Domínios de Lingu@gem

Organização: Heliana Mello

2º Trimestre 2015
Volume 9, número 2

ISSN: 1980-5799

Expediente

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

Domínios de Lingu@gem, v. 9, n. 2, 2015, Uberlândia, Universidade Federal de Uberlândia, Instituto de Letras e Linguística, 2007-

Trimestral.
Modo de acesso:
http://www.seer.ufu.br/index.php/dominiosdelinguagem
Editoração: Guilherme Fromm.
Organização: Heliana Mello.

ISSN: 1980-5799

1. Linguística - Periódicos. 2. Linguística aplicada - Periódicos. I. Universidade Federal de Uberlândia. Instituto de Letras e Linguística.

CDU: 801(05)

Participaram dessa edição como pareceristas *ad hoc*

Barbara Malveira Ofarnó (UFMG)

Ivanir Azevedo Delvizio (UNESP)

Sumário

Expediente 2

Sumário 5

Apresentação 6

Linguística de *Corpus*, metodologias e interfaces – Heliana Mello (UFMG) 6

Artigos 11

Comparando anotações linguísticas na Gramateca: filosofia, ferramentas e exemplos – Diana Santos (Universidade de Oslo/Linguateca), Rui Marques (Universidade de Lisboa), Cláudia Freitas (PUC/Rio), Alberto Simões (Universidade do Minho/Linguateca), Cristina Mota (Linguateca) 11

A relevância da Web como *corpus* para a identificação de padrões de lexicalização: o caso de "bater+SN" no português brasileiro - Milena de Uzeda Garrão (UFRRJ) 27

Extração automática de candidatos a termos do *Curso de Linguística Geral* com apoio de recursos da Linguística de *Corpus* e do Processamento de Linguagem Natural - Maria José Bocorny Finatto (UFRGS), Lucelene Lopes (PUC/RS), Alena Ciulla (UFRGS) ... 40

Diretrizes para a criação de um recurso lexical multilíngue a partir da semântica de *frames*: a experiência turística em foco - Maucha Andrade Gamonal (UFJF), Tiago Timponi Torrent (UFJF) 56

Designing English Teaching Activities Based On Popular Music Lyrics From A *Corpus* Perspective - Maria Claudia Nunes Delfino (FATEC/PG) 76

O uso de *corpora* comparáveis na pesquisa terminológica bilíngue - Marina Araújo Vieira (UFU), Silvana Maria de Jesus (UFU) 96

Ilocuções comissivas em dicionários híbridos italiano>português brasileiro: proposta de dicionarização a partir do uso de *corpora* - Renato Railo Ribeiro (USP) 125

Tradução e Retradução de *The Picture of Dorian Gray*, de Oscar Wilde: um estudo de *corpus* com foco na apresentação do discurso - Lívia Cremonez Domingos (UFU), Igor A. Lourenço da Silva (UFU) 150

Apresentação

Linguística de *Corpus*, metodologias e interfaces

É com prazer que apresentamos ao público leitor da Revista Domínios de Lingu@gem o seu primeiro número temático de 2015, voltado para a pesquisa em Linguística de *Corpus*, ou aquela que utiliza a metodologia proposta por este campo disciplinar da Linguística como ferramenta de investigação. Conforme apontado por Fromm e Novodvorsky (2015), a Linguística de *Corpus* vem ganhando espaço nos periódicos científicos brasileiros nos últimos anos, notadamente nas seguintes publicações temáticas, inteiramente dedicadas aos relatos de avanços na pesquisa da área: Veredas 13:2 (2009), Revista Brasileira de Linguística Aplicada 11:2 (2011) e Letras e Letras 30:2 (2014). É chegada a hora de a Domínios de Lingu@gem publicar um novo número temático sobre o assunto, acompanhando publicação análoga pela Revista de Estudos da Linguagem 23:3 (2015).

O presente número da Domínios de Lingu@gem, composto por oito artigos, traz trabalhos apresentados durante o XII Encontro de Linguística de *Corpus* (ELC) e da VII Escola Brasileira de Linguística Computacional (EBRALC) sob a temática *Corpus*, Tecnologia e Cultura, além de artigos pertinentes à área recebidos pela revista. Estes eventos ocorreram em novembro de 2014, no Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia (UFU). A organização coube aos Profs. Ariel Novodvorsky e Guilherme Fromm, ambos membros do Programa de Pós-Graduação em Estudos Linguísticos (PPGEL/ILEEL/UFU), vinculados ao Grupo de Pesquisas e Estudos em Linguística de *Corpus* – GPELC, da UFU.

Os artigos que compõem este número da Domínios de Lingu@gem cobrem variados focos temáticos da Linguística de *Corpus* e suas interfaces, como será relatado a seguir. As pesquisas discutidas pautam-se no desenvolvimento e utilização de ferramentas computacionais para a análise de dados linguísticos, a utilização da web como *corpus*, criação de atividades didáticas baseadas em *corpora*, incremento de informações pragmáticas extraída de *corpora* a dicionários bilíngues e interfaces da pesquisa terminológica e tradutória com metodologias baseadas em *corpora*.

No primeiro artigo, *Comparando anotações linguísticas na Gramateca: filosofia, ferramentas e exemplos*, Diana Santos, Rui Marques, Cláudia Freitas, Alberto Simões e

Cristina Mota apresentam a Gramateca, ambiente para estudos da gramática da língua portuguesa baseados em *corpus*, bem como alguns estudos já desenvolvidos em seu âmbito. O foco desta equipe multiinstitucional é nos procedimentos metodológicos possibilitados pela Gramateca, destacando especialmente o sistema Rêve, que permite revisar e partilhar anotações linguísticas. Dentre os aspectos de destaque apontados no artigo está a possibilidade de cruzamento de anotações humanas com o processamento automático ou semiautomático de dados linguísticos, o que garante o detalhamento qualitativo e a validação quantitativa necessários aos estudos linguísticos contemporâneos.

Milena de Uzeda Garrão, em *A relevância da web como corpus para a identificação de padrões de lexicalização: o caso de "bater + SN" no português brasileiro*, argumenta a favor da utilização da web como recurso empírico para pesquisas linguísticas. A autora explora as inovações de uso do verbo "bater", notadas tanto no Brasil quanto em Portugal, através das quais esse verbo passa a ter ocorrências transitivas diretas, com sentido de "vencer", aliadas a ocorrências aparentemente de verbo suporte como "bater um medo". A pesquisa desenvolvida incialmente, a partir de textos jornalísticos, levou a autora a exploração ampliada através da web como *corpus* para a sua verificação. Garrão conclui o seu artigo com uma defesa epistemológica favorável ao uso de grandes volumes de dados, como ferramenta necessária a que o pesquisador possa efetivamente contemplar mudanças, ampliações e volatilidades do sistema linguístico em uso.

Em *Extração automática de candidatos a termos do "Curso de Linguística Geral" com apoio de recursos da Linguística de Corpus e do Processamento da Linguagem Natural*, Maria José Borcony Finatto, Lucelene Lopes e Alena Ciulla exploram ferramentas computacionais e comparam sua eficácia para a identificação de sintagmas nominais relevantes ao se tomar a famosa obra póstuma de Ferdinand de Sausurre como *corpus* de análise. As pesquisadoras descrevem os passos metodológicos da pesquisa, indicando a anotação morfossintática do *corpus* através do anotador Palavras como elemento um dessa trajetória; a seguir, faz-se a extração de SNs através da ferramenta ExATOlp, a qual faz uso de técnicas estatísticas e linguísticas. Os resultados são consolidados a seguir através de procedimentos de visualização. Os resultados obtidos nesta etapa dos procedimentos são comparados àqueles obtidos com a utilização da ferramenta computacional AntConc, de acesso livre à comunidade. As autoras concluem,

destacando a funcionalidade de ambas as ferramentas testadas, porém destacando o potencial específico dos resultados obtidos via ExATOlp.

Na sequência, Maúcha Andrade Gamonal e Tiago Timponi Torrent apresentam *Diretrizes para a criação de um recurso lexical multilíngue a partir da semântica de frames: a experiência turística em foco*. Neste artigo os autores exploram os passos metodológicos adotados na confecção do Dicionário FrameNet Brasil da Copa do Mundo, um dicionário trilíngue (português-inglês-espanhol) de acesso grátis, lançado para atender às demandas turísticas com o evento da Copa do Mundo celebrada no Brasil em 2014. O projeto, que se beneficia da larga experiência da matriz FrameNet desenvolvida na Universidade de Berkeley, utiliza a Semântica de Frames como marco teórico que orienta suas propostas de entradas no dicionário, orientadas por ocorrências de uso em *corpora*. O artigo apresenta o domínio turístico para ilustrar as funcionalidades da ferramenta descrita.

Maria Cláudia Nunes Delfino aborda em *Designing English teaching activities based on popular music lyrics from a corpus perspective* o potencial das ferramentas disponíveis via Linguística de *Corpus* na composição de atividades que possam fomentar a atuação didática de professores de inglês. A pesquisadora compilou um *corpus* de 150.000 palavras a partir de letras de músicas votadas por seus alunos e o tratou através da ferramenta AntConc. Na sequência, a pesquisadora cotejou os resultados obtidos com listas de palavras mais frequentes no *corpus* de referência de língua inglesa COCA e elaborou atividades aplicadas a seus alunos. A aplicabilidade e impacto do material didático desenvolvido foram testados através de diários reflexivos elaborados pelos alunos, que corroboraram a intuição de que materiais criados a partir de *corpora* propiciam condições favoráveis de aprendizagem.

Ampliando a discussão sobre o estudo terminológico para o domínio bilíngue, Marina Araújo Vieira e Silvana Maria de Jesus, autoras de *O uso de corpora comparáveis na pesquisa terminológica bilíngue*, exploram a interface estudos terminológicos e tradução, via metodologia de exploração de *corpora*. O domínio temático do estudo é o espiritismo, abordado através de obras originalmente escritas em português e traduzidas para o inglês. Termos espíritas, notadamente relativos à mediunidade, foram extraídos via *corpora* bilíngues comparáveis e paralelos. Os resultados levaram à elaboração de fichas terminológicas e seleção de termos multivocabulares a fim de se criar uma amostra de glossário bilíngue. As autoras concluem que a riqueza vocabular oferecida no campo explorado na língua inglesa não foi contemplada nas obras traduzidas através das escolhas

terminológicas dos tradutores, desta forma apontando lacunas a serem preenchidas na formação de tradutores profissionais.

Explorando a interface entre a metodologia da linguística de *corpus* e os estudos do léxico, Renato Railo Ribeiro, no artigo *Ilocuções comissivas em dicionários híbridos italiano>português brasileiro: proposta de dicionarização a partir do uso de corpora*, investiga o enriquecimento de verbetes em dicionários bilíngues através de inserção de informações ilocucionárias em ambas as línguas representadas. Para tal fim, o autor apresenta os critérios adotados para a exploração de valores ilocucionários em *corpora*, indicando a pesquisa exploratória a partir de cinco verbos em ambas as línguas enfocadas, e como se daria a inserção de tais informações em dicionários bilíngues. O autor conclui apontando os ganhos que a introdução de dados pragmáticos em verbetes de dicionários bilíngues poderiam acarretar ao aprendiz que deles faz uso.

Também explorando a interface Linguística de *Corpus* e Estudos da Tradução, o artigo que fecha este número da Domínios de Lingu@gem, *Tradução e Retradução de "The Picture of Dorian Gray", de Oscar Wilde: um estudo de corpus com foco na apresentação do discurso*, de Líbia Cremonez Domingos e Igor A. Lourenço da Silva, investiga um *corpus* compilado a partir de trechos da obra original "The Picture of Dorian Gray", de sua primeira tradução para o português e duas retraduções. O propósito dos autores é avaliar a hipótese de que a primeira tradução de uma obra literária é "incompleta e domesticadora". A partir da metodologia instaurada, que utilizou-se de etiquetagem do *corpus*, seu alinhamento e processamento semi-automático, os autores concluem que há indícios que corroboram a hipótese investigada.

Esperamos que este número temático da Domínios de Lingu@gem auxilie a divulgação da pesquisa que vem sendo desenvolvida no campo temático da Linguística de *Corpus*, bem como estimule a curiosidade de pesquisadores ainda não familiarizados com as metodologias empregadas nos estudos aqui relatados na busca de interfaces com seus próprios interesses de pesquisa.

Heliana Mello (UFMG)

Referências

Letras & Letras, 30:2, 2014. Disponível em: http://www.seer.ufu.br/index.php/letraseletras/issue/view/1217. Acesso em 17/12/2015.

FROMM, G.; NOVODVORSKY, A. Triangulando *corpus*, tecnologia e cultura: ELC e EBRALC na UFU. Apresentação. **Revista de Estudos da Linguagem**. 23:3, 2015, no prelo.

Revista Brasileira de Linguística Aplicada, 11:2, 2011. Disponível em: http://www.scielo.br/scielo.php?script=sci_issuetoc&pid=1984639820110002&lng=es&nrm=1. Acesso em 17/12/2015.

Revista de Estudos da Linguagem 23:3, 2015, no prelo. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/relin. Acesso em 17/12/2015.

Veredas – Revista de Estudos Linguísticos, 13:2, 2009. Disponível em: http://www.ufjf.br/revistaveredas/edicoes/2009-3/2009-2/. Acesso em 17/12/2015.

Artigos

Comparando anotações linguísticas na Gramateca: filosofia, ferramentas e exemplos

Comparing linguistic annotations in Gramateca: philosophy, tools and examples

Diana Santos[*]
Rui Marques[**]
Cláudia Freitas[***]
Alberto Simões[****]
Cristina Mota[*****]

RESUMO: Neste artigo apresentamos a filosofia geral da Gramateca – um ambiente para fazer uma gramática da língua portuguesa baseada em corpos – e alguns estudos no seu âmbito, nomeadamente o estudo (1) dos conectores condicionais, (2) das palavras referentes ao corpo humano e (3) das emoções na língua. A ênfase é na metodologia, e apresentamos detalhadamente o sistema Rêve para rever e partilhar anotações linguísticas. Ao descrever os vários estudos, indicamos também as metamorfoses e melhorias por que essa ferramenta passou, assim como o tipo de perguntas e de resultados que já conseguimos obter em áreas muito diversas.

PALAVRAS-CHAVE: Corpos. Anotação. Semântica. Ferramentas de *corpora*.

ABSTRACT: This paper presents the general philosophy of Gramateca, for corpus-based Portuguese grammar studies, by reporting on three different studies – conditional connectives, body terms, and emotions – emphasizing methodological aspects. It presents in detail the Rêve system, which allows revising and sharing annotations of Rêve's underlying corpora. While describing the different studies we also report on the improvement of the Rêve tool, and discuss the kinds of questions and results already available for diverse fields.

KEYWORDS: Corpus. Annotation. Semantics. Corpora tools.

1. Apresentação

Neste artigo apresentamos a filosofia geral da Gramateca[1] – um ambiente para fazer gramática com base em corpos[2]. Gramática, aqui, é entendida em sentido amplo e compreende

[*] Universidade de Oslo e Linguateca
[**] Universidade de Lisboa (FLUL)
[***] PUC-Rio e Linguateca
[****] Universidade do Minho e Linguateca
[*****] Linguateca

[1] http://www.linguateca.pt/Gramateca

[2] Utilizamos ao longo do artigo o termo "corpo", como já foi defendido em Santos (2008, p. 43): "Vejamos o próprio exemplo da linguística com corpos: este último objecto tem sido variadamente chamado *corpora* (plural *corpora*), *córpus* (plural *corpora* ou *córpus*), mas parece não ter sido sequer equacionado o uso duma palavra genuinamente portuguesa e semelhante, *corpo*, empregue aliás de forma análoga em linguagem legal: *corpo de delito*. Na acepção mais lata de corpo como colecção de textos, é usada naturalmente a palavra *acervo* no Brasil,

não apenas aspectos (morfo-)sintáticos, mas também áreas relacionadas ao sentido e ao léxico, bem como questões vinculadas aos gêneros textuais e aspectos culturais, por exemplo. A ênfase é na metodologia, visto que em outras publicações cobrimos outros aspetos da Gramateca (Santos, 2014; Simões & Santos, 2014; Santos, 2015). Especificamente, iremos ilustrar como uma das ferramentas desse ambiente, o Rêve, direcionou o estudo relacionado a (a) conectores condicionais; (b) léxico do corpo humano e (c) emoções na língua.

No âmbito da Gramateca, uma das maneiras de se estudar e/ou descrever uma língua é por meio do estudo de grandes quantidades de texto. Uma das fontes de inspiração para este projeto foi a gramática de Biber et al. (1999) para o inglês; mas, passados quinze anos do projeto inicial, é possível ir além: não só porque os corpos em que a Gramateca se baseia são públicos – diferentemente do trabalho de Biber et al. (1999) – e porque as técnicas estatísticas modernas estão bem mais desenvolvidas, mas também porque, em vez de um simples etiquetador, temos acesso à riqueza da anotação morfossintática fornecida pelo PALAVRAS (Bick, 2000), e também a alguma anotação semântica criada pela Linguateca[3], nomeadamente cor (Silva & Santos, 2012), roupa, corpo humano (Freitas, 2013) e emoções (Santos e Mota, 2015).

De fato, o AC/DC[4] – um serviço de acesso a *corpus* e disponibilização de corpos, iniciado pela Linguateca em 1999 – constitui a base sobre a qual se sustenta a Gramateca. Atualmente, graças ao grande número de projetos que nos autorizaram a disponibilizar seus corpos, temos material público de qualidade para dar início ao projeto. Como indicado na página da Gramateca, temos consciência de que uma gramática (ou qualquer estudo) baseado em corpos depende do material em que se baseia. Ainda que seja provável que não tenhamos (ainda) o corpo ideal para escrever **a** gramática, e conscientes de que o material de que dispomos refere-se majoritariamente à língua escrita[5], acreditamos que o que for produzido pode ser relevante como passo inicial de futuros estudos, nem que seja pelo aspeto metodológico.

Enfatizamos o aspecto metolodógico porque a Gramateca é também um laboratório para o estudo da língua portuguesa, no qual estão disponíveis (a) todos os corpos disponibilizados

mas aparentemente não no sentido técnico associado a corpos electrónicos, mais influenciado pelo inglês. Infelizmente o uso não consagrou a possível expansão desse termo, provavelmente por não ter semelhanças suficientes com as designações inglesas/latinas. Proponho assim usar *corpo* e *corpos*, na esperança de que esta portuguesificação (e não aportuguesamento) seja aceite."

[3] A Linguateca é um centro de recursos para o processamento computacional da língua portuguesa; abriga o projeto Gramateca.

[4] Disponível em: http://www.linguateca.pt/ACDC.

[5] Embora essa seja uma limitação, lembramos que, comparativamente, existem muito mais estudos sobre o português oral (do Brasil e de Portugal), como ilustram o projeto NURC e o projeto do Português Fundamental.

pelo AC/DC; (b) a anotação automática desses corpos; (c) ferramentas de visualização e de exploração dos corpos, como a própria interface do AC/DC, e os serviços mais recentes Comparador e Disparador, descritos em Simões & Santos (2014); (d) o Rêve, um serviço para anotação manual de subconjuntos dos corpos associado a uma plataforma de revisão e de comparação de diferentes análises, foco do presente artigo.

A anotação linguística, portanto, é um dos elementos fundamentais da Gramateca, que almeja, além de contribuir com a descrição e o estudo do português, incentivar a prática de compartilhamento (e, consequentemente, a possibilidade de reanálise) dos dados que permitiram os estudos e sistematizações realizados por meio da anotação. Em outras palavras, porque acreditamos que a prática de compartilhamento do material classificado linguisticamente poderá servir de base para mais estudos (e controvérsias) sobre a gramática da língua portuguesa, investimos na viabilização dessa prática. Concordamos, portanto, com Sampson (2001) e Archer (2012), para quem textos contendo material anotado constituem a matéria-prima da maior parte da investigação linguística moderna.

Nossa intenção é contribuir com a metodologia científica no campo da linguística: não só permitir a repetição de uma experiência (o que é uma das propriedades exigidas à metodologia científica), mas também partilhar diferenças de interpretação de um mesmo material. Ou seja, enquanto nas ciências naturais se espera que a mesma experiência leve aos mesmos resultados, nas ciências humanas é não só esperável, mas provável, que haja diferenças na interpretação quando algo é repetido por outros pesquisadores. Estranhamente, tal situação é em geral criticada, em vez de ser vista como uma propriedade que resulta precisamente de lidarmos com um tipo diferente de ciência (ou conhecimento).

Neste artigo (e no projeto da Gramateca) tratamos diretamente dessa questão, criando um ambiente para que diferenças de interpretação possam ser mais do que comentários em artigos ou notas de rodapé na documentação de um corpo anotado. Para justificar a ênfase em um ambiente que promove o compartilhamento e a discussão da anotação de corpos, valemo-nos dos seguintes pressupostos:

(i) entendemos a anotação como uma forma de estudar a língua, e não apenas como uma atividade mecânica capaz de prover material para sistemas que processam a língua automaticamente. A anotação possibilita um estudo linguístico empírico, desenhado à maneira clássica, no qual se criam hipóteses (categorias/etiquetas provisórias) que serão verificadas durante a anotação. Nesse processo, as hipóteses iniciais podem ser confirmadas ou os dados podem levar à reformulação das categorias iniciais, e o processo recomeça. Enfatizamos a anotação como um

procedimento que envolve interpretação, classificação e formalização do fenômeno em foco; e

(ii) aceitamos o conceito de vagueza como fundamental na linguagem natural, como proposto e advogado por Santos (2006). Daí que aceitamos, por princípio, a possibilidade de mais de uma interpretação – e, consequentemente, de classes de anotação – ao mesmo segmento de texto, o que não significa que qualquer interpretação seja aceitável, e aqui invocamos o conceito de "comunidades interpretativas" de Fish (1980). Não consideramos tais diferenças interpretativas anomalias, mas antes visões alternativas que enriquecem a discussão de/sobre uma língua.

Note-se que, de um ponto de vista prático, a discussão sobre as opções linguísticas de anotação do material da Linguateca sempre foi possível, visto que os corpos sobre os quais trabalhamos na interface AC/DC são públicos e, portanto, qualquer interessado pode escolher um subconjunto das análises para rever. No entanto, até agora não havia um serviço que permitisse comparar tais análises, fazer novas análises, propor análises de fenômenos ainda não estudados ou anotados e, além disso, partilhá-las com outros. Agora há, e é disso que trataremos ao longo deste artigo, que está dividido em quatro partes. Na primeira, tratamos de anotação; na segunda, tratamos de dados sobre conectores condicionais. Tais dados, tornados públicos, motivaram a construção do Rêve, ferramenta que permite explorar de maneira simples a anotação como forma de estudo linguístico. Na terceira parte, relatamos uma experiência com o Rêve na qual foram possíveis o refinamento e a validação de um esquema de anotação complexo, o Esqueleto, com ênfase em alguns dos pontos mais controversos dessa proposta de anotação. Na quarta parte, apresentamos uma parcela de um estudo mais amplo sobre as emoções na língua portuguesa, no qual o Rêve atuou como ambiente de validação de uma hipótese.

2. Rêve: por uma outra ênfase na anotação?

Em geral, atividades de anotação são feitas em larga escala, porque são vinculadas a necessidades do processamento automático de uma língua. Quando a anotação é feita privilegiando o estudo linguístico – nosso caso –, a exigência da quantidade perde a relevância. Claro que amostras maiores tendem a apresentar casos novos, trazendo potenciais desafios a um dado esquema de anotação, mas casos particulares podem ser exatamente o que o pesquisador deseja. Porque os interesses são distintos – estudar a língua se sobrepõe à criação de um recurso linguístico para o processamento automático de uma língua (PLN) –, a

metodologia também pode ser outra: poucas frases e poucos – ou mesmo um único – anotador, que almeja tirar proveito da metodologia da anotação para materializar, em segmentos da língua em uso, uma dada classificação do fenômeno em análise.

Esse tipo de atividade-análise, de configuração de cenário de pesquisa, já é possível com o que temos hoje: a compilação de textos para constituir um corpo, a posterior utilização de um editor de textos para marcação e a disponibilização desses dados em alguma página na internet permitem a criação de um ambiente nos moldes do que apresentamos aqui, mas não sem algum trabalho. Assim, o que propomos com o Rêve é possibilitar ao linguista um ambiente de teste de hipóteses *on-line*, gratuito e público, no qual (a) os textos – associados a alguma (boa)[6] anotação linguística – já estão disponíveis para uma seleção simples das frases de interesse; (b) a criação e a in*corpora*ção das classes testadas também são extremamente simples; e (c) a tabulação dos resultados é automática.

Em contrapartida, retomando a ideia de anotação em larga escala, é verdade que a "anotação massiva" tem sido bastante apregoada pelos defensores do "*crowd-sourcing*", que também a aplicam à linguística (Munro et al. 2010). Embora em alguns casos isso seja não só apropriado como a única maneira de o fazer, noutros casos – por exemplo, em Rumshisky (2009) – pensamos que se tem ido longe demais, como aliás, os próprios Rumshisky et al. (2012) parcialmente discutem, apontando os (novos) problemas que esse tipo de tarefa acarreta – por exemplo, devido à falta de confiança nos próprios anotadores. Isto acontece porque a anotação (a tarefa linguística) por meio de um sistema como o Amazon MTTurk[7] é feita por pessoas sem comprometimento com a tarefa, e não por colegas interessados no mesmo tipo de assuntos – peritos – e cujas análises, além disso, são tornadas públicas de forma identificada.

Outra questão muito debatida em projetos associados à anotação humana é a concordância entre anotadores[8], com a ideia implícita de que quantidade (na concordância) pode dizer alguma coisa sobre qualidade. Como Reidsma & Carletta (2008) demonstraram, quando diferentes anotadores "erram" na mesma direção, os erros acabam por transformar-se em concordâncias, o que sem dúvida é um problema. Por outro lado, a discordância na anotação

[6] O analisador PALAVRAS é reconhecidamente um dos melhores e mais ricos para a língua portuguesa.
[7] O Mechanical Turk é um sistema lançado pela empresa americana Amazon em 2005, no qual se contratam pessoas (público em geral, não especialistas) para realizar tarefas intelectuais relativamente simples, mas em casos em que seja preciso fazer muitas vezes para obter volumes quantitativamente satisfatórios. O Mechanical Turk não é necessariamente para pesquisa. Veja em: https://www.mturk.com/mturk/welcome.
[8] Veja-se Tinsley & Weiss (2000) para uma apresentação matemática e Artstein & Poesio (2008) para uma panorâmica extensa no PLN. Contudo, ainda existem bastantes vozes dissidentes, como Powers (2012) e Uebersax (s/d).

não é necessariamente um indicativo de baixa qualidade, sobretudo em áreas onde a interpretação e o conhecimento do contexto e do assunto podem ser mais importantes do que o "palpite".[9] No entanto, achamos que é conveniente quantificar casos em que a nossa própria interpretação vacila, porque isso pode precisamente apontar quer para deficiências da análise proposta, quer para casos de mudança sincrônica em progresso. Por essa razão, essa quantificação é uma das tarefas que propomos.

Para nós, um dos pontos de interesse quando deslocamos o foco da tarefa de anotação da construção de um recurso linguístico-computacional para uma maneira de estudar fenômenos linguísticos é explicitar as diferentes possibilidades de leitura, as diferentes interpretações, isto é, as discordâncias. Interessa-nos também, no estudo de uma língua, a divergência e a alteridade, pois delas também se alimenta a pesquisa, sobretudo nas letras e nos estudos sociais.

3. Conectores condicionais

A ideia base do estudo sobre os conectores condicionais é obter dados quantitativos sobre diferentes tipos de contextos em relação aos conectores condicionais "a + infinitivo" (que é usado com esse sentido sobretudo em português europeu), "se", "caso" e "no caso de". Duzentas ocorrências de cada tipo foram anotadas quanto à sua base modal (epistêmica ou circunstancial), assim como foi marcado se eram "condicionais de enunciação" ou não (cf. Lopes 2009). Simplificadamente, a anotação tem em conta se a relação entre a oração condicional e a frase matriz é uma relação do plano epistémico[10] (como em *se a rua está molhada, [então] choveu*), do plano ontológico (como em *se não houver oxigénio suficiente na água, os peixes acabam por morrer*) ou do plano da enunciação (como em *se tiveres sede, há cerveja no frigorífico*). Esses dados (texto e anotação) foram depois tornados públicos na página da Gramateca.

[9] Um exemplo anedótico é quando a interpretação de um verbo em contexto pode ser significativamente melhorada por conhecer o resto da entrevista ou mesmo conhecer pessoalmente o entrevistado, como descrito em Bacelar do Nascimento et al. (1993).

[10] A distinção entre o plano epistémico, ou dedutivo, o plano ontológico, ou de conteúdo, e o plano da enunciação é observada por Sweetser 1991. Simplificadamente, a construção condicional remete para o plano epistémico se o nexo entre o antecedente e o consequente for um nexo dedutivo (grosso modo, "se p, então q" será equivalente a "a verdade de p permite concluir a verdade de q"); remete para o plano ontológico se o nexo entre antecedente e consequente for próximo do de uma relação de causalidade (grosso modo, "se p, então q" será equivalente a "a situação descrita por p leva à situação descrita por q"); e remete para o plano da enunciação se, simplificadamente, a asserção do antecedente for uma justificação para a asserção do consequente. Para uma explicação mais detalhada sobre a distinção entre estes três planos, ver Sweetser 1991.

Além da anotação relativa ao tipo de condicional, foram também anotados os casos em que os conectores considerados não são operadores condicionais. Por exemplo, "se" pode ser também uma conjunção integrante (como em *ele perguntou se alguém telefonou)* e "a" pode ser um operador aspectual (como em *ele está a cantar*). Nessas ocorrências, a anotação do operador foi marcada como "OUTR". Os resultados teóricos desse estudo foram publicados em Marques (2014).

Esse trabalho anterior de análise, viabilizado no ambiente da Gramateca, foi precioso na possibilidade de permitir fixar os requisitos de especificação de uma ferramenta que permitisse anotar os corpos e torná-los públicos e revisáveis por outros interessados, permitindo tanto uma anotação inicial como anotações alternativas; portanto, anotações múltiplas. Entre as várias lições que aprendemos com esta atividade, destacamos a necessidade de levar em conta erros quer da anotação, quer dos próprios corpos[11], assim como a vantagem de poder obter *a posteriori* mais informação sobre cada ocorrência. Assim nasceu a ferramenta Rêve.

A Figura 1 mostra o Rêve na sua vertente de anotação, em que os revisores podem ter acesso a cada um dos exemplos obtidos e selecionar a categoria que lhes parece correta. Note-se que, de modo a que esses revisores não sejam influenciados por prévias anotações, o sistema não sugere qualquer categoria *por omissão* (ou seja, em inglês, *by default*). Note-se, também, a possibilidade de atribuição de mais de uma classe, simultaneamente, materializando, dessa forma, o conceito de vagueza na língua.

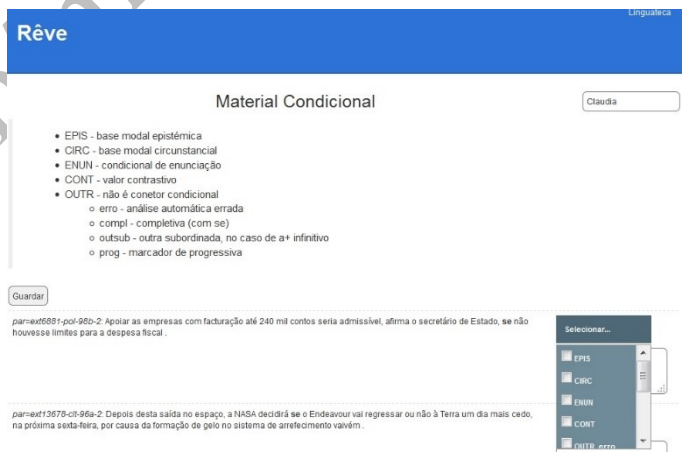
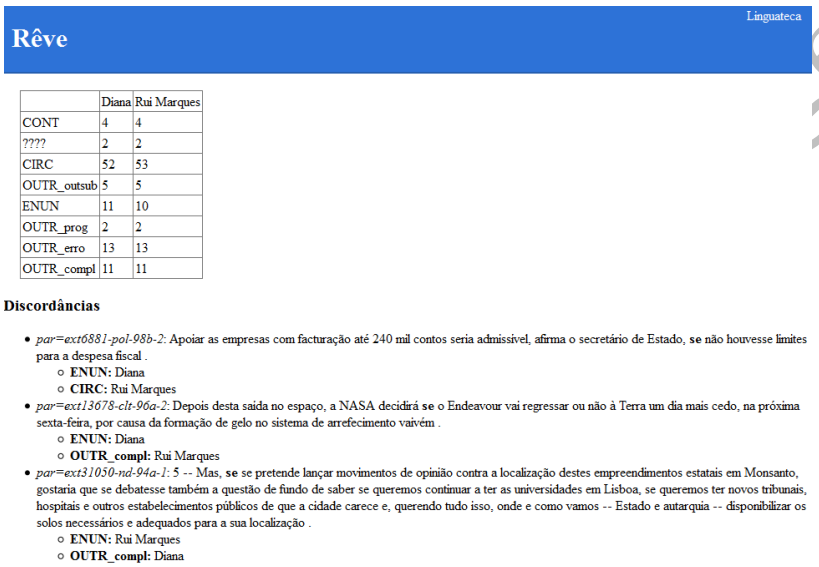


Figura 1. Anotação de exemplos no Rêve.

[11] Mais especificamente: alguns exemplos selecionados automaticamente tinham erros de digitação ou de ortografia, por exemplo.

Posteriormente, os resultados de anotação podem ser analisados e comparados. Se, por um lado, a anotação *concordante* entre diferentes revisores é útil para a obtenção de um recurso anotado; por outro lado, a informação *discordante* é igualmente útil para posterior discussão ou, simplesmente, para a obtenção daquelas situações duvidosas. A Figura 2 mostra o resultado de comparação das anotações relativas aos conectores condicionais, levando em consideração a contabilização independente, isto é, a distribuição de cada classe, por anotador. Veja-se Simões & Santos (2014) para outros exemplos.



Linguateca

Rêve

	Diana	Rui Marques
CONT	4	4
????	2	2
CIRC	52	53
OUTR_outsub	5	5
ENUN	11	10
OUTR_prog	2	2
OUTR_erro	13	13
OUTR_compl	11	11

Discordâncias

- *par=ext6881-pol-98b-2:* Apoiar as empresas com facturação até 240 mil contos seria admissível, afirma o secretário de Estado, **se** não houvesse limites para a despesa fiscal .
  - **ENUN:** Diana
  - **CIRC:** Rui Marques
- *par=ext13678-clt-96a-2:* Depois desta saída no espaço, a NASA decidirá **se** o Endeavour vai regressar ou não à Terra um dia mais cedo, na próxima sexta-feira, por causa da formação de gelo no sistema de arrefecimento vaivém .
  - **ENUN:** Diana
  - **OUTR_compl:** Rui Marques
- *par=ext31050-nd-94a-1:* 5 -- Mas, **se** se pretende lançar movimentos de opinião contra a localização destes empreendimentos estatais em Monsanto, gostaria que se debatesse também a questão de fundo de saber se queremos continuar a ter as universidades em Lisboa, se queremos ter novos tribunais, hospitais e outros estabelecimentos públicos de que a cidade carece e, querendo tudo isso, onde e como vamos -- Estado e autarquia -- disponibilizar os solos necessários e adequados para a sua localização .
  - **ENUN:** Rui Marques
  - **OUTR_compl:** Diana

Figura 2. Comparação de anotações discordantes.

O conjunto total das anotações de um ou de todos os revisores pode ser facilmente exportado para um arquivo de texto com valores separados por caracteres de tabulação, o que permite que esses dados possam ser subsequentemente processados por outras ferramentas de análise estatística, como o R (R Core Team, 2015).

4. Como anotar características de partes do corpo humano: o projeto Esqueleto

Em relação às partes do corpo, a nossa intenção foi usar o Rêve para estudar a possibilidade de consenso sobre algumas questões de interpretação de usos de palavras de partes do corpo humano. Em Freitas *et al*. (2015), foram propostas classes semânticas para as palavras do léxico do corpo humano. Em termos gerais, a proposta consiste em distinguir, por um lado, as palavras do corpo humano que fazem referência, no contexto da frase, ao corpo humano de fato, daquelas que se distribuem por outros campos semânticos. Para esses sentidos não

*corpora*is, propomos treze classes semânticas[12], além da classe genérica "outros". Se algumas classificações correspondem a estabilizações consensuais (por exemplo, a classe "corpo:medida" para *dois **dedos** de pinga*), em outros casos as estabilizações propostas podem não ser tão evidentes. Nesse sentido, o Rêve auxiliou a tarefa de avaliação da classificação proposta e a tarefa de validação dessas mesmas categorias. Para tanto, escolhemos um conjunto de categorias de anotação que nos pareceram especialmente finas e, portanto, com maior potencial para discordância, como: (a) a diferença entre parte de algo ou lugar (virtual), como em *O calibre real é a grandeza medida directamente na **boca** do cano, sem qualquer artefacto, ou no projéctil*.; e (b) a diferença entre expressões que envolvem sentimento ou opinião, como em *Não é daqueles vinhos que se possa **torcer o nariz***). Selecionamos um conjunto de exemplos e pedimos a pessoas interessadas no assunto, mas não necessariamente envolvidas com o projeto de anotação, para os reanotarem, apenas com base nas nossas descrições genéricas das classes envolvidas.

Dois grupos de dados foram assim criados[13]: o Esqueleto 1 e o Esqueleto 2, com cerca de 100 frases cada um, a partir de procura no AC/DC de exemplos que já haviam sido anotados, mas que tinham dado origem a discussões e incertezas, junto com outros casos que pareciam menos complicados. Cada grupo lidava com diferentes classes. Para ver se o contexto da tarefa influenciava o julgamento/escolha dos participantes, alguns exemplos foram incluídos nos dois conjuntos – precisamente 23 exemplos.

Os resultados[14] da anotação, por um lado, confirmaram a complexidade da proposta, o que ficou evidente pela grande discordância; por outro lado, esses mesmos resultados levaram à reformulação de uma das classes – as classes partedeobjeto e membro foram reunidas na classe parte – o que motivou uma segunda rodada de anotação. Na segunda rodada, a concordância obtida subiu para cerca de 85%, um número bastante razoável quando se trata de anotações semânticas. No entanto, para esse cálculo, consideramos também concordância a sobreposição de classes[15], alternativa possível porque já havíamos determinado que uma

[12] As classes semânticas são: sentimento, vegetal, parte, lugar, doença, opinião, posição, animal, movimento, faculdade, grupo, medida e centralidade.
[13] Disponíveis em: http://www.linguateca.pt/reve/stats/8 e http://www.linguateca.pt/reve/stats/10.
[14] Os resultados teóricos provenientes dessas anotações encontram-se em Freitas et al.2015.
[15] Para contabilizar a concordância, que neste caso é parcial, consideramos a sobreposição de classificações como concordante, mas consideramos que cada maneira de classificar corresponde a uma interpretação diferente. O Rêve também calcula outra medida de concordância, a concordância estrita, que só considera concordância se duas (ou mais) classificações/interpretações forem exatamente iguais.

palavra poderia receber mais que uma análise. Assim, a análise da palavra "nervos", na frase abaixo, foi considerada concordante:

> A equipe não controlou os **nervos**, errou passes e insistiu nas bolas altas, e os principais jogadores estiveram muito apáticos.
> **SENTIMENTO:** anotador1, anotador2
> **SENTIMENTO_FACULDADE:** anotador3

A análise das discordâncias, por sua vez, revelou que, em quase todos os casos, uma das categorias selecionadas era a categoria "outros", evidenciando a dificuldade de enquadramento de certas palavras/sentidos nas classes propostas, por um lado, e a sobreposição de sentidos, por outro.

É importante mencionar ainda que casos considerados simples pela equipe do projeto revelaram-se também alvo de controvérsia quando tornados públicos no ambiente do Rêve, indicando que mesmo esses deveriam ser repensados ou ter a sua explicação melhorada.

Especificamente com relação ao Esqueleto, a utilização do Rêve não foi especialmente original , uma vez que uma das intenções era validar o esquema de anotação proposto. Na próxima seção, trataremos de outro tipo de utilização, cujo principal objetivo é ilustrar a divergência.

5. Um estudo das emoções com a Gramateca: os diferentes sentidos de *admirar*

O lado emotivo da língua é uma área que tem recebido algum interesse nos últimos tempos[16], ou seja, o fato de que uma língua não é apenas um veículo de comunicação de fatos, mas também, ou sobretudo, de emoções e de valores. A esse respeito consideramos que foi muito infeliz, para a linguística, a separação logocêntrica entre verdade objetiva e opinião, que pôs a ênfase nos valores de verdade e na informação transmitida[17]. O interesse renovado sobre as emoções na língua não significa que existam soluções ou consenso (ainda), e a Gramateca pretende dar um contributo à investigação desse tema em relação ao português. Usaremos, pois, esse nosso interesse no estudo das anotações emotivas com a intenção de verificar os diferentes sentidos da emoção que podemos representar pela palavra *admirar* e outras que podem significar tanto surpresa quanto apreciação. (Mota e Santos, 2015).

[16] Ver Pang & Lee, 2008, para uma compilação de trabalhos recentes na área.
[17] Opinião essa também corroborada, por exemplo, por Arrojo (1992).

Um dos temas relevantes na análise de sentimentos ou de opiniões é até que ponto essas são consensuais ou se dependem do utilizador/leitor de dada obra ou texto. Em Santos e Mota (2015), estudamos a diferenciação entre duas "emoções" associadas ao lexema *admirar*, uma envolvendo respeito (que consideramos positiva) e outra, surpresa (que pode ser positiva ou negativa). Embora à partida essa diferença seja fácil de compreender, o estudo detalhado e a anotação por várias pessoas diferentes revelaram que não são poucos os casos em que as interpretações podem ser tanto divergentes como ambíguas ou vagas. O exercício com o Rêve foi desenhado para verificar precisamente essa questão: até que ponto uma mesma emoção (admiração) pode carregar dois sentidos com polaridade antagônica? Até que ponto a diferença entre *surpresa* e *respeito* é consensual entre falantes de português? Adicionalmente, seria possível que uma mesma ocorrência em contexto desse margem a interpretações distintas? Nossa aposta era de que a anotação, por diferentes pessoas, de frases selecionadas evidenciasse discordâncias trazendo assim questões teóricas relevantes para o quadro dos estudos de emoção e sentimentos.

Para estudar essas questões, selecionamos 108 casos que continham palavras como *admirar* e *maravilhar.* Desses 108, em 63 houve discordâncias, quantidade considerável. Com relação a *surpresa* e *respeito*, em 20 casos alguma das anotadoras considerou a presença de ambos os sentidos, e apenas eles, na frase (concordância estrita, ver nota 15). Em apenas um caso, todas as anotadoras concordaram quanto à presença de ambos os sentidos, simultaneamente:

> 167581: -- De tudo isso, aventurou Raimundo, o que mais me **admira** é a sua memória: o senhor com efeito tem uma memória de anjo.

Esses dados, facilmente obtidos com o Rêve, corroboram a hipótese inicial sobre a dificuldade, em certos casos, de distinção entre esses sentimentos.

Adicionalmente, a tarefa com o Rêve nos mostrou outros pontos para o estudo dos sentimentos em língua portuguesa:

(1) existem ainda mais sentidos associadas ao lexema *admirar*, o que nos fez redesenhar o experimento e incluir a classe/etiqueta *apreciar*, empregada consensualmente – e exclusivamente[18], em 6 casos, como em:

> 237065: Ele ficou algum tempo a contemplá-la naquela posição, que a fazia mais

[18] Apenas o sentido de "apreciar" foi atribuído.

bonita, e, perdido em saudosas reminiscências da sua mocidade, **admirava** a curva macia dos seios, palpitantes, sob a compressão.

(2) existem casos em que nenhum dos três sentidos está em jogo;

(3) existem casos em que os três sentidos são possíveis. Especificamente, em sete casos, alguma das anotadoras considerou a presença simultânea dos três sentidos, mas não houve concordâncias em quaisquer desses casos; e

(4) não são poucos os casos em que cada avaliador interpretou de uma maneira distinta.

A Figura 3 apresenta a distribuição das classes atribuídas por anotadora, considerando as classificações múltiplas. O cálculo é feito automaticamente pelo Rêve.

Contabilização Combinatória

Nesta contabilização, cada classificação múltipla é considerada como uma categoria independente.

	Claudia	Cristina	Diana
APRECIAR	21	18	9
APRECIAR|NENHUM	1	0	0
APRECIAR|RESP	11	9	20
APRECIAR|RESP|SURP	1	2	4
APRECIAR|SURP	11	2	3
NENHUM	4	3	2
NENHUM|SURP	2	0	0
RESP	17	29	36
RESP|SURP	3	11	6
SURP	25	34	28

Figura 3. Distribuição das discordâncias no estudo das emoções.

Diferentemente do Esqueleto, no caso das emoções, mais do que buscar, com várias rodadas de anotação, a concordância das interpretações em todos os casos, interessou-nos observar a variabilidade de interpretações que palavras de emoção podem ter.

6. Observações finais

O Rêve é uma ferramenta criada para explorar a ideia de pesquisa e aprendizado (humano) por meio da anotação. Está inserido no contexto da Gramateca, que tem como propriedade essencial partilhar publicamente os seus dados e discussões sobre gramática – em sentido amplo – do português. Uma vez públicos, os dados estão disponíveis para que sejam

discutidas correlações e controvérsias, assim, fomentando, por exemplo, o compartilhamento e replicação de experiências.

Partilhamos conhecimento por meio de artigos e capítulos de livros, mas não temos o costume de partilhar dados, especificamente anotações. Como argumenta Xiao (2009) em defesa da anotação, analisar linhas de concordância, classificando-as de acordo com a intuição, é também um processo de anotação, só que um processo implícito. É justamente esse caráter velado, que torna a classificação irrecuperável e, portanto, bem menos confiável que a anotação explícita.

Por isso, o objetivo da ferramenta está menos vinculado ao resultado "convencional" da anotação – o material anotado – que ao processo de anotação propriamente. Pelo mesmo motivo, o Rêve não é, primordialmente, uma ferramenta para medir a concordância entre anotadores, embora isso também seja possível, mas antes um facilitador para a revisão e reanotação de materiais linguísticos variados, cujo objetivo é não apenas torná-los acessíveis, mas também permitir que a sua manipulação e a medição de diferenças conceituais não demandem conhecimentos informáticos muito específicos.

Estamos conscientes de que existem vários ambientes para corrigir anotações ou fazer anotação a partir do zero[19], mas queríamos desenvolver algo que pudesse interagir da melhor maneira possível com a Gramateca, não só quanto à escolha inicial de dados, mas também quanto às informações associadas às frases que já se encontram no AC/DC, e com isso possibilitar o cruzamento de informações diferentes ou estudos de replicação. Neste último caso, pretendemos que, se nova informação – ou simplesmente informação mais atualizada – vier a ser in*corpora*da nos corpos,[20] possamos fazer uso dela no Rêve sem ter de reanotar tudo outra vez. Em outras palavras, após um pesquisador ter passado várias semanas a caracterizar/anotar um conjunto de casos e tê-lo tornado público através da Gramateca e, portanto, passível de ser investigado por outros pesquisadores, também com base em outros critérios, seria certamente desolador que, na próxima versão dos corpos subjacentes à Gramateca, essa informação deixasse de poder ser utilizada ou necessitasse da repetição do trabalho anterior.

[19]Alguns exemplos são o Brat Rapid Annotation Tool (http://brat.nlplab.org), o GATE (https://gate.ac.uk/demos/movies.html#section-1.2.5) e o EtiquetHAREM (http://www.linguateca.pt/aval_conjunta/HAREM/ManualUtilEtiquetHAREM.pdf).

[20] Os problemas que a existência de corpos dinâmicos (melhorando com o tempo) e a anotação de versões antigas desses corpos podem acarretar são discutidos por exemplo por Santos (2014c).

De maneira a garantir a integração do Rêve na Gramateca, um dos nossos critérios de desenho é poder cruzar o estudo da anotação humana com os resultados do processamento automático ou semiautomático. Assim, juntamos estudos qualitativos detalhados, exigindo profundo conhecimento linguístico, a estudos quantitativos abrangendo grandes quantidades de dados, cuja análise humana é impossível.

Referências Bibliográficas

ARCHER, D. *Corpus* annotation: a welcome addition or an interpretation too far? In: TYRKKÖ, J.; KIPIÖ, M.; NEVALAINEN, T.; RISSANEN, M. (Org.). Outposts of historical *corpus* linguistics: from the Helsinki *corpus* to a proliferation of resources. **Studies in Variation, Contacts and Change in English**. Vol. 10, 2012. Disponível em http://www.helsinki.fi/varieng/series/volumes/10/archer

ARROJO, R. **O signo desconstruído**. São Paulo: Pontes, 1992.

ARTSTEIN, R.; POESIO, M. Inter-coder agreement for computational linguistics. **Computational Linguistics**, v. 34, n. 4, p. 555-596, 2008. crossref http://dx.doi.org/10.1162/coli.07-034-R2

BACELAR DO NASCIMENTO, M. F.; MENDES, A.; SANTOS, D. O *corpus* e a classificação sintáctica dos verbos. **Actas do 1.o Encontro de Processamento de Língua Portuguesa (escrita e falada) - EPLP'93** (Lisboa, 25-26 de fevereiro de 1993), pp. 125-129.

CRAGGS, R.; WOOD, M. M. Evaluating Discourse and Dialogue Coding Schemes. **Computational Linguistics** 31, No. 3, September 2005, pp. 289-296. crossref http://dx.doi.org/10.1162/089120105774321109

FISH, S. E.. **Is There A Text in This Class? The Authority of interpretive communities**. Cambridge: Harvard University Press. 1980.

FREITAS, C. de. **Esqueleto**: anotação das palavras do corpo humano. Primeira edição: 15 de novembro de 2013. Disponível em: http://www.linguateca.pt/acesso/Esqueleto/Esqueleto.html

FREITAS, C. de; SANTOS, D.; CARRIÇO, B.; JANSEN, H.; MOTA, C. Investigação do léxico do corpo humano e anotação semântica de *corpus*. **Revista de Estudos da Linguagem** v. 23, n. 3, 2015 (no prelo).

LOPES, A. C. M. Contributos para o estudo de construções condicionais não-canónicas em Português europeu contemporâneo. **Diacrítica, Ciências da Linguagem**, 23/1, pp. 149-169, 2009.

MARQUES, R. Modalidade e condicionais em português. **Revista Virtual de Estudos da Linguagem - ReVEL,** edição especial, vol. 12, n. 8, pp. 106-130, 2014.

MOTA, C.; SANTOS, D. **Emotions in natural language: a broad-coverage perspective**. Janeiro de 2015. Disponível em: http://www.linguateca.pt/acesso/emotionsBC.pdf.

MUNRO, R.; BETHARD, S.; KUPERMAN, V.; TZUYIN LAI, V.; MELNICK, R.; POTTS, C.; SCHNOEBELEN, T.; TILY, H.. Crowdsourcing and language studies: the new generation of linguistic data. In **Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10).** Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 122-130.

