ISSN 1980-5799

Domínios de Lingu@gem

Revista do Programa de Pós-Graduação em Estudos Linguísticos
Universidade Federal de Uberlândia

Tratamento Computacional do Português Brasileiro

Heliana Mello, Fernanda Farinelli
org.

Universidade Federal de Uberlândia

v. 16, n. 4
Outubro/Dezembro 2022

Domínios de Lingu@gem

Tratamento Computacional do Português Brasileiro

Organização: Heliana Mello (UFMG), Fernanda Farinelli (UnB)

4º Trimestre 2022
Volume 16, número 4
ISSN: 1980-5799

Expediente

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

Domínios de Lingu@gem, / Universidade Federal de Uberlândia, Instituto de Letras e Linguística, 2007-
V. 1 -

Trimestral.
ISSN: 1980-5799
Modo de acesso: http://www.seer.ufu.br/index.php/dominiosdelinguagem
A partir de 2020 a Revista é de responsabilidade do Programa de Pós- Graduação em Estudos Linguísticos

1. Linguística - Periódicos. 2. Linguística aplicada - Periódicos. I. Universidade Federal de Uberlândia. Instituto de Letras e Linguística. III. Universidade Federal de Uberlândia. Programa de Pós-Graduação em Estudos Linguísticos.

CDU 801(05)

Participaram da edição 52 como pareceristas *ad hoc*

Antonio Carlos Silvano Pessotti (UNICAMP)
Daniel Alves (UFPB)
Fabíola Aparecida Sartin Dutra Parreira Almeida (UFC)
João Antônio Moraes (UFRJ)
Odair Luiz Nadin (UNESP/Araraquara)
Rafaela Rigaud Peixoto (PUC/RJ)
Waldemar Ferreira Netto (USP)

Sumário

DOI: 10.14393/DL52-v16n4a2022-1

Tratamento computacional do português brasileiro

Computational treatment of Brazilian Portuguese

Heliana MELLO[*]
Fernanda FARINELLI[**]

1 O tratamento computacional das línguas naturais

O tratamento computacional de dados linguísticos tem estado na agenda de linguistas e cientistas da computação há no mínimo cinco décadas; entretanto, apenas nas últimas duas décadas tal movimento ganhou impulso no cenário brasileiro. Este movimento conta com a adesão de pesquisadores de diversas áreas do conhecimento, que progressivamente, através das novas tecnologias e formações acadêmicas mais sintonizadas com as necessidades do tratamento de línguas naturais via procedimentos computacionais, vão ganhando visibilidade.

É relevante que destaquemos aqui o quão importante a formação dos jovens graduandos, sobretudo na área de estudos linguísticos, esteja alinhada às pautas de pesquisa e inovações metodológicas que a área de tratamento computacional de línguas naturais exige. Por isso, somos fortes defensoras do ensino de programação e estatística na formação linguística e da promoção de interação com os conhecimentos oriundos das áreas informáticas e da computação.

A linguística de *corpus* de chegou ao Brasil há cerca de duas décadas, à época, com uma predominância de atuações voltadas para as subáreas do ensino de línguas estrangeiras, sobretudo língua inglesa, e estudos da tradução. Os estudiosos de

[*] Doutora em Linguística (CUNY, EUA), Professora Titular da Faculdade de Letras, UFMG. ORCID: https://orcid.org/0000-0003-0267-9005. hmello@ufmg.br
[**] Doutora em Ciência da Informação (ECI-UFMG), Professora Adjunta da Faculdade de Ciência da Informação, UnB. ORCID: https://orcid.org/0000-0003-2338-8872. fernanda.farinelli@unb.br

fraseologia e lexicografia também se envolveram com os estudos de *corpora* e formação de bancos de dados linguísticos.

No Brasil temos pesquisadores que poderiam ser chamadas de precursores da linguística de *corpus* e do tratamento computacional do português brasileiro (PB). Na frente da linguística, pesquisadores como Tony Berber Sardinha, Violeta Quental, Bento Dias da Silva, Leonel Figueiredo, dentre muitos outros; bem como as equipes do NILC USP-São Carlos, da UFRS e da PUC-RS, dentre outras, na frente da Ciência da Computação e da Informática, nos vêm imediatamente à mente quando tentamos traçar o percurso do tratamento computacional do PB.

A tradição criada pelos nossos pioneiros é generosa, com o compartilhamento de muitos *corpora* e ferramentas computacionais gratuitos e passíveis de serem utilizados pela comunidade livremente.

Para além dos pioneiros brasileiros, devemos mencionar o portal LINGUATECA, liderado por Diana Santos, que baseado em Portugal, tem disponibilizado gratuitamente, ao longo de décadas, tanto *corpora* quanto ferramentas computacionais para o estudo de diversas variedades do português, inclusive o PB. Agregando usabilidade e camadas de anotação a diversos *corpora* disponíveis na LINGUATECA, está o *parser* PALAVRAS, desenvolvido por Eckhard Bick (BICK, 2000).

A nossa intenção ao propor a temática deste volume à Revista Domínios de Lingu@gem foi a de oferecer aos leitores um mapa, mesmo que incompleto, da riqueza da produção da área e a necessidade cada vez mais urgente de formarmos jovens aptos a lidar com as grandes perguntas metodológicas, que se colocam para o linguista e o cientista da computação, e outras formações profissionais, que desejam acompanhar o inexorável fluxo do desenvolvimento tecnológico e seu impacto na pesquisa linguística. Todos nós, que trabalhamos na área, compartilhamos bases empíricas e

metodológicas complexas e múltiplas, que exigem o trabalho em colaboração e a troca rápida e eficiente de saberes.

O tratamento computacional das línguas naturais pode, grosso modo, ser subdivido em três grandes campos que não guardam, absolutamente, fronteiras rígidas entre si:[1] compilação e tratamento de *corpora*, estudos pautados pela linguística computacional (metodologias para o tratamento computacional de línguas naturais) e processamento de linguagem/língua natural. Esses três campos contam com profissionais de diversas áreas do saber, mas a colaboração entre linguistas e cientistas da computação é a interface de competências predominante. Para os dois primeiros campos, há uma primazia se não de liderança de linguistas, pelo menos de perguntas de pesquisa de ordem linguística, já que o foco principal, mesmo que intermediado por código e ferramentas computacionais, é linguístico; enquanto no terceiro, a liderança geralmente é de cientistas da computação, interessados cada vez mais em inteligência artificial e seus métodos, que exigem *big data* para extrair informações de línguas naturais. Este número temático seguirá essa divisão tripartite, admitidamente difusa, no grupamento de seus doze artigos.

A base comum às três subáreas mencionadas acima é, obviamente, a utilização de métodos computacionais para o tratamento das línguas naturais. Tal tratamento pode lidar com dados linguísticos não-estruturados (textos, como os conhecemos), semiestruturados (representação heterogênea, textos formatados via XML ou JSON, por exemplo) ou estruturados (em formato de tabela, como num banco de dados relacional). Igualmente, a intervenção humana no tratamento de tais dados varia entre dois polos: intervenção humana ativa (como no processo de anotação manual, mesmo que com utilização de uma interface computacional) e processo totalmente

[1] Essa subdivisão não é acatada por todos que atuam na área. Há visões distintas sobre o assunto e a nossa proposta parte de uma preocupação pedagógica sobre o tratamento computacional de línguas naturais, sem uma adesão rígida a qualquer corrente teórica da área. Ver Othero e Menuzzi (2005, p. 17) para uma visão que divide a linguística computacional entre linguística de corpus e PLN.

automatizado e conduzido pelas máquinas (como em aprendizado profundo ou redes neurais profundas).

O primeiro *corpus* linguístico computacional foi o Brown Corpus of Standard American English[2], compilado em 1961, a partir de uma variedade de gêneros textuais, contendo 1 milhão de palavras e, posteriormente publicado (1964) e apresentado, com aplicações, por Francis e Kučera (1967). Já este trabalho precursor sinalizava um caminho de muito interesse para a compilação de dados linguísticos habilitados ao tratamento computacional e à sua usabilidade para diversos tipos de análises linguísticas.

Outro *corpus* de importância histórica no cenário internacional é o London-Lund Corpus of Spoken British English[3], contendo cerca de 500 mil palavras (SVARTVIK 1990). Pela primeira vez, dados de fala foram coletados e compilados em um formato de *corpus*. Este é um *corpus* incluído no que se chama usualmente de "primeira onda da linguística de *corpus*", em que *corpora* orais disponibilizavam as transcrições da fala, ainda sem a possibilidade de se conectar o sinal sonoro à sua correspondente transcrição gráfica.

Outro passo significativo na área foi a ampliação nos tamanhos dos *corpora* compilados, passando-se à escala dos milhões e, posteriormente, bilhões de palavras. Num momento histórico em que *big data* tornou-se uma colação frequente em diversas áreas do saber, também nos estudos de *corpora* a escalada dos dados é algo que vem ocorrendo e que se faz necessário em diversos tipos de aplicações linguísticas, como no exemplo clássico da composição de dicionários, além de ser crucial para tarefas computacionais variadas, realizadas através de aprendizado de máquina na área de processamento de língua natural (PLN).

[2] http://icame.uib.no/brown/bcm.html
[3] http://korpus.uib.no/icame/manuals/LONDLUND/INDEX.HTM

A linguística de *corpus* ganhou, desde os seus anos iniciais, uma enorme variedade de métodos para a compilação de dados, os quais, se inicialmente eram representativos da língua escrita, ampliaram-se para incluir a língua falada e a gestualidade, havendo hoje os *corpora* multimodais, em que gestos, fala e transcrição são alinhados sincronamente, permitindo ao pesquisador uma experiência realisticamente empírica na análise de dados.

Para além de diamesias variadas, os *corpora* disponíveis atualmente cobrem um impressionante espectro de épocas históricas, áreas de especialização, línguas, propósitos de pesquisa, formatos, tipologias de anotação e escalabilidade.

Esta mesma tendência pode ser observada nos *corpora* dedicados ao estudo do PB, conforme os exemplos a seguir. O Corpus Brasileiro[4], traz textos de diversas tipologias, inclusive textos transcritos de fala, perfazendo 1 milhão de palavras, podendo ser acessado gratuitamente através do portal LINGUATECA[5] ou através do seu site próprio (ver nota 4). O portal de ferramentas e recursos do NILC[6] também oferece uma grande variedade de *corpora* escritos, além de ferramentas computacionais e bases de dados para PLN. O Corpus NILC, com mais de 40 milhões de palavras, é disponibilizado através do portal do NILC e também da LINGUATECA[7]. O PB conta hoje com *corpora* orais com alinhamento síncrono sinal sonoro-transcrição, como o C-ORAL-BRASIL[8] e o NURC Digital[9], além de *corpora* especializados voltados para a tradução, a fraseologia, a lexicografia, o ensino de línguas, dentre outras aplicações, como pode ser apreciado no Portal do Projeto COMET[10].

[4] http://corpusbrasileiro.pucsp.br/cb/Inicial.html
[5] https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS
[6] http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources
[7] https://www.linguateca.pt/acesso/corpus.php?corpus=SAOCARLOS
[8] https://www.c-oral-brasil.org/
[9] https://fale.ufal.br/projeto/nurcdigital/
[10] https://comet.fflch.usp.br/projeto

Passando agora, à linguística computacional, tomamos aqui a visão de Freitas (no prelo, p. 4)[11], que entende que o termo linguística computacional pode ser utilizado quando a dimensão linguística de uma pesquisa está em evidência, mesmo que o trabalho computacional que a subjaz esteja muito alinhado com PLN. Assim, a linguística computacional desenvolve métodos computacionais para responder a perguntas científicas da linguística. Caberiam aqui trabalhos que tentam formalizar e investigar, através de metodologias computacionais, questões relacionadas ao conhecimento linguístico, aquisição e uso linguístico, distribuição do léxico em estruturas linguísticas, gramáticas para dados de língua natural, dentre outros. Destacam-se neste escopo, adicionalmente, ferramentas computacionais desenvolvidas especificamente para atender às necessidades analíticas oriundas da linguística. Num certo sentido, como compartilhado por Jason Eisner[12], bons modelos em linguística computacional, que mais e mais deem conta da competência linguística humana, acarretarão melhores modelos em PLN, se estes se basearem nos frutos gerados pela pesquisa que leva a sério as perguntas fundamentais da investigação linguística.

Os avanços em linguística computacional internacionalmente, assim como aqueles em PLN, são tantos e tão rápidos, que não caberia discuti-los aqui. Dentro da visão que adotamos para a linguística computacional, é válido, entretanto, mencionar ferramentas ou conjuntos de ferramentas computacionais que têm grande empregabilidade nos estudos linguísticos. Dentre eles, os anotadores PoS e os *parsers* sintáticos são provavelmente as ferramentas mais bem conhecidas e absolutamente necessárias para o trabalho com *corpora*. Adicionalmente os lematizadores e *stemmers* são ferramentas básicas à investigação linguística computacional. Mas, na verdade,

[11] A autora considera linguística computacional aproximadamente como PLN, que por sua vez é parte de inteligência artificial, logo, um campo dentro da ciência da computação (*op. cit*., p. 3).
[12] https://www.cs.jhu.edu/~jason/

qualquer tipo de anotação automática com alto nível de acurácia, seja ela morfossintática, semântica ou morfológica é extremamente útil no tratamento computacional de dados linguísticos.

Para além dos diversos tipos de anotadores, as ferramentas atualmente desenvolvidas para a identificação de elementos de impacto linguístico que transcendam o texto escrito, como o software de análise acústica PRAAT[13], também ocupam um espaço crucial nas investigações linguísticas.

Conjuntos de ferramentas gratuitas, desenvolvidas por linguistas computacionais, como as oferecidas por Lawrence Anthony[14], são de grande utilidade e acessibilidade, auxiliando tanto linguistas, quanto professores e estudantes de línguas.

Para o tratamento computacional do PB há um longo histórico de ferramentas computacionais oferecidas pela equipe do NILC[15], para além das dissertações e teses desenvolvidas nesse núcleo de pesquisa. Também de destaque são iniciativas de desenvolvimento de ferramentas, de pesquisadores que não necessariamente atuam em equipes institucionais de grande porte, como as de Leonel Figueiredo de Alencar Araripe[16], da UFC, desenvolvedor do *parser* Aelius[17]. Notória também é a contribuição de Plínio de Almeida Barbosa[18] para a pesquisa das ciências da fala, com o desenvolvimento de ferramentas computacionais junto ao seu grupo de colaboradores, como o recente Alinha-PB[19].

Movendo-nos, agora, para a contribuição de PLN, entramos em um universo em que a escalabilidade dos dados e os modelos de aprendizado de máquina, cada vez

[13] https://www.fon.hum.uva.nl/praat/
[14] https://www.laurenceanthony.net/software.html
[15] http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources
[16] https://scholar.google.com.br/citations?user=H7HoA0AAAAAJ&hl=pt-BR
[17] https://github.com/LR-POR/aelius
[18] https://www.researchgate.net/profile/Plinio-Barbosa-2
[19] https://conversoralinhador.herokuapp.com/about

mais sofisticados e com resultados impressionantes, se tornam a prática da área. Para além da contribuição de linguistas e cientistas da computação, essa área conta com matemáticos, estatísticos, físicos, engenheiros – sendo, assim, uma área verdadeiramente multidisciplinar.

É historicamente difícil dizer quando foi o início dessa área de pesquisa. Há opiniões divergentes (cf. CHURCH; LIBERMANN, 2021). Mas, uma referência seminal é Turing (1950), quando o famoso Teste de Turing é proposto. Neste teste, um avaliador humano deve decidir se respostas a perguntas em língua natural foram produzidas por um humano ou por uma máquina. No momento presente, 2022, ainda há diferenças significativas na capabilidade de uma máquina emular seja a produção ou a compreensão de uma língua natural. Entretanto é inegável o imenso avanço da área e as inumeráveis aplicações de PLN já alcançados. Os progressos da área se devem a uma conjunção de fatores: a disponibilidade de volumes cada vez maiores de dados linguísticos em formato digital, avanços na pesquisa linguística, crescimento exponencial do poder de computação, e o desenvolvimento de novos métodos, principalmente em aprendizado de máquina.

Tanto no contexto acadêmico quanto no industrial, um grande repertório de tarefas são desenvolvidas, como a classificação de textos, extração de informação, análise de sentimento, tradução automática, síntese de fala, reconhecimento de fala, geração de textos, sumarização, *chatbots* e agentes virtuais, dentre outras em um rol imenso de produtos.

O investimento que a indústria dedica à área pode ser testemunhado facilmente através do acesso a qualquer dos serviços oferecidos pelas empresas tecnológicas internacionais via *world wide web*. Os nossos computadores pessoais tornaram-se um observatório privilegiado para a chamada *BigTech*, que espiona com muito sucesso toda a produção linguística veiculada pelas nossas interações com a máquina, extraindo as informações que lhe são de interesse comercial.

As aplicações de PLN estão também presentes em contextos que afetam a vida da humanidade em todo o planeta, como na medicina, ou na gestão do trânsito nas grandes metrópoles, ou até mesmo nas aplicações forenses, cruciais tantas vezes na resolução de atividades criminosas.

Em um artigo de opinião recente, Kavinski (2022) afirma que podemos estar próximos do momento em as línguas naturais deixarão de ser exclusivamente humanas e passarão a fazer parte da nossa interação com as máquinas e os objetos. Parece-nos bastante ousada essa afirmação, mas é fato que não sabemos qual será o limite para PLN e questões éticas têm sido progressivamente mais discutidas pelos praticantes da área (cf. os recursos disponibilizados em https://aclweb.org/aclwiki/Ethics_in_NLP).

No contexto brasileiro, acompanhando a tendência internacional, sabemos que há interesse e investimento em PLN pelas grandes empresas que atuam em diversos setores de serviços. Isso, interessantemente, tem criado novas possibilidades de trabalho para jovens linguistas já formados dentro de programas de estudo que consideram a programação como algo necessário para a atuação profissional.

Na academia, praticamente em todas as universidades em que há núcleos de pesquisa em ciência da computação, informática, matemática e estatística aplicada, há também pesquisadores envolvidos com projetos em PLN, muitas vezes voltados para aplicações industriais, médicas ou comerciais. A produção científica na área cresce logaritmicamente, passando por um momento de muita efervescência e projetos de muito interesse social. A título de ilustração, sugerimos a verificação de artigos listados pela OMS relacionados ao coronavírus, que envolvem PLN como base metodológica (cf. https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/?lang=pt&q=mh:%22Processamento%20de%20Linguagem%20Natural%22).

