C1 de 0,9 pontos. As ocorrências do tipo 'afeto', ligado aos sentimentos emotivos, ranquearam em 2º lugar em ambos os *subcorpora* (C1- 15,3%; C2- 13,0%), com uma diferença também pró-C1 de 2,3 pontos. Quanto às ocorrências do tipo 'julgamento', que envolvem sentimentos éticos, seu ranqueamento ocupa a 3ª posição relativamente aos dois *subcorpora* (C1- 8,7%; C2- 9,3%), com uma diferença de 0,6 pontos pró-C2. Excertos ilustrativos são apresentados no Quadro 3.

Quadro 3 - Exemplos de ocorrências avaliativas atitudinais do tipo 'afeto', 'julgamento' e 'apreciação' indicadas pelas etiquetas de C1 e C2

EXCERTOS [25]
N-234) sa barbe est broussailleuse</t>. Il est plus âgé, le <C1_ATIT_AFETO>**visage fermé**</t>. <(4')> Le conducteur regarde son
N-102) très haut au milieu d'un nuage de froufrous. Adriana tape dans ses mains. Elle <C2_ATIT_AFETO>**sourit à pleines dents**</t>. <(11:18:02)> Les filles
N-107) <(Bonsoir, messieurs.)> <(03:16:16 Enfin, surtout lui.)> Philippe <C1_ATIT_JULG>**rigole en silence**</t>. <///> Philippe et Driss sont au premier
N-77) <(10:35:15 Et là, tu peux bien faire l'amour de nouveau.)> Hemingway <C2_ATIT_JULG>**transperce Gil du regard**</t>. <(Réfléchis à ça.)>
N-238) sourit. <(Rire.)> <(03:16:05 Il est marrant, lui.)> <(Très vite.)> Une <C1_ATIT_APREC>**jolie**</t> ouvreuse attend. <(Bonsoir, messieurs.)>
N-27) dans les <C2_ATIT_APREC>**beaux**</t> quartiers, les pelouses

Fonte: Os autores.

[25] Extraídos por meio do concordanciador do *Wordsmith Tools 5.0*, os excertos exibem a indicação numérica (N) de cada etiqueta atribuída pelo programa e o cotexto no qual a mesma foi inserida.

No Quadro 3, as escolhas avaliativas atitudinais da tradutora/audiodescritora manifestam: **'afeto'** relativo à 'insatisfação' como em "*le visage fermé*" (a cara fechada) e quanto à 'felicidade' como em "*elle sourit à pleines dents*" (ela sorri de orelha a orelha); **'julgamento'** (do comportamento) relativo à 'estima social' como em "*rigole en silence*" (se diverte em silêncio) e relativo à sansão social como em "*Hemingway transperce Gil du regard*" (Hemingway olha Gil penetrantemente); e **'apreciação'** quanto à 'reação' (provocada pela entidade avaliada) como em "*une jolie ouvreuse attend*" (uma funcionária bonita aguarda) e "*dans les beaux quartiers, les pelouses...*" (nos lindos bairros, os gramados...).

4.2 Segunda Pergunta

Na segunda pergunta de pesquisa, a questão que levantamos procura saber como se caracterizam os roteiros das ADs de C1 e C2 quanto à presença ou ausência de neutralidade operacionalizada pela ausência ou presença, respectivamente, de marcas de posicionamento avaliativo/interpretativo de 'engajamento', quanto aos TIPOS DE ENGAJAMENTO 'monoglossia' ou 'heteroglossia'.

Face a Tabela 1, ambos os *subcorpora* caracterizam-se pela presença de 'engajamento' (C1- 9,1%; C2- 7,6%), tendo ocorrido seus dois tipos. As ocorrências do tipo 'monoglossia' – ligado à voz autoral manifestando-se ou via desvio descritivo categórico ou via inferência descritiva categórica (para o registro geral 'roteiro de AD' segundo Praxedes Filho e Magalhães, 2013b) –, apresentam percentual superior ao tipo 'heteroglossia' tanto em C1 (6,9%) quanto em C2 (5,7%), com uma diferença pró-C1 de 1,2 pontos. No tocante às ocorrências avaliativas por 'heteroglossia', que dizem respeito ao reconhecimento de vozes ou pontos de vista externos, os percentuais em C1 (2,2%) e C2 (1,9%) são muito semelhantes, com uma diferença de 0,3 pontos. Excertos ilustrativos dessas ocorrências seguem no Quadro 4.

Quadro 4 - Exemplos de ocorrências avaliativas por engajamento do tipo 'monoglossia' e 'heteroglossia' indicadas pelas etiquetas de C1 e C2.

EXCERTOS
N-7) Yvonne mange un mini-éclair</t> <C1_ENG_MONOGL>**au chocolat**</t>.
N-28) C'est <C2_ENG_MONOGL>**le déluge**</t> devant les cinémas
N-30) vous plaît... S'il vous plaît ! +1')> Le serveur arrive du fond de la salle <C1_ENG_HGL>**presque vide**</t>. Driss renverse son verre. <(Gling.)>
N-8) <(11:22:58 Parce que la vie est toujours un peu insatisfaisante.)> Adriana <C2_ENG_HGL>**ne sourit plus**</t>. <(C'est ça, le problème, avec les écrivains.

Fonte: Os autores.

No Quadro 4, as ocorrências avaliativas demonstram o 'engajamento' do tipo: **'monoglossia'** por inferência descritiva categórica como em "*un mini-éclair au chocolat*" (um docinho de chocolate), visto que a cena não dá nenhuma indicação sobre o sabor da guloseima, e por desvio descritivo categórico como em "*le déluge devant les cinémas*" (o dilúvio em frente aos cinemas), pelo fato de a voz da audiodescritora referir-se a uma chuva comum como se fosse torrencial; e **'heteroglossia'** como em "*presque vide*" (quase vazia) e "*ne sourit plus*" (não sorri mais), em que a 'contração' do espaço dialógico, no primeiro caso, é percebida pelo rompimento de expectativas ('contraexpectativa') e, no segundo caso, pela negativa.

4.3 Terceira Pergunta

Na terceira pergunta, levantamos a questão de como se caracterizam os roteiros das ADs de C1 e C2 quanto à presença ou ausência de neutralidade operacionalizada pela ausência ou presença, respectivamente, de marcas de posicionamento avaliativo/interpretativo de 'gradação', quanto aos TIPOS DE GRADAÇÃO 'força' e/ou 'foco'.

Conforme a Tabela 1, C1 e C2 caracterizam-se pela presença de 'gradação' (C1- 44,1%; C2- 48,2%), com ocorrências de seus dois tipos. As ocorrências dos posicionamentos do tipo 'força', que abrangem avaliações escalares relativas à 'quantificação' ou 'intensificação', evidenciam percentuais superiores tanto em C1 (39,4%) quanto em C2 (43,6%), com uma diferença pró-C2 de 4,2 pontos. Quanto às ocorrências do tipo 'foco' – que tratam do nível de prototipicalidade de entidades, processos e fenômenos –, o percentual em C1 (4,7%) é praticamente igual ao percentual em C2 (4,6%), com uma diferença pró-C1 de apenas 0,1 pontos. Excertos ilustrativos são exibidos no Quadro 5.

Quadro 5 - Exemplos de ocorrências avaliativas por gradação do tipo 'força' e 'foco' indicadas pelas etiquetas de C1 e C2.

EXCERTOS
N-227) <(Bon, on réessaye la casquette.)> Il porte un gilet gris et une veste bleue <C1_GRAD_FORÇA>**foncée**</t>. Driss lui met une casquette. <(Ça, c'est pas
N-14) Gil regarde des boucles d'oreille <C2_GRAD_FORÇA>**anciennes**</t> dans une vitrine.
N-50) Il a une petite <C1_GRAD_FOCO>**moustache à la Hitler**</t>
N-26) <//> La lumière est faible et <C2_GRAD_FOCO>**orangée**</t>.

Fonte: Os autores.

No Quadro 5, as ocorrências por 'gradação' indicam que a tradutora/audiodescritora ajustou suas escolhas avaliativas do tipo: **'força'** por intensificação como em "*foncée*"

(escuro) que incide sobre a qualidade "*bleue*" (azul) e por quantificação como em "*anciennes*" (antigos) que modifica o grupo nominal "*boucles d'oreille*" (brincos); e **'foco'** por referir-se a um exemplar genuíno ou prototípico de uma determinada categoria como em "*moustache à la Hitler*" (bigode do tipo Hitler) ou por falar de um exemplar que está na periferia de uma categoria como em "*orangée*" (alaranjada), que faz com que a "*lumière*" (luz) seja localizada na periferia de "orange" (laranja).

4.4 Quarta Pergunta

Aqui questionamos se os roteiros de C1 e C2 são neutros ou avaliativos/interpretativos e, se avaliativos/interpretativos, assim são de modo semelhante ou diferente.

As respostas às três primeiras perguntas indicam a presença de escolhas avaliativas nos roteiros de AD de C1 e C2 e demonstram, pois, que os mesmos são avaliativos/interpretativos em todas as áreas de significados avaliativos segundo os pressupostos da TA. Portanto, a neutralidade inexiste tanto em C1 quanto em C2.

Em vista dos dados exibidos na Tabela 1, o ranqueamento das escolhas avaliativas/interpretativas no 1º nível de delicadeza ficou: para C1 → 'atitude' (46,7%) > 'gradação' (44,1%) > 'engajamento' (9,1%); para C2 → 'gradação' (48,2%) > 'atitude' (44,1%) > 'engajamento' (7,6%).

Nesse primeiro nível, C1 e C2 são avaliativos/interpretativos de modo parcialmente semelhante. Contudo, a diferença não é profunda por duas razões: 1) o termo/escolha 'engajamento' ranqueou em terceiro lugar em C1 e C2; 2) as diferenças percentuais quanto às ocorrências dos termos/escolhas 'atitude' e 'gradação entre *subcorpora* são, respectivamente, de 2,6 pontos pró-C1 e 4,1 pontos pró-C2, as quais não representam diferenças grandes.

Ainda conforme os dados da Tabela 1, a sequência do ranqueamento das escolhas avaliativas/interpretativas no 2º nível de delicadeza foi: para C1 e C2 → 'força' > 'apreciação' > 'afeto' > 'julgamento' > 'monoglossia' > 'foco' > heteroglossia'.

Dado que depreende-se do segundo nível, por detalhar o primeiro, padrões mais reveladores do comportamento avaliativo da voz autoral, ousamos defender, com base nesse argumento, que C1 e C2 são avaliativos/interpretativos de modo semelhante, o que evidencia um forte indicativo da representatividade dos *subcorpora* relativa à prática tradutória da audiodescritora.

Com larga margem, os resultados encontrados indicaram maior ocorrência das escolhas avaliativas/interpretativas gradacionais de 'força' (C1- 39,4%; C2- 43,6%), que têm a ver com a

quantidade ou intensidade das avaliações atitudinais e de engajamento, seguidas das atitudinais de 'apreciação' (C1- 22,7%; C2- 21,8%), que veiculam sentimentos estéticos. Estamos tratando de ADs de filmes do gênero cinematográfico ficção e parece que a audiodescritora sentiu a necessidade de descrever avaliativamente a aparência de elementos da narrativa fílmica como personagens (atributos como aspecto físico e vestuário) e ambientações (cenários, adereços, iluminação, cores) (cf. JIMÉNEZ HURTADO et al., 2010), sem deixar de graduar, para mais ou para menos, as aparências descritas. Como há personagens, ela também parece ter se sentido impelida a falar de suas emoções e seus comportamentos e o fez de maneira avaliativa, pois as avaliações de 'afeto' (C1- 15,3%; C2- 13,0%) e 'julgamento' (C1- 8,7%; C2- 9,3%) são as seguintes no ranqueamento. Logo atrás, aparecem as ocorrências relativas a trechos avaliados por 'monoglossia' (C1- 6,9%; C2- 5,7%), significando o fato de a audiodescritora, nesses trechos, ter atingido o auge de sua capacidade interpretativa porque ou se desviou da cena descrita ou fez inferência sobre ela. O 'foco' – com percentuais de ocorrência de 4,7% e 4,6%, respectivamente –, certamente contribuiu, com menor presença, para graduar as aparências dos personagens e das ambientações. Por último, em ambos os *subcorpora*, ranqueou a avaliação por 'heteroglossia' (C1- 2,2%; C2- 1,9%), aspecto relativo ao qual os resultados ora reportados se diferenciam dos resultados reportados em Praxedes Filho e Magalhães (2013b), segundo os quais, para os *corpora* em português e inglês, o ranqueamento de 'heteroglossia' é o quarto e o terceiro, respectivamente (Ver Subseção 2.3). Em grande medida, as avaliações por 'engajamento' heteroglóssico ocorrem via modalizações e modulações, cujas realizações lexicogramaticais demandam estruturas com maior quantidade de palavras. Como os espaços destinados à inserção de AD em filmes – intervalos de silêncio entre uma fala e outra –, são geralmente muito pequenos, o que não ocorre relativamente a ADs de pinturas, a necessidade de mais palavras pode ser uma causa plausível para essa diferença entre os roteiros de AD de filmes e os roteiros de AD de pinturas.

Vale observar, ainda, que, em Praxedes Filho e Magalhães (2013a,b), os resultados mostraram que as duas maiores ocorrências avaliativas incidiram, tal como no presente estudo, em 'força' e 'apreciação'. Esses resultados convergentes podem apontar para a indicação de uma tendência de padrão avaliativo predominante relativa ao registro mais geral 'roteiro de AD'.

5. Considerações finais

O presente estudo dedicou-se a uma das modalidades da TAV-Ac, a que cuida da acessibilidade sociocultural junto aos DVs, e tratou do parâmetro de neutralidade quanto ao

registro 'roteiro de AD de filmes de longa-metragem' instanciado por dois filmes franceses, buscando investigar, com o auxílio da LC-ETBC, a presença ou ausência de interpretação por parte do tradutor/audiodescritor, segundo os fundamentos da TA-LSF.

Na interface com a LC, procedimentos de tratamento dos *subcorpora* e de análise dos dados puderam ser adotados, via s*oftware Wordsmith Tools 5.0*, que viabilizaram o bom êxito metodológico, tendo em vista tratar-se de estudo com números elevados: 10 termos/escolhas tomados da TA que geraram 14 etiquetas a serem inseridas em 10.387 palavras, o que levou a 2.422 etiquetagens. Em outras palavras, acreditamos que a LC garantiu que os objetivos fossem alcançados e as perguntas respondidas de modo satisfatório.

Pela abordagem teórico-analítica, amparada nos pressupostos da TA-LSF, estudamos os *subcorpora* na abrangência dos dois primeiros níveis de delicadeza da rede de sistemas de avaliatividade. Assim, evidenciamos e descrevemos as marcas de posicionamento da tradutora/audiodescritora dos pontos de vista de suas atitudes avaliativas, de seu engajamento com sua voz e com outras vozes avaliativas e de como ela gradua suas atitudes e seus posicionamentos de engajamento, o que viabilizou postularmos pela **in**existência de neutralidade nos roteiros. Nessa perspectiva, esta pesquisa contribui com os Estudos da Tradução, pois, ao descrever a existência de avaliação/interpretação em textos traduzidos, se alia aos EDT para investigar a neutralidade, ainda muitas vezes prescrita ao trabalho do tradutor/audiodescritor.

Os resultados demonstraram, em ambos os *subcorpora*, que há presença de posicionamentos avaliativos/interpretativos da perspectiva dos sete tipos de significados avaliativos no segundo nível de delicadeza. Dado que as ocorrências avaliativas/interpretativas gradacionais de 'força' e atitudinais de 'apreciação' ranquearam nas duas primeiras posições, tal como visto no estudo pioneiro de Praxedes Filho e Magalhães (2013a,b), essa 'coincidência' pode ser indicativa de um provável padrão de uso.

Em decorrência dessa convergência, deixamos a seguinte sugestão para futuras pesquisas: aprofundar a interface entre TAV-Ac/AD e LSF via TA e entre esta e os ETBC via LC, descrevendo um *corpus* de extensão média – que, para Berber Sardinha (2004), tem tamanho entre 250 mil e 1 milhão de palavras –, de roteiros de AD de filmes de longa-metragem elaborados pela mesma audiodescritora, o que deve ser possível por se tratar de profissional há cerca de 20 anos no mercado, com o objetivo de estudar seu padrão avaliativo/interpretativo ou sua a assinatura avaliativa, levando-se em conta os sistemas da rede de avaliatividade até o último nível de delicadeza. Outra sugestão diz respeito à compilação de um *corpus* composto por roteiros de AD de um mesmo filme de longa-

metragem escritos por diferentes audiodescritores franceses, com o objetivo de estudar descritivamente o padrão avaliativo/interpretativo do registro 'roteiro de AD de filmes de longa-metragem' ou seu estilo avaliativo.

Para além da relevância de suas interfaces, acreditamos que a contribuição maior deste estudo reside no fato de seus resultados – somados aos resultados de Jiménez Hurtado (2007), Holland (2009) e Praxedes Filho e Magalhães (2013a,b) –, libertarem os profissionais de AD do parâmetro prescritivo de neutralidade. Não duvidamos que esses resultados também impactarão a pedagogia de formação de audiodescritores. Visto que este estudo abordou questão relacionada à atividade do tradutor, esperamos, ainda, que desperte e estimule o interesse pela tradução, fomente reflexões, diálogos e pesquisas nas modalidades da TAV-Ac em interface com a LC, contribuindo com a acessibilidade sociocultural, a formação dos profissionais da área e com a valorização do trabalho do tradutor/audiodescritor.

Referências bibliográficas

ADERALDO, M. F. **Proposta de parâmetros descritivos para Audiodescrição à luz da interface revisitada entre Tradução Audiovisual Acessível e semiótica social – multimodalidade**. 2014. 206 f. Tese (Doutorado em Estudos Linguísticos). Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2014.

ARAÚJO, V. L. S.; ADERALDO, M. F. **Os novos rumos da pesquisa em Audiodescrição no Brasil.** Org. Araújo e Aderaldo. 1 ed. Editora CRV. Curitiba, PR. 2013, p. 8.

BAKER, M. Corpus-based translation studies: the challenges that lie ahead. In: Somers, H. (Ed.). **Terminology, LSP and translation:** studies in language engineering in honour of Juan C. Sager. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1996, p. 177-186. crossref http://dx.doi.org/10.1075/btl.18.17bak

BASSNETT, S. **Estudos de tradução**. Tradução de Vivina de Campus Figueiredo. Lisboa: Gulbenkian, 2003.

BEDNAREK, M. **Emotion talk across corpora**. Hampshire: Palgravre Macmillan, 2008. crossref http://dx.doi.org/10.1057/9780230285712

_____. **Glossary attitude** (brochura no minicurso 'Appraisal and Corpus Linguistics, ministrado no VI Congresso da Associação de Linguística Sistêmico- Funcional da América Latina, realizado na UECE, em Fortaleza-CE, de 05 a 09/10/2010). Manuscrito, 2010.

BENECKE, B. **Audio-Description**. Meta, vol. 49, nº 1, 2004, p. 78-80. crossref http://dx.doi.org/10.7202/009022ar

BERBER SARDINHA, T. **Lingüística de Corpus.** Barueri, SP: Manole, 2004.

BERNENGO, C. **L'audiodescription fait son cinéma:** état des lieux et perspectives. Mémoire de fin d'étude. França, 2012, jun. 166f. École Nationale Supérieure Louis Lumière. França. 2012. Disponível em : <http://www.ens-louis-lumiere.fr/fileadmin /recherche/2012 Bernengo-son.pdf.>. Acesso em dez. 2013.

DÍAZ CINTAS, J. In search of a theoretical framework for the study of audiovisual translation. In: Orero, P. (org.). **Topics in Audiovisual Translation.** Amsterdam e Philadelphia: John Benjamins, 2004, p. 19-32. crossref http://dx.doi.org/10.1075/btl.56.06dia

_____. **Traducción Audiovisual y accesibilidad**. In: Jiménez Hurtado, C. Traducción y accesibilidad. Subtitulación para sordos y audiodescripción para ciegos: nuevas modalidades de traducción audiovisual. (Ed.) Frankfurt AM Main: Peter Lang, 2007. p. 9-23.

FRANÇA. Charte de Qualité de l'Audiodescription. **L'Audiodescription:** principes et orientations. France: [s.n.], 2008. 13f. Não paginado. Disponível em: <http://www.enaparte. org/audiodescription/ La_Charte_files/La_Charte. html>. Acesso em: 30 jul.2012.

FRANCO, E. P. C.; ARAÚJO, V. L. S. Questões terminológico-conceituais no campo da tradução audiovisual (TAV). In: FROTA, M. P.; MARTINS, M. A.P. (Orgs.). **Tradução Audiovisual**. Revista, nº 11, 2011, p. 1-23. Disponível em: < http://audiodescricao.com/ site/files/2010/02/18884.pdf>. Acesso em: 16 jul. 2013.

FRANCO, E.; SILVA, M. Audiodescrição: breve passeio histórico. In: Motta, L.; Filho, P. R. **Audiodescrição:** ***transformando imagens em palavras***. (Ed.) São Paulo, 2010: Secretaria dos Direitos da Pessoa com Deficiência do Estado de São Paulo. p. 19-36.

HALLIDAY, M. A. K. **An introduction to functional grammar**. London: Edward Arnold, 1985.

_____. **An introduction to functional grammar**. 2 ed. London: Edward Arnold, 1994.

HALLIDAY, M. A. K; MATTHIESSEN, C. **An introduction to functional grammar**. 3 ed. New York: Arnold, 2004.

HERMANS, T. **Translation in systems.** Descriptive and systemic approaches explained. Manchester: St. Jerome, 1999.

HOLLAND, A. Audio description in the theatre and the visual arts: images intowords. In: Anderman, G. & Díaz-Cintas, J. Eds. **Audiovisual Translation:** language transfer on screen. Basingstoke; New York: Palgrave MacMillan, 2009.

HOLMES, J. The Name and the Nature of Translation Studies. **Translated!** Papers on Literary Translation and Translation Studies, Amsterdam, Rodopi, 1988/1972, p. 67-80.

INTOUCHABLES. Direção: Eric Toledano e Olivier Nakache. Produção: Nicolas Duval Adassovsky, Yann Zenou e Laurent Zeitoun. (Co)Produção: Quad, Gaumont, TF1 films production, TEN films, Chaocorp, Canal plus e Cinécinéma. Copyright@2011 Splendido, Gaumont, TF1 films

production, TEN films e Chaocorp. Origem: produção francesa. 2011. 1 Filme DVD, 1h52 min, son., color. Audiodescrito: VOSTF para Médiadub International. Francês. 2011.

JAKOBSON, R. Aspectos linguísticos da tradução. Trad. Izidoro Blikstein. In: Jakobson, R. **Linguística e comunicação**. São Paulo: Cultrix, 1995/1959, p.63-86.

JIMÉNEZ HURTADO, C. Una gramática local del guión audiodescrito. Desde la semántica a la pragmática de nuevo tipo de traducción. In: Hurtado, C. J. **Traducción y accessibilidad:** subtitulación para sordos y Audiodescripción para ciegos: nuevas modalidades de Traducción Audiovisual. Amsterdã: Peter Lang, 2007, p. 55-80.

_____; RODRÍGUEZ, A.; SEIBEL, C. **Un corpus del cine**. Teoría y practica de la Audiodescriptión. Granada: Tragacanto, 2010.

MACKEN-HORARIK, M. Interacting with the multimodal text: reflections on image and verbiage in ArtExpess. **Visual Communication**, v. 3, n. 1, p. 5-26, 2004. crossref http://dx.doi.org/10.1177/1470357204039596

MARTIN, J. R.; ROSE, D. **Working with discourse:** meaning beyond the clause. 2 ed. New York: Continuum, 2007.

MARTIN, J. R; WHITE, P. R. R. **The Language of evaluation:** appraisal in English. London: Palgrave/Macmillan, 2005.

MASCARENHAS, R. O. **A Audiodescrição da minissérie policial *Luna Caliente*:** uma proposta de tradução à luz da narratologia. Salvador, 2012. 285f. Tese. Universidade Federal da Bahia, Salvador, 2012.