PANG, B.; LEE, L.. Opinion mining and sentiment analysis. **Foundations and Trends in Information Retrieval** vol. 2, pp.1-135, 2008. crossref http://dx.doi.org/10.1561/1500000011

POWERS, D. M. W. The Problem with Kappa. In **Proceedings of EACL 2012**, pp. 245-355, 2012.

R Core Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2015. Disponível em: http://www.R-project.org.

REIDSMA, D.; CARLETTA, J. Reliability Measurement without Limits. **Computational Linguistics**, 34(3), pp. 319-326, 2008. crossref http://dx.doi.org/10.1162/coli.2008.34.3.319

RILOFF, E.; WIEBE, J.; PHILIPS, W. Exploiting subjectivity classification to improve information extraction. **Proc. 20th National Conference on Artificial Intelligence (AAAI-2005)**, volume 3, pp. 1106-1111, 2005.

RUMSHISKY, A. Crowdsourcing Word Sense Definition. In **Proceedings of the Fifth Linguistic Annotation Workshop (LAW-V)**. *ACL-HLT 2011*. Portland, Oregon, 2011.

RUMSHISKY, A.; BOTCHAN, N.; KUSHKULEY, S.; PUSTEJOVSKY, J. Word Sense Inventories by Non-Experts. In CALZOLARI, N.; CHOUKRI, K.; DECLERCK, T.; DOGAN, M. U.; MAEGAARD, B.; MARIANI, J.; ODIJK, J.; PIPERIDIS, S.. **Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'12)** (Istambul, 23-25 de Maio de 2012).

SAMPSON, G. **Empirical Linguistics**. London: Continuum, 2001.

SANTOS, D. What is natural language? Differences compared to artificial languages, and consequences for natural language processing. Palestra convidada, **SBLP2006** e PROPOR'2006, Itatiaia, RJ, Brasil, 15 de Maio de 2006.

SANTOS, D. Corporizando algumas questões. In: TAGNIN, S. E. O..; VALE, O. (Orgs.). **Avanços da Linguística de *Corpus* no Brasil**. São Paulo: Editora Humanitas/FFLCH/USP, 2008, pp.41-66.

SANTOS, D. Podemos contar com as contas?. In ALUÍSIO, S.; TAGNIN, S. E. O.. **New Language Technologies and Linguistic Research:** A Two-way Road. Cambridge Scholars Publishing, 2014, pp. 194-213.

SANTOS, D. Gramateca: *corpus*-based grammar of Portuguese. In BAPTISTA, J.; MAMEDE, N.; CANDEIAS, S.; PARABONI, I.; PARDO, T. A. S.; VOLPE NUNES, M. das G. **PROPOR 2014**, LNAI 8775. Springer, Heidelberg (2014), pp. 214-219. crossref http://dx.doi.org/10.1007/978-3-319-09761-9_24

SANTOS, D. PoNTE: apontando para corpos de aprendizes de tradução avançados. **Linguamática** 6, 1, 2014, pp. 69-86.

SANTOS, D. Comparando corpos orais (transcritos) e escritos na Gramateca. In: BARDEL, C. **Proceedings from the conference Parler les langues romanes/Parlare le lingue romanze/Hablar las lenguas romances/Falando línguas românicas GSCP 2014**, University Press Università di Napoli L'Orientale, 2015.

SANTOS, D.; MOTA, C. A admiração à luz dos corpos. In SIMÕES, A.; BARREIRO, A.; SANTOS, D; SOUSA-SILVA, R.; TAGNIN, S. E. O. **Linguística, Informática e Tradução:** Mundos que se Cruzam. Homenagem a Belinda Maia, OSLa, Vol 7, No 1, 2015, pp. 57-77.

SILVA, R.; SANTOS, D. **Arco-íris**: notas sobre a anotação do campo semântico da cor em português. Disponível em: http://www.linguateca.pt/acesso/arcoiris.pdf.

SIMÕES, A.; SANTOS, D.. Nos bastidores da Gramateca: uma série de serviços. **Workshop on Tools and Resources for Automatically Processing Portuguese and Spanish**, at PROPOR 2014, São Carlos, Brasil, 9 de outubro de 2014, pp. 97-104, 2014.

SWEETSER, E. **From Etymology to Pragmatics**: Metaphorical and Cultural Aspects of Semantic Structure. Cambridge University Press, 1991.

TINSLEY, H. E. A.; WEISS, D. J. Interrater Reliability and Agreement. In: TINSLEY, H. E. A.; BROWN, S. D. (org), **Handbook of Applied Multivariate Statistics and Mathematical Modeling**, San Diego, CA: Academic Press, 2000, pp. 95-124. crossref http://dx.doi.org/10.1016/B978-012691360-6/50005-7

UEBERSAX, J. **Kappa Coefficients**: A Critical Appraisal. Disponível em: http://www.john-uebersax.com/stat/kappa.htm.

XIAO, R. Theory-driven *corpus* research: Using *corpora* to inform aspect theory. In: LÜDELING, A.; KYTÖ, M. ***Corpus* Linguistics**: An International Handbook. Berlin: De Gruyter, 2009, volume 2, pp. 987-1.008.

Artigo recebido em: 29.03.2015
Artigo aprovado em: 20.04.2015

A relevância da Web como *corpus* para a identificação de padrões de lexicalização: o caso de "bater+SN" no português brasileiro

The usefulness of the Web as *corpus* for the identification of lexical patterns: The case of "bater+NP" in Brazilian Portuguese

Milena de Uzeda Garrão*

RESUMO: Nesse estudo longitudinal, e inicialmente com base em *corpora* jornalísticos, caracterizamos padrões de lexicalização do tipo Bater+SN, uma vez que: i) identificamos um crescente status de transitividade direta do verbo "bater" tanto no Português Brasileiro, quanto no Português Europeu (como construções do tipo bater o oponente, um caso em que o verbo ganharia uma semântica de "derrotar"); ii) identificamos uma tipologia sintático-semântica do padrão Bater +SN bastante peculiar, como "Bater um medo", que se distanciava de combinação livre ou de uma coocorrência sintática aleatória (descrita em i); iii) curiosamente, constatamos que esse padrão, que defendemos ser análogo a expressões com verbo-suporte, também se distanciava do fenômeno comumente rotulado como expressão fixa, cristalizada, caracterizado por revelar um alto grau de opacidade semântica (como bater as botas). Também questionamos o nível de previsibilidade da metodologia utilizada para a identificação das expressões fixas, (descritas em iii). Já em 2013, com base na Web como *corpus*, pudemos notar uma mudança da restrição semântica do padrão descrito em ii. Portanto, nesse estudo, caracterizamos o quanto tanto a descrição como a metodologia de identificação de padrões lexicais V+SN merecem ser revistas com base no poder empírico da Web como *corpus*.

PALAVRAS-CHAVE: Web como *Corpus*. Lexicologia. Expressões Fixas. Bater+Sintagma Nominal.

ABSTRACT: Based on a Brazilian Portuguese jornalistic corpus, we described in 2001 and 2003 lexicalization degrees of "Bater+Noun Phrase" pattern. Three reasons motivated our study: i) a new transitivity status of verb "Bater" both in Brazilian and European Portuguese (eg.: "bater o oponente"– "to beat the opponent" – where the verb would take in the meaning of "to defeat"); ii) a particular semantic pattern of "Bater +NP", such as "bater um medo" ("to get frightened") which would not conform to a random syntactic cooccurrence, described in i); iii) this pattern, which we claim to behave as support-verb expressions, would also not conform to what is commonly labeled as fixed expressions, which are claimed to have a high degree of semantic opaqueness (such as "bater as botas" or "to kick the bucket"). We also claim that this framework had drawbacks if we rely on corpus evidences. Then, in 2013, in a Web corpus based approach, we could spot a semantic change regarding the pattern described in ii). Therefore, in this longitudinal study, we take a critical stance of the framework chosen along these years for multi-word expressions identification and description, and we portray the Web as corpus as a powerful empirical tool for that purpose.

KEYWORDS: Web as Corpus. Lexicology. Multi-word Expressions. "Bater"+Noun Phrase.

* Doutora em Estudos da Linguagem (PUC-Rio), Professora Adjunta do Instituto Multidisciplinar, UFRRJ.

1. Nosso percurso

O interesse por expressões com padrão "Bater+SN" surgiu em função de uma visão inclusiva de um léxico marginalizado para o seu tratamento automático. A motivação inicial do estudo, entre os anos de 2001 (em Garrão e Dias) e 2003 (em Basílio, Oliveira e Garrão), era a de identificação de estruturas lexicalizadas com esse padrão para que pudessem figurar em um dicionário eletrônico. A opção pelo verbo *bater* se justificava em função da detecção do seu alto grau de ocorrência encabeçando o que se convencionou chamar de expressões idiomáticas (como *bater as botas, bater perna, bater pino, bater boca, bater os olhos, bater o martelo, bater ponto, bater papo*, entre outras).

Curiosamente, no ano de 2001, em uma das buscas em *corpora* jornalísticos, até então ainda não etiquetados, foi possível identificar o seguinte fato: no Português do Brasil (PB, doravante) havia, além das expressões supracitadas, um tipo de expressão encabeçada por esse verbo, mas que, no entanto, não deveria ser considerada propriamente uma expressão idiomática ou cristalizada, tampouco uma unidade lexical. Tratava-se de construções como *bater um medo, bater uma fome*, que, por sua vez, apresentavam um comportamento sintático-semântico diverso, uma vez que seus constituintes pareciam apresentar uma função composicional para o todo. Em outras palavras, as partes da expressão pareciam contribuir para o seu significado global. E o que gerou uma maior motivação na descrição desse padrão foi o fato de figurarem em textos escritos com registro semiformal.

Além desse novo padrão, também notamos, nessa mesma época, que o verbo *bater* vinha ocupando, não somente no PB, mas também no português europeu, um status de transitividade direta muito notório (como sinonímia de superar; ex: *bater o concorrente*). Portanto, dentro desse quadro empírico, passamos a conceber a possibilidade de detecção de níveis de lexicalização de expressões com o padrão Bater+SN.

Como fonte para a identificação desse padrão, utilizamos em 2001 o primeiro jornal brasileiro a ter sua versão online, *o Jornal do Brasil*, e já em 2003, as versões online do jornal *O Globo* e a da revista *Veja*. Nesse mesmo ano, já contávamos com um *corpus* jornalístico etiquetado disponibilizado pelo grupo NILC-São Carlos: o *corpus* da Folha de São Paulo.

O padrão Bater +SN que nos interessou na época, portanto, revelava-se como construções extremamente comuns em PB, encontradas não somente em discurso oral como também em textos escritos com registro semiformal (como *bater uma fome, bater um desespero, bater uma dúvida*). Na época, propusemos que esse tipo de construção tinha um comportamento análogo

ao de expressões com verbo-suporte (como *dar um riso; fazer compras*), uma vez que pudemos identificar um comportamento sintático-semântico comparável àquelas construções quando submetidas aos testes de níveis de lexicalização. Traçamos, para tanto, um vetor de nível de lexicalização das expressões Bater+SN, com base nos *corpora* e em uma metodologia de detecção de expressões lexicalizadas proposta por Neves (1999).

Como os testes descritos em Neves (adaptados para o português de Radford, 1988:90) se propunham a distinguir níveis de lexicalização tendo como base o padrão V+SN; isto é, visavam contrastar mais especificamente construções livres, construções com verbo-suporte e construções fixas, cristalizadas, lançamos mão dessa mesma metodologia para aferir os diferentes graus de lexicalização de Bater+SN. Portanto, com nítida inspiração no artigo de Neves, pudemos esquematizar o seguinte vetor de nível de lexicalização:

Quadro 1. Identificação de graus de lexicalização do padrão Bater+SN com base em *corpora* jornalísticos.

+ sintaxe		+ léxico
construções livres	***construções com verbo-suporte***	***expressões cristalizadas***
(bater transitivo- "derrotar")	(Bater+SN sentimento de falta)	(Bater+SN fixo)
Bater o concorrente	*Bater uma saudade*	*Bater as botas*
	(Bater+SN sentimento negativo)	*Bater perna*
	Bater um medo	

Fonte: elaboração própria.

Na extrema esquerda, teríamos combinações com verbos plenos e sintagmas nominais complementos, que são completamente livres, onde os dois elementos exercem papéis independentes na estrutura argumental (*bater o concorrente*); na extrema direita, expressões que constituem um significado unitário, em que "nem mesmo parece ser possível postular um SN em posição de objeto" (Neves, 1999:99), (como *bater as botas*); e entre estes dois graus extremos de construção, há aquelas construções intermediárias, constituídas dos chamados verbos-suporte, que, por sua vez, recebem certo grau de esvaziamento do sentido lexical, porém, semanticamente, contribuiriam para o significado total da construção.

1.1 Mas o que viria a ser o grau intermediário de lexicalização?

Em relação ao vetor de nível de lexicalização, pudemos constatar que, embora as expressões que se encontrassem na sua extrema esquerda fossem consideradas livres e até imprevisíveis, a distinção entre os dois outros tipos de construções traria uma certa hesitação pelo fato de ambas se situarem no domínio da convencionalidade ou, nas palavras de Neves (1999:103), no domínio "das estruturas recorrentes que o falante escolhe com reduzida liberdade quanto ao modo de composição".

Portanto, as construções com verbo-suporte foram consideradas por Neves como intermediárias por ora se situarem mais próximas de construções livres, ora mais próximas de expressões cristalizadas. Isto é, por vezes mais próximas de um; por outras vezes mais próximas de outro extremo do vetor. Compõem-se de: um verbo com determinada natureza semântica básica que funciona como instrumento morfológico e sintático na construção do predicado; e um SN que entra em composição com o verbo para configurar o sentido do todo, bem como para determinar os papéis temáticos da predicação.

Em conformidade com essa característica, observamos do ponto de vista sintático-semântico, uma restrição do SN posposto ao verbo bater, a saber: a) determinante intensificador + sentimento de falta (como, por exemplo, *uma fome, a maior saudade, uma dúvida*) e b) determinante intensificador + sentimento negativo *(*como *uma baita tristeza, o maior desespero, um medo).*

Note-se que se a análise for feita sob um aspecto puramente formal, as expressões Bater+SN que figuram no meio do vetor apresentado na Figura 1, não teriam um comportamento exato do que é chamado tradicionalmente de construções com verbo-suporte (como *dar um riso, fazer compras, dar uma olhada*), pois teoricamente a sua construção sintática seria *bater em alguém uma dúvida / um medo*. Contudo, propusemos em 2003 que o seu comportamento semântico seria análogo ao dessas construções com base nos seguintes argumentos:

De acordo com Cruse (1986) e Radford (1988), os critérios definidores para uma suposta construção sintática tender a um alto grau de lexicalização, ou seja, se enquadrar na extrema direita do vetor, são a sua resistência aos seguintes rearranjos sintáticos: substituição das partes, posposição, coordenação, inserção de constituinte e elipse. Tendo tais critérios como pressuposto, e se focarmos o meio do vetor, fica bastante evidente que as construções do tipo *bater um medo*, por exemplo, não se comportariam como cristalizadas, já que parecem admitir

coordenação (*bate um medo e uma dúvida*), posposição (*um medo bate*), inserção de constituintes (*bateu uma pontinha de dúvida*), assim como o SN pode funcionar como fragmento da oração ou um constituinte sintagmático (ex.: *bate uma dúvida? Não, um medo*). E, seguindo Neves (1999), é exatamente esse comportamento que revelam as expressões com verbo-suporte, como *fazer compras* e *dar um riso*.

2. Nossos percalços: implicações do método utilizado

Até agora vimos que os instrumentos mais seguros para determinar a estrutura dos constituintes de expressões fixas seriam a resistência aos seguintes movimentos sintáticos: a distribuição, a posposição, a coordenação, a intercalação de advérbios, a elipse.

O semanticista britânico David Cruse (1986) atribui ao grau máximo de lexicalização, ou seja, às expressões que propusemos pertencer à extrema direita do vetor, o rótulo de "expressão idiomática". O autor afirma que os itens que a compõem não contribuem para o significado total da expressão. A checagem desse nível máximo de opacidade semântica é possível através dos testes descritos acima para se comprovar a impossibilidade de se compreender a expressão sem que todos os seus itens estejam presentes na ordem original (é o caso do exemplo clássico *bater as botas*).

Dentro dessa perspectiva, aplicamos os testes de Neves (1999) para diferenciar as construções cristalizadas das construções com característica de verbo-suporte do tipo *Bater+SN*. Contudo, identificamos que as expressões que seriam consideradas pelos testes como indevassáveis, ou seja, cristalizadas, admitiam, segundo os dados analisados dos *corpora O Globo, Veja* e *Jornal do Brasil on-line*, inserção de constituinte. Propusemos, portanto, um outro teste a fim de demonstrar que somente os testes de Neves não dariam conta de um padrão revelado pelos *corpora*: era o caso de *bater (muita) perna, bater(um) papo, bater (uma) bola*. Chegamos à conclusão, em contrapartida, de que outras expressões cristalizadas como *bater os olhos, bater as botas* pareciam não admitir a inserção de outros constituintes. Percebemos, na época, que isso parecia estar intimamente relacionado ao perfil semântico destas expressões, que apresentavam um aspecto **pontual**. Portanto, o perfil pontual dessas expressões parecia bloquear a possibilidade de marcação de frequência ou quantificador.

Portanto, já em 2006, argumentamos que mesmo esses casos considerados mais nitidamente impermeáveis do ponto de vista semântico não são tão indivisíveis se recorrermos a evidências de *corpus*. Ou seja, os testes de opacidade semântica não seriam teoricamente

conclusivos porque: 1) os avaliadores são linguistas e não falantes sem pretensões ou interesses em relação a uma teoria de expressões cristalizadas; 2) a noção de *opacidade* versus *transparência semântica* é escorregadia e também carece de uma delimitação teórica precisa e incontroversa.

2.1 Quando o *corpus* fala mais alto

Em 2006, em uma busca ao sistema de Recuperação de Informação *Google*™, ou da Web como *corpus*, pudemos constatar um fato curioso. A expressão *bater a caçuleta* também é utilizada no PB (principalmente no nordeste do país) com o mesmo sentido de *bater as botas*, o que se faria supor que o SN da expressão não seria tão fixo quanto se imagina: "E o Doutor Morte finalmente bateu a caçuleta", (*Casseta & Planeta online,* acesso em junho de 2005*).* O mesmo pudemos questionar em relação às expressões *bater perna* e *dar o braço a torcer.* Se, assim como *bater as botas*, fossem, de fato, expressões indecomponíveis, a inserção de qualquer constituinte tornaria as expressões literais, mas os exemplos a seguir parecem contra-argumentar tal afirmação:

> Resumindo, quem quiser economizar, ou fica em casa, ou vai ter que **bater muita perna**_____para achar onde comer e onde ficar. (http://www.bemtevibrasil.com.br/diarioviagem18.htm)

> Para não dizerem que sou um fanático *apenas* pela aviação militar, **dei meu braço a torcer** e consegui alguns interessantíssimos anúncios de companhias americanas. (http://www.jetsite.com.br/aviacao_mkt.asp)

Ambas as expressões acima teriam comportamento idiomático nos testes, mas sua utilização pelo usuário da língua parece bem mais flexível. No ponto de vista de Cruse (1986), a expressão idiomática é uma unidade lexical elementar: “embora consista em mais de uma palavra, apresenta uma coesão interna de palavras simples” (p. 38). Embora o autor considere Expressões Idiomáticas, Metáforas Cristalizadas e Colocações como tipos de expressões cristalizadas distintas, reconhece que há casos limítrofes. Mas como, então, teorizar sobre um fenômeno que é escorregadio?

Vale (2001, p. 16), numa proposta de tipologia de *expressões cristalizadas* para o PB, também expõe sintomaticamente a arbitrariedade da intuição do pesquisador em relação aos testes de composicionalidade. A sua argumentação deixa clara a falta de força teórica distintiva

entre opacidade e transparência semântica. Ao explicar a aplicação dos testes, recorre ao uso de "asterisco para inaceitabilidade; ponto de interrogação para aceitabilidade duvidosa; dois pontos de interrogação para aceitabilidade ainda mais duvidosa do que a precedente; três pontos de interrogação para aceitabilidade no limite da inaceitabilidade". Sua tentativa parece ser um sintoma de que não há como teorizar sobre as noções de opacidade/transparência semântica.

Trata-se, portanto, de intuição linguística do pesquisador; mas não de uma regra. Dentro dessa análise, pode-se dizer também que, quando Neves (1999) avalia que a expressão "tomar partido", em "Valéria *tomou partido* da tia" (p.99), seria uma expressão cristalizada, está, na verdade, desconsiderando o fato de a construção admitir intercalação de advérbio dependendo do teor aspectual da frase em que se insere. Segundo a autora, a expressão não admite inserção de nenhum tipo de constituinte. O mesmo ela afirma em relação à expressão "ter cabeça" em "O capitão Aparício *tem cabeça* para tudo". O Google, contudo, recupera contra-exemplos para o que a autora propõe:

> A Quarta é um meio termo, uma sinfonia que parece não **tomar <u>muito</u> partido** desta relação, pois está montada sobre um afresco extremamente original de ... (www.mnemocine.com.br/filipe/ensaios.htm)
>
> Tem que ser um exame de nível nacional para entrar quem **tem <u>mais</u> cabeça**, ...(www.museudapessoa.net/MuseuVirtual/hmdepoente/depoimentoDepoente.do?action=ver&idDepoente=63&key)

Portanto, ao longo de nosso estudo, notamos que muitas das construções do tipo Bater+SN por nós identificadas, seriam alocadas na extrema direita do vetor de lexicalização por serem consideradas pelos testes de composicionalidade propostos pela autora — substituição, coordenação, posposição, elipse e inserção de constituinte — como uma unidade indevassável. O que nos surpreendeu foi o fato de esse conjunto de testes ser insuficiente para dar conta de uma boa parte dessas construções, já que algumas delas parecem admitir intercalação de intensificador ou marcador de frequência, conforme detectamos até mesmo antes de recorrer à Web como *corpus*, mas ainda com base nos *corpora* etiquetados a) Cetenfolha e b) Portugal Natura Publico, respectivamente:

> a) Quem pretende ter peixe à mesa durante a Semana Santa precisa **bater <u>muita</u>**

perna.
b) Para os comunistas, o grave é que não estão em condições de **bater <u>demasiado</u> o pé**.

Os exemplos acima demonstram que há expressões supostamente indevassáveis que admitem inserção de advérbio, mais especificamente, de um marcador de frequência ou um intensificador, o que, ao menos, nos faria ampliar os testes geralmente feitos para detectar o teor de fixidez dessas construções. É importante ressaltar, ainda, que a sua alegada opacidade semântica parece não ser definidora do nível de indivisibilidade da expressão, visto que há expressões que admitem marcador de frequência e cujos constituintes não parecem ter o que se costuma chamar de papel composicional (como *bater (muita) perna, bater (muita) boca, dar (muita) trela, fazer (muita) questão).*

Pudemos constatar também que o aspecto verbal da expressão como um todo parece ser muito mais preditivo em relação à sua fixidez do que a sua suposta opacidade, uma vez que tais construções verbais com aspecto pontual tenderiam a um grau de fixidez elevado (*bater o martelo/?bater muito martelo*) e aquelas com aspecto durativo seriam menos rígidas (*bater boca/ bater muita boca*). Mas observamos também que isso parece ser um padrão de comportamento, uma tendência, e não uma regra.

3. O que fazer, então?

As conclusões a que chegamos em 2.1 nos impulsionaram a concluir que caracterizar deterministicamente uma expressão como cristalizada ou fixa é elevar um olhar dedutivo, baseado na intuição, a uma supremacia que talvez não mereça. Enquanto se priorizar uma abordagem dedutiva do pesquisador, que se pretende capaz de caracterizar a expressão tendo por base a sua própria intuição, estaremos ignorando o fato de que é o discurso do falante desavisado, sem pretensões nem comprometimentos teóricos, a fonte mais segura para tanto.

Portanto, em Garrão e Dias (2006), tendo como inspiração o artigo *"I don´t believe in word senses*" (2000), de Adam Kilgarriff, um estudioso do léxico do ponto de vista computacional, e seguindo uma visão semântica dependente de *corpus*, discutimos o problema sintomático de profusão de rótulos para identificação de fraseologias do tipo V+SN, como as terminologias recorrentes na área de semântica lexical: colocações, expressões fixas, expressões idiomáticas, metáforas cristalizadas, em função da total dependência dessas rotulações à intuição semântica do pesquisador, muitas vezes, falha. Utilizamos os próprios

contra-exemplos dos *corpora*, apresentados em 2.1, para refutar a abrangência e demonstrar a parcialidade dos testes propostos na literatura fraseológica.

Defendemos, além disso, que o problema de explosão terminológica no domínio multivocabular seria sintoma da impossibilidade de se chegar a um modelo teórico que pudesse dar conta desse conceito (ou não-conceito).

3.1. Web como *corpus* e a mudança semântica de construções Bater+SN suporte

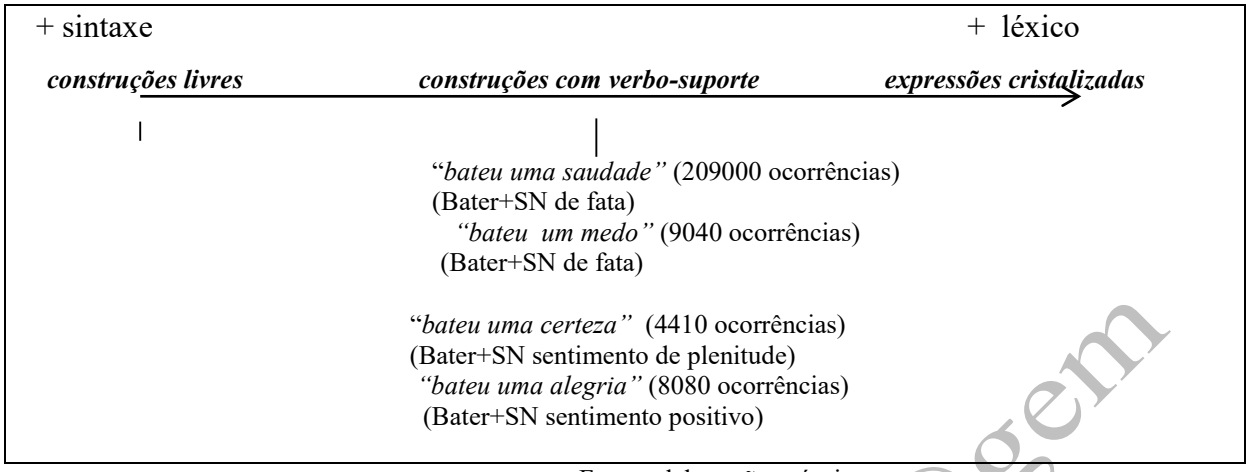
Já em 2013, cientes de que não podemos fazer aferições sobre o comportamento dessas expressões sem verificação em *corpora* e com base nos pressupostos de Stefan Diemer em seu contundente artigo *Corpus linguistics with Google?* (2011), discutimos a tão polêmica utilização de *corpora* abertos para considerações de ordem descritiva na língua. Para tanto, levamos em consideração suas implicações e benefícios (cf. Davies, 2011), tendo como foco de discussão agora, o status intermediário do vetor esboçado na seção 1 ou, mais especificamente, o que propusemos como expressão com verbo-suporte com padrão Bater+SN.

Com o intuito de verificar a manutenção do status de restrição semântica do SN posposto ao verbo nessa construção, que foi nosso objeto de estudo em 2001 e 2003, partimos para a observação de sua ocorrência em *corpora* abertos.

Ou seja, apenas com a mais corriqueira e popular ferramenta de busca, identificamos, dessa forma, a importância de uma busca em *corpora* abertos ou da Web como *corpus* para a revisão da restrição sintático-semântica do SN posposto ao verbo "bater" proposta anteriormente, caracterizada como: a) determinante intensificador + sentimento de falta (como, *uma fome, a maior saudade, uma dúvida*) e b) determinante intensificador + sentimento negativo *(*como *uma baita tristeza, o maior desespero, um medo).*

Observamos, em relação ao que tínhamos detectado em 2001 e 2003, contra-exemplos do SN posposto ao verbo. Ou seja, já existem evidências sólidas de que os falantes ampliaram a restrição em a) para determinante intensificador + sentimento de falta **mas também de plenitude** (com ocorrências como *bater uma certeza*, *bater um insight)* e b) determinante intensificador + sentimento negativo **mas também positivo** *(*como *bater uma alegria, bater uma felicidade*). Portanto, os dados recuperados pela Web como *corpus* revelam que a restrição semântica claramente evidenciada pelos *corpora* jornalísticos nos nossos primeiros estudos não procede mais 10 anos depois.

Quadro 2. Uso de *corpora* abertos para ratificação da mudança semântica detectada primeiramente em discurso oral informal.



Fonte: elaboração própria.

3.2 Em defesa do uso da Web como *corpus*

Alguns resultados da investigação implementada por Diemer (2011, p. 5) em relação às implicações linguísticas do *corpus* aberto foram as seguintes: i) imprecisão do tamanho do *corpus* (um comprometimento quantitativo); ii) organização dos dados (uma vez que existe claramente uma dificuldade de busca por lema ou *POS* – partes do discurso), iii) linguagem lúdica (o que o autor chama de "*playful use of language"*) e iv) facilitação sintática. Estas duas últimas características de escolhas linguísticas inovadoras calcadas na oralidadede, encontradas principalmente em ambientes digitais como Blogs, Facebook e Twitter.

Contudo, também concordamos com Diemer (2011; p. 11) quando avalia que o *corpus* mais revelador e confiável é a Web em detrimento a *corpora* fechados sempre que o foco de estudo for o uso real dos falantes. Portanto, para o objeto de estudo dessa etapa da pesquisa, o uso da Web como *corpus* serviu para ratificar que o falante não utiliza a mesma restrição de negatividade ou falta para expressões Bater+SN suporte. A Web foi também fundamental para a ratificação de que a flexão preferencial do verbo que encabeça a expressão em análise é a terceira pessoa do singular no pretérito perfeito, um padrão flexional que havia sido identificado informalmente em linguagem oral.

Dessa forma, passamos a entender que o uso da Web como *corpus* é extraordinariamente mais fiel àquilo que pode ser considerado como dados da intuição de uma comunidade linguística. Diemer (2011, p. 3) ressalta que para muitas buscas, o *corpus* etiquetado é desnecessário, como identificação de novos prefixos. Concluímos, também, que para outras caracterizações empíricas da língua, a Web como *corpus* é valiosa, como:

i) Para identificação de padrão negativo, ou seja, para detecção ou não de um suposto fenômeno linguístico, como implementado em 2.1, em relação às supostas expressões fixas ou cristalizadas do tipo Bater+SN.

ii) Para corroborar evidências esparsas ou isoladas não advindas de *corpus*, como implementado em 3.1, em relação às construções com verbo-suporte do tipo Bater+SN.

Defendemos, portanto, a ideia de que o domínio aparentemente labiríntico da Web como *corpus* é, indubitavelmente, revelador e equivocadamente menosprezado por muitos pesquisadores que se dedicam ao campo da descrição linguística. Acreditamos, assim, que não existem verdades absolutas ou caminhos únicos, irrevogáveis, na Linguística de *Corpus*. É o foco de análise linguística que recorta o caminho mais produtivo para determinado fim; e o nosso foco de análise procurou legitimar o uso da Web como *corpus*.

Portanto, podemos também distinguir aqui, e de forma contundente, a força do pensamento saussureano uma vez que o uso de *corpus* é a concretização da aplicabilidade de duas de suas célebres constatações: a de que "a linguagem é multiforme e heteróclita" e que, portanto, se deixa avaliar de diferentes formas e, primordialmente, a de que "é o ponto de vista que cria o objeto".

Por fim, ratificamos também que a Linguística de *Corpus* revela claramente a perspicácia do pragmatismo radical da virada linguística de Wittgenstein (1953) quando afirma em Investigações Filosóficas que "o significado está no uso". Conforme argumentamos em Garrão (2006), por uma perspectiva wittgensteiniana da linguagem, podemos enxergar a fusão de dois domínios da Linguística: a Semântica e a Pragmática, ou conhecimento linguístico e conhecimento enciclopédico. E por essa visão negar "a vocação representacionista da linguagem tão defendida por filósofos como Platão, Aristóteles e Locke" (Garrão, 2006:136), ou de uma visão entitativa do significado, podemos supor que o único caminho legítimo para verificação da prática desse uso é o seu registro exaustivo; ou seja, é a Linguística de *Corpus*.
E é notável como esse domínio sintetiza de forma curiosa, mas ao mesmo tempo eloquente, um diálogo entre pensadores da linguagem que se debruçaram sobre o fenômeno da significação, mesmo que de forma díspare. Portanto, não seria nenhum exagero afirmar que se trata de um novo paradigma para a ciência da linguagem.

Referências bibliográficas

AIRES, R. V. X.; ALUÍSIO, S. M. Criação de um *corpus* com 1.000.000 de palavras etiquetado morfossintaticamente. **Série de Relatórios do NILC**, NILC-TR-01-8. 2001

BASÍLIO, M.; OLIVEIRA, C.; GARRÃO, M. U. A Não-Delimitação das Unidades Lexicais. In: Claudio Cezar Henriques. (Org.). **Linguagem, Conhecimento e Aplicação**. 1ed. Rio de Janeiro: Europa, 2003, v. Vol. 1, p. 137-148.

CRUSE, D. **Lexical Semantics**.Cambridge, Inglaterra: Cambridge University Press. 1986.

DAVIES, M. **The *Corpus* of Contemporary American English (COCA) and Google / Web as *Corpus*.** 2011 Disponível em http://view.byu.edu/coca/compare-google.asp Acesso em 14 de julho de 2013.

DIEMER, S. ***Corpus* Linguistics with Google?** Saarland University, Alemanha. 2011. Disponível em http://www.bu.edu/isle/files/2012/01/Stefan-Diemer-*Corpus*-Linguistics-with-Google.pdf . Acesso em 20 de julho de 2013.

GARRÃO, M. U.; DIAS, M.C.P. Um Estudo de Expressões Cristalizadas do tipo V+SN e sua Inclusão em um Tradutor Automático Bilíngüe (português/inglês). **Cadernos de Tradução** (UFSC), Florianópolis, v. v.2, n.VIII, 2001, p. 165-182.

GARRÃO, M. U.; DIAS, M.C.P. The *corpus* never lies: a statistical approach for the identification of verbal collocations. In: Marja Nenonen, Simikka Niemi. (Orgs.). **Studies in Language** 41 - Collocations and Idioms 1. Joensuu: Unversity of Joensuu, 2006, v. 41, p. 354-362.

GARRÃO, M. U. Lingüística de Córpus: o lugar da fusão entre semântica e pragmática. **Calidoscópio** (UNISINOS), v. 04, p. 135-140, 2006.

JACKENDOFF, R. **The Architecture of lhe Language Faculty**. Cambridge, Massachusetts: MIT Press. 1997.

KILGARRIFF, A. **I don't believe in word senses.** 2000. Disponível em http://www.kcl.ac.uk/humanities/cch/seminar/99-00/seminar_kilgarriff.html Acesso em 8 de outubro de 2005.

NEVES, M. H. M. A delimitação das unidades lexicais: o caso das construções com verbo-suporte. In Basílio, M. (org.) A delimitação de unidades lexicais**. Palavra** n° 5. Rio de Janeiro: Departamento de Letras da PUC, 1999, p.98-114.

RADFORD, A. **Transformational grammar**: a first course. Cambridge, Inglaterra: Cambridge University Press, 1988. crossref http://dx.doi.org/10.1017/CBO9780511840425

SAUSSURE, F. (1916) **Curso de Linguística Geral.** Cultrix: São Paulo, 1975.

VALE, O. **Expressões Cristalizadas no português do Brasil**: uma proposta de tipologia. Tese de Doutorado, Araraquara: UNESP. 2002.

WITTGENSTEIN, L. (1953) Investigações Filosóficas. **Coleção Os Pensadores**, São Paulo: Abril Cultural, 1979.

Artigo recebido em: 30.03.2015
Artigo aprovado em: 21.06.2015

Extração automática de candidatos a termos do *Curso de Linguística Geral* com apoio de recursos da Linguística de *Corpus* e do Processamento de Linguagem Natural[1]

Automatic extraction of term candidates from *Course in General Linguistics* with resources from *Corpus* Linguistics and Natural Language Processing

Maria José Bocorny Finatto[*]
Lucelene Lopes[**]
Alena Ciulla[***]

RESUMO: Este trabalho apresenta um estudo em que técnicas de Processamento de Linguagem Natural (PLN) e de Linguística de *Corpus* (LC) são utilizadas para extrair e estruturar termos relacionados a conceitos importantes de Saussure no texto em português do *Curso de Linguística Geral* (CLG). Tomando o CLG como um *corpus*, busca-se um método de representação automática de conteúdo através de ferramentas computacionais. Uma vez submetido ao *parser* PALAVRAS, um etiquetador morfossintático para a língua portuguesa, o *corpus* do CLG é processado pela ferramenta extratora de sintagmas nominais relevantes, denominada ExATOlp, que implementa diversas técnicas de PLN de base linguística e de base estatística. Em seguida, são geradas listas e gráficos hierarquizados dos sintagmas nominais do CLG, elencados pela ferramenta como os mais específicos/relevantes do *corpus* em questão. Esses resultados são comparados com dados gerados pela ferramenta AntConc, ferramenta de acesso livre bastante empregada em trabalhos de LC, aplicada ao mesmo *corpus*. Os resultados mostram o potencial da ferramenta ExATOlp para trabalhos em LC e para o levantamento de dados lexicais para estudos terminológicos, para a mineração de dados e para a geração de ontologias em língua portuguesa.

ABSTRACT: This paper presents a study based on Natural Language Processing techniques (PLN) and Corpus Linguistics (CL) approaches to extract terms related to important saussurean concepts in the Brazilian Portuguese edition of the *Course in General Linguistics*. Taking the CGL as a corpus, we aim at an automatic representation method of content through computer tools. Once submitted to the parser PALAVRAS, a morphossyntatic tagger, the corpus is processed by ExATOlp, a tool implementing various linguistic and statistically based NLP techniques. The tool generates hierarchical lists and charts of noun phrases, which are organized according to their specificity / relevance in the target corpus. These lists are then compared to data generated by AntConc - a free access tool quite used in LC approaches - applied to the same corpus. The results show the potential of ExATOlp in works on LC and in colleting lexical data for terminology studies, data mining and generation of ontologies in Portuguese.

[1] Este trabalho é uma extensão de outros relatos de resultados relacionados ao mesmo projeto de pesquisa sobre representação automática de conteúdo de textos científicos, tal como vemos em Ciulla, Finatto (2013).

[*] Docente do PPG-Letras-UFRGS e pesquisadora CNPq.
[**] Professora colaboradora da FACIN-PUCRS e pós-doutoranda DOCFIX-FAPERGS/CAPES.
[***] Professora visitante do Departamento de Linguística, Filologia e Teoria Literária e do PPG- Letras-UFRGS e pós-doutoranda DOCFIX-FAPERGS/CAPES.

PALAVRAS-CHAVE: Extração automática de termos. Curso de Linguística Geral. ExATOlp.

KEYWORDS: Automatic extraction of terms. Course in General Linguistics. Saussure.

1. Introdução

1.1 A extração automática de termos em *corpora*

A extração automática de termos, conforme Di Felippo (2013, p. 66) consiste na identificação e na recolha, a partir de um *corpus* especializado, de expressões linguísticas que tenham um potencial terminológico. No Brasil, essa extração tem sido procedida no âmbito do Processamento de Linguagem Natural (PLN), uma especialidade da Ciência da Computação – cm um caráter bastante aplicado, e também no âmbito da Linguística de *Corpus* (LC), que se integra à Linguística Aplicada, com um caráter descritivo e de auxílio a pesquisas de Terminologia e Terminografia.

Em que pesem as enormes vantagens dessa automatização, ainda assim, o verdadeiro estatuto terminológico dessas unidades deverá ser posteriormente confirmado, seja em novos levantamentos ou em novos contrastes em *corpora* ou por especialistas do domínio. Assim, a extração ainda é um trabalho de máquina, que precisaria ser validado, para que haja uma identificação propriamente dita de unidades terminológicas.

A extração automática, cujas melhores margens de acerto encontram-se ainda em torno de 27% (conforme TEIXEIRA, 2011) para *corpora* em português do Brasil (PB), vem se desenvolvendo, especialmente desde a década de 80 para diferentes idiomas. Nela têm sido empregadas abordagens estatísticas e índices como *tf-idf* e *log likelihood*. Vale lembrar que são ainda muito discutidas as medidas ou pontos de corte de frequência ou de distribuição de uma dada expressão ou palavra ao longo de diferentes textos que integram um *corpus* para que uma unidade ou expressão multipalavra possa se enquadrar como um "candidato a termo".

Ferramentas informatizadas para essa extração partem de cálculos robustos do ponto de vista estatístico e matemático, mas, em geral, não fazem uso de informações linguísticas ou de descrições de linguagens especializadas, o que limita a obtenção de melhores resultados, de acordo com Lopes (2012) e Teixeira (2011). Os melhores resultados, em termos de trabalho automático, em diferentes idiomas, são os de abordagens que utilizam regras linguísticas em seu algoritmo. Contudo, a quase totalidade desses trabalhos são destinados ao inglês, alemão ou francês, sendo raro encontrar abordagens voltadas ao português.

Outra dificuldade da extração automática é a especificidade de certas abordagens que são dependentes de domínios bem pontuais, tal como o trabalho de Bui e Slot (2010), que se aplica à área do comportamento biológico de proteínas. Assim, algumas metodologias de extração tornam-se muito específicas de uma língua ou de um domínio e tendem a não poderem ser replicadas para *corpora* de diferentes áreas de conhecimento.

1.2 Foco deste trabalho

Este trabalho relata parte dos resultados da pesquisa *Recuperação da informação em representação do conhecimento em bases de textos científicos de Linguística e de Medicina*, que associa Processamento de Linguagem Natural (PLN), Linguística de *Corpus* (LC) e Estudos do Texto e do Discurso. Neste artigo, relatamos um estudo com o *corpus* da área de Linguística, representado pelo Curso de Linguística Geral (CLG). Na pesquisa ampla são explorados dois tipos de *corpora* de textos científicos em português brasileiro (PB): a) um *corpus* da área de Medicina, sobre Pneumopatias Ocupacionais – com textos de artigos científicos, dissertações, teses e textos de popularização para leigos; e b) um *corpus* da área de Linguística, representado pelo todo do texto em português do CLG. Esses dois *corpora*, bastante heterogêneos, foram escolhidos, deliberadamente, por representarem, respectivamente duas situações distintas:

a) um domínio e gêneros textuais e discursivos recorrentemente tratados em pesquisas[2] de PLN ou de LC e também em estudos bibliométricos, que visam traçar um perfil da produção científica em diferentes tópicos ou em temas das Ciências da Saúde ao longo de um dado período de tempo, tipo de publicação ou de um periódico específico;

b) um domínio (Linguística) e um gênero (o manual acadêmico ou livro-texto universitário) ainda pouco explorados por pesquisas de LC[3] e de PLN especialmente dedicadas ao PB.

Ambos os *corpora* são tratados linguística e computacionalmente com vistas à identificação das melhores metodologias de representação automática do seu conteúdo e à

[2] Como exemplo de um estudo bibliométrico, ver SANTIN *et al* (2015). Para trabalhos de PLN, entre inúmeros, sugere-se ver LOPES *et. al* (2009). Na LC, cabe citar Di-Felippo (2013). Hoje em dia, especialmente na área de mineração de dados e de recuperação de informação (subáreas da Ciência da Computação), há inúmeros trabalhos que se ocupam de organizar, reunir ou modelar o conteúdo disperso de trabalhos publicados em periódicos de ciências da saúde disponíveis na internet.

[3] Um trabalho a ser citado sobre a linguagem e as terminologias da Linguística, que contempla o PB, é Fromm e Yamamoto (2013).

sistematização, com apoio informatizado, de sua informação lexical, terminológica e textual. De acordo com os princípios da LC, tal como apresentada no Brasil por Berber Sardinha (2004), os *corpora* sob estudo devem ser processados e comparados com outros, na sistemática *corpus* de estudo *versus corpus* de referência, na proporção de 1 para 3 ou 5 vezes o número de palavras de cada um. Essa comparação estatística visa destacar as *keywords* ou palavras-chave, específicas de um *corpus* (*corpus* de estudo) em relação a um *corpus* genérico (*corpus* de referência). Essa metodologia de LC está detalhada no artigo didático de Kader e Richter (2013), seguindo indicações de Berber Sardinha (1999).

Pela ótica do PLN[4], para esse mesmo fim, podem ser adotados métodos estatísticos distintos e técnicas variadas. Para identificação de palavras ou expressões típicas de um dado *corpus*, podem ser feitos, por exemplo. apenas contrastes de *corpora* de uma área de conhecimento e gênero textual com outros *corpora* também da mesma área e gênero textual. E ainda sem se levar em conta apenas a relação de tamanho de dois *corpora* para contrastes, conforme observam Vecchia *et al* (2014) e Ferreira (2012). Um exemplo de comparação de *corpora* de domínios[5] pode ser acessado em http://vhflabs.com.br/nontax/uso.php.

Assim, tomando o texto em português do CLG como um *corpus* de estudo*,* este artigo relata a busca de um método de representação automática de conteúdo através de ferramentas e de técnicas de PLN e de LC que são então comparadas em seus rendimentos. Dar-se-á maior destaque para uma ferramenta de PLN, o ExATOlp - a seguir detalhada, visto que é pouco conhecida entre pesquisadores linguistas. O objetivo do experimento foi extrair termos relacionados a conceitos importantes do CLG que pudessem se enquadrar na condição de "candidatos a termos", conforme Teixeira (2011).

2. Metodologias do trabalho com o CLG: PLN e LC

Uma vez submetido ao *parser* PALAVRAS, um etiquetador morfossintático para a língua portuguesa, o *corpus* do CLG - devidamente escaneado, preparado e salvo em formato .txt - que apresenta 7.606 *types* e 73.586 *tokens*, passou pela ferramenta extratora de sintagmas nominais relevantes, denominada ExATOlp (LOPES, 2012). O nome ExATOlp corresponde a

[4] Para detalhes sobre a natureza e escopo do PLN, ver Dias-da-Silva et al. (2007).
[5] *Corpus de domínio* é como, no PLN, são chamados os *corpus* que reúnem textos de uma mesma área de conhecimento e, normalmente, de mesmo gênero textual. Assim, artigos acadêmicos da área de Medicina, por exemplo, pertenceriam a um mesmo *domínio*.

Extrator Automático de Termos para Ontologias em Língua Portuguesa, recurso que implementa diversas técnicas de PLN de base linguística e de base estatística, tendo sido desenvolvido pelo grupo de PLN da PUCRS (http://www.inf.pucrs.br/linatural/) para trabalhos de construção automática de ontologias de domínio com o PB. Essas ontologias servem, entre outras coisas, para a representação automática de conteúdo de textos ou de *corpora.*

Infelizmente, ainda não há acesso livre e *online* a esse sistema, que necessita que o *corpus* a ser examinado seja pré-processado com algum tipo de *parser* – assim como também os *corpora* que se tomem para contraste. Em seguida ao processamento pelo ExATOlp, foram geradas listas e gráficos hierarquizados de elementos lexicais do CLG, incluindo-se sintagmas nominais e verbos recorrentemente associados a esses sintagmas.

Em uma segunda etapa, o mesmo *corpus* foi submetido à ferramenta AntConc (ANTONY, 2013), uma ferramenta de acesso livre bastante empregada em trabalhos de LC e pouco utilizada em PLN. Nessa ferramenta, utilizamos as funcionalidades *Wordlist, Clusters* e *Keywords* para obtenção de palavras ou expressões mais típicas do CLG, tendo sido tomado como *corpus* de referência o Lácio-Ref (disponível em http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm). Foram também adotados os procedimentos básicos descritos em trabalhos de LC, como, por exemplo, os procedimentos apontados no trabalho de Teixeira (2011), no qual se fez uma análise do desempenho de extratores automáticos de terminologias em *corpora*, embora não se tenha tratado do AntConc ou do ExATOlp.

3. Estudo comparativo

Após serem convertidos do formato .pdf para .txt, os arquivos de texto do CLG foram corrigidos manualmente, para que as unidades lexicais pudessem ser corretamente processadas pelos *softwares*, tanto o AntConc, como o ExATOlp.

Utilizando o AntConc, foram extraídas listas das palavras-chave. O *corpus* utilizado para contraste foi o Lácio-Ref, um *corpus* aberto e de referência do Projeto Lácio-Web, desenvolvido pela FFLCH da USP e pelo NILC-USP, composto de textos de vários gêneros em português brasileiro, tendo como característica serem escritos respeitando a norma culta.

No experimento com a ferramenta ExATOlp, foram extraídas listas dos SNs (sintagmas nominais) mais relevantes do *corpus*, geradas a partir de um processo de extração híbrido, que utiliza heurística de base linguística, para detectar os termos, e cálculos estatísticos, para estimar

a relevância de cada termo para o domínio. Para essa estimativa de relevância, foi utilizado o mesmo *corpus* de contraste utilizado no experimento com a ferramenta AntConc.

Além disso, um aspecto diferenciado da identificação de SNs relevantes feita pelo ExATOlp está na sua representação visual, que conta com diagramas, nuvens e árvores hiperbólicas dinâmicas. É importante observar que, se a partir do AntConc, a proposta é a de extrair palavras-chave de um domínio, com o ExATOlp, a funcionalidade é um pouco diferente, pois foi projetado para extrair SNs que sejam relacionados a conceitos específicos ao domínio do *corpus* de estudo. Acreditamos, no entanto, que se possa estabelecer uma relação estreita entre esses conceitos específicos representados por SNs e as palavras-chave de um *corpus*. Nos dois casos, trata-se de elencar termos ou expressões que designem temáticas de importância maior para um determinado *corpus*.

4. Resultados

4.1. SN candidatos a termos

A Tabela 1 apresenta os 20 unigramas[6] mais relevantes obtidos através de cada um dos processos. Numa primeira análise das listas, percebe-se que o processo estatístico, representado na tabela pela ferramenta AntConc, apresenta palavras gramaticais e outras que não poderiam ser candidatas a termos, já que não designam conceitos. Essas palavras estão grifadas. Já o processo de metodologia híbrida, representado na tabela pela ferramenta ExATOlp, gera uma lista em que todas as primeiras 20 palavras são termos específicos da área de Linguística. Esse aspecto já aponta para uma confirmação do melhor desempenho, por parte do processo híbrido, de identificar automaticamente a relevância dos termos.

[6] *n- gramas* é a expressão usada para referir itens que são coletadas de textos de um *corpus*. Um n-grama de 1 item é chamado de unigrama, de 2 itens, bigrama e assim por diante.

Tabela 1. Listas dos 20 primeiros unigramas extraídos a partir de processo híbrido (ExATOlp) e a partir de processo estatístico (AntConc) com a funcionalidade *keywords*.