No contexto brasileiro, mencionaremos apenas um grande projeto que nos parece particularmente relevante quando se discute PLN para o português brasileiro:

o NLP2[20], liderado por Marcelo Finger e Thiago Pardo. Esse projeto guarda-chuva faz paz parte do Center for Artificial Intelligence e agrega pesquisadores de diversas instituições, com o propósito de desenvolver recursos que avancem PLN em português. Dentre suas frentes de trabalho estão o processamento do português via análise sintática e *parsing*, tarefas de anotação para reconhecimento e síntese da fala e a criação de um *corpus* geral do português brasileiro contemporâneo.

Temos hoje no Brasil equipes que trabalham desenvolvendo recursos e ferramentas linguístico-computacionais que acompanham o estado da arte no cenário internacional – uma tendência que, esperamos, continue a se desenvolver e a apoiar cada vez mais a nossa compreensão da língua portuguesa e das aplicações de impacto social que a partir dela podem ser criadas.

Também crescente é o número de associações profissionais e eventos acadêmicos (cf. STIL, JDP) que têm dado espaço a trabalhos voltados para o processamento computacional do português brasileiro.

Por fim, gostaríamos de mencionar o grupo Brasileiras em PLN[21], fundado em 2020 por Helena Caseli e Brielen Madureira. O grupo reúne mais de 100 mulheres brasileiras que atuam em PLN sob vários formatos de inserção, e tem uma agenda que destaca a atuação das mulheres brasileiras neste campo científico.

Passaremos, a seguir, à apresentação dos doze artigos que compõem este volume temático e que foram agregados de acordo com as três vertentes temáticas discutidas nesta seção 1.

2 Estudos sobre Tratamento Computacional do Português Brasileiro

Os artigos apresentados a seguir, de uma maneira ou de outra abordam questões relativas a *corpora* do português brasileiro, seja ao tratar aspectos da

[20] https://c4ai.inova.usp.br/pt/pesquisas/#NLP2_port
[21] https://sites.google.com/view/brasileiras-pln/in%C3%ADcio

composição de *corpora*, sua anotação e sua exploração para a abordagem de problemas linguísticos específicos, seja no tratamento via ferramentas computacionais de dados de *corpora* ou na exploração de algoritmos e técnicas de aprendizado de máquina. Os *corpora*, como discutido na ampla literatura da área, são a base e o ponto de partida para o tratamento computacional das línguas naturais. Alguns dos artigos trazem revisões bibliográficas sobre o desenvolvimento da linguística de *corpus*, da linguística computacional e do processamento de línguas naturais. Há visões distintas sobre as áreas envolvidas, bem como adoção terminológica variada.

Conforme já mencionado, com o propósito de melhor orientar o leitor, os doze artigos que compõem este volume foram agrupados em três eixos temáticos principais: (a) compilação, tratamento e exploração de *corpora*, (b) estudos pautados pela linguística computacional e (c) processamento de língua/linguagem natural (PLN). Os cinco primeiros trabalhos (Santos, Kauffmann, Gomide *et al*., Cruz, Ferrari e Cunha) representam o eixo sobre aplicações de linguística de *corpus*. O segundo eixo temático, metodologias da linguística computacional para a análise linguística, aglutina mais cinco artigos (Rodrigues *et al*., Raso *et al*., Santos Junior e Isquerdo, Alencar e Rademaker, Bick). E, finalmente, o terceiro eixo temático, sobre processamento de língua natural (PLN) reúne os dois trabalhos finais deste volume (Duran *et al*., Souza e Di Felippo).

O volume é aberto pelo estudo de Diana Santos (Universidade de Oslo e Linguateca) intitulado **"A Gramateca e a Literateca como macroscópios linguísticos"**. Este estudo apresenta uma breve contextualização sobre os ambientes da Gramateca e da Literateca, que integram a Linguateca, trazendo exemplos de perguntas de pesquisa acerca do português que estes ambientes são capazes de responder. A autora apresenta uma discussão sobre as diferentes potencialidades e funcionalidades destes ambientes para apoiar a pesquisa em língua portuguesa, principalmente o português brasileiro, demonstrando o papel destes ambientes como um observador macroscópio da língua

do ponto de vista semântico e morfossintático. Adicionalmente reflete sobre estes ambiente no ensino do português, na leitura distante de textos literários e na extração de informação em língua portuguesa.

O segundo artigo intitulado **"Cognição e variação linguística de gêneros/registros jornalísticos: um estudo baseado em corpus"** escrito por Carlos Henrique Kauffmann (PUC-SP) investiga o processo de identificação e reconhecimento de gêneros/registros jornalísticos na diferenciação entre textos. Para isso, o autor descreve sua metodologia que combina os recursos metodológicos da linguística de *corpus* a uma pesquisa qualitativa, a fim de medir o grau de concordância de um grupo de especialistas quanto aos gêneros/registros existentes nos textos jornalísticos. O *corpus* avaliado foi formado por textos de duas edições do jornal "Folha de São Paulo" impresso, totalizando 1.431 textos. Os resultados obtidos foram tabulados pelo grau de concordância entre os classificadores, levando em conta as dimensões do texto jornalístico identificadas pelo autor em um estudo de 2005, demonstrado uma tendência das classificações para as categorias reportagem e notícia.

Na sequência, as autoras Andressa Rodrigues Gomide (Universidade de Coimbra), Taíse Simioni (UNIPAMPA) e Aden Rodrigues Pereira (UNIPAMPA) exploram "**O fenômeno do desfocamento do agente uma discussão sobre a importância dos recursos computacionais para os estudos linguísticos"**. O artigo, trata das possibilidades de uso de ferramentas da linguística de *corpus* (LC) e de um *corpus* de escrita acadêmica em português (CoPEP) para explorar o fenômeno do desfocamento do agente em artigos acadêmicos publicados no Brasil e em Portugal. As autoras discutem o papel da LC para análises linguísticas sintetizando alguns dos seus aspectos teóricos e pressupostos mais importantes para medir a precisão da ocorrência do fenômeno linguístico, foco de sua pesquisa. Para sustentar a proposta de diferenciação entre as construções frasais e seu lugar no contínuo de desfocamento do agente, as autoras apresentam resultados de dois estudos empíricos sobre o tema

para o português brasileiro (PB) e suas implicações. Por fim, o tema desfocamento do agente em um *corpus* de escrita acadêmica é abordado através da descrição do processo de preparação do *corpus* (tratamento via TreeTagger e Spacy), buscas via CQPweb e apresentação de resultados de análises preliminares.

O estudo **"As contribuições da Linguística de Corpus e do Processamento de Linguagem Natural na elaboração do protótipo do Dicionário Ideológico de Locuções"**, de autoria de Thyago José da Cruz (CPTL/UFMS) inicialmente apresenta uma revisão bibliográfica teórica sobre a linguística de *corpus* e o processamento da linguagem natural. O autor demonstra o uso de recursos da linguística de *corpus* e de uma ferramenta computacional por meio do software *FieldWorks Language Explorer* (FLEx) para a prototipação do Dicionário Ideológico de Locuções, de caráter monolíngue, com locuções tanto onomasiológicas quanto semasiológicas. A extração das unidades fraseológicas para compor o dicionário adotou o *Corpus Brasileiro* e a *Web* como fontes primárias, e também o Tesouro do Léxico Patrimonial Galego-Português e o Dicionário de Expressões Idiomáticas como fontes secundárias. O artigo ainda descreve os passos executados no FLEx para a elaboração do corpo do dicionário, com a construção da parte sinóptico-analógica, da parte analógica e da parte alfabética.

O próximo artigo deste número temático, que fecha o primeiro eixo temático, **"Reflexões metodológicas sobre *datasets* e linguística de *corpus*: uma análise preliminar de dados legislativos"**, foi escrito por Lúcia de Almeida Ferrari (UFMG) e Evandro Landulfo Teixeira Paradela Cunha (UFMG). Inicialmente, os autores analisam a diferença entre a utilização de *corpora* e de *datasets* na linguística, apontando as potencialidades e limitações de cada um. Na sequência, demonstram as possibilidades de uso de um *dataset* para pesquisa linguística ao realizar uma análise quantitativa e exploratória de caracterização dos dados, além de uma análise lexical avaliando as mudanças ao longo do tempo em uma análise do tipo diacrônica. O estudo linguístico exploratório, parte de uma etapa de pré-processamento no *dataset*

da Base de Normas Jurídicas Brasileiras visando sua compreensão, o que apontou a existência de erros, trazendo a reflexão dos autores para a importância desta etapa principalmente para análises linguísticas quando os dados não contaram com a participação de um/a linguista em sua preparação. Em suas considerações finais, os autores trazem as principais observações acerca do *dataset* analisado e discutem tanto as questões metodológicas empregadas quanto o uso de ferramentas e métodos computacionais para a análise linguística diacrônica.

O segundo grupamento temático, estudos pautados pela linguística computacional, é aberto pelo sexto artigo, intitulado **"Bases lexicais verbais do português brasileiro",** de autoria de Roana Rodrigues (UFS), Marcella Lemos-Couto(UFSCar), Francimeire Leme Coelho(UFSCar), Isaac Souza de Miranda Junior(UFSCar) e Oto Vale(UFSCar). O estudo é uma contribuição sobre o estado da arte no que se refere às bases lexicais verbais do português brasileiro (PB). Os autores focam na descrição, sistematização e classificação dos verbos do PB com ênfase nos recursos descritivos que podem ser utilizados nas tarefas de PLN. Foram descritas e comparadas três bases de dados verbais do PB - VerbNet.Br, VerboWeb e Verbo-Brasil - com extensão superior a 1.000 lexemas verbais, com acesso disponível on-line e gratuitamente, e que sofreram atualização nos últimos 10 anos. O artigo traz uma análise crítica sobre as bases de dados verbais do PB, discutindo os seus pontos comuns e divergentes, e explora as informações sintático-semânticas presentes em cada base.

Em **"*Modeling the prosodic forms of Discourse Markers* (Para uma modelagem das formas prosódicas dos Marcadores Discursivos)"** os autores Tommaso Raso (UFMG), Albert Rilliard (Université Paris Saclay) e Saulo Mendes Santos (UFMG), motivados pela imprevisibilidade na determinação de um item lexical ser um Marcador Discursivo (MD), e pela falta de clareza na identificação da função específica de um item lexical, mesmo se já determinado como um MD, trazem uma dupla

contribuição. A primeira é uma proposta de nova solução linguística, baseada em parâmetros prosódicos, para a identificação dos MDs e as funções específicas desempenhadas pelos diferentes tipos de MDs. Para isso, os autores demonstram a natureza prosódica inerente às marcas formais dos MDs, distinguindo diferentes funções de natureza interacional veiculadas por eles. Como segunda contribuição, os autores apresentam a metodologia aplicada à descrição das diferentes pistas prosódicas e os passos adotados na modelagem computacional dos MDs, de forma a possibilitar a extração automática dos diferentes tipos de MDs a partir de novos dados, discutindo diferentes estratégias e os prós e contras de cada uma delas.

Em **"A construção de um banco de dados lexicográficos em XML a partir de dados dialetais: o Processamento Automático de Linguagem Natural (PLN)"**, Jorge Luiz Nunes dos Santos Junior(UFMS) e Aparecida Negri Isquerdo (UFMS) demonstram o uso de ferramentas informáticas voltadas para a criação e o gerenciamento de *corpora*, na tarefa de extração automática de dados. Para isso, partem de aspectos teóricos da Lexicografia, da Dialetologia e da Linguística Computacional para estabelecer parâmetros para a construção de uma base de dados em *XML* (*Extensible Markup Language)*. Adicionalmente, os autores demonstram os passos metodológicos que vêm sendo utilizados na coleta de dados em áudio do *corpus* dialetal do Projeto Atlas Linguístico do Brasil (ALiB). A estruturação dos dados resultante deste estudo permitirá a construção de uma aplicação web que servirá de suporte para a elaboração de um vocabulário dialetal *on-line*. Por fim, os autores refletem sobre a importância do planejamento da arquitetura usada para organizar e estruturar os dados provenientes da extração automática de um *corpus* dialetal.

Na sequência, o nono artigo deste volume, **"Modelação da valência verbal numa gramática computacional do português no formalismo HPSG"** de autoria de Leonel Figueiredo de Alencar (UFC-EMAp/FGV) e Alexandre Rademaker (EMAp/FGV), versa sobre a implementação da valência verbal na PorGram, uma nova

gramática computacional do português a partir da teoria gramatical formal lexicalista HPSG. Através de uma rica discussão demonstrativa de resultados comparativos entre análises sintáticas rasas e as análises sintáticas profundas, e diferentes modelos de análise, nota-se um ganho na adoção do modelo HPSG - um analisador sintático profundo - dada sua capacidade de integrar a descrição sintática e a descrição semântica num único nível de representação. A justificativa ao desenvolvimento da PorGram configura-se em ser esta uma alternativa de software livre e de código aberto em contraponto à LXGram. Os autores apresentam todo o arcabouço teórico-prático inerente ao desenvolvimento da primeira parte da PorGram, oferecendo seus principais indicadores com resultados satisfatórios.

Em seu artigo, **"PFN-PT: *A Framenet Annotator for Portuguese* (Anotação semântica automática: um novo Framenet para o português)"**, Eckhard Bick (*University of Southern Denmark*) apresenta um novo recurso *framenet* para a anotação semântica automática do português. Partindo de uma discussão da relação entre a complexidade linguística e do desempenho das ferramentas de anotação, nota-se que as decisões sobre o design do *framenet* dependem do idioma a que está vinculada a anotação, devido às variações dos seus aspectos sintáticos e semânticos. O *Parsing Framenet for Portuguese* (PFN-PT) configura-se como um recurso para a anotação semântica automática de Português capaz de delinear tal ponte sintático-semântica. O PFN-PT consiste em duas partes complementares, o *framenet* e o anotador de *frames*, ligados por uma representação de dependência semanticamente informada e orientada por valência fornecida por um analisador morfossintático. O artigo apresenta o processo de construção do PFN-PT abrangendo questões como tamanho, cobertura e granularidade do léxico, parâmetros diferenciadores de frames como valência, sintaxe e classe semântica, e por fim, são introduzidas as questões vinculadas à anotação automática de *frames*, que foca em *parsing* e *tagger* para *frames* e papéis semânticos baseados em regras. Rumo à conclusão do artigo, são apresentadas

estatísticas de distribuição de categorias e a avaliação doeste novo recurso computacional.

Abrindo o eixo temático final, processamento de língua natural, o artigo **"Manual de anotação como recurso de Processamento de Linguagem Natural: o modelo *Universal Dependencies* em língua portuguesa"** dos autores Magali Sanches Duran (UNESP), Maria das Graças Volpe Nunes (USP São Carlos), Lucelene Lopes (PUC-RS) e Thiago Alexandre Salgueiro Pardo (USP São Carlos) contribui para os estudos em PLN ao apresentar o manual de anotação fundamentado no modelo internacional *Universal Dependencies* (UD). Os autores ainda refletem brevemente sobre a necessidade de *corpora* anotados e sua aplicação como bases de treinamento para os modelos de Aprendizagem de Máquina voltados ao PLN. Na introdução os autores trazem um percurso sobre o papel do linguista ao longo da evolução do PLN, passando pela ascensão dos *corpora* anotados como recurso valioso para os novos métodos de PLN e terminando por discorrer sobre a importância dos manuais de anotação como parte indissociável de um esquema de anotação. Na sequência apresentam fundamentações teóricas sobre anotação sintática de *corpus* de língua portuguesa e sua relação com o desenvolvimento de *parsers* do português. Seguindo, eles apresentam o modelo UD, seu esquema de anotação, seus conjuntos de etiquetas morfossintáticas e relações sintáticas, trazendo as principais decisões tomadas na instanciação de suas diretrizes no português brasileiro. Por fim, exploram questões relacionadas ao desenvolvimento de manuais para a anotação de *corpora* em português brasileiro segundo o modelo internacional UD.

Concluindo o volume temático, o artigo de Jackson Wilke da Cruz Souza (UNIFAL-MG) e Ariani Di Felippo (UFSCar) intitulado **"*Evaluating a typology of signals for automatic detection of complementarity* (Avaliação de uma tipologia de sinais para a detecção automática da complementaridade)"** apresenta uma tarefa de validação da taxonomia de sinais (textuais) proposta anteriormente pelo primeiro

autor para a detecção automática das relações de complementaridade CST (*Cross-Document Structure Theory*) em um *corpus* multidocumental de notícias em português brasileiro. Inicialmente, os autores fornecem uma introdução ao CST, apresentando sua estrutura principal para a análise e uma visão geral da noção de complementaridade. Em seguida, apresentam o *corpus CSTNews* e a tipologia (ou taxonomia) de sinais usada na avaliação. A avaliação foi realizada utilizando-se algoritmos de diferentes paradigmas de Aprendizado de Máquina supervisionados disponíveis no software *Weka*. Seus resultados apontaram para um alto índice de acurácia geral (superior a 90%), indicando o potencial dos algoritmos usados na detecção automática das relações de complementaridade.

3 Palavras finais

Gostaríamos de agradecer ao editor-chefe da Revista Domínios de Lingu@gem, Guilherme Fromm, por ter aceitado a nossa proposta temática para compor o conjunto de temas em votação pelo conselho editorial da revista para o ano de 2022, bem como por todo o trabalho que efetuou para que esse volume se concretizasse.

Igualmente, agradecemos aos autores pelo interesse em compartilhar suas pesquisas e seus artigos, e ao corpo de pareceristas que avaliou e fez sugestões para o engrandecimento dos trabalhos submetidos.

Esperamos que esse volume ajude a divulgar os trabalhos sobre o tratamento computacional do PB em toda a sua riqueza temática e metodológica, e desperte o interesse e curiosidade dos leitores para investigar mais sobre o assunto e descobrir o vasto e promissor campo que se abre à nossa frente.

Referências Bibliográficas

BICK, E. **The Parsing System "Palavras":** Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus, Denmark: Aarhus University Press, 2000.

CHURCH, K.; LIBERMAN, M. The future of computational linguistics: on beyond alchemy. **Frontiers in Artificial Intelligence**, v. 4, 2021. Disponível em: https://www.frontiersin.org/articles/10.3389/frai.2021.625341. Acesso em: 02 ago. 2022. DOI https://doi.org/10.3389/frai.2021.625341

FRANCIS, W. N.; KUČERA, H. **Computational Analysis of Present-Day American English**. Providence: Brown University Press, 1967.

FREITAS, M. C. **Linguística computacional**. São Paulo: Parábola, no prelo.

KAVINSKI, A. A corrida pelo processamento da linguagem natural. **MIT Technology Review**. 12 jan. 2012. Disponível em: https://mittechreview.com.br/a-corrida-pelo-processamento-de-linguagem-natura/. Acesso em: 02 ago. 2022.