MATTHIESSEN, C. **Lexicogrammatical cartography**: English systems. Tokyo: International Language Sciences Publishers, 1995.

MINUIT à Paris. Direção: Woody Allen. Produção: Pontchartrain. (Co)Produção: Helen Robin, Raphaël Benoliel, Letty Aronson, Stephen Tenenbaum e Jaume Roures. Copyright@2011 Mediaproducción, Versátil Cinema e Gravier Productions. Origem: produção américo-hispânica. 2011. Filme dublado francês. 1 Filme DVD, 1h34 min, son., color. Audiodescrito: VOSTF para Médiadub International. Francês. 2011.

NAVARRO, F. **Appraisal toolkit**: sobreviviendo a la Teoría de la Valoración (versión 12.04). Disponível em: <http://discurso.wordpress.com/2012/01/28/appraisal-toolkitsobreviviendo- a-la-teoria-de-la-valoracion/>. Acesso em: 04 fev. 2012.

NAVARRO, J. J. **A inclusão social dos deficientes visuais e a publicidade brasileira:** um breve panorama. 06/20815. Brasília, 2012. 61f. Monografia (Bacharel em Publicidade e Propaganda). Departamento de Audiovisuais e Publicidade da Faculdade de Comunicação Social da Universidade de Brasília. Brasília. 2012. Disponível em: <http://bdm.unb.br/handle/10483/4259 ou http://bdm.bce.unb.br/bitstream/10483/4259/ 1/2012Juliana JobimNavarro.pdf>. Acesso em: 18 mar. 2013.

PAGANO, A; VASCONCELLOS, M. L. Estudos da tradução no Brasil: reflexões sobre teses e dissertações elaboradas por pesquisadores brasileiros nas décadas de 1980 e 1990. **DELTA**, vol.19, 2003.

PRAXEDES FILHO, P. H. L.; MAGALHÃES, C. M. A neutralidade em Audiodescrições de pinturas: resultados preliminares de uma descrição via Teoria da Avaliatividade. **Os novos rumos da pesquisa em Audiodescrição no Brasil.** Org. Araújo e Aderaldo. 1 ed. Editora CRV. Curitiba, PR. 2013a, p. 73-87.

PRAXEDES FILHO, P. H. L.; MAGALHÃES, C. M. **A Audiodescrição de pinturas é neutra?** um estudo descritivo via Teoria da Avaliatividade. 2013b. 367f. Relatório de Estágio Pós-Doutoral (Programa de Pós-Graduação em Estudos Linguísticos). Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2013b.

RAI, S.; GREENING, J.; PETRÉ, L. **A comparative study of audio description guidelines prevalent in different countries**. London: Media and Culture Department, Royal National Institute of Blind People, 2010.

SALDANHA, G. Principles of corpus linguistics and their application to translation studies research. **Revista Tradumática**. Barcelona. n.7, dez. 2009. Disponível em: <http:// www.raco.cat/index.php/Tradumatica/.../ 206722 >. Acesso em: 15 jan. 2013.

SILVA, C. F. da. **Aspectos relevantes na tradução da charte francesa para Audiodescrição.** Fortaleza-CE, 2013. 93f. Monografia (Especialização em Formação de Tradutores). Universidade Estadual do Ceará, Fortaleza, 2013. Disponível em: <http://www.uece.br/posla/ dmdocuments/Cristiene%20Ferreira%20TCC>. Acesso em: set. 2014.

SILVA, F. T. dos S.; BONA, V. de; SILVA, A. da N. A.; CARVALHO, I.; SILVA, E. V. da. Refexões sobre o pilar da áudio-descrição: "descreva o que você vê". **Revista Brasileira de Tradução Visual**, 2010, v. 4, n. 4, p. 1-19.

SINCLAIR, J. **Corpus, concordance, collocation**, Oxford: Oxford University Press. 1991.

STANDARDS for audio description and code of professional conduct for describers. Disponível em: <http://audiodescriptioncoalition.org/adcstandards090615.pdf.>. Acesso em: mar. 2013.

TOURY, G. The Nature and Roles of Norms in Translation. In: **Descriptive Translation Studies - and beyond.** Amsterdam-Philadelphia: John Benjamins, 1995. p. 53-69. crossref http://dx.doi.org/10.1075/btl.4

VIAN JR., O.; SOUZA, A. A. de; ALMEIDA, F. S. D. P. (orgs.). **A linguagem da avaliação em língua portuguesa:** estudos sistêmico-funcionais com base no sistema de avaliatividade. São Carlos, SP: Pedro & João, 2010.

VILARONGA, I. A Dimensão formativa do cinema e a audiodescrição: um outro olhar. In: **Anais do II Encontro Nacional de Estudos da Imagem**. Londrina, 2009. p.1060.

Disponível em: <http://www.uel.br/eventos/eneimagem/anais/trabalhos/pdf/ Rodrigues Iracema% 20Vilaronga .pdf>. Acesso em: 18 mar.2013.

Artigo recebido em: 15.10.2014
Artigo aprovado em: 14.12.2014

Convergência lexical entre letras de música e inglês geral: um estudo baseado em *corpus*

Lexical convergence between pop song lyrics and general English: a *corpus*-based study

Patrícia Bértoli*

RESUMO: Este trabalho apresenta um estudo de convergência lexical entre letras de música gravadas originalmente em inglês e o inglês geral. O objeto do estudo é um *corpus* composto por cerca de 1 milhão de palavras, advindas de 5.962 letras de musicas diferentes. As letras de música foram consideradas por suas características linguísticas, ou seja, como texto (BÉRTOLI-DUTRA, 2002). Do *corpus* de estudo foram extraídas listas de palavras individuais e de feixes lexicais de três palavras (trigramas). Essas listas foram contrastadas com listas extraídas de dois *corpora* de referência, os quais são demonstrativos do inglês geral; as versões disponibilizadas online do BNC e do ANC. A pesquisa encontra-se fundamentada na Linguística de *Corpus*, que considera a linguagem como um sistema probabilístico, argumentando que o uso de características e traços linguísticos varia de acordo com o contexto imediato (as palavras que a antecedem e sucedem) e com o registro. A análise contrastiva apontou proximidade linguística entre letras de música e o inglês geral, no que diz respeito a ocorrência e frequência de palavras isoladas e trigramas, o que indica proximidade entre a linguagem usada nas letras das músicas analisadas e o inglês coloquial, ou seja, o inglês informal em uso geral (BIBER, 1988)

PALAVRAS-CHAVE: Letras de música. Inglês geral. Linguística de *corpus*. Convergência. Frequência.

ABSTRACT: This paper presents a study of lexical convergence between pop song lyrics and General English. The corpus is a compilation of roughly 1 million words from 5.962 different pop song lyrics. For the purpose of this study, 'Lyrics' is understood as a text type/register. Lists of individual words and lexical bundles of three words (trigrams) were extracted from the corpus. The lists were then contrasted to lists extracted from two reference *corpora* representative of General English: the BNC and the ANC. Grounded on Corpus Linguistics, which views language as a probabilistic system, the study assumes that the use of words–varies according to its immediate context (preceding and following words), as well as according to the register. The contrastive analysis showed strong linguistic similarities between pop song lyrics and General English, respectively to the occurrence and frequency of individual words and trigrams. This result suggests proximity between the language used in pop song lyrics and colloquial English, that is general informal English (BIBER, 1988).

KEYWORDS: Song Lyrics. General English.Corpus Linguistics. Convergence. Frequency.

* Doutora em Linguísica Aplicada e Estudos da Linguagem. Instituto de Letras – ILE – LAG, UERJ.

1. Introdução

Na qualidade de registro/tipo textual (BÉRTOLI-DUTRA, 2002; 2012), letras de música têm recebido pouco valor como fonte textual de pesquisa, especialmente, por serem consideradas como um registro não espontâneo e de cunho popular. Grande parte dos estudos sobre letras de música tem como foco suas características poéticas, temáticas ou semióticas (e.g., TATIT, 2001; COSTA, 2002; ROSSATO, 2002; DAMAZO, 2004; CROSSLEY, 2005; MATTE, 2005); ou de suas características sonoras e para o desenvolvimento de atividades de ensino (MURPHEY, 1990; MORA, 2000; ARLEO, 2000; BÉRTOLI-DUTRA, 2002). Este trabalho propõe uma outra abordagem ao investigar similaridades e diferenças de uso lexical entre o discurso de letras de música popular produzida originalmente em língua inglesa e outros registros em inglês, que compõem o que se considera inglês geral. Segundo Berber Sardinha (2004), um *corpus* de geral de uma língua deve "incluir o maior número possível de registros encontrados na língua alvo, e cada registro, por sua vez, deve ter o maior número possível de exemplares." (BERBER-SARDINHA, 2004, p. 18)

A noção de música popular utilizada nesta pesquisa segue a proposta de Starr e Waterman (2007) que a consideram como toda forma musical produzida e divulgada em massa pelos meios de comunicação, a fim de atingir o maior número de consumidores possíveis. Essa perspectiva foi adotada pois considera-se que esse é o tipo de música que chega rapidamente ao ouvido do consumidor – em especial do consumidor de música internacional que é motivação constante de pesquisa por ser um dos meios de praticar a aprendizagem em língua estrangeira ao considerar suas letras.

O estudo segue também o princípio de que letras de música não podem ser consideradas apenas por suas características poéticas, ou seja, como poesias musicadas, pois embora os dois registros compartilhem características semelhantes como métrica, rima, etc. (COSTA, 2002), também apresentam características diferentes, tanto por seu componente sonoro quanto pelas escolhas lexicais que as caracterizam (BÉRTOLI-DUTRA, 2002, 2012; GRIFFITHS, 2003). Ademais, o foco deste estudo reside exatamente no uso lexical.

A fim de alcançar o objetivo principal deste estudo exploratório a fim de detectar similaridades e diferenças lexicais entre o discurso de letras de música popular escrita em inglês e o inglês geral , o estudo foi orientado especificamente pela Linguística de *Corpus* (BERBER SARDINHA, 2004; BIBER, 1988) e guiado pelos seguintes procedimentos: (a) levantamento de frequências das palavras usadas no *corpus* de estudo; (b) descrição de

padrões léxico-gramaticais do *corpus* de estudo; e (c) análise contrastiva dos padrões encontrados no *corpus* de estudo em oposição aos padrões existentes na língua inglesa, conforme representados pelo *corpus* de referência.

Este artigo está dividido da seguinte forma, primeiramente serão apresentados os princípios teóricos que fundamentaram a pesquisa, em sendo lugar, serão descritos os *corpora* utilizados e os procedimentos metodológicos e analíticos. Em seguida, estarão dispostos os resultados e as conclusões.

2. Base Teórica

A seleção de letras de música como fonte de pesquisa encontra sustento teórico na relevância de sua característica social, conforme argumentado por Starr e Waterman (2007) e Frith (1993). Outrossim, o estudo ora descrito partiu do pressuposto de que letras de música podem ser consideradas como representantes de um registro específico, fazendo, portanto, jus a análise linguística (BÉRTOLI-DUTRA, 2002; 2012).

As representações transferidas pela música, sejam sonoras ou referentes às mensagens presentes nas letras ou o conteúdo imagético apresentado por seus intérpretes, integram o conjunto de características com as quais nos identificamos e, consequentemente construímos nossas personalidades, nossas identidades particulares e partilhadas. A natureza sociológica da música popular e sua contribuição para a formação da cultura e da identidade dos indivíduos (STARR; WATERMAN, 2007; FRITH, 1993) faz da música um relevante objeto de estudo, ainda que sejam observadas especificamente suas características linguísticas dissociadas de seus componentes musicais, como feito aqui.

É importante esclarecer que a concepção de gênero utilizada neste trabalho, assim como a de registro, segue a perspectiva de Biber e Conrad (BIBER, 1988; BIBER; CONRAD, 2009, p. 2) para quem o nível do gênero demonstra os propósitos comunicativos e situacionais de um texto, enquanto o nível de registro apresenta as características funcionais, as quais são representadas pelas propriedades sintáticas, morfológicas e semânticas de um texto.

A análise de letras de música como texto observa suas características no nível de registro, consequentemente, considera-a como pertencente ao gênero oral, tal como sermões e discursos políticos, os quais são preparados antecipadamente a serem apresentados ao ouvinte, mas não envolvem interação direta e imediata entre os interlocutores. Segundo o

sociólogo Frith (1998), as letras fazem parte do gênero coloquialismo conversacional, trazendo consigo inúmeras variações linguísticas, que permitem a exploração de diferentes línguas, modos de falar, de palavras enquanto símbolos sonoros, entre outros, podendo, portanto, ser relacionada a contextos situacionais específicos, que possuem propósitos comunicativos específicos, como se espera de um gênero (BERTOLI-DUTRA, 2012).

Sendo assim, neste trabalho, pressupomos a existência do gênero letras de música, que se caracteriza principalmente por sua peculiaridade sonora, cujos textos compartilham o propósito de terem sido escritos de modo que pudessem ser integrados a estruturas melódicas. Ademais, Frith (1998) argumenta que letras não são simplesmente poesias cantadas, mas uma forma de retórica e oratória, num relacionamento de persuasão entre cantor e ouvinte. Dessa forma, consideraram-se letras de música como fonte de conhecimento linguístico, vez que se entende a transferência das palavras que ouvimos nas canções para outras formas do discurso, como, por exemplo, uma discussão oral sobre música, num ambiente social (MOORE, 2003). Em suma, a presença da música e suas características linguísticas foram fatores preponderantes para a consideração de letras de música enquanto texto, uma forma de registro que merece ser investigado. Sendo assim, o estudo teve como fundamentação teórica principal a Linguística de *Corpus* (BERBER-SARDINHA, 2004; BIBER, 1988)

A Linguística de *Corpus* é uma área que estuda os fenômenos da língua por meio da observação de grandes quantidades de dados empíricos armazenados eletronicamente, por computador. De modo simples, pode-se dizer que um *corpus* é uma coletânea de dados linguísticos autênticos, que podem ser processados por programas computacionais, cuja coleta segue critérios rigorosos, que estejam de acordo com o propósito da pesquisa e que sejam representativos (BERBER SARDINHA, 2004). Segundo Berber Sardinha, as quatro características fundamentais para a coleta dos dados formadores de um *corpus* computadorizado são textos em linguagem natural, autenticidade, critério rigoroso e representatividade. No que diz respeito a representatividade é um pouco mais delicada, enquanto a extensão do *corpus* deve ser considerada, vez que a visão probabilística da língua subjaz que "quanto maior a quantidade de palavras, maior a probabilidade de aparecerem palavras de baixa frequência" (BERBER SARDINHA, 2004, p. 22-23), por outro lado, é preciso considerar-se o quê se pretende representar. Em suma, o tamanho de um *corpus* de estudo é dependente daquilo que se pretende estudar.

O *corpus* coletado para esta pesquisa encaixa-se nas características fundamentais

apontadas por Berber Sardinha, vez que advêm de fonte natural, pois foram produzidos por falantes nativos para outros fins que não o da pesquisa; são autênticos, pois estiveram sujeitos à interação natural com o leitor/ouvinte (são letras de músicas realmente gravadas e veiculadas pela mídia); e são representativos de letras de música, vez que correspondem a uma amostra significativa desse universo linguístico, mais de 1 milhão de palavras, sendo, portanto, classificado como um *corpus* médio-grande.

Vale ressaltar que os linguistas do *corpus* consideram a língua como um sistema probabilístico, o que está enraizado na observação empírica da língua, ou seja, na observação natural da ocorrência de dados linguísticos em situações reais de uso, seguindo as propostas de Halliday, Sinclair e Firth (FIRTH, 1957; HALLIDAY; HASAN, 1989; HALLIDAY; 1991; SINCLAIR; 1991; HALLIDAY; WEBSTER; 2002). Essa visão probabilística assume que, embora muitas escolhas e combinações lexicais sejam possíveis, dentro de um determinado contexto (e na língua em geral), elas não ocorrem da mesma forma nem com a mesma frequência (BERBER SARDINHA, 2004, p. 30). A variação de escolhas linguísticas não ocorre aleatoriamente, ela segue uma certa padronização representativa de cada gênero. Pode-se dizer, portanto, que as combinações lexicais têm maior ou menor probabilidade de ocorrerem, ou são mais ou menos frequentes, de acordo com o contexto. A observação empírica vai permitir uma descrição mais completa e mais precisa da língua, pois evidenciará aquilo que é mais típico (HALLIDAY; WEBSTER, 2002, p. 10) em um determinado contexto. Nesse sentido, vale lembrar que a noção de gênero e registro empregadas para esta pesquisa foge da visão tradicional no que se "refere a diferentes perspectivas de variedade textuais", onde o registro é analisado pela combinação de características linguísticas funcionais comuns, como pronomes e verbos, com a análise da situação de uso da variedade textual. Já a perspectiva de gênero abarca a descrição dos propósitos comunicativos e situacionais, como a de registro, mas a análise linguística "concentra-se nas estruturas convencionais usadas para construir um texto completo dentro de uma variedade, por exemplo, a maneira convencional como uma carta começa e termina" (BIBER; CONRAD, 2009 p.2).

A Linguística de *Corpus* também considera a análise de características e traços linguísticos (BIBER, 1988; BERBER SARDINHA, 2004) como representativos da língua, vez que sua utilização não obedecem a um critério arbitrário de escolha, ou seja, existe variação, ou seja, os falantes podem utilizar características linguísticas (palavras, estruturas,

etc.) diferentes para dizer as mesmas coisas. A variação linguística pode ser associada às funções comunicativas do texto. Essa ideia permeia o trabalho descrito aqui, uma vez que procura-se observar as semelhanças entre diferentes representações da língua (diferentes registros pertencentes a gêneros variados).

A análise de características linguísticas pode ser feita de diversas maneiras. Para este estudo, consideramos palavras individuais e trigramas (conjunto de três palavras seguidas), por suas ocorrências e frequências, defendendo suas características como marcadores textuais. Um n-grama é um feixe lexical (*lexical bundle*) que contém um numero de palavras representado pela letra "n". Esta pesquisa utilizou trigramas, ou seja, um feixe lexical de três palavras. Esse feixe deve ser analisado por seu sentido ou função como conjunto, pois entende-se que os agrupamentos não acontecem por acaso ou de forma randômica, mas carregam um significado específico quando em bloco. Segundo Sinclair (1991) e Hyland (2008) grande parte das palavras não tem um significado independente e, portanto, a interpretação dos feixes, ou n-gramas, auxilia no reconhecimento de gêneros discursivos.

Dados os fundamentos que norteiam esse estudo, passa-se à descrição dos *corpora* e da metodologia adotada.

3. *Corpora*

O objetivo do estudo aqui descrito foi extrair, relacionar e contrastar padrões léxico-gramaticais (BÉRTOLI DUTRA, 2002; BERBER SARDINHA, 2004; HUNSTON & FRANCIS, 2000) entre um *corpus* de letras de música gravadas originalmente em inglês coletado especificamente para esta pesquisa e dois *corpora* de inglês geral: as versões disponíveis online do British National Corpus (BNC), com mais de 100 milhões de palavras, e do American National Corpus (ANC) com aproximadamente de 15 milhões de palavras.

A versão online do BNC, também conhecida como BNC World Edition[1], está dividida em 3.144 textos escritos e 910 textos falados, computando 100.467.090 palavras. Os textos escritos advêm de fontes como: jornais e periódicos voltados para o público de todas as idades e interesses, textos e ensaios acadêmicos, obras ficção popular, cartas, memorandos, entre outros. A porção falada é composta por transcrições de conversas informais e formais (e.g. conversas telefônicas, reuniões de negócios ou governamentais, programas de rádio).

[1] Cf. http://www.natcorp.ox.ac.uk/*corpus*/index.xml.ID=numbers

O ANC, no momento de coleta, contava com 22 milhões de palavras[2]. Trata-se de um projeto cujo objetivo é oferecer fonte de pesquisa do inglês americano para fins educacionais, linguísticos, lexicográficos, e para o desenvolvimento tecnológico. Foi utilizado neste estudo por estar disponível online para download na forma de listas de frequência por palavras individuais, ou por associadas (i.e., bigrams e trigramas) .

Para compor o *corpus* de estudo foram coletadas 5.962 letras de música compostas originalmente em inglês, computando 1.078.882 ocorrências (*tokens*), resultantes do uso de 22.136 formas (*types*), ou palavras diferentes. A coleta foi realizada seguindo rigorosamente os seguintes critérios: (1) As letras de música foram coletadas e categorizadas como sub*corpora* por artista (banda ou intérprete) na totalidade de suas músicas gravadas (foram descartadas músicas apenas instrumentais); (2) músicas gravadas por mais de um artista foram consideradas apenas uma vez; (3) os artistas foram selecionados por sua representatividade (marcas de sucesso em determinado período) e pela longevidade (fizeram ou fazem sucesso por muitos anos); (4) um dos sub*corpora* deveria favorecer músicas mais tocadas na época da pesquisa. Tais critérios foram considerados a fim de não se favorecer músicas ou artistas de apenas uma determinada época. Em outras palavras, artistas como Frank Sinatra foram escolhidas por sua presença nas paradas de sucesso por mais de sessenta anos (cf. *Billboard Charts*[3]), mesmo após sua morte, ao mesmo tempo que a banda Greenday, por exemplo, representa o grupo dos que fazem sucesso nos anos 2000 (id.). A versão final do *corpus* de estudo contém 5.962 letras de músicas gravadas por 30 artistas diferentes e pode ser melhor visualizada no quadro a seguir:

Quadro 1: composição do *corpus* de estudo.

ARTISTA	CANÇÕES
Aerosmith	56
Beatles	250
Bon Jovi	195
Creed	69
Def Leppard	143
Elton John	326
Foo Fighters	109
Frank Sinatra	1.224
Greenday	169

[2] Cf. http://www.americannational*corpus*.org/
[3] Para ver detalhes sobre paradas de sucesso em diversas épocas e variadas categorias, sugerimos consultar os sites: http://www.billboard.com/ e http://www.musicimprint.com/Chart.aspx?id=C000002

Iron Maiden	215
The Killers	52
Led Zeppelin	79
Lenny Kravitz	125
Madonna	168
My Chemical Romance	61
Metallica	119
Nickelback	79
Nirvana	98
Paramore	50
Pearl Jam	146
Pink Floyd	164
Queen	172
Queensryche	112
Ray Charles	206
Red Hot Chilly Peppers	169
Rolling Stones	432
Simple Plan	49
Teen	202
TOP 100	121
U2	170

Vale ressaltar que o sub*corpus* TOP 100 é composto pelas músicas mais tocadas pela MTV Brasil e outras rádios brasileiras em agosto de 2008, e o sub*corpus* TEEN é representado por músicas de sucesso em 2008 voltadas para o público adolescente, sendo composto por letras das trilhas sonoras dos filmes High School Musical I, II e III e de artistas como Hannah Montana, The Jonas Brothers.

Uma vez coletados e definidos os *corpora* a serem utilizados no estudo, passou-se aos procedimentos de análise de convergência lexical entre os mesmos.

4. Análise

A análise contrastiva partiu da obtenção (*download*) de listas de palavras do ANC e da extração de listas de palavras do *corpus* de estudo e do BNC. Essas últimas foram extraídas utilizando-se o software de análise lexical WordSmith tools[4] (SCOTT, 1998) que processa textos apresentando recursos como listas de palavras, de palavras chave e de concordâncias. Em seguida, as 500 palavras mais frequentes no *corpus* de estudo foram investigadas quanto a estarem ou não presentes nos *corpora* de referência. Optou-se pelo contraste das 500 formas

[4] disponível em < http://www.lexically.net/wordsmith/>

mais frequentes por se tratar de uma análise manual e por se considerar como uma amostra representativa da totalidade do *corpus* de estudo em relação ao inglês geral. A lista das palavras mais frequentes no *corpus* de estudo pode ser vista no anexo 1.