	ExATOlp	**AntConc**
1	Signos	**não**
2	Linguística	**é**
3	Latim	**se**
4	Acento	língua
5	Plural	**são**
6	Linguistas	**que**
7	Francês	**à**
8	Significação	linguística
9	Sintagma	palavra
10	Sânscrito	**uma**
11	Grego	**etc**
12	Fonologia	sons
13	Fonema	francês
14	Desinência	**só**
15	Gramáticos	**um**
16	Surdos	latim
17	Sílaba	**também**
18	Adjetivo	**assim**
19	Genitivo	**por**
20	Ortografia	signo

Usando a funcionalidade *stoplist*, da ferramenta AntConc, podemos eliminar palavras que sabemos, de antemão, não serem relevantes, conforme o objetivo do estudo. Assim, aplicamos essa funcionalidade para evitar palavras, como artigos, preposições, numerais, pronomes, conjunções e alguns verbos, conforme a *stoplist* para português do Brasil disponível em http://www.ranks.nl/stopwords/brazilian. Na Tabela 2, então, mostramos a nova lista de palavras do AntConc, agora gerada com uma *stoplist*, novamente em cotejo com os resultados do ExATOlp.

Tabela 2. Listas dos 20 primeiros unigramas extraídos a partir de processo híbrido (ExATOlp) e a partir de processo puramente estatístico (AntConc) com as funcionalidades *keywords* e *stoplist*.

	ExATOlp	**AntConc**
1	signos	língua
2	linguística	linguística
3	latim	palavra
4	acento	sons
5	plural	francês
6	linguistas	latim
7	francês	signo
8	significação	elementos
9	sintagma	fatos
10	sânscrito	unidade
11	grego	escrita
12	fonologia	linguagem
13	fonema	dizer
14	desinência	princípio
15	gramáticos	ponto
16	surdos	analogia
17	sílaba	ideia
18	adjetivo	fenômeno
19	genitivo	relação
20	ortografia	formas

Observamos na Tabela 2 que, com a *stoplist*, o AntConc apresenta uma lista de palavras relevantes para o CLG, sem dúvida. Contudo, nem todas específicas da obra ou da área da Linguística, como é o que observamos na lista gerada pelo ExATOlp. Além disso, a lista gerada pelo ExATOlp possui uma ordenação que reflete a relevância dos termos, ou seja, um cálculo que leva em conta tanto a especificidade, quanto a frequência dos termos. Assim, podemos dizer que a precisão do ExATOlp é maior. Além disso, o resultado gerado pelo ExATOlp é totalmente automático, enquanto que, para o AntConc, foi preciso um trabalho parcialmente manual de confeccionar a *stoplist*.

Um segundo aspecto sobre as listas é o fato de que, ainda que todas as palavras da ferramenta híbrida possam ser consideradas pertencentes à Linguística, elas não são, necessariamente, termos relacionados a conceitos importantes do CLG, isso sob o olhar de um especialista em Saussure. É necessário ressaltar que, como sabemos, as palavras têm seu sentido atribuído somente quando em uso, nas relações de oposição e combinação com as outras palavras do texto. Assim, palavras como "arbitrariedade" ou "diacronia" - fundamentais no

CLG - só assumem o seu valor como termos importantes nessa obra quando em relações de combinação.

Consequentemente, para encontrar termos específicos desse domínio, é necessário extrair também os termos compostos. Na Tabela 3 apresentamos os resultados dos principais 20 bi- e trigramas mais relevantes extraídos pelo processo híbrido (ExATOlp), o que, por si só, já é uma vantagem. Observamos que as listas em n-gramas do AntConc dizem respeito à frequência simples de *clusters* e, por isso, não serão colocados aqui em cotejo.

Tabela 3. Listas dos 20 primeiros bigramas e trigramas extraídos a partir do ExATOlp.

	ExATOlp	
	Bigramas	**Trigramas**
1	imagem acústica	estado de língua
2	mudanças Fonéticas	arbitrariedade do signo
3	signos linguísticos	evolução de sons
4	cadeia falada	fato de gramática
5	fenômenos fonéticos	sistema de valores
6	fatos diacrônicos	mecanismo de língua
7	indo europeu	noção de valor
8	palavra francesa	contraparte de imagem
9	impressão acústica	estudo da linguagem
10	som laríngeo	fatos de língua
11	igual modo	interior de língua
12	linguística diacrônica	linha de conta
13	signo gráfico	objeto da Linguística
14	sistema linguístico	partida de xadrez
15	alto alemão	passagem de ar
16	imagem auditiva	sequência de sons
17	unidade linguística	valor do termo
18	alfabeto grego	vida da língua
19	aparelho vocal	arbitrário do signo
20	elo implosivo	ciência da língua

Na listagem de bigramas do ExATOlp, já aparecem mais termos que podem ser reconhecidos como específicos da Linguística saussuriana, como "imagem acústica", "signos linguísticos", "linguística diacrônica" e "sistema linguístico". E, com exceção de "igual modo", todos os bigramas são termos específicos do domínio da Linguística. Nos trigramas, são colocados em relevo outros tantos termos importantes e específicos da teoria de Saussure, como "estado de língua", "arbitrariedade do signo" e "objeto da Linguística". Aparentemente, bi- e trigramas são menos propensos à ambiguidade do que unigramas. A combinação de termos simples em compostos concede especificidade ao conceito, caracterizando o SN como termo.

Além disso, há uma tendência, e essa é uma hipótese a ser confirmada em nossos *corpora*, em trabalhos futuros, de que termos compostos sejam mais frequentes em textos especializados.

A extração automática dos candidatos a termos do CLG proporcionou o delineamento de uma ontologia - conjunto de termos a partir dos quais são especificados, formalmente, conceitos importantes contidos nessa obra. A metodologia híbrida de extração mostrou-se como a mais produtiva, como se pode observar pelas listas, tanto nos unigramas, como nos bigramas e trigramas. Esses termos podem ser visualizados, a partir do ExATOlp, também em nuvens, como mostram as Figuras 1 e 2.

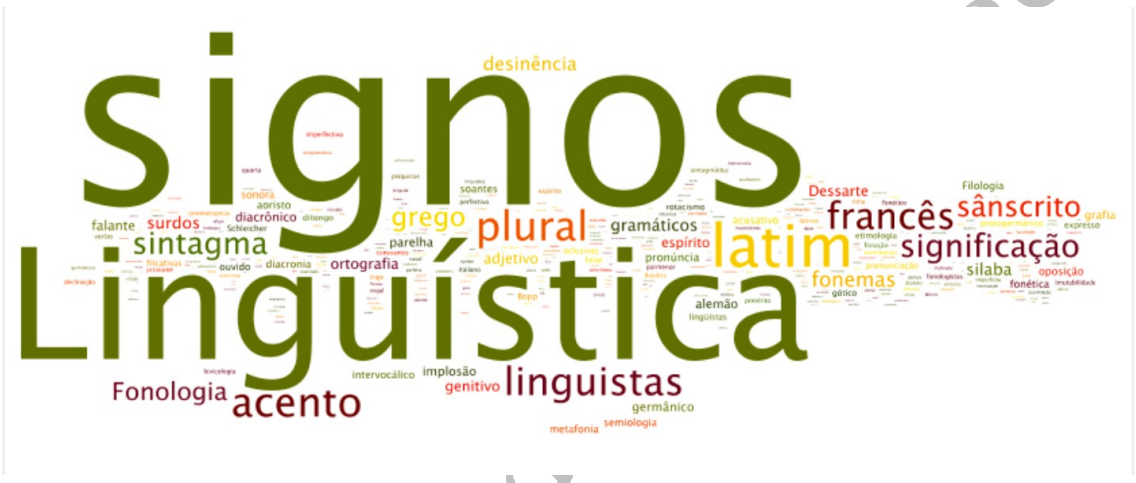


Figura 1. Nuvem de unigramas do CLG extraídos a partir do ExATOlp.

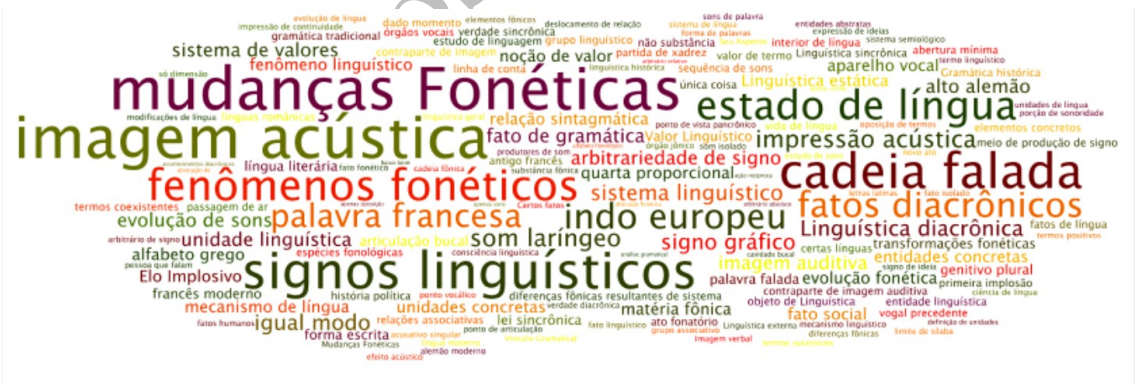


Figura 2. Nuvem de bigramas do CLG extraídos a partir do ExATOlp.

Este estudo com o CLG revela que a aplicação de métodos automáticos para a extração de informação, especialmente os híbridos, com regras linguísticas e estatísticas mais sofisticadas, como o ExATOlp, pode ser muito útil para o trabalho do terminólogo, identificando de maneira rápida itens lexicais que podem ser considerados efetivamente como

termos. Essa mesma tarefa, considerando-se o trabalho humano, demandaria muito mais tempo, especialmente numa obra complexa como o CLG.

Um aspecto que salta aos olhos, em especial nas nuvens, é a grande importância dos termos relacionados à Fonologia sugerida pelo resultado do ExATOlp. Não é surpresa que o tema da Fonologia seja parte da obra, mas mais frequentemente, na literatura, são mencionadas outras questões, como, por exemplo, os estudos sincrônicos e as relações sintagmáticas.

4.2.Explorando outras possibilidades do ExATOlp

Posto que a extração de bigramas e trigramas indicou uma qualidade superior do ExATOlp, fizemos a extração de quadrigramas e pentagramas do *corpus* CLG. Na Tabela 4 estão os 20 quadrigramas e 20 pentagramas mais relevantes segundo a extração do ExATOlp. Cabe salientar que a metodologia de extração para estes termos é a mesma empregada anteriormente, ou seja, identificam-se sintagmas nominais, aplicam-se heurísticas linguísticas e calcula-se o índice de relevância.

Tabela 4. Lista dos 20 quadrigramas e 20 pentagramas extraídos do *corpus* CLG com a ferramenta ExATOlp

	ExATOlp	
	quadrigramas	**pentagramas**
1	contraparte de imagem auditiva	meio de produção de signo
2	interior de mesma língua	aspectos diversos de mesmo fato
3	alteração de signos linguísticos	contato com capa de água
4	aspecto paradoxal de questão	gramática tradicional de francês moderno
5	cadeia falada em sílabas	imobilidade de latim em época
6	correspondência exata de valores	leis estranhas a sua função
7	definição de unidade linguística	papel de esses mesmos órgãos
8	domínio fechado existente próprio	radical coincidente com oposição gramatical
9	gramática tradicional de francês	sistema de fonemas de russo
10	idêntico estado de coisas	vínculo entre ideia e signo
11	imagem acústica com conceito	acumulação de consoantes em alemão
12	interior de cada signo	análise de partes de sintagma
13	jogo de diferenças fônicas	atenção concedida a língua literária
14	número determinado de letras	ausência de todo suporte material
15	parte conceitual de valor	bosquejo de sistema de Fonologia
16	ponto de vista pancrônico	caráter fônico de signo linguístico
17	próprio de instituição linguística	compostos mais próximos de palavras
18	questão de aparelho vocal	condições de vida de línguas
19	sistema de elementos sonoros	condição essencial de signo linguístico
20	série de palavras análogas	condição mecânica com efeito acústico

Ao que parece, para o CLG, os quadri- e pentagramas expressam a importância de alguns contextos de termos importantes, mas não exatamente termos relacionados a conceitos importantes.

Contudo, ressaltamos que uma facilidade adicional do ExATOlp é justamente identificar, junto com os termos extraídos, os seus contextos de utilização. Dentre outras informações, o ExATOlp disponibiliza para cada termo extraído o predicado para o qual o termo exerce a função de sujeito ou de objeto. Desta forma, é possível extrair as ocorrências de verbos mais relevantes de um *corpus* e, a exemplo do que é feito para os termos, é possível utilizar *corpora* contrastantes que permitam associar um índice de relevância de cada predicado para o *corpus* de estudo. Aplicando esta funcionalidade da ferramenta ExATOlp ao *corpus* CLG, identificamos os 30 predicados mais relevantes, conforme apresenta a Tabela 5. Note-se que, devido ao uso da técnica de *corpora* contrastante, foi possível filtrar predicados muito frequentes, mas que não são particularmente significativos para o *corpus* CLG, como é o caso dos verbos "ser", "ter", "poder" e "estar".

Tabela 5. Os 30 predicados mais relevantes do *corpus* CLG segundo extração feita pelo ExATOlp

	predicados relevantes
1	exprimir
2	intervir
3	pronunciar
4	apagar
5	evocar
6	assinalar
7	suscitar
8	confundir
9	supor
10	transtornar
11	designar
12	unir
13	afrouxar
14	ater
15	negligenciar
16	poder empregar
17	poder ser chamar
18	recobrir
19	atribuir
20	acarretar
21	repousar
22	dever trazer

23	parecer favorecer
24	suceder
25	equiparar
26	reger
27	continuar ser
28	suprimir
29	distinguir
30	precipitar

No entanto, como a extração é feita por predicados, há predicados que contém estes verbos com a função de auxiliar, como é o caso do décimo-sétimo predicado mais relevante "poder ser chamar", onde os verbos "poder" e "ser" são utilizados como auxiliares, por exemplo na expressão "pode ser chamado". De toda maneira, os predicados apresentados na Tabela 5 deixam clara a importância de verbos como "exprimir", "intervir" e "pronunciar", que são ao mesmo tempo abundantes no *corpus* CLG e consideravelmente menos frequentes nos *corpora* de contraste empregados. A Figura 3 ilustra a relevância dos verbos extraídos através de uma nuvem de predicados, em que o tamanho de cada predicado é proporcional à sua relevância.

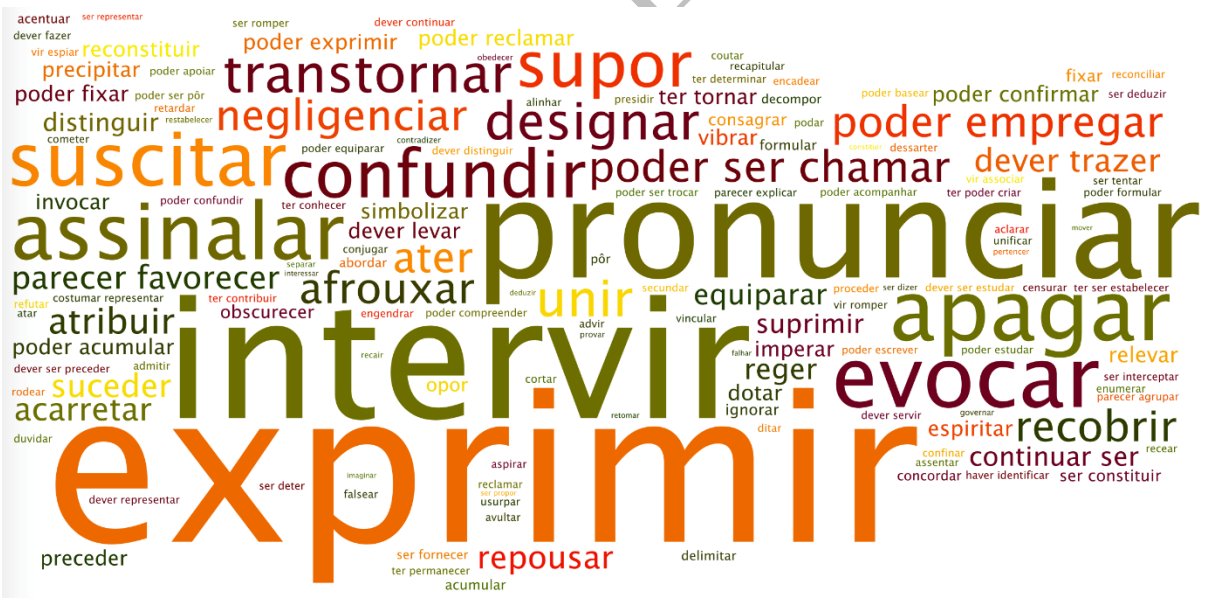


Figura 3. Predicados mais relevantes do *corpus* CLG segundo extração feita pelo ExATOlp

Os verbos, além do papel essencial na gramática das línguas, oferecem, via predicação, elementos para a representação de diferentes relações e podem corresponder a termos ou conceitos de um domínio, como podemos especular, aqui, ao visualizar as nuvens. Assim, o

estudo e a descrição dos verbos podem ser importantes também para subsidiar uma série de recursos de representação e recuperação de informação.

5. Considerações finais e trabalhos futuros

Na comparação do desempenho das duas ferramentas de extração automática de candidatos a termos, quais sejam, o AntConc e o ExATOlp, concluímos que ambas auxiliam no trabalho de levantamento de dados lexicais para estudos terminológicos. Contudo, o ExATOlp traz a vantagem de realizar todo o trabalho automaticamente, além de, pelo menos para o CLG, apresentar candidatos a termos que são mais específicos da teoria saussuriana. Além disso, o ExATOlp lista também automaticamente os bi- e trigramas mais relevantes do *corpus* de estudo, enquanto que o AntConc, mesmo com o auxílio de *stoplists*, lista os bi- e trigramas apenas dos *clusters* mais frequentes. E é justamente nos bi- e trigramas que encontramos os termos mais específicos dos temas tratados pelo CLG.

Observamos também a grande relevância, que pode ser facilmente visualizada nas nuvens geradas pelo ExATOlp, de termos relativos à Fonologia no CLG. Tal resultado sugere uma maior importância do tema em Saussure, que, no entanto, muitas vezes fica em segundo plano, cedendo lugar à discussão de outras ideias saussurianas, como as dicotomias e a arbitrariedade do signo. Assim, outra sugestão de estudos futuros é a de investigar mais a fundo os conceitos de Saussure sobre Fonologia, a partir dos principais candidatos a termos elencados pelo ExATOlp.

Outra pesquisa que se apresenta como sugestão, a partir dos resultados da extração automática de termos, é sobre as recategorizações, ou seja, com que termos e tipos de construções os conceitos são designados e retomados anaforicamente e quais as consequências dessas escolhas. Ainda que se trate de um trabalho mais voltado para os estudos linguísticos do texto, ele pode ser útil no sentido de aperfeiçoar sistemas de extração automática e a análise de seus resultados.

Referências bibliográficas

ANTHONY, L. (2013) **AntConc** (Version 3.4.1) [Computer Software]. Tokyo, Japan: Waseda University. Disponível em : http://www.antlab.sci.waseda.ac.jp/

BERBER SARDINHA, T. UsingKeyWords in text analysis: Practical aspects. **DIRECT Papers**, 42. LAEL, Catholic University of Sao Paulo, Brazil / AELSU, University of Liverpool, England, 1999. Disponível em : www.direct.f2s.com

_____.**Linguística de *Corpus***. Barueri: Manole, 2004.

BUI, Q-C.; SLOOT, P.M.A. Extracting biological events from text using simple syntactic patterns. In: BioNLP Shared Task 2011 Workshop. **Proceedings of** BioNLP Shared Task, Association for Computational Linguistics, 2011, pp. 143–146.

CIULLA, A.; FINATO, M. J. B. O CLG e sua tradução para o português brasileiro - algumas questões sobre a reconstrução da noção de signo linguístico. **Revista Traduzires**. Brasília: Editora da UnB, número especial em homenagem ao centenário de Saussure, 2013, v.2, n.1.

DIAS-DA-SILVA, B.C.; MONTILHA, G.; RINO, L.H.M.; SPECIA, L.; NUNES, M.G.V.; OLIVEIRA JR., O.N.; MARTINS, R.T.; PARDO, T.A.S. Introdução ao Processamento das Línguas Naturais e Algumas Aplicações. Série de **Relatórios do NILC**, NILC-TR-07-10. São Carlos, SP, agosto, 2007, 121p.

DI FELIPPO, A. Extração automática de termos a partir de *corpus* e sua validação para a construção de *Wordnets* terminológicas em português do Brasil. In: Stella Tagnin; Cleci Bevilacqua. (Org.). ***Corpora* na Terminologia**. 1ed. São Paulo: HUB Editorial, 2013, v. 1, p. 63-86.

FERREIRA, V. H. **Uma proposta para descoberta automática de relações não-taxonômicas a partir de *corpus* em língua portuguesa**. 86 f. Dissertação (Mestrado em Ciência da Computação) – Faculdade de Informática, PUCRS. Porto Alegre, RS, 2012.

FROMM, G.; YAMAMOTO, M.I. Terminologia, Terminografia, Tradução e Linguística de *Corpus*: a criação de um vocabulário bilíngue sobre Linguística. In: Stella Tagnin; Cleci Bevilacqua. (Org.). ***Corpora* na Terminologia**. 1ed. São Paulo: HUB Editorial, 2013, v. 1, p. 129-152.

KADER, C.C.; RICHTER, M.G. Linguística de *corpus*: possibilidades e avanços**.** **Instrumento**: Revista de Estudo e. Pesquisa em. Educação. Juiz de Fora: Universidade de Juiz de Fora, v. 15, n. 1, jan./jun. 2013. p.13-23. Disponível em: http://instrumento.ufjf.emnuvens.com.br/revistainstrumento/article/download/2641/1903

LOPES, L. *et al.* Extração automática de termos compostos para construção de ontologias: um experimento na área da saúde. **Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, [S.l.], v. 3, n. 1, mar. 2009. ISSN 1981-6278. crossref http://dx.doi.org/10.3395/reciis.v3i1.244pt

LOPES, L. **Extração automática de conceitos a partir de textos em língua portuguesa**. 2012. 113f. Tese (Doutorado em Ciência da Computação) - Faculdade de Informática, PUCRS, Porto Alegre, RS, 2012.

SAUSSURE, F. **Curso de Linguística Geral**. São Paulo: Cultrix, 1975.

SANTIN, D. M.; NUNEZ, ZIZIL, A. G.; MOURA, A. M. M. de. Produção científica brasileira em células-tronco nos anos 2000 a 2013: características e colaboração internacional. **Revista**

Eletrônica de Comunicação, Informação & Inovação em Saúde, [S.l.], v. 9, n. 2, jun. 2015. ISSN 1981-6278. Disponível em: http://www.reciis.icict.fiocruz.br/index.php/reciis/article/view/965.

TEIXEIRA, R. de B. S. e. Análise do desempenho de extratores automáticos de candidatos a termos: proposta metodológica para tratamento de filtragem dos dados. **Tradterm**, [S.l.], v. 18, p. 297-319, dez. 2011. ISSN 2317-9511. Disponível em: http://www.revistas.usp.br/tradterm/article/view/36765.

VECCHIA, A. D. ; WILKENS, R. ; BOITO, M. Z. ; PADRO, M. ; VILLAVICENCIO, A. Size does not matter. Frequency does. A study of features for measuring lexical complexity. In: 14th Ibero-American Conference on Artificial Intelligence, 2014, Santiago. **Proceedings of the 14th Ibero-American Conference on Artificial Intelligence**, 2014. v. 1.

Artigo recebido em: 30.03.2015
Artigo aprovado em: 22.06.2015

Diretrizes para a criação de um recurso lexical multilíngue a partir da semântica de *frames*: a experiência turística em foco

Guidelines for the creation of a multilingual lexical resource based on Frame Semantics: the tourist experience in focus

Maucha Andrade Gamonal[*]
Tiago Timponi Torrent[**]

RESUMO: Este artigo apresenta as diretrizes utilizadas para o desenvolvimento do Dicionário FrameNet Brasil da Copa do Mundo, dicionário eletrônico trilíngue (Português – Inglês – Espanhol) para os domínios da Copa, do Futebol e do Turismo. Caracterizada como uma teoria linguística que enfatiza a estreita relação entre sistema linguístico e experiência humana, possibilitada através de *corpus*, a Semântica de *Frames* desenvolve molduras que são evocadas pelas palavras na construção de seus significados. A FrameNet, rede semântica em constante desenvolvimento para a língua inglesa no International Computer Science Institute, em Berkeley, e em processo de extensão para outras línguas em diversos países, como o Brasil, fornece a metodologia necessária para a criação deste recurso. O produto desenvolvido a partir deste aporte teórico-metodológico, disponível online gratuitamente, é aqui apresentado através do domínio turístico.

PALAVRAS-CHAVE: FrameNet Brasil. Semântica de *Frames*. Lexicografia Computacional. Dicionários Eletrônicos. Experiência turística.

ABSTRACT: This paper presents the guidelines used for the development of the FrameNet Brazil World Cup Dictionary, a trilingual electronic dictionary (Portuguese – English – Spanish) for the domains of the World Cup, Football and Tourism. Emphasizing the close relation, attested by corpora, between language systems and human experience, Frame Semantics develops frames that are evoked by words for constructing their meanings. FrameNet, a semantic network being developed for English at the International Computer Science Institute in Berkeley, and in process of extension to other languages in several countries, such as Brazil, provides the methodology necessary for the creation of this resource. The product developed from this theoretical and methodological basis, freely available online, is presented here from the perspective of the tourist domain.

KEYWORDS: FrameNet Brazil. Frame Semantics. Computacional Lexicography. Eletronic Dictionaries. Touristic Experience.

1. Introdução

Uma imensa quantidade de pesquisas é financiada mundo afora com o objetivo de desenvolver mecanismos computacionais inteligentes capazes de manipular a linguagem

[*] Doutoranda em Linguística pela Universidade Federal de Juiz de Fora. Bolsista de Doutorado Sanduíche do Programa Ciência sem Fronteiras junto ao International Computer Science Institute e à University of California Berkeley.
[**] Docente do Programa de Pós-Graduação em Linguística da Universidade Federal de Juiz de Fora.

humana. Por mais que a diferença linguística não deva ser fator que limite a comunicação entre pessoas de todo o mundo, este ainda é o panorama. Tendo isso em vista, o linguista que orienta seus estudos através de *corpus* e apresenta interesse em contribuir para os estudos voltados ao Processamento de Linguagem Natural assume papel importante neste cenário.

A rede de *frames* FrameNet[1] (BAKER, 2008; RUPPENHOFER *et al*, 2010) é um projeto que utiliza *corpus* como fonte de respostas empíricas para as propriedades lexicais da língua inglesa. Em desenvolvimento desde 1997 no International Computer Science Institute, na cidade de Berkeley, Califórnia, a FrameNet explora o conceito de *frames* na construção de um recurso lexical para a língua inglesa através da Semântica de *Frames* de Charles Fillmore (1982). Para Fillmore (1985), os sentidos são possibilitados através de palavras inseridas em contextos, os *frames*, ou seja, molduras cognitivas que nos permitem fazer as devidas correlações de sentidos.

Com intuito de estender tal recurso para o português brasileiro, a professora Margarida Salomão investiu na implementação da FrameNet Brasil – http://www.ufjf.br/framenetbr/ – (SALOMÃO, 2009), que vem sendo desenvolvida desde então na Universidade Federal de Juiz de Fora. O retorno instigante possibilitado pelos *insights* da Semântica de *Frames* e a tentativa de criar um recurso lexical nos termos de tal teoria levou a FrameNet Brasil a investir na criação de um recurso lexical trilíngue – Português, Inglês, Espanhol – para os domínios do Turismo, do Futebol e da Copa do Mundo, o Dicionário FrameNet Brasil da Copa do Mundo, disponível para consulta através do endereço http://dicionariodacopa.com.br/.

O presente trabalho tem como intuito apresentar, através do domínio do turismo, as principais decisões teórico-metodológicas adotadas para este dicionário. Assim, as duas principais questões que orientam a estruturação deste texto são: como a Semântica de *Frames* e a FrameNet[2] podem atuar no desenvolvimento de dicionários eletrônicos para usuários não especializados e qual é a vantagem de se utilizar *corpus* para viabilizar tal empreendimento.

2. A Semântica de *Frames* no desenvolvimento de dicionários eletrônicos

Inserida nos estudos da Linguística Cognitiva, a Semântica de *Frames* surgiu através de Charles J. Fillmore a partir do desenvolvimento da Gramática de Casos (FILLMORE, 1968a,

[1] https://framenet.icsi.berkeley.edu
[2] Quando o intuito for fazer referência ao projeto mãe desenvolvido em Berkeley, apenas o nome Framenet com a inicial maiúscula será utilizado. Ao se referir a suas extensões, será acrescentado o país de origem.

1968b), hipótese para representação semântica tendo por base as relações existentes entre predicador e seus complementos, o que ele chamou de "casos". Tanto casos sintagmáticos como semânticos foram abordados por esta investigação e a regularidade em tais combinações foi o que Fillmore considerou como os *case frames*. O desenvolver da pesquisa mostrou que os casos atribuídos não eram suficientes para dar conta de diferenças semânticas importantes, o que o fez optar por funções microtemáticas (FILLMORE, 2003), desenvolvendo, assim, uma Semântica de *Frames*, que, nas palavras de seu próprio criador pode ser assim definida:

> um programa de pesquisa em linguística empírica e uma metodologia descritiva para apresentar os resultados de tal pesquisa (...) pelo termo *frame*, eu tenho em mente qualquer sistema de conceitos relacionado de tal forma que, para entender qualquer um deles, você tem de entender toda a estrutura na qual ele se encaixa; quando um dos conceitos em dada estrutura é introduzido dentro de um texto ou de uma conversa, todos os outros são automaticamente disponibilizados (...) (FILLMORE, 1982, p.111).[3]

Um clássico exemplo é o *frame* que trata da Transação_comercial (FILLMORE, 1977). Há uma pessoa – o VENDEDOR – interessada em repassar mercadorias em troca de dinheiro para outra – o COMPRADOR – que aceita trocar DINHEIRO por MERCADORIAS[4]. Destacando tal experiência em termos verbais, podemos dizer que "comprar", "pagar" e "vender", sem dúvida, fazem referência a tal evento, mas perfilam perspectivas distintas, uma será a do comprador sobre a mercadoria, outra do comprador sobre o dinheiro necessário para a obtenção de mercadorias e a outra será do vendedor sobre a mercadoria. Vejam-se, na Tabela 1, exemplos do *corpus* Copa 2014 FRAMENET BRASIL, constituído para o desenvolvimento do dicionário.

[3] Texto original: "(...) a research program in empirical semantics and a descriptive framework for presenting the results of such research (...) By the term 'frame' I have in mind any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits; when one of the things in such a structure is introduced into a text, or into a conversation, all of the others are automatically made available".

[4] Ao longo do texto, os nomes dos *frames* aparecem com a fonte `Courier` enquanto os elementos que os compõem, em letra maiúscula. Já os predicadores considerados potenciais evocadores de *frames* são destacados em negrito e letra maiúscula.

Tabela 1. Exemplos de sentenças que instanciam os diferentes *frames* da transação comercial.

Frame	**Exemplo**
Comércio_comprar	**COMPRAMOS** uma bacia de camarão por cinco reais.
Comércio_vender	As barracas **VENDEM** comida, bebida e artesanato típico de cada região.
Comércio_pagar	Teríamos pago menos da metade do que nós **PAGAMOS.**

Fillmore quis salientar que a compreensão de uma palavra só é possível caso o *frame* seja mentalmente acessado. Compreender, por exemplo, o sentido de "comprar", "vender" ou "pagar" nos exemplos dados significa conhecer molduras de conhecimento como Comércio_comprar, Comércio_vender e Comércio_pagar. Por tal motivo, o significado linguístico é relativizado a *frames* (FILLMORE, 1977, p.59).

Enquanto isso, em outras áreas, pesquisadores também escolheram o mesmo termo para desenvolverem seus estudos. Na sociologia, Goffman explorava a palavra *frame* para enfatizar a moldura de conhecimento necessária para a compreensão de intenções, perspectivas, rituais e padrões que os indivíduos estabelecem na interação cotidiana. Com a publicação do livro "Frame Analysis: An Essay on the Organization of Experience", em 1974, ele discorreu sobre a organização social das experiências humanas, revelando como as interações sociais são definidas em torno de molduras específicas que orientam as ações dos indivíduos na sociedade.

Dentre suas variadas analogias, ele considerava que a atuação do Homem na sociedade pode ser comparada a uma peça de teatro, em que vários papéis sociais são assumidos a depender da função da interação estabelecida. Bastante ilustrativa, nesse sentido, é a ideia do uso de máscaras no legado teatral grego. Elas eram utilizadas para representar personagens, trocar de máscara era trocar de papel. Na vida cotidiana, num mesmo dia, o comportamento varia diversas vezes, ora somos filhos, ora pais; ora estudantes, ora professores; podemos ser vendedores, mas, certamente, também somos consumidores. Esse fluxo constante nos mostra que ocupamos diversos papéis, a troca de "máscaras" é uma necessidade diária, não uma escolha, pois o "cenário" e "os personagens" variam.

Já Minsky, um dos fundadores do laboratório de Inteligência Artificial do Instituto de Tecnologia de Massachusetts, demonstrou interesse pelos estudos da cognição humana com a proposição de estruturas de dados estereotipadas que representam as situações. Segundo ele, o conhecimento não deve ser visto como uma coleção de fragmentos simples e desconexos, e, sim, como estruturas complexas, denominadas *frames***,** definidas, por ele, como

> uma estrutura de dados para a representação de situações estereotipadas, tais como estar em certo tipo de sala de estar ou ir a um aniversário de criança. Anexos a cada *frame* existem diversos tipos de informações. Algumas dessas informações dizem respeito ao modo de uso do *frame*. Algumas concernem ao que se espera que aconteça em seguida. Algumas tratam do que fazer caso essas expectativas não se confirmem. (MINSKY, 1975, p.1)[5]

Um exemplo utilizado é o da festa de aniversário, com ele, o autor pondera que as definições de dicionário nunca dizem o suficiente. Por mais que qualquer pessoa saiba que esse tipo de evento envolve mais que um encontro para comemorar mais um ano de vida completado por alguém, nenhuma definição breve é capaz de mostrar a complexidade desse evento. No Brasil, como em vários outros países, há o costume de acender as velas postas no bolo durante a música de comemoração. Dessa forma, se, antes deste momento, o anfitrião lamenta que se esqueceu da vela, dificilmente, alguém irá questionar se a luz acabou, o que aponta para o fato de que todos compartilham das mesmas expectativas no que diz respeito à experiência com a festa de aniversário.

O resultado comum pelo termo "*frame"* não foi uma mera coincidência dos diferentes objetos de estudo no final do século XX, mas confirma a necessidade de investigar tais estruturas e a necessidade de investir mais no diálogo científico. O interesse pelo comportamento social e pelas maneiras de conceitualizar a linguagem humana em termos linguísticos e computacionais corrobora a nossa opção pela Semântica de *Frames,* abordagem que inspirou e baseou o desenvolvimento do Dicionário FrameNet Brasil da Copa do Mundo. A pretensão esteve em testá-la como mecanismo único de organização de um dicionário para não especialistas (TORRENT et al, 2014), diferentemente, por exemplo, do recurso lexical *online* do domínio do Futebol, o Kicktionary (SCHMIDT, 2006; 2007; 2008; 2009), disponível *online* em http://www.kicktionary.de, que, apesar de fazer uso também de aspectos da Semântica de *Frames*, confia mais nas WordNets para seu desenvolvimento.

Nesse sentido, passamos, a seguir, à apresentação da metodologia adotada para o desenvolvimento do Dicionário FrameNet Brasil da Copa do Mundo.

[5] Texto original: "A frame is a data-structure for representing a stereotyped situation like being in a certain kind of living room or going to a child's birthday party. Attached to each frame are several kinds of information. Some of this information is about how to use the frame. Some is about what one can expect to happen next. Some is about what to do if these expectations are not confirmed".

3. A metodologia da FrameNet adaptada aos propósitos de um dicionário eletrônico para usuários não-especialistas

Considerada uma extensão teórica da Semântica de *Frames* para o domínio da lexicografia computacional, o objetivo inicial da FrameNet de criar um recurso lexical para descrever a língua inglesa num viés semântico e sintático passou também a ser útil para profissionais ligados à lexicografia, na estruturação de dicionários, e para a Linguística Computacional, em tarefas ligadas ao Processamento de Linguagem Natural (PLN).

Como extensão à ideia original, atualmente, a framenets estão sendo desenvolvidas em outros lugares do mundo, além do Brasil, como Alemanha, Coréia do Sul, Espanha, Japão e Suécia. O intuito é desenvolver bancos de dados em suas próprias línguas e, futuramente, viabilizar empreendimentos multilíngues. No Brasil, a FrameNet vem sendo explorada desde 2007, na Faculdade de Letras da Universidade Federal de Juiz de Fora. O sítio da FrameNet Brasil, disponibiliza, com acesso livre, a rede semântica para o Português do Brasil, tanto de vocabulário específico como de vocabulário genérico.

3.1. A FrameNet

Tendo todas as análises lexicais ancoradas em evidências em *corpora*, a estrutura de uma framenet que siga os princípios do projeto-mãe se alicerça em três grandes tipos de dados: *frames*, unidades lexicais e sentenças anotadas.

Os *frames* são vistos como uma modelagem com interesse computacional de uma estrutura de conhecimento reconhecível em uma dada cultura. São definidos em torno de seus constituintes, os Elementos de *Frame* (EF), que podem ser atores, ferramentas ou circunstâncias, por exemplo.

As unidades lexicais (ULs) são entendidas como o pareamento de uma forma, com todas as suas flexões, a um significado específico, são essas palavras associadas a sentidos específicos que evocam os *frames*. Os verbos assumem destaque especial, pois são predicadores por natureza, mas nomes, adjetivos, advérbios e também preposições são evocadores de *frames*.

As anotações lexicográficas giram em torno de sentenças, elas fornecem evidência empírica para as análises que levaram à constituição dos *frames* e à definição das ULs. São analisadas tanto sintaticamente quanto semanticamente. Assim, dada uma UL, as sentenças em que ela se instancia têm seus constituintes sintagmáticos anotados tanto para os EFs que circundam a UL (anotação semântica), quanto para as funções gramaticais e tipos sintagmáticos

que caracterizam o material linguístico que manifesta os EFs (anotação sintática). As funções gramaticais e os tipos sintagmáticos utilizados para o português do Brasil foram definidos por Torrent & Ellsworth (2013).

Como pode se ver, nesse contexto, a semântica é o centro para a explicação da gramática e as regularidades combinatórias abstraídas das anotações, ou seja, os padrões de valência, assumem destaque primordial. Considerando-se as três estruturas de dados principais apresentadas, framenets se caracterizam como esforços de lexicografia prática que têm como tarefas, segundo Fillmore (2008):

i) descrever Unidades Lexicais a partir dos *frames* evocados, bem como descrever os respectivos *frames*;

ii) descrever os Elementos de *Frame* que compõem cada *frame*;

iii) extrair sentenças de *corpora* para validar as análises das Unidades Lexicais;

iv) selecionar, dentre as sentenças extraídas, aquelas que sejam representativas das diversas possibilidades de valência das Unidades Lexicais;

v) disponibilizar os resultados na forma de entradas lexicais que resumem os padrões de valência sintático-semântica das Unidades Lexicais;

vi) definir uma rede de relações entre *frames* e apresentá-la graficamente.

Na subseção seguinte, as etapas descritas acima serão ilustradas com os dados do Copa 2014 da FrameNet Brasil através de adaptações necessárias no que diz respeito à metodologia.

3.2. Tarefas linguísticas para a constituição de um dicionário baseado em *frames*

A primeira tarefa realizada no projeto do dicionário foi a compilação de *corpora* específicos. Nessa tarefa, foram fundamentais as contribuições de Sardinha (2004), no que tange aos requisitos para a criação e caracterização dos *corpora*, e de Calvi (2010), no que tange aos gêneros textuais a serem incluídos neles.

Começando pelas primeiras, Sardinha (2014) propõe que textos a serem incluídos em *corpora* devem ser autênticos, em linguagem natural e não desenvolvidos com o propósito de servir a uma pesquisa linguística. Numa perspectiva multilíngue, *corpora* autênticos de mais de uma língua podem ser comparáveis ou paralelos. Enquanto estes se caracterizam por conter versões traduzidas de um mesmo texto, aqueles são compostos por textos de um mesmo gênero, porém, sem o compromisso de representação de um mesmo conteúdo. O Dicionário FrameNet Brasil da Copa do Mundo explorou ambos os tipos de *corpora*, uma vez que o tamanho dos

textos paralelos encontrados (c. 375.000 tokens por idioma) não se mostrou suficiente para atestar adequadamente as ULs levantadas.

Além desse aspecto, vários outros critérios presentes na literatura da Linguística de *Corpus* (modo, tempo, seleção, conteúdo, autoria e finalidade) foram especificados no que tange aos *corpora* levantados para o domínio do turismo. Assim, os *corpora* coletados sobre turismo são: escritos (modo); sincrônicos/contemporâneos, por designarem o período corrente, atual (tempo); dinâmicos, pois podem ser aumentados e/ou diminuídos (seleção); de domínios especializados (conteúdo); e, por último, são textos de falantes nativos (autoria).

Já no tangente aos gêneros textuais selecionados para compor os *corpora*, seguiram-se as considerações de Calvi (2010, p.19), para quem os gêneros textuais cujos objetivos comunicativos são os de descrever e promover destinos turísticos são os mais representativos do vocabulário turístico. Assim sendo, os *corpora*, nas três línguas alvo do dicionário, são compostos por guias de turismo (textos paralelos com c. 375.000 tokens por idioma), *sites* governamentais de fomento a atividades turísticas (textos comparáveis com c. 585.000 tokens por idioma) e *blogs* de viagem (textos comparáveis, com c. 40.000 tokens por idioma), totalizando cerca de 1.000.000 de tokens por idioma.

Todos os *corpora* foram pré-processados sintaticamente, utilizando-se os *parsers* PALAVRAS (BICK, 2000), para o português, e TreeTagger (SHMID, 1994), para o inglês e o espanhol. Posteriormente, foram armazenados, compilados e acessados através da ferramenta SketchEngine (http://sketchengine.co.uk). Criada a infraestrutura para o trabalho com os *corpora*, a próxima tarefa centrou-se na criação dos *frames*.

Na etapa de estruturação de *frames*, priorizou-se o método *bottom-up*, que partia do mais básico, as ULs, para aquilo mais abstrato, os *frames*. Assim, o anotador parte dos dados para criar o *frame*. A primeira tarefa é selecionar, intuitivamente, um agrupamento de lexemas que se relacionam semanticamente, por exemplo, *visitar*, *turista*, *apreciar*, *visitante* e *atração*. Estes lexemas, ainda que apresentem particularidades quanto ao sentido e não pertençam todos às mesmas classes de palavras, participam de um mesmo domínio semântico, o das atividades turísticas.

Num segundo momento, parte-se para a pesquisa dos itens lexicais selecionados nos *corpora*, tendo em mente um possível *frame* para investigação. Após esse estudo, selecionam-se algumas sentenças com esses itens lexicais com o objetivo de analisar o comportamento desses predicadores no que tange à valência sintático-semântica. Com isso em mãos, o analista

procura regularidades tanto semânticas quanto sintáticas que permitam a estruturação de uma situação específica, essa etapa é a definição do *frame*. Nela, são determinadas as ULs, previamente selecionadas pelo analista, são especificadas a nuclearidade dos Elementos de *Frame* (EFs) bem como as relações entre os EFs. Quando o *frame* estiver estruturado, é possível relacioná-lo com outros, caso existam.

É importante ressaltar que, por serem funções microtemáticas, EFs – e, por consequência, os *frames* que deles se compõem – podem ser bastante específicos em sua definição. A título de exemplo, observem-se (1) e (2):

(1) [Jô Soares VISITANTE] VISITA [a presidente Dilma Rousseff ENTIDADE], [em Brasília LUGAR].[6]

(2) [Dilma TURISTA] VISITA [praia deserta ATRAÇÃO] [na Ilha dos Frades, na Bahia LOCAL].[7]

Veja-se que, tanto em (1) quanto em (2), poderíamos assumir a mesma estrutura sintática NP V NP. Todavia, os dois enunciados referem-se a experiências distintas. Em um caso, há elementos que sugerem um tipo de visita que não se adequa ao esperado para atividades turísticas. Assim, as duas sentenças evocariam *frames* de *background* distintos na construção do sentido de *visitar* em cada uma: o de Visitar, para (1) e o de Turismo_por_turista, para (2). Porém, como a atividade turística não deixa de compartilhar traços genéricos de uma visita, na base de dados, esses dois *frames* estão relacionados entre si.

A análise de dados acontece pela anotação lexicográfica da FrameNet Brasil, majoritariamente, em três camadas, Elemento de Frame (no inglês FE, Frame Element), Função Gramatical (no inglês GF, Gramatical Function), e Tipo Sintagmático (no inglês PT, Phrase Type). A Figura 1 mostra a anotação de uma sentença que instancia a UL *visitar.v*, evocadora do *frame* Turismo_por_turista.

[6] Manchete acessada em 23 set. 2015. http://www.ofuxico.com.br/noticias-sobre-famosos/jo-soares-visita-a-presidente-dilma-rousseff-em-brasilia/2015/05/18-239130.html
[7] Manchete acessada em 23 set. 2015. http://correiodopovo.com.br/Noticias/545102/Dilma-visita-praia-deserta-na-Ilha-dos-Frades,-na-Bahia

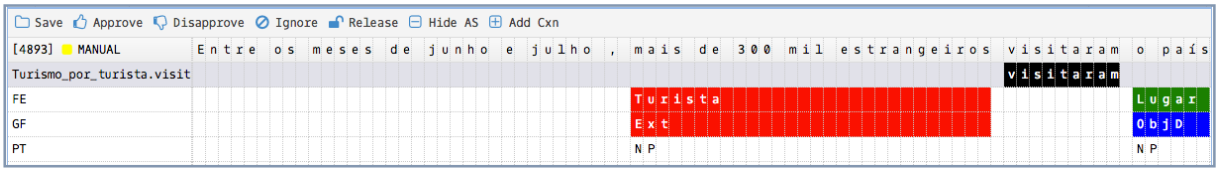


Figura 1. Anotação de uma sentença na FrameNet Brasil

Como as palavras que evocam *frames* são predicadores, como verbos, substantivos e adjetivos, especifica-se, a partir de um conjunto de anotações, a valência desses itens lexicais tanto em relação à sintaxe quanto à semântica.

É válido destacar que, diferentemente da WordNet (FELLBAUM, 1998), recurso lexical que organiza o léxico por uma perspectiva paradigmática, através dos chamados *synsets*, a rede semântica FrameNet e o recurso ora apresentado não sistematizam as informações de maneira semelhante. Ainda que haja o empenho em descrever as informações combinatórias dos itens lexicais, não se pode afirmar que a FrameNet assuma um viés apenas sintagmático. Ao conceber o conhecimento lexical através de molduras intrinsicamente relacionadas às experiências humanas, os *frames* são os responsáveis pelo agrupamento. É possível analisar os dados por um viés paradigmático, por mais que isso não esteja sistematizado no recurso. Por outro lado, as relações combinatórias entre os participantes, os Elementos de *Frame*, desfrutam de grande atenção. Todas as sentenças são analisadas sintaticamente e, depois, são fornecidos os padrões de tais combinações.

As Tabelas 2 e 3 destacam os resultados do chamado processo de anotação, fazendo referência à palavra *apreciar*. Os *integrantes* da cena a que está vinculada (os Elementos de *Frame*) aparecem em destaques coloridos. A Tabela 2 enfatiza os dados em relação aos participantes e as suas realizações sintáticas, já a Tabela 3 mostra os padrões de valência. Ou seja, a partir das realizações sintáticas encontram-se os padrões combinatórios das análises feitas em *corpora*.

Tabela 2. Padrões de realização dos EFs – *apreciar*

Elemento de Frame [Frame Element]	Número Anotado [Number Annotated]	Realizações [Realization(s)]
Atração	(17)	NP.Ext (1) NP.ObjD (16)
Lugar	(11)	NP.Dep (3) 2nd.-- (1) DNI.-- (7)
Maneira	(4)	PP.Dep (3) NP.Dep (1)
Turista	(17)	CNI.-- (6) NP.Ext (10) DNI.-- (1)

Tabela 3. padrões de valência - *apreciar*

Número Anotado [Number Annotated]	Patterns				
3 TOTAL	Atração	Lugar	Lugar	Turista	
(1)	NP Ext	NP Dep	NP Dep	CNI --	
(1)	NP ObjD	NP Dep	NP Dep	CNI --	
(1)	NP ObjD	NP Dep	NP Dep	NP Ext	
1 TOTAL	Atração	Lugar	Maneira	Turista	
(1)	NP ObjD	DNI --	PP Dep	NP Ext	
1 TOTAL	Atração	Lugar	Maneira	Turista	Turista
(1)	NP ObjD	2nd --	PP Dep	NP Ext	NP Ext
4 TOTAL	Atração	Lugar	Turista		
(1)	NP ObjD	DNI --	CNI --		
(1)	NP ObjD	DNI --	DNI --		
(2)	NP ObjD	DNI --	NP Ext		
2 TOTAL	Atração	Lugar	Turista	Turista	
(2)	NP ObjD	DNI --	NP Ext	NP Ext	
2 TOTAL	Atração	Maneira	Turista		
(1)	NP ObjD	NP Dep	CNI --		

No Dicionário FrameNet Brasil da Copa do Mundo, três conjuntos diferentes de etiquetas foram utlizados para a camada GF e outros três para a camada PT, uma vez que nessas camadas registram-se as especificidades morfossintáticas dos idiomas cobertos pelo recurso. Já os EFs foram traduzidos do português para os dois outros idiomas (inglês e espanhol) para fins de interface, sendo os *frames* os mesmos para as três línguas[8].

[8] Vide Gamonal e Torrent (GAMONAL;TORRENT, 2014) para uma discussão aprofundada desta questão, à qual este artigo retornará na próxima seção.

4. O Dicionário FrameNet Brasil da Copa do Mundo: o Turismo em foco

O Dicionário FrameNet Brasil da Copa do Mundo encontra-se disponível gratuitamente, na forma de um *web app*, tendo sido lançado em junho de 2014, semanas antes da Copa do Mundo FIFA Brasil 2014. Por ser voltado a não especialistas, no dicionário, a terminologia adotada pelas framenets sofreu pequenas alterações: *frames* passaram a ser chamados de cenas, EFs de participantes, ULs de palavras e assim por diante.

A Figura 2 apresenta a interface inicial ao usuário, na qual ele deve selecionar o idioma através do qual pretende interagir com o aplicativo. O recurso pode ser explorado através dos seguintes comandos: *buscar por palavra*, *digitar texto*, *ver significado* e *explorar a rede*, mostrados na tela de acesso principal, na Figura 3.

Figura 2. Seleção do idioma de interface

Figura 3. Tela de acesso aos sistemas de busca

Clicando em *Buscar palavra*, o usuário é levado a uma lista de palavras na língua de sua escolha. Ao clicar sobre qualquer uma delas, por exemplo, *apreciar.v*, o aplicativo apresenta, na primeira tela de resultados – Figura (4) –, a cena evocada pela palavra: Fazer Turismo. Em seguida, há uma definição do item lexical (glosa) bem como equivalências para as demais línguas do dicionário. Tais equivalências são calculadas automaticamente pelo próprio aplicativo, com base nos padrões de valência armazenados na base de dados.

Arrastando a tela de resultados para o lado – Figura (5) –, o usuário pode visualizar os participantes envolvidos na cena evocada pela palavra, os quais, através do código de cores, serão marcados na tela que traz as sentenças de exemplo – Figura (6).