OTHERO, G. A.; MENUZZI, S. M. **Linguística computacional:** teoria e prática. São Paulo: Parábola, 2005.

SVARTVIK, J. (org.). **The London Corpus of Spoken English:** Description and Research. Lund Studies in English 82. Lund: Lund University Press, 1990.

TURING, A. M. I.—Computing Machinery and Intelligence. ***Mind***, vol. 59, n. 236, p. 433–460, 1950. Disponível em: https://academic.oup.com/mind/article/LIX/236/433/986238. Acesso em: 02 ago. 2022. DOI https://doi.org/10.1093/mind/LIX.236.433

DOI: 10.14393/DL52-v16n4a2022-2

A Gramateca e a Literateca como macroscópios linguísticos

Gramateca and *Literateca* as language macroscopes

Diana SANTOS[*]

RESUMO: Neste artigo exploramos várias potencialidades que os ambientes da Gramateca e da Literateca permitem aos usuários interessados na pesquisa em língua portuguesa. Por um lado, apresentamos estes ambientes dando conta de novas funcionalidades acessíveis; por outro, trazemos dez exemplos de perguntas de pesquisa para demonstrar a utilidade da existência destes serviços, que pretendem ser uma espécie de macroscópio para observar a língua, nas vertentes semântica e morfossintática, assim como para a leitura distante de textos literários e a extração de informação em português.

PALAVRAS-CHAVE: Corpos. Visualização. Linguística com corpos. Estudos literários. Português.

ABSTRACT: This paper demonstrates several features of *Gramateca* and *Literateca*, which are environments for linguistic and literary research on top of being a large richly annotated corpora in Portuguese. The paper presents them and pinpoints several new functionalities; it additionally provides ten examples of research questions that demonstrate the usefulness of this kind of Web-based service, that can be conceived as language macroscopes. Examples concern semantics, morphosyntax, distant reading of literary texts, and information extraction.

KEYWORDS: Corpora. Visualization. Corpus linguistics. Literary studies. Portuguese.

1 Introdução

Há cerca de 20 anos a Linguateca proporciona, a todos os linguistas e cientistas da computação interessados no processamento do português, uma plataforma para buscar e obter dados linguísticos, através do projeto AC/DC, desde a publicação dos artigos seminais Santos e Ranchhod (1999) e Santos e Bick (2000). Ao longo do tempo,

[*] Doutora, Universidade de Oslo e Linguateca. ORCID: https://orcid.org/0000-0002-3108-7706. d.s.m.santos@ilos.uio.no

muito mais material de vários gêneros foi incorporado, graças à adesão da comunidade, o que nos permitiu disponibilizar e tratar corpos[1] não criados por nós, sempre em regime de não exclusividade. Da mesma forma, vários grupos uniram esforços com a Linguateca para criar novos recursos em parceria, muitos dos quais em constante crescimento, como é, por exemplo, o caso do OBras (SANTOS *et al*., 2018). Isso implica que as descrições do material e dos recursos acessíveis necessitem de alguma atualização desde o artigo Santos (2014), que buscou oferecer um histórico da atividade com corpos na Linguateca. Além disso, devido ao nosso constante uso dos corpos, é de se esperar que tenhamos desenvolvido também novas informações a eles associada, e novas formas de as visualizar, que merecem ser descritas à comunidade linguística e de processamento de linguagem natural (PLN).

2 Gramateca

A Gramateca, conceitualizada em 2014, foi apresentada em Santos (2014) como um ambiente para estudar a gramática da língua portuguesa. Ela contém, além dos corpos propriamente ditos, a potencialidade de criar formulários de interação com vários falantes para testar casos complicados, através do Rêve – unindo, assim estudos quantitativos com qualitativos. Santos *et al*. (2015) descreve em mais detalhe essa nova forma de fazer gramática, que utiliza textos reais, diferentes gêneros, e, adicionalmente, agrega a capacidade de levar em conta diversos informantes. Esse artigo exemplificou os conectores condicionais, a descrição fina do corpo humano, e os diferentes sentidos do verbo *admirar.*

Mas outras questões foram também sendo identificadas como pertinentes para uma visualização do conteúdo de um corpo, como o artigo de Santos (2015) e a

[1] Para a motivação de usar o termo *corpo* e *corpos* em vez de *corpus* e *corpora*, veja-se Santos (2008).

apresentação de Santos (2018) ilustram: linhas de tempo, formas distintas de identificação de uma mesma entidade, nuvens de palavras etc.

O que significa que, associadas à Gramateca, se encontram várias ferramentas que permitem manipular e apresentar resultados de forma mais intuitiva. Tais ferramentas ajudam o pesquisador a explorar o material e serão exemplificadas na sequência, depois de apresentarmos a Literateca.

3 A Literateca

De fato, foi o início da consideração dos textos literários como um gênero por si só, com requisitos específicos, e com bastante informação associada a cada texto (os metadados, que incluem autor, data, gênero do autor, gênero literário, escola literária, canonicidade...), que levou a uma explosão de novas ferramentas associadas aos corpos, e que mereceu, portanto o nome de Literateca – que é uma especialização da Gramateca para textos literários.

Assim, além de tudo o que já era possível fazer nos outros corpos, definiram-se distribuições por período, assim como a anotação de localizações, e de personagens, permitindo a construção de redes de personagens e da sua presença na obra ao longo do tempo, como foi descrito em Santos e Freitas (2019), e o desenho automático de mapas, cf. Santos e Alves (2022). Foi também possível utilizar ferramentas de redução de dimensionalidade para oferecer uma visão de conjunto usando muitas características diferentes, como ilustrado em Santos *et al*. (2020).

É importante, contudo, deixar claro que os desenvolvimentos na Literateca são consequência do que já existe na Gramateca, e que algumas anotações iniciadas na Literateca foram depois expandidas a outros corpos, ou seja, ultrapassaram o domínio restrito dos textos literários. E, de qualquer maneira, a existência da Literateca como um subconjunto da Gramateca com características especiais se beneficia grandemente da comparação com os corpos todos, para identificar o que é sobretudo literário ou

pertence à língua em geral. Esperamos poder ilustrar isso de forma convincente no resto do artigo.

4 Algumas explorações em textos literários

Começaremos por perguntas relativamente fáceis, e vamos aumentando a sofisticação linguística, de forma a vermos como estes ambientes, ou infraestruturas de pesquisa, permitem estudar o português e o português brasileiro em especial.

Mas, não podemos deixar de destacar que, mesmo que as perguntas sejam fáceis, está em geral subentendido muito trabalho, além da reflexão sobre a anotação à qual as perguntas recorrem.

Por outro lado, é sempre necessário não se esquecer a base sobre a qual se pergunta e que determina o universo explorado. Em relação aos textos literários, e devido à problemática dos direitos autorais, contamos na sua maioria apenas com obras já no domínio público, e daí as perguntas literárias que faremos se referirem a livros publicados há mais de 80 anos. Apresentamos, a seguir, as perguntas de pesquisa.

#1 *Quais autores do século XIX incluem mais nomes de localizações nos seus textos?*

O que é um autor do século XIX? Escolhemos aqui operacionalizar esta questão como "autores de obras publicadas no século XIX", e só contar com essas obras, mas evidentemente outras escolhas seriam possíveis. Veja-se, a seguir, a Tabela 1, na qual são listados autores do século XIX e a incidência de topônimos em suas obras.

Tabela 1 – A presença de topónimos em autores com obras publicadas no século XIX.

Autor	**Loc.**	**Palavras**	**loc/pal*1000**
Marquês de Fronteira e d'Alorna	1236	67 031	18,44
Manuel Pinheiro Chagas	1103	77 178	14,29
Zeferino Norberto Gonçalves Brandão	862	60 869	14,16
Joaquim Nabuco	1073	80 144	13,39
Alexandre Herculano	17867	1 484 356	12,04
Ramalho Ortigão	788	67 535	11,67
Alexandre de Serpa Pinto	2167	192 543	11,25
Alberto Pimenta	1743	158 688	10,98
Teixeira de Vasconcellos	1033	95 742	10,79
António Nobre	295	28 936	10,19
Matilde Isabel de Santana e V. M. Bettencourt	423	42 502	9,95
Adolfo Caminha	1911	193 735	9,86
Luciano Cordeiro	1498	160 322	9,34
António Francisco Barata	1850	198 275	9,33
Candido de Figueiredo	166	17 871	9,29
Bernardino Pereira Pinheiro	465	50 803	9,15
Alfredo Campos	291	33 916	8,58
Tomaz de Melo	539	64 151	8,4
Inglês de Sousa	996	119 746	8,32
Inácio Pizarro de Morais Sarmento	346	45 312	7,64

Fonte: elaborada pela autora.

Uma rápida análise dos resultados na Tabela 1 permite-nos observar que os romances históricos, e aqueles que se passam em zonas em princípio desconhecidas do leitor, como é o caso da Amazônia para Inglês de Sousa, são os que têm maior densidade toponímica. A presença de muito mais autores portugueses do que brasileiros nesta lista deve-se ao fato de a própria amostra apenas conter (por enquanto) 28 autores brasileiros deste período, num universo de 117 autores no total.

#2 Quais obras brasileiras têm mais referências a etnicidade?

Se procurarmos agora apenas entre as 203 obras brasileiras presentes na Literateca, a Tabela 2 ilustra as que têm mais menções ao campo semântico da etnicidade.

Tabela 2 – A presença da etnicidade em obras brasileiras.

Obra	Etn.	Pal.	Freq. Rel. em 10000 palavras
Rei negro	430	67053	64,13
Uma tragédia no Amazonas	114	28530	39,95
O Uraguai	57	14684	38,81
O Guarani	449	124417	36,08
As Vítimas-Algozes	374	112716	33,18
Nove noites	16	5077	31,51
Banzo	99	32136	30,81
A escrava	16	5274	30,34
Iracema, lenda do Ceará	102	34788	29,32
O sonho das esmeraldas	81	30375	26,67
Pausa	2	754	26,53
Os irmãos Leme	91	35276	25,8
A viagem maravilhosa	335	133116	25,17
Vidros quebrados	5	2004	24,95
Turbilhão	217	88164	24,61
O Bom-Crioulo	110	45303	24,28

Fonte: elaborada pela autora.

Se em alguns casos das 16 obras com proporcionalmente mais termos relativos a etnicidade isso seria previsível pelo seu título, convém chamar a atenção para os dois contos com apenas duas ou cinco palavras remetendo para esse domínio, de Moacyr Scliar e de Machado de Assis respetivamente, que chegaram à lista devido ao seu tamanho reduzido. Para evitar isto, poderíamos ter escolhido apenas romances e novelas, mas este é mais um caso em que a pergunta poderia ter sido interpretada de maneira diferente.

#3 *Entre médicos e boticários por um lado, e padres e frades por outro, quais são as profissões mais presentes na literatura lusófona?*

Mais uma vez, esta pergunta pode ser operacionalizada de diferentes formas: simplesmente contando as vezes que cada uma das profissões aparece no total das obras, ou identificar esse assunto por obra: quantas obras só têm referência a uma profissão ou a outra, e em quantas obras uma supera a outra?

Num universo de 860 obras, em 472 obras há referência a um médico ou boticário, enquanto em 582 há referência a padres, frades, freiras ou monges.

E a conclusão é inescapável. Embora a medicina e a profissão médica sejam muito presentes na literatura lusófona – veja-se Santos (2019) e Langfeldt (2021) –, na amostra que temos a profissão religiosa ainda é mais conspícua. Com efeito, das 587 obras em que uma ou ambas as ocupações, ou vocações, ocorrem, 96% (561 obras) falam mais da classe religiosa, sejam heróis ou vilões. É também interessante notar que apenas 92 obras mencionam ambos os tipos de profissões.

Um exemplo que demonstra inequivocamente este predomínio é a obra *A Mortalha de Alzira.* Sendo esta a obra que contém mais referências às palavras *médico* e *boticário* em valores absolutos, 40, apresenta mesmo assim 54 menções a palavras descrevendo religiosos.

#4 Quais são as ações mais associadas a uma ou a outra profissão?

Em 1.133 casos de um médico sujeito (de uma frase ativa ou passiva)[2], os seguintes verbos são mais frequentes do que para a média dos sujeitos: *receitar*, *recomendar*, *aconselhar*, *declarar, examinar*, *chamar* e *tomar*. Em 3.020 casos de um religioso sujeito (de uma frase ativa ou passiva), são os seguintes os verbos mais frequentes do que a média: *curar*, *rezar*, *erguer-se, levantar* e *olhar*.

Enquanto alguns destes verbos concordam imediatamente com a nossa intuição, outros há que merecem investigação: em particular, não conseguimos achar explicação para a alta frequência de *olhar* associada a religiosos, sobretudo se considerarmos que uma quantidade significativa dos casos de *levantar* se refere precisamente a *olhos* ou *olhar*. Já quanto a *tomar*, são as expressões *tomar o pulso*, *tomar notas* e *tomar injeção* que explicam a co-ocorrência preferencial com a profissão médica.

[2] A expressão de busca para obter os verbos foi: [lema="médico|facultativo|boticário|farmacêutico" & pos="N.*"] [pos=".*AUX.*|ADV"]* @[pos="V.*" & pos!=".*AUX.*"]

Mas o mais surpreendente foi o verbo *curar* estar mais associado a religiosos do que a médicos. Ao analisar esses casos, descobrimos um erro da análise sintática, que considerava incorretamente *cura* em *padre cura* como forma do verbo *curar*. Se o mencionamos aqui, é porque é sumamente importante não confiar cegamente nos resultados, e investigar tudo o que vá contra a nossa intuição de falantes de uma língua. Afinal de contas, não é possível garantir que a informação associada a mais de um bilhão de palavras esteja 100% correta. O que fazemos é tentar melhorar, e corrigir semiautomaticamente, os casos que vamos encontrando, usando a filosofia de interação entre pessoas e máquinas apresentada em Santos e Mota (2010).

#5 Em termos de co-ocorrência de emoções, existe distinção entre texto literário e texto jornalístico (por exemplo)?

Dado que a menção de emoções é algo muito corrente em português, talvez a maneira mais fácil de responder a esta pergunta seja tentar visualizar as emoções em conjunto através da sua co-ocorrência, na Literateca, e nos corpos todos, e comparar as duas visualizações.

As figuras (feitas com o programa igraph (CSARDI; NEPUSZ, 2006) do R (R Development Core Team, 2008)) ilustram um grafo de coocorrências entre todas as palavras que estavam marcadas com o campo semântico da emoção, segundo os grupos previamente identificados (veja-se SANTOS; SIMÕES; MOTA (2021) para a documentação deste esforço).

Figura 1 – Grafo das emoções na Literateca.

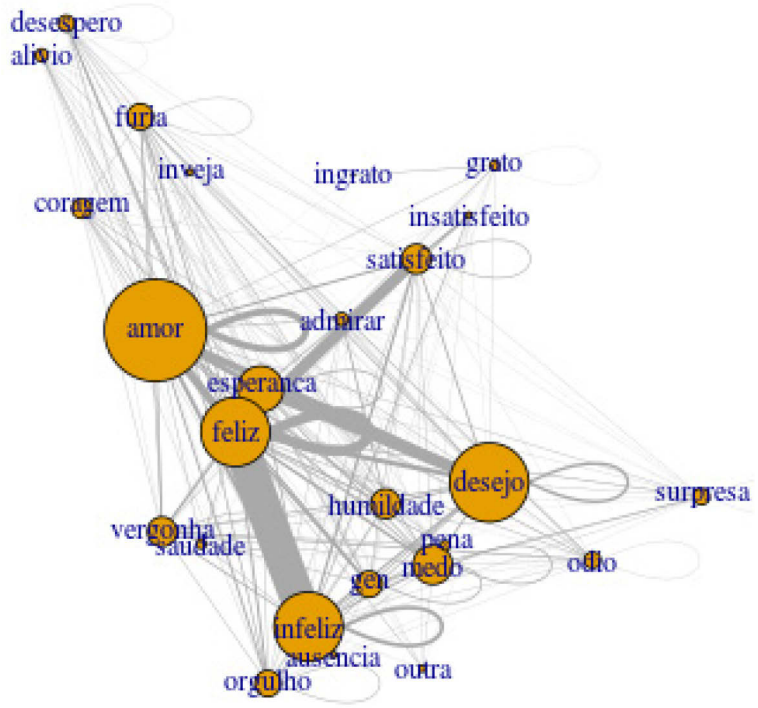
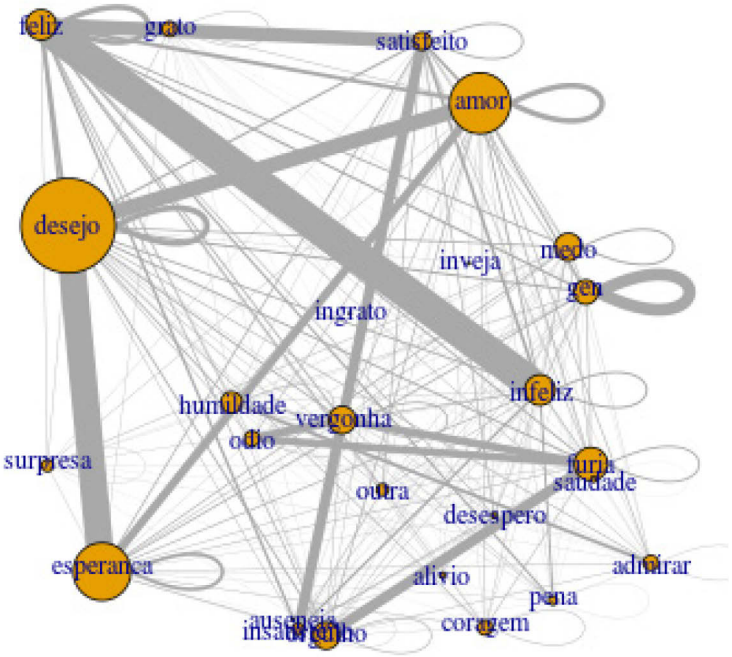


Figura 2 — Grafo das emoções em todos os corpos.



As diferenças entre os dois grafos – que, sendo desenhados aleatoriamente, não mantêm, infelizmente, a mesma posição espacial das emoções – mostram que os

campos semânticos do AMOR e da (IN)FELICIDADE são bastante mais proeminentes nos textos literários (Figura 1), enquanto que nos corpos todos o DESEJO (que aqui seria mais natural chamar intenção e/ou volição) e a ESPERANÇA são mais usados (Figura 2).

Para ilustrar também a riqueza lexical de um destes campos, mostramos a nuvem de palavras associada a AMOR na Figura 3[3].