Todas as 500 palavras aparecem no BNC e no ANC, como era de se esperar. Portanto, passou-se a observar a frequência, ou seja, o número de ocorrência de cada uma delas em cada um dos *corpora*. A fim de que os dados pudessem ser comparados, as ocorrências nos três *corpora* foram normalizadas por 100. A normalização um procedimento que permite que as frequências compartilhem o mesmo parâmetro de comparação e é feita da seguinte maneira, divide-se o número de frequência pelo número de palavras do *corpus* e o resultado é multiplicado por 100. Dessa forma, a palavra "*the*" (a mais frequente nos três *corpora*), por exemplo, aparece 43.428 vezes no *corpus* de estudo, tendo frequência normalizada para 4,025. Já no BNC, a frequência bruta de 6.055.105 é normalizada para 6,027 e no ANC de 1.204.816 para 5,44. Em suma, se cada um dos *corpora* tivesse apenas 100 palavras, o "*the*" teria aparecido cerca de 4 vezes no CE, 6 vezes no BNC e 5 vezes no ANC.

A fim de ampliar o escopo da análise, foram também considerados agrupamentos de três palavras, também chamados trigramas (e.g. "*I want you*"; "*I love you*"; "*I don't know*"), fundamentando-se na premissa de que o que determina o uso de determinada palavra é sua associação com as palavras que a antecedem e as que a sucedem (LAFFERTY; SLEATOR; TEMPERLY, 1992). Além disso, segundo os mesmos autores, trigramas refletem tanto sintaxe, quanto semântica e pragmática, de forma simultânea. Esta análise também considerou as 500 primeiras formas mais frequentes nos três *corpora*. Assim como para as palavras individuais, os dados foram normalizados para se tornarem comparáveis.

De posse das listas das palavras individuais e trigramas mais frequentes nos *corpora* de estudo e de referência, os dados foram contrastados a partir do *corpus* de estudo. Em outras palavras, procurou-se nos *corpora* de referência a ocorrência concomitante das palavras presentes no *corpus* de estudo. Suas frequências foram contrastadas. O mesmo procedimento foi dado à análise de trigramas nos três *corpora*.

5. Resultados

Na análise de palavras individuais, foi observado que todas as 500 palavras mais frequentes no *corpus* de estudo também estão presentes no BNC e no ANC. Ademais, houve concomitância entre as 6 primeiras palavras mais frequentes, as quais são exatamente as

mesmas nos três *corpora*, dadas as devidas diferenças de número de ocorrência, ainda que normalizadas. Por exemplo, "*the*", aparece 4,02 no *corpus* de estudo, 6,02 no BNC e 5,44 no ANC. "*You*" aparece 3,33 no *corpus* de estudo, 0,58 no BNC e 0,80 no ANC. Ou seja, em comparação ao inglês geral, embora "the" seja a palavra mais frequente, ela é menos utilizada em letras de música; enquanto que "you" é muito mais utilizada em letras de música que no inglês geral. A título de ilustração, as 15 palavras mais frequentes nos três *corpora* estão dispostas na tabela seguinte:

Tabela 1: contraste de frequência das palavras mais frequentes no *corpus* de estudo com o BNC e o ANC.

palavra	frequência *Corpus* de estudo	frequência BNC	frequência ANC
1. THE	4,02	6,02	5,44
2. YOU	3,33	0,58	0,80
3. I	3,33	0,73	0,85
4. TO	2,36	2,58	2.40
5. AND	2,28	2,61	2,68
6. A	2,14	2,17	2,21
7. ME	1,59	0,13	0,15
8. MY	1,35	0,14	0,24
9. IN	1,29	1,93	1,84
10. IT	1,21	0,91	1.15
11. OF	1,17	3,03	2,73
12. YOUR	0,99	0,13	0,11
13. ON	0,91	0,72	0,63
14. THAT	0,87	1,04	0,76
15.ALL	0,80	0,27	0,23

A observação da tabela indica que embora as palavras estejam presentes nos três *corpora*, há variação de frequência, isto é, de uso. O quadro mostra que nas letras de música presentes no *corpus* de estudo há presença marcante de pronomes pessoais de primeira e segunda pessoas *(I; you*), demonstrando preferência por discurso interpessoal, cujos participantes devem ser a pessoa que canta (eu) e o ouvinte (você), o que corrobora com o achado de Bértoli-Dutra (2002).

Foi considerado como foco de análise particularmente o uso de léxico com funções básicas, como substantivos, verbos, pronomes, etc. O contraste quantitativo entre as frequências dessas palavras nos *corpora* indicou a presença das formas "*'m*" e "*are*" sugerindo que acompanhem o uso pronominal mais frequente, de primeira e segunda pessoa, especialmente para "*I*", cuja frequência apresenta-se muito mais elevada no *corpus* de letras de música. Já "*are*" pode aparecer acompanhando outras palavras, não estando

exclusivamente associada a um pronome, diferente de "*'m*", que só aparece quando acompanhado de "*I*" e, por isso, sua frequência aparece de forma mais equilibrada, quando contrastada aos outros *corpora*.

O uso não flexionado do verbo "*be*" também aparece de forma equilibrada, assim como "*is*", "*are*" e "*we*". Palavras como "*love", "no", "like", "do", "can", "got", "if", "one", "up", "time", "never", "see"* e "*baby"* apresentam sobreuso no *corpus* de estudo, com destaque para "*one*" e "*baby*", sendo que a última aparece cerca de 40 vezes mais. A tabela a seguir representa uma amostra da análise realizada:

Tabela 2: amostra de contraste de frequência de itens lexicais do *corpus* de estudo com o BNC e o ANC.

palavra	Frequência no *corpus* de estudo	frequência BNC	frequência ANC
'm	0,82	0,062	0,15
be	0,76	0,64	0,48
is	0,75	0,96	1,02
love	0,70	0,02	0,01
we	0,54	0,29	0,42
know	0,53	0,11	0,21
like	0,44	0,14	0,02
do	0,42	0,17	0,09
can	0,41	0,21	0,21
got	0,41	0,08	0,06
if	0,40	0,25	0,25
one	0,38	0,28	0,0018
get	0,37	0,95	0,10
down	0,36	0,91	0,05
go	0,36	0,86	0,021
up	0,36	0,20	0,017
time	0,35	0,15	0,010
are	0,33	0,45	0,47
never	0,32	0,05	0,059
was	0,32	0,85	0,71
see	0,31	0,11	0,104
baby	0,31	0,0079	0,008

A análise quantitativa de palavras individuais (uma única palavra) indicou que a escolha lexical em letras de música apresenta grande semelhança com a escolha realizada em outros gêneros da língua inglesa, ou seja, em outros registros[5] que não letras de música. Tal constatação sugere que a linguagem usada em letras de música pode ser também classificada

[5] Ver Biber; Conrad, 2009 para concepção de gênero e registro usada neste estudo, já mencionada anteriormente

como coloquial, pois está simultaneamente presente em outros gêneros da língua inglesa, tais como cartas, conversas informais e textos jornalísticos – lembrando que os *corpora* de referência não apresentam letras de música em sua composição.

A fim de não restringir tal constatação e a análise ao uso de palavras individuais, o mesmo procedimento de análise foi dado à ocorrência e frequência de trigramas presentes nas letras de música de forma concomitante aos *corpora* de inglês geral. A extração de trigramas resultou na ocorrência de 129.117 trigramas diferentes para o *corpus* de estudo, 5.431.734 para o BNC e as listas do ANC apresentaram 1.453.050 para sua porção falada e 4.236.030 para sua porção escrita, cujos 30 mais frequentes podem ser vistos no quadro a seguir:

Quadro 2: 30 primeiros trigramas no *corpus* de estudo contrastados com trigramas dos *corpora* de referência.

trigramas CE	trigramas BNC	trigramas ANC falado	trigramas ANC escrito
1. I WANT TO 540	ONE OF THE 20760	I I DO 997	THE UNITED STATES 2600
2. I LOVE YOU 505	THE END OF 11604	AND I AM 993	THE NEW YORK 2477
3. I DON'T KNOW 371	AS WELL AS 11502	YOU THINK OF 99	NEW YORK TIMES 2427
4. LA LA LA 333	PART OF THE 10182	YEAH YEAH OH 99	ONE OF THE 2217
5. I DON'T WANT 294	THERE IS A 9655	YEAH BUT UM 99	THE THE THE 2141
6. DON'T WANT TO 285	OUT OF THE 9162	WAS NICE TALKING 99	THE OF THE 1635
7. OH OH OH 256	SOME OF THE 9142	UM THERE IS 99	IT IS A 1554
8. YEAH YEAH YEAH 254	EQUO HE SAID 8922	TIME AND I 99	THE IN THE 1505
9. I KNOW THAT 239	A NUMBER OF 8358	THAT IS SO 99	I DO NOT 1313
10. I DON'T WANNA 228	END OF THE 7931	TAKING CARE OF 99	IT IS NOT 1213
11. I GOT A 214	THERE WAS A 7741	SIT DOWN AND 99	THE THE A 1156
12. IN LOVE WITH 207	IT WAS A 7685	RIGHT BUT UH 99	IN THE THE 1124
13. I WANT YOU 205	THE FACT THAT 7673	NOT KNOW THE 99	THE A THE 1112
14. NA NA NA 202	THERE IS NO 7457	IT THAT WAY 99	A THE THE 1061
15. WANT TO BE 200	BE ABLE TO 7298	IT FOR A 99	A IN THE 1056
16. YOU AND ME 195	IN ORDER TO 7110	IT A LITTLE 99	OF THE NEW 1049
17. YOU WANT TO 191	TO BE A 6992	IS I AM 99	THERE IS A 1006
18. YOU AND I 187	IT IS NOT 6893	I UNDERSTAND THAT 99	AS WELL AS 980
19. IN YOUR EYES 184	PER CENT OF 6841	I THINK ABOUT 99	A LOT OF 958
20. I WANNA BE 183	A LOT OF 6574	I REALLY I 99	THE WHITE HOUSE 948
21. I DON'T NEED 181	IT IS A 6377	I LISTEN TO 99	THE END OF 934
22. IF YOU WANT 180	EQUO BQUO I 6291	I I UH 99	A A A 879
23. AND I DON'T 178	HE SAID BQUO 5929	I HAD THE 99	THERE IS NO 871
24. IN MY HEART 177	IN TERMS OF 5816	I GUESS THERE 99	THE THE IN 865
25. AND I KNOW 176	IT WOULD BE 5764	I AM LOOKING 99	THE THE OF 818
26. THIS IS THE 175	AT THE END 5608	HUM UM I 99	OF THE THE 815
27. CAN'T YOU SEE 174	MOST OF THE 5403	HUH AND UM 99	THE A A 812
28. TO BE A 169	ON THE OTHER 5391	GOOD THAT IS 99	SOME OF THE 795
29. YOU KNOW THAT 167	THAT IT IS 5389	AND UH MY 99	PART OF THE 782
30. ON AND ON 166	THE NUMBER OF 5372	AND THERE ARE 99	THE TO THE 775

O quadro anterior representa apenas uma amostra dos trigramas analisados, vez que foram considerados os 500 trigramas mais frequentes do *corpus* de estudo, os quais foram manualmente observamos por sua presença nos *corpora* de referência. Esse primeiro

levantamento considerou simplesmente a ocorrência concomitante dos trigramas. Dessa forma, constatou-se que apenas 12 dos 500 trigramas mais frequentes no *corpus* de letras de música não estão presentes em um ou outro dos *corpora* de referência. Os 222 primeiros trigramas mais frequentes no *corpus* de estudo estão também presentes no BNC e ANC. A partir do 223º trigrama (*ooh ooh ooh*), os outros onze começam a aparecer exclusivamente no *corpus* de letras de música, sendo: "*c'mon c'mon c'mon"; "oooh oooh oooh"; "oo oo oo"; "good to ya"; "aah aah aah"; "wanna be your"; "need your love"; "love love love"; "all I wanna"; "ah ah ah"; "be your man*".

A análise desses onze trigramas demonstra que sete dessas expressões apresentam variações da escrita em relação à forma culta (i.e: fogem dos padrões gramaticais estabelecidos), como a repetição de letras (e.g. *"aah"; "oooh")* que é muito usada em letras de música, normalmente indicando o prolongamento da emissão do som (no cantar); da mesma forma, apresenta formas contraídas como "*wanna*" e "*ya*" (para *you*); e a repetição de palavras (*c'mon c'mon*).

O trigrama "*need your love*", surpreendentemente não aparece nos *corpora* de referência, onde aparecem expressões como "*need your advice/passport/help/expertise/attention*". Tal fato pode se justificar pelo sobreuso da palavra "*love*", conforme observado na análise de palavras individuais. O mesmo aconteceu com o trigrama "*be your man*", que não foi encontrado nos *corpora* de referência, os quais apresentaram variações como "*be your dream/friend/guide/home/idea/you"*. Pode-se inferir que essa expressão parece ser de uso mais comum em letras de música por também apresentar referência a amor e sexualidade.

No que diz respeito aos 22 trigramas que estão presentes nos três *corpora*, o trigrama "*I want to*", que ocorre 540 vezes no *corpus* de estudo, sendo o mais frequente, refere-se a 0,5% das vezes em contraste ao BNC, onde aparece 2.110 vezes, equivalendo a 0,02%. Embora pareça uma grande diferença, a presença do trigrama já confirma seu uso na língua inglesa em geral, não apenas nas músicas. Interessantemente, expressões como "*la la la*" aparecem nos três *corpora*. Isto é, embora possam parecer de uso exclusivo da música, também são usadas no inglês geral.

6. Conclusão

O estudo exploratório aqui apresentando buscou traços de convergência lexical entre

um *corpus* de letras de música e dois *corpora* de inglês geral, o BNC (de inglês britânico) e o ANC (de inglês americano). A constatação de que as palavras individuais mais frequentes no *corpus* de estudo também estão entre as mais frequentes nos *corpora* de referência, seja americano ou britânico, confirma o identificado pela pesquisa de Bértoli-Dutra (2002), que ressalta a proximidade entre a língua utilizada em letras de música e o inglês coloquial. Por outro lado, este estudo também observou a mesma tendência de uso quando consideradas três palavras juntas, ou trigramas.

O sobreuso de certas palavras, ou seja, a maior ocorrência dessas palavras no *corpus* de estudo do que nos *corpora* de inglês geral, sugere que exista uma espécie de conversa romântica[6] (entre pessoas que se amam) nas letras de música, de forma muito mais acentuada do que no inglês geral[7], dadas as frequências de palavras como "*baby*" e "*love*", ou seja, evidencia um tipo de "fala à pessoa amada". O que também pode ser confirmado pelo uso exclusivo dos trigramas "*be your man*" e "*need your love*". Deve-se considerar, contudo, a possibilidade de que parte das ocorrências de "*baby*" nos *corpora* de referência refira-se a bebê, e não a tratamento de afeto, o qual é a forma de uso mais frequentemente no *corpus* de estudo.

Este estudo indicou também que se uma análise puramente quantitativa de frequências normalizadas pode apontar convergências com o inglês geral, ao mesmo tempo não parece ser suficiente para apontar as características das letras de música como um todo (o que não era objeto deste estudo, mas que parece ser bastante interessante). Um dos fatores que contribuiria para uma análise mais profunda das letras de música seria a etiquetagem do *corpus*. Sendo assim, uma das limitações do estudo refere-se a, por exemplo, não se poder separar os usos de palavras como "*love", "like"* ou *"know*", bastante frequentes no *corpus* de estudo, em relação ao seu uso como morfossintático, o que, poderia nos dar uma ideia melhor de como as letras se comportam. Isto é, se as letras usam a forma "*know*", por exemplo, para indicar o que se sabe ou conhece da mesma forma que no inglês geral ou se está sendo usada mais frequentemente para preencher espaços conversacionais, com expressões como "*you know*".

A análise de trigramas possibilitou uma nova perspectiva de aproximação do discurso geral (utilizado em variados registros) com o das letras de música e, portanto, foi considerada

[6] Os temas das canções são variados, há canções de amor, Guerra, morte, paz, justiça, religião, etc. (ver Bértoli-Dutra, 2014)
[7] Entende-se por inglês geral a língua utilizada nos mais variados registros falados e escritos

de forma satisfatória para justificar o uso de letras de música em sala de aula para fins de ensino de línguas, não simplesmente para práticas auditivas.

As duas esferas de análise revelaram que letras de música estão muito próximas do inglês geral, ao mesmo tempo que provocaram a necessidade estudos futuros que venham a descrever mais detalhadamente o discurso da música.

Referências

ARLEO, A. Music, song and foreign language teaching. **Les Cahiers de l'APLUIT**, vol. XIX, no. 4. 2004 p. 5-19.

BERBER SARDINHA, A. P. **Linguística de Corpus**. Barueri: Manole, 2004.

BÉRTOLI-DUTRA, P. **Explorando a linguística de *corpus* e letras de música na produção de atividades pedagógicas**. Dissertação Inédita (Mestrado em LAEL), PUC-SP, 2002.

___. Song Lyrics and Speech: similarities, differences and multidimensional analysis of Song Lyrics from 1940 to 2009. In: MELLO, H.; PETTORINO, M.; RASO, T. **Proceedings of the VII GSCP International Conference**: Speech and Corpora. Firenze: Firenze University Press, 2012.

___. Multi-Dimendional analysis of pop songs In: BERBER SARDINHA, T; VEIRANO PINTO, M. **Multi-Dimensional Analysis, 25 years on**: A tribute to Douglas Biber. 1 ed.Amsterdam : John Benjamins, 2014, p. 151-177.

BIBER, D. **Variation across speech and writing**. Cambridge: Cambridge University Press, 1988. crossref http://dx.doi.org/10.1017/CBO9780511621024

COSTA, N. B. da. As letras e a letra: o gênero canção na mídia literária. In: DIONÍSIO, A. P.; MACHADO, A. R.; BEZERRA, M. A.. (orgs.). **Gêneros textuais e ensino**. Rio de Janeiro: Lucerna, 2002.

CROSSLEY, S. A. Metaphorical conceptions in hip-hop music. **African American Review**, vol. 39, no. 4, Winter, 2005. p. 501-512.

DAMAZO, F.A.F.T. **O canto do povo de um lugar**: uma leitura das canções de João do Vale. Tese de doutorado apresentada ao Programa de Estudos Literários da Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP: SJ Rio Preto, 2004.

FIRTH, A. The discursive accomplishment of normality on 'língua franca' English and conversation analysis. **Journal of Pragmatics** 26. 1996 (237-259). crossref http://dx.doi.org/10.1016/0378-2166(96)00014-8

FRITH, S. Music and identity. In: HALL, S.; Du GAY, P. (Eds). **Questions of Cultural Identity**, London, UK: Sage publications, 1993. p. 108-127.

_____. Performin Rites. **On the value of popular music**. Cambridge, Massachusetts, USA: Harvard University Press, 1998.

GRIFFITHS, D. From Lyrics to anti-lyric: analyzing the words in pop songs. In: A. F. MOORE. (ed). **Analysing Popular Music**. Cambridge: Cambridge University Press, 2003. (p. 39-59). crossref http://dx.doi.org/10.1017/CBO9780511482014.003

HALLIDAY, M. A. K. Corpus studies and probabilistic grammar. In: K. Aijmer; B. Altenberg (org.) **English Corpus Linguistics**: Studies in honour of Jan Svartvik. (30-43). London: Longman, 1991.

HALLIDAY, M.A.K; HASAN ; R. **Language, context, and text**: aspects of language in a social-semiotic perspective. 2nd edition. Deakin University Press/Oxford University Press, 1989.

HALLIDAY, M.A.K; WEBSTER, J. (ed.). **On grammar:** By Michael Alexander Kirkwood Halliday. New York: Continuum, 2002.

HUNSTON, S.; FRANCIS, G. **Pattern grammar:** a corpus-driven approach to the lexical grammar of English. Amsterdam/Phildelphia: John Benjamins, 2000. crossref http://dx.doi.org/10.1075/scl.4

HYLAND, K. As can be seen: lexical bundles and disciplinary variatuon. **ESP**. Vol. 27, 2008, p. 4-21.

LAFFERTY, J.; SLEATOR, D.; TEMPERLEY, D. Grammatical trigrams: A probabilistic model of link grammar, in **Proceedings of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language**, Cambridge, MA, October 1992.

MATTE, A. C. F. Gostar de música: percurso de uma paixão. **Significação**: revista brasileira de semiótica. 23. São Paulo: ANNABLUME. Junho, 2005. p. 72-92.

MOORE, A. F. (ed). **Analysing Popular Music**. Cambridge: Cambridge University Press, 2003.

MORA, C. F. Foreign language acquisition and melody singing. **ELT Journal**, vol. 54, no. 2, 2000, p. 146-152. crossref http://dx.doi.org/10.1093/elt/54.2.146

MURPHEY, T. The song stuck in my head phenomenon: a melodic din in the lad? **System**, vol. 18, vol. 1, 1990, p. 53-64.

ROSSATO, E. V. A brasilidade na poesia contemporânea de Caetano Veloso. In: III SEMINÁRIO DE ESTUDOS SOBRE LINGUAGEM E SIGNIFICAÇÃO - **SELISIGNO**: DISCURSO E REPRESENTAÇÃO E IV SIMPÓSIO DE LEITURA DA UEL, 2002, Londrina. Caderno de Resumos do IV Simpósio de Leitura da UEL: III Selisigno: Discurso e Representação, 2002. v. 1. p. 93-93.

SCOTT, M. **WordSmith Tools.** Version 3. Oxford University Press: England, 1998.

SINCLAIR, J. *Corpus*, **Concordance, Collocation**. Oxford: Oxford University Press, 1991.

STARR, L.; WATERMAN, C.. **American Popular Music**. From minstrelsy to MP3. 2nd. ed. New York: OUP, 2007.

TATIT, L. **O cancionista:** composição de canções no Brasil. São Paulo: EDUSP, 1996.