A Figura (5) mostra os participantes da cena Fazer Turismo. A compreensão desta experiência está vinculada, necessariamente, à existência do Turista e também da Atração. Outro participante é o Acompanhante, aquele que compartilha a experiência com o Turista, o que não significa que não possa assumir o papel de turista, mas, nas evidências em *corpora*, não é o participante em destaque. Veja que a nomenclatura escolhida tem o intuito de ser transparente ao usuário.

Na Figura (6), as sentenças que exemplificam os itens lexicais advêm de *corpora* coletados durante a elaboração do dicionário. O usuário pode optar por acessar o sítio no qual as sentenças foram encontradas ao clicar nelas.

Figura 4. Verbete: cena, glosa e traduções

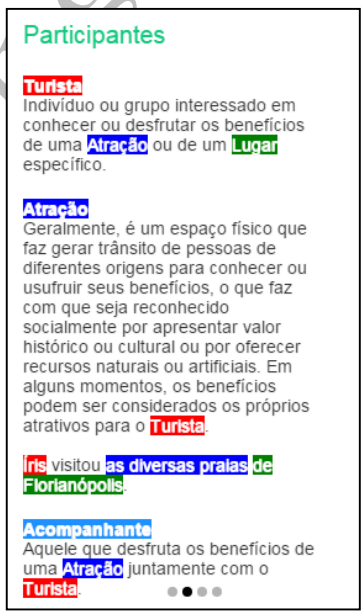


Figura 5. Verbete: participantes da cena

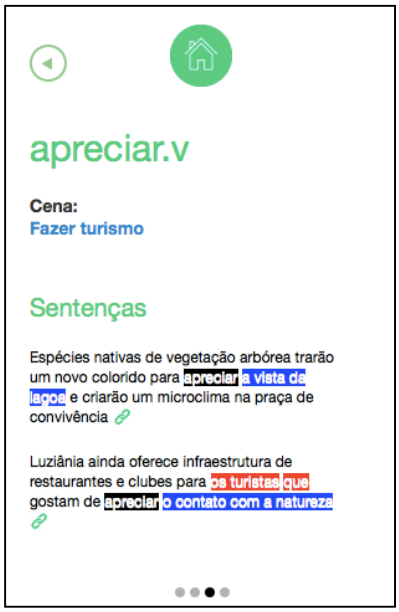


Figura 6. Verbete: sentenças de exemplo

Figura 7. Verbete: mais palavras

Por fim, na última tela de resultados, Figura (7), são mostradas as demais palavras que evocam a mesma cena, tais como *conhecer.v*, *desfrutar.v* e *tour.n*. Após o recurso *Digitar frase*, ferramenta de busca para acessar o significado das palavras no contexto de uso, a próxima opção é *Ver significado*. Nela, o usuário encontra várias cenas vivenciadas na atividade turística. A Figura (8) ilustra essa ferramenta de busca através da cena Fazer Turismo. Além da definição e apresentação dos participantes, há descrição das palavras que evocam tal experiência tanto na língua portuguesa quanto nas outras duas línguas.

Um recurso muito importante no desenvolvimento da rede semântica FrameNet é a designação das relações estabelecidas entre os *frames*, tais relações projetam como as experiências e os eventos se conectam no estabelecimento das relações de sentido. Objetivou-se reproduzir esse propósito no recurso ora apresentado. A Figura (9) apresenta as relações descritas. A Figura (10) destaca a partir do Cenário da Chegada as relações entre as cenas descritas no dicionário.

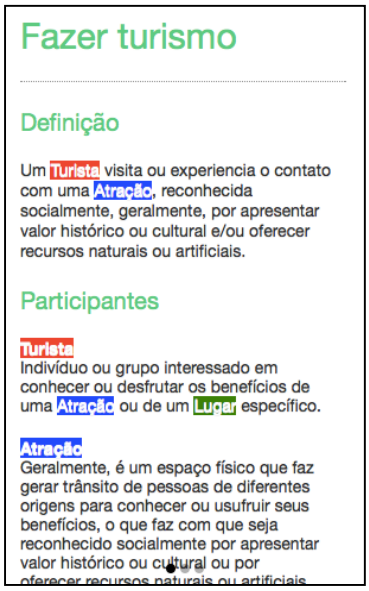


Figura 8. Ver significado - resultado

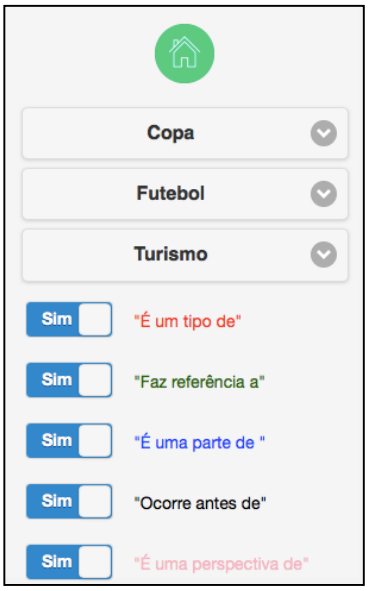


Figura 9. Explorar rede – busca

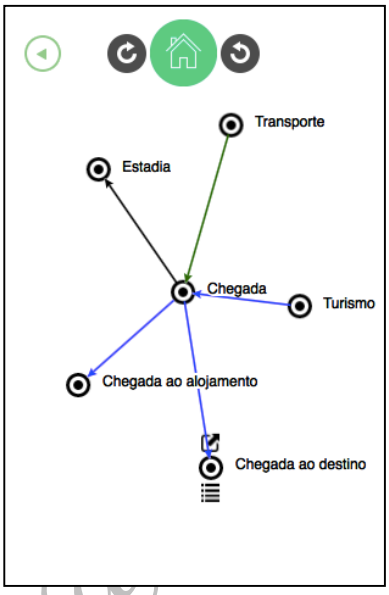


Figura 10. Explorar rede: grafo

Ao primeiro trabalho concluído em torno desse projeto, coube formular as diretrizes para a constituição do dicionário, o que foi feito com a criação de *frames* para o domínio do Turismo, bem como através das análises semânticas e sintáticas dos itens lexicais incluídos no recurso. Várias perguntas surgiram no início do estudo e, logo, se tornaram o objeto de pesquisa, dentre elas: em que medida os *frames* do domínio turístico modelados com *corpora* compilados da língua portuguesa do Brasil servem para representar os *frames* do Turismo para as demais línguas do dicionário?

Responder essa pergunta considerava a reiteração de que *frames* podem atuar como modelagens da conceptualização humana. Através de leituras na literatura da área, mas centrando-se na experiência individual e, ao mesmo tempo, coletiva, manifesta na linguagem, conclui-se que o evento turístico pode ser considerado dotado de uma estrutura transcultural. Qualquer pessoa com o intuito de se juntar à prática turística domina todo o *background* envolvido. Tal fato pôde ser comprovado na medida em que a estrutura de modelagem aplicada ao Cenário do Turismo, mostrado parcialmente na Figura 10, para o português brasileiro serviu igualmente para o tratamento lexicográfico do inglês americano e do espanhol europeu (variantes dos textos que compõem os *corpora* dos demais idiomas do dicionário).

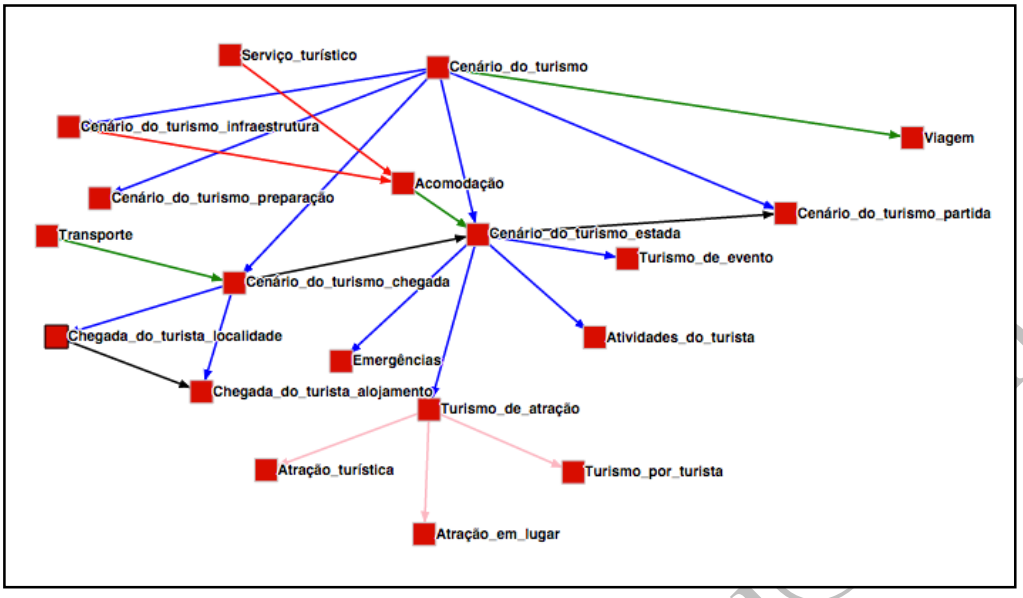


Figura 10. Os *frames* no Cenário do Turismo[9]

A afirmação de que se trata de uma estrutura de *frames* transcultural significa que tal mapeamento é reconhecível por todos, independentemente da diversidade cultural/linguística. Há o conhecimento compartilhado por todos os envolvidos, de que se trata de uma atividade com data marcada, ainda que não haja um dia definido, em que o retorno acontecerá, normalmente, ao local de origem. E assim, os *frames* em tal domínio vão de moldando e sendo esmiuçados.

Não se pode, entretanto, estender essa mesma afirmação para os correspondentes de tradução. Não necessariamente, as palavras do dicionário terão um correspondente perfeito em todas as línguas. Em Peron-Corrêa (2014), isso é detalhado. Uma de suas importantes contribuições neste sentido foi mostrar como aqueles considerados os melhores equivalentes de tradução por diversos dicionários conceituados, na verdade, não evocam o mesmo *frame*. No caso do *frame* de Turismo_por_turista, cujos dados são mostrados no Quadro (1), enquanto algumas Unidades Lexicais cognatas apresentam-se como equivalentes de tradução para o domínio do turismo, *apreciar*/*apreciar*, *desfrutar*/*disfrutar* e *visitar*/*visitar*, outras, como *conhecer*, em espanhol, *conocer*, não permitem correspondência direta.

[9] Para mais informações sobre o a estruturação deste cenário, ver Gamonal (2013).

Quadro 1. Sumariamento dos usos de Unidades Lexicais nas línguas portuguesa e espanhola
Fonte: Peron-Côrrea (2014)

UL em Português	***Frame* evocado**	***Frame* secundário**	**UL em Espanhol**	***Frame* evocado**	***Frame* secundário**
conhecer	Turismo_por_turista	Conhecimento	conocer	SEM EQUIVALÊNCIA	
apreciar	Turismo_por_turista	Exp._em_foco	apreciar	Turismo_por_turista	Exp._em_foco
desfrutar	Turismo_por_turista	Exp._em_foco	disfrutar	Turismo_por_turista e/ou Atração_em_lugar	Exp._em_foco
visitar	Turismo_por_turista	Visitante	visitar	Turismo_por_turista	Visitante

Entretanto, o fato de não haver equivalência entre palavras cognatas para o domínio do turismo entre duas línguas não invalida a proposta de que se use a Semântica de *Frames*, através da FrameNet, como princípio organizador de dicionários multilíngues. Pelo contrário, isso reitera a importância da Semântica de *Frames* como meio para tratar adequadamente das implicaturas culturais envolvidas no processo tradutório. Uma vez que os equivalentes de tradução propostos pelo Dicionário FrameNet Brasil da Copa do Mundo se baseiam na comparação das valências sintático-semânticas dos itens lexicais, as quais, por sua vez, são oriundas de *corpora*, tal propositura pode se relevar muito mais ancorada na realidade do uso linguístico do que aquelas feitas *ad-hoc* ou com base exclusivamente na etimologia.

5. Considerações finais

A Semântica de *Frames* se coloca como sendo a hipótese que fundamenta as principais abordagens circunscritas na Linguística Cognitiva. Dessa forma, nada mais plausível do que explorar métodos de aplicação dessa teoria. O desenvolvimento do Dicionário FrameNet Brasil da Copa do Mundo correspondeu ao grande objetivo do projeto: construir um recurso lexical online multilíngue a partir da Semântica de *Frames* e da metodologia da FrameNet. A atividade turística apresentada em diversos *frames* ilustra um domínio transcultural, o que não significa que haverá sempre equivalentes perfeitos de tradução. Por outro lado, o recurso possibilita encontrar palavras dentro de um mesmo campo semântico com base em evidências oriundas de *corpora*.

Tendo isso em vista, percebe-se que a teoria linguística e a metodologia explorada tanto para a Lexicografia Computacional quanto no desenvolvimento de tarefas no âmbito do Processamento de Linguagem Natural podem gerar diferentes contribuições. A continuação deste trabalho acontecerá através da abordagem de outros domínios do conhecimento, que nos permitam explorar mais fenômenos linguísticos.

Referências Bibliográficas

BAKER, C. FrameNet, Present and Future. In: WEBSTER, J.; IDE, N. & CHENGUY FANG, A. (Eds.). **The First International Conference on Global Interoperability for Language Resources**. Hong Kong: City University, 2008.

BICK, E. **The parsing system PALAVRAS:** automatic gramatical analysis of Portuguese in a constraint grammar framework. 2000. 505p. Tese de Doutorado em Filosofia, Arhus University, Arhus, 2000.

FELLBAUM, C. (Ed.). **WordNet: An Electronic Lexical Database**. Cambridge, MA: MIT Press, 1998.

FILLMORE, C. J. The case for case. In: BACH, E. & HARMS, R. T. (Eds.). **Universals in linguistic theory**. New York: Rinehard and Winston, 1968a, p. 1-88.

_____. Lexical Entries for Verbs. **Foundations of Language**, v.4, n.4, 1968b, p. 373-393.

_____. The case for case reopened. **Syntax and semantics**, v. 8, 1977, p. 59-82.

_____. Frame semantics. In: THE LINGUISTICS SOCIETY OF KOREA (Ed.). **Linguistics in the Morning Calm**. Seul: Hanshin Publishing Co., 1982, p.111-137.

_____. Frames and the semantics of understanding. **Quaderni di Semantica**. v.6, n.2, 1985, p. 222-254.

_____.; CALLEJAS, C. M. B. Entrevista a Charles J. Fillmore. **Odisea**, n. 4, 2003, p. 41-48.

_____. Border Conflicts: FrameNet Meets Construction Grammar. In: **Proceedings of EURALEX 13**. Barcelona, 2008, p. 49-68.

GAMONAL, M. A. **Copa 2014 Framenet Brasil: diretrizes para a constituição de um dicionário eletrônico trilíngue a partir da análise de frames da experiência turística.** 2013. 146p. Dissertação de Mestrado em Linguística, Universidade Federal de Juiz de Fora, Juiz de Fora, 2013.

_____.; TORRENT, T. T. Frames como Interlíngua na Estruturação de Dicionários Eletrônicos Multilíngues de Domínios Especializados. **Revista da ANPOLL**, n. 37, 2014, p. 247-261.

GOFFMAN, E. **Frame Analysis:** An Essay on the Organization of Experience. New York: Harper & Row, 1974.

MINSKY, M. A framework for representing knowledge. In: WINSTON. P. (Ed.). **The Psychology of Computer Vision**. New York: McGraw-Hill, 1975, p. 211-277.

PERON-CORRÊA, S. R. **Copa 2014 FrameNet Brasil: *frames* secundários em unidades lexicais evocadoras da experiência turística em português e em espanhol**. 2014. 147p. Dissertação de Mestrado em Linguística, Universidade Federal de Juiz de Fora, Juiz de Fora, 2014.

RUPPENHOFER, J.; ELLSWORTH, M.; PETRUCK, M. R. L.; JOHNSON, C & SCHEFFCZYK, J. **FrameNet II**: Extended theory and practice. Berkeley: International Computer Science Institute, 2010.

SARDINHA, T. B. **Linguística de *Corpus***. São Paulo: Manole. 2004.

SALOMÃO, M. M. M. FrameNet Brasil: um trabalho em progresso. **Calidoscópio**, São Leopoldo: UNISINOS, v. 7 n. 3, 2009. p. 171-182.

SCHMID, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: **Proceedings of International Conference on New Methods in Language Processing**. Manchester, UK, 1994.

SCHMIDT, T. Interfacing Lexical and Ontological Information in a Multilingual Soccer FrameNet. In: **Proceedings of OntoLex 2006 – Interfacing Ontologies and Lexical Resources for Semantic Web Technologies**. Itália, 2006.

_____. The Kicktionary: A Multilingual Resource of the Language of Football. In: REHM, G., WITT, A. & LEMNITZER, L. (Eds.). **Data Structures for Linguistic Resources and Applications**. Tübingen: Gunter Narr, 2007.

_____. The Kicktionary: Combining *Corpus* Linguistics and Lexical Semantics for a Multilingual Football Dictionary. In: LAVRIC, E. et al. (Eds.). **The Linguistics of Football**. Tuebingen: Gunter Narr, 2008, p. 11–23.

_____. The Kicktionary – a multilingual lexical resource of football language. In: BOAS, H. (Ed.). **Multilingual FrameNets – Methods and Applications**. Berlin/New York: Mouton de Gruyter, 2009, p. 101-132.

TORRENT, T. T.;ELLSWORTH, M. Behind the labels: criteria for defining analytical categories in FrameNet Brasil. **Veredas**, v.17, n.1, 2013, p. 44–65.

_____.; SALOMÃO, M. M. M.; CAMPOS, F. C.; BRAGA, R. M.; MATOS, E. E.; GAMONAL, M. A.; GONÇALVES, J.; GOMES, D. S.; SOUZA, B. C. P. & PERON-CORREA, S. R. Copa 2014 FrameNet Brasil: a frame-based trilingual electronic dictionary for the Football World Cup. In: **Proceedings of the 25th International Conference on**

Computational Linguistics (COLING 2014) – System Demonstrations. Dublin, 2014, p. 10-14.

Artigo recebido em: 30.03.2015
Artigo aprovado em: 23.06.2015

Designing English Teaching Activities Based On Popular Music Lyrics From A *Corpus* Perspective

Desenvolvimento de Atividades de Ensino de Inglês com Base em Letras de Músicas Inglesas e Americanas: uma Perspectiva com Base em Linguística de *Corpus*

Maria Claudia Nunes Delfino*

RESUMO: O trabalho teve como objetivo principal utilizar uma abordagem baseada em Linguística de *Corpus* (LC) na produção de atividades didáticas baseadas em letras de música, a fim de estimular a construção do conhecimento do aprendiz de inglês como língua estrangeira. Para tanto utilizamos na análise dois *corpora*, sendo um denominado *corpus* de estudo, composto por aproximadamente 150.000 palavras provenientes de 585 letras de música em inglês de artistas britânicos e norte-americanos, e outro, de referência (COCA), que foi comparado ao *corpus* de estudo a fim de que as características léxico gramaticais mais salientes das letras de música fossem levantadas. A partir dos Itens Lexicais (ILs) encontrados, montou-se um banco de dados com vários tipos de exercícios, para que o professor possua uma variedade de exercícios a serem aplicados em sala de aula. Os referidos exercícios seguiram os seguintes critérios de elaboração: (1) os exercícios devem ser replicáveis, (2) de fácil adaptação e (3) pouco tempo de preparação para o professor, (4) com conteúdo fixo e variável, (5) além de serem divertidos, ou seja, que levem os alunos a se sentirem motivados a realizá-los e aumente seu interesse pelo aprendizado da língua com música.

PALAVRAS-CHAVE: Linguística de *Corpus*. Atividades didáticas. Itens lexicais. Música.

ABSTRACT: This work had as its main goal to focus on a *corpus*-based approach upon Corpus Linguistics (CL) for the production of teaching activities based on lyrics, so as to stimulate students to build their own learning. We analyzed two corpora, a study corpus, made of around 150,000 words from 585 lyrics from British and American singers and by a reference corpus (COCA), which was compared to the study corpus so that the most outstanding lexical grammar characteristics could be pulled up. A taxonomy of exercises was then developed, making up a variety of exercises, so that the teacher can have a range of exercises to be used in the classroom. Such exercises followed a list of criteria for corpus-based material design which included a number of recommendations, such as: (1) exercises should be replicable, (2) easily adapted, (3) they should not be too time-consuming (as this would limit the production of such materials and subsequently use in class), (4) some of the exercises must have fixed and variable content, (5) they should be fun, so that students feel motivated to complete them, increasing their interest in the learning of the language through songs.

KEYWORDS: Corpus Linguistics. Classroom activities. Lexical items. Music.

* Mestranda do Programa de Linguística Aplicada e Estudos da Linguagem (LAEL), PUC-SP / Professora da Faculdade de Tecnologia (Fatec-PG).

1. Introduction

English teachers have been facing a challenge for some time, which is how to teach the language in a pleasant way, knowing that the students are not in touch with the language outside the classroom. With the presence of new technologies, there is no reason for the teacher to be constrained by the book, which most of times, is not a suitable material and, even when it is, the classes can always be improved with movies, sitcoms, internet websites, games, newspapers and music.

Popular music has been used as a tool in the teaching of foreign language for a long time (Bertoli-Dutra, 2014), but it is usually seen as an extra material to be used when the teacher has some free time during the class or as an extracurricular activity. So, the question that many teachers and students have is how to contextualize, how can a song become relevant, how can an exercise be useful to the student, how can a class be fun and, at the same time, clear enough to help the understanding of a topic which is unclear in the book?

In the research reported here, we argue that popular music can be the central element in language teaching; in fact, in our proposal, all the exercises were based on popular music and on texts that draw on topics related to popular music. At the same time, our goal was not to teach "pop song" English, but current spoken English. To meet this goal, the analysis of the song lyrics was used as a starting point for the materials. The patterns found in the songs as well as their register characteristics were then used as search criteria in the reference English corpus and the patterns resulting from these searches were incorporated into the teaching materials.

In short, our proposal argues for the need of a blend of register-specific and general English corpus sources. After the elaboration of the exercises, they were implemented in a course of English as a foreign language for elementary students enrolled at a Technology College in the state of São Paulo. In this paper, we will report on the design of the corpora used in the research, present the main findings from the analysis of these corpora, and give examples of exercises which were made up based on the findings.

The questions which guide this work are (1) What are the lexical grammar patterns with highest frequency (Biber et al. 1998; Sinclair, 1991) in the pop song lyrics corpus (PSLC)? (2) How can teaching materials be built in a way that the patterns resulted from the research can be used as the starting and main point in an English class? (3) How does the teacher see the process of using these corpus-based materials?

1.1 Corpus **Linguistics**

Corpus Linguistics is an area that studies the language through samples of natural language; its analyses is usually carried out with specialized software programs on a computer. It is thus a method to obtain and analyze data quantitatively through description of features.

But what is a corpus? According to Berber Sardinha (2004), a corpus is a package of linguistic data (belonging whether to the spoken or written usage of language, or to both), which is systematized according to some criteria, being wide and deep enough, in such a way that it becomes representative of the totality of the linguistic use or of a part of it in a way that can be processed by computer, with the aim to promote useful results for description and analysis.

Among the different corpora which can be collected, we can mention the general corpus and the specialized corpus. A general corpus is a collection of texts used to explore the language and answer specific questions about vocabulary, grammar or discourse structure of a language. A specialized corpus is developed to fulfill the specific needs of a determined research work. The researcher can put the result of both corpora in contrast, and then observe and highlight his corpus particularities in relation to other linguistic genres which are present on the studied language. (Berber Sardinha, 2004). This approach was used on this work.

Halliday (1993) considers the language a probabilistic system, in which the words associate to one another, defining their use and functions. According to Corpus Linguistics, lyrics show patterns of regularities that can be used to teach the language. (BERTÓLI-DUTRA, 2010).

1.1.1 Corpus **Linguistics and Teaching-Learning of Languages**

From time to time, we look at language and language learning in a “new” way, which directly influences the way languages are taught. In our proposal, we highlight that language is made up by patterns which need to be taught to the learners of a Foreign Language (FL) because these patterns involve group of words which systematically co-occur and give a specific meaning to the texts where they are.

With the ever increasing popularity of studies in CL, tools and their application, there has been a huge interest of researchers in using CL on language teaching, because it can give relevant elements concerning word frequency, occurrence and co-occurrence of some lexical items. The teacher or researcher can count on the direct observation of linguistic phenomena,

which assures trust to the work, since the events are represented the way they occur and not as they are believed that can occur.

Huston (2002) states that the corpus has a direct impact on the FL teacher's professional activity in two different ways: first, it changes the way language is noticed and, secondly, it can be explored to produce material to be used in teaching, making a bank for the planning of new contents and methodology.

In 2010, Berber Sardinha made up a proposal of how to prepare teaching materials for FL with corpus, which goes beyond the use of concordance (relation of all the occurrences of a search word with its context in a corpus), but using tools such as word lists, key-words and clusters (group of words), beginning to work not only with concordance in written texts, but also focusing in music and video; he named this approach multimedia/multigenre, considering the fact that students of FL are in contact with the language in different types of media, that can be written (newspaper text), heard (news), prepared to be seen as if it were spontaneous (movies), among others. And among the media, without mentioning the traditional ones, like newspaper, magazine, book, telephone, music, radio and TV, learners are specially in contact with the new ones which appeared on the digital environment like podcast, Twitter, Youtube, email, WhatsApp, among others. The amount of genre and media is really huge and tends to grow and becomes more complex, with the genre transfer from one media to the other. As the society is each time more technologic, the classroom has to make the students able to perform well on these genres, or else, it has a great chance to become obsolete and boring.

1.1.2 Corpus **Linguistics and the English Teaching in Brazil**

Acunzo (2010) states that the combination of lexical items as lexical-grammar patterns and collocations in the teaching of a FL still is approached in a superficial and traditional way in Brazil, with little or no contextualization and exploration of the language in use. Thus, we do not know much what happens during the class when materials produced with corpora are used. Our research also focus on this subject, since reflexive journals were produced by the teacher/researcher to check the process of using CL and music to teach English.

This part of research is important because as Berber Sardinha (2010) states even with the development of research in the teaching area, it is very complicated to get the teachers, coordinators and principals' cooperation to get the data of such researches. This cooperation does not happen for many reasons, such as lack of interest in participating of the academic

research in general, they are afraid that the research shows weakness of methodology, teachers' lack of time, little opportunity to "escape" from the course program when the collect demands a specific frame or theme, among others. Because of this, it is important to develop (semi) automatic preparation tools and to encourage a philosophy which blends the reuse of material with the creation of a material bank, besides making a "modular" design (with interchangeable parts), to increase the reuse of materials.

This is one of the aims of this work, to create a data bank of materials to be used by teachers for the teaching of English based upon music and CL. One set of exercises developed in this bank of materials can be seen in the Attachment.

1.2 Music and English Teaching as a Foreign Language

One of the main points of this research, which in fact is one of the principles of investigation, is the use of the language that the student is learning in meaningful and relevant situations. For this situation to happen in the classroom, some conditions must be fulfilled (PELIZZARI et al., 2002). The first one is that the student has to be keen to learn, because if he does not want to learn or only wants to memorize contents, the learning will not happen. The second one is that the content has to be potentially meaningful, in other words, the logical meaning must depend only on the nature of the content, if it is aligned to the learner's needs and the psychological meaning is a unique experience that each person has. Each learner filters the contents which have or do not have meaning for himself. Finally, it is suggested that the students make meaningful learning for themselves, i.e., that each student becomes a researcher responsible for his own knowledge.

And how can one fit music in this context? In English, many songs are available with vocabulary from all levels, showing complexity or simplicity in the language, with a recurrent theme or story, offering authentic examples of colloquialism and a wide source of modern linguistic data.

The level of satisfaction and fun expressed by students and teacher, the rapport, in other words, the harmonious environment obtained through music makes up the integration, motivation and consequently, knowledge.

According to Silva (2015), the contact with songs (authentic productions in English), besides stimulating the desire to obtain more vocabulary/structure, makes the learner realize that the acquired knowledge can be applied in many other textual genres, written and orally.

Through songs, it is possible to expose students to the different kinds of English. British and American English are widely represented through modern songs and their respective accents. (MORENO, 2011).

Two studies, Domoney and Harris (1993) and Little (1983), investigated the prevalence of popular music in the FL learners' lives. Both studies show that music is the highest source of English outside the classroom. Paiva (1998), in another study, indicates that from each ten students interviewed, four listened to music to learn English. Therefore, we can infer that music, being part of people's daily activities, is an important tool, even knowing its purpose is not to be a teaching tool, pedagogical characteristics can be seen in songs, which can neither be ignored nor avoided in the classroom.

Music can mean knowledge, because for the student understands the lyrics, many times he has to infer, through the language he has learned previously and understands what he is listening. So through cognates, for example, he can make comparisons with his native language. Examples exercises which explore this can be seen on the Attachment (Exercises 7 and 8, where they are invited to find the answers by themselves on the lyrics, relying on their mother tongue and their previous knowledge of English).

The exposition to the authentic language on a fun way can stimulate the students' learning. The use of songs in English has different targets. Besides the linguistic and cultural aspect of the work with music in the FL teaching, we have to take into consideration the learner's emotional aspect and, according to Krashen (1982), for the effective learning to happen, it is necessary that the person has his affective filter "down", in other words, the person has to be relaxed and motivated. The affective filter takes emotional and attitudinal factors like motivation, self-confidence, anxiety and fear. For the author, motivated students who are full of self-confidence can show a better result than the ones who are anxious or afraid to expose themselves to their peers. The student who is able to express himself without being afraid of making a mistake usually has more chances of having a concrete learning than the one who does not speak for being insecure and consequently, misses opportunities to practice the language.

One can say that music, instead of helping can become a source of stress for the student, because singing in the classroom (in public) is more stressful than talking to a friend, or even one may consider that listening to a song can be harder to understand than listening to a person on a conversation without the "interference" of the musical instruments. But, for adult learners

to engage spontaneously on a conversation in another language, it is necessary that they have high confidence concerning his knowledge on the language, which is not always the reality of colleges students in Brazil, who usually do not have previous knowledge of the English language and feel intimidated when asked to talk; nevertheless, when the activity given by the teacher involves a song the students know or even a sitcom that they usually watch at home, they may feel more comfortable concerning the theme and risk to sing the song and/or repeat the dialogues of the sitcom, sometimes getting involved in conversations about their topic, maybe because the chosen topics are in the context the learners know, what can put the fear of making mistakes down or even abolishing it.

Murphey (1990) states that the language which is learned through music can occur in more quantity and with better fixation since it motivates people to learn and provides a bridge between the school and the world. Harmer (2013), referring to music, recognizes a relationship between music and the practice of the language, stating that music can be used to create a rapport, to stimulate the imagination and to make the learner to talk about the songs he appreciates. He still adds that it can be used to explore vocabulary and grammar.

Music in the classroom makes the contents become more dynamic and meaningful, being able to ease the learning. Nevertheless, the English teacher and/or learner will hardly ever see the lyrics on the book used in the classroom. The absence of lyrics in books can be explained due to its high cost, which would increase even more the cost of materials, just like McCarthy (1998) states.

Despite economical problems, the constant exposition to the lyrics outside the classroom, or alternatively, inside the classroom, but independently from the book, can influence the memorization of some lexical grammar items. Songs, when automatized, can become an important linguistic reference because they have the potential to contribute to the transfer of input (listening, reading) for output (writing, speaking), on the moment the students are put in communicative use.

The practice with songs can have a positive impact not only on the oral understanding but also on the oral production. After some time using sounds and rhythms found in music as models, it is possible to recognize them in other songs and produce them in communicative situations. (SILVA, 2015)

1.3 Reflexive Journals

Another point on this research is to know about the process of using materials based on Corpus Linguistics and music as a means to teach English as a foreign language. This process was analyzed by the researcher through reflexive journals produced by the researcher/teacher on the course of the classes.

But, what is a reflexive journal? It is the act of writing experiences, where the user talks about his feelings and opinions, instead of only describing the facts, which happened during the day (Machado, 1998). According to the author, through the writing of individual feelings, new experiences can be built and improved. The aim of this journal is to improve the learning through the process of thinking and writing about the learning experiences, the good and the bad ones.

Besides the teacher, the students also have a benefit from this journal because, from the experiences in the classroom the teacher, many times, rethinks his behavior and reflection means change in pedagogical practice, according to Liberali (1999).

2. Methodology

2.1 Corpus

The students were asked to vote on their favorite bands and singers, so that the teacher could make a research based on music and the creation of teaching materials involving Corpus Linguistics. The most voted bands were Beatles, Bon Jovi and Maroon 5, and the singer Bruno Mars.

All the lyrics from the mentioned bands and singer (585 lyrics, which made up about 150,000 words) were collected from their official websites and saved in plain texts, without the title of each song, so that they could be used on the Ant Conc program.

Chart 1. Corpus Design

Bands	Number of Songs	Number of Tokens
Beatles	202	34,888
Bon Jovi	220	59,964
Bruno Mars	92	30,593
Maroon 5	71	21,211

Source: Delfino (2015) for this research.

According to chart 1 we can see that the Band Bon Jovi has more lyrics than the others. This is because they have been more time on the road and, consequently, have recorded more than the others. About Beatles, this corpus only comprises what the band recorded as a group and not as individual singers. The corpus was collected in February 2014 and, by the end of this year Bruno Mars and Maroon 5 recorded more songs, which are not on the corpus.

Lists of words were made on the Ant Conc program so that we could pull up the 100 most frequent words from the Corpus of Songs. These words were compared to the 100 most frequent words from the Corpus of Contemporary American English (COCA).

After that, concordances lines were run for each of the words that were present in both corpora.

2.2 Design of the Activities

A list of criteria for the design of the exercises was developed by the researcher, so that the exercises could gain more focus. It comprised 28 items and it was decided that each exercise had to fulfill at least 03 of these criteria, as follows:

Chart 2 – List of Criteria for Development of Corpus Activities

1. The exercise uses corpus
2. The exercise has clear goals
3. The exercise has as main focus the study of meaning and not of form
4. The exercise is ethical
5. The exercise is replicable
6. The exercise is participative
7. The exercise is collaborative
8. The teacher is a facilitator and not a distributor of knowledge
9. The student is a discoverer, researcher and not a recipient of knowledge
10. The exercise does not demand excessive preparation time from the teacher
11. The exercise deals with the concept of patterning
12. The exercise deals with the concept of frequency
13. The exercise deals with the concept of variation
14. The exercise deals with the concept of textual varieties
15. The exercise includes different media
16. The exercise includes concordances
17. The exercise includes texts
18. The exercise includes word frequency lists
19. The exercise includes key words
20. The exercise enables students to work directly with the corpora
21. The exercise includes diagrams and different ways of visualization
22. The exercise can be easily adapted

23. The exercise is motivator
24. The exercise uses authentic language
25. The exercise has a relevant content for the student
26. The exercise has a suitable difficulty level
27. The exercise develops autonomy
28. The exercise teaches and not only tests.

Source: Delfino (2015) for this research.

A taxonomy of exercises was also developed, with the aim of focus on the kinds of exercises the researcher wanted to test on the students, using music and corpus linguistics:

Chart 3 – Taxonomy of Exercises to Be Used with Corpus Linguistics and Music

1. blank filling
2. singing
3. concordance analysis
4. awareness
5. patterns relation
6. charts analysis
7. register analysis
8. texts comprehension
9. writing sentences
10. word clouds
11. listening

Source: Delfino (2015) for this research.

From the results of the pattern analysis and the use of the lists of the 100 most frequent words in both corpora, activities comprising the criteria and taxonomy above were produced and tested on the students.

2.3 Reflexive Journal

At the end of each class, the teacher wrote the impressions she had from the class, concerning the process of teaching English based on Corpus linguistics and music. Some of the findings were important to the improvement of the design of the exercises present in the activity. An example of it can be seen on the following sentences from the journal:

> …"One thing I have to keep is the use of COCA website in the classroom for the students to check different meanings of a word on the concordance lines. Besides doing this today, they also got to the conclusion that the word SO is frequent on the music corpus because it is very close to a conversation. They checked SO on the chart section of the website and could see this word is very frequent on the Spoken section, but not very common on the written one. From this point we started working with other words, like CONSEQUENTLY, which was also researched."

This was the starting point for a reflection which led the researcher to create, in all the lessons, an exercise involving the charts section from COCA, which is very simple to be made, not spending too much time from the teacher on the preparation and students love it. Such exercise can be seen on the Attachment section (Exercise 10).

3. Results

About the Lexicogrammatical analysis, we can see that some items co-occur with others making patterns which can be used to teach the language. For example, the lexical item GET, which is the 34th most frequent word in the corpus of songs and the 92nd most frequent word on COCA, has a high "attraction" for prepositions, as we can see on Chart 4.

Chart 4: GET + preposition; example with frequency on the corpus:

Get to – 96 times
Get up – 33 times
Get in - 09 times
Get back – 60 times
Get by – 20 times
Get down – 08 times

Source: Delfino (2015) for this research.

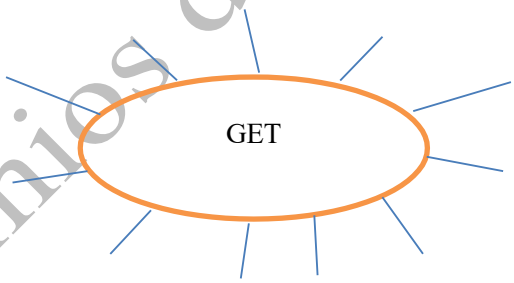
Concerning the design criteria and taxonomy, each exercise was fit into the list of criteria and at least one of the taxonomy. We can see this on the following exercise, which meets the following criteria from the list (Chart 2): (1) Criteria 1 - the exercise uses corpus, because for the researcher to develop it, it was necessary to go to the music corpus; (2) Criteria 11 – the exercise deals with the concept of patterning, since students are supposed to look at patterns to be able to complete the exercise; (3) Criteria 21 – The exercise includes diagram and different ways of visualization, because it is not on the written form only. The researcher wanted the students to deal with different ways of visualization and not only written texts; (4) Criteria 26 – The exercise has a suitable difficulty level, which is important students who are learning a new language can not be frustrated by receiving exercises they are not able to do because they do not know the lesson, or worse, a barrier to their learning learning process can be created because of this, which is very hard to break.

About the taxonomy, this same exercise fell within the following 02 types (Chart 3): (1) Type 03 – Concordance analysis, because through the concordance analysis the students can

get to conclusions concerning the patterns; (2) Type 05 – Patterns Relations, which is connected with the concordance.

> **Exercise 2**: On this song we saw some meanings of the word GET. Which meaning do you think is the most common? One of the greatest ways to learn English is to know how to find the answers on your own. One way to find the answers on your own is to research on the corpus. So, how about checking this information? Let's check these concordance lines, taken from the corpus PSLC and make a word map with the collocates of GET, i.e., the words which make company to GET most frequently.

1	I would gladly hit the road **get** up and go if I knew That's
2	Why do we let the pressure **get** into our heads? Your broke
3	you better do what she said **Get** to the barber shop and get
4	your dreams before they rust **Get** what you can and hope it's
5	u know I would beg and plead, **get** down on my knees Do most a
6	to show you got bills to pay, **get** out my way it's time to
7	streets No time for praying **get** up off your knees There's
8	If you and me should **get** together, Who knows baby
9	lookin' fine ain't she? … **get** out the way! No, no, no
10	don't need you any longer So **get** off your knees Your words
11	That I've **got** the key Oh So **get** in the car We can ride it
12	That I've **got** the key Oh So **get** in the car We can ride it



4. Final Considerations

From the findings in this research, we could get to some conclusions. Firstly, concerning the lexicogrammar analysis, we can say that the most frequent words found in the corpus of lyrics use common pattern found in spoken English and that songs can be used to learn spoken patterns.

About the design criteria and taxonomy, exercise design should be based on explicit criteria because it gains a better focus and can be used to teach the linguistic characteristics associated with the patterns found on the corpus of lyrics.

And finally, through the reflective journal we could see interesting points on the process of using corpus linguistics and music to teach English, such as: teaching with corpus is challenging, because it's a novelty for most students and teachers, but they are mostly receptive and are keen to use the technology they learn, science students are the most receptive and are glad to explore the quantitative side of language research; keeping a journal is "more work", but at the same time helps teachers learn to teach.

Reference

ACUNZO, C. M. *Corpora* no Ensino de Línguas Estrangeiras Como usar a Linguística de Corpus no ensino de língua estrangeira. In: VIANA, V.; TAGNIN, S.E. O. (org.) ***Corpora*** **no Ensino de Línguas Estrangeiras**. São Paulo: Hub Editorial, 2010.

BERBER SARDINHA, T. **Linguística de *Corpus***. São Paulo: Manole, 2004

_____. Como usar a Linguística de *Corpus* no ensino de língua estrangeira. Por uma Linguística de *Corpus* Educacional Brasileira. In: VIANA, V.; TAGNIN, S. E. O. (org.) ***Corpora*** **no Ensino de Línguas Estrangeiras**. São Paulo: Hub Editorial, 2010.

BERTÓLI-DUTRA, P. **Linguagem da Música Popular Anglo-Americana de 1940 a 2009**. PhD Thesis. PUC-SP, 2010

_____. Multi-Dimensional analysis of pop songs. In: BERBER SARDINHA, T.; VEIRANO PINTO, M. (editors). **Multi-Dimensional Analysis, 25 years on**. A Tribute to Douglas Biber. Amsterdam, Philadelphia: John Benjamins, 2014, pp. 149-176. crossref http://dx.doi.org/10.1075/scl.60.05ber

BIBER, D.; CONRAD, S.; REPPEN, R. ***Corpus*** **linguistics** – Investigating language structure and use. Cambridge: Cambridge University Press, 1998. crossref http://dx.doi.org/10.1017/CBO9780511804489

DOMONEY, L.; HARRIS, S. Justified and ancient: Pop music in EFL classrooms. **ELT Journal**, 47, pp. 234-41, 1993. crossref http://dx.doi.org/10.1093/elt/47.3.234

HALLIDAY, M. A. K. Quantitative studies and probabilities in grammar. In: HOEY, M. (Ed.), **Data Description Discourse** - Papers on the English language in Honour of John McH Sinclair on his Sixtieth Birthday pp.1-25. London: HarperCollins, 1993.

HARMER, J. Interview. **Revista New Routes**. São Paulo, n. 50, p. 10-13, 2013. Available in : http://www.disal.com.br/newr/nr50/nrlogin.asp?anterior=nr50&A1=5683938012171572585220&A2=C. Acessed on December 22, 2014.

KRASHEN, S. D. **Principles and Practices in Second Language Acquisition**. Oxford: 1982.

LIBERALLI, F. C. **O diário como ferramenta para a reflexão crítica**. PhD Thesis. Pontifícia Universidade Católica de São Paulo. São Paulo, 1999.

LITTLE, J. Pop and rock music in the ESL classroom. **TESL Talk**, 14, 1983, pp. 40-4.

MACHADO, A. R. **O Diário de Leituras.** A Introdução de um novo Instrumento na Escola. São Paulo: Martins Fontes. 1998

MCCARTHY, M. **Interview**. Available in: http://www.cambridge.org.br/authors-articles/interviews?michael-mccarthy-ii&id=2460. Acessed on December 22, 2014.

MORENO, T. A. O ensino da língua inglesa através das músicas e das tecnologias. In **Webartigos**. Available in: http://www.webartigos.com/artigos/o-ensino-da-lingua-inglesa-atraves-das-musicas-e-das-tecnologias/66290/. 2011. Accessed on August 05, 2014.

MURPHEY,T. **Music and Song**. Oxford: Oxford University Press, 1990.

PAIVA, V. L. M. O. Estratégias individuais de aprendizagem de língua inglesa. **Letras e Letras**. v. 14, n. 1, jan/jul. 1998. pp. 73-78.

PELIZZARI, A. *et al.* Teoria da Aprendizagem Significativa segundo Ausubel. In: **Rev. PEC**, Curitiba, v.2, n.1, p.37-42, jul.2001 – jul.2002.

SILVA, E. B. **Proposta de uma Classificação de Letras de Música em Língua Inglesa a partir da Léxico-Gramática**. Classificando Músicas em Inglês pela Léxico-Gramática. (Personal Communication), 2015.

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

Attachment

An example of a whole activity involving 18 exercises:
Exercise 1: Read the title of the song. What may be the topic of today's class?
a-) Born to be bad b-) Born to be rich c-) Born to be my baby
d-) Born to be a star

Born To Be My Baby- Bon Jovi

Rainy night and we worked all day
We both got jobs cause there's bills to pay
We got something they can't take away
Our love, our lives

Close the door, leave the cold outside
I don't need nothing when I'm by your side
We got something that'll never die
Our dreams, our pride

My heart beats like a drum (All night)
Flesh to flesh, one to one (And it's alright)
And I'll never let go 'cause
There's something I know deep inside

You were born to be my baby
And baby, I was made to be your man
We got something to believe in
Even if we don't know where we stand

Only God would know the reasons
But I bet he must have had a plan
'Cause you were born to be my baby
And baby, I was made to be your man

Light a candle, blow the world away
Table for two on a TV tray
It ain't fancy, baby that's ok
Our time, our way

So hold me close better hang on tight
Buckle up, baby, it's a bumpy ride
We're two kids hitching down the road of life
Our world, our flight

If we stand side by side (All night)
There's a chance we'll get by (And it's alright)
And I'll know that you'll live
In my heart till the day that I die

'Cause you were born to be my baby
And baby, I was made to be your man
We got something to believe in
Even if we don't know where we stand

Only God would know the reasons
But I bet he must have had a plan
'Cause you were born to be my baby
And baby, I was made to be your man

And my heart beats like a drum (All night)
Flesh to flesh, one to one (And it's alright)
And I'll never let go 'cause
There's something I know deep inside

Exercise 2: This song works with a new pattern, which is present on its title. Let's complete the concordance lines with it?

1	_____rock, but nonetheless basically a quiet, unassuming guy.
2	With that name he was_____direct the Spider Man movie.
3	Farmers were work in the light and air.
4	My first boy was hunt it.
5	It is the kind of labor I was .
6	Like I was do this one thing.
7	Humans were manipulate.
8	Clean Bottle was_____stem the tide of these landfill-bound vessels.
9	Whose debute CD as Lana Del Rey,_____Die, comes out this month.
10	As if she were_____sit and write beautiful, perfect poems.
11	I was be a model.
12	I was be rich.

Exercise 3: Let's go to the website http://en.wikipedia.org/wiki/Born_to_Be_My_Baby to answer:

When was the song released? _____

Who is (are) its composers? _____

Did this song reached the top parade? Where? _____

Who appears on the videoclipe? Was it a very expensive videoclipe to be produced? _____

Exercise 4: Read the lyrics of the song and answer: What's the song about?

Exercise 5: Answer True (T) or False (F) to the following sentences according to the text:

1- Born to Be My Baby was the only song from the album *New Jersey* to chart in the Top 10. ( )
2- Born to Be My Baby was included in the greatest hits *Cross Road* album. ( )
3- The video for the song is very colorful. ( )
4- The video for the song was very expensive. ( )
5- Bon Jovi's wife is in the video. ( )

Exercise 6: Choose 05 words on the lyrics thtat you don't know the meaning. Let's look them up on the online dictionary?

_____

_____

Exercise 7: Now choose 05 words on the lyrics which are similar to words in Portuguese. Let's look them up on the online dictionary too? Do they have a different meaning from the one you first imagined?

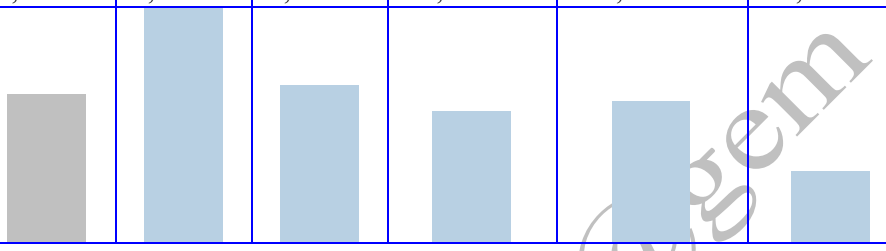
_____

_____

Exercise 8: Pronunciation: In English, when a word finishes on the letter "e" and the next one starts with the letter "a" the final "e" is not pronounced. So, "take away" is pronounced | **teɪk** əˈ**weɪ** | and "like a drum" is pronounced | ˈ**laɪk ə drʌm** |.

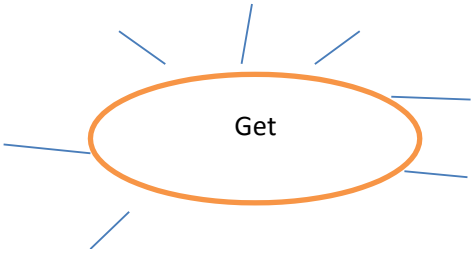
Exercise 9: Now go back to the song, find and highlight **GET + Complement.** Don't forget that GET is the lemma, so we have to consider all its forms present on the song (GET / GOT).

Exercise 10: Now let's observe the COCA charts. Where is GET more frequent?

SECTION	ALL	SPOKEN	FICTION	MAGAZINE	NEWSPAPER	ACADEMIC
FREQ	1549477	504741	322331	285342	291052	146011
PER MIL	3,336.97	5,281.65	3,564.45	2,986.04	3,173.35	1,603.35
SEE ALL SUB-SECTIONS AT ONCE						

Exercise 11: On this song, we saw some meanings of the word GET. Which meaning may be the most common? One of the greatest ways to learn English is to know how to find the answers by your own. One great way to find the answers on your own is to research on the corpus. So, let's check this information? Let's look these concordance lines, taken from the corpus CoEL and make a word map with the collocates from GET, i.e, which words keep company to GET with more frequency.

1	I would gladly hit the road **get** up and go if I knew That's
2	Why do we let the pressure **get** into our heads? Your broke
3	you better do what she said **Get** to the barber shop and get
4	your dreams before they rust **Get** what you can and hope it's
5	u know I would beg and plead, **get** down on my knees Do most a
6	to show you got bills to pay, **get** out my way it's time to
7	streets No time for praying **get** up off your knees There's
8	If you and me should **get** together, Who knows baby
9	lookin' fine ain't she? … **get** out the way! No, no, no
10	don't need you any longer So **get** off your knees Your words
11	That I've **got** the key Oh So get in the car We can ride it
12	That I've got the key Oh So **get** in the car We can ride it



Exercise 12: Let's learn a little more, observing other patterns and classifying them according to its **pattern** and meaning? One of the meanings will repeat.

Column A	Column B
1. **Beatles:** But when I get home to you (*A Hard Day's Night)* 2. **Beatles:** To get you money to buy you things (*A Hard Day's Night)* 3. **Beatles:** I'll get you anything my friend. (*Can't Buy me Love)* 4. **Beatles:** I get by with a little help from my friends. (*With a Little Help from my Friends)* 5. **Beatles:** I get high with a little help from my friends. (*With a Little Help from my Friends)* 6. **Bon Jovi:** We both got jobs cause there's bills to pay (*Born to Be my Baby)* 7. **Maroon 5:** The taste of her breath, I'll never get over (*Won't Go Home without You)* 8. **Maroon 5:** You and I get sick, yeah, I know that we can't do this no more. (*One more Night)*	a-) superar b-) possuir, ter c-) ficar doente d-) conseguir e-) ficar louco f-) chegar g-) sobreviver

Exercise 13: Now repeat what you did on exercise 11 with the reference corpus COCA, on the Academic section. Are the words the same? Why do you think this happen?