Figura 3 — Lemas que foram (em contexto) associados à emoção AMOR.

5 Algumas explorações na "língua toda"

Por língua toda queremos indicar a variedade dos corpos todos que temos, e que não incluem, nem podem incluir, evidentemente, toda a língua, mas que contêm

[3] Agradeço a um revisor anônimo pelo seguinte comentário, com que concordo absolutamente: as nuvens de palavras, embora cada vez mais usadas devido ao seu apelo visual, tornam difícil recuperar impressões precisas a partir delas, donde devem ser feitas as devidas ressalvas.

muitos gêneros e tipos de discurso, desde o oral ao religioso, ao acadêmico e à literatura farmacêutica. Feita esta explicação, vejamos então a próxima pergunta:

#6 Que verbos requerem completivas com subjuntivo? Como explicar a variação nos casos em que tanto indicativo quanto subjuntivo são usados?

A questão do uso do subjuntivo, muito frequente em português em comparação com outras línguas românicas, exige que este seja ensinado logo em níveis básicos de língua, e compreendido por quem o ensina. E o que é certo é que existem muitas variáveis que influenciam o seu uso. Em seguida veremos como identificar propriedades relevantes para a escolha de um modo, e criar materiais de ensino em relação às completivas finitas regidas por um verbo (um processo semelhante seria seguido se estivéssemos interessados em completivas regidas por um substantivo, ou por um adjetivo).

Vejamos, por exemplo em texto jornalístico (no corpo CHAVE), qual a distribuição do modo nas completivas finitas associadas aos verbos *esperar*, *acreditar*, *crer*, *considerar*, *julgar*, *achar* e *pensar*[4], indicada na Tabela 3.

Tabela 3 — A distribuição no modo das completivas finitas.

verbo	**SUBJ**	**%**	**IND**	**%**	**Coef. SUBJ/IND**
pensar	149	7.80%	1742	92.12%	0.09
esperar	899	98.60%	12	1.32%	74.9

[4] Um exemplo de como obter estes valores para *achar* seria: [lema="achar"] [word="que"] [pos!="V.*"] [pos="V.*" & temcagr=".*IND.*"] e [lema="achar"] [word="que"] [pos!="V.*"] [pos="V.*" & temcagr=".*SUBJ.*"]

acreditar	380	30.72%	857	69.28%	**0.44**
crer	108	22.60%	370	77.40%	**0.29**
considerar	32	2.46%	1269	97.54%	0.03
julgar	12	4.96%	230	95.04%	0.05
achar	130	3.94%	3165	96.05%	0.04

Fonte: elaborada pela autora.

Vemos que há dois verbos em que existe de fato variação no modo das completivas, nomeadamente *acreditar* e *crer*. Os outros casos ou exigem subjuntivo, como *esperar* (o que era esperado), ou indicativo, embora sejam versões um pouco menos assertivas de *ter certeza (de) que*. A maioria dos casos em que aparece o subjuntivo com esses verbos corresponde ao verbo negado.

Vejamos então com mais atenção os casos de *acreditar* e *crer*, que, embora apresentem o verbo da completiva majoritariamente no indicativo, têm uma fatia considerável no subjuntivo. Por meio da análise dos 488 casos de subjuntivo, verificamos que 271 são negados, 12 correspondem a perguntas, 17 à existência de uma outra oração na completiva (geralmente com *se*) e, portanto, não são casos de subjuntivo devido a *acreditar* ou *crer*, e 30 correspondem a casos da forma *é X acreditar/crer que*, fórmula essa que costuma induzir o subjuntivo (cf. *é bom/provável/difícil que ele venha*) e que pode, portanto, provocá-lo depois do verbo no infinitivo. Outros casos identificados são as expressões *levar a crer*, *ser de crer*, *custar a crer*, assim como casos com o sujeito formado com o pronome *poucos.*

A Figura 4 ilustra a variedade de tempos e modos que são possíveis com a expressão *levar a crer,* nomeadamente o indicativo, o subjuntivo, mas também o futuro e o condicional, o que significa que não faz sentido tentar criar regras para ensinar, visto que as diferenças parecem puramente estilísticas.

Figura 4 — Excerto de concordâncias relativas a *levar a crer que.*

par=FSP951103-063-725: Folha -- Mas tudo **leva a crer que** você é culpada .
par=FSP951116-079-799: Segundo Vieira, tudo **leva a crer que** o ferimento tenha sido feito «com técnica» .
par=FSP950207-008-81: Em face da crise instalada na assistência pública, e sem saber exatamente como ela se processa, o contribuinte pode ser **levado a crer que** terá, com o PAS, algo mais que a desassistência atual .
par=FSP940921-080-811: Na ocasião, Zé Roberto surpreendeu ao anunciar o corte de Pinha, pois tudo **levava a crer que** ele faria a opção entre Jorge Édson e Douglas .

Por outro lado, esta análise mais fina permitiu-nos identificar construções interessantes para serem ensinadas explicitamente, e para treino, por exemplo com o Ensinador (SIMÕES; SANTOS, 2011), um programa de criação de exercícios didáticos baseados nos corpos do AC/DC. Na Figura 5 apresentamos uma possibilidade de exercício, focando o sujeito *poucos* e a construção *é X acreditar*, em que a intenção é que aluno aprenda os verbos apresentados entre parênteses no subjuntivo, preenchendo as lacunas.

Figura 5 — Um exercício associado a completivas objeto de *crer* e *acreditar*.

Coloque o verbo no tempo e modo correto

1. *par=FSP940807-181-2343*: Em ` True Romance ' é impossível acreditar que _____ ele. (ser)
2. *par=PUBLICO-19950311-180-1946*: Era difícil acreditar que _____ passado apenas dez anos desde que Gorbatchov subiu ao poder após a morte de três dinossauros decrépitos. (ter)
3. *par=FSP941106-079-1036*: Poucos acreditam que _____ realista, até porque a inflação será pressionada pela desvalorização do rublo, que encarece as importações. (ser)
4. *par=FSP951219-141-1375*: Penn pode ter amadurecido, mas é difícil acreditar que _____ se livrado das confusões. (ter)
5. *par=PUBLICO-19951128-044-420*: Se é admissível que haja quem 'teja disposto a comprar enchidos ou bebidas a preços baixos, já é difícil acreditar que _____ receptadores de fatias de bolo de chocolate. (existir)
6. *par=FSP950208-008-56*: Com o PIB crescendo 5 % ao ano e o juro real da dívida pública crescendo 35 %, é difícil acreditar que _____ no caminho da «sustentabilidade»! (estar)
7. *par=PUBLICO-19950112-048-328*: Poucos acreditam que _____ sido um negócio destinado a favorecer o «núcleo duro» de accionistas da instituição, apesar de assumirem que o volume e o preço levantam suspeitas. (ter)

#7 Qual a posição preferida de orações participiais?

Outro tópico interessante na sintaxe da língua portuguesa é a existência frequente de orações participiais, ou seja, com o primeiro verbo na forma de particípio passado, e que geralmente caracterizam um elemento da oração, comportando-se como orações adjetivas, mas podendo encontrar-se a distância significativa dele, antes ou depois, como ilustram os exemplos da Figura 6.

Figura 6 — Orações participiais no corpo NILC/São Carlos, mostrando a que entidade se referem.

par=8477: A festa da première de Missão impossível, um dos mais badalados filmes da temporada de verão deste ano — que estréia dia 12 de julho no Brasil — lotou, na noite de segunda-feira, o teatro Mann Village**, comandada** pelo astro principal, Tom Cruise.
par=Mais—94b-1: Mas é preciso ainda lembrar a sua posição a respeito da função social da ciência**, baseada** num ponto de vista democrático que elaborou a partir da sociologia durkheimiana .
par=Agrofolha—94b-2: **Formada** pelos municípios de Lucas do Rio Verde, Tapurah, Nova Mutum e Sorriso, a região (a 350 km de Cuiabá) está produzindo 1,1 milhão de toneladas de soja em cerca de 400 mil hectares .
par=9369: **Decorado** de verde e amarelo e com preços aumentados em até 100 % -- o caldinho de feijão, por exemplo, subiu de R$ 1, na sexta, para R$ 2, na segunda --, o bar Coringa virou um pequeno Maracanã .
par=Agrofolha—94a-2: No total, **computadas** as coberturas e os animais, o faturamento bruto foi de R$ 338,7 mil .

Dos 368.108 particípios passados do corpo NILC/São Carlos, 76.404 funcionam como adjetivos pospostos e 13.831 como adjetivos antepostos. Além disso, 19.310 funcionam como nomes, e temos 130.868 orações participiais que seguem o conceito que modificam[5], e 14.472 que o precedem.

#8 As orações gerundivas são mais frequentes no português do Brasil?

Visto que uma das marcas diferenciais do português do Brasil em comparação com o de Portugal é o seu uso da progressiva com gerúndio, contrastando com a progressiva com *a*+ infinitivo, seria de esperar que as orações gerundivas também fossem significativamente mais usadas em português brasileiro. Usando o corpo CONDIV (Soares da Silva, 2008), contudo, vemos que tal fato não é incontroverso: a porcentagem é um pouco mais elevada, mas não decididamente: com efeito, 5,19 em 1000 palavras em português de Portugal contra 6,25 em português do Brasil são gerúndios numa oração gerundiva.

[5] Seguem: [temcagr=".*PCP.*" & func="ICL-(<.*|N<)"] e antecedem: [temcagr=".*PCP.*" & func="ICL-.*>"]

Se observarmos esta característica ao longo das três décadas contidas no corpo, respectivamente a década de 1950, a de 1970 e a de 2000, ilustradas na Tabela 4, vemos que, embora em português brasileiro a frequência seja sempre superior, a situação parece convergir para uma maior semelhança entre as duas variantes, ambas diminuindo o número de orações gerundivas.

Tabela 4 — A proporção de gerúndios de uma oração gerundiva por 1.000 palavras no corpo. CONDIV.

Variante	**ger.**	**Total**	**%**	**ger.**	**Total**	**%**	**ger.**	**Total**	**%**
	década de 1950			**década de 1970**			**década de 2000**		
PT	7225	1279115	5,6	6101	1204692	5,1	3921	837203	4,7
BR	5906	821426	7,2	4877	843826	5,8	5276	970864	5,4

Fonte: elaborada pela autora.

#9 Quais os gentílicos mais mencionados nos jornais brasileiros?

Considerando agora algo menos gramatical, podemos tentar identificar, nos jornais a que temos acesso, quais as palavras referentes a nacionalidade ou região mais usadas nos jornais do Brasil (nas décadas em que temos material: 1950, 1970, 1994-1995, 2000, 2010). Naturalmente que este é apenas um estudo exploratório, para demonstrar as capacidades da Gramateca, e não pretende apresentar resultados definitivos sobre a comunicação social ou o Brasil em geral. Também é um estudo que exemplifica uma das muitas informações fornecidas pelo PALAVRAS, o analisador sintático subjacente (BICK, 2000, 2007, 2014) [6].

Apresentamos os (primeiros) resultados na Tabela 5 como uma lista por ordem decrescente, e como uma nuvem de palavras na Figura 7, aproveitando para esclarecer que a *Folha de São Paulo* é uma das principais fontes do material jornalístico da Gramateca, o que explica naturalmente a grande quantidade de menções a *paulistas* e

[6] [classe="jorn" & sema=".*nat.*" & variante="BR"]

paulistanos. Convém também relembrar que as palavras *português* e *inglês* podem referir-se à nação ou à língua. Neste caso, as suas posições altas referem-se majoritariamente às línguas e não aos países[7].

Figura 7 — Nuvem de palavras gentílicas.

Fonte: elaborada pela autora.

[7] Existe além disso um clube de futebol chamado a *Portuguesa*, que contribui significativamente para a contagem: 9.105 dos 23.538 casos a ele se referem: 38,7%!

Tabela 5 — Gentílicos no texto jornalístico.

gentílico	**freq.**
brasileiro	96956
paulista	25903
português	23538
inglês	22307
índio	22157
norte-americano	21287
francês	20019
habitante	18681
alemão	15921
americano	13664
brasileira	13398
argentino	13035
italiano	12236
estrangeiro	11335
paulistano	10905
palestino	10320
guarani	10279
fluminense	9508
russo	9008
sérvio	8415
japonês	7767
carioca	7261

Fonte: elaborada pela autora.

5 Algumas explorações em outros corpos

#10 Quais as localizações mais frequentes no DHBB?

Finalmente, podemos aproveitar corpos específicos para fazer perguntas que só façam sentido nesse contexto. Já que temos o privilégio de disponibilizar o corpo Dicionário Histórico-Biográfico Brasileiro, uma obra que compreende muitos verbetes sobre personalidade e eventos relacionados com a história política moderna do Brasil, podemos, usando padrões de busca na Gramateca, identificar as relações familiares entre políticos (Higuchi et al., 2019), ou os locais de nascimento e de morte dos

verbetados, apresentados na Tabela 6. Em negrito se apresentam os locais onde mais políticos acabam a sua vida do que começam.

Tabela 6 — Locais de nascimento e de falecimento dos verbetados no DHBB.

Local	**Nascimentos**	**Mortes**
Rio de Janeiro	1039	1506
São Paulo	126	257
Recife	211	81
Salvador	201	81
Porto Alegre	163	86
Belo Horizonte	95	104
Fortaleza	131	43
Niterói	102	68
Curitiba	86	82
Campos	137	14
Belém	113	28
Brasília	6	122
Maceió	51	33
São Luís	65	19
Cuiabá	65	18
Petrópolis	50	32
Manaus	58	19
Natal	50	25
Aracaju	54	20
Teresina	45	28
Juiz de Fora	48	16
Florianópolis	34	26
Campinas	51	8
Goiânia	26	25
Pelotas	41	7
Pernambuco	41	5
Rio Grande do Sul	39	6
João Pessoa	24	19
Paris	7	29

Fonte: elaborada pela autora.

Embora evidentemente estejamos fazendo uma grande simplificação, podemos identificar os centros de poder político no Brasil observando as cidades em que muitos mais acabam a vida ao invés de lá nascerem: o Rio de Janeiro, São Paulo, Belo

Horizonte e Paris (!). Paris foi identificado por este método, mas claro que a conclusão neste caso tem de ser outra, não só por Paris não ser obviamente um centro do poder político brasileiro (e as mortes lá poderem se dever a um exílio, ou a um cargo diplomático), mas evidentemente porque muito poucos brasileiros nascem em Paris.

Em relação ao número de filhos[8] dos políticos, e não obstante a diferença entre os sexos em termos do conteúdo do DHBB ser muito marcada (só 204 verbetes se referem a mulheres, contra 6.457 sobre homens), verificamos que apenas 51% das mulheres verbetadas tinham filhos, contra 64% dos homens.

6 Considerações finais

Num volume sobre o tratamento computacional do português brasileiro, pode parecer estranho à primeira vista apresentar dois ambientes computacionais que se dedicam à língua portuguesa em todas as suas variedades, mas penso que ficou claro que essa é uma vantagem mesmo que o interesse primordial do pesquisador seja pela variante brasileira do português. Isto porque é possível selecionar só texto em português brasileiro, e ao mesmo tempo se pode comparar os resultados com mais material em português.

No artigo, escolhemos apresentar as potencialidades da Gramateca e da Literateca sobretudo a partir de exemplos de pesquisas relevantes para diferentes tipos de leitores, mas podemos também fazer uma sistematização aqui, até porque nem todas as possibilidades puderam ser exemplificadas. Mencionamos a criação de exercícios (enunciado e sua solução) para ensino da língua portuguesa, por meio do Ensinador, mas não a possibilidade de comparação entre duas buscas, para identificar diferenças entre distribuições, através do Comparador (este, e o Distribuidor, estão descritos em Simões e Santos (2014)). O Distribuidor faculta a obtenção de dados

[8] Para buscar o número de filhos no DHBB: ([word="uma?"]|[pos="NUM.*"]) [classe="biográfico" & lema="filh[oa]"]

extensivos em formato de planilha, para processamento subsequente em programas externos; se os dados forem apenas de distribuição (três colunas, portanto), podem se obter através da interface original, escolhendo-se a opção "Resultados em formato separado por ponto e vírgula". Ilustramos e comentamos a obtenção de nuvens de palavras como um resultado adicional, mas não referimos a criação automática de mapas, no caso de os corpos inquiridos terem geolocalização[9].

Além disso, não documentamos a ligação dos corpos com a possibilidade de criar formulários de consulta a informantes, por meio do Rêve, nem nos debruçamos sobre a criação de documentação de boas práticas de interligação com a linguagem R para a criação de outras visualizações mais complexas do material.

Independentemente de todas estas funcionalidades, é importante salientar que novas informações de anotação foram incluídas nos corpos (como os exemplos de emoção, etnicidade, relações familiares e geolocalização comprovam), e que, em última análise, é o aumento de informação, e a sua melhoria, que tornam estes ambientes úteis para a pesquisa.

Terminamos o artigo indicando que, apesar de termos, ao longo dos anos, desenvolvido ampla documentação e páginas de ajuda [veja-se por exemplo, FREITAS *et al*. (2011), em constante atualização], estamos sempre acessíveis para perguntas e pedidos de auxílio para permitir que todos possam fazer uso do material, assim como estamos abertos a sugestões de cooperação e de criação conjunta de novos recursos ou funcionalidades.

As páginas "oficiais" da Gramateca e da Literateca são respetivamente https://www.linguateca.pt/Gramateca/ e https://www.linguateca.pt/Literateca/,

[9] Esta funcionalidade não foi aqui ilustrada porque até agora apenas os corpos literários portugueses têm essa informação, devido ao financiamento do projeto BILLIG, mas a nossa intenção é continuar esse preenchimento para a parte brasileira num futuro próximo.

embora a interface principal continue a ser a do projeto AC/DC, https://www.linguateca.pt/ACDC/.

Agradecimentos

Agradeço sinceramente a Heliana Mello a sua gentil adaptação do texto para o português do Brasil.

Agradeço também a toda a equipe da Linguateca a existência do projeto e dos recursos, à FCCN – Fundação para a Computação Científica Nacional (Portugal), o alojamento da Linguateca nos seus servidores, e ao UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway pelos recursos computacionais.

Referências Bibliográficas

BICK, E. **The Parsing System "Palavras"**: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Dissertação (PhD), Aarhus University. Aarhus University Press, 2000.

BICK, E. Automatic Semantic Role Annotation for Portuguese. *In*: **TIL**, V Workshop em Tecnologia da Informação e da Linguagem Humana. Rio de Janeiro, RJ, 30 de junho a 6 de julho de 2007. p. 1715-1719.