ANEXO 1

LISTA DAS 500 PALAVRAS MAIS FREQUENTES NO CORPUS DE ESTUDO

1 THE 43.428	126 THEN 1.419	251 MYSELF 598	376 DEAR 388
2 YOU 35.995	127 HOME 1.402	252 REMEMBER 596	377 ITS 387
3 I 35.964	128 ALWAYS 1.359	253 THOUGHT 596	378 UNTIL 386
4 TO 25.485	129 ABOUT 1.356	254 WAIT 589	379 LATE 382
5 AND 24.662	130 STILL 1.340	255 SOUL 587	380 FINE 381
6 A 23.138	131 HAD 1.333	256 LIVING 584	381 MORNING 379
7 ME 17.209	132 EVERY 1.327	257 COLD 581	382 YEARS 379
8 MY 14.635	133 KEEP 1.321	258 FEELING 580	383 ANYTHING 374
9 IN 13.966	134 WOULD 1.305	259 ONCE 575	384 TRYING 374
10 IT 13.118	135 OVER 1.287	260 SIDE 575	385 DOES 373
11 OF 12.685	136 US 1.261	261 PAIN 574	386 MISS 373
12 YOUR 10.694	137 EVER 1.255	262 THOSE 574	387 PRETTY 370
13 ON 9.951	138 HEAR 1.233	263 HANDS 566	388 TOUCH 370
14 THAT 9.471	139 AN 1.225	264 CHORUS 562	389 LAY 369
15 I'M 8.891	140 WERE 1.216	265 DONE 562	390 MONEY 369
16 ALL 8.729	141 YOU'LL 1.209	266 TILL 561	391 SUCH 366
17 BE 8.205	142 AM 1.185	267 HELP 559	392 NOBODY 365
18 IS 8.199	143 MUCH 1.146	268 LOOKING 559	393 MEET 364
19 FOR 7.759	144 MIND 1.145	269 DAYS 558	394 REASON 359
20 LOVE 7.606	145 BETTER 1.089	270 TOWN 553	395 RUNNING 358
21 DON'T 7.217	146 BELIEVE 1.087	271 WHAT'S 552	396 EASY 357
22 SO 6.492	147 OLD 1.085	272 BRING 550	397 GROUND 357
23 IT'S 6.220	148 NOTHING 1.083	273 GOODBYE 550	398 GUESS 355
24 BUT 6.157	149 INTO 1.074	274 NA 543	399 SOON 353
25 WE 5.865	150 THINGS 1.065	275 REAL 541	400 CHILD 351
26 JUST 5.815	151 GONE 1.057	276 FOUND 538	401 BYE 348
27 KNOW 5.712	152 HOLD 1.037	277 FORGET 535	402 MATTER 347
28 NO 5.488	153 LIVE 1.018	278 THEY'RE 531	403 LISTEN 346
29 WITH 5.459	154 ALONE 1.011	279 LOSE 530	404 SOUND 346
30 WHEN 5.307	155 TRY 991	280 GOES 520	405 WENT 343
31 WHAT 4.956	156 I'D 988	281 KNEW 518	406 SAD 341
32 OH 4.905	157 GOTTA 964	282 ALRIGHT 517	407 UNDERSTAND 339

33 THIS 4.879	158 DREAM 963	283 MAKES 517	408 MUSIC 338
34 LIKE 4.842	159 THEM 963	284 BEST 515	409 NEXT 338
35 DO 4.550	160 LEAVE 946	285 START 512	410 RED 338
36 CAN 4.520	161 FACE 942	286 DEAD 507	411 SHAKE 338
37 GOT 4.447	162 RUN 940	287 WOMAN 505	412 READY 337
38 IF 4.386	163 YOU'VE 927	288 HEAVEN 504	413 PART 336
39 NOW 4.379	164 ANOTHER 925	289 LONELY 504	414 SEA 336
40 YOU'RE 4.253	165 LIGHT 915	290 ROCK 503	415 WORK 336
41 OUT 4.194	166 PLEASE 915	291 RAIN 500	416 FEAR 335
42 ONE 4.122	167 DID 911	292 DOOR 497	417 FUN 334
43 GET 4.024	168 SUN 907	293 CLOSE 496	418 BURNING 331
44 YEAH 4.024	169 HAS 906	294 WISH 496	419 TRUTH 331
45 DOWN 3.906	170 BEFORE 905	295 BECAUSE 495	420 ABOVE 330
46 GO 3.889	171 GOING 904	296 SMILE 492	421 FEET 330
47 UP 3.883	172 NEW 902	297 SEEN 490	422 GIRLS 330
48 TIME 3.820	173 SOMETHING 902	298 BLACK 489	423 LADY 330
49 ARE 3.613	174 REALLY 898	299 MAYBE 488	424 TOMORROW 330
50 NEVER 3.546	175 CALL 891	300 MEAN 485	425 WONDER 330
51 WAS 3.485	176 HEAD 884	301 KIND 484	426 BOYS 329
52 SEE 3.438	177 YES 882	302 TALK 484	427 EVERYONE 327
53 BABY 3.387	178 HAND 879	303 EVERYBODY 478	428 WANTS 327
54 SHE 3.387	179 INSIDE 876	304 CRAZY 477	429 TRIED 326
55 WILL 3.374	180 TONIGHT 871	305 HAPPY 475	430 WHITE 324
56 CAN'T 3.270	181 LAST 860	306 ROLL 475	431 MAMA 321
57 HAVE 3.260	182 PLACE 842	307 FLY 474	432 BROKEN 320
58 COME 3.249	183 STAY 840	308 ROUND 471	433 STRONG 318
59 NOT 3.194	184 OFF 838	309 OPEN 469	434 BABE 316
60 I'LL 3.181	185 THAN 837	310 TOLD 466	435 LIGHTS 316
61 WANT 3.103	186 THING 837	311 YOURSELF 465	436 FALLING 314
62 SAY 3.070	187 HIM 804	312 AH 462	437 SAYS 314
63 FROM 3.041	188 STOP 798	313 SOMEBODY 461	438 SET 312
64 TAKE 2.973	189 THEIR 791	314 SURE 461	439 WANTED 312
65 HER 2.933	190 LET'S 788	315 MANY 460	440 DARK 311
66 WAY 2.851	191 DIE 784	316 USED 460	441 C'MON 306
67 THEY 2.822	192 WALK 780	317 CHANCE 459	442 BORN 305
68 BACK 2.734	193 WE'LL 780	318 CAME 457	443 COULDN'T 305
69 GONNA 2.733	194 MINE 778	319 ENOUGH 457	444 ASK 304
70 AS 2.731	195 FALL 777	320 LINE 457	445 MOTHER 302
71 MAKE 2.729	196 CRY 774	321 FRIEND 456	446 WHOA 302
72 AWAY 2.689	197 MADE 774	322 TEARS 454	447 LAND 299
73 THERE 2.661	198 OOH 774	323 MOON 453	448 WHO'S 298
74 AT 2.642	199 TWO 770	324 SOMETIMES 452	449 WILD 298
75 LET 2.575	200 TRUE 768	325 FRIENDS 451	450 LORD 297
76 I'VE 2.425	201 TURN 768	326 FOREVER 450	451 SEEM 297
77 DAY 2.382	202 SHOULD 767	327 FIGHT 448	452 GAME 296

78 HOW 2.320	203 SOMEONE 766	328 ARMS 444	453 KILL 292
79 NIGHT 2.306	204 WRONG 760	329 RIDE 444	454 LOW 291
80 HEART 2.260	205 SWEET 758	330 WORDS 444	455 BEAUTIFUL 290
81 HE 2.223	206 LA 757	331 LIE 443	456 STANDING 290
82 LIFE 2.221	207 HARD 756	332 GETTING 441	457 HATE 288
83 FEEL 2.178	208 WITHOUT 756	333 TIMES 440	458 WHOLE 288
84 TELL 2.079	209 HE'S 751	334 STAR 439	459 CITY 287
85 RIGHT 2.061	210 MUST 737	335 KNOWS 438	460 TAKES 286
86 COULD 2.055	211 DANCE 736	336 TODAY 438	461 SHINE 284
87 WELL 2.051	212 LOST 733	337 YOUNG 438	462 WAR 283
88 HERE 2.046	213 EVERYTHING 729	338 ANY 437	463 HEARTS 282
89 CAUSE 2.041	214 LEFT 726	339 HOPE 437	464 HOT 282
90 THERE'S 1.938	215 COMING 725	340 ALIVE 435	465 HIT 281
91 TOO 1.937	216 FREE 725	341 HIDE 434	466 LIPS 279
92 MORE 1.932	217 SONG 720	342 HELL 432	467 WALKING 279
93 WHERE 1.871	218 COMES 718	343 BLOOD 430	468 DEATH 278
94 BY 1.845	219 END 710	344 STARS 429	469 MEN 277
95 MAN 1.836	220 SING 710	345 EACH 428	470 PARTY 277
96 LITTLE 1.831	221 PEOPLE 709	346 SLEEP 426	471 GUN 276
97 GOOD 1.823	222 SHOW 703	347 MIGHT 424	472 HA 275
98 NEED 1.820	223 CARE 699	348 AFTER 422	473 DOESN'T 274
99 BEEN 1.814	224 WHILE 698	349 DIDN'T 420	474 GETS 274
100 HEY 1.775	225 THESE 695	350 LIES 420	475 SINCE 273
101 OR 1.727	226 BAD 689	351 MOVE 420	476 BURN 272
102 GIVE 1.694	227 HIGH 683	352 ALONG 418	477 GAVE 272
103 THAT'S 1.663	228 SAME 679	353 WIND 414	478 HURT 272
104 THINK 1.655	229 WAITING 676	354 SEEMS 411	479 AIR 270
105 WORLD 1.652	230 OWN 670	355 BEAT 409	480 LIVES 267
106 OUR 1.643	231 PLAY 668	356 TOOK 409	481 SOMEWHERE 267
107 EYES 1.639	232 SKY 659	357 SAVE 408	482 TIRED 267
108 AGAIN 1.632	233 BREAK 658	358 HONEY 407	483 LEARN 266
109 ONLY 1.628	234 KISS 658	359 STREET 402	484 SOMEDAY 266
110 THROUGH 1.614	235 PUT 657	360 THOUGH 402	485 BED 263
111 WHO 1.604	236 NAME 649	361 YOU'D 402	486 BIT 259
112 WANNA 1.553	237 CHANGE 647	362 WATCH 401	487 PAST 259
113 SAID 1.547	238 DREAMS 645	363 UNDER 400	488 UPON 258
114 LONG 1.545	239 BOY 639	364 FOOL 397	489 TIGHT 256
115 WHY 1.542	240 BLUE 633	365 VERY 397	490 DARLING 255
116 GIRL 1.534	241 GOD 619	366 DEEP 396	491 FULL 255
117 WE'RE 1.533	242 MAY 618	367 BEHIND 395	492 BLUES 254
118 AROUND 1.517	243 STAND 618	368 ROOM 395	493 CONTROL 252
119 SOME 1.517	244 BIG 614	369 WORD 395	494 LOVED 251
120 HIS 1.496	245 TOGETHER 606	370 SAW 394	495 USE 251
121 SHE'S 1.494	246 FAR 604	371 WE'VE 391	496 STEP 250
122 AIN'T 1.478	247 OTHER 604	372 ELSE 390	497 MOMENT 249

123 FIND 1.471	248 EVEN 602	373 FIRST 390	498 EYE 248
124 WON'T 1.471	249 YA 602	374 HEARD 389	499 HELLO 248
125 LOOK 1.437	250 FIRE 600	375 ROAD 389	500 LOVING 248

Artigo recebido em: 15.10.2014
Artigo aprovado em: 29.11.2014

Fotografia técnica de documentos para formação de *corpora* digitais eletrônicos: o método desenvolvido no Lapelinc

Technical Photography of documents for construction of digital corpora: the method developed in Lapelinc

Jorge Viana Santos[*]
Giovane Santos Brito[**]

RESUMO: O presente artigo apresenta descritivamente etapas do método fotográfico que vem sendo desenvolvido e utilizado, desde 2008, no Lapelinc (*Laboratório de Pesquisa em Linguística de Corpus* - UESB). Nos limites deste trabalho, por recorte metodológico, destaca-se um dos instrumentos desenvolvidos especificamente para uso neste método, no processo de transposição de documentos manuscritos históricos do tipo jurídico para formação de *corpora* linguísticos eletrônicos: a *Mesa Cartesiana*.

PALAVRAS-CHAVE: Linguística de *Corpus*. Fotografia. Corpora Eletrônicos. Documentos Históricos.

ABSTRACT: This paper shows steps of Lapelinc Photographic Method, developed since 2008, at the Lapelinc (Corpus Linguistics Research Laboratory). Methodologically, we emphasize the Cartesian Table (Mesa Cartesiana), a instrument developed by us for material transposition of historical documents to construction of digital linguistic corpora.

KEYWORDS: Corpus Linguistics. Photography. Electronic Corpora. Historical Documents

1. Introdução

Como explicam Brito e Santos (2013), ao depararmos com documentos antigos, podemos encontrar respostas para fenômenos linguísticos complexos relacionados a períodos remotos e presentes das línguas, a exemplo do que foi feito por Santos (2008). Investigações desse tipo podem ajudar pesquisadores na busca da compreensão, descrição e explicação de fenômenos concernentes a uma língua, tanto em seus aspectos históricos e filológicos quanto no que se refere à própria gramática.

[*] Doutor em Linguística pela Universidade de Campinas (Unicamp). Professor do Departamento de Estudos Linguísticos e Literários da Universidade Estadual do Sudoeste da Bahia (DELL/UESB). Professor do quadro permanente do Programa da Pós-Graduação em Linguística (PPGLIN/UESB). Professor colaborador do Programa de Pós-Graduação em Memória, Linguagem e Sociedade (PPGMLS/UESB). Pesquisador da Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB). Membro do Grupo de Pesquisa em Estudos da Linguagem (UESB/Cnpq), do Grupo de Pesquisa Lugares de Enunciação e Processos de Subjetivação (Unicamp/Cnpq), do Grupo de Pesquisa Mulheres em Discurso (Unicamp/Cnpq) e do Grupo de Pesquisa PROHPOR (UFBA/UESB/Cnpq).

[**] Mestrando em Linguística pelo Programa da Pós-Graduação em Linguística da Universidade Estadual do Sudoeste da Bahia (PPGLIN/UESB). Pesquisador bolsista da Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB). Membro do Grupo de Pesquisa em Estudos da Linguagem (UESB/Cnpq).

Tal realidade, porém, encontra algumas dificuldades: acesso e disponibilização de documentos históricos para fins de pesquisa; e, no caso de reprodução e/ou cópia de documentos desse tipo, a garantia de fidedignidade necessária para a pesquisa científica.

Neste sentido, o presente artigo objetiva apresentar descritivamente etapas do método fotográfico quem vem sendo desenvolvido e utilizado no Lapelinc (*Laboratório de Pesquisa em Linguística de Corpus* - UESB), desde 2008, destacando especificamente, a *Mesa Cartesiana*, um dos instrumentos desenvolvidos para uso no processo de transposição, para o formato digital, de documentos manuscritos históricos do tipo jurídico para formação de *corpora* linguísticos eletrônicos.

Considerando a questão "Qual a viabilidade do uso da fotografia para a captação fidedigna de documentos para compor *corpora* digitais, visando estudos linguísticos e científicos?", no Lapelinc postulamos por hipótese que, desde que metodicamente controlada em suas fases de captura, catalogação, edição, armazenamento, e leitura, a Fotografia apresenta-se como forma altamente viável e producente de digitalização, permitindo à Linguística, ou outra ciência, acessar imageticamente, de modo confiável, o documento não disponível no local da pesquisa.

Para tanto, nos limites deste trabalho, consideraremos, para comprovação, dois itens: em primeiro, lugar apresentamos as etapas do *Método Lapelinc*; em segundo, descrevemos exemplificadamente a *Mesa Cartesiana*, instrumento de fotografia, desenvolvido para tal método.

2. Fotografia técnica de documentos: o método desenvolvido no Lapelinc

O *Método Lapelinc* é um método de fotografia cientificamente controlada que, desenvolvemos e temos aplicado e aperfeiçoado desde 2008 no processo de transposição[1] de documentos manuscritos originais em papel para o formato digital, conforme descrito nos trabalhos de Santos (2008, 2010a, 2010b, 2013a, 2013b), Namiuti, Santos e Leite (2011), Namiuti-Temponi, Santos, Costa e Farias (2013), Brito, Santos e Namiuti-Temponi (2013),

[1] Sobre o conceito de transposição, ver Santos e Namiuti (2013).

com vistas a integrar *corpora* eletrônicos anotados, a exemplo do *Corpus Dovic*[2] *Beta*. Possui três características:

a) Pressupõe domínio da Fotografia (*Photography*) enquanto linguagem e enquanto técnica.

b) Necessita de equipamento e aparato técnico auxiliar específicos, a exemplo da *Mesa Cartesiana* (cf. item 2).

c) Visa à construção, sobretudo, de *corpora* manuscritos para uso científico: Linguística, História, Direito, Memória, dentre outras.

Em conjunto, com estas três características, como postulado em Namiuti-Temponi, Santos, Costa e Farias (2013, p. 12), objetivamos algo muito importante: transformar a Fotografia não num simples meio de reprodução de um documento, uma fotografia pragmática, que serve apenas a uma pesquisa e não tem compromisso de futuro, mas sim praticá-la com método científico de reprodução digital, como se vê no gráfico 1:

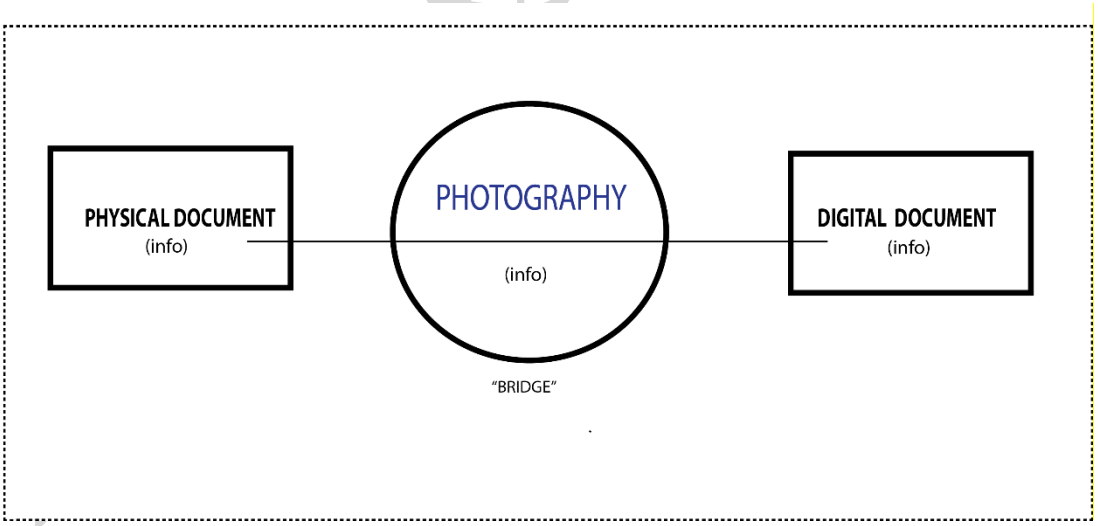


Gráfico 1: Fotografia praticada com método científico de reprodução digital: a *ponte* entre Documento Físico (DF) e Documento digital (DD)[3]
Fonte: *Material transposition of Documents for a digital corpus implementation* (SANTOS; NAMIUTI, 2013b)

[2] Dovic (Documentos Oitocentistas de Vitória da Conquista e região) (SANTOS; NAMIUTI, 2014) é um corpus eletrônico anotado, desenvolvido no âmbito do Projeto Corpora Digitais Para a História do Português Brasileiro – região Sudoeste da Bahia: Aliança PHPB – Tycho Brahe (FAPESB: 6171/2010) (SANTOS; NAMIUTI, 2010).
[3] Gráfico originalmente apresentado em inglês em Santos e Namiuti (2013b).

> No gráfico [...], a Fotografia funciona como uma espécie de *ponte* entre DF (documento físico) e o DD (documento digital). Mas para isso, defendemos que é preciso que ela registre, na própria imagem, dados/informações que façam com que a imagem gerada não perca o vínculo com o documento que lhe deu origem (NAMIUTI-TEMPONI; SANTOS; COSTA; FARIAS, 2013, p. 15)

Tal método, em sua aplicação, apresenta cinco etapas principais:

1) Controle: etapa da captura de informações da fonte (por exemplo, catalogação de dados de um livro a ser fotografado);

2) Captura fotográfica da imagem do original: fotografia sequenciada dos documentos utilizando equipamentos adequados, inserindo na imagem a quantidade necessária de dados que garanta a sua relação com o objeto que a originou. Ou seja: fotografa-se o DF para se formar o DD;

3) Catalogação no *Database* Dovic das folhas-imagens componentes do documento;

4) Edição;

5) Criação de imagens de uso co-indexadas à imagem-original.

Dentre essas etapas, destaque-se que para a uso na 2, desenvolvemos no Lapelinc um equipamento com vistas a viabilizar de modo controlado e cientificamente padronizado a Fotografia: a *Mesa Cartesiana*. Vejamos.

3. Um instrumento do método Lapelinc: a *mesa cartesiana*

3.1 O projeto

A *Mesa Cartesiana* foi idealizada e concebida para ser um equipamento que possibilite, junto com o registro do documento, o registro visual cientificamente controlado de suas características físicas, tais como: tom, cor, tamanho, acervo/arquivo do original físico, tipologia, paginação e sequenciação.

3.2 O *layout*

No tocante ao *layout,* a fim de alcançar tal objetivo, a *Mesa Cartesiana* apresenta-se como na figura 1, em que se observam: escala de tom (1), escala de cores (2), instrumentos de medição calibrados (3, 4, 5), informações catalográficas (6), paginação (7) e sequenciação (8).

FIGURA 1: *Mesa Cartesiana* (lay out)
Fonte: *Corpus Dovic Beta*

3.3 O funcionamento

Assim configurada, a *Mesa Cartesiana* possui as condições técnicas necessárias para contribuir para a garantia de fidedignidade necessária à pesquisa científica. Isto porque, durante a captura fotográfica digital (cf. item 2), os elementos componentes funcionam do seguinte modo:

> a) Escala de tom (1) e escala de cores (2): sendo escalas científicas elaboradas para o controle fotográfico, possui amostras de tons e cores com parâmetros que, podem ser interpretados por programas e *softwares* de edição e leitura de imagem, capazes por isso de, por exemplo, recuperar numa tela de computador as tom/cores originais de um documento, independente da leitura que o olho humano faça.
>
> b) Instrumentos de medição (3, 4, 5): sendo escalas científicas elaboradas para controle milimétrico, do modo como estão dispostas,

formam um perfeito plano cartesiano, capaz de matematicamente permitir o cálculo preciso das medidas de quaisquer documentos (livros, folhas...), independente da sua posição.

c) Informações catalográficas (6), paginação (7)[4], sequenciação (8)[5]: garantem um vínculo permanente entre o DF e o DD.

Os formatos das imagens capturadas atendem basicamente à orientação do CONARQ (Conselho Nacional de Arquivos) (CONARQ, 2010, p. 14-15), que sugere a captura de uma matriz no formato *Raw*[6] que, em nosso caso, tem a dimensão de 4256 x 2832 *pixels*, com profundidade de 14 bits por canal (RGB), gerando um arquivo não comprimido (*uncompressed*) de 12.1 *megapixels*. Juntamente com esse arquivo, a câmara gera um arquivo em JPEG[7], com resolução mínima de 300 *dpi*. O mesmo arquivo em *Raw* possibilita, ainda, a criação de arquivos de alta resolução para armazenamento no formato TIFF[8], outra recomendação do CONARQ.

Destaque-se que, por se tratar de um método que se desenvolve em função da realidade do objeto para qual se destina, e considerando que os livros notariais antigos normalmente eram numerados em apenas um lado da folha (no anverso), ficando o verso sem o respectivo número, equivalendo, pois, ao verso de mesmo número do anverso – como se vê nas figuras 2 e 3 abaixo –, após o uso de um numerador manual para indicar a paginação com codificação de cores: vermelho, para folhas frente, e verde, para folhas verso, combinado com uso de um sequenciador (em azul) para indicar a ordem da imagem no original, já está em desenvolvimento no setor de informática do Lapelinc um numerador e um sequenciador eletrônicos, para serem utilizados nas coletas a partir de 2015, com vistas a eliminar as possibilidades de erro humano.

---

[4] Numerador com números sequenciais que seguem a ordem de numeração das folhas (frente e verso) conforme constam no livro físico. Folhas não numeradas no original, por regra, são indicadas como "000", caso, por exemplo, de uma capa.

5 Sequenciador com números sequenciais, de 1 (capa) a "n" (verso da última capa ou folha do documento), que indicam a ordem das folhas-imagens, isto é, de cada imagem capturadas para fins de edição.

6 Raw, arquivo digital com dados não processados, que é gerado pela câmara, e não pode ser diretamente manipulado, apenas possibilitando a geração de outro arquivo em formato diverso, a exemplo de JPEG ou TIFF (cf. LONG, 2004, p. 26)

7 JPEG (Joint Photographic Experts Group), arquivo digital comprimido (cf. LONG, 2004, p. 27)

8 TIFF, arquivo digital não comprimido, o que, em comparação, por exemplo, ao JPEG, o torna superior em qualidade de imagem (cf. LONG, 2004, p. 26).