1	Such as open the door or **get** key.
2	And **get** as transitive verb with aliases grab
3	The mechanism and network state required to **get** a context chunk are sufficient to also get all of its keys.
4	The mechanism and network state required to get a context chunk are sufficient to also **get** all of its keys.
5	We **get** their key any time we receive content from them
6	Can easily **get** the key of parc.com
7	When an Interest does not **get** a response and times out.
8	So to **get** them
9	So that we can make choices before we **get** to the check out
10	So the South Asian community **gets** diabetes more commonly
11	Generally people with type 2 diabetes do not **get** any benefit
12	Having to **get** up at night to urinate

Exercise 14: Let's go to the program AntConc, on the CLUSTERS section and we will find the list below. Which words make company to GET? Are they the same ones from the song we studied? And from the COCA? What is similar? What is different? Talk to your friends and write your findings down. This is a great way to investigate the language in use and learn a new language on a new and efficient way.

Exercise 15: Complete the sentences from COCA with the clusters from the chart above:

1-_____here. Leave me alone.

2- We had to_____our hands and knees to try to find peace.

3- When you_____hospital.

Exercise 16: In all the sentences below (taken from the SketchEngine program) the verb GET is being used. Pick two of them and replace GET + complement for another verb, without changing the meaning of the sentence:

a-) GET rid of it.

b-) When the Belgian songwriters' association GOT involved.

c-) But he probably GOT a big welcome at the airport when he arrived.

d-) Her money note at the end still GOT a big cheer.

e-) He GOT just six points.

f-) But Gary still GOT twice as many points as Norway.

g-) Though I was only able to GET tickets to the Friday evening dress rehearsal.

h-) The following week I GOT a parcel containing the famous tracksuit.

i-) Visit here often and GET your baseball news from Thomas Harding

_____

_____

Exercício 17: Musical Bingo for Born to Be my Baby by Bon Jovi

Write the following expressions on the bingo chart below, in any order. So cross them out as you listen to them on the song.

Take away	Leave the cold outside	Beats like a drum
Flesh to flesh	Don't make it bad	Hang on tight
Side by side	We'll get by	It ain't fancy
Don't need nothing	Let her under your skin	Buckle up
Our way	Deep inside	Table for two
Never let go	That's ok	Hold me close

	Born to Be my Baby	

Exercise 18: How about singing the song on the karaoke? Let's go to the address http://www.youtube.com/watch?v=T6oyujbaw1E

Artigo recebido em: 30.03.2015
Artigo aprovado em: 22.06.2015

O uso de *corpora* comparáveis na pesquisa terminológica bilíngue

Using comparable *corpora* for biligual terminology research

Marina Araújo Vieira[*]
Silvana Maria de Jesus[**]

RESUMO: Este trabalho insere-se na interface entre os Estudos Terminológicos e os Estudos da Tradução baseados em *corpora*, com foco no uso de *corpora* comparáveis. Sendo o Espiritismo uma religião bastante desenvolvida no Brasil, é grande o número de traduções produzidas a partir de obras espíritas brasileiras, sendo que essas traduções são exportadas para vários países do mundo. Nesse sentido, este trabalho objetivou analisar termos espíritas, sobretudo referentes à mediunidade, utilizados em obras espíritas brasileiras (Português Original, PO), e seus equivalentes em obras traduzidas para o inglês (English Translated, ET). Para validar o uso desses equivalentes, analisou-se, ainda, sua ocorrência em obras escritas originalmente em inglês (English Original, EO). Com esse intuito, a metodologia empregada foi a Linguística de *Corpus*, que permitiu o uso de *corpora* comparáveis e paralelos. A metodologia consistiu na (a) seleção dos termos específicos dessa área, (b) na elaboração de fichas terminológicas e na seleção de termos multivocabulares, a partir dos colocados, para (c) criação de uma amostra de glossário bilíngue. Os resultados apontaram para a existência, em inglês original, de opções que não foram contempladas pelos tradutores nas traduções para o inglês.

PALAVRAS-CHAVE: Terminologia. Estudos da Tradução baseados em *corpora*. *Corpora* comparável e paralelo. Glossário. Espiritismo.

ABSTRACT: This work falls within the scope of Terminology and Corpus Based Translation Studies, as it analyses the spiritist terminology in Portuguese and English making use of comparable corpora. Spiritism (also known as Spiritualism in Anglo-Saxon cultures) is a well-developed religion in Brazil and many of the Brazilian spiritist books are translated into many languages. For that reason, this research aimed to analyze spiritist terms related to mediumship used in Brazilian spiritist books and their equivalents in books translated into English. In order to verify the use of these equivalent terms, they were compared to the ones used in texts originally written in English. Corpus Linguistics was used as the basic methodology for corpus analyses, which involved both comparable and parallel corpora. The methodology consisted of the following steps: (a) selection of the specific terms related to mediumship, mainly the multi-words, i.e., terms with more than one word, and their equivalent in English identified in the parallel and comparable corpora; (b) elaboration of terminological records with the terms in original and translated English; and (c) building of a bilingual sample glossary. The results indicated that many terms used in translation were not used in non-translated English.

KEYWORDS: Terminology. Corpus based Translation Studies. Comparable and parallel corpora. Sample glossary. Spiritism/Spiritualism.

[*] Bacharel em Tradução. Curso de Tradução, ILEEL, Universidade Federal de Uberlândia (UFU).
[**] Profa Dra. do Curso de Tradução, ILEEL, Universidade Federal de Uberlândia (UFU).

1. Palavras iniciais

Em virtude do volume de traduções de obras espíritas brasileiras para o inglês, este trabalho buscou investigar os termos relacionados ao campo semântico que se pode denominar "mediunidade" em obras espíritas originalmente escritas em inglês, referidas neste trabalho como EO (English Original), e em obras traduzidas para esse idioma, referidas como ET (English Translation). Com base nessa análise, as autoras elaboraram um glossário bilíngue[1] com os termos selecionados, contrastando os equivalentes[2] encontrados em textos traduzidos e não traduzidos em inglês.

O Espiritismo é uma filosofia de caráter científico e religioso considerado, nesta pesquisa, como uma linguagem de especialidade por apresentar uma terminologia[3] específica, cujos termos fazem referência a conceitos e realidades próprios desse contexto religioso, os quais são utilizados para estabelecer a comunicação e proporcionar o entendimento entre os praticantes.

Esta pesquisa iniciou-se com um trabalho de Iniciação Científica (VIEIRA; JESUS, 2013), no qual se analisou um *corpus* paralelo, ou seja, buscou-se mapear os termos referentes ao mundo espiritual na obra *Nosso Lar*, de Chico Xavier (1992) e identificar os equivalentes usados na tradução para o inglês, *The Astral City* (XAVIER, 2000). Nesse trabalho, observou-se a necessidade de validar os termos encontrados na tradução em inglês por meio da análise de textos espíritas não traduzidos em inglês, ou seja, observar se os equivalentes utilizados na tradução seriam também recorrentes em textos escritos originalmente em inglês.

Essa abordagem foi possibilitada pela metodologia proposta pelos Estudos da Tradução baseados em *corpora* (BAKER, 1993, 1995, 1996, 2004), que permitiu estudar o uso de termos do Espiritismo em inglês em obras traduzidas e não traduzidas; pela Linguística de *Corpus* (BERBER SARDINHA, 2004), para processar os textos e analisar os termos, buscando as equivalências; e pelo referencial teórico da Terminologia (KRIEGER; FINATTO, 2004) para estudar os termos na linguagem de especialidade, bem como elaborar as fichas terminológicas e a amostra do glossário mediúnico bilíngue (inglês/português).

[1] O glossário elaborado na pesquisa que deu origem a este texto apresenta 36 termos, mas não foi reproduzido aqui por questão de espaço. Destes, apenas três termos foram explorados com mais detalhe neste artigo.
[2] Em Terminologia, há distinção entre *equivalente* e *correspondente*, mas esse ponto não será abordado aqui. Ver Silveira (2005).
[3] O uso de terminologia e Terminologia segue o critério de Barros (2007), que afirma: "A Terminologia é o estudo científico dos termos usados nas línguas de especialidade", enquanto a terminologia é o "conjunto de termos de uma área especializada" (BARROS, 2007, p. 11).

A análise desenvolveu-se no sentido de responder à seguinte pergunta: há variação na terminologia espírita encontrada nos textos traduzidos e não traduzidos? As traduções utilizam os mesmos termos usados em inglês original, aproximando-se da convencionalidade e da idiomaticidade do inglês, ou fazem uso de opções mais literais, mais próximas dos termos que aparecem em português original?

Os resultados encontrados apontaram para a literalidade e podem ser relacionados com o conceito de "tradutor ingênuo" proposto por Tagnin (2005), o qual não é capaz de identificar expressões fixas da língua, "fugindo" da convencionalidade, e muitas vezes traduzindo os termos literalmente. Cabe observar que por tradução literal entende-se o procedimento de tradução "em que se mantém uma fidelidade semântica estrita, adequando, porém, a morfo-sintaxe às normas gramaticais da LT [língua da tradução]" (AUBERT, 1987 *apud* BARBOSA, 1990, p. 65). Esse tipo de tradução, usado no ET, diferencia-se das opções encontradas em EO, como pode ser observado na seção de análise e discussão dos dados.

2. A interface: Tradução, Terminologia e Linguística de *Corpus*

A Linguística de *Corpus* tem crescido enquanto abordagem teórico-metodológica utilizada na interface com a Terminologia e os Estudos da Tradução (VIANA; TAGNIN 2010; TAGNIN; BEVILACQUA, 2013). A Linguística de *Corpus* oferece a metodologia para a Terminologia organizar o vocabulário, bilíngue ou monolíngue, de uma dada área do conhecimento, de forma que esse produto fique disponível para uso do tradutor. É nessa interface que se posiciona esta pesquisa, cujo objetivo foi mostrar o caminho para a elaboração de um glossário bilíngue com base em *corpora* comparáveise paralelos. Essa associação vem apresentando resultados satisfatórios, como apontado por Almeida e Correia (2008, p. 72): "para a sistematização de terminologias, sobretudo em projetos que visam à elaboração de produtos, tais como dicionários, glossários, vocabulários, ontologias, bases terminológicas, etc., é fundamental a utilização de *corpus*".

A seguir serão apresentados os conceitos e os preceitos teóricos de cada área – Terminologia e Estudos da Tradução baseados em *corpora* – com suas devidas interseções.

2.1 Terminologia

A Terminologia é a área do saber que estuda o vocabulário das linguagens de especialidade. Segundo Aubert (1996, p. 27):

> por linguagem de especialidade entende-se, genericamente, o conjunto de marcas lexicais, sintáticas, estilísticas e discursivas que tipificam o uso de um código linguístico qualquer em ambiente de interação social centrado em uma determinada atividade humana.

Cada componente do vocabulário de uma linguagem de especialidade é chamado de termo, que Barros (2007, p. 11) define como "unidade lexical que designa um conceito de um domínio de especialidade. É também chamado de unidade terminológica". Cada unidade terminológica representa um conceito específico da área de especialidade, que pode ser diferente do significado conhecido em língua geral. Em uma situação bilíngue, como é o caso deste estudo, é preciso identificar os termos específicos tanto na língua A quanto na língua B.

Nesse sentido, a Terminologia vai ao encontro dos Estudos da Tradução de forma a estudar esses termos e produzir materiais terminológicos bilíngues para uso do tradutor, o que tende a garantir maior confiabilidade para o material traduzido. Como afirmam Krieger e Finatto (2004, p. 72), "[embora] a Terminologia não seja um requisito sem o qual a prática tradutória não se efetue, [...] [ela] é uma forma de tornar seu ofício mais consciente e facilitado". A seguir serão apresentadas as questões teóricas referentes aos termos mono e multivocabulares.

2.1.1 Termos monovocabulares e multivocabulares

A busca por termos da mediunidade neste trabalho voltou-se para a identificação de termos multivocabulares. Essa terminologia é usada por Aubert (1996), que define termo monovocabular como termo composto por apenas uma palavra e termo multivocabular como termo composto por duas ou mais palavras. Aubert ainda afirma que, nos domínios específicos, mais da metade dos termos é multivocabular, e que esse fato aponta para uma maior especificidade da terminologia analisada. Bowker e Pearson (2002, p. 168) também comentam o assunto: "Embora um termo possa ser constituído de apenas uma palavra, os termos são frequentemente compostos por mais de uma palavra".[4]

No entanto, quando o terminólogo não é iniciado na área de conhecimento da qual sistematiza a terminologia, torna-se mais difícil delimitar os termos multivocabulares, ponto em que a Linguística de *Corpus* auxilia a Terminologia. Segundo Aubert (1996), uma das

[4] "Although it is possible for a term to consist of a single word, terms are frequently composed of multiple words." (tradução nossa)

formas de identificar os termos multivocabulares é pela frequência estatística do termo. Nesse caso, o trabalho com *corpora* é relevante, como ocorreu nesta pesquisa, pois a ferramenta de processamento dos textos permite a verificação da frequência de combinações ou coocorrências, que podem ser identificadas como termos. Além disso, o programa utilizado nesta pesquisa, o *AntConc* (ANTHONY, 2011), aponta as palavras que ocorrem com maior frequência ao redor do termo pesquisado.

Nesse ponto, há uma convergência de conceitos que deve ser esclarecida: o uso de colocação e termo multivocabular. A busca por termos multivocabulares vai ao encontro da pesquisa por colocações. Dayrell (2005) desenvolve um trabalho dentro da Linguística de *Corpus* que busca colocações em obras de autoajuda e ficção em português brasileiro original e traduzido, ou seja, um *corpus* comparável. A definição de colocação utilizada pela autora é a de Sinclair (1991 *apud* DAYRELL, 2005): "ocorrência de duas ou mais palavras próximas umas das outras dentro de um texto"[5].

A busca feita nesta pesquisa iniciou-se pelos termos monovocabulares, por meio da análise da lista de palavras-chave, por exemplo, *médium*. Em seguida, buscaram-se as palavras que coocorreram com os termos, de forma a identificar colocações, como *médium psicofônico*. No entanto, por se tratar de uma pesquisa terminológica, pode-se afirmar que uma colocação, nesse caso, pode ser um termo no domínio de especialidade. Daí o emprego de colocação, coocorrência e termo multivocabular para fazer referência aos agrupamentos de duas ou mais palavras que podem constituir termos. Em outras palavras, dois itens lexicais que coocorrem frequentemente são considerados colocações, mas somente constituem um termo se designarem um conceito específico de uma área de especialidade.

2.2 Estudos da Tradução baseados em *corpora*

A Linguística de *Corpus* é um ramo da Linguística que se ocupa da

> coleta e da exploração de *corpora*, ou conjuntos de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador (BERBER SARDINHA, 2004, p. 3).

[5] "[colocation is] the occurrence of two or more words within a short space of each other in a text." (tradução minha).

O trabalho com *corpora* nos Estudos da Tradução tem suas bases em Baker (1993, 1995, 1996, 2004), que apresenta sugestões de pesquisa com *corpora* paralelos e comparáveis como forma de analisar a tradução em si, como uma variedade linguística, em vez de limitar sua análise à comparação entre texto original e texto traduzido.

De acordo com Baker, *corpus* é "um conjunto de textos eletrônicos, de fontes diversas, reunidos a partir de critérios e finalidades específicos, passível de receber tratamento automático ou semiautomático" (BAKER 1995 *apud* VIEIRA; JESUS, 2013, p. 2). Os *corpora* podem ser classificados em três tipos (BAKER, 1995): i) os *corpora* paralelos são compostos por textos originais e suas traduções; ii) os *corpora* multilíngues são um conjunto de textos originais monolíngues, agrupados por um critério comum; e iii) os *corpora* comparáveis são compostos por textos originais em uma dada língua e textos traduzidos para essa mesma língua.

Para caracterizar os textos usados em um *corpus* comparável, Baker utiliza os termos "traduzido" e "não traduzido", pois acredita que considerar o texto fonte da tradução como "original" implica desconsiderar outros tipos de textos como "originais" (OLOHAN, 2004). Neste trabalho, no entanto, usou-se tanto "não traduzido" quanto "original" para fazer referência a textos produzidos originalmente no idioma em questão.

Neste trabalho foram utilizados *corpora* paralelos e comparáveis, com ênfase no comparável, uma vez que o objetivo concentrou-se na comparação dos termos encontrados em inglês original e em inglês traduzido. A pesquisa com *corpora* comparáveis começou a ganhar força com o trabalho de Baker (1995, p. 235): "o acesso a *corpora* comparáveis permite-nos identificar padrões que são restritos a textos traduzidos ou que ocorrem com uma frequência significativamente maior ou menor em textos traduzidos do que em não traduzidos".[6]

A pesquisa com *corpora* comparáveis pode mostrar o que de fato acontece no processo de tradução e auxiliar a tarefa do tradutor, uma vez que não se concentra apenas na análise do texto original e de sua tradução.

O trabalho com *corpora* comparáveis permite olhar para o texto traduzido em si, não em comparação com o original, mas com o texto não traduzido. Esse tipo de análise pode revelar padrões específicos da linguagem usada na tradução e nos textos produzidos originalmente no idioma. Com o auxílio do *corpus* paralelo, a análise da tradução em contraste com seu original revela estratégias de tradução, que, por sua vez, podem ser comparadas com os dados de um

[6] "access to comparable *corpora* should allow us to capture patterns which are either restricted to translated text or which occur with a significantly higher or lower frequency in translated text than they do in originals" (tradução nossa).

corpus comparável para verificar diferenças e/ou semelhanças entre a escrita traduzida e a escrita não traduzida em determinada língua e conferir confiabilidade aos termos encontrados nas traduções.

Na Terminologia, o uso de *corpus* paralelo permite verificar as opções terminológicas usadas na tradução, o que pode ser vantajoso ou não. Por um lado, a busca por termos já traduzidos em *corpora* paralelos não garante a confiabilidade do uso, uma vez que o tradutor nem sempre é especialista na área e talvez não disponha de capacidade linguística para, se necessário, criar os termos na língua de chegada (AUBERT, 1996). Por outro lado, se há dificuldade em encontrar algum termo na língua de chegada, talvez haja pouco ou nenhum material produzido originalmente na língua sobre o assunto pesquisado, sendo necessário, às vezes, recorrer a textos traduzidos. Nesse caso, a pesquisa com *corpora* paralelos torna-se inevitável, apesar de requerer cautela e, se possível, uma consulta a especialistas da área para verificar a confiabilidade da fonte traduzida (AUBERT, 1996).

Neste projeto, a pesquisa com *corpora* comparáveis e paralelos visou identificar termos espíritas relacionados à mediunidade como meio de produzir um material terminológico confiável para tradutores e especialistas da área.

3. O Espiritismo traduzido no mundo

O Espiritismo (*Spiritism* ou *Spiritualism*, em inglês) é uma doutrina estudada e praticada em vários países do mundo.[7] Seus seguidores a consideram uma religião, uma filosofia e uma ciência, e acreditam na sobrevivência do espírito após a morte e na reencarnação. O Espiritismo também acredita na possibilidade de comunicação dos homens com os espíritos desencarnados por meio da mediunidade.

A mediunidade, como define Barbosa (2002, p. 118),

> é a fonte primordial dos ensinamentos da Doutrina [espírita], e suas tarefas constituem, hoje, sem dúvida, importante contribuição dos espíritas que a elas se dedicam, à consolidação da fé raciocinada e ao retorno, à normalidade, das condições psíquicas alteradas daqueles que, enleados nas tramas da obsessão disfarçada e tenaz, procuram, agoniados, os centros espíritas, ou são a eles encaminhados. A comunicação entre os dois mundos, o *corpora*l, material ou visível e o incorpóreo, imaterial ou invisível, é uma premissa básica do Espiritismo, que seria apenas um espiritualismo irreal e duvidoso, se a negasse ou a repudiasse.

[7] Ver número de traduções para diversas línguas realizadas pela FEB (Federação Espírita Brasileira). Se há tradução, é porque há demanda. Disponível em: <http://www.feblivraria.com.br/livros-em-outros-idiomas.html>. Acesso em: 27 jan. 2014.

Os primeiros estudos científicos sobre o fenômeno mediúnico, principalmente dentro da Psiquiatria, começaram entre os séculos XIX e XX, período em que Frederic Myers fundou a *Society for Psychical Research* em 1882 no Reino Unido (TYMN, 2011; ALMEIDA; LOTUFO NETO, 2004), e William Moses conduziu estudos sobre mediunidade na Inglaterra e atuou também como médium, produzindo as obras *Spirit Teachings*, em 1883, e *More Spirit Teachings*, em 1892 (TYMN, 2011).

Allan Kardec[8], o codificador do Espiritismo, também realizou suas pesquisas e observações acerca da mediunidade nesse período, publicando em 1861 o *Le livre des médiums (O livro dos médiuns, em português, e The mediums' book, em inglês)*, na França, com instruções sobre os tipos de mediunidade, os fenômenos mediúnicos e as reuniões mediúnicas.

O livro dos médiuns faz parte do pentateuco, que abarca os preceitos básicos da doutrina espírita, escrito por Kardec. Segundo o autor, a filosofia espírita foi ditada pelos espíritos, daí ser chamada de *Spiritisme* (Espiritismo) (TYMN, 2011). No entanto, o trabalho desenvolvido por ele na França assemelhava-se ao realizado por outros pesquisadores na Inglaterra e nos Estados Unidos, os quais usavam o termo *Spiritualism* (Espiritualismo) (TYMN, 2011).

Em inglês, o termo mais usado é *Spiritualism*, mas a confusão entre Espiritualismo[9] e Espiritismo[10] é inevitável. Tal diferença terminológica pode ser observada na tradução da obra *History of Spiritualism* (DOYLE, 1926) para o português, *História do Espiritismo* (DOYLE, 2008), sendo este o termo consagrado no Brasil, presente nas traduções das obras kardequianas (do francês para o português). Espiritismo também foi a opção usada para traduzir *The scientific basis of Spiritualism* (SARGENT, 1881), intitulado *Bases científicas do Espiritismo* (SARGENT, s.d.) em português.

Ao se realizar um levantamento rápido, feito por meio de buscas na Internet, com o termo *Spiritism*, foram encontradas várias obras espíritas brasileiras traduzidas para o inglês.

[8] Allan Kardec foi o pseudônimo adotado pelo professor Hippolyte Léon Denizard Rivail, nascido em 3 de outubro de 1804, em Lyon, na França (e desencarnou em 31 de março de 1869, na cidade de Paris) para realizar a tarefa, missionária, de codificar, isto é, apresentar em livros, metódica, didática e logicamente organizados, comentados e explicados, os postulados da Doutrina Espírita" (BARBOSA, 2002). É o autor das obras básicas do Espiritismo, a saber: *O livro dos espíritos* (1857), *O livro dos médiuns* (1861), *O Evangelho segundo o Espiritismo* (1864), *O céu e o inferno* (1865) e *A gênese* (1868) (BARBOSA, 2002).

[9] "Doutrina que consiste na afirmação da existência ou realidade substancial do espírito, e de sua autonomia, diferença e preponderância em relação à matéria; qualquer doutrina ocultista ou religiosa que acredita na existência de espíritos imateriais." (ESPIRITUALISMO, 2009)

[10] "Doutrina de cunho filosófico-religioso voltada para o aperfeiçoamento moral do homem por meio de ensinamentos transmitidos por espíritos desencarnados que se comunicam com os vivos especialmente através de médiuns." (ESPIRITISMO, 2009)

Entretanto, a busca em enciclopédias *on-line*, como a *Wikipedia*, mostra que o assunto aparece em inglês não traduzido sob o termo S*piritualism*. Ainda assim, grupos de brasileiros que moram em países de língua inglesa e criaram grupos de estudo espíritas buscam utilizar mais *Spiritism* para fazer uma associação direta ao trabalho de Kardec.

O Espiritismo de Kardec surgiu oficialmente na França, em 1857 (publicação de *Les livre des esprits* (KARDEC, 1860)), e chegou ao Brasil em 1866, quando Teles de Menezes traduziu para o português a *Introdução ao Estudo da Doutrina Espírita* da edição francesa de *O Livro dos Espíritos* (BAIO, s.d.). A doutrina espírita teve grande expansão no país, sendo que muitas obras de médiuns brasileiros se tornaram conhecidas mundialmente por meio das traduções feitas pela FEB (Federação Espírita Brasileira).

De acordo com Barbosa (2005), a publicação de textos religiosos traduzidos tem crescido no Brasil, o que aponta um campo interessante para pesquisa. O Espiritismo destaca-se como importante fatia nesse setor, principalmente com obras de médiuns conhecidos, como Chico Xavier, que publicou mais de 450 livros, com 50 milhões de exemplares vendidos (ARRAIS, 2010), e Divaldo P. Franco, que psicografou cerca de 250 livros e vendeu mais de 8 milhões de exemplares no Brasil e no mundo (DIVALDO..., 2012).

Como consequência, essa alta produção espírita brasileira gera a participação do país na produção de material espírita mundial, uma vez que essas obras têm sido traduzidas para diversas línguas. Algumas das obras de Divaldo Franco foram traduzidas para 17 idiomas (DIVALDO..., 2012), ao passo que as obras de Chico Xavier foram traduzidas para inglês, francês, alemão, espanhol, russo, húngaro, italiano, tcheco e grego (MATOS, s.d.). De acordo com o Conselho Espírita Internacional (CEI), que reúne cerca de 40 países,[11] o Brasil é o que apresenta maior número de adeptos. Por esse motivo, a tradução das obras espíritas brasileiras contribui para a difusão da doutrina em outros países.

Em consequência dessa expansão internacional, surgiu o questionamento sobre a terminologia utilizada nos textos espíritas traduzidos e não traduzidos em inglês, considerando-se o Espiritismo como uma linguagem de especialidade. Optou-se pelo tema da mediunidade em virtude das publicações encontradas em inglês não traduzido, produzidas no período de surgimento do Espiritismo, que são verdadeiros estudos científicos acerca dos fenômenos mediúnicos.

[11] PAÍSES membros do CEI. Disponível em: <http://cei.spirite.org/pt/about/member-countries-of-the-isc/>. Acesso em: 2 fev. 2014.

Embora o foco deste trabalho tenha sido a descrição de uma metodologia de pesquisa terminológica com *corpora* paralelo e comparável e a construção de um glossário bilíngue sobre mediunidade, também se propôs a alertar o tradutor para questões terminológicas e culturais que ocorrem nos textos espíritas. A seguir serão apresentados os passos percorridos nesta pesquisa para selecionar os termos analisados e elaborar o glossário bilíngue.

4. Procedimentos metodológicos

A metodologia deste trabalho pode ser dividida em duas grandes etapas: i) a compilação e o processamento do *corpus* e ii) a elaboração das fichas terminológicas e da amostra de glossário. Essas etapas serão resumidas a seguir.

4.1 Compilação e processamento do *corpus*

A primeira etapa englobou vários passos, desde a seleção e a compilação dos *corpora*, incluindo todo o processo de limpeza e formatação dos textos, o alinhamento do *corpus* paralelo e a extração de dados dos *corpora* paralelo e comparável, com a utilização do programa *AntConc*, até a seleção dos termos, a busca dos equivalentes e a análise comparativa dos termos em inglês traduzido (ET) e inglês original (EO), também com o apoio do *AntConc*.

4.2 O *corpus*

O *corpus* coletado para esta pesquisa foi composto por textos espíritas sobre o tema mediunidade. São textos publicados nos primeiros 150 anos do Espiritismo – de 1850 a 2000, considerando-se o surgimento dessa religião na França (em 1857), nos Estados Unidos (a partir de 1848) e sua proliferação no Brasil (a partir de 1866).

O *corpus* da pesquisa foi formado por três sub*corpora*: PO, ET e EO. As obras em inglês original (EO) referem-se às primeiras pesquisas científicas sobre mediunidade nos Estados Unidos e na Inglaterra e usam o termo *Spiritualism* para se referir ao conjunto de ideias e princípios filosóficos e científicos (sobretudo no que se refere à mediunidade) que mais tarde, na França, Allan Kardec chamaria de Espiritismo. As obras em inglês traduzido (ET) são mais recentes, sendo a maioria publicada na primeira década do século XXI, mas os originais em

português (PO), publicados no Brasil, são da segunda metade do século XX. Os textos estão apresentados no Quadro 1 (com o número de *tokens*[12]).

Os *corpora* PO e ET são parcialmente paralelos. Conforme o Quadro 1, os textos PO5 e ET5 não são paralelos. Eles foram inseridos para balancear o tamanho dos *corpora*, para que tivessem dimensões semelhantes, facilitando a comparação da frequência dos termos entre ET e EO.

Na compilação dos *corpora*, utilizaram-se livros inteiros, bem como extratos de alguns livros. Além disso, para atender ao critério da dimensão[13], alguns textos tiveram trechos removidos, principalmente prefácios, sumários e índices; quando necessário, foram removidos também alguns capítulos de livros, buscando manter os que tratavam especificamente sobre mediunidade.

Os textos paralelos foram alinhados para facilitar a identificação das opções de tradução para os termos analisados. O alinhamento foi realizado com o programa *Plus tools*[14].

[12] Número total de palavras do texto (BERBER SARDINHA, 2004).
[13] A metodologia da Linguística de Corpus recomenda que os *corpora* sejam do mesmo tamanho para sua comparabilidade.
[14] O programa pode ser obtido no site: <http://www.wordfast.net/index.php?whichpage=plustools&lang=ar01>. Acesso em: 26 ago. 2012. Mas é necessário baixar o *Wordfast* (programa de memória de tradução, gratuito para testar), que pode ser obtido em <http://www.wordfast.net/index.php?whichpage=plustools&lang=ar01>. Acesso em: 26 ago. 2012.

Quadro 1. Dados do *corpus* de estudo.

Corpora	**Título**	**Sigla**	**Autor/Tradutor**	**1ª edição**	**Edição utilizada**	**Tamanho dos *corpora***
Inglês original (EO)	*A guide to mediumship and psychical unfoldment*	EO1	Edward Walter Wallis e Minnie Harriot Wallis	1903	1903	143.046
	Life after death: problems of the future life and its nature	EO2	James H. Hyslop	1918	1918	
	Psychics and mediums: a manual and bibliography for students	EO3	Gertrude Ogden Tubby	1935	1935	
	Psychography: a treatise on one of the objective forms of psychic or spiritual phenomena	EO4	William Stainton Moses	1878	1878	
Inglês traduzido (ET)	*Disobssession: preparation for a counseling meeting*	ET1	Tânia Stevanin, Jussara Korngold e Marie Levinson	2003	2003	141.887
	In the realms of mediumship: life in the spirit world	ET2	Darrel W. Kimble e Ily Reis	2011	2011	
	Obsession, passes, counselling	ET3	Jussara Korngold e Marie Levinson	2004	2004	
	We are all mediums	ET4	Públio Lêntulus Vicente Coelho	2001	2001	
	The mediums' book	ET5	Anna Blackwell	1986	1986	
Português original (PO)	*Desobsessão*	PO1	Francisco Cândido Xavier e Waldo Vieira (André Luiz)	1964	1964	120.703
	Nos domínios da mediunidade	PO2	Francisco Cândido Xavier (André Luiz)	1955	2000	
	Obsessão, passe e doutrinação	PO3	J. Herculano Pires	1979	2008	
	Somos todos médiuns	PO4	Carlos A. Baccelli (Odilon Fernandes)	1993	1993	
	Médiuns e mediunidades	PO5	Cairbar Schutel	1923	1923	

4.3 Seleção dos termos

A seleção dos termos para análise e elaboração das fichas terminológicas ocorreu pela criação da lista de palavras-chave (*Keyword List*) do *corpus* de português original (PO). Partiu-se desse *corpus* porque foi com base nele que as traduções para os termos foram identificadas no ET e, em seguida, comparadas com os termos usados em inglês original (EO).

Para gerar a *Keyword List* foi necessário usar um *corpus* de referência, que é um *corpus* de língua geral utilizado para contrastar com o *corpus* de especialidade, de forma que se espera que as palavras mais frequentes da *Keyword List* sejam as da linguagem de especialidade. Segundo Berber Sardinha (2004), o *corpus* de referência deve ser cerca de cinco vezes maior que o *corpus* de estudo. Para este trabalho, utilizou-se um *corpus* de referência (em português original)[1] formado por textos jornalísticos, acadêmicos, literários, turísticos e textos da *web*, composto por 746.054 *tokens* (sete vezes maior que o PO, que tem 120.703 *tokens*).

Após carregar os *corpora* no *AntConc*, gerou-se a *Word List* com *stoplist* e, em seguida a *Keyword List* do PO. Das palavras-chave listadas, foram selecionados os termos que comporiam as fichas terminológicas e o glossário bilíngue, partindo-se dos colocados. Optou-se por escolher substantivos que fossem específicos da área da mediunidade; então, com base nas primeiras cinco palavras-chave, foram selecionados três termos para análise: *médium*, *mediunidade* e *espírito*. Em seguida, foram analisados os colocados de cada um deles, ou seja, selecionaram-se termos multivocabulares, os quais compuseram o glossário produzida nesta pesquisa.

A busca por termos multivocabulares vai ao encontro da proposta de Aubert (1996, p. 64), que afirma que a seleção de termos multivocabulares é uma forma de aumentar a especificidade da terminologia pesquisada. Sendo assim, esta pesquisa buscou as coocorrências das palavras-chave analisadas para montar um glossário mais específico do campo da mediunidade.

4.4 Identificação de termos multivocabulares

Com o objetivo de analisar termos mais específicos da área da mediunidade, buscou-se identificar colocados para as palavras selecionadas para análise na *Keyword List*. Para isso,

[1] Esse *corpus* de referência foi compilado a partir dos *corpora* Cordiall e Klapt!, detalhados em Jesus (2008) e Jesus e Nunes (2014).

verificaram-se as coocorrências das palavras-chave na aba *Clusters* do *AntConc*, a qual apresenta as palavras que ocorrem com mais frequência com o termo escolhido.

Para buscar os termos multivocabulares neste trabalho, optou-se por manter o tamanho do *cluster* em dois itens, esperando encontrar formações do tipo substantivo+adjetivo/ adjetivo+substantivo, mas aumentou-se a frequência mínima para três, conforme proposto por Jones e Sinclair (1974 *apud* DAYRELL, 2005), para considerar a coocorrência significativa para a pesquisa. Esse procedimento foi aplicado aos três termos – *médium*, *mediunidade* e *espírito* – considerando-se os termos no singular/plural e também as formas em caixa alta/baixa (usando a função *Treat all data as lowercase*).

As colocações encontradas e consideradas como termos foram anotadas nas fichas terminológicas no campo Colocados, e depois foram inseridas no glossário. A busca pelas colocações foi realizada nos três sub*corpora* individualmente, com vistas a identificar os termos equivalentes para a amostra de glossário.

4.5 Busca pelos equivalentes

Após a seleção dos termos para as fichas e dos termos multivocabulares para o glossário, passou-se para a busca dos equivalentes utilizados no ET e no EO.

No caso do inglês traduzido (ET), utilizou-se o *Concordance* para analisar o *corpus* alinhado PO-ET e identificar como os termos do PO foram traduzidos no ET. Com os arquivos salvos como memória de tradução, buscou-se o termo no *Concordance*, que mostra todas as linhas em que o termo ocorre. Para verificar a opção de tradução dada, basta clicar no termo, que ele automaticamente aparece destacado no texto na aba *File View*. O pesquisador, então, localiza o equivalente no segmento alinhado.

Essa etapa possibilitou não só encontrar a tradução dos termos, mas também identificar seus sinônimos, quando havia mais de uma opção de tradução para o termo do PO. Como consequência desse rol maior de opções, a busca no EO foi mais direcionada, sendo feita com base nos equivalentes encontrados em ET.

Outra forma utilizada para encontrar os equivalentes foi a função *Sort*, que marca os coocorrentes do termo pesquisado e organiza as linhas de concordância em ordem alfabética, facilitando a identificação dos termos no EO, uma vez que a busca nesse *corpus* guiou-se, inicialmente, pelos termos encontrados no ET, que nem sempre ocorreram em EO.

Como exemplo desse uso, pode-se citar o caso de *médium iniciante*, que apareceu 8 vezes no PO, e seu sinônimo, *médium principiante*, que ocorreu 11 vezes. Esse termo foi traduzido no ET como *novice medium* (16 ocorrências) e *beginner* (21). No EO, *beginner* apareceu 3 vezes, mas *novice medium* não ocorreu. Então, partiu-se para a análise das linhas de concordância de *medium* com a ferramenta *Sort*, chegando-se ao termo *young medium*.

Buscou-se o termo *medium* e selecionou-se como *Level 1* – o nível que determina a ordem alfabética das ocorrências – 1L, ou seja, a primeira palavra à esquerda de *medium*, como forma de identificar um adjetivo que coocorresse com o termo e que significasse algo semelhante a *médium iniciante*. Nessa busca, foi identificada uma estrutura adjetivo+substantivo para o termo, *young medium*. O mesmo procedimento foi feito com *psychic*, sinônimo de *medium* no EO, para encontrar outros possíveis termos, encontrando-se *psychic student*, com 2 ocorrências.

Nas buscas pelos equivalentes no EO, verificou-se uma variação na classe em que alguns termos apareceram (como adjetivo, substantivo ou verbo) e no sentido que apresentaram no texto. Um dos sinônimos de *medium* encontrados no EO foi *psychic*. Esse termo ocorreu 331 vezes, mas apareceu tanto como substantivo, *the psychic*, quanto como adjetivo, *psychic force*. Além disso, o termo apresentou duas acepções: *psíquico*, "relativo ao que ocorre na esfera mental do indivíduo"[2], e *mediúnico*, relativo à capacidade de comunicação do médium com os espíritos.

Para contabilizar o uso de *psychic* como substantivo e adjetivo no EO, e também as ocorrências nas duas acepções, seria necessário etiquetar[3] o *corpus*, ou seja, analisar todas as 331 linhas de concordância, identificando a classe gramatical e o significado, o que demandaria mais tempo para a pesquisa. Portanto, na análise dos dados, os casos como o de *psychic* foram marcados com um * após a frequência do termo, como forma de indicar que o valor é bruto, ou seja, o termo apresentou mais de um significado e/ou mais de uma classe gramatical no *corpus*, e, nesse caso, as linhas de concordância não foram analisadas individualmente para separar as ocorrências. Todas as informações encontradas no *corpus* sobre cada termo foram anotadas nas fichas terminológicas, cuja organização está detalhada a seguir.

[2] PSÍQUICO. In: DICIONÁRIO Aulete. Disponível em <http://www.aulete.com.br/ps%C3%ADquico>. Acesso em: 26 jul. 2014.
[3] Segundo Berber Sardinha (2004), a etiquetagem é um procedimento que insere códigos no *corpus* que indicam a classe gramatical das palavras. Dessa forma, seria possível separar as ocorrências de *psychic* como adjetivo e como substantivo.

4.6 Elaboração das fichas terminológicas

Como foi dito, foram selecionados três termos para a elaboração de fichas terminológicas – *médium, mediunidade* e *espírito* – e o glossário foi feito com os colocados desses três termos. Selecionaram-se, então, alguns termos multivocabulares para discussão e análise.

4.6.1 As fichas terminológicas

Conforme aponta Aubert (1996), as fichas terminológicas servem para uniformizar os procedimentos de análise e registro das informações coletadas na pesquisa terminológica. Neste estudo, as fichas foram fundamentais para organizar os dados e anotar as observações feitas durante a análise. Foram feitas três fichas terminológicas, com os termos *médium*, *mediunidade* e *espírito comunicante*.

A ficha utilizada nesta pesquisa foi baseada no modelo adotado na Iniciação Científica realizada por uma das autoras, cujos campos seguiram a proposta de Costa Filho (2008). A ficha foi dividida em três colunas, PO, ET e EO, e cada coluna continha nove campos, que foram preenchidos com as informações referentes ao termo de cada *corpus* separadamente. Os campos foram os seguintes: Termo, Morfossintaxe, Definição, Contexto, Variantes, Ocorrências, Colocados, Sinônimos e Notas/comentários.

Em todas as fichas, o campo Termo da coluna do PO foi preenchido com os termos selecionados na lista de palavras-chave, isto é, *médium*, *mediunidade* e *espírito comunicante*. No caso da ficha de *espírito comunicante*, optou-se por usar o termo multivocabular em vez do mono, *espírito*, como forma de trabalhar com um termo mais específico da mediunidade. Dentre os colocados de *espíritos*, identificou-se *comunicante* como mais adequado, visto que representa uma das partes da comunicação mediúnica. No ET, o campo Termo foi preenchido com a tradução mais recorrente dada para o termo. No EO, colocou-se a opção mais frequente encontrada para o termo com base na análise da ficha.

Durante o período de análise, outras informações e observações foram anotadas no campo Notas/comentários, as quais, posteriormente, compuseram a seção de análise dos dados desta pesquisa. Foram anotadas, principalmente, questões relativas à ocorrência e à não ocorrência de termos no ET e no EO, as quais foram utilizadas para mostrar as diferenças entre inglês original e inglês traduzido, além de auxiliar na busca pelos equivalentes de cada termo

do PO em inglês original. A organização das informações nesse modelo de ficha mostrou-se bastante eficiente e auxiliou na sistematização dos dados encontrados.

5. Análise de termos espíritas em *corpora* paralelos e comparáveis

A pesquisa completa analisou todas as ocorrências e coocorrências dos termos escolhidos para compor as fichas terminológicas e o glossário deste trabalho: *médium*, *mediunidade* e *espírito*. Conforme explicitado na metodologia, partiu-se desses três itens retirados da lista de palavras-chave, para os quais foram investigados os colocados e seus equivalentes em inglês. O foco do estudo é a análise comparativa entre os termos utilizados nas traduções (*corpus* de inglês traduzido, ET) e os termos utilizados nos textos originais em inglês (EO).

A lista de palavras-chave foi criada a partir do *corpus* de textos em português original (PO), e, dentre as cinco primeiras palavras, foram escolhidos três termos e seus colocados para a análise, os quais foram colocados no glossário. A seguir, será apresentada uma análise mais detalhada de um desses termos e de alguns de seus colocados.

5.1 Médium

O termo *médium* foi a primeira palavra da lista de palavras-chave extraída do *corpus* PO. Ocorreu 901[4] vezes no PO, e seu equivalente no inglês, *medium*, ocorreu 1.561 vezes no ET e 598 no EO. Nota-se a alta frequência do termo nas traduções (ET), o que pode ser justificado pela presença do ET5, que não é paralelo, como dito na metodologia, e uma frequência menor de *medium* no EO, o que talvez possa ser explicado pelo uso de sinônimos no *corpus*, como será discutido nesta seção.

A definição encontrada para o termo foi:

> No sentido expresso da palavra, **médium** quer dizer intermediário, agente, instrumento. [...] Da mesma forma que a Física, a Química, a Botânica, a Astronomia têm os seus aparelhos apropriados, segundo a necessidade dos seus estudos, o Espiritismo tem um aparelho, um instrumento, o **médium**, com o qual estuda a alma e suas manifestações. É com este auxiliar indispensável

[4] A busca pelos termos e por suas frequências ocorreu pelas formas singular e plural, sendo que os números de ocorrências apresentados no texto e nas fichas terminológicas equivalem à soma das ocorrências do termo nas duas formas. Como dito na metodologia, essa busca foi feita pelo uso de asterisco no final da palavra singular, ex.: *espírito**, em que a ferramenta buscou todas as terminações da palavra dada, ou pela busca individual dos termos que têm formação plural irregular, como *médium* e *médiuns*.

> que penetra no labirinto da Psicologia e da Parapsicologia para a descoberta do Novo Mundo, e o estreitamento de relações com os seus habitantes. (PO1).

A tradução para *médium* encontrada no ET foi *medium*. Já no EO, apesar de *medium* apresentar mais ocorrências, encontrou-se também o uso de *psychic* e *sensitive* como sinônimos de *medium*. Os exemplos a seguir ilustram essas ocorrências:

> Dificilmente o **médium** precisará com nitidez quando estará sendo intuído ou inspirado a dizer palavras ou tomar atitudes que mudem o rumo dos acontecimentos dos quais participe. (PO4)

> On rare occasions, a **medium** will be able to clearly know whether he has received intuition or has been inspired, to say words, or take action,that could change the way of things in which he takes part. (ET4)

> There is a "more excellent way" of approaching the people of the unseen realm whereby good, not evil, accrues to both **sensitive** and spirit. The co-operative association of **medium** and spirit on the plane of thought and purpose, emotion and motive, ethics and inspiration, results in the education and elevation of the **sensitive**, and the increase of the knowledge of the operator as to the conditions on this side. (EO1)

> As distance on the spirit side is more a matter of state than geography, the **psychic** must strive to attain a higher degree of lucidity to get away from the plane of haunting, vicious, earthly or vindicative spirit people by rising above it, so as to be unaffected by those denser vibrations and respond to the more subtle and spiritual forces. (EO1)

Nota-se que *psychic* teve 331*[5] ocorrências em EO e 46 em ET, ao passo que *sensitive* teve 141* ocorrências em EO e 12 em ET. O termo *psychic* foi bastante utilizado no EO (331* ocorrências) e apareceu tanto como substantivo, como apresentado acima, quanto como adjetivo. O termo parece ser usado principalmente como adjetivo, para caracterizar *force, nature, mind, energy, perception, condition, power, faculty*, entre outros elementos envolvidos na comunicação com os espíritos, como mostra este exemplo:

> The failure to obtain results under such impossible conditions is a proof of the genuine **psychic nature** of the powers of the mediums. If they were pretenders they would succeed in doing something under any circumstances and in spite of such adverse **psychic conditions**. (EO1).

[5] Cabe relembrar aqui que o asterisco depois da frequência indica que o valor apresentado é bruto, ou seja, as diferentes acepções e classes gramaticais não foram separadas em análise individual das linhas de concordância, visto que esse não é o foco da pesquisa.

Destaca-se, entretanto, que *psychic*, como adjetivo, também tem o significado de *psíquico*, ou seja, faz referência ao que ocorre na esfera mental do médium, independentemente da influência dos espíritos. Como sinônimo de *medium*, ou seja, como substantivo, a frequência de *psychic* pareceu ser baixa no *corpus* em inglês original (EO).

No ET, não houve ocorrência de *psychic* como substantivo, o que foi possível verificar devido a sua baixa ocorrência nesse sub*corpus*. Mas, o termo aparece como adjetivo, caracterizando a influência psíquica ou mental do médium na comunicação. O trecho a seguir ilustra essa afirmação:

> **Psychic** Process or Animism: The influence of the medium is really of great importance, particularly in the involuntary substitution of his own ideas for those which the communicating spirits endeavor to suggest. It is also important in the formulation of baseless and fantastic theories, in accordance with his own opinions or prejudices, whether as a product of his own mind, or derived from the suggestions of ignorant or mocking spirits. (Nota do tradutor, ET4).

O adjetivo *psychical* também ocorreu em EO, com 33 ocorrências. Novamente, uma análise de cada ocorrência seria necessária para verificar o sentido desse adjetivo, que se refere tanto a *mediúnico* quanto a *psíquico*. *Psychical* não apareceu no ET.

Após a identificação do termo *médium* e de suas equivalências no ET e no EO, buscaram-se as colocações com o termo por meio da ferramenta *Clusters*. Conforme explicado na metodologia, pela limitação de tempo, foram selecionados três colocados para a análise. Para o termo *médium*, discutiremos os seguintes termos multivocabulares: *médium passista*, *médium psicógrafo* e *médium iniciante*. Os outros termos não foram apresentados na análise, mas aparecem no glossário.

Os termos discutidos na análise são os que apresentaram maior divergência no contraste entre ET e EO, ou seja, os casos que apontam para o uso de opções literais de tradução em ET, e que tendem a não ocorrer nos textos em inglês original (EO).

5.1.1 Médium passista

Tabela 1. Frequências do termo *médium passista*, seus sinônimos e equivalentes em ET e EO.

PO	ET	EO
médium passista – 15 **médium curador – 11**	pass giver (pass-giver) – 16 healing medium – 4 healer – 4	pass giver (pass-giver) – 0 healing medium – 2 healer – 9

O termo *médium passista* apareceu 15 vezes no *corpus* PO, e o termo sinônimo, *médium curador*, ocorreu 11 vezes. Por serem considerados sinônimos, os termos foram colocados na mesma entrada no glossário.

Apesar de *médium curador* aparecer apenas 1 vez no *corpus* alinhado (PO-ET), foi possível verificar que ele é sinônimo de *médium passista* por meio do termo *mediunidade curadora*, que foi traduzido como *healing mediumship*, mesma tradução de *mediunidade passista*, como mostram os exemplos a seguir:

> Nem os mais estudiosos seriam capazes de saber dos verdadeiros prodígios ocultos efetuados pela **mediunidade curadora**. (PO4)
>
> Not even the most literate of people would be able to comprehend the truly hidden prodigious results produced by the **healing mediumship**. (ET4)
>
> A **mediunidade passista** é, ainda, a força mantenedora do equilíbrio e da paz de quantos médiuns se sintam, por este ou aquele motivo, impedidos de abraçar uma mediunidade mais ostensiva, que lhes exija compromissos mais regulares. (PO4)
>
> **Healing mediumship** is the supporting power which brings balance and peace to those who, for any reason, are restrained in developing a more arduous mediumship, seeing that it would demand from them more serious commitments. (ET4)

A única ocorrência de *médium curador* no *corpus* alinhado (PO-ET) não apresenta o adjetivo na tradução para o inglês, podendo-se considerar que *healing* é omitido ou fica implícito.

> A primeira pessoa que a mediunidade curadora cura é o **médium curador**. (PO4)
>
> The first person who reaches the benefit of cure is **the own medium**. (ET4)

O termo encontrado para *médium passista/curador* no ET foi *healing medium*, com 4 ocorrências, ou simplesmente *healer*, com 4 ocorrências. *Médium passista* também foi traduzido como *pass giver*, que ocorreu 16 vezes no ET, e levantou outro questionamento: o uso de *pass* como equivalente de *passe*, uma vez que *pass giver* não aparece no EO. No *corpus* alinhado (PO-ET), *passe* aparece traduzido como *pass*/*passes*, além de *magnetic pass*, *healing* e *laying on of hands*.

> O **passe** é simplesmente a imposição das mãos, ensinada por Jesus e praticada por Ele. É uma doação humilde e não uma encenação, dança ou ginástica. (PO3)
>
> The **pass** is simply the imposition of hands as taught by Jesus and practiced by Him. It is a humble donation and not a motive for drama, dances or gymnastics. (ET3).

No EO, o termo usado foi *passes* também, mas sempre no plural. Também foram encontrados os termos *healing*, *magnetic healing* e *laying on of hands* no EO.