BICK, E. PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese. *In*: BERBER SARDINHA, T.; FERREIRA, T. L. S. B. (ed.). **Working with Portuguese Corpora**. London/New York: Bloomsbury Academic, 2014. p. 279-302.

FREITAS, C.; SANTOS, D.; GONÇALVES, A. **Perguntas já respondidas sobre o AC/DC**: desde como começar até uso complexo de funcionalidades poderosas. Primeira edição: 15 de Outubro de 2011. Disponível em: https://www.linguateca.pt/ACDC/

HIGUCHI, S.; SANTOS, D.; FREITAS, C.; RADEMAKER, A. Distant reading Brazilian history. *In*: NAVARRETA, C.; AGIRREZABAL, M.; MAEGARD, B. (ed.). **Proceedings of the Digital Humanities in the Nordic Countries 4th Conference** (Copenhagen, Denmark, March 5-8, 2019), 2019. p. 190-200.

LANGFELDT, M. C. Entre médicos e charlatães: A ascensão da medicina na formação da literatura brasileira. Apresentação no **III Encontro Nacional de Estudos Linguísticos e Literários (ENAELL), I Encontro Internacional de Pesquisas em Letras (ENIPEL)**. UEMA, 25-27 de maio de 2021. Disponível em: https://www.youtube.com/watch?v=XFEVaZC_ibU

SANTOS, D. Corporizando algumas questões. *In*: TAGNIN, S. E. O.; VALE, O. A. (org.). **Avanços da Lingüística de Corpus no Brasil**. Editora Humanitas/FFLCH/USP, São Paulo, 2008. p. 41-66. Disponível em https://www.linguateca.pt/Diana/download/Santos2008livroStellaOtofinal.pdf

SANTOS, D. Corpora at Linguateca: Vision and roads taken. *In*: BERBER SARDINHA, T.; FERREIRA, T. L. S. B. (ed.). **Working with Portuguese Corpora.** London/New York: Bloomsbury Academic, 2014. p. 219-236.

SANTOS, D. Gramateca: corpus-based grammar of Portuguese. *In*: BAPTISTA, J.; MAMEDE, N.; CANDEIAS, S.; PARABONI, I.; PARDO, T. A.S.; NUNES, M. G. V. (ed.). **Computational Processing of the Portuguese Language, 11th International Conference, PROPOR 2014**, São Carlos/SP, Brazil, October 6-8, 2014, Proceedings. LNAI 8775. Heidelberg: Springer, 2014. p. 214-219.

SANTOS, D. Um novo corpo e seus desafios. *In*: FREITAS, C.; RADEMAKER, A. (ed.). **STIL 2015**: X Brazilian Symposium in Information and Human Language Technology and Collocated Events, Proceedings of the Conference, November 4 to 7, 2015. Natal, Rio Grande do Norte. p. 39-43.

SANTOS, D. José Mariano Gago - O Ministro da Língua. Apresentação no Encontro **Caminhos do conhecimento**, Leiria, 16 de Maio de 2018. Disponível em: https://www.linguateca.pt/Diana/download/LeiriaJMG.pdf

SANTOS, D. Doctors in lusophone literature. **Blog post in Digital Literary Stylistics** (SIG-DLS). 2019. Acessível de: https://dls.hypotheses.org/952

SANTOS, D. Explorando o vestuário na literatura em português. **TradTerm**, v. 37, n. 2, p. 622-643, 2021. DOI https://doi.org/10.11606/issn.2317-9511.v37p622-643

SANTOS, D.; ALVES, D. **Placing GIS and NLP in literary geography**: experiments with literature in Portuguese. Em apreciação. Disponível em: https://www.linguateca.pt/Diana/download/SantosAlves2022subm.pdf

SANTOS, D.; BICK, E. Providing Internet access to Portuguese corpora: the AC/DC project. In: GAVRILIDOU, M. et al. (ed.). **Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000.** Athens, 31 May-2 June 2000. p. 205-210.

SANTOS, D.; FREITAS, C. Estudando personagens na literatura lusófona. *In*: **STIL 2019** — XII Symposium in Information and Human Language Technology and Collocates Events, October 15-18, 2019, Salvador, BA, Proceedings of conference, 2019. p. 48-52.

SANTOS, D.; FREITAS, C.; BICK, E. OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in the public domain. *In*: **OpenCor**, Canela, RGS, Brasil, 24 de setembro de 2018. Disponível em: https://www.linguateca.pt/Diana/download/CorLex.pdf

SANTOS, D.; MARQUES, R.; FREITAS, C.; SIMÕES, A.; MOTA, C. Comparando anotações linguísticas na Gramateca: filosofia, ferramentas e exemplos. **Domínios de Linguagem,** v. 9, n. 2, abr./jun. 2015. p. 11-26. DOI https://doi.org/10.14393/DL18-v9n2a2015-2

SANTOS, D.; MOTA, C. Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. *In*: CALZOLARI, N.; CHOUKRI, K.; MAEGARD, B.; MARIANI, J.; ODIJK, J.; PIPERIDIS, S.; ROSNER, M.; TAPIAS, D. (ed.). **Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010**, 17-23 May 2010, Valletta, Malta. European Language Resources Association, 2010. p. 1437-1444.

SANTOS, D.; RANCHHOD, E. Ambientes de processamento de corpora em português: Comparação entre dois sistemas. *In*: **Actas do IV Encontro sobre o Processamento Computacional da Língua Portuguesa (Escrita e Falada), PROPOR**, Évora, 20-21 de Setembro de 1999. p. 257-268.

SANTOS, D.; SIMÕES, A.; MOTA, C. Broad coverage emotion annotation. **Language Resources and Evaluation**, 2021. DOI https://doi.org/10.1007/s10579-021-09565-1

SIMÕES, A.; SANTOS, D. *Ensinador*: corpus-based Portuguese grammar exercises. **Procesamiento del Lenguaje Natural,** v. 47, septiembre de 2011, p. 301-309.

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

SOARES DA SILVA, A. O corpus CONDIV e o estudo da convergência e divergência entre variedades do português. In COSTA, L.; SANTOS, D.; CARDOSO, N. (org.). **Perspectivas sobre a Linguateca**: Actas do encontro Linguateca : 10 anos. Linguateca, 2008. p. 25-28. Disponível em: http://www.linguateca.pt/LivroL10/Cap04-Costaetal2008-Silva.pdf

Artigo recebido em: 31.10.2021 Artigo aprovado em: 02.05.2022

DOI: 10.14393/DL52-v16n4a2022-3

Cognição e variação linguística de gêneros/registros jornalísticos: um estudo baseado em *corpus*

Linguistic variation and cognition of press genres/registers: a corpus-based study

Carlos Henrique KAUFFMANN[*]

RESUMO: As páginas de um jornal congregam diversos gêneros/registros linguísticos especializados, seja qual for o meio de acesso para a sua leitura. Nem sempre, porém, percebe-se distinção clara entre gêneros/registros jornalísticos, mesmo entre os produtores desses textos, o que leva a especular sobre a conformação e estabilidade linguística de determinados gêneros/registros. O presente estudo investiga essa questão, ao analisar um *corpus* formado por textos de duas edições da "Folha de S.Paulo", classificados em termos de gêneros/ registros por especialistas. Os resultados foram tabulados por grau de concordância entre classificadores. A maioria das classificações recaiu sobre as categorias reportagem e notícia. Em seguida, escores que refletem o consenso na determinação do gêneros/registro foram mapeados segundo as dimensões de variação do texto jornalístico de Kauffmann (2005), verificando o quão separados linguisticamente estão os gêneros/registros menos consensuais.

ABSTRACT: The newspaper is a repository of specialized genres/registers that are present in a daily basis on its pages, by any medium of access. However, as the clear distinction among press genres/registers is sometimes blurry, even by the writers of those texts, some questions about the linguistic stability of some genres/registers could be made. This study will analyze language variation from a corpus of the "Folha de S.Paulo" broadsheet newspaper, classified in terms of press genres/ registers by four expert classifiers. This task generated a database of texts classified by journalistic genres/registers, in which the majority of the texts were labeled as reportage and/or news report. Scores that reflect the degree of consensus around press registers were then plotted along the dimensions of variation of the journalistic text identified in Kauffmann (2005). This allowed to verify to what extent the average partial agreement scores of the main news reports are linked to their respective agreement categories.

[*] Doutor em linguística aplicada e estudos da linguagem pela Pontifícia Universidade Católica de São Paulo (PUCSP). ORCID: https://orcid.org/0000-0001-8792-7375. chkauffmann@corpuslg.org

PALAVRAS-CHAVE: Linguística de *Corpus*. Análise multidimensional. Gênero jornalístico. Registro.

KEYWORDS: Corpus Linguistics. Multi-dimensional analysis. Press genre. Register.

1 Introdução

A língua é sistematicamente utilizada por meio de formas discursivas que coabitam os espaços de produção linguística de qualquer sociedade, a qualquer tempo. Modernamente, um desses espaços produtivos provém da veiculação de material noticioso. Seja em formato impresso ou digital, jornais, blogs e sites informativos reúnem em suas páginas diversos gêneros, ou registros – tais como reportagem, notícia, artigo, editorial etc. – convivendo cotidianamente, cada qual possuidor de características próprias capazes de o distinguir socialmente como um grupo singular.

O registro jornalístico, visto sob um prisma mais ampliado, representa uma das referências fundamentais do uso real da língua, ao lado dos discursos falado, acadêmico e literário (BIBER *et al*., 1999). Esta visão permitiu o desenvolvimento de estudos comparativos e investigações de natureza diacrônica de registros com base em pesquisa empírica, a partir de *corpora* especializados – isto é, bases de textos coletados criteriosamente e armazenados digitalmente, na maioria das vezes etiquetados sob diversos níveis de análise linguística, de forma automática, semiautomática ou manual (BERBER SARDINHA, 2004). Por outro lado, quando observados em maior detalhe, entram em ação processos de identificação e reconhecimento de gêneros/registros jornalísticos, capazes de diferenciar um texto de outro sob esse aspecto. Tais processos são considerados habilidades importantes na educação, formação e conscientização de estudantes, uma vez que auxiliam no aprimoramento da cidadania e estimulam uma visão crítica de mundo, em conformidade com o que é recomendado pelos Parâmetros Curriculares Nacionais (BRASIL, 1998).

Não há, porém, um entendimento pacífico sobre a atribuição de um determinado texto da imprensa a um gênero ou registro – termos aqui empregados de

forma equivalente, pois as duas definem variedades linguísticas com certos propósitos comunicativos reconhecidos socialmente, passíveis de serem submetidas a uma análise linguística empírica (BIBER; CONRAD, 2009). A classificação em termos de gênero/registro é comumente feita após a publicação do texto jornalístico e independe do tema ou tópico tratado por ele. Tampouco existe delimitação clara de gêneros/registros jornalísticos, exceto por alguns explicitamente distintos, como o editorial, que por vezes é reconhecido por uma apresentação gráfica diferenciada.

Esta pesquisa buscou identificar em textos jornalísticos quais são os gêneros/registros jornalísticos considerados mais consensuais – e os mais conflitantes –, em termos de seu reconhecimento, entre observadores especialistas. Para tal, associa os recursos metodológicos da Linguística de *Corpus* (BERBER SARDINHA, 2004), abordagem que investiga padrões de língua com a utilização de *corpora* auxiliados por computador, a uma pesquisa qualitativa, que aferiu a cognição de um grupo em relação aos gêneros/registros existentes no jornal. O grau de concordância dessa classificação, tomado em conjunto, foi analisado em termos de sua chance em relação ao acaso com uma medida estatística – o kappa de Fleiss, explicado a seguir – adequada ao desenho da pesquisa. Os textos mais e menos concordes em termos de gêneros/registros foram analisados e comparados de acordo com a Análise Multidimensional (BIBER, 1988) efetuada por Kauffmann (2005), que identificou as dimensões atuantes na variação linguística do texto da imprensa no modo escrito em língua portuguesa. Desse modo, buscou-se interpretar em que medida as diferenças entre gêneros/registros jornalísticos, influenciadas pelos graus de concordância obtidos, podem ser explicadas por sua natureza linguística.

2 Pressupostos teóricos

Diversas áreas do conhecimento têm estudado a natureza e a tipologia dos gêneros/registros jornalísticos, como a de Comunicação Social (MARQUES DE MELO;

ASSIS, 2010), Linguística Aplicada (MARCUSCHI, 2002; AITCHISON; LEWIS, 2003), Análise Crítica do Discurso (VAN DIJK, 1988; FAIRCLOUGH, 1995), bem como manuais de jornalismo de caráter utilitário e profissional (FOLHA DE S. PAULO, 2018; EDITORA ABRIL, 1990; O GLOBO, 1998; O ESTADO DE S. PAULO, 1990).

Marques de Melo e Assis (2010) reúne estudos sobre diversos gêneros/registros jornalísticos na perspectiva comunicacional, tratando-os pela denominação de "formatos" (cf. COSTA, 2010), por sua vez inscritos em categorias de "gêneros" – opinativo, informativo, interpretativo, diversional ou utilitário. Marques de Melo e Assis (2016), revisitando esse modelo, considera que a imprensa escrita tenha originado tal estrutura de classificação e ainda influencia os gêneros/registros jornalísticos da atualidade, mesmo em suportes eletrônicos ou digitais.

Na área linguística, Marcuschi (2002) pôs em perspectiva uma visão de cunho bakhtiniano para os gêneros textuais, definidos como "entidades sociodiscursivas e formas de ação social incontornáveis em qualquer situação comunicativa" (MARCUSCHI, 2002, p. 19), capazes de assumir diversidade de formas e uma dinâmica histórica. O autor inscreve os gêneros textuais em grupos maiores, os domínios discursivos, que designariam "uma esfera ou instância de produção discursiva ou de atividade humana" (MARCUSCHI, 2002, p. 22) – por exemplo, o domínio jornalístico incluiria todos os gêneros/registros tratados no presente estudo. Nesse sentido, o conceito de domínio discursivo aproxima-se da definição mais ampla de registro, no sentido empregado por Biber (BIBER; CONRAD, 2009). Já os tipos de texto expressariam formas reconhecíveis no aspecto linguístico, que seriam utilizados em proporções diversas na produção de gêneros textuais, numa "construção teórica definida pela natureza linguística de sua composição" (MARCUSCHI, 2002, p. 21). São exemplos de tipos de texto a narração, a argumentação, a exposição, a descrição e a injunção (ou instrução).

Outras perspectivas, na área de Comunicação, são propostas por Chaparro (1997) e Seixas (2009). Com base em uma análise dos papéis identitários discursivos dos textos, Seixas (2009) propõe uma classificação alternativa de duas categorias gerais, embora tenha alocado nelas, em sua maioria, gêneros/registros jornalísticos já conhecidos. Chaparro (1997), por sua vez, analisa os gêneros/registros do jornal segundo esquemas de superestruturas e macroestruturas (VAN DIJK, 1988), que resultam em uma tipologia que diferencia gêneros/registros agrupados sob esquemas narrativos ou esquemas argumentativos.

Os manuais de redação publicados na imprensa brasileira descrevem características de gêneros/registros jornalísticos olhando-os sob o viés prático da produção textual especializada. Na maioria das publicações técnicas, há a distinção entre o jargão profissional que designa o texto ("matéria", entre outros termos) e a tipologia de gêneros/registros jornalísticos estabelecidos. As diversas edições do manual do jornal "Folha de S. Paulo" (FOLHA DE S.PAULO, 1984; 1987; 1992; 2001; 2018) apresentam os gêneros/registros mais conhecidos, com maior ou menor visibilidade. Por exemplo, apesar de uma edição mais recente do manual (FOLHA DE S.PAULO, 2018) não mais destacar o gênero/registro notícia como um item de um verbete, seja dedicado aos gêneros jornalísticos (FOLHA DE S. PAULO, 2001), ou como verbete próprio (FOLHA DE S.PAULO, 1992), sua presença ainda é efetiva no capítulo voltado à prática jornalística, conforme a publicação qualifica situações como o uso de "notícia exclusiva" ou, ainda, de *fake news* ("notícia falsa").

Estudos fundamentais sobre gêneros/registros jornalísticos na Linguística Aplicada originaram-se da área de Análise Crítica do Discurso (VAN DIJK, 2010; FAIRCLOUGH, 1995; BONINI, 2012), embora outras abordagens (BELL, 1991; AITCHINSON; LEWIS, 2003) tenham também expandido a discussão.

Com um viés cognitivo, no propósito de investigar o conhecimento de gêneros/registros jornalísticos entre uma comunidade linguística que lida com o texto

de imprensa profissionalmente, Bonini (2002) apresentou trechos e textos de alguns gêneros/registros a jornalistas, para efetuar uma classificação, sem que tivessem prévio conhecimento metalinguístico sobre os gêneros/registros. O tipo de avaliação utilizada levou a uma lista aberta de gêneros/registros, e ao uso de jargão por alguns observadores. Outros trabalhos do mesmo autor investigaram as designações ou rótulos dos gêneros/registros, entre eles Bonini (2003). Por sua vez, Bonini (2009) analisou a distinção existente entre os gêneros/registros notícia e reportagem com a metodologia de Swales (1990), que envolve a análise de movimentos retóricos, concluindo que há um *continuum* fluido interligando os dois gêneros/registros.

A Análise Multidimensional (BIBER, 1988; BERBER SARDINHA, 2000) foi a metodologia escolhida para mensurar a variação linguística entre gêneros/registros efetuada pela pesquisa de classificação por observadores especialistas, grupo composto por profissionais da área jornalística com formação acadêmica de nível superior. Originalmente, a Análise Multidimensional (AMD) centrou-se no estudo de variação entre os vários gêneros/registros que compõem a língua, nos modos escrito e oral, em inglês e outras línguas (BIBER, 1995), e sempre incluiu gêneros/registros ligados ao jornalismo. Na AMD realizada para o português do Brasil (BERBER SARDINHA et al., 2014b), por exemplo, entre os 48 gêneros/registros estudados, sete estão ligados à mídia: reportagem, revista de notícias, revista de celebridades, editoriais, crônicas (modo escrito), e entrevistas publicadas e notícias de TV (modo oral). Com base no estudo foi possível concluir que os gêneros/registros de mídia são letrados, não procedurais e com foco no passado; tomados individualmente, os gêneros/registros editorial, crônica e entrevista são mais argumentativos que os demais (KAUFFMANN, 2015).