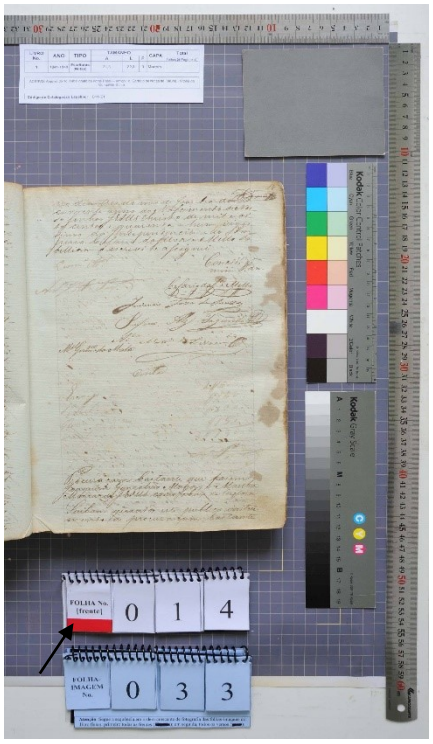


FIGURA 2: folha frente.
Fonte: *Corpus Dovic Beta.*

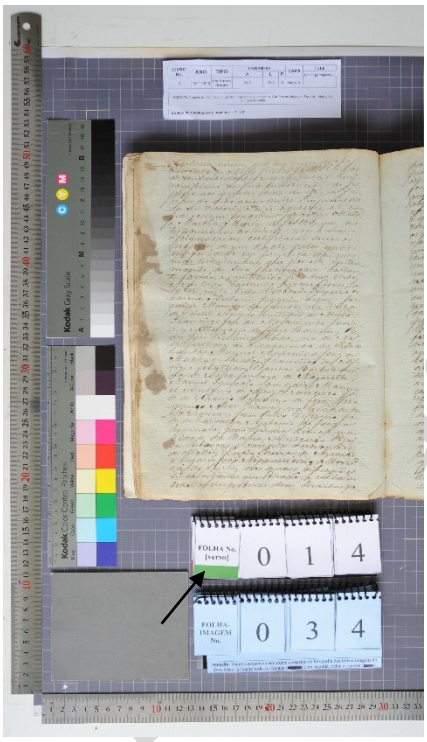


FIGURA 3: folha verso.
Fonte: *Corpus Dovic Beta.*

Tais informações são de suma importância para o controle da edição, pois, como se vê nas figuras 2 e 3, as imagens das folhas são capturadas uma a uma, com excesso de borda suficiente para serem editadas posteriormente, quando se recortam todas as informações inseridas na *Mesa Cartesiana*, para obtermos a imagem de uso, contendo apenas o documento, como exemplificam as figuras 4 e 5:

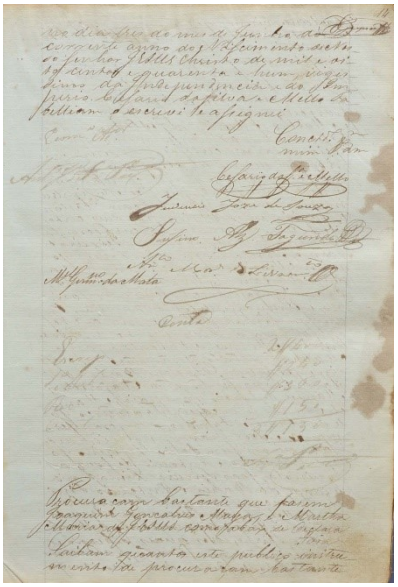


FIGURA 4: folha frente editada.
Fonte: *Corpus Dovic Beta.*

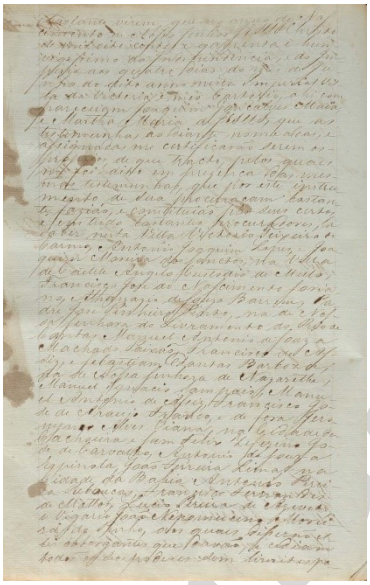


FIGURA 5: folha verso editada.
Fonte: *Corpus Dovic Beta.*

Cabe ressaltar que o *Método Lapelinc* se propõe como método que visa à digitalização por Fotografia e não por escaneamanento. Isto devido à natureza do objeto que compõe o acervo com que lidamos: livros notariais manuscritos que, dada a sua idade/datação (Século XIX) e grande tempo de manuseio (muitos ainda estão em uso), apresentam-se hoje em estado de extrema fragilidade, sendo impraticável a sua digitalização por *scanner*. O processo de escaneamento por mais moderno que seja é fixo e exige que o objeto/documento se adeque a ele em termos de tamanho, iluminação, etc. Diferentemente, a Fotografia, enquanto linguagem e enquanto técnica, dada à sua flexibilidade, devida a quase infinita possibilidade de variação de parâmetros, se adequa à realidade do objeto e não o inverso.

4. Considerações finais

Pelo exposto, verifica-se que o *Método Lapelinc,* no processo de transposição de documentos manuscritos históricos para formação de *corpora* linguísticos eletrônicos, configura-se de fato como passível de transpor um documento do formato papel, um DF, para um formato digital, um DD, com fidedignidade requerida pela Ciência.

Isto porque, ao pô-lo ao adotá-lo, torna-se necessário, não só postular, mas também praticar a Fotografia de modo cientificamente controlado, utilizando equipamentos especificamente criados para esse fim, tal como a *Mesa Cartesiana*.

É, portanto, a Fotografia viabilizando a ciência, preservando a história, preservando a memória e – muito importante – contribuindo para o desenvolvimento da Linguística de *Corpus*.

Referências

BRITO, G. S.; SANTOS, J. V. (2013). **A designação dos sobrenomes de escravos como estigma da escravidão em fontes judiciais do Sudoeste baiano**. Projeto de pesquisa. Fapesb. 2013.

BRITO, G. S.; SANTOS, J. V.; NAMIUTI-TEMPONI, C. Photograph(y) as a means for transposition of paper to digital text. Poster Session. **Workshop on construction and use of large annotated corpora**. Unicamp. Campinas: 2013.

CONARQ – Conselho Nacional de Arquivos. **Recomendações para digitalização de documentos arquivísticos permanentes**. Rio de Janeiro: Arquivo Nacional, 2010.

LONG, B. **Complete digital photography**. Hingan: Charles River Media, 2004.

NAMIUTI, C.; SANTOS, J. V.; LEITE, C. M. B. Propostas e Desafios dos Novos Meios das Antigas Fontes: A Preservação da Memória pela Linguística de Corpus. In: X Colóquio Nacional e II Colóquio Internacional do Museu Pedagógico UESB, 2011, Vitória da Conquista. **Anais do X Colóquio Nacional e II Colóquio Internacional do Museu Pedagógico** UESB. Vitória da Conquista: UESB, 2011. v. 1. p. 1-11.

NAMIUTI-TEMPONI, C.; SANTOS, J. V.; COSTA, A.; FARIAS, I. S. Computação e Linguística: importante diálogo para pesquisas e preservação da memória nos novos meios das antigas fontes. **RBBA**, v. 2, n. 1, p. 9-34, jul./2013.

SANTOS, J. V. (2008). **Liberdade na escravidão**: uma abordagem semântica do conceito de liberdade em cartas de alforria. Tese (Doutorado em Linguística) – Instituto de Estudos da Linguagem da UNICAMP, Campinas, 2008.

SANTOS, J. V. **Apresentação de meios para o transporte**: digitalização de documentos manuscritos e impressos. Conferência ministrada na I Oficina de Linguística de Corpus da Bahia (UEFS, UESB, UFBA). Feira de Santana: UEFS, 2010a.

SANTOS, J. V. **Técnicas de transporte do texto manuscrito para o meio digital.** Conferência ministrada na I Oficina de Linguística de Corpus da Bahia (UEFS, UESB, UFBA). Feira de Santana: UEFS, 2010b.

SANTOS, J. V.; NAMIUTI, C. **Corpora Digitais Para a História do Português Brasileiro - região Sudoeste da Bahia**: Aliança PHPB - Tycho Brahe. Projeto de pesquisa. UESB, Vitória da Conquista, 2010. (FAPESB: 6171/2010)

SANTOS, J. V.; NAMIUTI, C. **DOVIC (Documentos Oitocentistas de Vitória da Conquista e região)**. Corpora Eletrônico. UESB. Vitória da Conquista, 2014.

SANTOS, J. V.; NAMIUTI, C. Material transposition of Documents for a digital corpus implementation. Lecture. **Workshop on construction and use of large annotated corpora**. Unicamp. Campinas: 2013b.

SANTOS. J. V. **Um método de Fotografia técnica documental para formação de corpora digitais de documentos históricos manuscritos**. 2013. Uesb. Vitória da Conquista. 2013a. (Curso).

Artigo recebido em: 15.10.2014
Artigo aprovado em: 15.12.2014

Centro e margem dos discursos sobre *sustentabilidade*: da ecologia linguística ao ecossistema social

Center and margins of discourse about sustainability: from the social ecology to the linguistic ecosystem

Cláudio Márcio do Carmo*

RESUMO: Conforme Kennedy (1998), *ecologia linguística* é uma das quatro áreas de trabalho em Linguística de *Corpus*, ocupando-se da análise de padrões lexicais de que um determinado item faz parte, tendo como objetivo descrever sentidos a que um item se associa, em quais estruturas ele aparece e qual correlação existe entre seu uso e o sentido a ele atribuído. A partir disso, procura-se ter acesso a seu valor na organização do texto. Tendo esse pressuposto como base, trabalhando na interface entre estudos de *corpora*, Linguística Sistêmico-Funcional e Análise Crítica do Discurso, pretende-se analisar o item *sustentabilidade* em um *corpus* de pequena dimensão (SINCLAIR, 2001) de textos coletados na *internet*, para procurar entender o que é central tematicamente e o que margeia os discursos a ele associados, com foco no entendimento do *ecossistema social* (LEMKE, 2003). Entende-se, apoiando-se em Martin (1992) e Eggins (1994), que as relações lexicais são relevantes para a compreensão de fenômenos linguísticos e, em Williams (1976), que existem itens culturalmente relevantes. Mas essa relevância também é produzida sócio-historicamente; logo, por consequência disso, hoje se torna fundamental especialmente por causa do processo de globalização, em que novos (ou a recontextualização/rearticulação de velhos) discursos passaram a ter maior alcance devido às novas temporalidades e espacialidades permitidas pelas novas tecnologias de informação. Descrevendo a

ABSTRACT: As Kennedy (1998), *linguistic ecology* is one of the four desktops in Corpus Linguistics, minding the analysis of lexical patterns that a particular item is part, aiming to describe the way an item is associated, in which structures it appears and the correlations between their use and the meaning attributed to it. From this, we seek to have access to its value in the organization of the text. Taking this assumption as a basis, working into the interface between corpus studies, Systemic Functional-Linguistics and Critical Discourse Analysis, we intend to analyze the *sustainability* item on a small corpus (SINCLAIR, 2001) of texts collected in internet, searching for understand what is central and bordering thematically and the discourses associated with it, with a focus on understanding the *social ecosystem* (LEMKE, 2003). Understand, relying on Martin (1992) and Eggins (1994), the *lexical relations* as relevant to the understanding of linguistic phenomena, and on Williams (1976), which are culturally relevant items. However, this relevance is also produced socio-historically, so by consequence, becoming especially relevant today because of globalization, where new (or recontextualization/rearticulation of old) discourses have been given greater reach due to new temporalities and spatiality allowed by new information technologies. Describing the linguistic ecology of the item *sustainability*, we try to enter the social organization that surrounds the subject in the global context.

* Doutor em Estudos Linguísticos pela Universidade Federal de Minas Gerais (UFMG), Pós-Doutor trabalhando sobre a relação entre Antropologia, Linguística e Semiótica pela Universidade de São Paulo (USP). Professor Associado de Linguística e Língua Portuguesa na Universidade Federal de São João del-Rei (UFSJ), atuando na graduação e no Mestrado em Letras.

ecologia do item *sustentabilidade*, procura-se adentrar na organização social que cerca o tema no contexto global.

PALAVRAS-CHAVE: Linguística de *Corpus*. Análise Crítica do Discurso. Ecologia Linguística. Ecossistema Social. Sustentabilidade.

KEYWORDS: Corpus Linguistics. Critical Discourse Analysis. Linguistic Ecology. Social Ecosystem. Sustainability.

1. Introdução

Diversos autores, dentre os quais Biber *et al.* (1998) e Berber Sardinha (2004), enfatizam a importância da Linguística de *Corpus* por ela ser capaz de fornecer um suporte metodológico e recursos relevantes para triagem e análise de dados linguísticos, razão pela qual, quando utilizada, sobreleva e potencializa seu uso nas mais diversas áreas da Linguística.

De forma estrita, isso decorre de uma visão de linguagem como um sistema probabilístico, em que se colocam em relevo padrões e traços linguísticos recorrentes, os quais apontam para diversos níveis da organização textual, momento quando pode ser conectada à Linguística Sistêmico-Funcional, conforme Halliday (1978, 1985, 1991, 1994) e Halliday e Matthiessen (2004, 2014).

Nessa perspectiva, "os padrões de uma palavra podem ser definidos como todas as palavras e estruturas que são regularmente associadas com a palavra e que contribuem para o seu significado/sentido" (HUNSTON; FRANCIS, 2000, p. 37), ocorrendo relativamente de forma frequente. Dessa maneira, podem ser feitos mapeamentos desses padrões, os quais mostram traços linguísticos contextuais e situacionais como apontando para outros traços contextuais, que, por sua vez, podem ser associados a questões sociais e culturais relevantes numa dada conjuntura para o estudo da estrutura social e de mudanças sociais. Por isso, aqui, pode-se fazer nova associação devido aos princípios regentes da Análise Crítica do Discurso (ver, por exemplo, FAIRCLOUGH, 1989, 1992, 2006, 2010; CHOULIARAKI; FAIRCLOUGH, 1999).

Retornando à abordagem da Linguística de *Corpus*, de forma bastante específica, neste trabalho, interessa de perto o estudo da chamada *ecologia linguística*, a qual, conforme Kennedy (1998), é uma de suas quatro áreas, que se ocupa da análise de padrões lexicais de que um determinado item faz parte, tendo como objetivo descrever sentidos aos quais esse

item se associa, em quais estruturas ele aparece, qual correlação existe entre seu uso e o(s) sentido(s) a ele atribuído(s). Logo, por essa via, procura-se ter acesso ao seu valor na organização do texto.

Tendo esse pressuposto como base e trabalhando na interface entre estudos de *corpora* e Análise Crítica do Discurso (ACD), pretende-se analisar o item *sustentabilidade* em um *corpus* de textos coletados na *internet*, para procurar entender o que é central tematicamente e o que margeia os discursos a ele associados.

Entende-se, apoiando-se nos sistemicistas Martin (1992) e Eggins (1994), que as *relações lexicais* são relevantes para a compreensão de fenômenos linguísticos e, em Williams (1976), que existem itens culturalmente relevantes. Mas essa relevância também é produzida sócio-historicamente; logo, por consequência disso, hoje se torna crucial especialmente por causa do processo de globalização, em que novos (ou a recontextualização/rearticulação de velhos) discursos passaram a ter maior alcance devido às novas temporalidades e espacialidades permitidas pelas novas tecnologias de informação.

Nesse sentido, a perspectiva da ACD é extremamente importante por mostrar-se capaz de contribuir com pesquisas sociais e engajadas sobre diferentes tipos de discurso e sobre o discurso midiático dentro desse contexto (FAIRCLOUGH, 1995, 2006). Assim, com apoio na Linguística de *Corpus*, na Linguística Sistêmico-Funcional e em pressupostos da ACD de Norman Fairclough, mas com destaque para a ACD proposta por Jay Lemke, analisa-se um *corpus de pequena dimensão* (SINCLAIR, 2001), descrevendo a ecologia do item *sustentabilidade* para adentrar na organização social que o cerca no contexto global.

Nessa direção, destaca-se a importância da Linguística de *Corpus* por possibilitar o estudo das relações lexicais formadas com e a partir do item lexical *sustentabilidade* e por permitir focalizar as regularidades lexicais de forma sistemática no *corpus*, descrevendo seus contextos de uso e mostrando suas associações mais frequentes.

Estas reflexões estão organizadas nas seguintes seções, quais sejam:(1) Linguística de *Corpus*, Linguística Sistêmico-Funcional e Análise Crítica do Discurso; (2) O *corpus* e a metodologia: um percurso de pesquisa; e (3) Análise dos dados: centro e margem dos discursos sobre sustentabilidade, às quais seguem as considerações finais e as referências.

2. Linguística de *Corpus*, Linguística Sistêmico-Funcional e Análise Crítica do Discurso

Hoey (1993) explica que o desenvolvimento do computador, com cada vez mais capacidade de memória, representa para a Linguística o que o desenvolvimento do microscópio com lentes mais poderosas representou para a Biologia, justamente por permitir que se amplie o conhecimento dos padrões linguísticos e, em consequência, de seus valores quando instanciados nos textos, o que pode ser inferido também em Hunston e Francis (2000).

Esse pensamento encontra eco quando se destaca que "a análise de um *corpus* pode revelar, e frequentemente revela, fatos a respeito de uma língua, os quais nunca se tinha pensado em procurar" (KENNEDY, 1998, p. 9). Leech (1992) explica que a Linguística de *Corpus* (doravante LC) "não é somente uma nova metodologia emergente para o estudo da linguagem, mas uma nova empreitada de pesquisa, e na verdade uma nova abordagem filosófica" (p.106). Isso corrobora o pensamento de Hoey (1997) ao considerá-la, sobretudo, não exatamente um ramo da Linguística, mas uma nova rota ou um novo caminho para ela.

Hunston e Francis (2000) destacam a LC como uma área de investigação linguística que emprega computadores e *corpora* na investigação da padronização do léxico, com o objetivo de compreender a léxico-gramática que engendra os textos. Aqui, percebe-se que a noção de padronização está estritamente ligada à de recorrência, remontando à raiz probabilística que a sustenta e que a faz, por esse motivo, conectar-se à visão probabilística também presente na Linguística Sistêmico-Funcional, em conformidade com Halliday (1978, 1985, 1991, 1994) e Halliday e Matthiessen (2004, 2014), o que já estava presente nos estudos do linguista John Rupert Firth, professor de Halliday e uma de suas grandes influências (cf. FIRTH, 1957). Isso significa que, se existe um padrão de uso linguístico que terá maior probabilidade de ocorrência, isso será parte importante e integrante do processo de semantização e produção dos sentidos textuais em seus diversos níveis de organização e instanciação.

Sendo assim, uma contribuição central para este estudo advinda do pensamento de Halliday foi o fato de ele ter buscado, para a formulação de sua teoria, o pensamento antropológico, especialmente de Bronislaw Malinowski, de quem tomou os conceitos de *contexto de situação* e *contexto de cultura* a partir da leitura proposta por Firth.

Para a Linguística Sistêmico-Funcional, a interpretação do contexto social inclui a análise do *contexto de situação* (imediato) e do *contexto de cultura* (concebido pelo grupo). Este, o *contexto de cultura*, é percebido como derivado de uma rede ampla e complexa dos

gêneros dos discursos usados por uma determinada cultura sempreassociado com o *contexto de situação*. É dessa maneira quese procura justificar o valor desta abordagem, enquanto pensamento fundacional, quando se advoga a possibilidade de partir da *ecologia linguística*do item *sustentabilidade* para uma análise do *ecossistema social* que o circunda, na íntima relação que precisa ser estabelecida entre *contexto de situação* e *contexto de cultura*, a qualésempre mediada pela linguagem, do micronível (grafológico ou fonético-fonológico) ao macronível, representado pelos gêneros textuais-discursivos, em que sucessivamente se percebe o processo de instanciação das categorias linguísticas no texto.

Dessa maneira, instanciação pode ser compreendida como um processo a partir do qual uma instância é um polo da escala de instanciação e a instanciação é a própria escala que vai daquilo que é potencial no sistema para a instância textual, pois a escala se estende daquilo que é potencial para a instância num percurso que vai do sistema para o texto (MATTHIESSEN; TERUYA; LAM, 2010, p. 121).

Por outro lado, a Linguística Sistêmico-Funcional influenciou sobremaneira o projeto da chamada Linguística Crítica (cf. FOWLER *et al.* 1979), que, tendo os recursos fornecidos pela Gramática Sistêmico-Funcional, tornou-se relevante aporte teórico para o estudo da linguagem e da ideologia e constituiu uma teoria de transição para a chamada Análise Crítica do Discurso (ACD), sobretudo na proposta do linguista Norman Fairclough (cf. FAIRCLOUGH, 1989, 1992, 1995, 2010). Embora a ACD, tal qual proposta por Fairclough, seja fundacional em muitos sentidos aos quais se aventou no início deste artigo, será tomada a proposta de Jay Lemke, físico do Departamento de Educação da Universidade de Michigan, como aporte para estas considerações, sobretudo sua visão analítica sobre o *ecossistema social* ao qual será associada a descrição da *ecologia linguística* do item *sustentabilidade* no *corpus*.

Esta observação que toma como primazia a relação da linguagem em relação à sociedade constitui o cerne da Análise Crítica do Discurso e requer que os textos sejam vistos sob um viés performativo, ou seja, como forma de ação social, pois eles representam e possuem um papel constitutivo na estrutura social. E, sem embargo de todo o desenvolvimento da teoria, Lemke (2003) afirma que ainda é necessário desenvolver uma perspectiva teórica para entender o papel material e semiótico desses textos no interior dessa estrutura.

Parte-se, então, da premissa de que o convívio social também requer um aparato de análise de suas minúcias discursivas, que deem vazão à análise dos princípios regentes das relações interpessoais as quais se dão no âmbito da sociedade com todas as suas mazelas e sutilezas.

Lemke (1998) lembra que analisar a interpessoalidade nos textos é importante porque eles apresentam dimensões avaliativas que podem ser trabalhadas de forma a permitir o acesso às várias vozes que os constituem e que fazem parte de determinada comunidade que as incorpora em seus discursos, como sendo uma maneira de ter acesso a uma relação semântica complexa constitutiva do que ele chama de *formações discursivas*, termo originalmente proposto pelo filósofo francês Michel Foucault (cf. FOUCAULT, 1997).

Para essa análise, em termos de materialidade textual, o autor busca subsídios na léxico-gramática dos textos com base na perspectiva sistêmico-funcional de maneira que possa recuperar os sentidos sociais neles presentes a partir da própria interação travada nos textos a partir das vozes discursivas emanadas. Segundo Lemke (1995a), isso permite que se reconheçam as formações discursivas de uma subcomunidade específica em relação semântica com outras.

Quando Lemke (1995b) trata da visão restrita sobre os discursos ideológicos relacionados a práticas de dominação, ele pretende esclarecer que um discurso só funciona ideologicamente se as relações sociais e políticas em que os discursos são produzidos estiverem unidas na produção dos sentidos, pois a construção de sentido é orientada pelas visões de mundo e interesses sociopolíticos. Mas também há os construtos culturais mantenedores das noções, juízos e hierarquias, dentre outros, que são veiculados por meio da linguagem e de sua carga simbólica. Para o autor, existe uma interdependência entre processos de atribuição de sentido e as posições políticas e sociais que são ocupadas, razão pela qual é possível considerar todos os textos e discursos como ideológicos. Entretanto, do ponto de vista da mudança e transformação social e com base no que foi exposto, não pode haver visão única ou certa de mundo, mas múltiplas visões de mundo.

Tudo isso é o pano de fundo ideal para que o autor proponha o termo *ecossistema social* no interior de sua abordagem (ver, especialmente, LEMKE, 2003). De forma geral, dentro da biologia, ecossistema é uma comunidade de organismos que interagem entre si e com o meio ambiente ao qual pertencem. Assim, pensar a sociedade pelo viés de um ecossistema é fortalecer os laços entre comunidades e culturas que a constroem e dão sentido

às diferentes linguagens como sistemas simbólicos reconhecíveis. Assim, torna-se uma produtiva maneira de analisar os discursos e os processos sociais, pois, nesse sentido, como uma unidade funcional, a análise do ecossistema social poderá dar acesso à própria forma de organização de uma sociedade. E, aqui, aloca-se esta análise dos discursos que estão no centro e na margem do que é dito sobre *sustentabilidade*, termo que tem se tornado chave nas relações sociopolíticas do mundo atual (cf. SOARES; VIEIRA, 2013; HENRIQUES; SANT'ANA, 2013).