O termo encontrado para *médium curador/passista* no EO também foi *healing medium*, com 2 ocorrências. Contudo, o termo *healer* foi mais usado, com 9 ocorrências, das quais 7 fazem referência ao *médium passista*, e as outras 2 referem-se aos espíritos (*spirit healer*). Os trechos abaixo ilustram o uso de *médium passista* nos três sub*corpora*:

> O **médium passista**, igualmente, para estar em sintonia com os Benfeitores Espirituais que operam por seu intermédio, não carece de qualquer tipo de agitação no ato do passe. (PO4)
>
> Equally, the **healing medium** in order to be tuned with the Spiritual Benefactors, does not need any kind of agitation when giving healing. (ET4)
>
> Every very successful **healer**, whether Spiritualist or "Christian Scientist," has his own mental and spiritual power reinforced by a "spirit band," or battery, if we may so speak. (EO1)

Com base nessa análise, foi possível destacar algumas variações no uso dos termos em ET e EO. Embora o uso de *healing medium* seja comum, há uma preferência no EO pelo uso de *healer*, ao passo que em ET a frequência de *pass giver* é maior, sendo que esse termo não aparece em EO. Observa-se ainda que o termo *pass* apresentou uma variação de número, pois aparece no singular e no plural no ET, enquanto no EO apenas a forma plural é usada, *passes*. A seguir, será apresentada a análise do termo *médium psicógrafo*.

5.1.2 Médium psicógrafo

Tabela 2. Frequências do termo *médium psicógrafo*, seus sinônimos e equivalentes em ET e EO

PO	ET	EO
médium psicógrafo – 8 **médium escrevente – 5**	psychography medium – 3 automatic writing medium – 1 mediums with the gift of psychography – 1 mediums of psychography – 1 writing medium – 13	automatic writer – 3 psychographic medium – 1 slate-writing medium – 1 writing medium – 0

O termo *médium psicógrafo* teve 8 ocorrências e *médium escrevente* teve 5. Os termos foram considerados sinônimos dado o esclarecimento encontrado neste trecho: "As formas mais comuns de mediunidade intelectual ou espiritual, são a falante e a **escrevente**. [...] Psicofonia e **psicografia**, respectivamente" (PO5). Por esse motivo, os termos foram colocados na mesma entrada no glossário.

Os equivalentes encontrados para os termos no ET foram *psychography medium* (3 ocorrências), *automatic writing medium* (1), *mediums with the gift of psychography* (1) e *mediums of psychography* (1). Foram encontradas também 13 ocorrências do termo *writing medium* no ET, mas todas estavam no texto não alinhado, o ET5. O trecho abaixo mostra o contexto dos termos utilizados no PO e no ET:

> Se, porventura, apesar de sua dedicação e empenho, o **médium psicógrafo** notar, depois de alguns anos de exercício, que a sua mediunidade não apresenta a evolução desejada, não há nada demais que ele abdique da psicografia e canalize os seus recursos medianímicos para, por exemplo, a transmissão de passes, para a oratória, enfim, para uma tarefa que não lhe tome tanto tempo improficuamente. (PO4)
>
> If, perchance, through his dedication and effort, the **psychography medium** realises after years of exercise that his mediumship does not show the wished improvement, then there is nothing wrong in giving up this kind of mediumship and directing his mediumistic resources to another task. Such as healing or speech and that brings him more positive results. (ET4)

Nenhum desses termos ocorre em EO, apesar de os termos *psychography* (18 ocorrências) e *automatic writing* (12) serem utilizados para se referir a *psicografia/mediunidade escrevente*, assim como no ET:

> Sem dúvida, embora não possa ser considerada mediunidade especial, aliás, como

> nenhuma outra o pode, pela sua própria natureza, a **mediunidade psicográfica** pode beneficiar um grande número de pessoas, através dos esclarecimentos que presta, permitindo que os comunicados fixados no papel sejam analisados e meditados. (PO4)
>
> Undoubtedly, **psychography**, or **automatic writing**, as well as the other mediumistic faculties cannot be considered special. It can benefit a large number of people, as the communications written on a paper can be analysed and meditated upon. (ET4)
>
> This subject of **Psychography**, or writing without the intervention of ordinary human agency, is by no means new, though it has of late attracted greater attention. It has been familiar to all investigators of Psychic Phenomena, and has been called variously Direct or Independent Writing. (EO4)

No EO, o termo mais usado para *médium escrevente/psicógrafo* foi *automatic writer*, (3 ocorrências), seguido de *psychographic medium* (1). Apareceu também uma vez o termo *slate-writing medium*, em que *slate* refere-se à pedra sobre a qual aconteciam os fenômenos de psicografia, que funcionava como um quadro negro, como explica esse exemplo:

> The side of the **slate** that was being written upon was pressed by us against the table. Our second hands were linked together, and lay upon the table. While this position was preserved, the writing proceeded without pause. (EO4).

No ET não houve ocorrência de *slate* ou variantes.

Com base nas análises feitas, observou-se que os termos utilizados em ET e EO não são idênticos, apresentando pequenas variações na forma – *automatic writing medium* x *automatic writer* e *psychography medium* x *psychographic medium. Psychography* e *automatic writing* não foram utilizados no EO na função de adjetivo, modificando *medium*, conforme empregado no ET. A seguir, será apresentada análise do termo *médium iniciante*.

5.1.3 Médium iniciante

Tabela 3. Frequências do termo *médium iniciante*, seus sinônimos e equivalentes em ET e EO

PO	ET	EO
médium iniciante – 8 **médium principiante – 11**	novice medium– 16 beginner – 21* young medium – 0 novice – 1 novice in mediumship – 0	novice medium – 0 beginner – 3 young medium – 7 novice – 1 novice in mediumship – 1 psychic student – 2

No PO, o termo *médium iniciante* ocorreu 8 vezes, e *médium principiante*, 11. Dessas 19 ocorrências, 16 foram traduzidas por *novice medium* e 3 por *beginner,* como mostram os exemplos retirados do *corpus* alinhado PO-ET:

> Todo **médium principiante** carece exercitar-se, e isto porque nenhum médium nasce pronto. (PO4)
>
> Every **novice medium** needs to exercise himself, for no medium is born ready to work. (ET4)
>
> Segundo Kardec, o **médium iniciante** deve considerar-se feliz por manter intercâmbio com os espíritos considerados inferiores, e não com os levianos. (PO4)
>
> According to Kardec, a **novice medium** should consider himself fortunate in having contact with inferior spirits, and not frivolous ones. (ET4)
>
> "O escolho da maioria dos **médiuns iniciantes** é ter relações com espíritos inferiores, e devem se considerar felizes quando não o sejam senão espíritos levianos." (Cap. XVII, Segunda Parte, Item 211) (PO4)
>
> "The great stumbling-block of the majority of **beginners** is, in fact, their liability to be drawn in to hold conversation with inferior spirits. They may usually consider themselves fortunate, if they only come into contact with spirits who are merely frivolous, and not positively wicked." ("The Mediums' Book", Chapter 17, second part, item 211). (ET4)

Considerando o ET como um todo (não apenas a parte do *corpus* que foi alinhada), *beginner* ocorreu 21* vezes, e *novice medium*, 16.

Beginner também apareceu no EO, com 3 ocorrências, mas o termo mais frequente para *médium principiante/iniciante* foi *young medium*, com 7 ocorrências; também foi encontrado *psychic student*, com 2 ocorrências. *Novice medium* não ocorreu no EO, mas apareceram *novice* (1) e *novice in mediumship* (1). *Novice* tem uma ocorrência no ET.

Novamente, observaram-se diferenças no uso dos termos em ET e EO. Os termos usados em ET foram *novice medium* e *beginner*, sendo que o primeiro não ocorreu em EO, e o segundo teve baixa frequência. Em EO, o termo mais frequente foi *young medium*.

6. Palavras finais

A pesquisa terminológica apresentada aqui mostrou resultados satisfatórios na medida em que comprovou a eficiência da metodologia de pesquisa com *corpora* comparável e paralelo

para elaboração de glossários bilíngues, bem como contribuiu para a solução das perguntas de pesquisa levantadas na introdução: há variação na terminologia espírita encontrada nos textos traduzidos e não traduzidos? As traduções utilizam os mesmos termos usados em inglês original, aproximando-se da convencionalidade e da idiomaticidade do inglês, ou fazem uso de opções mais literais, mais próximas dos termos utilizados em português original?

Com base nos dados organizados no glossário, foi possível observar que a maioria dos termos em ET foi total ou parcialmente diferente dos encontrados no EO. Além disso, ao comparar os termos do ET com os termos em PO, percebeu-se que há uma semelhança morfológica, indicando a tradução literal dos termos.

A diferença no uso dos termos espíritas no inglês original (EO) e traduzido (ET) ilustra a fala de Tagnin (2005) sobre o "tradutor ingênuo", que não é capaz de identificar expressões fixas da língua, "fugindo" da convencionalidade. Conforme dito no início deste trabalho, buscou-se verificar se os termos usados em inglês traduzido foram usados em inglês original na área da mediunidade, com vistas a verificar se os termos seriam traduzidos literalmente, ou seja, se a opção dos tradutores estaria mais próxima do texto original. Essa opção, de certa forma, pode ser caracterizada como ingênua, pois não valida o uso dos termos em inglês não traduzido.

Esse tipo de tradução literal, no entanto, não torna o texto inadequado, mas evidencia uma marca do texto/da língua de partida que pode gerar dois efeitos: (a) causar estranhamento no leitor da tradução, que pode identificar o termo, mas este não lhe soar natural; ou (b) modificar, de forma gradual, o vocabulário da língua de chegada, de modo que esta se adapte aos novos termos, como ocorre com frequência na língua portuguesa, como os empréstimos da língua inglesa.[6]

Outro ponto importante nessa análise é o fato de que as obras espíritas brasileiras são com frequência traduzidas para o inglês por brasileiros, o que poderia justificar o uso de opções mais literais, gerando as questões destacadas acima. É possível cogitar que a tradução desses textos por nativos da língua inglesa poderia conferir um aspecto mais natural e idiomático às obras.

Contudo, para verificar o impacto das traduções na língua de chegada, seria necessário um estudo de recepção que colocasse falantes de língua inglesa diante dos textos traduzidos e

[6] Para mais informações sobre os procedimentos técnicos de tradução, ver Barbosa (1990).

verificasse o nível de estranhamento identificado nas obras, bem como proporcionasse a validação dos termos encontrados nesta pesquisa. Essa, portanto, é uma sugestão para trabalhos futuros, visto que o estudo de recepção dos termos analisados não faz parte dos objetivos desta pesquisa.

Além disso, para propor um glossário mais abrangente e com opções de tradução mais específicas, seria necessário aumentar o *corpus* de inglês não traduzido, possibilitando a análise de um número maior de ocorrências dos termos e dos equivalentes. Na pesquisa terminológica bilíngue,

> para assegurar uma razoável cobertura das noções e termos levantados a partir do córpus na primeira língua, convém, na segunda língua, aumentar em duas vezes o volume [...]. Mesmo com esta providência, porém, a experiência demonstra que a cobertura obtida na segunda língua corresponderá, em média, a 80% das noções e termos coligidos na primeira língua, a cobertura dos termos remanescentes tendo, geralmente, de ser efetuada por meio de consulta aos especialistas de área (AUBERT, 1996, p. 60).

O objetivo deste trabalho foi apresentar a metodologia de pesquisa terminológica com *corpora* comparáveis e paralelos, para contrastar o inglês traduzido com o inglês não traduzido e verificar se a terminologia mediúnica usada nos textos seria semelhante, ou se, conforme o questionamento inicial, a tradução apresentaria opções mais literais e próximas do original. Como resultado da metodologia, elaborou-se um glossário, que pode ser ampliado e adequado em pesquisas futuras para uso do tradutor.

Referências

ALMEIDA, G. M. de B.; CORREIA, M. Terminologia e *corpus*: relações, métodos e recursos. In: TAGNIN, S. E. O.; VALE, O. A. (Org.). **Avanços da Linguística de *Corpus* no Brasil**. São Paulo: Humanitas, 2008. p. 67-94.

ALVES, F.; TAGNIN, S. E. O. *Corpora* no ensino de tradução: o papel do automonitoramento e da conscientização cognitivo-discursiva no processo de aprendizagem de tradutores novatos. In: VIANA, V.; TAGNIN, S. E. O. (Org.). ***Corpora* no ensino de línguas estrangeiras**.1. ed. São Paulo: HUB Editorial, 2010.p. 189-203. v. 1.

ALMEIDA, A. M.; LOTUFO NETO, F. A mediunidade vista por alguns pioneiros da área mental. **Rev. Psiq. Clin.**, v. 31, n. 3, p. 132-141, 2004. crossref http://dx.doi.org/10.1590/S0101-60832004000300003

ANTHONY, L. **AntConc** *(Windows, Macintosh OS X, and Linux)*: Build 3.2.4. 2011. Disponível em: http://www.antlab.sci.waseda.ac.jp/software/README_AntConc3.2.4.pdf. Acesso em: 18 jul. 2014.

ARRAIS, A. **Médium brasileiro ajudo u a projetar Espiritismo internacionalmente**. *G1*, 3 fev. 2010. Disponível em: http://g1.globo.com/Noticias/Mundo/0,,MUL1554524-5602,00-MEDIUM+BRASILEIRO+AJUDOU+A+PROJETAR+ESPIRITISMO+INTERNACIONALMENTE.html. Acesso em: 2 fev. 2014.

AUBERT, F. H. A tradução literal: impossibilidade, inadequação ou meta? In: *Ilha do desterro*. Florianópolis: UFSC, 1987 *apud* BARBOSA, H. G. **Procedimentos técnicos da tradução**: uma nova proposta. Campinas: Pontes, 1990.

AUBERT, F. H. **Introdução à metodologia de pesquisa terminológica bilíngue**. São Paulo: CITRAT/FFLCH/USP, 1996.

BAIO, M. F. P. **A chegada de O livro dos espíritos no Brasil**. s.d. Disponível em: http://www.oclarim.org/site/_pages/ler.php?idartigo=3189. Acesso em: 26 jun. 2014.

BAKER, M. *Corpus* Linguistics and Translation Studies: implications and applications. In: _____*et al*. (Ed.). **Text and technology**: in honour of John Sinclair. Amsterdã; Filadélfia: John Benjamins Publishing Company, 1993. crossref http://dx.doi.org/10.1075/z.64.15bak

BAKER, M. *Corpora* in translation studies: an overview and some suggestions for future research. **Target**, Amsterdam, v. 7, n. 2, 1995, p. 223-243. 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**: studies in language engineering in honour of Juan C. Sager. Amsterdã; Filadélfia: John Benjamins Publishing Company, 1996. p. 175-186. crossref http://dx.doi.org/10.1075/btl.18.17bak

BAKER, M. A *corpus*-based view of similarity and difference in translation. Manchester, **International Journal of *Corpus* Linguistics**, v. 9, n. 2, 2004, p. 167-193. crossref http://dx.doi.org/10.1075/ijcl.9.2.02bak

BARBOSA, H. G. **Procedimentos técnicos da tradução**: uma nova proposta. Campinas: Pontes, 1990.

BARBOSA, H. G. Tradução, mercado e profissão no Brasil. **Confluências**: revista de tradução científica e técnica, n. 3, nov. 2005. Disponível em: http://confluencias.net/cfl/category/n-3/. Acesso em: 21 set. 2010.

BARBOSA, P. **O espiritismo básico**. 5. ed. Rio de Janeiro: FEB, 2002.

BARROS, L. **Conhecimentos de Terminologia geral para a prática tradutória**. São José do Rio Preto: Nova Graf, 2007.

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

COSTA FILHO, J. E. **Elementos para um glossário bilíngue (português e inglês) de termos-chave da Teoria da Metáfora Conceitual**. 2008. 148 f. Dissertação (Mestrado em Linguística Aplicada) – Departamento de Letras, Universidade Estadual do Ceará, Fortaleza.2008.

DAYRELL, C. **Investigating lexical patterning in translated Brazilian Portuguese**: a *corpus*-based study. 2005. Thesis (Doctor of Philosophy) – Faculty of Arts, University of Manchester, Manchester, 2005.

DIVALDO F. **Biografia**. 2 jan. 2012. Disponível em: http://www.divaldofranco.com.br/biografia.php. Acesso em: 2 fev. 2014.

DOYLE, A. C. **History of Spiritualism**. London: Cassell and Company, Ltd., 1926.

DOYLE, A. C. **História do Espiritismo**. 2008. Disponível em: http://www.luzespirita.org.br/leitura/L10.html. Acesso em: 24 fev. 2014.

ESPIRITISMO. In: HOUAISS, A. **Dicionário eletrônico Houaiss da língua portuguesa**. Rio de Janeiro: Objetiva, 2009.

ESPIRITUALISMO. In: HOUAISS, A. **Dicionário eletrônico Houaiss da língua portuguesa**. Rio de Janeiro: Objetiva, 2009.

KARDEC, A. **Le livre des esprits**. 4e éd. Paris: Didier et Cie, 1860.

KRIEGER, M. da G.; FINATTO, M. J. B. **Introdução à Terminologia**: teoria e prática. São Paulo: Contexto, 2004.

JESUS, S. M. de. **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, Belo Horizonte, 2008.

JESUS, S. M. de; NUNES, L. P. K. *Corpus* design and compilation. In: KUNZ *et al*. (Ed.). **Caught in the middle**: language use and translation – a Festschrift for Erich Steiner on the occasion of his 60th birthday. Saarbrucken: Universaar, 2014. p. 177-194.

MATOS, M. V. **Biografia de Chico Xavier**. s.d. Disponível em: http://www.100anoschicoxavier.com.br/biografia-de-chico-xavier. Acesso em: 21 fev. 2014.

SARGENT, E. **The scientific basis of Spiritualism**. 2. ed. Boston: Colby and Rich Publishers, 1881.

SARGENT, E. **Bases científicas do Espiritismo**. [S.l.: s.n., s.d.]. Disponível em: http://www.autoresespiritasclassicos.com/Autores%20Espiritas%20Classicos%20%20Divers

os/Epes%20Sargent/Epes%20Sargent%20-%20Bases%20Cient%C3%ADficas%20do%20Espiritismo.htm. Acesso em: 22 jul. 2014.

SILVEIRA, F. de A. **As equivalências terminológicas e o caso dos epônimos no domínio da Dermatologia**: estudo comparado português-inglês em um conjunto terminológico. 2005. Dissertação (Mestrado em Análise Linguística) – Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista Júlio de Mesquita Filho, São José do Rio Preto, 2005.

SINCLAIR, J. ***Corpus* concordance and collocation**. Oxford: Oxford University Press, 1991 *apud* DAYRELL, C. Investigating lexical patterning in translated Brazilian Portuguese: a *corpus*-based study. 2005. Thesis (Doctor of Philosophy) – Faculty of Arts, University of Manchester, Manchester, 2005.

TAGNIN, S. E. O. **O jeito que a gente diz**: expressões convencionais e idiomáticas. São Paulo: Disal, 2005.

TAGNIN, S.; BEVILACQUA, C. ***Corpora* na Terminologia.** São Paulo: HUB Editorial, 2013.

TYNM, M. **The afterlife revealed**: what happens after we die. Guilford: White Crow Books, 2011. Appendix D.

VIANA, V.; TAGNIN, S. ***Corpora* no ensino de línguas estrangeiras**. São Paulo: Hub Editorial, 2010.

VIEIRA, M. A.; JESUS, S. M de. Tradução de textos religiosos: um *corpus* paralelo do livro *Nosso Lar*, de Chico Xavier. **Horizonte científico**, Uberlândia, v. 7, n.1, set. 2013. Disponível em: http://www.seer.ufu.br/index.php/horizontecientifico/article/view/22484. Acesso em: 24 fev. 2014.

XAVIER, F. C. **Nosso Lar**. Pelo Espírito André Luiz. 40. ed. Rio de Janeiro: Editora da FEB, 1992.

XAVIER, F. C. **The Astral City**: the story of a doctor's odyssey in the spirit world. Pelo Espírito André Luiz. Trad. Antônio Leite e GEAE (Grupo de Estudos Avançados de Espiritismo). 1.ed. eletrônica. 2000. Disponível em: http://www.geae.inf.br/en/books/ac/. Acesso em: set. 2011.

Artigo recebido em: 30.03.2015
Artigo aprovado em: 22.06.2015

Ilocuções comissivas em dicionários híbridos italiano>português brasileiro: proposta de dicionarização a partir do uso de *corpora*

Commissive illocutions in Italian>Brazilian Portuguese hybrid dictionaries: a proposal for dictionarizing through *corpora*

Renato Railo Ribeiro*

RESUMO: O objetivo é expor os resultados de um projeto de mestrado desenvolvido entre 2012-2015 cuja proposta fora a de sugerir um modo de inserção, em dicionários híbridos italiano>português-brasileiro, de informações acerca da dimensão ilocucionária de ambas as línguas, a partir de pesquisa feita em *corpora* eletrônicos. O texto foi assim estruturado: apresentação dos pressupostos teóricos; dos critérios adotados para investigação de ilocuções em *corpora* (e resultados obtidos); da sugestão de inserção de ilocuções em dicionários híbridos italiano>português-brasileiro; da análise lógico-conceitual que a sustenta. Justifica-se a pesquisa na medida em que são poucos os estudos que se debruçaram sobre a presença e/ou inclusão de informações ilocucionárias em dicionários híbridos. A proposta final foi a de inserir: no interior dos verbetes, marcas de uso referentes às classes de ilocução e de remissivas que possam conduzir o leitor a uma seção externa à nomenclatura, com explicações referentes às classes de ilocuções, lista com sua respectiva tipologia disposta frequencialmente, e exemplos de uso retirados de *corpora*. Concluiu-se que a inserção de tais informações pode ser de utilidade a estudantes brasileiros que pretendam (re)conhecer tal aspecto pragmático da língua italiana.

PALAVRAS-CHAVE: Dicionários híbridos. Atos de Fala. Linguística de *Corpus*.

ABSTRACT: The paper aims to expose the results of a Master's project, developed from 2012 to 2015, which has proposed a way to insert, in Italian>Brazilian Portuguese hybrid dictionaries, information about the illocutionary dimension of both languages through electronic corpora. The paper was organized as follows: presentation of theoretical assumptions, the procedures adopted for investigating illocutions in corpora (and their results), a suggestion for inserting illocutions in Italian>Brazilian Portuguese hybrid dictionaries, and a conceptual analysis that justifies it. The reasons for these works are the fact that there are few studies that have looked into the presence and/ or inclusion of illocutionary information in hybrid dictionaries. The suggestion for insertion was as follows: within the entries, inclusion of usage features concerning illocution classes and cross references which conduct the reader into a section that is external to the nomenclature; within a that section, inclusion of a section that contains explanations concerning illocution classes, a list of their conventionally recurring species arranged according to frequency in corpora and usage examples of such illocutions, taken from corpora. The conclusion is that such insertion may be helpful to every Brazilian intending to know/ to recognize such pragmatic aspect of Italian language.

KEYWORDS: Hybrid dictionaries. Speech acts. Corpus Linguistics.

* Mestre em Letras (FFLCH/ USP). Graduado em Filosofia (USJT) e em Biblioteconomia e Ciência da Informação (ECA/ USP). Email: renatorailo@yahoo.it.

1. Introdução

Alguns pesquisadores apresentam os dicionários gerais de língua como pontos de encontro entre as várias práticas discursivo-comunicativas existentes, apontando assim sua função de auxiliar no aprendizado de regras e usos linguísticos imprescindíveis para a mútua compreensão entre sujeitos inseridos em comunidades linguísticas semelhantes e/ou diversas. Van Hoof (1998) é um deles, além de Krieger (2007: 301), para quem "os dicionários de língua são instrumentos potenciais para o aprendizado e desenvolvimento da leitura, da redação e da comunicação em geral".

Um tipo específico de dicionário geral de língua é o *dicionário híbrido*, aquele que segundo Welker (2004: 202) está "entre os bilíngues e os monolíngues". Um caso concreto de híbrido é o dicionário *Parola Chiave* (2012) (doravante, *PC*), composto por um dicionário monolíngue da língua italiana (o *Dizionario Italiano per stranieri*, da Editora Giunti) e enriquecido com a tradução de seus verbetes para o português brasileiro.

Considera-se tal acréscimo como fundamental para o aprendizado do italiano por parte de um estudante brasileiro. Porém, concebe-se também que para a aquisição/ aperfeiçoamento da competência comunicativa de uma língua estrangeira é necessário ir além do (re)conhecimento de aspectos semânticos da língua estudada, buscando identificar também aspectos *pragmáticos* específicos a tal língua, i. e., aqueles que dizem respeito às regras da língua em uso interacional e/ou comunicativo concreto. A *teoria dos atos de fala* (doravante, TAF) é uma das abordagens possíveis de investigação e produção de conhecimento acerca da língua sob perspectiva pragmática, pois segundo Guerra Salas e Gómez Sánchez (2005: 355) possui como característica-chave a possibilidade de representar, por meio de categorias, aquilo que um falante realiza "ao proferir determinado enunciado, levando em conta os meios linguísticos dos quais se utiliza".

O presente artigo foi elaborado com base em uma pesquisa de mestrado[1] desenvolvida entre os anos de 2012 e 2015 e que se direcionou pelas seguintes perguntas: (1) considerando-se a possibilidade de um dicionário híbrido italiano>português brasileiro (tal como o *PC*) conter informações acerca dos atos de fala, de que maneira esse tipo de informação poderia ser

[1] Atos de fala em dicionários híbridos italiano>português-brasileiro: sugestão para dicionarização de ilocuções via *corpora*. Dissertação (Mestrado). Orient. Profª. Drª. Angela Maria Tenório Zucchi. Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo (FFLCH/ USP), São Paulo, 2015. A dissertação pode ser obtida no Banco de Teses da USP (http://www.teses.usp.br/) ou neste outro endereço: http://minhateca.com.br/Railo-Ribeiro.

disponibilizado? (2) considerando-se a Linguística de *Corpus* (doravante, LC) como metodologia de pesquisa linguística de grande valia para a Lexicografia – pelo fato de permitir "a identificação das unidades convencionais da língua" (TAGNIN, 2005: 21) –, de que modo tal metodologia poderia contribuir para a tarefa de dicionarização de ilocuções? Vale dizer que o objetivo geral da referida pesquisa foi, assim, o de responder a tais perguntas, de modo a, uma vez respondidas, oferecer uma sugestão de inserção de ilocuções em dicionários híbridos italiano>português brasileiro.

Logo, o que se pretende com o presente artigo é apresentar o caminho percorrido ao longo de tal investigação, percurso esse dividido em quatro tópicos: apresentação da fundamentação teórica; análise de *corpora* e resultados obtidos; sugestão de inserção de ilocuções em dicionários híbridos italiano>português brasileiro; análise lógico-conceitual que sustenta tal sugestão. Neste sentido, justifica-se o trabalho na medida em que até hoje foram poucos os estudos que se debruçaram sobre a presença e/ou inclusão de informações pragmático-ilocucionárias em dicionários híbridos, tanto quanto foram poucos aqueles que se propuseram a examinar modos de buscar, reconhecer, estabelecer e/ou extrair ilocuções de *corpora* com vistas a uma posterior dicionarização. Com isso, pretendeu-se oferecer uma possibilidade de construir um instrumento capaz de favorecer o aprendizado (conhecimento, uso) de padrões ilocucionários de duas línguas a partir de dicionários.

2. Fundamentação teórica

A fundamentação teórica será apresentada em três partes: (2.1) a teoria dos atos de fala de Austin (1990[1975]), enquanto objeto de estudo da Pragmática Linguística; (2.2) os dicionários híbridos, tais como o *PC* (2012), enquanto objetos de estudo da Lexicografia; (2.3) os *corpora* eletrônicos, tais como o *corpus* italiano *Paisà* e o *Corpus do Português*, enquanto fontes de pesquisa linguística elaboradas por procedimentos característicos da Linguística da *Corpus.*

2.1. A TAF de Austin e as ilocuções comissivas

A TAF é a teoria que permite a compreensão da linguagem como forma de ação (KOCH, 1992: 19). Isso significa, segundo Bianchi (2008: 57), que expressões linguísticas, ao serem proferidas, criam novos fatos, i. e., modificam o contexto comunicativo em questão. As ações, neste caso, de acordo com Sbisà (2009 [1989]: 40), devem ser concebidas como disposições

que, em razão de dado proferimento, "operam transformações em seu contexto" – o que conduz à noção de contexto de fala como instante "construído pelo ato linguístico" (SBISÀ, 2009 [1989]: 45). Não por acaso a TAF é objeto de estudo da Pragmática Linguística, disciplina que segundo Bianchi (2008: 10-11) se ocupa do uso da linguagem, daquilo que um falante comunica, do contexto, do significado em contexto e do significado nas interações sociais.

Uma vez que as ilocuções constituem os contextos de fala, esses últimos devem ser entendidos como "rede de crenças, intenções, atividade dos interlocutores, e contribui para a determinação de suas intenções comunicativas" (BIANCHI, 2008: 24). Isso significa que a língua é, como propõe Bazzanella (2009: 105), mais do que uma estrutura abstrata, pois é principalmente um instrumento para interação social. Em outras palavras, segundo a perspectiva pragmática a língua é correlata às necessidades comunicativas, e, assim sendo, o termo *pragmática* designa também uma *competência* a partir da qual o falante que vem a possuí-la se torna capaz de reconhecer o "uso funcional dos recursos linguísticos (produção de funções linguísticas, atos de fala)" (QECR, 2001: 35). A competência pragmática, deste modo, é vista como uma das subcompetências da competência comunicativa, aquela que permite a um indivíduo "agir utilizando especificamente meios linguísticos" (QECR, 2001: 29).

Para sugerir a inserção de ilocuções em dicionários híbridos, adotou-se uma TAF específica: aquela proposta pelo filósofo inglês John L. Austin e exposta em sua obra póstuma *Quando dizer é fazer: palavras e ações* (1990[1975]). Para Austin, existem três atos subjacentes a qualquer proferimento: locucionário, ilocucionário e perlocucionário. Grosso modo, enquanto o ato locucionário é o ato *de* dizer algo, o ilocucionário é o ato que se realiza *ao* dizer algo, sendo que o perlocucionário, por sua vez, é o ato que se realiza *por* se ter dito algo.

Austin distingue cinco classes de atos ilocucionários ou ilocuções:

> o *vereditivo* é um exercício de julgamento, o *exercitivo* é uma afirmação de influência ou exercício de poder, o *comissivo* é assumir uma obrigação ou declarar uma intenção, o *comportamental* é a adoção de uma atitude e o *expositivo* é o esclarecimento de razões, argumentos e comunicações (1990[1975]: 131).

Afora isso, ofereceu uma lista com espécies de ilocuções que pertencem a tais classes:

Tabela 1. Classes de força ilocucionária e algumas de suas espécies, segundo Austin (1990[1975]: 123-130).

Classes/ Espécies	Vereditivos	Exercitivos	Comissivos	Comportamentais	Expositivos
	absolvo, condeno, constato, determino	ordeno, mando, concedo, designo	prometo, compactuo, comprometo-me, contrato	peço desculpas, agradeço, deploro, compadeço-me	afirmo, nego, observo, aceito

Percebe-se a associação que o filósofo propõe entre ilocuções e verbos (os chamados verbos ilocucionários, responsáveis por realizar determinada ilocução segundo dada convenção linguística). A preferência[2] de Austin pelos verbos ilocucionários se deu em função do fato de que as ações constituídas por esses são *explícitas* – ou seja, mais facilmente reconhecíveis. Pode-se, assim, apontar tal associação como o primeiro motivo pela escolha da TAF de Austin, pois como se pretendeu investigar ilocuções em *corpora* escritos de modo a se valer de ferramentas de natureza estatística[3] oferecidas pela LC para legitimar determinadas escolhas lexicográficas futuras, julgou-se necessário um critério para reconhecê-las. Outro motivo por se ter escolhido a TAF austiniana foi o critério por ele utilizado para obtenção de ilocuções, a saber, *a forma verbal na primeira pessoa do singular do presente do indicativo na voz ativa* (1990[1975]: 122) uma vez que, por constituírem ilocuções explicitamente reconhecíveis, tal forma verbal permite a criação de uma sintaxe de busca padronizada a ser utilizada para a investigação de ilocuções em *corpus* eletrônico; elimina a necessidade de se analisar todo[4] o *corpus* na busca por casos indubitáveis de ilocuções.

Além disso, restringiu-se a pesquisa ao se optar por *ilocuções comissivas* –aquelas que "empenham o falante a certa linha de ação" (SBISÀ, 2009[1989]: 71). As espécies de ilocuções comissivas investigadas foram aquelas ligadas aos verbos italianos *promettere, assicurare, garantire, giurare, impegnarsi*, e aos seus equivalentes semânticos em português brasileiro *prometer, assegurar, garantir, jurar, comprometer-se*. A peculiaridade de tais ilocuções reside no fato de que os verbos a essas associados possuem certa sinonímia entre si, segundo pesquisa

[2] O que não significa que o autor tenha considerado *apenas* os verbos como expedientes linguísticos utilizados para realizar ações. Veja-se, p. ex., seu artigo *A Plea for Excuses* (1979[1955-6]), no qual oferece uma análise de advérbios.
[3] De acordo com as quais os contextos de proferimento são ou negligenciados, ou considerados de modo restrito, o que prejudicaria o seu reconhecimento-extração caso fossem buscadas ilocuções implícitas.
[4] Já que as ilocuções, por não se reduzirem à determinada categoria sintático-gramatical, tampouco equivalerem ao significado de dado item lexical-sentença, de certo modo transcendem a expressão linguística, ao passo que os *corpora* escritos são constituídos por dados eminentemente linguísticos – o que demandaria uma investigação que os considerasse em sua totalidade, caso o critério apontado não fosse estabelecido.

feita ao dicionário *Sinonimi e contrari* (2009), entrevendo-se, de início, a possibilidade de que, após a pesquisa em *corpora*, fosse possível apontar se tais verbos também são pragmaticamente equivalentes.

2.2. Dicionários híbridos e suas divisões internas

Por dicionário geral de língua entende-se determinada compilação de itens lexicais de uma ou mais línguas, organizada de determinado modo (em geral, em ordem alfabética das entradas) e a partir da qual são fornecidas diversas informações linguísticas, em termos de estrutura e uso, de acordo com os objetivos a que se pretende alcançar com tal obra (em consideração com seu público-alvo). Segundo Welker (2004: 93), tais compilações apresentam os itens lexicais de uma língua em sua (quase) totalidade, fornecendo informações de natureza fonológica, sintático-gramatical, semântica etc.

Um tipo de dicionário geral de língua é o dicionário híbrido, entendido como meio-termo entre um monolíngue (aquele que informa a "estrutura e funcionamento da língua" (BORBA, 2003: 16)) e um bilíngue (aquele que "trata da equivalência das unidades lexicais de duas línguas" (HÖFLING et al, 2004: 2)). De acordo com Welker (2004: 202), os híbridos possuem duas características principais: (1) uma base monolíngue (língua-fonte); (2) equivalentes semânticos na língua-alvo.

Os híbridos são construídos e examinados pela Lexicografia, segundo Krieger (2008: 170) a área científica que cobre diversos aspectos de registros lexicais, problematiza a constituição e tratamento de unidades simples e complexas (além de outras faces do léxico geral, quando registrados em dicionários de língua) e discute aspectos metodológicos da produção dicionarística. De acordo com Welker (2004: 11), a Lexicografia pode ser dividida em dois campos: Lexicografia prática (técnica, prática de elaborar dicionários) e Metalexicografia (estudo de problemas relacionados à elaboração, críticas e pesquisas de uso de dicionários), enquanto que Duran (2004: 17) apresenta uma divisão dessa última: Metalexicografia empírica (realiza pesquisas de campo acerca da interação dicionário-usuário) e teórica (realiza análises lógico-metodológicas de dicionários). Vale dizer que a mencionada pesquisa de mestrado inseriu-se no campo da Metalexicografia teórica, pois seu objetivo foi o de examinar um aspecto particular da confecção de dicionários (a inserção de ilocuções).

Os dicionários (e os híbridos não fogem à regra) são compostos de três partes: (1) macroestrutura, que corresponde à organização do dicionário como um todo – em geral,

dividido em duas grandes subpartes: nomenclatura (conjunto de entradas) e material externo (as demais seções: introdução, lista de abreviaturas, seção de informações gramaticais etc); (2) microestrutura, que corresponde às informações dispostas no interior dos verbetes das entradas; (3) medioestrutura, que corresponde ao sistema de remissivas utilizadas para conduzir o consulente de uma parte a outra do dicionário. Julgou-se relevante apresentar e examinar a divisão tripartite dos dicionários porque não faria sentido propor qualquer sugestão de inserção de informações ilocucionárias que não estivesse em conformidade com a estrutura-padrão dos dicionários gerais de língua (incluindo aí os híbridos)[5].

O *PC* (2012), aqui adotado como dicionário-base, possui um material externo nomeado de *Gramática de uso da língua italiana*, que por sua vez possui uma subseção chamada *Usos e regras pragmáticas*, responsável por expor fórmulas de cortesia e de cumprimento, marcadores discursivos, interjeições etc. Nesta subseção, os atos de fala não são indicados diretamente, mas não seria equivocado dizer ser possível apontá-los como implícitos a alguns tópicos abordados na subseção, a saber: na definição do que são regras pragmáticas (PC, 2012: 1010), do que são formas de tratamento (2012: 1011), do que são interjeições (2012: 1016), entre outros. Neste sentido, tendo em vista a divisão tripartite dos dicionários híbridos, a princípio vislumbrou-se a possibilidade de que seções externas à nomenclatura pudessem conter informações explícitas sobre atos de fala.

Quanto à sua microestrutura, veja-se o caso do verbete da entrada *promettere* (2012: 587):

> **prométtere**
> *v.tr.*[coniugato come *mettere*]. 1 Impegnarsi a fare o a dare qualcosa o a comportarsi in un certo modo: *mi ha promesso di venire, di aiutarmi, di non disturbarmi più; promettere un regalo, un lavoro*. 2 ♣ Far sperare, far prevedere: *un cielo che promette bel tempo; il suo atteggiamento non promette niente di buono; un ragazzo che promette bene*. □ **Prometer**.

Tal verbete pode ser assim decomposto: (1) **prométtere**: entrada (com indicação da sílaba tônica); (2) *v.tr.*: categoria sintático-gramatical; (3) [coniugato come *mettere*]: conjugação do verbo; (4) **1** Impegnarsi a fare o a dare qualcosa o a comportarsi in un certo modo: *mi ha promesso di venire, di aiutarmi, di non disturbarmi più; promettere un regalo, un*

[5] A não ser que fosse proposta uma nova estruturação – algo, porém, que ultrapassaria os limites traçados pelos objetivos desta investigação.

lavoro: primeira acepção seguida de exemplos de uso; (5) **2 ♣** Far sperare, far prevedere: *un cielo che promette bel tempo; il suo atteggiamento non promette niente di buono; un ragazzo che promette bene*: segunda acepção seguida de exemplos de uso; (6) ♣: símbolo que indica uso figurado; (7) □ **Prometer**: equivalência semântica na língua-alvo. Percebe-se aqui que não há referência aos (nem mesmo a possibilidade de se inferir presença implícita dos) atos de fala – diferentemente do *Diccionario Salamanca de la Lengua Española* (1996), que em sua microestrutura oferece *marcas de uso pragmáticas*, ou seja, categorias cujo objetivo é fornecer informações relativas a algum aspecto do uso concreto que convencionalmente se faz de determinado item lexical/ vocábulo.

Veja-se o verbete da entrada *siempre* (1996: 1460), como exemplo:

> **siempre**
> *adv. temp.* 1 En todo momento o todo el tiempo. *El escritor siempre fumaba puros*. *Te echaré de menos siempre*. OBSERVACIONES: Normalmente se refiere al periodo de tiempo del que se habla o a uno que se da por consabido. (…) 4 AFIRMACIÓN. Resalta una cosa que se dice o refuerza una afirmación que no admite duda: *Siempre vivirás mejor solo que mal acompañado*.

O *Salamanca* (1996: 1460) informa na microestrutura que tal item lexical "ressalta algo que se diz ou reforça uma afirmação que não admite dúvidas", ao lado de um exemplo ("*é sempre melhor viver só do que mal acompanhado*") e do rótulo *Afirmação*. Dito de outro modo: para tal dicionário, a ilocução de *siempre* é a de afirmação, ou seja, o ato de fala convencional/ padrão ligado ao proferimento do item lexical próprio da língua espanhola *siempre* é o *ato de afirmar* algo. Outra possibilidade a princípio vislumbrada, então, para a inserção de ilocuções e tendo em vista a divisão tripartite dos dicionários híbridos, é a inserção de marcas de uso.

Por fim, quanto à medioestrutura do *PC*, veja-se um caso concreto: o da entrada *sonare* (2012: 734):

> **sonàre** vedi **suonàre**.

Vedi é o indicador de remissiva e conduz o leitor à entrada *suonare*. Ao que parece, todas as remissivas do *PC* estão na microestrutura e conduzem o leitor a outra entrada (e seu respectivo verbete), não existindo, contudo, alguma que faça remissão aos materiais externos do dicionário. Além disso, em tal dicionário tais expedientes não foram utilizados com o objetivo de remeter o consulente a informações de cunho pragmático-ilocucionário da língua,

e sim semânticos e/ou ortográficos (ou seja, estabelecendo relações entre itens lexicais sinônimos e entre diversas possibilidades gráficas de se realizar a transcrição fonética da fala, respectivamente)[6]. Contudo, ao final do artigo será apresentada uma sugestão de inclusão de ilocuções que possa se valer da relevante função desempenhada pelas remissivas (medioestrutura).

2.3. LC, representatividade linguística e os *corpora Paisà* e *do Português*

Muitos autores apontam a utilidade dos *corpora* para a Lexicografia. Béjoint (2000: 97-99), Welker (2004: 89) e Krieger (2008: 171) são exemplos – além de Faber et al (1999: 177), segundo os quais a investigação baseada em *corpora* permite novos métodos de estudo em várias áreas de investigação da linguagem, como Análise do Discurso, Pragmática e Lexicografia. De acordo com eles,

> [c]om a introdução do uso de *corpora* textuais informatizados, multiplicaram-se formidavelmente as possibilidades de análise linguística a serem empreendidas por lexicógrafos interessados na compilação das entradas. A *Linguística de Corpus* tornou patente a importância de se derivar descrições linguísticas detalhadas de uma língua tal como esta é usada naturalmente, já que tal estudo pode ajudar a revelar regularidades (e irregularidades) em nosso uso que antes desconhecíamos, ou ainda ajudar a observá-las de maneira uniforme, sob uma perspectiva ampla e com índices de frequência mais confiáveis (FABER et al, 1999: 185-6).

Isso se deve ao fato de a LC, segundo definição proposta por Tagnin e Teixeira (2004: 321), poder ser vista também como metodologia de pesquisa segundo a qual a exploração da linguagem é feita por meio de evidências empíricas, extraídas por meio do uso de ferramentas computacionais de um *corpus* de linguagem natural autêntica, criteriosamente reunido e disponível eletronicamente.

Neste sentido, segundo Berber-Sardinha (2004: 18) *corpus* é um conjunto de dados linguísticos (do uso oral e/ou escrito da língua) suficientemente extenso, passível de ser processado por computador e utilizado para reconhecimento e extração de padrões linguísticos. Vale dizer que uma das principais ferramentas para análise de *corpus* é a lista de frequência, que segundo Gutierrez Gonzalez (2007: 15) "revela a ocorrência de cada palavra naquele

[6] Motivo pelo qual a entrada *suonare* não foi examinada, já que neste ponto apenas a função e o modo de apresentação das remissivas estão em questão.

corpus específico, além de listar todas as palavras que o compõem". Novamente com Berber-Sardinha (2000: 351), a padronização nesse caso é evidenciada pela recorrência de um item lexical, podendo significar um caso de padrão lexical. O mesmo diz Tagnin (2002) ao reforçar que um *corpus* oferece "a forma mais usual na língua sob investigação" – evidenciando assim sua utilidade para a Lexicografia, já que ao se pensar em elaborar um dicionário representativo da língua geral deve-se pensar em critérios que justifiquem tais escolhas e que essas últimas de fato representem a língua em geral.

Para tal justificação/ representatividade, Berber-Sardinha (2000: 342-343) aponta a necessidade de se valorizar "a extensão do *corpus*, o que significa em termos simples que para ter representatividade o *corpus* deve ser o maior possível" – sobretudo porque quanto mais extenso um *corpus,* maior a probabilidade de que evidenciem-se itens lexicais desempenhando a mesma função. Além disso, sendo difícil de se estabelecer o tamanho ideal de uma amostra genuinamente representativa da língua, quanto mais extenso for um *corpus*, mais essa representatividade aproximar-se-á do ideal.

O critério de representatividade em função da extensão norteou a opção que aqui se fez pelo *Corpus Paisà* e pelo *Corpus do Português* como fontes de pesquisa. Segundo informações de seu site (2014), o *Corpus Paisà* é composto de "uma ampla coleção de textos autênticos da língua italiana extraídos da internet", totalizando cerca de 250 milhões de *tokens* – número também apontado por Lyding et al (2014: 36-37). De acordo os últimos autores (2014: 36-37), o *Paisà* contém cerca de 388 mil documentos, retirados de 1067 sites diferentes (em geral os da Fundação Wikimedia). Por sua vez, o *Corpus do Português* (2014) possui 45 milhões de *tokens* – 15 milhões pertencendo a documentos dos séculos 1200 a 1400, 10 milhões aos séculos 1500 a 1700, e 15 milhões aos séculos 1800 a 1900. Dos anos 1700 em diante, os textos podem ser divididos entre os de Portugal e os do Brasil, e para os de 1900 em diante podem ser divididos também entre textos falado, fictício, jornalístico e acadêmico.

Outro critério para escolha de tais *corpora* foi o de representatividade da língua *standard*, já que, objetivando sugerir a inserção de ilocuções em dicionários gerais de língua, considerou-se relevante minimizar traços de regionalismo – para tal, optando-se por *corpora* majoritariamente escritos. Além disso, tais *corpora* escritos também foram escolhidos por permitirem acesso remoto e gratuito.

Por fim, vale dizer que a abordagem de investigação utilizada foi a *corpus-based* (abordagem baseada em *corpus*), uma vez que se pretendeu investigar padrões linguísticos pré-determinados (as ilocuções comissivas).

3. Análise de *corpora* e resultados obtidos

Os critérios para a investigação de atos de fala em *corpora* linguísticos foram os seguintes: (a) adoção de cinco verbos ilocucionários italianos (*Promettere, Assicurare, Garantire, Giurare, Impegnarsi*) da classe austiniana das ilocuções comissivas; (b) adoção dos cinco verbos correspondentes em português brasileiro (*Prometer, Assegurar, Garantir, Jurar, Comprometer-se*), tal como fornecidos pelo *PC* (2012); (c) adoção da forma verbal na primeira pessoa do singular do presente do indicativo da voz ativa como sintaxe de busca para investigação em *corpora*; (d) adoção do *Corpus Paisà* e do *Corpus do Português* como fonte de pesquisa.

Em seguida, realizou-se a pesquisa de tais verbos em *corpora*. Vale dizer que além da sintaxe de busca adotada, refinou-se a busca em ambos os *corpora* ao se adotar a anotação gramatical *verbo*. Assim, obteve-se o seguinte resultado no caso dos verbos italianos:

Tabela 2. Nº de ocorrências totais de ilocuções comissivas pesquisadas no *Corpus Paisà*.

Verbo ilocucionário/ ilocução	Nº de ocorrências totais
Assicuro	785
Giuro	318
Garantisco	158
Prometto	114
Mi impegno	31

Feito isso, todas as ocorrências de todos os verbos foram salvos em um arquivo Word, de modo a se poder realizar posteriormente uma análise de natureza qualitativa capaz de verificar se em todas as ocorrências tais verbos desempenharam de fato a função ilocucionária esperada. Tal desempenho não ocorreu, de modo que alguns casos foram excluídos da contagem real – a saber, casos de: homonímia (p. ex., *juro* como substantivo e não como verbo na primeira pessoa...); negação ("*non prometto*..." etc); repetição de ocorrências; ininteligibilidade. Os números reais, então, foram os seguintes:

Tabela 3. Nº de ocorrências reais de ilocuções comissivas pesquisadas no *Corpus Paisà*.

Verbo ilocucionário/ ilocução	Nº de ocorrências reais
Assicuro	704
Giuro	272
Garantisco	143
Prometto	89
Mi impegno	23

O mesmo procedimento foi feito em relação às ilocuções do português brasileiro: sintaxe de busca, anotação gramatical e critérios de exclusão idênticos. A única diferença foi a utilização de outro filtro, aquele que restringe a pesquisa a textos do português brasileiro. Eis os resultados totais e reais:

Tabela 4. Comparação entre o nº de ocorrências totais e reais de ilocuções comissivas pesquisadas no *Corpus do Português*.

Verbo ilocucionário/ ilocução	Nº de ocorrências totais	Nº de ocorrências reais
Juro	5055	388
Garanto	505	421
Prometo	404	292
Asseguro	157	145
Comprometo-me / me comprometo[7]	71	60

Percebe-se uma diferença considerável entre "Juro" e os outros verbos em termos de números totais – deve-se ao fato de que o *corpus* trouxe inúmeros casos de "Juro" como termo da área de Economia (apesar da anotação gramatical *verbo* ter sido feita previamente). Após a exclusão desses e demais casos anômalos, o número de ocorrências se equilibrou entre todos os verbos.

Eis, agora, a comparação entre os verbos ilocucionários do italiano e do português brasileiro:

[7] Vale dizer que quanto à forma verbal de primeira pessoa do singular do presente do indicativo na voz ativa de *comprometer*, a pesquisa feita no *Corpus do Português* considerou tanto "comprometo-me" quanto "me comprometo". Mesmo somando as ocorrências das duas possibilidades, esse verbo foi o menos frequente entre todos.

Tabela 5. Comparação entre o nº de ocorrências reais de ilocuções comissivas pesquisadas no *Corpus Paisà* e no *Corpus do Português*.

Verbo ilocucionário/ ilocução PT	Nº de ocorrências reais	Verbo ilocucionário/ ilocução IT	Nº de ocorrências reais
Garanto	421	Assicuro	704
Juro	388	Giuro	272
Prometo	292	Garantisco	143
Asseguro	145	Prometto	89
Comprometo	60	Mi impegno	23

Vale ressaltar que tanto no caso do italiano como no caso do português brasileiro as ilocuções pesquisadas se mostraram frequentes – com exceção, talvez, dos casos de *comprometer-se* e *impegnarsi*. Neste caso, considera-se que os resultados das análises de *corpora* justificam a inserção de tais ilocuções em dicionários, em função de tal recorrência, sendo possível, por outro lado, a exclusão daqueles verbos mencionados.

4. Sugestão de inserção de ilocuções em dicionários híbridos italiano>português brasileiro

Finalmente, sugere-se o seguinte modo de inserção de ilocuções em dicionários híbridos italiano>português brasileiro, a partir da consideração da tradicional divisão tripartite dos dicionários (microestrutura, medioestrutura e macroestrutura) e dos números de frequência obtidos via *corpora*:

(i) inserção, no interior dos verbetes, de marcas de uso referentes às classes de ilocução (microestrutura);

(ii) inserção, no interior de verbetes, de remissivas capazes de conduzir o leitor a uma seção externa à nomenclatura (medioestrutura);

(iii) inserção, como material externo, de uma seção que contenha: (a) explicações referentes às classes de ilocuções e lista de respectivas espécies convencionalmente recorrentes do italiano, dispostas segundo frequência em *corpora*; (b) lista de espécies de ilocuções convencionalmente equivalentes do português brasileiro, dispostas segundo frequência em *corpora*; (c) exemplos de uso, retirados dos *corpora*, de tais ilocuções (macroestrutura).