Por meio da AMD, é possível também estudar subgrupos de gêneros/registros, como os de cunho jornalístico (KAUFFMANN, 2005) ou acadêmico (BIBER, 2006), além de gêneros/registros isolados, como por exemplo filmes norte-americanos

(VEIRANO PINTO, 2013), músicas pop (BÉRTOLI-DUTRA, 2010) e reportagens de capa da revista "Time" (SOUZA, 2012). Outros estudos utilizaram corpora de textos jornalísticos para análise de variação da linguagem (BIBER *et al*., 1999; BEDNAREK, 2006; BIBER; CONRAD, 2009, p. 116-7).

As perguntas de pesquisa que motivaram o presente trabalho poderiam ser elencadas da seguinte forma:

> 1. Qual é a medida de concordância geral obtida na classificação de gêneros/registros jornalísticos por meio de observadores especialistas?
> 2. Quais são os gêneros/registros jornalísticos considerados em maior e menor grau de concordância na classificação de gêneros/registros jornalísticos por meio de observadores especialistas?
> 3. Como os gêneros/registros jornalísticos e seus diversos subgrupos de concordância estão mapeados segundo as dimensões de variação do texto jornalístico de Kauffmann (2005)?

Para respondê-las, serão apresentados previamente detalhes da metodologia da AMD, do coeficiente utilizado nas medidas de concordância da classificação e do desenho do *corpus*. A seguir, serão apresentados os resultados da pesquisa e, concomitantemente, sua discussão em relação a outros estudos da área.

3 Metodologia

A consecução de uma AMD pode ser total (BIBER, 1988; BERBER SARDINHA et al., 2014b) ou parcial, como no caso deste estudo, em que é realizada uma adição de gêneros/registros a uma AMD total (BERBER SARDINHA *et al*., 2019). A AMD envolve algumas etapas básicas (BERBER SARDINHA; VEIRANO PINTO, 2019), entre elas: 1. a coleta de um *corpus* extenso e balanceado da língua ou variedade de língua em uso, de modo a criar uma amostra de textos que representam os vários gêneros/registros que a compõem; 2. o uso de software capaz de anotar cada palavra em termos de suas características morfossintáticas, e de algoritmos que permitem contar e processar os dados resultantes do primeiro procedimento, para tabulá-los em matrizes em que são

levadas em conta algumas variáveis selecionadas, a partir de análise da anotação inicial; 3. o alinhamento dos tamanhos dos textos que compõem o *corpus* de estudo, por meio da normalização das frequências absolutas para uma frequência relativa, por mil palavras; 4. a análise fatorial, técnica estatística multivariada que busca investigar no *corpus* a existência de fatores latentes – grupos de variáveis coocorrentes que indicam uma influência não identificada originalmente na pesquisa –, ou dimensões; 5. o cálculo da padronização das variáveis e dos escores médios de gênero/registro nas dimensões, medidas que permitem a análise e a interpretação da variação linguística encontrada nas dimensões. Neste estudo, apenas a última etapa da AMD será executada, de modo a incorporar os dados relativos à concordância de gêneros/registros jornalísticos à pesquisa original de AMD efetuada em Kauffmann (2005).

O *corpus* coletado para representar a variedade de gêneros/registros jornalísticos no modo escrito foi composto por sete edições integrais da edição impressa do jornal "Folha de S.Paulo" (idêntica à versão digital, oferecida no formato pdf, e muito semelhante à edição online), de modo a possibilitar a construção de uma semana construída (KENNEDY, 1998, p. 75; KAUFFMANN, 2005) com um total de 1.431 textos. Todos foram processados pelo etiquetador on-line VISL, para a língua portuguesa (BICK, 2005), uma versão simplificada do PALAVRAS (BICK, 2014), posteriormente desenvolvido. Das 19 variáveis de ordem lexicogramatical que inicialmente compuseram a AMD, foram selecionadas por fim 13 variáveis na extração final da análise fatorial, com peso significativo (acima do valor de 0,3). A matriz padrão expressa na Tabela 1 reúne as variáveis coocorrentes da solução de dois fatores em Kauffmann (2005), interpretados como dimensões: Narrativo versus Expositivo (Dimensão 1) e Argumentativo versus Informativo (Dimensão 2). Quando uma variável tem um valor negativo, a relação de co-ocorrência é inversa: enquanto as

variáveis positivas do fator se manifestam em um texto, as negativas tendem a ficar ausentes.

Tabela 1 – Composição e peso das variáveis nas dimensões da AMD.

Variável	**Dimensão 1**	**Dimensão 2**
Pretérito perfeito	0,72	-0,36
Verbos / pronomes 3ª pessoa singular	0,64	
Conjunções subordinativas	0,52	0,40
Verbos públicos (falar, afirmar, dizer, etc.)	0,44	
Pretérito imperfeito	0,37	
Advérbios	0,35	
Verbos / pronomes 1ª pessoa singular	0,33	
Substantivos	-0,35	
Presente do indicativo		0,55
Pronomes demonstrativos		0,48
Quantidade de palavras		0,34
Nomes próprios		-0,58
Números cardinais		-0,35

Fonte: Kauffmann (2005, p. 92).

O peso que cada variável teve no cálculo de escores das dimensões é indireto. O escore é produto da soma das frequências relativas de cada texto nas variáveis de maior valor entre as dimensões (BIBER, 1988, p. 93), e não do peso relativo de cada variável na composição da dimensão. Os escores médios por gênero/registro foram resultantes da média das frequências padronizadas das observações (textos) agrupadas nos gêneros/registros. As análises de variância ANOVA (Tabela 2) realizadas *a posteriori* mostraram que as dimensões encontradas são significativamente relevantes para explicar a variação encontrada no *corpus* em relação aos diversos gêneros/registros jornalísticos.

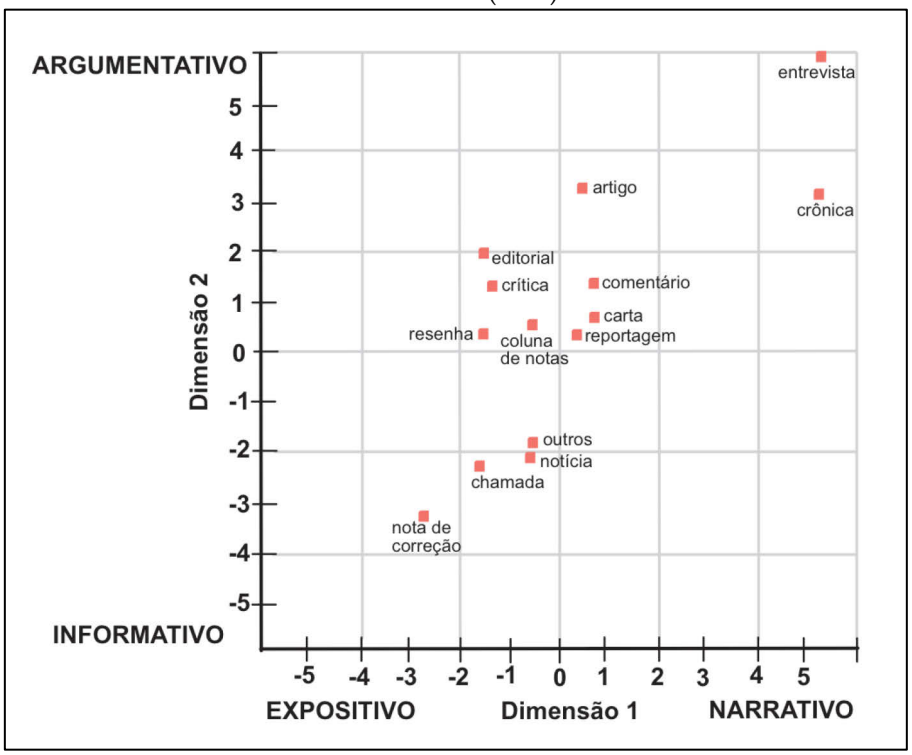
Tabela 2 – ANOVA entre gêneros/registros, por dimensão (valores de F, *p* e R^2).

Dimensão	F	*p*	R^2
Dimensão 1	6,7	0,00000	5,8%
Dimensão 2	50,4	0,00000	31,6%

Fonte: Kauffmann (2005, p. 109).

As dimensões resultantes em Kauffmann (2005) assemelham-se às dimensões de variação D2 e D5 do português (BERBER SARDINHA *et al.*, 2014b). A Dimensão 1 de Kauffmann (2005) tem, no polo narrativo, variáveis de tempos verbais que coincidem com o posicionamento para o passado da Dimensão 5 de Berber Sardinha et al. (2014b) – pretéritos perfeito e imperfeito; já a Dimensão 2 de Kauffmann (2005) é semelhante no propósito argumentativo, embora ambas não compartilhem as variáveis que caracterizam essa dimensão (cf. discussão em BERBER SARDINHA *et al.*, 2014a).

Gráfico 1: Distribuição dos escores médios de gêneros/registros jornalísticos nas dimensões de Kauffmann (2005).



Fonte: adaptado de Kauffmann (2005, p. 115).

Com o cálculo dos escores médios por gênero/registro, foi possível posicionar graficamente cada gênero/registro ao longo dos eixos representados pelas dimensões resultantes (Gráfico 1). Os gêneros/registros entrevista, artigo, crônica, comentário, carta e reportagem pertencem aos quadrantes "Narrativo-Argumentativo"; em quadrante oposto, os registros chamada, notícia e nota de correção estão presentes no quadrante "Expositivo-Informativo". O quadrante "Expositivo-Argumentativo" concentra, em uma terceira região, os gêneros/registros editorial, resenha, crítica e coluna de notas (KAUFFMANN, 2005, p. 115).

A presente análise da classificação de gêneros/registros jornalísticos por especialistas, aqui relatada, originalmente apoiou a pesquisa conduzida em Kauffmann (2005), para confirmar a hipótese de que uma classificação efetuada segundo uma criteriosa revisão na literatura a respeito dos gêneros/registros na imprensa diária escrita do Brasil, por um só especialista (no caso, o autor), seria largamente coincidente com uma classificação de gêneros/registros jornalísticos feita por outros observadores especialistas, de forma independente. Foi logo percebido, porém, que os resultados tiveram a capacidade de iluminar alguns aspectos relativos à cognição de gêneros/registros jornalísticos por observadores especialistas, ultrapassando os propósitos iniciais do levantamento, motivo pelo qual justificou-se uma análise mais detida sobre os dados gerados.

A pesquisa desenvolvida lançou ao escrutínio de mais três especialistas uma amostra significativa do *corpus* de pesquisa de Kauffmann (2005), gerando dados inéditos relativos ao grau de concordância entre os gêneros/registros jornalísticos. O *corpus* abrangido pela pesquisa é composto por duas edições completas da "Folha", escolhidas aleatoriamente, perfazendo um total de 425 textos (136.330 palavras) por classificar – equivalentes a 29,6% do total de textos estudados em Kauffmann (2005). Cada um deles realizou a sua classificação individualmente, com instruções, planilha de preenchimento, cópias impressas de cada texto e a edição impressa onde foram

originalmente publicados, acompanhado de uma lista de gêneros/registros jornalísticos com 14 itens previamente estabelecida, e material de apoio com suas respectivas definições e referências, provenientes de fontes da literatura, como livros acadêmicos da área de comunicação e jornalismo, manuais de redação e estilo e dicionários especializados. Em comum, os observadores especialistas apresentavam experiência profissional jornalística acumulada por vários anos nas áreas editorial e de documentação em empresas de comunicação ligadas à produção de jornais. Nenhum deles teve contato com os demais para discutir assuntos ligados à classificação efetuada. Como resultado, chegou-se a um *corpus* de textos jornalísticos classificado de acordo com a tipologia usual de gêneros/registros jornalísticos identificada anteriormente pela literatura, porém graduada pela convergência – total ou parcial – ou divergência na atribuição de determinado gênero/registro a cada texto, entre os quatro especialistas.

O cômputo das menções de gênero/registro dadas aos textos do *corpus*, mostrado adiante, foi tabulado para refletir o grau de concordância/discordância entre gêneros/registros. A fim de interpretar os resultados derivados da pesquisa, um modo encontrado de analisar e sintetizar a opinião de um grupo de quatro classificadores, com um número razoável de alternativas à escolha, é proposto no Quadro 1.

Quadro 1 – Graus de concordância entre quatro observadores.

Grau de concordância	**Menções de gênero(s)/registro(s) dadas a um texto do *corpus***
Concordância total	Quatro menções de um mesmo gênero/registro
Concordância predominante	Três menções de um mesmo gênero/registro e uma menção única de gênero/registro diverso
Concordância parcial	Duas menções duplas de um mesmo gênero/registro (Divisão 2 – 2) ou Uma menção dupla e duas menções únicas de gêneros/registros diversos (Divisão 2 – 1 – 1)
Discordância	Quatro menções únicas de gêneros/registros diversos

Fonte: elaborado pelo autor.

A distinção permite uma primeira análise dos casos concordes (aqueles com graus de concordância total e predominante) para a identificação de dimensões de linguagem capazes de distinguir entre os gêneros/registros jornalísticos mais frequentes do *corpus*, de forma avalizada.

Para avaliar de forma mais consistente os resultados, o recurso de mensuração da concordância interobservadores utilizado nesta pesquisa é o coeficiente kappa de Fleiss, uma medida de associação mais precisa que o percentual de concordância, pois ele verifica em que medida há concordância considerando as chances de acaso entre vários observadores (BIBER; EGBERT; DAVIES, 2015, p. 20). Foi utilizada uma variação do kappa de Fleiss, que considera o fato de que os classificadores não têm um número de casos fixado para cada categoria (RANDOLPH, 2005). Um valor de kappa igual ou acima de 0,7 é normalmente utilizado para indicar que houve concordância entre os observadores. Um valor a partir de 0,8 é considerado uma concordância "quase completa" (MIOT, 2016, p. 91).

4 Resultados

A classificação realizada pelo grupo de observadores especialistas resultou em vários conjuntos de textos, que apresentam níveis de concordância quanto ao gênero/registro, em maior ou menor grau: grupos de gêneros/registros com textos consensualmente percebidos pelos observadores de forma plena ou parcial, grupos de textos com divergência de opinião sobre gênero/registro etc. Todavia, tomados em conjunto, a classificação efetuada pelos observadores especialistas apresentou um coeficiente de concordância de 0,76 (kappa de Fleiss), indicando uma significativa concordância do grupo em relação às designações dos gêneros/registros jornalísticos atribuídos aos textos.

A Tabela 3 apresenta os resultados da pesquisa, organizados pelo número de textos atribuídos pelos classificadores de acordo com o grau de concordância, envolvendo todos os gêneros/registros jornalísticos mencionados na pesquisa. Mais de

86% dos textos do *corpus*, correspondentes à soma dos textos em que houve concordância total ou predominante, foram considerados pelo grupo de especialistas como portadores de características patentes que os fazem integrantes de um determinado gênero/registro jornalístico reconhecido socialmente.

Tabela 3 – Concordância para gêneros/registros jornalísticos de 425 textos.

Concordância total	Concordância predominante	Concordância parcial		Discordância
		Divisão 2 – 2	Divisão 2 – 1 – 1	
267	100	30	26	2
62,8%	23,5%	7,1%	6,1%	0,5%

Fonte: elaborada pelo autor.

Para caracterizar em maior detalhe como se dá a distribuição de gêneros/registros concordes no meio jornalístico, a Tabela 4 apresenta a proporção que possui cada gênero/registro concorde em relação ao total de textos, com base na soma das avaliações de concordância total e predominante. A mesma tabela exibe também, abaixo da linha destacada, o total de textos e a proporção atingida pelos textos que obtiveram classificação de concordância parcial ou discordância.

Tabela 4 – Frequência de gêneros/registros jornalísticos concordes no *corpus*.

Gênero/Registro	Número de textos	Percentual
Reportagem	201	47,3%
Notícia	52	12,2%
Crítica	22	5,2%
Chamada	20	4,7%
Artigo	19	4,5%
Carta	18	4,2%
Coluna de notas	15	3,5%
Editorial	7	1,6%
Entrevista	5	1,2%
Outros	4	0,9%
Crônica	2	0,5%
Nota de correção	2	0,5%
Resenha	0	0,0%
Comentário	0	0,0%

Concordância parcial de gêneros/registros	56	13,2%
Discordância	2	0,5%
Total	425	100%

Fonte: elaborada pelo autor.

Em relação à proporção de gêneros/registros encontrada em Kauffmann (2005, p. 105) e Chaparro (1997, p. 139), o percentual atribuído à notícia discrepa à primeira vista – 20% nos estudos anteriores, em contraste com os 12,2% desta pesquisa –, enquanto a proporção atribuída ao gênero/registro reportagem, de 50%, é semelhante, e o mesmo ocorre em relação aos demais gêneros/registros. No entanto, como será demonstrado a seguir, a proporção do gênero/registro notícia alcança valores próximos aos dos estudos anteriores, se for levada em consideração sua presença hegemônica na categoria de "Concordância parcial de registros".

Cabe analisar em detalhe, portanto, os casos em que houve dissenso na classificação. Biber, Egbert e Davies (2015) decompuseram a concordância parcial sobre gêneros/registros quando analisaram a classificação de quatro observadores a respeito de gêneros/registros em textos da internet, dividindo-a em dois grupos: o de duas menções duplas de dois gêneros/registros; e o de uma menção dupla de um gênero/registro, com duas menções únicas de gêneros/registros diversos. No primeiro caso, os autores aventam a possibilidade de haver gêneros/registros híbridos que possuem características comuns dos dois gêneros/registros para conformar um tipo de gênero/registro híbrido. Nesta pesquisa, igualmente foram mantidos os subgrupos da categoria de concordância parcial, no intuito de investigar em que medida as diferenças de concordância parcial se refletem na mensuração dos grupos de textos, em relação às dimensões de variação do texto jornalístico de Kauffmann (2005), e se é possível admitir uma ideia de gênero/registro híbrido com as evidências disponíveis.

Os textos que apresentaram uma classificação de concordância parcial do tipo Divisão 2 – 2 totalizam 30 textos. Seus gêneros/registros componentes estão

apresentados na Tabela 5. Evidencia-se a tensão terminológica entre os gêneros/registros reportagem e notícia, com a presença de 22 textos no *corpus* com menções duplas a esses gêneros/registros, somando dois terços dos casos nessa situação de concordância. Destaca-se secundariamente a divisão de opiniões entre reportagem e crítica, existente em quatro textos.

Tabela 5 – Gêneros/registros com concordância parcial (Divisão 2 – 2).

Gêneros/registros com concordância parcial (Divisão 2– 2)		Nº de textos
Reportagem	Notícia	22
Reportagem	Crítica	4
Notícia	Chamada	1
Artigo	Crítica	1
Artigo	Comentário	1
Artigo	Crônica	1
Total		30

Fonte: elaborada pelo autor.