Na perspectiva de Lemke (2003), os significados são produzidos e interpretados de acordo com as convenções de uma comunidade. Os textos, como artefatos semiótico-materiais e culturais, podem carregar diversos significados em diferentes momentos dentro de uma comunidade, podendo contribuir em larga escala para a organização social. E as tecnologias dessa organização são mediadas por eles, que reproduzem ou tentam alterar significações em variadas escalas de tempo (curtas ou longas) e espaço, ligando-se, portanto, às questões históricas e aos princípios regentes dos processos de globalização que derrubaram as barreiras de tempo e espaço da forma tradicional como era conhecida. Os textos, então, desempenham um papel chave dentro dos processos de mudança social.

Isso tem a ver com o fato de que a textualidade na vida contemporânea ganhou novas tecnologias – hipertextos e *internet*, por exemplo –, que tanto podem colaborar para a manutenção de um ponto de vista como para alterá-lo, ligando-se às formas de controle no que se chama globalização, "um processo de integração global, definindo-se como a expansão, em escala internacional, da informação, das transações econômicas e de determinados valores políticos e morais" (SILVA; SILVA, 2010, p. 169). Sendo assim, como frisa Lemke (2003), textos e práticas sociais já existentes medeiam novas formas de controle social por novos caminhos.

3. O *corpus* e a metodologia: um percurso de pesquisa

Com o objetivo de analisar o item *sustentabilidade* em um *corpus* de textos coletados na *internet*, para procurar entender o que é central tematicamente e o que margeia os discursos a ele associados, o primeiro passo foi delimitar a forma de coleta.

Sinclair (2001, p. xi) explica que há *corpora* de pequena e de grande dimensão, sendo o primeiro constituído por um corpo de evidências relevante e confiável que precisa ser também pequeno o suficiente para ser analisado manualmente ou processado por computador

com ferramentas específicas, tendo sido projetado para intervenção humana inicial (EHI – *early human intervention*); e o segundo, de grande dimensão, projetado para intervenção humana tardia (DHI – *late or delayed human intervention*), por meio dos recursos computacionais, como a exemplo do programa *WordSmith Tools* (SCOTT, 1997, 2001, 2008).

Este trabalho foi feito com um *corpus* de pequena dimensão, para permitir a intervenção humana inicial, por isso rápida, a partir de um critério que levasse em consideração a forma de busca em sítios específicos, o que geralmente norteia as buscas na *internet* de forma geral via mecanismos e sítios como Google e UOL, dentre outros.

A partir desse pensamento, foram coletados os 100 primeiros textos listados, quando se digitava a palavra *sustentabilidade* no portal UOL, nos anos 2008 e 2010, período escolhido pelo fato de o tema ter entrado em discussões em diferentes áreas.

Nos 100 primeiros textos listados na busca do sítio da UOL, foram analisadas 167 ocorrências exatas do nódulo (não houve lematização), equivalentes a 0,35% do *corpus*, que continha 47.175 palavras. A categoria de base foram as *relações lexicais* estabelecidas com o nódulo como indicadoras do que é tido como de relevância sócio-histórica em momento de expansão da globalização (pela recontextualização/rearticulação de discursos) e, por isso, considerado como de exploração de novas temporalidades e espacialidades permitidas pelas novas tecnologias de informação. A partir disso, procurou-se focalizar a organização social que cerca o tema no contexto global.

Tomando *corpus* como um conjunto computadorizado de textos compilado e triado para efeitos de análise linguística e a visão da Linguística de *Corpus* conforme Berber Sardinha (2004), para quem ela é uma área que "se ocupa da coleta e exploração de *corpora*, ou conjuntos de dados linguísticos textuais que foram coletados criteriosamente com o propósito de servirem para a pesquisa de uma língua ou variedade linguística" (p. 3), fez-se a triagem dos dados de forma a procurar a *ecologia linguística* (cf. KENNEDY, 1998) do item *sustentabilidade*, com o uso das ferramentas oferecidas pelo programa *WordSmith Tools*.

O foco, então, foi primeiramente encontrar os padrões lexicais, para descrever sentidos a ele associados, em quais estruturas ele aparecia e, por último, a correlação existente entre o uso que se faz dele e o sentido a ele atribuído, isto é, buscou-se descrever sua *ecologia linguística*. Depois, descreveram-se as estruturas de que esse item fazia parte, bem como seu valor na organização do texto.

Quanto à categoria de acesso aos dados e expansão ao texto como unidade semântica, adotou-se o conceito de *relações lexicais* da Linguística Sistêmico-Funcional, na esteira de Eggins (1994, p. 113), em sua leitura do trabalho de Martin (1992), isto é, como forma de acesso à dimensão experiencial do discurso.

Com o auxílio das ferramentas oferecidas pelo programa *WordSmith Tools*, buscou-se, primeiramente, fazer uma lista de palavras para encontrar a frequência de uso do nódulo da pesquisa. A partir disso, foram feitas: (1) a confecção da lista de concordâncias com a palavra *sustentabilidade* para análise do contexto típico de sua ocorrência; (2) a confecção da listagem dos agrupamentos lexicais, a fim de se examinarem possíveis padrões dos quais a palavra *sustentabilidade* faça parte no *corpus*; (3) a confecção de tabelas de colocados e padrões de colocados para observação dos padrões colocacionais da palavra *sustentabilidade*; e (4) a verificação, a partir das tabelas de colocados e padrões de colocados, das principais relações lexicais e colocações, construídas com e/ou a partir da palavra *sustentabilidade*.

Os instrumentos empregados dentro do pacote *WordSmith Tools* foram: (a) a lista de palavras (*WordList*); (b) as linhas de concordância (C*oncord*); (c) a listagem de agrupamentos lexicais (*clusters*); (d) a lista de colocados (*colocates*); e (e) a tabela de padrões de colocados.

Tais procedimentos permitiram: visualizar as principais relações lexicais formadas com a palavra *sustentabilidade*; verificar, a partir das linhas de concordância, os campos a ela associados; e analisar os discursos que são centrais e os que margeiam o tema *sustentabilidade* como um aspecto discursivo da globalização, partindo das relações lexicais, dos campos a ele associados e dos gêneros em que foi veiculado.

A análise, de fulcro qualiquantitativo, obedeceu, quando da avaliação qualitativa dos dados, aos padrões da Análise Crítica do Discurso de Jay Lemke (2003). A partir dos dados, vendo o texto como unidade semântica e artefato de significação em seu todo, buscou-se: (1) esboçar um modelo complexo de sistema da semiótica mediado pelos ecossistemas sociais, discutindo os modelos gerais de texto que produzem a coerência de tais sistemas através do tempo e do espaço; (2) delinear a história das mudanças em que os textos foram mediados em diferentes formas de controle social; (3) caracterizar formas emergentes de textualidade como novas ordens de formação de significados nas experiências da sociedade contemporânea; e (4) analisar a produção de significados através da mediação dos signos ou símbolos, em diversos níveis de organização, em diferentes escalas de tempo e suas possíveis transformações pela circulação dos artefatos.

4. Análise dos dados: centro e margem dos discursos sobre *sustentabilidade*

O termo *sustentabilidade*, associado à necessidade de encontrar soluções para os problemas ligados ao desenvolvimento, numa clara relação com as assimetrias cada vez mais profundas entre pessoas, povos, países e regiões originadas pelo processo da globalização que afeta a todos, ganha maior expressividade a partir do grande *boom* que gerou o conceito de *desenvolvimento sustentável* (NATTRASS; ALTOMARE, 1999). Este, por sua vez, apareceu, pela primeira vez, em 1987, utilizado e definido pela *World Commission on Environment and Development* (WCDE – Comissão Mundial para o Ambiente e o Desenvolvimento), compreendido como uma forma de desenvolvimento que responda às necessidades do presente sem comprometer a capacidade de as gerações futuras responderem às suas próprias necessidades, conforme dito e acentuado de maneira geral na *web*.

Por isso, conforme Jacobi (2005), os problemas relacionados à *sustentabilidade* assumiram no final do século XX um papel central na reflexão em torno das dimensões do desenvolvimento e das opções que se configuram como propícias à manutenção de tudo que se mostra imperativo à sobrevivência do planeta como um todo e, por isso, um problema para o campo educacional.

A ideia de *sustentabilidade*, então, corrobora e se sustenta numa percepção de "sociedade de risco", conforme propõe Beck (1992). Assim, engloba muitos discursos, razão pela qual Henriques e Sant'Ana (2013) afirmam que "incorporado à lógica do mercado, o termo *sustentabilidade* ganha também urgência, sendo associado a questões civilizatórias" (p. 74). Ou seja, pressupõe-se um aspecto pedagógico em sua proposição que leve a população, as organizações e as empresas, dentre outros, a buscarem formas alternativas para o desenvolvimento que mantenha os recursos naturais para as gerações futuras. Assim, o trabalho desses autores demonstra os discursos sobre sustentabilidade como ideias conciliatórias entre negócios e interesses econômicos numa relação estreita com as pressões legais e sociais por responsabilidade ambiental.

Neste trabalho, parte-se do foco que assume os discursos ligados à *sustentabilidade* como ganhando cada vez mais notoriedade, uma vez que também é cada vez mais expressivo o tema em ambiente midiático. Isso traz o questionamento que faz atentar para o fato de haver diferentes relacionamentos e enfoques do tema e perguntar: que tipo de relação lexical estabelece ou ajuda no estabelecimento de sua significação, de conexão discursiva ou de visões de mundo em escala global?

Nesse sentido, muitos trabalhos foram produzidos nos domínios da comunicação organizacional (HENRIQUES; SANT'ANA, 2013), das políticas culturais e das transformações sociais, sob o viés da Administração (CAMINHA, 2013), e da própria Análise Crítica do Discurso, considerando a multimodalidade (RAFAEL, 2013; SOARES; VIEIRA, 2013), a gestão e o *marketing* ambiental (DIAS, 2009, 2011), dentre outros.

Por isso, pode-se afirmar que este trabalho se sustenta numa perspectiva que Lemke (2003) denomina *traversal* (transversal, que atravessa), que se define pela liberdade para misturar o que é diferente, criando as hibridizações de forma interdisciplinar nas teorias e suas possíveis aplicações, gerando, talvez, a transdisciplinaridade. Isso caminha para o que ele almeja ao afirmar que a Análise Crítica do Discurso precisa sempre encontrar caminhos novos para o estudo das relações entre os discursos estandardizados, gêneros e formas textuais, e suas práticas institucionais, e para os estudos que levam em consideração os interesses que os sustentam, os modos emergentes de mediação, juntamente com suas implicações no controle social e nas possibilidades de mudança nas relações entre os indivíduos.

Procurando descrever a *ecologia linguística* do item lexical *sustentabilidade*, após verificação de sua frequência, buscaram-se seus padrões de uso levando em consideração, em conformidade com Stubbs (2002), que esses padrões são "esquemas semânticos" e "esses esquemas semânticos podem ser modelados como agrupamentos lexicais (nódulos e colocados), gramaticais (coligação), semânticos (preferências por palavras de um campo lexical particular) e pragmáticos (conotações ou prosódias discursivas)" (p. 96). Todas as ocorrências foram avaliadas qualitativamente a partir das linhas de concordância que foram ampliadas o quanto foi necessário para a análise. A Tabela a seguir ilustra as linhas de concordância:

Tabela 1– Linhas de concordância com a palavra *sustentabilidade*.

N	Concordance
1	Votos: Tags: "planeta em equilíbrio" ***sustentabilidade*** "lixo eletrônico" computadores placas
2	. 20/08/2010. 10:29. Bienal Brasileira de ***Sustentabilidade*** e design se encontram em bienal.
3	de 2009 11:11 TI VERDE Bandeira da ***sustentabilidade*** Silas Scalione - Estado de Minas O resi
4	Outros Casatual 2009 terá como temas a ***sustentabilidade*** e a ousadia Fechar [x] Fechar [x] Parte
5	ar a Copa de 2014, quanto para garantir a ***sustentabilidade*** econômica e social do Conjunto após o
6	s, nos dias 29 e 30 de outubro, no Fórum ***Sustentabilidade*** Fingerplus. Os projetos vão participar,
7	e garantir a valorização do trabalho com ***sustentabilidade*** socioambiental. Para a Central, a classe
8	m e propagam a urgente e tão perseguida ***sustentabilidade***. O Axé Brasil deste ano também aderiu
9	to. A primeira é a Think green, focada na ***sustentabilidade***. A ideia é reduzir o impacto ambiental
10	ra apresentar a palestra "Cidade, cultura e ***sustentabilidade***". O tema dá continuidade aos assuntos

Tirando as palavras gramaticais como *que*, *com*, *sobre* e *para*, dentre outras, o item *sustentabilidade* possui como colocados os seguintes itens com suas frequências entre parênteses: *tema* (14), *fechar* (12), *Brasil* (10), *meio* (10), *ambiental* (7), *ambiente* (7), *conceitos* (7), *economia* (6), *agência* (5), *econômica* (5) e *preservação* (5).

Seus agrupamentos lexicais (*clusters*) são *conceitos de sustentabilidade* (5), *a sustentabilidade não* (4), *para a sustentabilidade* (4), *em prol da sustentabilidade* (4), *sobre sustentabilidade* (4), *sustentabilidade ambiental* (3), *a sustentabilidade do* (3), *ambiente e sustentabilidade* (3), *ações de sustentabilidade* (3), *conceito de sustentabilidade* (3), *de sustentabilidade* (3), *que a sustentabilidade* (3), *sobre a sustentabilidade* (3) e *sustentabilidade e preservação* (3).

Os padrões do nódulo *sustentabilidade* são formados, tirando-se as palavras gramaticais, num horizonte de cinco palavras à esquerda [L5, L4, L3, L2 e L1] e cinco palavras à direita [R1, R2, R3, R4 e R5], conforme padrão do *WordSmith Tools*, mantendo-se o nódulo no meio, com os itens e suas posições seguidas do número de ocorrências: *tema* (L4, L2, L1 e R1, uma vez cada; R2 e R4, duas vezes cada), *Brasil* (L5, L4, L3, R2, R4 e R5, uma vez cada), *meio* (L3, L2, R2, R4 e R5, uma vez cada), *conceitos* (L4, L2 e R5, uma vez cada), *ambiental* (L5, L2 e R1, uma vez cada), *ambiente* (L5, L2, L1 e R5, uma vez cada), *economia* (L4, L5, R2, R3 e R5, uma vez cada), *econômica* (L2, R1 e R3, uma vez cada), *agência* (R1, R3 e R4 uma vez cada) e *preservação* (R2 e R4, uma vez cada).

Essa padronização demonstra uma conotação sempre positiva para o item *sustentabilidade* por uma associação predominantemente ligada a desenvolvimento sustentável seja no âmbito do crescimento econômico, seja no ambiental. Os padrões também denotam força coesiva colocacional, uma vez que, mesmo não sendo utilizadas imediatamente antes ou após o item *sustentabilidade*, ocorrem com proximidade no mesmo ambiente. Para essa afirmação, ancorou-se na explanação de Halliday e Hasan (1976, p. 288) ao explicarem que o efeito da coesão lexical por colocação é sutil e difícil de estimar, pois todos os itens lexicais podem entrar em relação coesiva, mas não carregam em si nenhuma indicação se estão funcionando coesivamente ou não.

Isso tem implicação nos campos associados: economia, verificados no uso conjunto com os itens *economia*, *econômico* e *agência*; meio ambiente, a partir dos itens *ambiental*, *ambiente*, *preservação* e *meio* (indicando lugar, posição); e divulgação, evocando a mídia e outras formas de propagação de informação, nos itens *tema* e *conceitos* e no uso do item

Brasil como país que se insere no conjunto global para produção mundial sustentável. Essa análise pode ser verificada nos exemplos a seguir:

1) Os brasilienses que compareceram ao Eixão do Lazer neste domingo puderam aprender um pouco mais sobre *sustentabilidade* e preservação do meio ambiente. Esse era o tema do evento, promovido pela Administração Regional de Brasília.
2) 'Plano MTV' destacará meio ambiente e *sustentabilidade*.
3) Economia brasileira demonstra firmeza e *sustentabilidade*, afirma Lula.
4) Dr. Ladislau Dowbor discute a 'Perspectiva Econômica da *Sustentabilidade'*
5) A Perspectiva Econômica da *Sustentabilidade* será tema do quarto encontro do Fórum de Investidores Sociais 2009.
6) No Ano Internacional da Biodiversidade, Estação Pátio Savassi discute *sustentabilidade*.
7) *Sustentabilidade*– você sabe o que isso significa? O prefixo vem de sustentar: conservar, manter, alimentar física ou moralmente, de conseguir dar sustento a alguma coisa ou alguém.
8) Quando a gente entende que *sustentabilidade* é cumprir o ciclo da vida, as coisas ficam mais claras.
9) Bienal Brasileira de Design, que será realizada em Curitiba, tem como tema a *sustentabilidade*.

Entretanto, é importante verificar que, numa leitura estendida das linhas de concordância, o espectro relacional de sustentabilidade do ponto de vista dos temas é bem maior, associando-se também com lazer, *design*, eventos culturais, lançamentos de livros, *shows* musicais, arquitetura, turismo, urbanismo, saúde, esporte, publicidade, gastronomia, carnaval, tecnologia, arte e política.

Isso demonstra que existem temas centrais conectados à sustentabilidade, como os da economia e do meio ambiente; e outros que ficam à margem, como os imediatamente já citados. Dentre estes, era perceptível no *corpus*, de forma recorrente, mas periférica, seu uso dentro de um discurso promocional seja ligado a lazer, arte, arquitetura, turismo ou promoção de eventos, remontando à sua conotação positiva e capaz de positivar ações e eventos e de promover credibilidade:

10) *Sustentabilidade* e design se encontram em Bienal
11) Balneário Camboriú recebe evento a favor da *sustentabilidade*
12) *Sustentabilidade* é tema de Mostra Cultural do Colégio Sagrado Coração de Maria
13) Nova campanha da Coca-Cola aborda a *sustentabilidade*

14) *Sustentabilidade* é tema da Morar Mais Brasília 2009
15) Diferentemente de todos os estádios brasileiros já construídos, o Complexo Mineirão está implantado numa região que dispõe de fartas áreas livres para abrigar todas as funções esportivas e outras complementares, exigidas tanto para sediar a Copa de 2014, quanto para garantir a *sustentabilidade* econômica e social do Conjunto após o evento.
16) Profissionais aderem cada vez mais à *sustentabilidade* na hora de fazer seus projetos

Numa análise dos textos do ponto de vista dos gêneros textuais-discursivos, percebe-se que os gêneros utilizados quando o tema da sustentabilidade estava presente eram *notícia* (38), *reportagem* (33), *nota* (13), *propaganda* (15) e *artigo-verbete* (1).

É fundamental frisar que o objetivo não era delinear gêneros, por isso tomou-se o conceito de gênero de Fairclough (1992), para quem gênero é "um conjunto relativamente estável que é associado com, e parcialmente representa, um tipo de atividade socialmente aprovado, como a conversa informal, comprar produtos em uma loja, uma entrevista de emprego, um documentário de televisão, um poema ou um artigo científico"(p. 126).

Quanto ao ponto de vista classificacional, trabalhou-se a partir de uma perspectiva de profundidade temática, uma vez que a reportagem é um gênero que se caracteriza pela prioridade informativa em sua constituição, cujo propósito básico é prover o leitor de uma descrição objetiva e, às vezes, de uma interpretação dos fatos, como demonstra Martins (2004). E, partindo de uma escala, a notícia deveria prover o leitor de informação, apenas com a prioridade informativa, sendo, por isso, menor que a reportagem. A nota, por sua vez, é um texto bem mais curto que os anteriores, noticiando algo de forma mais particular.

O gênero propaganda foi tomado em seu sentido comum, isto é, um texto cujo objetivo é a propagação de algo (eventos e acontecimentos importantes, publicação de livros, lançamentos de CDs etc.), em que se destacam os caráteres publicitário, conativo e/ou persuasivo.

Por fim, o que foi denominado *artigo-verbete* partiu do conceito de artigo como sendo um gênero em que o autor – jornalista ou colaborador do jornal – assina o texto de caráter argumentativo e expõe seu ponto de vista, como destaca Cunha (2002), mas com o diferencial do verbete, porque tinha caráter metalinguístico, ao ter como motivação conceituar sustentabilidade, e uma linguagem bastante informal, coloquial e dialogal, como se verifica no trecho a seguir:

17) *Sustentabilidade*– você sabe o que isso significa? O prefixo vem de sustentar: conservar, manter, alimentar física ou moralmente, de conseguir dar sustento a alguma coisa ou alguém. É o que parece para a maioria das pessoas. Mas, na verdade, essa palavrinha tão em moda, que vive na boca de empresários, políticos e ativistas ambientais, ainda carece de mais explicação. A *sustentabilidade* do planeta Terra.

Pelos gêneros, percebe-se que o tema não estava sendo abordado com a profundidade que, tanto do ponto de vista teórico quanto sociopolítico, deveria, uma vez que seu impacto recai sobre diversos campos da sociedade, associando-se central ou marginalmente. Ou seja, à exceção do gênero artigo-verbete, pela preocupação conceitual e informalidade que poderia facilitar a compreensão do sentido do termo, partia-se do ponto de vista de que o tema era de conhecimento amplo e notório, pois notícias, reportagens e notas não são gêneros para adensamento de discussão.

O valor positivo do item *sustentabilidade* encontra eco especialmente no gênero propaganda, por ser um tema apelativo a uma sociedade com inúmeros problemas que tendem a se agravar, sendo que um desenvolvimento sustentável parece ser a melhor solução; logo, associações várias tendem a ser promovidas.

O uso de apenas um artigo com preocupação conceitual e assinado pode:(1) demonstrar essa análise e a existência de um pretenso conhecimento do assunto; e (2) apontar para uma estratégia de esquiva de qualquer veículo quanto à responsabilidade por um conceito e também pelas várias relações conflitivas que podem ser desenvolvidas ao seu redor.

Dessa maneira, percebe-se um engendramento discursivo que se mostra harmonioso nos textos e que se delineia a partir dos padrões linguísticos da palavra *sustentabilidade*, instanciados nos textos e que poderiam ser ampliados pela noção de gêneros como uma forma de ação social ou mesmo pelo adensamento da discussão a seu respeito.

É relevante verificar que tudo o que foi dito encontra suporte numa conjuntura de globalização em que os diversos países tornam-se partícipes da construção de um mundo melhor e, por isso e para isso, sustentável. Como definem Chouliaraki e Fairclough (1999), conjunturas são "conjuntos relativamente estáveis de pessoas, materiais, tecnologias e práticas – em seu aspecto de permanência relativa – em torno de projetos específicos" (p. 22). Sustentabilidade pode ser vista aqui como esse projeto específico de produção de um mundo sustentável, pois as barreiras temporais e espaciais não mais existem por causa da globalização que se expande por todos os campos da sociedade global em que se vive, cujos

impactos ambiental e econômico não dizem mais respeito ao universo das particularidades, pois se construiu um grande ecossistema social.

E aqui fica em evidência uma possibilidade de análise do contexto social, em termos sistêmico-funcionais, dividido no seu *contexto de situação* e no *contexto de cultura*, partindo da *ecologia linguística* do item *sustentabilidade*, ampliada para o texto visto como forma de ação social; por isso um gênero textual-discursivo, como apontando para uma cultura em transformação numa dimensão superior local e temporalmente, ao ligar-se a uma visão global dos problemas a serem enfrentados futuramente. E fica destacado, sobretudo, o papel de cada país do ponto de vista econômico, chamando-se atenção aqui para o Brasil por ter uma economia em expansão e para um desenvolvimento sustentável, talvez como a solução para o problema da sustentabilidade no contexto global.

A Análise Crítica do Discurso demonstrou sua produtividade quando o contexto social pôde ser teorizado e avaliado partindo das *relações lexicais* do item *sustentabilidade* como demonstrativo de um complexo sistema semiótico mediado por *ecossistemas sociais*, que foram abalados em função da dimensão globalizada que envolve o tema; logo, transformando os problemas relacionados à sustentabilidade num problema global.