Os pontos (i) e (ii) podem ser visualizados abaixo, estando em vermelho as sugestões propostas com base nos verbetes extraídos do *PC* (2012):

giuràre
v.tr. **1** Dichiarare, promettere qualcosa con un giuramento: *giuro di dire la verità; giurami che non mi tradirai; giurare a Dio, sulla Bibblia, sul proprio onore.* □ **jurar** Comissivo - vedi *r* | *Giurare il falso* = mentire sotto (spec. in un processo) □ **prestar falso-testemunho** ◊ Affermare con certezza **ṡ** assicurare: ti giuro che non lo sapevo □ **jurar** | *Non ci giurerei* = non ne sono sicuro □ **não apostaria nisso** | *Giurarla a qualcuno* = fare propositi di vendetta nei suoi confronti □ **jurar vingança contra alguém**.

garantìre
v.tr.[*garantìsco, garantìsci*]. **1** Assicurare sotto la propria responsabilità il rispetto di un impegno proprio o altrui: *garantire la restituzione di un prestito* ◊ Assicurare la qualità e il perfetto funzionamento di una merce, impegnandosi con l'acquirente a ripararla o a sostituirla gratuitamente quallora si guasti entro un certo periodo: *la fabbrica garantisce questo televisore per un anno* **2** Dare per certo, assicurare: *ti garantisco che ha deto la verità; la buona pensione gli garantisce una vecchiaia tranquilla* □ **garantir** Comissivo - vedi *r* ♦ **garantirsi** *v.pr.* Assicurarsi, prendere delle precauzioni contro eventuali danni o rischi □ **garantir-se**.

assicuràre
v.tr. **1** Rendere sicuro **ṡ** garantire: *risparmia per assicurare l'avvenire alla famiglia.* **2** Dare per certo; affermare con sicurezza: *lo assicurarono che non sarebbe successo nulla; assicura di non saperne niente.* □ **assegurar** Comissivo - vedi *r* **3** Fermare, fissare saldamente: *assicura le imposte prima di uscire* □ **fixar** **4** Proteggere un bene da eventuali rischi e danni facendo un'assicurazione: *assicurare l'auto contro il furto* □ **pôr no seguro** ♦ **assicurarsi** *v.pr.* **1** Accertarsi, controllare: *assicurati che le porte siano tutte chiuse* □ **assegurar-se** **2** Garantirsi da eventuali rischi e danni facendo un'assicurazione: *assicurarsi contro il furto* □ **fazer seguro**.

prométtere
v.tr.[coniugato come *mettere*]. **1** Impegnarsi a fare o a dare qualcosa o a comportarsi in un certo modo: *mi ha promesso di venire, di aiutarmi, di non disturbarmi più; promettere un regalo, un lavoro.* **2** ♣ Far sperare, far prevedere: *un cielo che promette bel tempo; il suo atteggiamento non promette niente di buono; un ragazzo che promette bene.* □ **Prometer** Comissivo - vedi *r*

E, abaixo, o ponto (iii), que para ser elaborado se valeu de trechos retirados de algumas das obras examinadas ao longo da pesquisa:

Seção *r* - Usos e regras pragmáticas

Nas situações reais de comunicação, além das regras de fonética, morfologia e sintaxe, são aplicadas as **regras pragmáticas** da língua. As regras pragmáticas regem os **usos da língua**: estão fortemente ligadas ao contexto de comunicação e aos sentidos das ações, presentes nas escolhas linguísticas. As regras pragmáticas refletem a condição social dos interlocutores, os objetivos da comunicação, os efeitos que se procura atingir sobre os interlocutores e sobre a situação de comunicação. Apresentamos alguns usos linguísticos e as relativas regras pragmáticas da língua italiana[8].

Atos de Fala

A teoria dos atos de fala é aquela que permite a compreensão da linguagem como forma de ação[9]. De acordo com essa abordagem, a linguagem, além de ser um meio utilizado para expressar pensamentos, descrever o mundo e transmitir mensagens/ informações, é também usada para realizar ações[10] - sendo que estas se constituem pelo próprio dizer. A ação levada a cabo por um enunciado é também conhecida pelo nome de ilocução.

Existem cinco classes de ilocução.

(i) Comissiva – abrange ilocuções que caracterizam promessas ou a assunção de algo; por outras palavras, ao serem enunciadas, comprometem o falante a determinada linha de ação[11]. Vejam-se as **ilocuções comissivas mais comuns** do italiano e do português brasileiro, segundo frequência de uso:

1	*Assicurare*	Ti **assicuro** che ho provato a farlo ma i risultati sono molto scadenti!
2	*Giurare*	Io **giuro** che se mai sarò libero, ti sposerò.
3	*Garantire*	Io ti **garantisco** che ero a scuola.
4	*Promettere*	D'ora in avanti **prometto** che starò più attento ai consigli dei miei dottori.

1	*Garantir*	No entanto, **garanto** ao povo brasileiro que quem está no Governo tem compromisso.
2	*Jurar*	E te **juro** pela alma de minha mãe que eu caso com Maria.
3	*Prometer*	Deixe-me ir e **prometo** não incomodá-lo mais
4	*Assegurar*	E eu lhe **asseguro** que o Exército não está dormindo.

As razões que justificam as sugestões feitas acima são explicitadas no tópico seguinte.

[8] Extraído da seção "Usos e regras pragmáticas" do *PC* (2012: 1010).
[9] Extraído de Koch (1992: 19).
[10] Extraído de Sbisà (2009[1989]:45).
[11] Extraído de Austin (1990[1975]: 123 e 12).

5. Análise lógico-conceitual para justificação da sugestão

Em termos de microestrutura, três possibilidades de inserção de ilocuções foram consideradas possíveis: (1) marcas de uso, (2) notas de uso e (3) exemplos de uso. As marcas de uso, como se viu, foram empregadas pelo *Salamanca*, de modo que também o poderiam ser pelo *PC*. Contudo, como restringiu-se a presente investigação aos verbos ilocucionários explícitos, a inserção de marcas referentes às espécies de ilocuções poderia ocasionar redundância, como se percebe abaixo:

> **assicuràre**
> *v.tr.* **1** Rendere sicuro **ś** garantire: *risparmia per assicurare l'avvenire alla famiglia*. **2** Dare per certo; affermare con sicurezza: *lo assicurarono che non sarebbe successo nulla; assicura di non saperne niente*. □ **assegurar Assegurar 3** Fermare, fissare saldamente: *assicura le imposte prima di uscire* □ **fixar 4** Proteggere un bene da eventuali rischi e danni facendo un'assicurazione: *assicurare l'auto contro il furto* □ **pôr no seguro ♦ assicurarsi** *v.pr.* **1** Accertarsi, controllare: *assicurati che le porte siano tutte chiuse* □ **assegurar-se 2** Garantirsi da eventuali rischi e danni facendo un'assicurazione: *assicurarsi contro il furto* □ **fazer seguro**.

Ou seja, ao lado da tradução para o português brasileiro do item lexical da língua italiana *assicurare* seria colocada uma marca de uso referente à ilocução em questão (*Assegurar*). Contudo, haveria redundância entre a palavra traduzida e a marca de uso, podendo gerar confusão entre semântica e pragmática.

Por essa razão, optou-se por inserir marcas que fizessem referência às *classes* de atos de fala em questão. Veja-se abaixo:

> **assicuràre**
> *v.tr.* **1** Rendere sicuro **ś** garantire: *risparmia per assicurare l'avvenire alla famiglia*. **2** Dare per certo; affermare con sicurezza: *lo assicurarono che non sarebbe successo nulla; assicura di non saperne niente*. □ **assegurar Comissivo 3** Fermare, fissare saldamente: *assicura le imposte prima di uscire* □ **fixar 4** Proteggere un bene da eventuali rischi e danni facendo un'assicurazione: *assicurare l'auto contro il furto* □ **pôr no seguro ♦ assicurarsi** *v.pr.* **1** Accertarsi, controllare: *assicurati che le porte siano tutte chiuse* □ **assegurar-se 2** Garantirsi da eventuali rischi e danni facendo un'assicurazione: *assicurarsi contro il furto* □ **fazer seguro**.

Aqui, como se pode observar, exclui-se a redundância e garante-se um acréscimo verdadeiramente informativo.

Outra possibilidade seria a de inserir notas de uso, que segundo Duran (2004: 110-111) são capazes de trazer informações acerca de aspectos pragmáticos:

> **assicuràre**
> *v.tr.* **1** Rendere sicuro **ṡ** garantire: *risparmia per assicurare l'avvenire alla famiglia*. **2** Dare per certo; affermare con sicurezza: *lo assicurarono che non sarebbe successo nulla; assicura di non saperne niente*. □ **assegurar** **Obs: ilocução da classe dos comissivos, que, ao ser enunciada, compromete o falante a determinada linha de ação.** **3** Fermare, fissare saldamente: *assicura le imposte prima di uscire* □ **fixar 4** Proteggere un bene da eventuali rischi e danni facendo un'assicurazione: *assicurare l'auto contro il furto* □ **pôr no seguro ♦ assicurarsi** *v.pr.* **1** Accertarsi, controllare: *assicurati che le porte siano tutte chiuse* □ **assegurar-se 2** Garantirsi da eventuali rischi e danni facendo un'assicurazione: *assicurarsi contro il furto* □ **fazer seguro**.

Tal possibilidade é visivelmente mais informativa do que a marca de uso. Entretanto, sua desvantagem é extensão, pois certamente aumentaria o tamanho do dicionário. Além disso, termos como "atos de fala", "verbo ilocucionário", "ilocuções", "força ilocucionária" e correlatos são menos conhecidos pelo público em geral do que termos como "verbo", "substantivo", "significado" etc, o que indica a necessidade de se fornecer por algum texto externo à nomenclatura informações auxiliares acerca daqueles.

Uma terceira possibilidade, como alternativa às marcas e notas de uso: exemplos de uso – que, inclusive, poderiam ser retirados do próprio *corpus* utilizado. Veja-se abaixo:

> **assicuràre**
> *v.tr.* **1** Rendere sicuro **ṡ** garantire: *risparmia per assicurare l'avvenire alla famiglia*. **2** Dare per certo; affermare con sicurezza: ***ti assicuro che non è facile tirare conclusioni; te lo assicuro tu non sai cogliere il punto***. □ **assegurar 3** Fermare, fissare saldamente: *assicura le imposte prima di uscire* □ **fixar 4** Proteggere un bene da eventuali rischi e danni facendo un'assicurazione: *assicurare l'auto contro il furto* □ **pôr no seguro ♦ assicurarsi** *v.pr.* **1** Accertarsi, controllare: *assicurati che le porte siano tutte chiuse* □ **assegurar-se 2** Garantirsi da eventuali rischi e danni facendo un'assicurazione: *assicurarsi contro il furto* □ **fazer seguro**.

O problema é que a simples inserção de exemplos de uso não garantiria que o consulente inferisse o ato de fala em questão – novamente, seria necessário um texto externo que fizesse tal *link*.

Neste ponto, voltou-se à possibilidade das notas de uso, mas dessa vez com o acréscimo de remissivas capazes de conduzir o consulente a um texto externo (aqui chamado de "Seção *r*")[12]. Veja-se abaixo como ficaria tal sugestão no verbete:

> **assicuràre** (2012: 49)
> *v.tr.* **1** Rendere sicuro **ś** garantire: *risparmia per assicurare l'avvenire alla famiglia*. **2** Dare per certo; affermare con sicurezza: *lo assicurarono che non sarebbe successo nulla; assicura di non saperne niente*. □ **assegurar Obs: ilocução da classe dos comissivos, que, ao ser enunciado, compromete o falante a determinada linha de ação; *vedi r*** **3** Fermare, fissare saldamente: *assicura le imposte prima di uscire* □ **fixar 4** Proteggere un bene da eventuali rischi e danni facendo un'assicurazione: *assicurare l'auto contro il furto* □ **pôr no seguro ♦ assicurarsi** *v.pr.* **1** Accertarsi, controllare: *assicurati che le porte siano tutte chiuse* □ **assegurar-se 2** Garantirsi da eventuali rischi e danni facendo un'assicurazione: *assicurarsi contro il furto* □ **fazer seguro**.

Neste ponto, deve-se então esclarecer melhor a confecção dessa seção externa. Como se disse anteriormente, essa seção (aqui chamada de "*r*") foi pensada como aquela que seria responsável por fornecer ao consulente de um dicionário híbrido informações variadas acerca da pragmática, dos atos de fala (ilocuções) etc de ambas as línguas. É de se supor que tal seção pudesse ser elaborada de *n* modos, contendo *n* informações – motivo pelo qual não seria desprovido dizer que as possibilidades de construção de uma tal seção seriam ilimitadas. Nesse sentido, a título de exemplo, a sugestão é a seguinte: aproveitar-se de algumas informações pragmáticas já contidas no *PC;* acrescentar a essas informações algumas definições e demais explicações contidas em algumas das obras utilizadas como fundamentação teórica do trabalho[13]; somar a tudo isso, por fim, informações obtidas pela análise de *corpora*. Com isso, seria possível, consequentemente, propor uma maneira de se compor uma seção externa dessa natureza – aqui, a fictícia "Seção *r*". .

Do *PC* (2012: 1010) aproveitou-se o seguinte trecho:

[12] O nome do texto externo (a fictícia "Seção *r*") foi posto no interior do verbete sem o termo "seção" para evitar repetições por considerar a possibilidade de um dicionário incluir várias remissivas, remetendo o leitor a várias seções do dicionário (com "seção", sua extensão aumentaria consideravelmente).
[13] O que não significa que um dicionário deva necessariamente se aproveitar de definições e explicações propostas por autores específicos. Optou-se por tal aproveitamento neste trabalho em razão de sua natureza crítico-metodológica (como se disse em 2.2, este é um trabalho que se insere no âmbito da Metalexicografia Teórica, não no da Lexicografia Prática).

> **Usos e regras pragmáticas**
> Nas situações reais de comunicação, além das regras de fonética, morfologia e sintaxe, são aplicadas as **regras pragmáticas** da língua. As regras pragmáticas regem os **usos da língua**: estão fortemente ligadas ao contexto de comunicação e aos sentidos das ações, presentes nas escolhas linguísticas. As regras pragmáticas refletem a condição social dos interlocutores, os objetivos da comunicação, os efeitos que se procura atingir sobre os interlocutores e sobre a situação de comunicação. Apresentamos alguns usos linguísticos e as relativas regras pragmáticas da língua italiana.

Seguindo-se a esse trecho da "Seção *r*", responsável por fornecer informações acerca da Pragmática, pensou-se que uma subseção intitulada "Atos de Fala", responsável por fornecer uma definição dessa teoria e informações específicas acerca de tal tópico, poderia ser incluída. Como exemplo, a definição a ser inserida poderia ser aquela proposta por Koch (1992: 19):

> **Atos de Fala**
> A teoria dos atos de fala é aquela que permite a compreensão da linguagem como forma de ação.

Após tal definição, outras informações poderiam ser dadas, como, p. ex., o trecho abaixo, retirado de Sbisà (2009[1989]:45):

> De acordo com essa abordagem, a linguagem, além de ser um meio utilizado para expressar pensamentos, descrever o mundo e transmitir mensagens/ informações, é também usada para realizar ações - sendo que estas se constituem pelo próprio dizer.

Como acréscimo a tal explicação e já indicando que em seguida seriam fornecidas informações acerca das classes e espécies de ilocuções, sugeriu-se o seguinte texto, baseado em Ribeiro (2015: 100):

> A ação levada a cabo por um enunciado é também conhecida pelo nome de ilocução.
> Existem cinco classes de ilocução.

Feito isso, seriam enumeradas as definições de cada uma das classes de ilocução. A título de exemplo, seria possível valer-se das definições estipuladas pelo próprio Austin (1990[1975]: 131):

> **(i) Comissiva** – abrange ilocuções que caracterizam promessas ou a assunção de algo; por outras palavras, ao serem enunciadas, comprometem o falante a determinada linha de ação.

Mais uma vez, na fictícia "Seção *r*", como se pode notar, apenas a definição dos comissivos foi oferecida – já que o presente trabalho limitou-se a trabalhar com algumas espécies dessa classe. É de se esperar, contudo, que um dicionário ofereça informações e definições acerca de todas as classes e de um número razoável de espécies correspondentes.

Por fim, vislumbrou-se a possibilidade de fornecer informações acerca das espécies mais recorrentes em italiano e em português brasileiro, segundo dados obtidos em *corpora*. Como se pode notar, as ilocuções enumeradas a seguir figuram segundo ordem de frequência – de acordo com os resultados da análise de *corpora* realizada – e em conjunto com exemplos de uso retirados dos próprios *corpora* (RIBEIRO, 2015: 102):

> Vejam-se as ilocuções comissivas mais comuns do italiano e do português brasileiro, segundo frequência de uso:

1	*Assicurare*	Ti **assicuro** che ho provato a farlo ma i risultati sono molto scadenti!
2	*Giurare*	Io **giuro** che se mai sarò libero, ti sposerò.
3	*Garantire*	Io ti **garantisco** che ero a scuola.
4	*Promettere*	D'ora in avanti **prometto** che starò più attento ai consigli dei miei dottori.

1	*Garantir*	No entanto, **garanto** ao povo brasileiro que quem está no Governo tem compromisso.
2	*Jurar*	E te **juro** pela alma de minha mãe que eu caso com Maria.
3	*Prometer*	Deixe-me ir e **prometo** não incomodá-lo mais
4	*Assegurar*	E eu lhe **asseguro** que o Exército não está dormindo.

Tal seria, então, a composição da "Seção *r*" (que pode ser visualizada integralmente nas páginas 13-13).

Deve-se lembrar, entretanto (ver pg. 16), que a possibilidade de inserção de um texto externo foi pensada em conjunto com a possibilidade de serem iseridas notas de uso no interior de verbetes. Contudo, impossível não perceber nesse ponto certa redundância entre aquilo que a "Seção *r*" (texto externo, macroestrutura) informa e o que informa o verbete de *assicurare*

(notas de uso, microestrutura), pois ambos – verbete e texto externo – fornecem explicações acerca do que é uma ilocução comissiva, com a diferença de que o texto externo amplia a explicação ao não se limitar a isso – pois também fornece explicações sobre Pragmática e TAF, fornece exemplos de uso etc. Isso permite a conclusão de que sua eficácia informativa é maior do que a das notas de uso, não havendo necessidade, então, de manter essas últimas.

Logo, em termos de microestrutura, voltou-se a considerar as marcas de uso referentes às classes de ilocuções, mas desta vez acrescidas de remissivas capazes de conduzir o leitor a informações detalhadas presentes na seção externa "*r*". Reveja-se (para visualizar os demais, ver pág. 13) abaixo um exemplo de verbete:

> **giuràre**
> *v.tr.* **1** Dichiarare, promettere qualcosa con un giuramento: *giuro di dire la verità; giurami che non mi tradirai; giurare a Dio, sulla Bibblia, sul proprio onore.* □ **jurar** Comissivo - vedi *r* *Giurare il falso* = mentire sotto (spec. in un processo) □ **prestar falso-testemunho** ◊ Affermare con certezza **s** assicurare: ti giuro che non lo sapevo □ **jurar** | *Non ci giurerei* = non ne sono sicuro □ **não apostaria nisso** | *Giurarla a qualcuno* = fare propositi di vendetta nei suoi confronti □ **jurar vingança contra alguém**.

Feita a remissão do leitor à seção externa à nomenclatura ("Seção *r*"), em tal texto o leitor encontraria informações a respeito da pragmática e dos atos de fala, bem como uma lista com as classes de ilocuções existentes e algumas de suas espécies, a saber, as mais recorrentes nos *corpora* do italiano e do português brasileiro pesquisado. Além disso, exemplos de uso concreto de tais ilocuções de ambas as línguas seriam fornecidas em tal seção, possibilitando ao leitor que percebesse que, ainda que verbos do italiano e do português-brasileiro possam ser considerados semanticamente equivalentes (p. ex., *assicurare* e *assegurar*), pragmaticamente falando (ou, de modo ainda mais específico, *ilocucionariamente* falando), não há tal equivalência – ao menos não em termos de frequência de uso, tal como se observou com a análise de *corpora*, que serviu de base para a confecção das duas tabelas acima que mostram que os usos não são equivalentes em italiano e português-brasileiro.

Um ponto importante a ser notado, entretanto, é que conjuntamente a tais exemplos retirados de *corpora* o lexicógrafo poderia fornecer análises linguísticas[14] capazes de explicitar as razões pelas quais um ou outro uso é frequentemente mais comum – ou seja, apontar

[14] Análises de tal natureza não são oferecidas aqui, devido às complexidades inerentes a tais investigações (se levadas a cabo, ultrapassariam os limites traçados pelos objetivos da presente investigação).

semelhanças e diferenças entre tais usos. Por exemplo, viu-se que a frequência de uso de *assicuro* é maior do que a de *prometto*; haveria uma razão para tal? Em caso afirmativo, qual? Quais seriam as características intrínsecas exclusivamente ao uso de *garanto* a ponto de esse ser mais frequente do que *garantisco*, seu equivalente semântico em italiano? Por que *assicuro* é, em italiano, o mais correntemente utilizado, ao passo que seu equivalente semântico em português brasileiro, *asseguro*, é aquele cujo número de ocorrências é o mais baixo entre todos os pesquisados de sua língua? Se um dicionário híbrido italiano> português brasileiro fornecesse respostas a tais perguntas, inegavelmente auxiliaria seu consulente a (re)conhecer aspectos (semelhantes e dessemelhantes) de uso de ambas as línguas, favorecendo-lhe a aquisição da competência pragmática e, consequentemente, o aprendizado da língua estrangeira como um todo.

A figura abaixo oferece uma visão panorâmica das sugestões (págs. 13-14) feitas e explicadas neste trabalho, a partir da qual se pode observar mais claramente a integração entre as três estruturas-padrão dos dicionários híbridos:

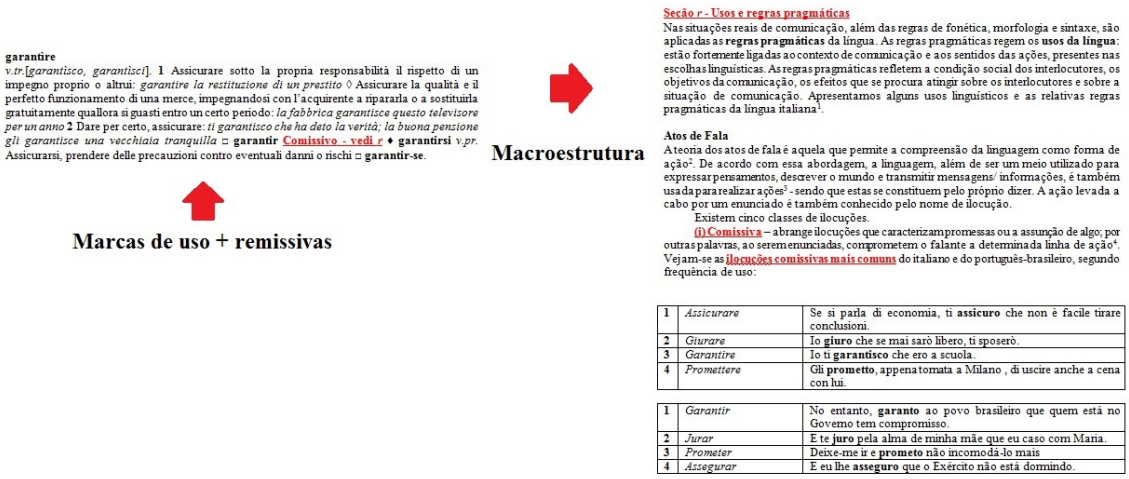


Figura 1. Integração entre as propostas para a micro, média e macroestrutura (RIBEIRO, 2015, p. 106).

Ou seja, supondo-se que dado consulente de um dicionário híbrido italiano>português-brasileiro estivesse interessado em obter informações acerca do verbo italiano *garantire*, ao buscar a entrada desse item lexical e ler o seu verbete, ele deparar-se-ia com a marca de uso, *Comissivo*, seguida por uma remissiva, *vedi r*. A partir de tal indicação, o leitor poderia então abrir o dicionário hipotético nessa seção fictícia, momento em que obteria todas aquelas

informações de cunho pragmático-ilocucionário sugeridas anteriormente e dispostas da maneira como previamente descrita e justificada.

6. Considerações finais

A partir da sugestão oferecida, e para responder as questões propostas na Introdução, concluiu-se que: (1) dicionários híbridos tais como o *PC* podem oferecer informações de natureza pragmático-ilocucionárias a partir da consideração de suas três estruturas: microestrutura (marcas de uso dadas nos verbetes referentes às classes de ilocuções), medioestrutura (remissivas dispostas após as marcas de uso dadas nos verbetes que conduzam a um texto externo) e macroestrutura (texto externo); (2) a LC se mostrou útil no momento em que, pela lista de frequência, apontou entre as ilocuções escolhidas aquelas que são mais recorrentes, justificando assim sua inserção (ou exclusão) em dicionários híbridos, além do fato de que, com tal lista, pode-se dispor ilocuções no dicionário de acordo com a frequência de cada uma.

Concluiu-se também que tal inserção pode ser de utilidade a estudantes brasileiros do italiano, aprendizes e/ou não, que pretendam conhecer/ reconhecer tal aspecto pragmático da língua italiana, de modo a compará-lo com sua língua materna e tornarem-se capazes de empregar-identificar tais regras convencionalmente estabelecidas em contextos de comunicação/ interação que possam vir a ter com a língua *standard* da terra de Dante.

Referências

AUSTIN, J. L. **Quando dizer é fazer**: palavras e ação [How to Do Things with Words, 2ª ed., 1975]. Trad. D. Marcondes. Porto Alegre: Artes Médicas, 1990.

_____. A Plea for Excuses. Proceeding of the Aristotelian Society, v. 57, 1956-7. In: _____. **Philosophical Papers**. 3ª ed. Ed. J. O. Urmson e G. J. Warnock. Oxford: Oxford University Press, 1979, p. 175-204. crossref http://dx.doi.org/10.1093/019283021x.003.0008

BAZZANELLA, C. **Linguistica e pragmatica del linguaggio**: un'introduzione. Roma-Bari: Laterza, 2009.

BÉJOINT, H. **Modern Lexicography**: An Introduction. Oxford: Oxford University Press, 2000.

BERBER-SARDINHA, T. Linguistica de *Corpus*: histórico e problemática. **DELTA**, v. 16, n. 2, 2000, p. 323-367.

_____. Linguística de *Corpus* e tradução: prosódia semântica. In: **Linguística de *Corpus***. Barueri: Manole, 2004, p. 233-250.

BIANCHI, C. **Pragmatica del linguaggio**. Roma-Bari: Laterza, 2008.

BORBA, F. S. **Organização de dicionários**: uma introdução à lexicografia. São Paulo: Ed. Unesp, 2003.

CORPUS do Português. Disponível em: http://www.*corpus*doportugues.org/. Acesso em: 01.10. 2014.

DURAN, M. S. **Dicionários bilíngues pedagógicos**: análise, reflexões e propostas. Dissertação (Mestrado). Orient. Profa. Dra. Claudia Maria Xatara. UNESP, 2004.

FABER, P.; PÉREZ HÉRNANDEZ, C.; MORENO ORTIZ, A. Lexicografía Computacional y Lexicografía de *Corpus*. **Volumen Monográfico**, 1999, pp. 175-213. Disponível em: tecnolengua.uma.es/doc2/resla98.pdf. Acesso em: 15.11.15.

GUERRA SALAS, L.; GÓMEZ SÁNCHEZ, E. Pragmática y lexicografía: análisis de las marcas pragmáticas en el Diccionario Salamanca de la lengua Española. **Actas del XVI Congreso Internacional de Ásele**, 2005, p. 353-362. Disponível em: http://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/16/16_0351.pdf. Acesso em: 29.04.12.

GUTIERREZ GONZALEZ, Z. M. **Linguística de *Corpus* na análise do internetês**. Dissertação (Mestrado). Orient. Prof. Dr. Antonio Paulo Berber Sardinha. São Paulo: PUC-SP, 2007.

HÖFLING, C.; SILVA, M. C. P. da; TOSQUI, P. O dicionário como material didático na aula de língua estrangeira. **Intercâmbio**, v. 13, 2004, p. 1-7.

KOCH, I. V. **A interação pela linguagem**. São Paulo: Contexto, 1992.

KRIEGER, M. G. O dicionário de língua como potencial instrumento didático. In: ISQUERDO, A. N.; ALVES, I. M. (orgs.). **As ciências do léxico**: lexicologia, lexicografia e terminologia, v. 3. Campo Grande: Ed. UFMS, 2007, p. 295-310.

_____. Lexicologia, lexicografia e terminologia: impactos necessários. In: ISQUERDO, A. N.; FINATTO, M. J. B. (orgs.). **As ciências do léxico**: lexicologia, lexicografia, terminologia, v. 4. Campo Grande: Ed. UFMS, 2008, p. 161-176.

LYDING, V. et al. The Paisà *Corpus* of Italian Web Texts. In: BILDHAUER, F; SCHÄFER, R. (eds.). **Proceedings of the 9th Web as *Corpus* Workshop (WaC-9).** Gothenburg: ACL, 2014, p. 36–43. crossref http://dx.doi.org/10.3115/v1/w14-0406

PAISÀ. ***Corpus* Italiano**. Disponível em: http://www.*corpus*italiano.it/. Acesso em: 30.08.2014.

PAROLA CHIAVE. **Dizionario di italiano per brasiliani**. Trad. C. A. Dastoli et al. São Paulo: Martins Fontes, 2012.

QECR. **Quadro Europeu Comum de Referência para as Línguas**: aprendizagem, ensino, avaliação. Porto: Edições ASA, 2001.

RIBEIRO, R. R. **Atos de fala em dicionários híbridos italiano>português-brasileiro**: sugestão para dicionarização de ilocuções via *corpora*. Dissertação (Mestrado). Orient. Profª. Drª. Angela M. T. Zucchi. Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo (FFLCH/ USP), São Paulo, 2015.

SALAMANCA. **Diccionario de la lengua española**. Ed. J. Gutiérrez. Madrid: Santillana-Univ. de Salamanca, 1996.

SBISÀ, M. **Linguaggio, ragione, interazione**. Per una pragmatica degli atti linguistici [1989]. Trieste: EUT Edizioni Università di Trieste, 2009.

SINONIMI e Contrari Minore. **Dizionario fraseologico delle parole equivalente analoghe e contrarie**. 3ª ed. Bologna: Zanichelli, 2009.

TAGNIN, S. E. O. Os *Corpora*: instrumentos de auto-ajuda para o tradutor. In: **Cadernos de Tradução IX**. Florianópolis, UFSC, 2002. Disponível em: http://www.cadernos.ufsc.br/online/9/stella.htm.

_____. **O jeito que a gente diz**. Barueri: DISAL, 2005.

_____; TEIXEIRA, E. D. Linguística de *Corpus* e tradução técnica: relato da montagem de um *corpus* multivarietal de culinária. **TradTerm,** 10, 2004, p. 313-358. crossref http://dx.doi.org/10.11606/issn.2317-9511.tradterm.2004.47184

WELKER, H. A. **Dicionários**: uma pequena introdução à lexicografia. 2ª ed. rev. e ampl. Brasília: Thesaurus, 2004.

VAN HOOF, H. Os tradutores e os dicionários. In: DELISLE, J.; WOODSWORTH, J. (orgs.). **Os tradutores na história**. Trad. S. Bath. São Paulo: Ática, 1998, p. 241-253.

Artigo recebido em: 30.03.2015
Artigo aprovado em: 24.06.2015

Tradução e Retradução de *The Picture of Dorian Gray*, de Oscar Wilde: um estudo de *corpus* com foco na apresentação do discurso

Translation and Retranslation of Oscar Wilde's The Picture of Dorian Gray: a *corpus*-based study focusing on discourse presentation

Lívia Cremonez Domingos[*]
Igor A. Lourenço da Silva[**]

RESUMO: Neste trabalho, analisa-se um *corpus* de pequenas dimensões constituído por excertos aleatoriamente extraídos do texto-fonte *The picture of Dorian Gray*, de Oscar Wilde (1891), e por seus respectivos correspondentes na primeira tradução e em duas das retraduções mais recentes da obra para o português brasileiro. O foco da análise incide sobre a apresentação do discurso e outros aspectos estilísticos a fim de testar a hipótese da retradução, segundo a qual a primeira tradução de uma obra literária é sempre incompleta e domesticadora. Adotando-se o arcabouço metodológico da Linguística de *Corpus*, seguiram-se diversas etapas, dentre as quais se destacam a etiquetagem do *corpus*, seu alinhamento e seu processamento semi(automático) utilizando o programa WordSmith Tools (versão 5.0). Os resultados apontam indícios a favor da hipótese da retradução quando se consideram os parâmetros de análise adotados nesta pesquisa.

PALAVRAS-CHAVE: Tradução. Retradução. Apresentação do discurso. Oscar Wilde. Linguística de *Corpus*.

ABSTRACT: This article analyses a small *corpus* of randomly extracted excerpts from Oscar Wilde's novel *The picture of Dorian Gray* (1891) and corresponding excerpts from its first translation and two of its most recent retranslations into Brazilian Portuguese. The analysis focuses on discourse presentation and other stylistic features with a view to testing the retranslation hypothesis, i.e., the first translation of a literary text is always incomplete and domesticating. Corpus Linguistics' methodological framework was used to, among other procedures, label the corpus, align it and process it semi(automatically) using software WordSmith Tools (version 5). The results point to evidence that confirms the retranslation hypothesis from the perspective of the parameters used in this study.

KEYWORDS: Translation. Retranslation. Discourse presentation. Oscar Wilde. Corpus Linguistics.

1. Introdução

O termo retradução, conforme definido por Tahir-Gürçağlar (2009) em um verbete da *Routledge encyclopedia of translation studies*, denota tanto o ato de traduzir uma obra que já foi traduzida para uma mesma língua quanto o resultado desse ato. Trata-se de um fenômeno

[*] Graduanda do Curso de Bacharelado em Tradução, Instituto de Letras e Linguística, Universidade Federal de Uberlândia. *E-mail*: liviashy@gmail.com.
[**] Professor do Curso de Bacharelado em Tradução, Instituto de Letras e Linguística, Universidade Federal de Uberlândia. *E-mail*: ials@ileel.ufu.br.

que existe de longa data, haja vista as diversas traduções de textos sagrados como a Bíblia e de textos da Grécia e da Roma antiga para os mais diversos idiomas. Contudo, mesmo sendo a retradução uma prática tão ou mais corrente que a própria tradução, devido ao grande interesse comercial e acadêmico pelos textos canônicos, ela ainda tem sido objeto de poucas análises (MILTON; TORRES, 2003), pois o tema "retradução" exige uma compreensão peculiar do seu vasto campo de possibilidades.

O interesse pela retradução como objeto de estudo ganhou importância na década de 1990, com a publicação de um artigo de Antoine Berman no qual foi postulada a chamada "hipótese da retradução". Segundo essa hipótese, a primeira tradução de um texto literário consiste em um ato "incompleto" cuja completude só pode ser alcançada – ou ao menos buscada – por meio de retradução. Se por um lado observações intuitivas parecem apontar inúmeros casos que confirmam essa hipótese (*e.g.*, BERMAN, 1990, 1999; BENSIMON, 1990; GAMBIER, 1994; CHESTERMAN, 2000), destacando o caráter domesticador da primeira tradução de uma obra, diversas pesquisas também apontam na direção contrária ou corroboram a hipótese apenas parcialmente (*cf.* KOSKINEN; PALOPOSKI, 2004, 2010).

Nesse contexto, este artigo apresenta os resultados de uma pesquisa que buscou testar a "hipótese da retradução" com base em uma investigação (semi)automática de um *corpus* constituído por excertos aleatoriamente extraídos do texto-fonte *The picture of Dorian Gray* (1891), de Oscar Wilde, da sua primeira tradução, *O retrato de Dorian Gray*, produzida por João do Rio, em 1923, e duas de suas retraduções mais recentes para o português brasileiro: uma de Paulo Schiller, publicada em 2012; e outra de Doris Goettems, publicada em 2014. Mais especificamente, buscou-se: (i) realizar uma análise contrastiva de aspectos estilísticos dos textos do *corpus*, com foco na apresentação do discurso; e (ii) verificar em que medida essa análise contribui para refutar ou corroborar a "hipótese da retradução".

Fundamentou essa proposta de pesquisa o fato de que estudos da retradução pautados em análises automáticas e semiautomáticas são – pela bibliografia levantada até o momento – incipientes na literatura, que em geral se baseia em cotejos qualitativos entre trechos do texto-fonte e dos textos (re)traduzidos ou entre textos (re)traduzidos, buscando identificar, dentre outras, as tendências deformadoras explicitadas por Bateman (1999), como a clarificação, o alongamento e a racionalização. Nessas circunstâncias, adotou-se uma metodologia baseada em uma adaptação daquela proposta em Dastjerdi e Mohammadi (2013), a fim de se desenvolver um estudo que se filia à Linguística de *Corpus* e, assim, distancia-se dessa prática de cotejo

assistemático entre os textos e procura uma perspectiva de análise mais automatizada e quali-quantitativa.

Uma pesquisa dessa natureza é relevante para os Estudos da Tradução porque se dedica a um fenômeno corrente na área, para o qual tem sido dispensada pouca atenção teórica no país, muito embora diversas pesquisas com *corpus* apresentem subsídios para testar a hipótese da retradução (*e.g.*, BAKER, 2000, 2004; ASSIS, 2009; BARCELLOS, 2011). Também traz contribuição para a Linguística de *Corpus* no que diz respeito a estudos focados em obras literárias, pois analisa de forma (semi)automática um romance que faz parte do cânone literário e para o qual existem inúmeras traduções, edições e adaptações para a língua portuguesa (MILTON; VASCONCELLOS, 2003).

Este artigo está estruturado em seis seções, incluindo esta Introdução. Na Seção 2, dispõe-se a fundamentação teórica, com enfoque na retradução e nos estudos da apresentação do discurso. Na Seção 3, tecem-se breves considerações acerca da obra analisada, *The picture of Dorian Gray*, e suas (re)traduções para a língua portuguesa, fazendo-se referência a um trabalho que já abordou essa obra sob a perspectiva da retradução. Na Seção 4, descreve-se a metodologia de coleta e análise dos dados, com foco no detalhamento de como foi realizada a etiquetagem do *corpus*. Na Seção 5, disponibilizam-se a análise dos dados e a discussão dos resultados, com apresentação de dados quantitativos e de exemplos ilustrativos. Na Seção 6, fazem-se as considerações finais acerca do trabalho, incluindo as limitações desta pesquisa e sugestões para estudos futuros.

2. Fundamentação teórica

Esta seção se subdivide em duas subseções. Na primeira, apresentam-se os pressupostos teóricos relativos à hipótese da retradução, que serviu de motivação para este estudo. Na segunda, revisa-se a literatura sobre aspectos estilísticos, os quais foram utilizados como parâmetros de análise do *corpus*, sobretudo no que tange à apresentação do discurso.

2.1 A hipótese da retradução

Para Berman (1990), a primeira tradução é "incompleta" porque "deforma" o texto-fonte para acomodá-lo aos padrões da cultura-alvo. Em outras palavras, enquanto a primeira tradução neutraliza o texto-fonte – é domesticadora, nos termos de Venuti (1995) –, as traduções subsequentes, as retraduções, permitem a expressão da alteridade do texto-fonte no texto

(re)traduzido, ou seja, são estrangeirizadoras, tomando novamente os termos de Venuti (1995). Por essa hipótese, a primeira tradução seria domesticadora para permitir a recepção do texto "estrangeiro" na cultura-alvo, enquanto as traduções subsequentes se beneficiariam com a crescente familiarização da cultura-alvo com a cultura-fonte, podendo se pautar, portanto, na técnica/procedimento de "tradução literal".

Como explicam Dastjerdi e Mohammadi (2013), com base em Vándor (2009), a "hipótese da retradução" traz consigo o paradigma do idealismo, pelo qual quanto mais nos distanciamos da época do texto-fonte e quanto mais o traduzimos, melhor será a tradução obtida. Seria possível, assim, chegar a um texto ideal, a uma tradução que se encontre, seguindo as palavras de Goethe, em "perfeita identidade com o original" (*apud* SCHULTE; BIGUENET, 1992).

Todavia, não se chegou até o momento a um consenso sobre as motivações para se retraduzir. Berman (1990) enfatiza o "envelhecimento" da tradução ao longo do tempo, enquanto o texto-fonte permanece eternamente "jovem". Pym (1998) destaca a existência de diferentes funções pedagógicas dos textos e de disputas pela posse do conhecimento contido no documento a ser traduzido. Milton (2001) aponta que a política das editoras também pode ser um forte agente motivador da retradução, haja vista interesses de vendas garantidas, baixa relação custo-benefício e prestígio. Venuti (2003) sublinha a reafirmação do poder ou autoridade de certas instituições sociais, como seria o caso de instituições acadêmicas e religiosas. Susam-Sarajeva (2003), baseada no fato de que retraduções são realizadas em curto espaço de tempo entre uma e outra, afirma que a retradução está mais relacionada com as necessidades e atitudes dentro do sistema receptor do que com as características do texto-fonte que o tornam mais suscetível a retraduções. Brownlie (2006) sugere que os principais motivos seriam as mudanças de contexto social e a evolução das normas de tradução na cultura-alvo.

Independentemente das motivações, pesquisas sobre retradução – em geral adotando o cotejo qualitativo entre texto-fonte e texto (re)traduzido – têm buscado corroborar ou refutar a hipótese da retradução. A título de exemplo, citam-se duas duplas de autores que apresentam resultados distintos entre si, cabendo também sublinhar que adotam metodologias distintas: Koskinen e Paloposki (2004, 2010) e Dastjerdi e Mohammadi (2013). Em uma investigação empírica de diversas retraduções no âmbito da Finlândia, Koskinen e Paloposki, em um artigo de 2004, não encontraram indícios suficientes a favor da hipótese da retradução e, em um artigo de 2010, salientam que são inúmeras as categorias ou parâmetros de análise que podem ser

adotados e que é tênue a linha que separa uma retradução de uma revisão e, em diversos casos, de um plágio ou "transpirataria" – termo de Tarvi (2005, p. 137)[1]. Por sua vez, Dastjerdi e Mohammadi (2013) apresentam indícios favoráveis à hipótese da retradução a partir da análise de um *corpus* composto por três capítulos do romance *Pride and prejudice* (1813), de Jane Austin, juntamente com seus correspondentes da sua primeira tradução (de 1955) e de uma retradução (de 2007) para o persa.

O trabalho de Dastjerdi e Mohammadi (2013) é de especial interesse porque a metodologia nele adotada serviu de base para o presente artigo. Para a análise do romance, da sua primeira tradução e de uma de suas retraduções, fez-se uma amostragem aleatória e compilou-se um *corpus* de 5.505 palavras com os capítulos 3, 44 e 55 do romance. Em seguida, analisou-se o *corpus* manualmente, considerando três aspectos estilísticos: a razão *type/token* (*i.e.*, palavras distintas/número total de palavras), o comprimento médio das sentenças e a apresentação do discurso. Observaram-se uma razão *type/token* maior e, portanto, uma maior variedade lexical na primeira tradução, o que, segundo os autores, provavelmente se deveu à influência das normas estilísticas da época ou ao estilo individual de escrita da primeira tradutora. Também se constatou um menor comprimento médio das sentenças na primeira tradução, provavelmente em função dos mesmos fatores citados anteriormente e da preferência da primeira tradutora por produzir um texto mais facilmente legível em persa. Ademais, utilizando-se categorias elaboradas por Short (1996), encontrou-se, em todos os textos, um predomínio, nesta ordem, da Representação da Fala pelo Narrador, da Fala Direta e da Fala Direta Livre; contudo, identificaram-se, na primeira tradução, um aumento no percentual de Representação da Fala pelo Narrador e uma consequente redução no percentual de Fala Direta, o que, segundo os autores, implicou menor grau de realismo e vivacidade na fala das personagens da primeira tradução. Associada a esse resultado da apresentação da fala, também se notou uma redução no percentual de Fala Direta Livre na primeira tradução, redução essa que os autores atribuíram a uma explicitação da primeira tradutora, que inseriu orações introdutórias de elocução em diversas instâncias do texto. Para todas as variáveis investigadas, os valores encontrados se mantiveram praticamente idênticos entre o texto-fonte e a retradução.

O estilo e a apresentação do discurso são tema da próxima seção.

1 No Brasil, destacam-se as colocações de Denise Bottmann, que, no mínimo, reforçam esse limite ténue entre plágio e retradução. Cf. BOTTMANN, D. Não gosto de plágio. Disponível em: <www.naogostodeplagio.blogspot.com>. Acesso em: 30 abr. 2015.

2.2 Estilo e apresentação do discurso

Existem diferentes tradições em se tratando de estilística. Para fins deste artigo, entende-se estilo como "a expressão de um conjunto de traços linguísticos distintivos", sendo que "[o]s recursos empregados em um texto, independentemente de seu efeito na produção de significados, representam o resultado de escolhas por determinadas realizações léxico-gramaticais em detrimento de outras" (BARCELLOS, 2011, p. 30). Parte-se do princípio de que as escolhas linguísticas dão significado ao texto e, portanto, um autor ou tradutor, ao desenvolver determinado estilo, "privilegia leituras e formas de ver a realidade enquanto suprime ou apaga outras" (SIMPSON, 1993, p. 8 *apud* BARCELLOS, 2011, p. 30). Em outras palavras, autores e tradutores podem representar de várias maneiras aquilo que foi dito, escrito ou pensado pelas personagens de um texto.

Em se tratando especificamente da apresentação do discurso, a estilística permite observar que as escolhas léxico-gramaticais do escritor ou tradutor têm impacto no aparente controle que o narrador exerce sobre o que é dito pelas personagens. Apoiando-se em autores como Semino e Short (2004), que revisitaram modelo inicialmente criado por Leech e Short (1981) e apresentado em Short (1996), pode-se observar um *continuum* que vai do controle total do narrador sobre o que é dito pelas personagens até a aparente ausência de controle.

O Quadro 1 apresenta as principais categorias de Semino e Short (2004), sendo que, em cada coluna, parte-se de uma aparente ausência de controle do narrador no topo e segue-se em direção a um total controle do narrador na última linha. Em seguida, essas categorias são mais bem explicadas e acompanhadas, quando possível, por exemplos extraídos do próprio *corpus* objeto deste estudo. Inicia-se pelas categorias de apresentação da fala, quais sejam: Fala Direta, Fala Direta Livre, Fala Indireta, Fala Indireta Livre, Representação do Ato de Fala pelo Narrador e Representação da Voz pelo Narrador. Em seguida, passa-se para as categorias de apresentação da escrita, que segue uma divisão semelhante àquela da apresentação da fala, a saber: Escrita Direta, Escrita Direta Livre, Escrita Indireta, Escrita Indireta Livre, Representação do Ato de Escrita pelo Narrador. Por fim, tem-se apresentação do pensamento, que similarmente é categorizada em: Pensamento Direto, Pensamento Direto Livre, Pensamento Indireto, Pensamento Indireto Livre, Representação do Ato de Pensamento pelo Narrador e Narração Interna.

Quadro 1. Principais categorias de apresentação do discurso segundo Semino e Short (2004).

FALA	ESCRITA	PENSAMENTO
FDL Fala Direta Livre	**EDL** Escrita Direta Livre	**PDL** Pensamento Direto Livre
FD Fala Direta	**ED** Escrita Direta	**PD** Pensamento Direto
FIL Fala Indireta Livre	**EIL** Escrita Indireta Livre	**PIL** Pensamento Indireto Livre
FI Fala Indireta	**EI** Escrita Indireta	**PI** Pensamento Indireto
RAFN Representação de Ato de Fala pelo Narrador	**RAEN** Representação de Ato de Escrita pelo Narrador	**RAPN** Repr. de Ato de Pensamento pelo Narrador
NV Representação de Voz pelo Narrador	**NE** Representação da Escrita pelo Narrador	**NI** Narração Interna

Fonte: Barcellos, 2011, p. 38. Adaptado.

2.2.1 Categorias de apresentação da fala

A Fala Direta (FD) ocorre quando tudo aquilo que foi veiculado pela personagem é representado *ipsis litteris* e se observa a presença de uma oração introdutória de elocução e de aspas ou travessão.

<FD> *"You don't understand me, Harry,"* *answered the artist*. </FD>

(Wilde)

⬇

(oração introdutória de elocução. Grifos inseridos.)

A Fala Direta Livre (FDL) ocorre quando a fala da personagem é reportada *ipsis litteris*, porém sem a presença de uma oração introdutória de elocução ou de aspas ou travessão.

The painter laughed. <FDL> *"I don't think there will be any difficulty about that."* </FDL>

(Wilde)

A Fala Indireta (FI) é aquela em que se veicula a fala de uma personagem sem reproduzir de forma exata o que foi dito. Nela existe uma oração introdutória de elocução e, em se tratando da língua portuguesa, há ainda um conectivo típico de Fala Indireta.

<FI> ... disse **que** lhe escreveria de Paris, se não o encontrasse no clube. </FI>
(Do Rio. Grifos inseridos.)

A Fala Indireta Livre (FIL) ocorre quando o que foi dito pela personagem é veiculado de forma indireta, sem reproduzir exatamente o que foi dito. Nela não há uma oração introdutória de elocução e os verbos são conjugados no passado. Não houve ocorrência dessa categoria no *corpus* analisado.

A Representação do Ato de Fala pelo Narrador (RAFN) é um resumo daquilo que foi dito, sem que seja possível recuperar fielmente o conteúdo original da mensagem.

As the door was closing behind him, <RAFN> he called him back. </RAFN>
(Wilde)

2.2.2 Categorias de apresentação da escrita

A Escrita Direta (ED) ocorre quando há uma reprodução fiel da mensagem original escrita, observando-se, ainda, a presença de aspas ou travessão e de uma oração introdutória de elocução. Não foram encontrados exemplos dessa categoria no *corpus* analisado.

A Escrita Direta Livre (EDL) ocorre quando há uma reprodução fiel da mensagem original escrita e, ao mesmo tempo, observa-se a presença de aspas, mas sem o acompanhamento de uma oração introdutória de elocução.

Then he took down the Blue Book from one of the shelves and began to turn over the leaves. <EDL> *"Alan Campbell, 152, Hertford Street, Mayfair."* </EDL>
(Wilde)

A Escrita Indireta (EI) não tem um compromisso com a reprodução exata do conteúdo escrito; pode tanto apresentá-lo *ipsis litteris* quanto parafraseá-lo. Observa-se nela a presença de uma oração introdutória de elocução ou das aspas ou travessão. Em língua portuguesa, tem-se ainda o uso de um conectivo. Não houve ocorrência dessa categoria no *corpus* analisado.

A Escrita Indireta Livre (EIL), que retrata o posicionamento de uma personagem ante alguma leitura, ocorre quando não há uma oração introdutória de elocução. Não houve ocorrência dessa categoria no *corpus* analisado.

A Representação do Ato de Escrita pelo Narrador (RAEN) ocorre quando há um resumo do ato da escrita, sem veicular o conteúdo original da mensagem ou parafraseá-la.

<RAEN> *Finally, he went over to the table and wrote a passionate letter to the girl he had loved, imploring her forgiveness and accusing himself of madness.* </RAEN>

(Wilde)

A Representação de Escrita pelo Narrador (NE) ocorre quando há alusões mínimas ao ato de escrita. Não houve ocorrência dessa categoria no *corpus* analisado.