Por sua vez, a análise dos resultados de concordância parcial que obedecem à Divisão 2 – 1 – 1, exibida na Tabela 6, revela um número relevante de textos em que os gêneros/registros reportagem e notícia estão presentes, como gênero/registro majoritário ou não, na classificação composta em que estão sempre presentes três gêneros/registros. Nessa categoria de baixa concordância, notícia e reportagem estão associados a crítica, "outros", coluna de notas, artigo, carta e chamada.

Tabela 6 – Gêneros/registros com concordância parcial (Divisão 2 – 1 – 1).

Gêneros/registros com concordância parcial (Divisão 2 – 1 – 1)			**Nº de textos**
Menção dupla	Menção única	Menção única	
Notícia	Reportagem	Outros	8
Artigo	Comentário	Coluna de notas	5
Reportagem	Notícia	Crítica	4
Reportagem	Resenha	Crítica	2
Reportagem	Notícia	Artigo	1

Reportagem	Notícia	Coluna de notas	1
Notícia	Reportagem	Carta	1
Notícia	Crítica	Outros	1
Notícia	Comentário	Outros	1
Chamada	Reportagem	Notícia	1
Artigo	Reportagem	Comentário	1
Total			26

Fonte: elaborada pelo autor.

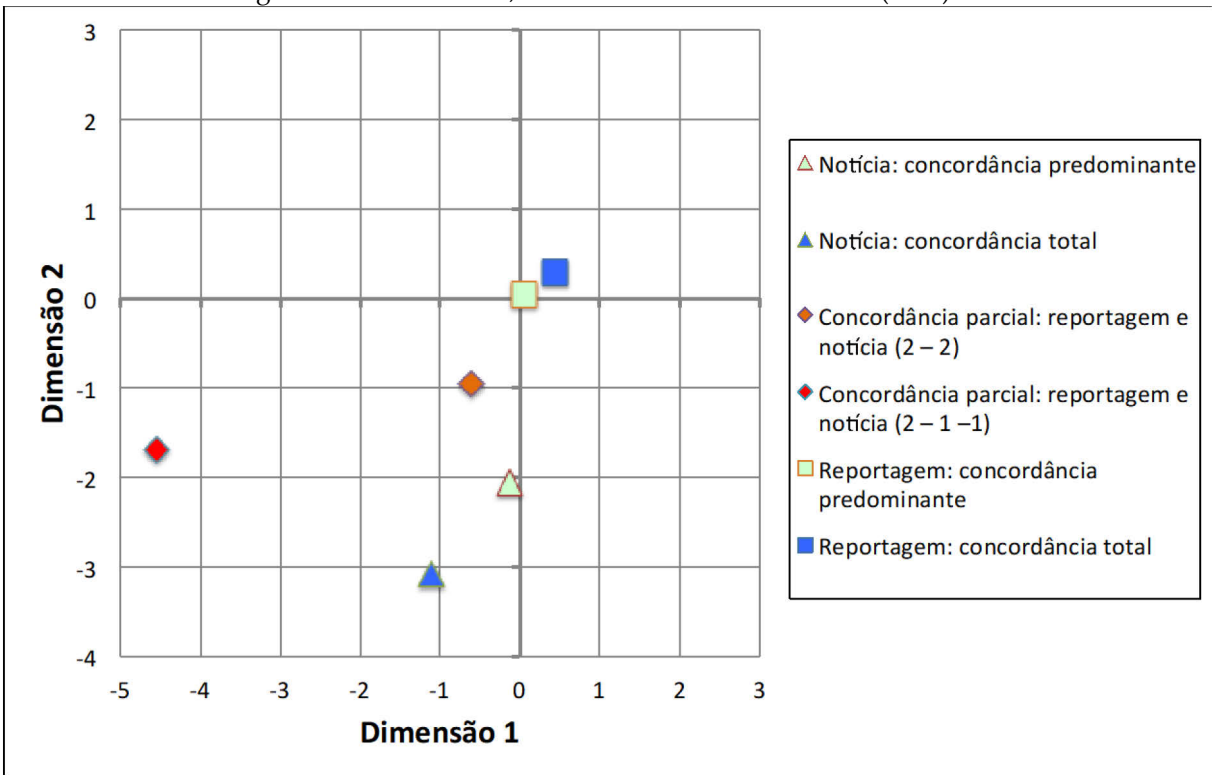
Quanto aos casos de discordância, conforme visto na Tabela 4, eles só ocorreram duas vezes no *corpus*. Devido à interpretação complexa e ao baixo número apresentado, não serão comentados.

Observou-se que a principal divisão de opiniões entre os observadores especialistas está concentrada na oposição reportagem *versus* notícia. É possível argumentar que a razão para ocorrer essa dicotomia é a proximidade formal entre os dois gêneros/registros, emblemáticos da atividade jornalística. Desse modo, o conflito mostrado nas Tabelas 5 e 6 influenciou significativamente a baixa proporção de textos concordes classificados como notícia, conforme mostrado anteriormente. Por sua vez, observou-se que o gênero/registro reportagem sofreu impacto atenuado de variação de proporção devido à concordância parcial, pelo fato de ser o gênero/registro de maior expressão no *corpus*.

Para examinar de uma forma empírica as diferenças encontradas nos gêneros/registros jornalísticos reportagem e notícia em termos de sua concordância, recorreu-se ao modelo de mensuração de variação linguística proporcionado pela Análise Multidimensional aditiva (BERBER SARDINHA *et al*., 2019). A Análise Multidimensional permite que se disponham nas suas dimensões as respectivas frequências manifestadas nos grupos de textos, em uma escala. Aqui, os grupos de textos submetidos à análise são compostos pelos gêneros/registros concordes de reportagem e notícia – ou seja, textos que obtiveram concordância total ou predominante na classificação por observadores especialistas – e por uma seleção dos grupos de textos que obtiveram concordância parcial, especificamente aqueles textos

classificados como notícia ou reportagem. Ou seja, dos 30 textos com concordância parcial (Divisão 2 – 2), foi selecionado o grupo de 22 textos que possui menções duplas simultaneamente de gêneros/registros reportagem e notícia; enquanto dos 26 textos na situação de concordância parcial (Divisão 2 – 1 –1), foi selecionado o grupo de 15 textos com menções duplas em reportagem ou notícia, com uma menção única do outro gênero/registro, em complementação. Os resultados estão mostrados no Gráfico 2.

Gráfico 2 – Mapeamento dos escores médios dos gêneros/registros reportagem e notícia, conforme grau de concordância, nas dimensões de Kauffmann (2005).



Fonte: elaborado pelo autor.

Os escores médios dos gêneros/registros notícia e reportagem que têm concordância total/predominante estão bem próximos dos resultados obtidos em Kauffmann (2005). Apesar de não apresentados graficamente, as contagens dos demais grupos de gêneros/registros jornalísticos de concordância total/predominante apresentaram posicionamento semelhante à disposição dos gêneros/registros nos quadrantes formados pelas dimensões de Kauffmann (2005), exceto talvez pelo

gênero/registro artigo, que mudou de quadrante, conforme mostram os dados apresentados no Apêndice.

Os escores médios dos grupos de gêneros/registros notícia e reportagem com concordância predominante estão próximos dos escores de concordância total de seus respectivos gêneros/registros, indicando uma afinidade de características compartilhadas. Esses escores, em alguma medida, têm pequenas diferenças que os afastam do grupo de concordância total, e seus pontos correspondentes no Gráfico 2 indicam que essas diferenças os aproximam na direção do gênero/registro oposto.

No grupo de concordância parcial em que notícia e reportagem têm opiniões divididas (modelo 2 – 2), o ponto que o representa no Gráfico 2 está em uma região praticamente equidistante das regiões ocupadas pelos gêneros/registros concordes de reportagem e notícia, nas Dimensões 1 e 2. Esse posicionamento sugere que as características linguísticas que os textos jornalísticos noticiosos carregam têm uma considerável participação na distinção de gêneros/registros efetuada por observadores especialistas. A configuração linguística se equilibra em uma zona intermediária entre os dois gêneros/registros. Um exemplo de texto nessa situação é mostrado a seguir:

Exemplo 1 – Concordância parcial entre reportagem e notícia (Divisão 2 – 2)

> Emenda do BC é promulgada por Sarney
> DA SUCURSAL DE BRASÍLIA
> Em rápida sessão do Congresso, realizada no plenário do Senado com a presença de apenas cinco parlamentares, o presidente do Senado, José Sarney (PMDB-AP), promulgou a emenda constitucional que permite a regulamentação do sistema financeiro nacional por várias leis complementares –e não mais por apenas uma lei, como a Constituição determinava, em seu artigo 192.
> A emenda, que abre caminho para a aprovação de uma proposta de autonomia do Banco Central, revogou todos os incisos e parágrafos desse dispositivo constitucional –que trata do sistema financeiro–, inclusive o que estabelecia o teto de 12% ao ano para as taxas de juros reais.
> A proposta original de alteração do artigo 192 foi apresentada em 1997 pelo então senador José Serra (PSDB-SP). Na CCJ (Comissão de Constituição e Justiça), recebeu substitutivo do senador Jefferson Péres (PDT-AM), que foi enviado à Câmara e aprovado pelos deputados neste mês de maio em segundo turno.

> A emenda mantém apenas o caput do artigo 192 com uma alteração, determinando que o sistema financeiro nacional será regulado em leis complementares. (CP5_0788.TXT)

Nos escores médios dos textos de concordância parcial no modelo de divisão 2 – 1 – 1, a Dimensão 2 parece indicar um papel mais importante que o da Dimensão 1 em relação à distinção dos gêneros/registros. A Dimensão 2, ligada a aspectos de argumentação *versus* informação, possivelmente seja um agente diferenciador mais eficaz entre os dois gêneros/registros, sendo a notícia mais vinculada com o aspecto informativo, enquanto a reportagem teria, por outro lado, um viés argumentativo. Em relação à Dimensão 1, esses grupos de concordância parcial de reportagem e notícia em composição com outros gêneros/registros se diferenciam dos demais grupos encontrados. Valores negativos altos na dimensão indicam que esse grupo possui textos com poucos verbos e complementos verbais, enquanto apresentam uma alta frequência de substantivos. O Exemplo 2 ilustra as características do grupo:

Exemplo 2 – Concordância parcial entre reportagem e notícia (Divisão 2 – 1 – 1)

> Como são transmitidas as informações
> ESPECIAL PARA A FOLHA
> São três as principais formas de conectar periféricos sem fios: infravermelho, bluetooth e radiofrequência.
> O infravermelho é comum em equipamentos portáteis. Basta apontar um equipamento para o outro a uma distância de até 40 cm para que eles se enxerguem. A velocidade de transmissão de dados chega a 4 Mbps.
> A tecnologia bluetooth funciona com ondas de radiofrequência. Seu alcance é de aproximadamente 10 metros, e a velocidade de conexão pode atingir uma taxa de até 720 Kbps.
> A radiofrequência (802.11) pode alcançar 11 Mbps, mas logo haverá placas que possibilitarão a transmissão de dados com velocidades superiores a 50 Mbps.
> Alguns fabricantes já vendem produtos com a tecnologia 802.11g, que deve alcançar uma velocidade de 54 Mbps, ou seja, quase cinco vezes mais rápida. (JAR) (CP3_0533.TXT)

Os gêneros/registros reportagem e notícia mantêm uma certa ambiguidade discursiva. Por um lado, subsiste uma separação clara entre os grupos concordes dos gêneros/registros na percepção de especialistas, que reconhece a existência de traços discerníveis na produção de uma notícia ou de uma reportagem. Porém, de outra parte, pode-se admitir que se manifestam concomitantemente grupos intermediários de textos que estão a meio caminho dos dois gêneros/registros, gerando divergências na sua classificação – por exemplo quando, em um texto jornalístico, a carga informativa concentrada da notícia equilibra-se com os fatos narrados pela ação da reportagem.

5 Considerações finais

Os dados apresentados permitiram observar os gêneros/registros que atuam na imprensa diária escrita no Brasil sob um ponto de vista abrangente, tanto de ordem qualitativa como quantitativa. Através dos métodos empregados na pesquisa, foi possível concluir que 86,3% (367 textos, de um total de 425) dos textos classificados pelo grupo de observadores especialistas são concordes. Os gêneros/registros mais bem diferenciados pelo grupo foram editorial, nota de correção, chamada, carta, entrevista e reportagem. Já os gêneros/registros crônica, crítica, notícia, comentário e resenha demonstraram ser aqueles que apresentaram maior divergência de opinião.

Os achados confirmam a pesquisa de Bonini (2009), que por outros meios teóricos também detectou que existe uma interseção entre as áreas que delimitam os gêneros/registros reportagem e notícia. Nesse aspecto, a presente pesquisa não mostrou que essas zonas intermediárias teriam autonomia linguística capaz de produzir novas terminologias de registro. Assim, a hipótese de existência de gêneros/registros híbridos, como proposta por Biber, Egbert e Davies (2015), parece não se confirmar com os dados encontrados no ambiente do jornal, certamente mais restrito que o universo da *web*. Pode-se aventar que seja efeito de textos com

componentes dos dois gêneros/registros que teriam levado à indecisão na classificação, mas seria preciso efetuar novas pesquisas para investigar esse processo. Espera-se, por fim, que os resultados obtidos possam vir a contribuir com outros trabalhos que abordem aspectos relativos à cognição de gêneros/registros da imprensa.

Referências bibliográficas

AITCHISON, J.; LEWIS, D. M. (org.). **New media language**. Abingdon; New York: Routledge, 2003. DOI https://doi.org/10.4324/9780203696965

BEDNAREK, M. **Evaluation in media discourse**: analysis of a newspaper corpus. London; New York: Continuum, 2006.

BELL, A. **The Language of News Media**. Oxford: Blackwell, 1991.

BERBER SARDINHA, T. Análise Multidimensional. **D.E.L.T.A.**, v. 16, n. 1, p. 99-127, 2000. DOI https://doi.org/10.1590/S0102-44502000000100005

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

BERBER SARDINHA, T. ET AL. A multi-dimensional analysis of register variation in Brazilian Portuguese. **Corpora**, v. 9, n. 2, p. 239-271, 2014 (a). DOI https://doi.org/10.3366/cor.2014.0059

BERBER SARDINHA, T. ET AL. Dimensions of register variation in Brazilian Portuguese. *In*: BERBER SARDINHA, T.; VEIRANO PINTO, M. (org.). **Multi-Dimensional Analysis, 25 years on**: a tribute to Douglas Biber. Amsterdam: John Benjamins, 2014 (b). p. 35-79. DOI https://doi.org/10.1075/scl.60

BERBER SARDINHA, T. *et al*. Adding registers to a previous Multi-Dimensional Analysis. *In*: BERBER SARDINHA, T.; VEIRANO PINTO, M. (org.). **Multi-Dimensional Analysis**: research methods and current issues. London: Bloomsbury, 2019. p. 165-186. DOI https://doi.org/10.5040/9781350023857.0017

BÉRTOLI-DUTRA, P. **Linguagem da música popular anglo-americana de 1940 a 2009**. 2010. 290 f. Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem) – LAEL, Pontifícia Universidade Católica de São Paulo, São Paulo, 2010.

BIBER, D. **Variation across speech and writing**. Cambridge: Cambridge University Press, 1988. DOI https://doi.org/10.1017/CBO9780511621024

BIBER, D. **Dimensions of register variation**: A Cross-linguistic Comparison. Cambridge: Cambridge University Press, 1995. DOI https://doi.org/10.1017/CBO9780511519871

BIBER, D. **University language**: a corpus-based study of spoken and written registers. Amsterdam; Philadelphia: John Benjamins, 2006. DOI https://doi.org/10.1075/scl.23

BIBER, D.; CONRAD, S. **Register, genre and style**. Cambridge: Cambridge University Press, 2009. DOI https://doi.org/10.1017/CBO9780511814358

BIBER, D. *et al*. **Longman grammar of spoken and written English**. London: Longman, 1999.

BIBER, D.; EGBERT, J.; DAVIES, M.. Exploring the composition of the searchable web: a corpus-based taxonomy of web registers. **Corpora**, n. 101, p. 11-45, 2015. DOI https://doi.org/10.3366/cor.2015.0065

BICK, E. Gramática constritiva na análise automática da sintaxe portuguesa. *In*: BERBER SARDINHA, T. (org.). **A língua portuguesa no computador**. Campinas: Mercado de Letras, 2005. p. 91-112.

BICK, E. PALAVRAS, a constraint grammar-based parsing system for Portuguese. *In*: BERBER SARDINHA, T.; FERREIRA, T. S. B. (org.). **Working with Portuguese corpora**. London; New York: Bloomsbury; Continuum, 2014.

BONINI, A. **Gêneros textuais e cognição**. Florianópolis: Insular, 2002.

BONINI, A. Os gêneros do jornal: o que aponta a literatura da área de comunicação do Brasil? **Linguagem em (Dis)curso**, v. 4, n. 1, jul./dez., 2003.

BONINI, A. The distinction between news and reportage in the brazilian journalistic context: a matter of degree. *In*: BAZERMAN, C.; BONINI, A.; FIGUEIREDO, D. (org.). **Genre in a changing world**. Perspectives on writing. Fort Collins, Colorado: The WAC Clearinghouse and Parlor Press, 2009. Disponível em: https://wac.colostate.edu/books/genre/. Acesso em: 23 out. 2021. DOI https://doi.org/10.37514/PER-B.2009.2324.2.10

BONINI, A. Análise crítica de gêneros jornalísticos. SBPJor – Associação Brasileira de Pesquisadores em Jornalismo. **10º Encontro Nacional de Pesquisadores em Jornalismo**. Pontifícia Universidade Católica do Paraná, Curitiba, 2012.

BRASIL. **Parâmetros Curriculares Nacionais** – 5ª a 8ª série do Ensino Fundamental. Brasília: SEF/MEC, 1998.

CHAPARRO, M. C. **Jornalismo, discurso em dois gêneros**. 1997. 261 p. Tese (Livre Docência) – Escola de Comunicações e Artes, USP, São Paulo, 1997.

COSTA, L.A. Gêneros jornalísticos. *In*: J. MARQUES DE MELO; F. de ASSIS (org.). **Gêneros jornalísticos no Brasil**. São Bernardo do Campo: Universidade Metodista de São Paulo, 2010. p. 43-83.

EDITORA ABRIL. **Manual de estilo**. Rio de Janeiro: Nova Fronteira, 1990.

FAIRCLOUGH, N. **Media discourse**. London: Edward Arnold, 1995.

FOLHA DE S.PAULO. **Manual geral da Redação**. São Paulo: Folha de S.Paulo, 1984.

FOLHA DE S.PAULO. **Manual geral da Redação**. São Paulo: Folha de S.Paulo, 1987.