Esta ideia da ligação entre *ecologia linguística* e *ecossistema social* é apontada em termos da coesão que o item dá em termos colocacionais ao unir diferentes temas, seja central, seja marginalmente, trazendo coerência a um sistema complexo de relações temáticas através do tempo e do espaço numa conexão com o processo de globalização.

Nesse sentido, reescreve-se a história a partir de (e já apontando para outras) mudanças em que os textos foram mediadores em diferentes formas de controle social, especialmente nas esferas econômica e ambiental. Há, pelos gêneros utilizados, uma forma mais padrão de textualidade sendo utilizada para propagação do tema, mas não formas profundas. No entanto, o *artigo-verbete* aponta para formas emergentes de textualidade que podem ser vistas como novas ordens de formação de significados nas experiências da sociedade contemporânea, pois alia um gênero já consolidado, como o *artigo*, ao gênero *verbete*, mas produzido com linguagem despojada, coloquial e dialogal, com objetivos tanto metalinguísticos quanto pedagógicos, mas superando os modos mais comumente usados pela educação formal para a qual o verbete de dicionários e enciclopédias normalmente são produzidos. E isso corrobora para que diferentes pessoas tomem contato com um conceito e um problema consideravelmente novo que impacta de forma geral na vida de todos.

Por outro lado e por fim, de forma geral, este tipo de análise também pode dar acesso, mesmo que de forma limitada, à produção de significados através da mediação dos signos ou símbolos, em diversos níveis de organização, em diferentes escalas de tempo e suas possíveis transformações pela circulação dos artefatos dentro do processo de instanciação do texto, no momento em que une Linguística de *Corpus*, para análise da *ecologia linguística* das palavras culturalmente relevantes; Linguística Sistêmico-Funcional, para permitir a teorização desses processos de instanciação das palavras no texto e também do contexto social (*contexto de situação* e *contexto de cultura*); e a Análise Crítica do Discurso, por prover importantes ferramentas para expansão, análise e aferição do papel da linguagem na sociedade, aqui demonstrado com especial destaque quando da avaliação do *ecossistema social* que envolve o tema da sustentabilidade numa dimensão global.

5. Considerações finais

Como se pôde perceber durante o percurso analítico, aliar os recursos quantitativos da Linguística de *Corpus* numa análise qualitativa em muito auxilia o analista a verificar hipóteses, antever temas a serem abordados ou como auxílio na triagem dos dados.

A *ecologia linguística* da palavra *sustentabilidade* no *corpus* aponta de forma contundente para os problemas relacionados a uma economia em expansão, no caso do Brasil, ou de manutenção de uma economia no mundo, a partir de um desenvolvimento sustentável, mas ligado à reserva e preservação dos recursos naturais para subsistência de tudo e de todos, por isso sua ligação com o meio ambiente. Esse é o ponto que conecta tematicamente os discursos tanto centrais, como os dois citados, quanto os marginais, como turismo, arquitetura, eventos e arte, dentre outros, que possuem relação com a *sustentabilidade*.

Economia, informação, administração, urbanismo e turismo, dentre diversos campos, se encontram com o tema sustentabilidade quando se trata de promover um mundo habitável e saudável, mas nunca se esquecendo da questão econômica e de seu crescimento que está ligado às relações internacionais de poder; logo, às várias tensões que ganham voz por meio dos discursos engendrados na sociedade e propagados transnacionalmente, sobretudo pela *internet*, em época de globalização. Isso aponta para um ecossistema social que só pode ser compreendido quando as barreiras de temporalidade e espacialidade são derrubadas, pois o

tema da *sustentabilidade* ecoa para além de qualquer forma de territorialização, tendo impacto em inúmeras esferas da vida humana.

A *ecologia linguística* do item *sustentabilidade* mostrou-se, então, produtiva para se alçar, qualitativamente sob a égide da Análise Crítica do Discurso sobre o ecossistema social transnacional que permeia o tema da sustentabilidade, um discurso de raízes profundas em diferentes campos da vida social, impactando diretamente desde uma perspectiva micro à macro de múltiplas iniciativas em torno de um problema global.

Sendo assim, conforme proposto, a partir da *ecologia linguística* do item *sustentabilidade*, procurou-se chegar ao ecossistema social como mediador de um modelo complexo de sistema semiótico que produz sua coerência, derrubando tempo e espaço por meio da globalização e das novas tecnologias de informação.

Ele também delineia uma história de mudanças especialmente porque os textos passaram a mediar formas de controle social numa escala muito superior devido à emergência de um tipo de textualidade que exibe novas ordens de formação de significados nas experiências da sociedade contemporânea. Está-se diante de outra ordem de produção de significados através da mediação dos signos ou símbolos, um nível de organização superior àqueles indicados pelas territorialidades, ou seja, em diferentes escalas de tempo e espaço, mas ainda em transformação pela circulação dos textos vistos como artefatos semiótico-materiais e culturais na organização do grande ecossistema social em que o mundo se tornou.

Referências

BECK, U. **Risk Society**. Beverly Hills: Sage, 1992.

BERBERSARDINHA, T. **Linguística de Corpus**. Barueri, SP: Manole, 2004.

BIBER, D. *et al.* **Corpus Linguistics**: investigating language structure and use. Cambridge: Cambridge University Press, 1998.

CAMINHA, D. O. Políticas culturais e transformação social: um estudo crítico do discurso do Ministério da Cultura do Brasil no início do século XXI. In: ANPAD, 37.,2013,Rio de Janeiro. **Anais...**Rio de Janeiro, 2013. p. 1-16.

CHOULIARAKI, L.; FAIRCLOUGH, F. **Discourse in late modernity**: rethinking Critical Discourse Analysis. Edinburgh: Edinburgh University Press, 1999.

CUNHA, D. A. C. O funcionamento dialógico em notícias e artigos de opinião. In: DIONÍSIO, A. P.; MACHADO, A. R.; BEZERRA, M. A. (Org.).**Gêneros textuais e ensino**. Rio de Janeiro: Lucerna, 2002. p. 166-179.

DIAS, R. **Gestão ambiental:** responsabilidade social e sustentabilidade. São Paulo: Atlas, 2009.

DIAS, R. **Marketing ambiental:** ética, responsabilidade social e competitividade nos negócios. São Paulo: Atlas, 2011.

EGGINS, S. **An introduction to systemic functional linguistics**. London: Pinter, 1994.

FAIRCLOUGH, N. **Language and power**. London and New York: Longman, 1989.

FAIRCLOUGH, N. **Discourse and social change**. Cambridge: Polity Press, 1992

FAIRCLOUGH, N. **Media discourse**. London: Longman, 1995.

FAIRCLOUGH, N. **Language and globalization**. London: Routledge, 2006.

FAIRCLOUGH, N. **Critical discourse analysis**: the critical study of language. 2nded. Longman: Pearson Education, [1995] 2010.

FIRTH, J. R. **Papers on linguistics, 1934-1945**. Oxford: Oxford University Press, 1957.

FOUCAULT, M. **A arqueologia do saber**. Rio de Janeiro: Forense Universitária, 1997.

FOWLER, R. *et al.* **Language and control.** London, Boston and Henley: Routledge & Kegan Paul, 1979.

HALLIDAY, M. A. K. **Language as social semiotic**: the social interpretation of language and meaning. Australia: Edward Arnold, 1978.

HALLIDAY, M.A.K. **An Introduction to Functional Grammar**. London: Edward Arnold, 1985.

HALLIDAY, M. A. K. *Corpus* studies and probabilistic grammar. In: AIJMER, K.; ALTENBERG, B. (Org.). **English corpus linguistics**: Studies in honour of Jan Svartvik. London: Longman, 1991. p. 30-43

HALLIDAY, M. A. K. **An introduction to Functional Grammar**. London: Edward Arnold, 1994.

HALLIDAY, M. A. K.; HASAN, R. **Cohesion in English**. London: Longman, 1976.

HALLIDAY, M. A. K.; MATTHIESSEN, C. **An introduction to Functional Grammar**. London: Edward Arnold, 2004.

HALLIDAY, M. A. K.; MATTHIESSEN, C. **An introduction to Functional Grammar**. London: Edward Arnold, 2014.

HENRIQUES, M. S.; SANT'ANA, L. F. Ideias-força evidenciadas no discurso organizacional sobre sustentabilidade. **Organicom**, São Paulo: USP, v. 10, p. 71-82, 2013.

HOEY, M. (Ed.). **Data, description, discourse** – papers on the English language in honour of John McH Sinclair on his sixtieth birthday. London: Harper Collins, 1993.

HOEY, M. From concordance to text structure: new uses for computer *corpora*. In: LEWANDOSWKA-TOMASZCZYK, B.; MELIA, P. J. (Org.). **PALC'97** – Practical Applications in Language Corpora. Lodz: Lodz University Press, 1997. p. 2-22.

HUNSTON, S.; FRANCIS, G. **Pattern grammar**: a *corpus*-driven approach to the lexical grammar of English. Amsterdam: John Benjamins Publishing Company, 2000.

JACOBI, P. R. Educação ambiental: o desafio da construção de um pensamento crítico, complexo e reflexivo. **Revista Educação e Pesquisa FE-USP**, São Paulo, v. 31, n. 2, p. 302-313, maio/ago. 2005.

KENNEDY, G. **An introduction to corpus linguistics**. New York: Longman, 1998.

LEECH, G. *Corpora* and theories of linguistic performance. In: SVARTVIK, J. (Org.). **Directions in Corpus Linguistics**: proceedings of Nobel Symposium 82. Berlin, New York: De Gruyter, 1992. p. 105-127.

LEMKE, J. L. Intertextuality and Text Semantics. In: FRIES, P. H.; GREGORY, Michael (Ed.). **Discourse in Society***:* Systemic Functional Perspectives – Meaning and Choice in Language: Studies for Michael Halliday, NJ: Ablex Publishing Corp., 1995a. p. 85-114.

LEMKE, J. L. **Textual politics**: discourse and social dynamics. London: Taylor & Frances, 1995b.

LEMKE, J. L. Resources for attitudinal meaning – Evaluative orientations in text semantics. **Functions of Language**, v. 5, n. 1, p. 33-56, 1998.

LEMKE, J. L. Texts and discourses in the technologies of social organization. In: WEISS, G; WODAK, R. **Critical discourse analysis**: theory and interdisciplinary. New York: Palgrave Macmillan Ltd., 2003. p. 130-149.

MARTIN, J. R. ***English Text:*** *system and structure*. Philadelphia/Amsterdam: John Benjamins, 1992.

MARTINS, A. R. N. **A polêmica construída**: racismo e discurso da imprensa sobre a política de cotas para negros. 2004. 201 f.(Doutorado em Linguística)-Departamento de Linguística, Línguas Clássicas e Vernácula, Universidade de Brasília, Brasília, 2004.

MATTHIESSEN, C. M. I. M.; TERUYA, K; LAM, M. **Key terms in systemic functional linguistics.** London and New York: Continuum, 2010.

NATTRASS, B.; ALTOMARE, M. **The natural step for business** – wealth, ecology and evolutionary corporation. Gabriola Island: New Society Publishers, 1999.

RAFAEL, R. R. **Marketing verde**– uma análise multimodal da construção do discurso da *sustentabilidade* em campanhas publicitárias empresariais. 2013. 155 f.(Mestrado em Linguística)-Departamento de Linguística, Línguas Clássicas e Vernácula, Universidade de Brasília, Brasília, 2013.

SCOTT, M. R. PC analysis of key words – and key words. **System**, Great Britain, v. 25, n. 2, p. 233-245, 1997.

SCOTT, M. R. Comparing *corpora* and identifying key words, collocations, frequency distributions through the WordSmith Tools suite of computer programs. In: GHADESSSY, M.; ROSEBERRY, A. H. R. L. **Small corpus studies and ELT**: theory and practice. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2001. p. 47-67.

SCOTT, M. R. **Wordsmith Tools v. 5.** Software for text analysis. Oxford: Oxford University Press, 2008.

SILVA, K. V.; SILVA, M. H. **Dicionário de conceitos históricos**. São Paulo: Contexto, [2005] 2010.

SINCLAIR, J. M. Preface. In: GHADESSSY, M.; ROSEBERRY, A. H. R. L. **Small corpus studies and ELT**: theory and practice. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2001. p. VII-XV.

SOARES, N. M. M.; VIEIRA, J. A. Representação multimodal dos atores sociais no discurso de marcas. **Signum**: Estudos da Linguagem, v. 16, p. 233-258, 2013.

STUBBS, M. **Words and phrases**: *corpus* studies of lexical semantics. Oxford, UK: Blackwell Publishing, 2002.

WILLIAMS, R. **Keywords***:* a vocabulary of culture and society. London: Fontana, 1976.

Artigo recebido em: 15.10.2014
Artigo aprovado em: 09.12.2014

Entrevista

Conversaciones con un lingüista de corpus: Profesor Dr. Giovanni Parodi

por Ariel Novodvorski* y Ana Fritz Herrera**

Giovanni Parodi es actualmente Director de los Programas de Postgrado en Lingüística en la Pontificia Universidad Católica de Valparaíso, Chile, y también es Editor en Jefe de la *Revista Signos. Estudios de Lingüística*. Al mismo tiempo, es Director para Chile de la Cátedra UNESCO en Lectura y Escritura. Sus principales áreas de interés científico son la lingüística de corpus, la psicolingüística del discurso y la lingüística del texto. En la actualidad dirige proyectos de investigación en el discurso escrito académico y profesional con especial atención a los géneros multimodales especializados. El foco central de sus investigaciones está en profundizar en el conocimiento de los procesos textuales y psicolingüísticos implicados en la comprensión y producción de textos escritos en diversos géneros discursivos y en la construcción de conocimientos disciplinares y llevar estos hallazgos científicos a contextos educativos de enseñanza-aprendizaje en diversos niveles y dominios especializados con apoyo de tecnologías informáticas. Proyectos tales como El Grial y LECTES son programas asistidos por computadores tanto para la investigación básica como para el desarrollo de estrategias de lectura y escritura de alumnos de diversos niveles educacionales. Sus publicaciones incluyen más de 50 artículos en revistas en español e inglés, 19 libros (como autor único, coautor o editor) y diversos capítulos de libros en español e inglés. Sus últimos libros son: *Comprensión de Textos Escritos. La Teoría de la Comunicabilidad* (2014), *Academic and professional discourse genres in Spanish*, *Lingüística de Corpus: de la teoría a la empiria*, *Saber Leer*, *Alfabetización académica y profesional en el siglo XXI: leer y escribir en las disciplinas*. Por sus méritos académicos y logros científicos, en el año 2008 fue nombrado Miembro Correspondiente de la Academia Chilena de la Lengua. El sitio en Internet del profesor Parodi es http://www.giovanniparodi.cl/.

* Doutor pela UFMG. Professor adjunto do ILEEL/UFU.
** Mestranda do PPGEL/UFU.

1. Profesor Parodi, ¿cuáles han sido sus actuaciones más destacadas como lingüista y, más específicamente, como lingüista de corpus?

R: Pensando como lingüista y como lingüista de corpus, y haciendo la conexión entre ambos, hay un asunto que me parece muy interesante y es pensar en el momento en que comenzamos a construir diferentes corpus, a construir bases de datos de textos escritos. Yo venía en ese entonces, y aún es una de mis áreas de investigación, desde la psicolingüística. Esto hace ya unos quince años atrás, y una de las primeras ideas que tuvimos fue tener un etiquetador morfosintáctico. Yo miraba con envidia (sana) los trabajos en inglés y cómo marcaban los textos. Para lograr este propósito, entonces, necesitábamos contar con grupos de documentos en formato *TXT*. Y en ese momento hubo una cosa que siempre comento, pues luego tendría gran impacto en un paso que daríamos años después. En ese momento contar con un gran corpus era un gran paso. Y decíamos, "Bueno, digitalicemos los textos y borremos de los textos todo lo que no sea sólo las palabras", y yo siempre insistía, "Dejemos las palabras; así les decía a mis asistentes, porque lo que queremos son las palabras. Solo denme las palabras". Y en ese momento lo que buscábamos obviamente era tener textos que fueran en formato plano para poder aplicarles algún tipo de etiquetador, y ese fue un desafío tremendo, aprendimos todos a digitalizar y vimos todo lo difícil que era llevar adelante esa tarea, y luego surgió la idea de: ¿y qué hacíamos con estos documentos digitales? ¿Dónde poníamos estos documentos? Y, entonces, tanto la idea de desarrollar un etiquetador morfosintáctico y la necesidad de un espacio de almacenaje y consulta, nos llevaron a un hito muy importante como fue la creación del sitio web www.elgrial.cl.

Debo reconocer, sin falsa modestia, que en mi carrera profesional y académica existen varios hitos y momentos maravillosos. Uno de ellos que marca un antes y un después, ciertamente lo constituye el momento en que lanzamos El Grial. Fue un trabajo mancomunado con muchos colegas y grupos de investigación. Por ello, cuando miro hacia atrás y observo las cosas que hemos realizado y veo a toda la gente que se ha unido y con la que hemos trabajado y seguimos juntos en torno a El grial y sus proyecciones; entonces veo que construir El Grial y desarrollar el sitio web donde se alojan todos estos corpus, solo puedo decir que estoy seguro ha sido un desafío. Ha sido algo muy potente para desarrollar esta línea de investigación. Así, logramos tener un etiquetador, y luego construir la interface. No tan sólo un sitio donde colgáramos los textos y tuviéramos un etiquetador, sino que construir una interface que fuera realmente una plataforma que permitiera visualizar los corpora que están

allí y permitiera también hacer consultas, recuperar la información que está alojada ahí, y luego ir avanzando en establecer diferentes tipos de consultas. Bueno, eso ha sido un trabajo largo en varias etapas y varias fases en que han intervenido diferentes equipos de trabajo. Y, a veces, hay quienes dicen, pero bueno, no necesitamos tener un etiquetador morfosintáctico, podríamos hacer análisis basados en otros elementos. En fin, pero yo sigo pensando que el contar con un etiquetador morfosintáctico es necesario.

Lo que nosotros echábamos de menos era una herramienta para poder trabajar con el español, y una herramienta que, además, como hemos visto en las conversaciones en estos días de congresos aquí en Uberlândia, existe consenso en esta necesidad, que es compartida con los investigadores aquí en Brasil, lo que buscamos son herramientas disponibles gratuitamente. Por ello, desde la psicolingüística es muy valioso contar con investigación empírica desde la LC. Yo había sido un psicolingüista que estaba fuertemente concentrado en los procesos de producción y comprensión de textos y, no había visualizado hasta ese momento, como comentaba ayer en el marco del congreso, el estudio de los corpus, y gracias a Douglas Biber, debo reconocer, al descubrir el libro de Biber de 1988, visualizo ahí un campo de acción muy diferente. Creo que eso ha fortalecido la investigación en que hacemos psicolingüística, porque podemos investigar fenómenos que estén mucho mejor documentados, fenómenos que han sido indagadas descriptivamente en textos y luego podamos generar entonces instrumentos que tienen un basamento empírico muy importante que permiten hacer investigación psicolingüística. Este círculo virtuoso es muy robusta, en mi opinión, para una línea de investigaciones interdisciplinarias. Entonces, desde la psicolingüística, hemos ido a la lingüística de corpus y hemos vuelto a la psicolingüística.

2. Como miembro de la Sociedad Chilena de Lingüística y de la Academia Chilena de la Lengua, cómo ve su participación desde el área de la lingüística de corpus en el ámbito de las organizaciones ya mencionadas. ¿Cómo lo ven sus pares? ¿Cuáles son sus aportes más específicos en esta área?

R: Es difícil saber cómo lo ven los otros a uno, aunque algunos dicen que nosotros nos conocemos a partir de cómo los otros nos dicen que somos; entonces, cómo se construye el imaginario de uno mismo, a partir de lo que los otros nos dicen o uno mismo tiene su propio yo, digamos, ¿mi yo cuánto se alimenta del otro? Bueno, eso es casi filosófico.

Tal vez es más fácil comentar algunas acciones que hemos emprendido al interior de la Sociedad Chilena de Lingüística. Una cosa que hemos ido haciendo, a la luz de lo que recién conversábamos, es ir creando espacios para la Lingüística de Corpus (LC) y los estudio en esta línea de investigación, de modo que muy rápidamente fuimos proponiendo mesas de trabajo, o talleres, o seminarios, que fueran complementando los temas que se trataban clásicamente, la Sociedad Chilena de Lingüística, en general es una sociedad que ha estado, digamos, siempre ha tenido un área de psicolingüística, un área de gramática, un área de discurso, un área de análisis del discurso (más nueva por supuesto), pero obviamente no tenía un área de lingüística de corpus y yo creo que cuando uno mira el programa del congreso bianual de la Sociedad Chilena de Lingüística, cuya última reunión fue el año pasado en Concepción, uno observa un cambio en las temáticas en la forma de hacer investigación, yo creo que ahí aparece el rol de la LC, y yo creo que está gatillado también, y da mucho gusto, uno lo dice con orgullo, ver que muchas de las personas que están allí hoy en día trabajando son ex alumnos de nuestros programas de pregrado y de nuestros postgrados de magister y de doctorado. Entonces uno ve como una idea que parece tan pequeñita en un comienzo va creciendo y creciendo, y hoy en día esas líneas de investigación son fuertes.

Ahora bien, dentro de la Academia Chilena de la Lengua, existe la Comisión de Lexicografía, y en su interior se ha ido fortaleciendo el trabajo a partir de corpus y de herramientas tecnológicas. Así, progresivamente se ha impulsado la mirada desde los corpus para las revisiones que se hacen. Como seguramente ustedes saben, hoy en día el diccionario, y la gramática y la ortografía, se construyen desde las academias, de todas las academias de la lengua española. Ha habido un cambio aquí muy importante y estos cambios se han debido básicamente a la creación de la Asociación de Academias de la Lengua (ASALE) de modo que, la Nueva Gramática, la Ortografía y el Diccionario de la Lengua en las últimas versiones que tienen, no son ahora de la RAE sino que son de la Asociación de Academias de la Lengua, que son veintidós, donde la RAE es una más que se sienta en igualdad de términos a trabajar en este sentido. Esto ha sido muy potente. De hecho, el Diccionario de la Lengua Española (2014) que acaba de salir en su vigésima tercera edición, y tuvimos el placer de presentarlo en Valparaíso, la semana pasada, tan recientemente así, en la Pontificia Universidad Católica de Valparaíso, en su estreno en Chile, porque recién esta semana es su presentación oficial en el marco de la Feria Chilena del Libro, pero nosotros adelantamos la presentación oficial, y una cosa relevante que dijimos allí es que se llama Diccionario de la

Lengua Española, o sea, se llama DLE, y hay que trabajar para posicionar este nombre, ya no tenemos que llamarlo más DRAE (Diccionario de la Real Academia Española), y es muy importante, las palabras sí importan y como lingüistas tenemos que saberlo. Este diccionario es comunitario y, en el estricto sentido de la palabra, está elaborado por todas las academias; entonces, aprovecho de comentar esto porque creo que es un elemento muy importante para los latinoamericanos en general y sobretodo que el DLE tiene claramente hoy en día, y cada vez más asentado, la marca panhispánica.

Ahora bien, al interior de la Comisión de Lexicografía de la Academia Chilena de la Lengua, donde se preparan las revisiones de las nuevas palabras que probablemente se incorporen al diccionario y se revisan las existentes; allí se ha ido acentuando una mirada desde los corpus y su empleo en el trabajo que se realiza cada quince días. Y no tan solo ya basados en la experiencia de algunos académicos, que sigue siendo una fuente valiosa de información fundamental, sino también desde corpora digitales y disponibles en línea. De este modo, el trabajo en la Academia se apoya en nuevos recursos y en nuevos métodos más tecnológicos.