2.2.3 Categorias de apresentação do pensamento

O Pensamento Direto (PD) ocorre quando o que foi pensado é representado na íntegra e se observa a presença de uma oração introdutória de elocução e de aspas ou travessão. Não houve ocorrência dessa categoria no *corpus* analisado.

O Pensamento Direto Livre (PDL) ocorre quando o pensamento da personagem é representado na íntegra, mas sem a presença de aspas ou travessão. A oração introdutória de elocução pode não aparecer ou estar deslocada para o final da sentença. Na ausência dessa oração, o contexto fica a cargo de indicar se se trata ou não de um pensamento.

<PDL> *Words!* </PDL> <PDL> *Mere words!* </PDL>

(Wilde)

O Pensamento Indireto (PI) ocorre quando o pensamento da personagem é reportado de forma indireta e, em se tratando da língua portuguesa, também pode haver o uso de um conectivo. Segundo Barcellos (2011), essa categoria parece indicar o conteúdo proposicional do que teria sido pensado pela personagem.

<PI> A audiência provavelmente pensou que se tratava de um dueto. </PI>

(Goettems)

<PI> *The audience probably thought it was a duet*. </PI>

(Wilde)

O Pensamento Indireto Livre (PIL) ocorre sem a presença de uma oração introdutória de elocução e de aspas ou travessão. O pensamento é reportado de forma indireta e, no caso da língua portuguesa, há a presença de um conectivo. No *corpus* de estudo desta pesquisa, verificou-se que, geralmente, os verbos se encontram no futuro do pretérito e que, com frequência, trata-se de perguntas e questionamentos que pairam na mente da personagem.

> <PIL> *Should he move it aside, after all?* </PIL> <PIL> *Why not let it stay there?* <PIL> </PIL> *What was the use of knowing?* <PIL> </PIL> *If the thing was true, it was terrible.* <PIL> </PIL> *If it was not true, why trouble about it?* <PIL> </PIL> *But what if, by some fate or deadlier chance, eyes other than his spied behind and saw the horrible change?* <PIL> </PIL> *What should he do if Basil Hallward came and asked to look at his own picture?* <PIL> </PIL> *Basil would be sure to do that.* </PIL> <PIL> *No; the thing had to be examined, and at once.* </PIL> <PIL> *Anything would be better than this dreadful state of doubt.* </PIL>
>
> (Wilde)

A Representação do Ato de Pensamento pelo Narrador (RAPN) não é um resumo do pensamento, como ocorre na RAFN e na RAEN – que sinalizam um resumo da fala e da escrita de uma personagem, respectivamente. Consideram-se RAPN ocorrências ou alusões ao ato de pensamento. Segundo Barcellos (2011), a Representação do Ato de Pensamento pelo Narrador se diferencia da RAFN e da RAEN pelo fato de os atos de pensamento não serem de natureza comunicativa. Não foram, contudo, encontrados exemplos dessa categoria no *corpus* em análise.

A Narração Interna (NI) ocorre em referência a estados mentais e mudanças de fenômenos cognitivos e afetivos, não incluindo um ato de pensamento específico.

> <NI> *He felt intensely interested*. </NI> <NI> *He was amazed at the sudden impression that his words had produced.* </NI>
>
> (Wilde)

> <NI> *She was transfigured with joy.* </NI> <NI> *An ecstasy of happiness dominated her.* </NI>
>
> (Wilde)

2.2.4 Pesquisas sobre estilo e tradução

Pesquisas sobre estilo já foram realizadas nos estudos da tradução (*e.g.*, BAKER, 2000, 2004), e algumas delas usam como suporte o modelo de Semino e Short (2004). Um exemplo é Barcellos (2011), que analisou um *corpus* composto por *Heart of darkness*, de Joseph Conrad, e duas de suas traduções para o português. A autora observou, no texto-fonte e nas traduções, um predomínio de ocorrências de apresentação da fala, seguidas por apresentação do pensamento e apresentação da escrita, exatamente nesta ordem. Ela ainda encontrou, em todos os textos, mais ocorrências de Fala Direta Livre, seguidas, nesta ordem, de Fala Direta,

Representação do Ato de Fala pelo Narrador, Representação de Voz pelo Narrador, Fala Indireta e Fala Indireta Livre.

Em relação ao estilo dos tradutores, a autora não encontrou um padrão provável de mudança de uma categoria para outra, mas identificou exemplos de alterações de categorias do pensamento para categorias da fala e mudanças em sinais de pontuação (*e.g.*, ponto final por ponto de exclamação). A autora aponta que houve, em relação ao texto-fonte, um decréscimo sutil no número total de ocorrências das categorias nas traduções, o que sugere que, no texto-fonte, o leitor tem maior acesso ao que é dito pelas personagens. Em termos de razão *type/token* e comprimento médio das sentenças, a autora avaliou que um dos tradutores (Flaksman) tende a explicitar elementos do texto-fonte, enquanto o outro (O'Shea) tende a simplificar elementos do texto-fonte. A autora, contudo, não tece qualquer consideração à luz da hipótese da retradução, cabendo aqui apontar que, em sua pesquisa, ambas as traduções datavam, na impressão, de 2008.

A próxima seção apresenta algumas considerações sobre a obra *The picture of Dorian Gray* e suas (re)traduções para a língua portuguesa. Nessa oportunidade, também se relatam os resultados de uma pesquisa que abordou essa obra à luz da hipótese da retradução, porém sem filiação à Linguística de *Corpus*.

3 *The picture of Dorian Gray* e suas (re)traduções para a língua portuguesa

The picture of Dorian Gray (1890/1891), único romance do escritor irlandês Oscar Wilde, está inserido no cânone literário e existem inúmeras traduções, edições e adaptações dessa obra para o português (MILTON; VASCONCELLOS, 2003). Essa obra, que se tornou "símbolo da juventude intelectual 'decadente' da época e despertou grande polêmica por seu conteúdo homoerótico", tem sido descrita como "um dos clássicos modernos da Literatura Ocidental".[2]

Em um levantamento de abril de 2014 usando o Google e em consulta ao catálogo da Biblioteca Nacional[3], foram levantados os dados apresentados no Quadro 2. Vale destacar que a primeira tradução, de João do Rio, foi republicada pela editora Hedra em 2006. Além dessa e

[2] WIKIPEDIA. *O retrato de Dorian Gray*. Disponível em: <http://pt.wikipedia.org/wiki/O_Retrato_de_Dorian_Gray>. Acesso em: 30 abr. 2015.
[3] FUNDAÇÃO BIBLIOTECA NACIONAL. Disponível em: <http://www.bn.br>. Acesso em: 30 abr. 2014.

das demais edições brasileiras mostradas no quadro, também foi localizada uma edição portuguesa, de Maria de Lurdes Sousa Ruivo, publicada em 2000 pela Controljornal.

Quadro 2. Traduções e adaptações de *The picture of Dorian Gray*.

Tipo	Ano	Tradutor	Editora
TRADUÇÃO	1923	João do Rio	Garnier
	1965	Lígia Junqueira	Biblioteca Universal Popular
	1972	Oscar Mendes	José Aguilar
	1996	Maria Cristina F. da Silva	Nova Cultural
	1998	Marina Guaspari	PubliFolha
	2001	José Eduardo Ribeiro Moretzsohn	L&PM
	2006	Pietro Nassetti	M. Claret
	2012	Marcella Furtado	Landmark
	2012	Paulo Schiller	Penguin
	2013	Jorio Dauster	Globo Livros
	2014	Doris Goettems	Landmark
ADAPTAÇÃO	1974	Clarice Lispector	Ediouro
	1994	Cláudia Lopes	Scipione
	2005	Ana Carolina Vieira Rodriguez	Rideel
	2011	Douglas Tufano e Renata Tufano Ho	Paulus
	2011	Caio Riter	Escala Educacional
	2011	Stanislas Gros / Carol Bensimon	Quadrinhos na Cia.

O Quadro 2 também permite confirmar uma observação de Susam-Sarajeva (2003), qual seja, retraduções são, sim, realizadas em curto espaço de tempo entre uma e outra – talvez até mesmo quase que simultaneamente. É por essa razão que se propõe aqui a investigação de duas retraduções, que são duas das mais recentes até o momento de início desta pesquisa – as de Paulo Schiller (2012) e Doris Goettems (2014). Essas duas retraduções são, em termos cronológicos, relativamente próximas da reedição da primeira tradução, em 2006.

Essa obra já foi objeto de pesquisa orientada a investigações da retradução. Mais especificamente, Milton e Vasconcellos (2013) analisaram trechos selecionados de onze edições brasileiras[4] seguindo metodologia de cotejo entre os textos tal qual proposta por Berman (1990). Os autores observaram que "como esperado, nenhuma tradução foi uniforme o

[4] Cumpre salientar que Milton e Vasconcellos (2013) afirmam fazer a análise de onze edições de traduções para o português. Não foi possível identificar se por edições os autores entendem apenas traduções distintas ou também tiragens distintas das traduções de um mesmo autor. Como os autores não disponibilizam quais foram essas edições, não foi possível identificar por que obtiveram um número maior de publicações que aquele apresentado no Quadro 2, que contém dez publicações até 2013 (onze considerando a versão de 2014).

suficiente para caracterizar um tipo único de tendência – como domesticadora ou estrangeirizadora" e que "algumas traduções foram vítimas de apropriações indevidas por parte de algumas editoras" (MILTON; VASCONCELLOS, 2013, [s.p.]).

4. Metodologia

Nesta seção, descreve-se a metodologia de coleta e análise de dados empregada para este estudo. Conforme já mencionado, a metodologia se baseou em Dastjerdi e Mohammadi (2013). Contudo, foram implementadas algumas alterações, as quais são destacadas e justificadas no decorrer desta seção.

O *corpus* de estudo é composto por dez porções textuais aleatórias do texto-fonte, *The picture of Dorian Gray* 1890/1891 e de seus correspondentes na tradução da obra para o português, de João do Rio (1923), e em duas de suas retraduções mais recentes, uma de Paulo Schiller (2012) e a outra de Doris Goettems (2014). O *corpus* tem, portanto, pequenas dimensões, ou seja, menos de 80.000 palavras (BERBER-SARDINHA, 2003) – neste caso, conta-se especificamente com 29.893 palavras (pouco mais de 7.000 palavras por livro, conforme mostra a Tabela 1, na Seção 5). Os livros foram digitalizados e convertidos em .docx utilizando o programa ABBYY, para que pudessem ser analisados e etiquetados pelos pesquisadores. A identificação das porções correspondentes nas (re)traduções foi realizada de forma manual, embora com o aporte da função "Localizar" do MS Word, com vistas a uma busca por possíveis/prováveis traduções de palavras "chave" no texto-fonte. Todos os textos foram relidos após a conversão em .docx para assegurar, com o auxílio do corretor ortográfico do MS Word, que não houvesse erros resultantes da conversão dos arquivos.

A escolha dessas dez porções textuais foi realizada por amostragem aleatória. Seguindo a metodologia adotada em Nunes (2010) e em Lima (2013), utilizou-se a fórmula "*randbetween*" do programa MS Excel para uma seleção aleatória de páginas do texto-fonte até que se obtivessem aproximadamente 1.000 *tokens* (*i.e.*, 1.000 palavras consecutivas). Em caso de parágrafos incompletos, admitiu-se a inclusão de porções da página imediatamente anterior e da página imediatamente posterior àquela definida pela fórmula.

Optou-se por essa amostragem, em vez de selecionar capítulos aleatórios conforme fizeram Dastjerdi e Mohammadi (2013), em razão de haver capítulos longos e diferenças substanciais entre os comprimentos dos capítulos em *The picture of Dorian Gray*. Além disso, apesar de a amostragem ter sido superior em número de *tokens* àquela adotada por Dastjerdi e

Mohammadi (2013), que contaram com pouco mais de 5.000 palavras para cada livro, cabe sublinhar (i) que se optou por um *corpus* de pequena dimensão em função da necessidade de etiquetagem manual das categorias de apresentação do discurso e (ii) que o valor adotado se baseia em uma ampliação do número mínimo de 1.000 e 3.000 *tokens* sugeridos, respectivamente, por Lima (2013) e por Biber (1990) para que se possa tecer considerações sobre tendências de repetição de alguns padrões representativos em um mesmo tipo de texto.

Quadro 3. Etiquetas para o rastreamento da apresentação do discurso no *corpus*.

Tipo	Nome da Categoria	Identificação da Categoria	Etiqueta (início da ocorrência)	Etiqueta (final da ocorrência)
FALA	Fala Direta Livre	FDL	<FDL>	</FDL>
	Fala Direta	FD	<FD>	</FD>
	Fala Indireta Livre	FIL	<FIL>	</FIL>
	Fala Indireta	FI	<FI>	</FI>
	Representação de Ato de Fala pelo Narrador	RAFN	<RAFN>	</RAFN>
	Representação da Voz pelo Narrador	NV	<NV>	</NV>
ESCRITA	Escrita Direta Livre	EDL	<EDL>	</EDL>
	Escrita Direta	ED	<ED>	</ED>
	Escrita Indireta Livre	EIL	<EIL>	</EIL>
	Escrita Indireta	EI	<EI>	</EI>
	Representação de Ato de Escrita pelo Narrador	RAEN	<RAEN>	</RAEN>
	Representação da Escrita pelo Narrador	NE	<NE>	</NE>
PENSAMENTO	Pensamento Direto Livre	PDL	<PDL>	</PDL>
	Pensamento Direto	PD	<PD>	</PD>
	Pensamento Indireto Livre	PIL	<PIL>	</PIL>
	Pensamento Indireto	PI	<PI>	</PI>
	Repr. de Ato de Pensamento pelo Narrador	RAPN	<RAPN>	</RAPN>
	Narração Interna	NI	<NI>	</NI>

Fonte: Barcellos, 2011, p. 67.

Em seguida, procedeu-se à etiquetagem do *corpus* com base nas categorias de apresentação do discurso propostas por Semino e Short (2004), em versão mais atualizada que aquela de Short (1996) adotada por Dastjerdi e Mohammadi (2013), os quais analisaram apenas a apresentação da fala. Para tanto, ampliou-se o referencial também para a apresentação da escrita e do pensamento, e adotaram-se, conforme exibido no Quadro 3, as siglas utilizadas por Barcellos (2011) para compor as etiquetas que sinalizam as ocorrências dessas categorias no *corpus*.

Utilizando-se como referencial o nível da sentença e adotando-se parênteses angulares como marcadores, foram inseridas etiquetas antes (*i.e.*, "< >") e depois (*i.e.*, "</ >") de cada

ocorrência. No caso de ocorrência de orações introdutórias, as etiquetas delimitaram tanto essas orações quanto o conteúdo da fala, da escrita ou do pensamento das personagens.

Apresentam-se, a seguir, alguns exemplos de etiquetagem retirados do *corpus*. Compete apontar que houve inserção de espaços, tabulações e outras etiquetas, além destas, para satisfazer condições de utilização do programa. Um exemplo consiste nas reticências, que, para evitar contagem equivocada de número de sentenças, foram introduzidas entre parênteses angulares: <rf> ... </rf>.

Exemplo 1:
<FD> *"I am in Lady Agatha's black books at present," answered Dorian.* </FD>
(Wilde)

Exemplo 2:
<PI> *He remembered wandering through dimly lit streets, past gaunt, black-shadowed archways and evil-looking houses. Women with hoarse voices and harsh laughter had called after him. Drunkards had reeled by, cursing and chattering to themselves like monstrous apes.* </PI>
(Wilde)

No Exemplo 1, o trecho extraído do *corpus* está identificado com a etiqueta "FD", referente a uma "Fala Direta". A etiqueta <FD> indica o início da ocorrência; e a etiqueta </FD>, o fim dela. No Exemplo 2, o trecho extraído do *corpus* está identificado com a etiqueta "PI", que indica "Pensamento Indireto". A etiqueta <PI> indica o início da ocorrência; e a etiqueta </PI>, o fim dela.

A etiquetagem iniciou-se pelo texto-fonte. Uma vez concluída essa etapa, procedeu-se à etiquetagem, nesta ordem, da primeira tradução, da retradução de Schiller e da retradução de Goettems. Em seguida, com o suporte da ferramenta *on-line* YouAlign[5], os textos (re)traduzidos etiquetados foram alinhados, cada um, com o texto-fonte previamente etiquetado. Os arquivos gerados com o alinhamento serviram de insumo para um cotejo entre as etiquetas e a identificação de possíveis discrepâncias nas análises. Em seguida, todos os quatro textos foram alinhados manual e simultaneamente entre si e procedeu-se, mais uma vez, a um cotejo para a identificação de possíveis erros nas análises. Todas as etapas foram conduzidas isoladamente pela autora principal, cabendo ao segundo autor revisar os arquivos ao final de cada etapa. Em reuniões, ambos os pesquisadores discutiam discrepâncias e buscavam consenso em casos que

[5] YouAlign. Disponível em: <http://www.youalign.com>. Acesso em: 10 abr. 2015.

gerassem dúvidas ou divergências. Todas os problemas identificados – marcados com *highlights* ou por meio de comentários do MS Word – foram sanados, e as alterações necessárias foram realizadas tanto nos arquivos com os textos alinhados quanto nos arquivos com os textos isolados. Ao cabo dessa etapa, todos os arquivos, livres de *highlights* e comentários, foram convertidos em .txt para que pudessem ser processados pelo WordSmith Tools (versão 5.0).

O Quadro 4 apresenta, lado a lado, algumas ocorrências do *corpus*. Adotando-se esse tipo de alinhamento antes do processamento do *corpus* pelo programa WordSmith Tools, foi possível ter uma ideia inicial das diferenças entre os textos e, igualmente, identificar erros de categorização resultantes de adoção equivocada de um critério para classificar determinada ocorrência e de outro critério para classificar ocorrência similar em outro texto.

Compete salientar que o cotejo manual supramencionado foi realizado com o intuito de evitar a identificação de diferenças entre os textos resultantes de erros de categorização. No entanto, todas as análises dos dados foram realizadas de forma (semi)automática com o aporte do programa WordSmith Tools (versão 5.0). Nesse ponto, há novamente uma diferença em relação ao trabalho de Dastjerdi e Mohammadi (2013): enquanto os autores realizaram análises e contagens manuais, o presente trabalho, por explicitamente se filiar à Linguística de *Corpus*, contou com o apoio computacional para o levantamento de todos os dados. Sendo assim, os arquivos .txt foram adicionados ao programa WordSmith Tools e processados pela ferramenta "Wordlist" para a identificação da relação *type/token* e do comprimento médio das sentenças e pela ferramenta "Concord" para a identificação das etiquetas com as categorias de apresentação do discurso. Quando da utilização da ferramenta "Wordlist", o programa foi configurado para ignorar as etiquetas, pois, do contrário, obter-se-ia um valor maior de *types* e *tokens*.

Quadro 4. Exemplo de anotação do *corpus* de análise.

WILDE (1891)	DO RIO (1923)	SCHILLER (2012)	GOETTEMS (2014)
His cool, white, flowerlike hands, even, had a curious charm. <RAFN> They moved, as he spoke, like music, and seemed to have a language of their own. </RAFN>	As mãos mesmo, suas mãos frescas e brancas, lembrando flores, possuíam um encanto curioso. Tal como a voz, elas pareciam musicais, pareciam ter uma linguagem particular. (SEM OCORRÊNCIA)	Também as mãos frias, brancas como flores, tinham um estranho charme. <RAFN> Elas se movimentavam à medida que ele falava, como música, e pareciam ter uma linguagem própria. </RAFN>	Até as mãos frias, brancas como flores, possuíam um estranho encanto. <RAFN> Elas se moviam enquanto ele falava, como música, e pareciam ter uma linguagem própria. </RAFN>
<FDL>You do anything in the world to gain a reputation. </FDL> <FDL> As soon as you have one, you seem to want to throw it away. </FDL>	<FDL> Revolveis o mundo para ganhar a reputação e, logo que a possuis, como que quereis desembaraçar-vos dela! </FDL>	<FDL> Fazem de tudo no mundo para adquirir uma reputação. </FDL> <FDL> Assim que a obtêm, parecem querer jogá-la fora. </FDL>	<FDL> Fazem qualquer coisa no mundo para adquirir uma reputação. </FDL> <FDL> Assim que a conseguem, parecem querer jogá-la fora. </FDL>
<FDL> But you never sat better. </FDL> <FDL> You were perfectly still. </FDL> <FDL> And I have caught the effect I wanted -- the half-parted lips and the bright look in the eyes. </FDL>	<FDL> E nunca posaste tão bem: estavas perfeitamente imóvel e eu colhi o efeito que buscava: os lábios semiabertos e os olhos iluminados. <rf>...<rf> </FDL>	<FDL> Mas você nunca posou melhor. </FDL> <FDL> Esteve perfeitamente imóvel. </FDL> <FDL> E eu capturei o efeito que desejava -- os lábios entreabertos, e o brilho nos olhos. </FDL>	<FDL> Mas você nunca posou melhor.</FDL> <FDL> Esteve perfeitamente imóvel. </FDL> <FDL> E eu capturei o efeito que queria, os lábios entreabertos e o brilho no olhar. </FDL
<NI> He hesitated for a moment, then he turned back and took it from the table. </NI>	<NI> Hesitou um instante, depois entrou de novo e apanhou-a sobre a mesa. </NI>	<NI> Hesitou por um momento, em seguida voltou e a tirou da mesa. </NI>	Ele voltou, e pegou-a de cima da mesa. (SEM OCORRÊNCIA)
<FDL> "I get you not to go." </FDL>	<FDL> -- Suplico-te! </FDL>	<FDL> "Eu imploro." </FDL>	<FDL> "Eu lhe imploro." </FDL>
<NI> Lord Henry elevated his eyebrows and looked at him in amazement. </NI>	<NI> Lorde Henry abriu mais os olhos, fitando-o com surpresa. </NI>	<NI> Lord Henry ergueu as sobrancelhas e olhou para ele espantado. </NI>	<NI> Lorde Henry ergueu as sobrancelhas e olhou para ele espantado. </NI>

A considerar a "hipótese da retradução", esperava-se: (i) que houvesse, entre as duas retraduções recentes, similaridade em relação a todos os aspectos analisados; e (ii) que os aspectos analisados fossem similares mais entre essas duas retraduções e o texto-fonte do que entre a primeira tradução e o texto-fonte. Partiu-se do pressuposto de que, confirmando-se esses dois pontos, seriam encontrados indícios a favor da "hipótese da retradução".

5. Análise de dados e discussão dos resultados

Utilizando a função *Wordlist* do WordSmith Tools, foi possível identificar, automaticamente, a relação *type/token* e o comprimento médio dos parágrafos. Esses, dentre outros dados automáticos, estão disponibilizados na Tabela 1.

Tabela 1. Dados gerais do *corpus* para *types*, *tokens* e sentenças.

Parâmetro \ Autor	Wilde (1891)	Do Rio (1923)	Schiller (2012)	Goettems (2014)
Tokens	7.898	7.105	7.466	7.424
Types	1.622	2.224	1.974	1.954
Razão *type/token* normal	20,55	31,34	26,45	26,36
Razão *type/token* padronizada	40,71	50,84	46,17	46,34
Nº sentenças	744	687	735	735
Comprimento médio das sentenças	10,62	10,34	10,17	10,10

As tendências no *corpus*, extraídas por meio do programa WordSmith Tools, indicam similaridades entre o texto-fonte e as duas retraduções recentes em todos os parâmetros expostos na Tabela 1. Igualmente, sinalizam diferenças substanciais entre essas retraduções e a primeira tradução em todos os quesitos, à exceção do comprimento médio das sentenças.

Em termos qualitativos, o número de *tokens* sugere que todas as (re)traduções apresentaram tendências de implicitação, haja vista que houve uma redução de pelo menos 5,46% (Goettems) no número de palavras e de, no máximo, quase o dobro desse percentual (Do Rio). A implicitação, em casos simples, pode ocorrer, por exemplo, em razão da omissão do sujeito, como se observa em "*You were perfectly still*." (Wilde) e "Esteve perfeitamente imóvel." (Schiller). Em instâncias mais complexas, a implicitação parece estar atrelada a informações contextuais, como a tradução de "*I beg you not to go.*" (Wilde) por "Suplico-te!" (do Rio), "Eu imploro." (Schiller) e "Eu lhe imploro." (Goettems). Em todas essas instâncias de retradução, o contexto fornece pistas que permitem a implicitação daquilo que se implora que seja feito (*i.e.*, "*not to go*").

Em se tratando de *types*, bem como de razões *type/token* normal (*i.e.*, considerando a simples divisão de um pelo outro) e padronizada (*i.e.*, considerando a relação entre eles a cada 1.000 palavras), observa-se maior variedade lexical em todas as (re)traduções, com destaque para a primeira tradução, que tem aproximadamente 37% mais palavras diferentes que o texto-fonte e 13% a mais de palavras diferentes em relação às retraduções. O fato de a primeira tradução ser aquela com menor número de *tokens* realça ainda mais as diferenças: a sua razão *type/token* normal é 52,8% maior que a do texto-fonte e mais de 18% maior que as dos textos retraduzidos. Desconsiderando-se as diferenças no total de *tokens*, contudo, o percentual se reduz, embora as diferenças ainda sejam nítidas: a razão *type/token* padronizada da primeira tradução é 24,9% maior que a do texto-fonte e mais de 10% maior que a das retraduções.

Essa diferença da primeira tradução em relação ao texto-fonte e também em relação às duas retraduções recentes, no que diz respeito aos *types* e *tokens*, pode estar atrelada ao estilo

de cada autor ou (re)tradutor. No entanto, a proximidade entre o texto-fonte e as duas retraduções recentes e a diferença destas em relação à primeira tradução parecem ser indícios favoráveis à hipótese da retradução. Nesse caso, pode-se cogitar que a tradução de João do Rio (1923) buscou "ajustar" o texto-fonte às convenções ou tendências da língua portuguesa, como é o caso, por exemplo, de recomendações contra repetições de palavras nos textos.

Por sua vez, considerando-se o número de sentenças, nota-se que todas as (re)traduções apresentam algum decréscimo. Contudo, esse decréscimo é pequeno nas retraduções recentes (pouco mais de 1%), mas notório na primeira tradução (7,7%). No entanto, seguindo Dastjerdi e Mohammadi (2013) e considerando o comprimento médio das sentenças, as diferenças entre os textos, caso se considere um arredondamento, seria de uma palavra (ou aproximadamente 9%) por sentença entre as (re)traduções e o texto-fonte e de nenhuma palavra entre as (re)traduções.

Em princípio, considerando Dastjerdi e Mohammadi (2013), o comprimento médio das sentenças seria uma medida indicativa de estilos. Porém, à luz da já apontada tendência de implicitação e da drástica diferença entre os números de *tokens* em João do Rio e em Oscar Wilde, essa medida proporcional precisa ser revisitada. Em um olhar mais qualitativo do *corpus*, como no Exemplo 3, é possível notar diferenças substantivas entre a tradução de João do Rio e as demais retraduções.

Exemplo 3:
<FDL> *But you never sat better*. </FDL>
<FDL> *You were perfectly still*. </FDL>
<FDL> *And I have caught the effect I wanted--the half-parted lips and the bright look in the eyes*. </FDL>

(Wilde)

<FDL> E nunca posaste tão bem: estavas perfeitamente imóvel e eu colhi o efeito que buscava: os lábios semiabertos e os olhos iluminados. <rf>...<rf> </FDL>

(Do Rio)

<FDL> Mas você nunca posou melhor.</FDL>
<FDL> Esteve perfeitamente imóvel.</FDL>
<FDL> E eu capturei o efeito que desejava -- os lábios entreabertos, e o brilho nos olhos.</FDL>

(Schiller)

<FDL> Mas você nunca posou melhor.</FDL>
<FDL> Esteve perfeitamente imóvel. </FDL>
<FDL> E eu capturei o efeito que queria, os lábios entreabertos e o brilho no olhar.

</FDL

(Goettems)

Como o Exemplo 3 evidencia, ao contrário dos outros (re)tradutores, João do Rio condensa sentenças em uma única, o que, inclusive, tem impacto na contagem das categorias de apresentação do discurso, objeto da próxima análise. Se, por um lado, em exemplos como esse, o comprimento da sentença aumenta substancialmente, em outros ocorre o contrário, como mostra o Exemplo 4.

Exemplo 4:
<NI> *Lord Henry elevated his eyebrows and looked at him in amazement.* </NI>
(Wilde)

<NI> Lorde Henry abriu mais os olhos, fitando-o com surpresa </NI>
(Do Rio)

<NI> Lord Henry ergueu as sobrancelhas e olhou para ele espantado </NI>
(Schiller)

<NI> Lorde Henry ergueu as sobrancelhas e olhou para ele espantado </NI>
(Goettems)

Nesse exemplo, contam-se onze palavras no texto-fonte e doze palavras em cada uma das retraduções, mas nove palavras na primeira tradução. A razão disso está na implicitação que João do Rio faz ao traduzir "and looked at him in amazement" como "fitando-o com surpresa". A implicitação, nesse caso, consiste na condensação de informações por meio do gerúndio, que não só apaga marcas de "quem" e "quando" fitou alguém, mas sobretudo não deixa clara a relação dessa oração com a anterior, ou seja, não informa se o ato de fitar foi, por exemplo, uma causa, uma finalidade, uma consequência ou uma explicação para o fato de o personagem Lorde Henry ter aberto os olhos.

Em um balanço do aumento e da redução do número de *tokens* e do número de sentenças, parece que o saldo médio é nulo, ou seja, o comprimento médio das sentenças permanece quase que idêntico entre as (re)traduções. Sendo assim, parece que, em vez do comprimento médio das sentenças, talvez o número de sentenças fosse uma medida mais representativa daquilo que de fato ocorreu no *corpus* em tela. A considerar, portanto, o comprimento médio das sentenças, não se encontram indícios a favor da hipótese da retradução; contudo, ao levar em conta o número de sentenças, os indícios parecem apontar que João do

Rio apresenta escolhas léxico-gramaticais mais orientadas para a cultura-alvo, haja vista que, em geral, a língua portuguesa admite mais facilmente sentenças mais longas do que a língua inglesa, sem que isso necessariamente afete a facilidade de leitura do texto – argumento utilizado por Dastjerdi e Mohammadi (2013) para justificar a redução do comprimento médio das sentenças na primeira tradução de *Pride and prejudice* para o persa.

Conforme já aludido, as diferentes escolhas do tradutor e dos (re)tradutores no que diz respeito às palavras e às sentenças também tiveram impacto na contagem das categorias de apresentação do discurso, uma vez que se adotou a sentença como nível de análise. Por essa razão, no Exemplo 3, observam-se três ocorrências de Fala Direta Livre (FDL) no texto-fonte e nas retraduções recentes, mas apenas uma ocorrência na primeira tradução.

A Tabela 2 sintetiza os dados encontrados para a apresentação do discurso no texto-fonte de Oscar Wilde, na primeira tradução (de João do Rio) e em duas das retraduções recentes (de Schiller e de Goettems). Ela aponta que, dentre as categorias de apresentação do discurso, aquelas referentes à fala são predominantes no *corpus* (71,2%), seguidas por aquelas referentes ao pensamento (27,8%) e à escrita (1%). Esses dados são condizentes com aqueles apresentados por Barcellos (2011) em sua pesquisa com *Heart of darkness*, cabendo lembrar que, em razão de diferenças entre os estilos dos autores e de cada obra original, não necessariamente se esperaria essa convergência de resultados. Tanto é verdade que, por exemplo, no caso das categorias de apresentação da fala, Dastjerdi e Mohammadi (2013) encontraram mais ocorrências de Representação da Fala pelo Narrador, seguidas, nesta ordem, por ocorrências de Fala Direta e Fala Direta Livre. Sendo assim, o narrador exerce maior controle sobre o que as personagens falam em *Pride and prejudice* do que em *The picture of Dorian Gray.*

A Tabela 2 expõe, ainda, que, no total e por categorias, não houve diferenças entre os quatro textos na apresentação da escrita. Em contrapartida, o número de ocorrências de apresentação da fala e do pensamento foi menor nas (re)traduções que no texto-fonte, com exceção da retradução de Goettems no que diz respeito à apresentação do pensamento. Essa redução geral no número de ocorrências em boa medida se deve à redução no número de sentenças tanto na primeira tradução quanto nas retraduções em relação ao texto-fonte. Além disso, como as reduções se deram em todas as categorias de fala, não é possível cogitar que houve um aumento ou redução no aparente controle do narrador em relação ao que as personagens falam ou pensam.

Tabela 2. Apresentação do discurso no texto-fonte, na primeira tradução e em duas retraduções.

Parâmetro \ Autor		Wilde (1891)	Do Rio (1923)	Schiller (2012)	Goettems (2014)
FALA	FDL	363	330	351	356
	FD	56	48	55	55
	FIL	–	–	–	–
	FI	–	–	–	–
	RAFN	4	3	4	4
	NV	–	–	–	–
	TOTAL	**423**	**381**	**410**	**415**
ESCRITA	EDL	1	1	1	1
	ED	–	–	–	–
	EIL	–	–	–	–
	EI	–	–	–	–
	RAEN	3	3	3	3
	NE	2	2	2	2
	TOTAL	**6**	**6**	**6**	**6**
PENSAMENTO	PDL	7	14	6	6
	PD	–	–	–	–
	PIL	70	53	70	79
	PI	5	5	5	5
	RAPN	8	8	8	8
	NI	72	67	72	67
	TOTAL	**162**	**147**	**161**	**165**
TOTAL GERAL		**591**	**534**	**577**	**586**

No entanto, observa-se que as maiores diferenças percentuais ocorrem na primeira tradução em relação tanto ao texto-fonte quanto às duas retraduções mais recentes, que estariam mais similares ao texto de Wilde, no que diz respeito à apresentação da fala e da escrita. Já no que tange à apresentação do pensamento, merecem destaques as diferenças encontradas tanto em João do Rio, primeiro tradutor, quanto em Doris Goettems, última (re)tradutora.

Em João do Rio, conforme já apontado, a redução no número total de ocorrências de apresentação da fala em relação ao texto-fonte está atrelada às escolhas do tradutor por aglutinar sentenças e, em alguns casos, implicitar informações. Somente na apresentação do pensamento é que se observa uma tendência do tradutor por escolhas estilísticas que implicam menor controle aparente do narrador em relação ao que as personagens pensam. Em termos quantitativos, o primeiro tradutor dobra o número de ocorrências de Pensamento Direto Livre (PDL) em relação ao texto-fonte (catorze contra sete) enquanto os (re)tradutores diminuem uma ocorrência (seis cada). O Exemplo 5 ilustra o ocorrido.

Exemplo 5:

<PDL> *Words!* </PDL> <PDL> *Mere words!* </PDL> <PIL> *How terrible they were!* </PIL> <PIL> *How clear, and vivid, and cruel! </PIL> <PIL>One could not escape from them.* </PIL> <PIL> *And yet what a subtle magic there was in them!* </PIL> *They seemed to be able to give a plastic form to formless things, and to have a music of their own as sweet as that of viol or of lute.* <PDL> *Mere words!* </PDL> <PIL> *Was there anything so real as words?* </PIL>

(Wilde)

<PDL> As palavras! </PDL> <PDL> As simples palavras! </PDL> <PDL> Como são terríveis! </PDL> <PDL> Quantas são límpidas, fulgurantes ou cruéis! </PDL> <PDL>Bem quiséramos evitá-las! </PDL> <PDL> No entanto, que sutil magia há nelas?. <rf>...<rf> </PDL> <PDL> Dir-se-ia que dão uma forma plástica às coisas informes e que possuem uma música própria, tão doce como a do alaúde ou a de um violino! </PDL> <PDL> As simples palavras! </PDL> <PDL> Que há de mais real que as palavras? </PDL>

(Do Rio)

<PDL> Palavras! </PDL> <PDL> Simples palavras! </PDL> <PIL> Como eram terríveis! </PIL> <PIL> Como eram claras, vívidas e cruéis! </PIL> <PIL> Não era possível fugir delas.</PIL> <PIL> E no entanto que magia sutil havia nelas! </PIL> Pareciam capazes de propiciar uma forma plástica a coisas sem forma, e de ter uma música própria, doce como a da viola ou da flauta. <PDL> Simples palavras! </PDL> <PIL> Existiria algo tão real quanto as palavras? </PIL>

(Schiller)

<PDL> Palavras! </PDL> <PDL> Simples palavras! </PDL> <PIL> Como eram terríveis! </PIL> <PIL> Como eram claras, vívidas e cruéis! </PIL> <PIL> Não se podia escapar delas. </PIL> <PIL> E, no entanto, que magia sutil havia nelas! </PIL> Pareciam capazes de dar uma forma plástica a coisas sem forma, e de ter uma música própria, doce como a da viola ou do alaúde. <PDL>Simples palavras! </PDL> <PIL> Existiria alguma coisa tão real quanto as palavras? </PIL>

(Goettems)

Como se nota no Exemplo 5, João do Rio é o único que adota o tempo presente para apresentar o pensamento das personagens. Tanto Wilde quanto os demais (re)tradutores empregam o tempo passado. A implicação disso é uma diminuição no aparente controle do narrador no que diz respeito ao pensamento das personagens. O leitor tem, com isso, a impressão de que tem acesso direto à integralidade do pensamento da personagem.

Doris Goettems, por sua vez, também parece diminuir o aparente controle do narrador em relação ao que as personagens pensam. Contudo, a (re)tradutora o faz, de um lado, reduzindo as ocorrências de Narração Interna (67 contra 72 no texto-fonte) e, de outro, criando ou

dividindo sentenças e, por conseguinte, aumentando as ocorrências de Pensamento Indireto Livre (79 contra 70 no texto-fonte). Os Exemplos 6 e 7 ilustram essa observação.

Exemplo 6:
<RAPN> *Then he remembered the lamp.* </RAPN> *It was a rather curious one of Moorish workmanship, made of dull silver inlaid with arabesques of burnished steel, and studded with coarse turquoises.* <PIL> *Perhaps it might be missed by his servant, and questions would be asked.* </PIL> **<NI>** ***He hesitated for a moment, then he turned back and took it from the table.*** **</NI>** *He could not help seeing the dead thing.* <PIL> *How still it was!* </PIL> <PIL> *How horribly white the long hands looked!* </PIL> <PIL> *It was like a dreadful wax image.* </PIL>

(Wilde. Grifos inseridos.)

<RAPN>Então se lembrou da lamparina. </RAPN> Era um modelo bastante curioso de manufatura mourisca, feita de prata fosca incrustada com arabescos de aço polido. <PIL> Talvez o criado desse por sua falta, e haveria perguntas. </PIL> **Ele voltou, e pegou-a de cima da mesa.** <PIL> Como o homem estava imóvel! </PIL> Como as longas mãos pareciam horrivelmente brancas! </PIL> <PIL> Era como uma assustadora imagem de cera. </PIL>

(Goettems. Grifos inseridos.)

Pelo que se observa no Exemplo 6, Goettems omite "He hesitated for a moment", o que implicou a diminuição de PIL em uma ocorrência. Nesse caso específico, o que se tem não é sequer uma diminuição no aparente controle do narrador em relação ao pensamento da personagem, mas sim um total inacesso ao pensamento da personagem pelo leitor.

Exemplo 7:
<PIL> No wonder Basil Hallward worshipped him. </PIL>

(Wilde)

<PIL> Como surpreender que Basil Hallward o estimasse de tal forma? <rf>...<rf> <PIL>

(Do Rio)

<PIL> Não era de admirar que Basil Hallward o venerasse. </PIL>

(Schiller)

<PIL> Não era de se admirar que Basil Hallward o adorasse. </PIL> <PIL> Ele fora feito para ser adorado. </PIL>

(Goettems)

Conforme se observa no Exemplo 7, Goettems desenvolve uma sentença nova que, em tese, não apresenta correspondente formal no texto-fonte, mas mantém a mesma categoria de

apresentação do discurso. Isso justificou o aumento de ocorrências de Pensamento Indireto Livre (PIL).

Tendo em vista esses resultados para a apresentação do discurso, não se observa, tal qual Barcellos (2011) na obra de Joseph Conrad, um padrão provável de mudança de uma categoria para outra. No entanto, enquanto a autora associa um decréscimo sutil no número total de ocorrências das categorias nas traduções a um menor acesso ao que é dito pelas personagens no texto-fonte, o presente trabalho parece apontar que o decréscimo no total de ocorrências está mais atrelado à opção metodológica desta pesquisa por identificar categorias através de sentenças. Em outras palavras, o menor número de ocorrências das categorias de apresentação da fala em todos os textos (re)traduzidos e das categorias de apresentação do pensamento na primeira tradução e na retradução de Schiller está relacionado a opções por sintetização de sentenças e implicitação ou, nos termos de Barcellos (2011), simplificação de elementos no texto-fonte.

6. Considerações finais

Conforme apresentado ao final da seção de metodologia, em uma releitura do objetivo geral explicitado na Introdução, este estudo, tomando como ponto de partida a "hipótese da retradução", esperava: (i) que houvesse, entre as duas retraduções recentes, similaridade em relação a todos os aspectos analisados; e (ii) que os aspectos analisados fossem similares mais entre essas duas retraduções e o texto-fonte do que entre a primeira tradução e o texto-fonte. Partiu-se do pressuposto de que, confirmando-se esses dois pontos, seriam encontrados indícios a favor da "hipótese da retradução".

Após a análise dos dados, tanto dos tipos de apresentação do discurso como das estatísticas gerais fornecidas pelo programa WordSmith Tools, constatou-se que existe uma similaridade muito maior entre o texto-fonte e duas das (re)traduções mais recentes da obra do que entre o texto-fonte e sua primeira tradução. Portanto, pode-se afirmar que, nesta análise, encontraram-se, tal qual em Dastjerdi e Mohammadi (2013), indícios a favor da "hipótese da retradução". Esses indícios foram identificados em termos de relação *type/token*, número de sentenças (em substituição à proposta original de avaliar o comprimento médio das sentenças) e um dos três tipos de apresentação do discurso (aquele referente à fala). Sendo assim, no geral, este trabalho – ainda que com base em parâmetros distintos – aponta para perspectivas distintas daquelas sugeridas por Milton e Vasconcellos (2013), segundo os quais não foi possível

caracterizar um tipo único de tendência (*i.e.*, domesticadora ou estrangeirizadora) entre onze traduções da obra *The picture of Dorian Gray.*

O único parâmetro, no presente caso, que parece não subsidiar a hipótese da retradução diz respeito à apresentação do pensamento, haja vista que a retradução mais recente (a de Goettems) apresentou um padrão diferente daquele registrado no texto-fonte. Nesse sentido, a presente proposta de analisar categorias de apresentação do discurso que não apenas aquelas referentes à fala, tal qual feito por Dastjerdi e Mohammadi (2013), parece ser mais robusta, por cercear o *corpus* em mais de uma frente e apontar novas vertentes. Além disso, as observações feitas em relação ao comprimento médio das sentenças podem servir de alerta para futuras pesquisas caso encontrem significativas reduções tanto no número de *tokens* quanto no número de sentenças de uma ou mais (re)traduções em relação ao texto-fonte.

Contudo, este trabalho não tem a intenção de ser conclusivo, cabendo aqui registrar suas limitações e apontar sugestões para pesquisas futuras. As análises ora apresentadas partiram de um *corpus* de pequena dimensão, o qual foi compilado com base em porções aleatoriamente selecionadas de uma obra e em relação ao qual se adotaram apenas alguns dentre os muitos critérios que poderiam ser empregados para a comparação entre os textos. Portanto, não foram esgotadas as possiblidades para o tema aqui abordado, seja sob o viés da hipótese da retradução, da Linguística de *Corpus*, do estilo do tradutor (cf. BARCELLOS, 2011; BAKER, 2000, 2004) ou das inúmeras categorias ou parâmetros de análise que podem ser adotados, como sugerem Koskinen e Paloposki (2010).

Compete apontar que, embora seja alta a probabilidade de as tendências registradas naquelas categorias que apresentam maior número de ocorrências serem aquelas presentes no texto-fonte como um todo, muito provavelmente categorias menos representativas poderiam ser mais bem contempladas quando da análise de todo o texto, haja vista a reduzida probabilidade de serem encontradas nos poucos excertos selecionados do texto-fonte. Além disso, apesar de o modelo de Semino e Short (2004) ser robusto o suficiente, nem sempre as categorias são facilmente aplicáveis no caso concreto e, portanto, é possível que, em um ou outro ponto, pesquisadores diferentes façam classificações diferentes das mesmas ocorrências. Contudo, acredita-se que a metodologia de validação das etiquetas desse *corpus* foi minimamente robusta o suficiente para garantir resultados confiáveis.

Novos estudos podem ser feitos para confirmar os resultados aqui apresentados considerando o texto-fonte e as (re)traduções na íntegra, bem como in*corpora*ndo outras

(re)traduções e até mesmo outros textos-fonte. O importante é que tanto o fenômeno da retradução quanto o arcabouço teórico-metodológico da Linguística de *Corpus* sejam explorados o máximo possível para testar hipóteses e descrever textos os mais diversos possíveis.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) a bolsa de iniciação científica concedida à primeira autora. Também agradecem a Carolina Barcellos a prontidão em ajudá-los na replicação de sua metodologia de etiquetagem do *corpus*. Igualmente expressam seus agradecimentos aos pareceristas, que contribuíram para a qualidade da versão final deste artigo.

Referências Bibliográficas

ASSIS, R. C. **A representação de europeus e de africanos como atores sociais em Heart of darkness (O coração das trevas) e em suas traduções para o português:** uma abordagem textual da tradução. 2009. 267 f. Tese (Doutorado em Estudos Linguísticos) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2009.

BAKER, M. Towards a methodology for investigating the style of literary translator. **Target**, v. 12, p. 241-266, 2000.

BAKER, M. A *corpus*-based view of similarity and difference in translation. *International* **Journal of *Corpus* Linguistics**. Amsterdã: John Benjamins, v. 9, n. 2, p. 167-193, 2004.

BARCELLOS, C. P. **O estilo dos tradutores:** apresentação do discurso no *corpus* paralelo Heart of Darkness/ (No) Coração das Trevas. 2011. 154 f. Dissertação (Mestrado em Estudos Linguísticos) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte. 2011.

BENSIMON, P. Presentation. **Palimpsestes**, v. 4, p. ix-xiii, 1990.

BERBER SARDINHA, T. Tamanho de *corpus*. **The ESPecialist**, São Paulo, v. 23, n. 2, p. 103-122, 2003.

BERMAN, A. La retraduction comme espace de traduction. **Palimpsestes***,* Paris, v. 13, p. 1-7, 1990.

BERMAN, A. **A tradução e a letra**. Rio de Janeiro: UFSC, 1999.

BIBER, D. M. Methodological issues regarding *corpus*-based analyses of linguistic variation. **Literary and Linguistic Computing**, Oxford, v. 5, n. 4, p. 257-269, 1990.

BROWNLIE, S. Narrative theory and retranslation theory. **Across Languages and Culture**, v. 7, p. 140-170, 2006.

CHESTERMAN, A. A casual model for translation studies. In: OLOHAN, M. (ed.). **Intercultural faultlines**. Manchester: St. Jerome, 2000. p. 15-27.

DASTJERDI, H. V.; MOHAMMADI, A. Revisiting "retranslation hypothesis": a comparative analysis of stylistic features in the Persian retranslations of "Pride and prejudice". **Open Journal of Modern Linguistics**, v. 3, n. 3, p. 174-181, 2013.

GAMBIER, Y. La retraduction, re tour et de tour. **Meta**, v. 39, p. 413-417, 1994.

KOSKINEN, K.; PALOPOSKI, O. Reprocessing texts: the fine line between retranslating and revising. **Across Languages and Cultures**, v. 11, p. 29-49, 2010.

KOSKINEN, K.; PALOPOSKI, O. Thousand and one translations: revisiting retranslation. In: HANSEN, G.; MALMKJAER, K.; GILE, D. (ed.). **Claims, changes and challenges**. Amsterdã: John Benjamins, 2004. p. 27-38.

LEECH, G.; SHORT, M. **Style in fiction:** a linguistic introduction to English fictional prose. 1. ed. Londres: Longman, 1983.

LIMA, K. C. S. **Caracterização de registro orientada para a produção textual no ambiente bilíngue:** um estudo baseado em *corpora* comparáveis. 2013. 251 f. Dissertação (Mestrado em Estudos Linguísticos) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2 013.

MILTON, J. Translating classic fiction for mass markets. The Brazilian clube do livro. **The Translator**, v. 7, p. 43-69, 2001.

MILTON, J.; TORRES, M. (Org.). Apresentação. **Cadernos de Tradução**, Florianópolis, v. 11, v. especial sobre tradução, retradução e adaptação, n. 1, p. 9-17, 2003.

MILTON, J.; VASCONCELLOS, R. A. As traduções de "The picture of Dorian Gray". In: SIMPÓSIO INTERNACIONAL DE INICIAÇÃO CIENTÍFICA DA USP, 21., 22-24 out. 2013, São Carlos. **Anais...** São Paulo: USP, 2013. [on-line]. Disponível em: https://uspdigital.usp.br/siicusp/cdOnlineTrabalhoVisualizarResumo?numeroInscricaoTrabalho=827&numeroEdicao=21. Acesso em: 30 abr. 2015.

NUNES, L. P. **As conjunções but e mas em textos ficcionais originais e traduzidos:** uma abordagem tridimensional com base na Linguística Sistêmico-Funcional. 2010. 187 f. Dissertação (Mestrado em Estudos Linguísticos) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2010.

PYM, A. **Method in translation history**. Manchester: St. Jerome, 1998.

SCHULTE, R.; BIGUENET, J. **Theories of translation:** an anthology of essays from Dryden to Derrida. Chicago: The University of Chicago Press, 1992.

SEMINO, E.; SHORT, M. ***Corpus* stylistics:** speech, writing and thought presentation in a *corpus* of English writing. Londres: Routledge, 2004

SHORT, M. **Exploring the language of poems, plays and prose**. Londres: Longman, 1996.

SIMPSON, P. **Language, ideology and point of view**. Londres: Routledge, 1993.

SUSAM-SARAJEVA, Ş. Multiple-entry visa to travelling theory: retranslations of literary and cultural theories. **Target**, v. 15, n. 1, p. 1-36, 2003.

TAHIR-GÜRÇAĞLAR, Ş. Retradução. In: BAKER, M.; SALDANHA, G. **Routledge encyclopedia of translation studies**. Nova Iorque: Routdlege, 2009. p. 233-236.

TARVI, L. The problems of managing the 'translation stock': freelancing vs. freebooting. In: REINTJES, M., TÅQVIST, M. (ed.). **Ethics and politics of translation**. Norwich: University of East Anglia, 2005. p. 125-140.

VÁNDOR, J. **Adaptation and retranslation**. 2010. Tese (Doutorado em Estudos da Tradução) – Eötvös Lorand University, Budapeste, 2010.

VENUTI, L. **The translation studies reader**. Londres: Routledge, 2003.

VENUTI, L. **The translator's invisibility**. Nova Iorque: Routledge, 1995.

***Corpus* analisado**

DO RIO, J. **O retrato de Dorian Gray**. São Paulo: Martin Claret, 2009 [1923].

GOETTEMS, D. **O retrato de Dorian Gray**. São Paulo: Editora Landmark, 2014.

SCHILLER, P. **O retrato de Dorian Gray**. São Paulo: Companhia das Letras, 2012.

WILDE, O. **The picture of Dorian Gray**. Londres: Penguin Classics, 1891.

Artigo recebido em: 30.03.2015
Artigo aprovado em: 26.06.2015