En un sentido más amplio, veo que en Chile se ha venido experimentado progresivamente un giro importante hacia los trabajos de corpus, lo que no quiere decir que tradicionalmente no hayan existido investigaciones que emplearan corpus, ya que siempre han estado presentes en la investigación lingüística. Me refiero al giro contemporáneo que la Lingüística de Corpus implica y también a la Lingüística Computacional. Esto se nota en las publicaciones que se incrementan en torno a una línea de mayor sustento empírico y con mayores desarrollos tecnológicos. También se aprecia, por ejemplo, en aumento sostenido de los proyectos con financiamiento externo a las universidades, tal como CONICYT y FONDECYT. Todo ello ha llevado a ampliar los textos en análisis y a superar estudios, por ejemplo, únicamente a partir de ejemplos ad hoc. Estas líneas de investigación son las que, entre otras, hemos venido impulsando fuertemente desde Valparaíso a través de diversos frentes de acción; en particular, desde la Escuela Lingüística de Valparaíso (ELV, www.elv.cl) y desde los Programas de Postgrado en Lingüística (www.linguistica.cl). En mi opinión, los enfoques desde la LC son temas muy relevantes que se van siendo cada vez más notorios y que, al mismo tiempo, va reforzando la idea de que los nuevos investigadores y estudiantes más jóvenes se ven muy motivados por estas líneas de investigación. Así, se va estableciendo con mucha fuerza que si tengo una hipótesis y quiero investigar un determinado

fenómeno lingüístico, ya no es estudiar un documento de dos páginas, sino que ver la posibilidad de construir un pequeño corpus o un gran corpus, o trabajar con un corpus disponible y hacer un estudio a partir de aquello; entonces yo creo que ha habido un cambio importante y creo que Chile está avanzando fuertemente en esto, al igual que otros países como Brasil y el mundo.

3. En sus palabras ¿Cómo definiría un Corpus y la Lingüística de Corpus? ¿Qué lugar le asignaría a la Lingüística de Corpus en el ámbito más amplio de la Lingüística?

R: ¿Qué es un Corpus? Hay tantas definiciones de qué es un Corpus ya en la bibliografía que no parece necesario elaborar una nueva definición de qué es un Corpus, más bien conviene recordar qué han dicho o que se ha escrito al respecto. Una cosa interesante es que un corpus no es un texto ni un documento sino que un corpus se supone que es una colección grande o pequeña o mediana, porque corresponderá a ciertos parámetros que lo podrá hacer variar, de modo que el tamaño no es necesariamente una variable creo yo, aunque desde la Lingüística de Corpus siempre se ha dicho "más es mejor", o sea está este principio donde sí incide el tamaño de un Corpus supuestamente, pero también está claro que hay corpus especializados y que por la naturaleza de esos textos puede ser reducido. Entonces, el tamaño, definitivamente, no es lo que define un corpus, aunque ojalá sea de la mayor extensión posible, nos da eso sí un tema más de representatividad posible. Pero el corpus tiene que ver con una selección de criterios que estimo son relevantes, como que podamos saber a qué responden, o sea, quiénes han producido estos textos, en qué contextos se han producido estos textos, qué nos llevó a recolectar estos textos. Hoy en día podríamos agregar la disciplinaridad como un tema importante, si es que son de alguna disciplina particular, si son orales o escritos, entonces hay una serie de parámetros que podríamos establecer para definir en este sentido qué es un corpus. Es una colección de documentos que debe estar alineada con algunos principios que tienen que ver con la recolección misma de este, de modo que no debiera tener una heterogeneidad en su carácter, sino que debiera tener algunos principios unificadores en este sentido y debiera, por ejemplo, responder a algunas preguntas como ya decíamos que tiene que hacerse el investigador que lo recolecta. Ojalá tuviéramos la mayor cantidad de información posible para poder saber qué es un corpus, o qué corpus es el que estamos juntando; también debiéramos tener algunos parámetros ecológicos, o sea, ojalá que los textos que conforman el corpus sean textos completos de la variedad que estamos recolectando, o

sea, si son conversaciones de algún tipo de registro oral, ojalá sean conversaciones completas, o sea que de acuerdo a los propósitos comunicativos que la guían, tengamos desde el principio al fin de esa conversación, donde los participantes, cuantos sean, sean los que están registrados. Si es un corpus de textos escritos, el ideal, sino la obligación, creo yo hoy, es que los textos sean completos o sea, sean textos donde está toda la información de principio a fin. Hay hoy en día una definición que hace rato da vuelta que es la que yo no creo mucho que es la de "elimine el paratexto", y establecen una distinción entre el texto y el paratexto, y suponen entre otros que la portada, que las referencias bibliográficas, que toda la primera parte, por ejemplo, la editorial, el país, los autores de un documento escrito no son parte integral de un texto, sino que el texto comenzaría cuando el documento en sí comienza en la primera línea propiamente dicha. Yo creo que esas distinciones no corresponden, creo que un texto es todo lo que tenga como texto, y si lo que estoy tomando es un documento por ejemplo como un libro, irá todo lo que tenga que ver con la editorial, el lugar de impresión, con el número del tiraje, y eso es parte del texto propiamente tal, o sea, no son elementos "paratextuales", si está una mini biografía en la solapa del autor, algunos consideran que eso no es parte integral del texto, pero a mí me parece que sí, si esta es una obra de un determinado autor eso es parte, entonces pensemos en una colección de textos donde los textos sean completos, tengan toda la "completitud" que corresponde y no lo sesguemos, sobre todo desde la teoría del género, que es fundamental pensar que los textos deben ser situados y completos en este sentido y que tengamos la mayor información respecto de qué los caracteriza. También hoy en día diríamos que un corpus debe ser digital. En general, un corpus no es una colección de textos en papel, un rasgo hoy, porque la LC está asociada al tratamiento de los corpus en términos computacionales, lo que no quiere decir que todo tratamiento de corpus deba -en mi opinión- ser obligatoriamente computacional. En mi opinión, una rica combinación puede darse entre investigación cualitativa y cuantitativa, pero sí es importante que los textos estén en formato digital, porque pensamos en que deben estar colgados en algún sitio, que sean accesibles a procesadores de algún tipo.

En mi opinión, no es un principio obligatorio que todo análisis de corpus deba obligatoriamente tener un tratamiento estadístico y computacional, yo creo que la LC también tiene que estar abierta a tratamientos de esos corpus que puedan obedecer a principios más cualitativos, a investigaciones que tengan otros fines, no la cuantificación per se como un único objetivo. Un ejemplo, cuando estudiamos desde la teoría del género, los rasgos retórico

funcionales desde la perspectiva de Swales, ese estudio normalmente es posible hacerlo cualitativamente, o sea es un estudio que tenemos que hacerlo más bien manualmente, y que lo que busca es determinar organizaciones funcionales que puedan reconocerse en los patrones retóricos de esos textos, como clásicamente la introducción, la discusión, la parte metodológica, en fin, en un artículo de investigación. Es cierto que también puede estudiarse mediante técnicas computacionales automatizadas, aunque también es posible que se haga en un trabajo manual, de modo que yo lo veo reñido el trabajo cualitativo manual con el trabajo computacional estadístico, cuantificable, creo que se nutren ambos, por supuesto que hay una Lingüística de Corpus estadística per se, lingüística cuantitativa como lo llaman algunos también, y que sólo concibe una mirada muy cuantificable de los elementos textuales, pero yo creo que esa mirada se enriquece desde otras perspectivas también.

4. ¿En qué lugar posicionaría la Lingüística de Corpus pensando en la gran área de la Lingüística?

R: Si se mira, en un sentido diacrónico, no cabe duda de que el giro hacia la LC ha alcanzado un impacto muy significativo. En un periodo de tiempo relativamente breve (treinta a cuarenta años), los desarrollos en términos de sociedades nacionales e internacionales, revistas indexadas, grupos de investigación, series de libros, espacios al interior de otras asociaciones tradicionales, proyectos de investigación nacionales e internacionales y estudios multilenguas han alcanzado un impacto muy significativo. Al mismo tiempo, la LC ha ido ganando un espacio de relevancia dentro de las áreas de los estudios del lenguaje, tal como fácil se comprueba en fondos de financiamiento de proyectos, en congresos, en revistas, entre otros. No es una opción encapsulada, lo cual es muy bueno. Ahora, todo ello depende mucho de cómo entendamos lo que es la LC. En mi opinión, la LC es una metodología muy potente, para mí la LC no es una teoría. Es una forma de trabajar, no tiene necesariamente un conjunto de principios teóricos, de modo que yo puedo hacer LC desde la lingüística sistémica funcional, desde la teoría del género, o desde otras perspectivas lingüísticas. Por tanto, cuando lo decimos así, emerge claramente la idea que la LC no es un paradigma lingüístico en el sentido kuhniano, o sea que tenga un objetivo, un objeto y un método específico, y que tenga unos principios teóricos que la orientan; creo que es más bien una forma de aproximarse al tratamiento de la información que se recolecta, y eso hace que sea más poderosa aun, porque puede estar transida en todas las áreas de investigación lingüística, a lo mejor la lingüística

generativa puede estar reñida porque tiene una idea de la idealización, de no estudiar el uso, y la LC está comprometida con el estudio de uso y la variación, y yo creo que entonces casi todos los otros movimientos lingüísticos no tienen problema en asociarse con la LC. Así, pienso que la LC ha tenido un proceso creciente y yo le augurio un futuro promisorio en ese sentido.

5. Según su experiencia como profesor universitario ¿Cuál es el espacio que ocupa la LC en los programas de estudio a nivel nacional en Chile e internacional?

R: En este momento, a pesar de lo que decía más arriba, soy de la idea de que en Chile aún el espacio de la LC es reducido. Siendo muy honesto, creo que todavía está en ciernes este tipo de trabajo y, sobre todo en Chile, hemos tenido un proceso largo de reacomodación de currículos universitarios, a través de la discusión de qué tipo de profesionales formamos en las carreras de pregrado, donde en nuestros ámbitos están la pedagogía en castellano, la pedagogía en inglés, la traducción, los periodistas, entre otros. La cuestión de fondo es si estamos formando profesionales o si estamos formando investigadores, y en Chile existe una cierta tensión entre la idea del investigador al alero de las licenciaturas y una relación pensada con el profesional que tiene que hacer en el mundo laboral. Entonces muchas veces hay una formación más pedagogizante, más didactizante en algunos casos y menos investigadora de la lingüística per se; ciertamente se puede hacer investigación acción, investigación en el aula, investigación didáctica, o de lingüística aplicada, pero mi visión en este momento es que todavía en Chile, visualizo los estudio de corpus a nivel de formación de grado están lentos y tienen una incorporación ingenua o inicial, donde estamos todavía tratando de generar un espacio en ese ámbito. El más fuerte es en el ámbito de las licenciaturas donde está más asentada la idea de formar investigadores y, por ende, hacer investigación. Paralelamente, en algunos programas de magister y doctorado la LC ha tenido otro derrotero y sí ha encontrado adeptos muy entusiastas. En estos nichos creo que sí ha calado dentro de las universidades, pero –al mismo tiempo- creo que todavía está en franco desarrollo.

A pesar de que en Chile existe un número importante de universidades en que se desarrollan estudios lingüísticos, hay muchas universidades en que creo que lamentablemente la LC no se visualiza. No obstante ello, a mí me parece que esas son grandes oportunidades y quiero verlo más bien como una cosa positiva, una gran oportunidad en la cual tenemos una tarea. A mí me encanta encontrar desafíos, encontrar tareas que hacer, porque entonces si

estuviera todo hecho, tendría que casi jubilarme, pero como veo que hay mucho que hacer, encuentro motivaciones, encuentro nuevas tareas, me doy nuevos objetivos, y por ejemplo creo que este es un gran tema, situar más fuertemente y pensar como uno puede incluir en la investigación de grado. Si bien en maestrías y doctorados tiene mucha más potencia, creo que en la formación de grado hay que sentar las bases, hay que motivar a los alumnos inicialmente; mostrarles las potencialidades que hay en estos estudios, y cuando lo hemos hecho, tenemos grandes logros. Yo, con mucho orgullo, tengo en los últimos años publicaciones con alumnos de grado, y puede parecer curioso que en estos últimos años no tenga publicaciones con alumnos de maestría y doctorado, pero sin embargo tengo publicaciones con alumnos de grado. Y lo valoro grandemente

5.1 ¿Y a nivel internacional?

R: Bueno, separemos Latinoamérica del mundo, creo que en la lingüística y tal como lo hemos discutido en este congreso aquí en Brasil, hay que mirar a la LC, desde diferentes ángulos, un ángulo de donde la podemos mirar creo yo es desde la lingüística, sólo desde la lingüística porque ciertamente la LC tiene otros desarrollos desde la lingüística computacional, desde el procesamiento natural del lenguaje, y eso no corresponde ciertamente tal vez a lingüistas, pero desde la lingüística solamente, creo que dentro de Latinoamérica su impacto es lento todavía, creo que nuevamente ahí tenemos mucho que hacer. Si uno toma una asociación importante en Latinoamérica, todavía es poco el impacto.

Los eventos donde se hace LC en Latinoamérica creo que son pequeños en términos cuantitativos, justamente en cuanto a número de personas, a cuántos somos los investigadores que estamos ahí y presentamos trabajos. En cuanto al trabajo con corpus y de recolectar grandes corpus y la investigación con apoyo tecnológico, veo que aún esta línea es escasa. Existen sí grupos de investigación; identifico claramente gente en Brasil, Argentina, México y Chile, pero menos en otros países latinoamericanos, nuevamente debo decir que eso me parece un desafío, creo que tenemos un compromiso con otros países latinoamericanos, creo que España ha avanzado bastante, hace diez años atrás no existía mucho en esta línea de investigación, hoy en día se ve bastante, en algunas reuniones en las que he estado he visto, hay un desafío grande con esto, veo que está cambiando y ojalá que cambie más.

En síntesis, en términos generales, tal como se decía más arriba, la investigación en LC en el nivel internacional también es variada. En algunos países de Europa, ha sido un

proceso muy dinámico y muestra impactos fuertes. Lamentablemente en otros espacios internacionales se detecta escaso o casi nulo desarrollo de la LC. En lengua española, ha habido en los años recientes un importante avance, aunque en algunos países se nota mucho mayor interés que en otros. En parte, como decíamos más arriba, esto más que ser negativo, debería ser una motivación para crear e impulsar nuevos espacios de interacción.

6. ¿Cómo proyecta el futuro de la Lingüística de Corpus de aquí a diez o quince años más?

R: La LC, como ya lo hemos comentado en este congreso aquí en Uberlândia, tiene varios desafíos, como partí diciendo al inicio de esta entrevista, cuando comencé a trabajar con LC yo decía "denme las palabras, y borren todo el resto", yo creo que el gran descubrimiento, más bien el redescubrimiento, o como algunos lo llaman: "el descubrimiento de lo obvio", es el descubrimiento de la multimodalidad de los textos, o sea que los textos no son sólo palabras. Esto puede parecer una cosa obvia: que los textos no son sólo palabras no es nada nuevo, porque claro, cuando yo decía "denme sólo las palabras y borren todo el resto", estaba aludiendo a que yo veía que en los textos habían otras cosas más que palabras, pero mi motivación y la de mi equipo eran tan grande en ese momento por construir corpus digitales marcados, por tener un etiquetador y una interfaz de consulta, que estábamos obsesionados con la idea de las palabras y de poder decir: este es un sustantivo y esto es un pronombre de primera persona singular; poder tener el conteo de las formas y de las categorías gramaticales, que cuando decíamos denme todas las palabras y borren todo el resto, visualizábamos que los textos eran más que palabras, pero nuestro cometido estaba con las palabras. Pero, hace ya algunos años, también se me enfrentó un texto que me persiguió y me hizo ver que lo que estábamos borrando era parte integral de los textos, por lo tanto los textos no son sólo palabras obviamente, pero sobre todo si nosotros pensamos que los seres humanos hablamos en textos y, más aun, hablamos en géneros y nos comunicamos en textos y en géneros, entonces los textos y los géneros tienen mucho más que palabras. Como sabemos, para los seres humanos el centro de la comunicación es la construcción y transmisión de significados, y los significados no los podemos construir sólo de palabras. Y esto quiere decir que si alguien está interesado como yo en los textos y los géneros, y por supuesto desde la LC, tenemos el gran desafío de poder dar cuenta de cómo son realmente los textos. Esto quiere decir que la LC va a tener que desafiarse a sí misma y tendrá que pensar en una LC

Multimodal, o una visión multisemiótica del lenguaje desde la LC. Así, el sistema verbal constituye sólo uno de los sistemas constitutivos de los textos, por lo tanto si la LC quiere hacerse efectivamente cargo de los textos, los textos no son sólo palabras. Tenemos, entonces, que poder desarrollar herramientas que logren identificar los otros sistemas, los sistemas gráficos, sistemas visuales, sistemas matemáticos, sistemas del color, que es tan fundamental por ejemplo si pensamos en los gráficos.

Todo esto de la multimodalidad quiere decir mucho para la LC. Así, si queremos dar cuenta del significado que está en los textos, yo creo que un desafío maravilloso para la LC es pensarse como una LC Multimodal, y eso creo que son unos diez o quince, o hasta treinta años por lo menos. Es cierto que ya hay especialistas buscando crear programas computacionales que logren identificar un gráfico, una tabla, una ilustración, de modo que puedan tener un marcaje; algunos piensan que el XML puede ser una forma de trabajo para ayudarse con esto. Pero la verdad es que el desafío es de proporciones y más vale tener claro el escenario. Aún no se logra dar cuenta efectiva de las relaciones semánticas entre oraciones de un texto y, por ende, automatizar este tipo de procesos constituye un desafío tremendo que no está logrado aún. Esa es mi opinión en este momento. Creo que aún no logramos capturar sistemáticamente los mecanismos textuales de, por ejemplo, la construcción de la coherencia textual. Por lo tanto, pensar en cuáles son las relaciones semánticas que conectan una ilustración con las palabras de un determinado texto, o sea como el sistema verbal interactúa con los sistemas gráficos, visuales, matemáticos, entre otros. Por ejemplo, un texto de economía donde mucho de lo que se dice está en fórmulas y gráficos, no puede analizar solamente las palabras. Se debe analizar los gráficos y las tablas y se debe analizar, además, como se construye el significado en la intersemiosis entre esos gráficos y esas tablas y el sistema verbal. Suena maravilloso, o sea, hay una tarea enorme, mucho trabajo por delante. El futuro de la LC es tremendo.

7. Con respecto al interfaz computacional El Grial, pensando en la perspectiva del aprendizaje y enseñanza de la lengua española, tanto como lengua materna como lengua extranjera, y en la investigación, ¿nos podría contar un poco como funciona dentro de este ámbito?

R: El Grial tiene fortalezas tremendas, pero en el contexto mundial, está pensado solo para una lengua que es el español y entonces pensando en las lenguas extranjeras o la investigación en otras lenguas El Grial no ofrece posibilidades en este momento, y tampoco puede hacer nada con respecto de los textos multimodales, o sea, sólo trabaja con las palabras, tal vez eso es una fortaleza tremenda, y que lo haga bien y que se haga cargo de eso, yo creo que es un tremendo desarrollo y hay que perfeccionarlo y mejorarlo, pero también es interesante que El Grial debe desafiarse a sí mismo.

Esta herramienta nos ha apoyado de manera muy importante y ha dado impacto a nuestras investigaciones en LC. Sin El Grial, muy probablemente no habríamos alcanzado los entusiasmos y compromisos de muchos colaboradores y estudiantes. Ha sido, como ya decía, un antes y un después en nuestra investigación en la Pontificia Universidad Católica de Valparaíso. Sin duda, también nos ha aportado mucho a la visualización de los que hemos hecho y hacemos en la ELV. Espero en sus siguientes fases, El Grial pueda ser una herramienta que también apoye los procesos de enseñanza aprendizaje del español y pueda acoger corpus de aprendientes, por ejemplo.

8. ¿Podía hacernos un breve resumen de qué se trata El Grial?

R: El Grial es una interfaz computacional que está en internet y que, desde uno de sus objetivos, se presenta como de acceso abierto y gratuito. Básicamente lo que hace es que permite etiquetar morfosintácticamente los textos que se suben en formato plano; también permite colgar estos textos allí dentro y permite acceder a estos textos a través de diferentes herramientas de búsqueda para recuperar información en formato de consultas específicas. Es posible hacerlo por medio de formas o de categorías, o de cadenas de categorías o de cadenas de formas, entro otras. O sea, permite a través de la interfaz, en lo que denominamos la Búsqueda Compleja o, a través de lo que llamamos El Manchador de Textos, buscar alguna forma determinada, por ejemplo puedo buscar la palabra “para” específicamente, pero puedo buscar la palabra “para” mezclada con un adjetivo y mezclada en la misma cadena, seguida de un adjetivo y de un sustantivo por ejemplo. Entonces estoy definiendo en la cadena de búsqueda una forma y dos categorías gramaticales. Esto solo a modo de ejemplo y muy resumidamente.

8.1. Pensando en las ventajas para el aprendizaje de lengua, específicamente del español como lengua materna y como lengua extranjera

R: En este momento yo diría que El Grial, tal y como está, no es una herramienta que potencie el aprendizaje y enseñanza de lenguas. Es una herramienta más bien de investigación; está proyectado, como decía anteriormente, que a finales del 2015 tenga alguna contribución en torno a los llamados corpus de aprendices o aprendientes, y vamos a comenzar a hacerlo con una disciplina que es la disciplina de economía, pero en este momento tal y como está es una herramienta para que el investigador pueda testar sus hipótesis, comprobar, corroborar o indagar sus hipótesis, tener datos empíricos y a partir de estos datos que recupera un investigador. En virtud de esto, un investigador puede crear material o tomar decisiones metodológicas de enseñanza aprendizaje, pero no es una herramienta en sí misma -en este momento- construida para que un aprendiz pueda acercarse a ella y obtener la información para el desarrollo de su lengua materna o español como lengua extranjera. Sin duda, existe entonces un nicho importante que se puede desarrollar y que estamos pensando comenzar a explorar muy tentativamente.

9. ¿Qué mensaje le dejaría a un estudiante que se comienza a interesar por la LC?

R: Estoy seguro de que un camino certero es mostrarles que existe un asunto que a mí me motivó a comenzar a estudiar lingüística. Yo era ayudante del curso de gramática del inglés, con mi maestra, y observé en la lengua inglesa la existencia de regularidades, de tendencias pero también de variaciones. Y esto me obsesionó, me apasionó y lo hace hasta el día de hoy. Creo que una cosa que les podemos enseñar a los estudiantes noveles, que le podemos mostrar a los alumnos que están recién ingresando, es cómo la LC, como una herramienta, como una metodología de trabajo, puede permitir descubrir la maravilla que hay en el lenguaje, en las tendencias, en las regularidades, pero al mismo tiempo de las variaciones y como ciertas tendencias y regularidades varían. Cambiar los registros, los géneros, las disciplinas, y así observar cómo se construyen los discursos en ciertas disciplinas, en ciertos contextos, cómo esos contextos estimulan la variación de los rasgos léxico-gramaticales en la construcción del significado. Estoy convencido de que si un estudiante que está recién ingresando a la investigación visualiza las regularidades, la idea de encontrar tendencias, pero a la vez encontrar también “no tendencias” o de encontrar variaciones entre las tendencias, yo creo que se ve un escenario muy interesante de investigación, y yo creo que eso le permite mostrar

la riqueza y la variedad de las lenguas, lo cual a mi me parece que es una cosa fundamental. Mostrar que cada lengua en todas sus dimensiones es muy rica, es inconmensurable. Aportar así a la formación de las nuevas generaciones de profesionales y de futuros investigadores constituye un desafío muy motivador en el que desplegar todas las potenciales en conjunto es una proyección potente.

Al concluir esta última pregunta, deseo agradecer de manera especial a Ariel Novodvorski por su generosidad y compañerismo al invitarme a Uberlândia a dos magníficos eventos académicos a través de los cuales he podido conocer personas magníficas y, al mismo tiempo, aprender muchas nuevas ideas acerca de la LC y de los desarrollos en Brasil en esta línea. También y no menor, le agradezco a Ariel haber organizado esta entrevista que me brinda la posibilidad de poner en palabras orales y luego por escrito algunas nociones y reflexiones en torno a un área de las ciencias del lenguaje, con la que tanto él como yo estamos muy comprometidos en nuestras investigaciones: la LC. Les dejo desde este foro, un abrazo fraterno.

Uberlândia, Brasil, 5 de noviembre, 2014.