FOLHA DE S.PAULO. **Novo manual da Redação**. São Paulo: Folha de S.Paulo, 1992.

FOLHA DE S.PAULO. **Manual da Redação**. São Paulo: Publifolha, 2001.

FOLHA DE S.PAULO. **Manual da Redação**. São Paulo: Publifolha, 2018.

KAUFFMANN, C. H. **O corpus do jornal**: variação linguística, gêneros e dimensões da imprensa diária escrita. 2005. 202 f. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) – LAEL, Pontifícia Universidade Católica de São Paulo, São Paulo, 2005.

KAUFFMANN, C. H. Caracterização linguística de gêneros textuais do português brasileiro. **VIII Simpósio Internacional de Estudos de Gêneros Textuais** (SIGET), 2015. Comunicação. São Paulo: 2015.

KENNEDY, G. **An Introduction to corpus linguistics**. London: Longman, 1998.

MARCUSCHI, L. A. Gêneros textuais: definição e funcionalidade. *In*: DIONISIO, A. P.; MACHADO, A. R.; BEZERRA, M. A. (org.). **Gêneros textuais e ensino**. Rio de Janeiro: Lucerna, 2002. p. 19–36.

MARQUES DE MELO, J.; ASSIS, F. (org.). **Gêneros jornalísticos no Brasil**. São Bernardo do Campo: Universidade Metodista de São Paulo, 2010.

MARQUES DE MELO, J.; ASSIS, F. Gêneros e formatos jornalísticos: um modelo classificatório. **Intercom** - RBCC São Paulo, v. 39, n. 1, p. 39-56, jan./abr., 2016. DOI https://doi.org/10.1590/1809-5844201613

MIOT, H. A. Análise de concordância em estudos clínicos e experimentais. **J. vasc. bras.**, v. 15, n. 2, p. 89-92, 2016. DOI https://doi.org/10.1590/1677-5449.004216

O ESTADO DE S.PAULO. **Manual de redação e estilo**. MARTINS, E. (org.). São Paulo: O Estado de S.Paulo, 1990.

O GLOBO. **Manual de redação e estilo**. GARCIA, L. (org.). São Paulo: Editora Globo, 1998.

RANDOLPH, J. J. 2005. Free-marginal multirater kappa: An alternative to Fleiss' fixed-marginal multirater kappa. Comunicação. **Joensuu University Learning and Instruction Symposium 2005**, Joensuu, Finland, October 14-15th, 2005. (ERIC Document Reproduction Service No. ED490661).

SEIXAS, L. **Redefinindo os gêneros jornalísticos**: proposta de novos critérios de classificação. Covilhã: LabCom, 2009.

SOUZA, R. C. **A revista Time em uma perspectiva multidimensional**. 2012. 330 f. Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem) – LAEL, Pontifícia Universidade Católica de São Paulo, São Paulo, 2012.

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

VAN DIJK, T. A. **News as discourse**. Hillsdale, New Jersey: Erlbaum, 1988.

VAN DIJK, T. A. **Discurso e poder**. São Paulo: Contexto, 2010.

VEIRANO PINTO, M. **A linguagem dos filmes norte-americanos ao longo dos anos**: uma abordagem multidimensional. 2013. 467 f. Tese (Doutorado em Linguística

Aplicada e Estudos da Linguagem) – LAEL, Pontifícia Universidade Católica de São Paulo, São Paulo, 2013.

Apêndice

Tabela 7 – Registros jornalísticos concordes: posicionamento nas dimensões de Kauffmann (2005).

Registro	D1	D2
Reportagem	0,32	0,22
Notícia	-0,80	-2,77
Artigo	-0,39	1,97
Editorial	-2,10	1,14
Coluna de notas	-0,89	0,00
Crítica	-1,30	0,95
Chamada	-1,36	-2,39
Carta	0,63	0,20
Entrevista	3,86	3,74

Fonte: elaborada pelo autor.

Artigo recebido em: 31.10.2021 Artigo aprovado em: 30.05.2022

DOI: 10.14393/DL52-v16n4a2022-4

O fenômeno do desfocamento do agente: uma discussão sobre a importância dos recursos computacionais para os estudos linguísticos

The phenomenon of agent defocusing: a discussion of the relevance of computational resources for linguistic studies

Andressa Rodrigues GOMIDE[*]
Taíse SIMIONI[**]
Aden PEREIRA[***]

RESUMO: Embora em expansão, a pesquisa linguística empírica da língua portuguesa ainda está longe de alcançar todo o seu potencial. Acreditamos que isso possa se dever, em parte, pelo desconhecimento de alguns investigadores de recursos já disponíveis gratuitamente. Neste artigo, apresentamos algumas ferramentas da Linguística de *Corpus* e um *corpus* de escrita acadêmica em português (CoPEP), e como eles podem ser utilizados para explorar o fenômeno do desfocamento do agente em artigos acadêmicos publicados no Brasil e em Portugal. Para isso, utilizamos recursos já existentes para anotar e disponibilizar de forma gratuita e online o CoPEP, um *corpus* de extrema utilidade para investigações linguísticas acerca do português acadêmico.

PALAVRAS-CHAVE: Linguística de *Corpus*. Escrita acadêmica. Ferramentas de investigação de *corpora*. Desfocamento do agente.

ABSTRACT: Although expanding, empirical linguistic research on the Portuguese language is still far from reaching its full potential. We believe that this might be due to some researchers lack of awareness of resources already available for free. In this article, we present some Corpus Linguistics tools and a corpus of academic writing in Portuguese (CoPEP), and how they can be used to explore the phenomenon of agent defocusing in academic articles published in Brazil and Portugal. For this, we use existing resources to annotate and make available, free of charge and online, the CoPEP, an extremely useful corpus for linguistic investigations on academic Portuguese.

KEYWORDS: Corpus Linguistics. Academic writing. Tools for corpus analysis. Agent defocusing.

[*] Doutora. Universidade de Coimbra. ORCID: https://orcid.org/0000-0002-1481-4748. andressa.gomide@fl.uc.pt
[**] Doutora. Universidade Federal do Pampa. ORCID: https://orcid.org/0000-0002-9778-7393. taisesimioni@unipampa.edu.br
[***] Doutora. Universidade Federal do Pampa. ORCID: https://orcid.org/0000-0002-2866-4218. adenpereira@unipampa.edu.br.

1 Introdução

A quantidade de recursos computacionais (dados e ferramentas) disponíveis ou em fase de desenvolvimento no âmbito do processamento computacional da língua portuguesa está em franco crescimento. Tal fenômeno é observado tanto na indústria quanto na academia. Na indústria, possivelmente isso é causado pelo aumento da classe consumidora de tecnologias em Estados membros da Comunidade dos Países de Língua Portuguesa (CPLP). Há um crescente interesse em desenvolvimento de produtos como assistentes de voz (ex.: Alexa, Cortana, Siri) e *chatbots*. Na academia, temos grandes projetos em andamento, como a criação do primeiro Dicionário do Português de Moçambique (DiPoMo)[1] e o projeto de Reconhecimento Automático de Fala e Síntese de Fala no Centro de IA (TaRSila)[2].

Contudo, o progresso no contexto acadêmico não é tão rápido como gostaríamos. Há ainda um número reduzido de estudos linguísticos empíricos que utilizam métodos e dados computacionais. Acreditamos que uma razão para o número reduzido de estudos linguísticos computacionais do português não está na escassez de ferramentas e dados, mas no desconhecimento de tais ferramentas, ou mesmo, na insegurança por parte dos pesquisadores em usá-las.

Com esta suposição em mente, temos como objetivos deste artigo (a) traçar um breve histórico da Linguística de *Corpus* e discutir suas contribuições para análises linguísticas (seção 2), (b) apresentar dois estudos empíricos feitos a respeito do fenômeno do desfocamento do agente no português brasileiro (PB) (seção 3), (c) e demonstrar como podemos expandir as pesquisas apresentadas em (b) com o uso de um *corpus* de escrita acadêmica junto ao uso de uma ferramenta de busca em *corpus* e à aplicação de testes de análise estatística.

---

[1] Mais informações disponíveis no endereço eletrônico: https://www.instituto-camoes.pt/sobre/comunicacao/noticias/mocambique-projeto-do-primeiro-dicionario-de-portugues-de-mocambique-arranca-com-formacao-em-maputo

[2] Mais informações disponíveis no endereço eletrônico: https://sites.google.com/view/tarsila-c4ai

2 Relevância da Linguística de *Corpus* para análises quali-quantitativas

Segundo Berber Sardinha (2004, p. xvii), a Linguística de *Corpus* (LC) é uma área que trata do uso de *corpora* computadorizados que se compõem de textos coletados, transcrições ou escritos da fala, sendo estes mantidos em arquivo de computador. Assim, a LC busca contestar os paradigmas linguísticos em prol de abrir novos caminhos para diversos estudiosos da área da linguagem, tais como linguistas, professores, tradutores e lexicógrafos, a título de exemplo, bem como outros profissionais, beneficiando essas áreas de modo a instrumentalizá-las quanto às pesquisas das línguas que o investigador ou investigadora pode realizar em alinhamento com a LC.

De acordo com Berber Sardinha (2004, p. 4), o primeiro *corpus* linguístico eletrônico, o *Brown University Standard Corpus of Present-day American English*, criado em 1964, contava com 1 milhão de palavras. A partir dele, os textos eram transferidos para o computador por meio de cartões perfurados um a um. O autor ainda acrescenta que o primeiro *corpus* de língua falada era composto por 200 mil palavras, as quais foram coletadas pelo estudioso Sinclair (1995).

Naquele período, a coleta de dados linguísticos era vista com desconfiança na academia ao mesmo tempo em que Chomsky lançava a Teoria do Gerativismo, com maior destaque ao que a mente podia processar para, a seguir, ser convertido em linguagem, ou seja, a ênfase era dada mais à competência do que ao desempenho do falante, segundo Berber Sardinha (2004).

No Brasil, em meados dos anos 2000, a LC estava em fase preliminar, voltada mais para a Lexicografia e a Linguística Computacional, ainda que entre essas áreas não houvesse conformidade recíproca sobre qual seriam as funções da LC, seja para coleta, análise e processamento dos dados, os quais poderiam beneficiar tais áreas como instrumentos de pesquisa, como esclarece Berber Sardinha (2004, p. 6).

Ainda assim, com o crescimento da busca por mais produtos tecnológicos que propiciem a melhoria da comunicação, tanto quanto a de bens e produtos, ademais da esfera acadêmica, foi perceptível o crescimento do interesse no meio corporativo sobre *corpus*. Desse modo, muitas parcerias entre as universidades e as empresas sucederam-se.

Em tempos mais recentes, vimos testemunhando o interesse progressivo no processamento automático de textos, assim como na informatização de bases de dados e na montagem de sistemas inteligentes de reconhecimento de voz e gerenciamento da informação. Isso, especialmente, em se tratando de empresas de telecomunicações e marketing digital que buscam investir nesses elementos característicos da área de computação.

Uma das ferramentas bastante utilizadas para o processamento de *corpora* linguísticos é o programa *WordSmith Tools*, que apresenta diversas versões (gratuitas e pagas). A partir desse recurso, pode-se investigar a frequência, os graus de fixidez e decomponibilidade que ocorrem nas palavras, expressões e sentenças dos mais variados gêneros textuais.

Berber Sardinha (2004, p. 22) destaca a importância da representatividade de um *corpus* em uma língua. Nesse mesmo sentido, Leech (1992, p. 120) traz o foco para as investigações de Biber, de modo a apontar a função representativa de um *corpus*, já que aquela tem o papel de caracterizá-lo, haja vista que "No design de corpus, a representatividade da linguagem é alcançada por progressão cíclica, empiricamente a partir do teste de adequação de corpora previamente elaborados" (p. 120)[3].

Assim, Leech (1999) e Sinclair (1995) entendem que, para se ter um *corpus* representativo, é fundamental o conhecimento acerca da população de onde ele é derivado, porque isso estaria correlacionado à possibilidade de se estabelecer

[3] Texto original: "[...] in corpus design, representativeness of the language is achieved by cyclic progression, based on empirically testing the adequacy of previously designed corpora".

associações entre traços que parecem ser mais ou menos comuns em certos cenários. Em outros termos, para Berber Sardinha (2004, p. 22), conhecer a probabilidade em que ocorrem os traços lexicais, estruturais, pragmáticos e discursivos nos mais diversos contextos está no âmago da LC.

Além da representatividade apontada por Berber Sardinha, segundo Rocha (2007, p. 20), há outras características que um *corpus* precisa apresentar a fim de ser constituído como tal: amostra, tamanho finito e formato legível em um computador e referência padrão. Isso porque, para esse autor, "[...] a representatividade determina quais generalizações em relação às características de uma determinada população são confiáveis, muitas vezes expressa em termos de populações às quais as generalizações se aplicam" (p. 21) [4].

Rocha (2007) destaca que Biber *et al.* (1998) debatem a noção de *corpus* equilibrado, categorizando aquilo que deve ser incorporado ao próprio *corpus* porque para o autor "[...] as técnicas típicas de amostragem utilizadas em estudos estatísticos só são úteis à linguística de forma limitada" [5], já que, por exemplo, "[...] uma amostra proporcional de uma língua, tal como registrada através de um grupo de usuários da língua em suas atividades diárias, resultaria em um *corpus* homogêneo" [6].

Por outro lado, excluir tais textos poderia comprometer a representatividade, já que eles são gêneros utilizados com frequência na sociedade (ROCHA, 2007). Assim, no presente artigo, buscou-se trabalhar com um *corpus* representativo do gênero acadêmico para que se pudesse mensurar com maior precisão a ocorrência do fenômeno do desfocamento do agente.

---

[4] Texto original: "[…] representativeness determines which generalisations regarding features of a given population are trustworthy, often expressed in terms of populations to which generalisations apply".
[5] Texto original: "[…] typical sampling techniques used in statistical studies are only useful to linguistics to a limited extent".
[6] Texto original: "[…] a proportional sample of a language, as registered through a group of language users in their daily activities, would result in a rather homogeneous corpus"

Atualmente, os estudos de Biber destacados por Berber Sardinha e Pinto (2014), a partir do 25º aniversário da publicação do livro seminal daquele autor, *Variation Across Speech and Writing*, em 1988, apontam para o estudo de Análise Multifuncional (AM), já que diversos pesquisadores têm como proposta ampliar o escopo da AM analisando um espectro de registros, períodos de tempo e contextos de uso dos diversos *corpora* existentes desde aquele período histórico do surgimento da LC.

A flexibilidade da abordagem da AM torna possível destacar a eficácia de sondar tanto contextos especializados quanto os mais gerais com eficiência expressiva, além de aferir o que já existe há anos ou décadas de uso da linguagem. Neste sentido, de acordo com Halliday (1993/2005 *apud* BERBER SARDINHA; PINTO, 2014), já é possível detectar uma gama de probabilidades de uso de uma língua ou de suas variedades em diversos contextos, levantando hipóteses contundentes resultantes das mais diversas experiências humanas e de suas interações comunicacionais a partir dos experimentos da AM.

Para Berber Sardinha e Pinto (2014), a AM é um método muito potente que permite ao pesquisador empreender uma análise da língua em uso, a partir da qual é possível fazer descrições que consigam capturar como os usuários da língua fazem suas escolhas linguísticas em contextos específicos, já que a AM se fundamente na vida real humana. No caso do presente artigo, é possível verificar as escolhas dos acadêmicos ao se referirem aos seus objetos de estudo e verificar como fazem essas escolhas a partir dos contextos que o *corpus* nos apresenta.

Assim, em trabalhos futuros, será possível investigar, na sequência da discussão feita no presente artigo, conforme apontam Berber Sardinha e Pinto (2014),

> relações intensas com a comunicação humana, observação atenta do contexto e das pessoas que vivem nas situações em que a linguagem é usada e o movimento analítico rápido de um registro para outro em

> um esforço para perceber as qualidades que os unem ou os separam (p. xv).[7]

Isso sempre levando-se em consideração que o olhar investigativo de quem analisa o objeto de estudo, por mais objetivo que seja o método, sempre parte de sua perspectiva vivencial também como usuário da língua, no caso, o português do Brasil aqui em questão.

Após essa breve discussão sobre pressupostos importantes para a LC, a próxima seção apresenta os resultados de dois estudos sobre o fenômeno do desfocamento do agente, o que nos permitirá, na seção 4, apontar para caminhos possíveis de expansão de tais pesquisas a partir da exploração de recursos computacionais.

3 O fenômeno do desfocamento do agente

Morais (2016) analisa o uso do clítico *se* em artigos acadêmicos a fim de verificar a sua atuação como uma estratégia de impessoalização ou de desfocamento do agente. O *corpus* de sua pesquisa faz parte do projeto SAL (*Systemics Across Languages*) e foi constituído por 1225 artigos produzidos em língua portuguesa e coletados na plataforma *Scielo*. A autora destaca o uso de ferramentas computacionais em seu trabalho, especificamente o programa *WordSmith Tools*, o que lhe permitiu a análise de uma grande quantidade de textos produzidos em situações reais de interação.

Como etapa metodológica, Morais (2016) organiza os diferentes usos do clítico *se* em grupos. Para tal, a autora se utiliza de testes de refraseamento. Como o fenômeno que nos interessa no presente trabalho é o do desfocamento do agente, nossa atenção se voltará para o que a autora classificou como grupo 2. Trata-se de usos do clítico *se*

[7] Texto original: "[...] intense dealings with human communication, thoughtful consideration of the context and the people living in those situations where the language is used, and swift analytical movement from one register to the next in an effort to perceive the qualities that bond them together or tease them apart"

que equivalem a construções com passiva analítica ou com primeira pessoa do plural. Os exemplos em (1) mostram tais equivalências[8].

(1)
(a) Observa-se que houve diferença significativa…
(b) Foi observado que houve diferença significativa…
(c) Observamos que houve diferença significativa…

Como esclarece Morais (2016), neste grupo, há uma predominância de processos materiais, na terminologia usada pela linguística sistêmico-funcional, perspectiva teórica adotada pela autora. As equivalências em (1) evidenciam a existência de um agente, de forma que, na construção (1a), "o clítico é um mecanismo importante para *apagar* o autor no texto, deixando, porém, um resquício de sua participação" (MORAIS, 2016, p. 79).

Este grupo é incluído por Morais (2016) na categoria em que o clítico *se* encontra-se em construções com desfocamento de participante. Morais (2013) propõe três graus de desfocamento de participantes: no alto grau, qualquer participante poderia estar envolvido; no médio grau, dois participantes estariam envolvidos (o autor do artigo e a comunidade acadêmica); no baixo grau, haveria um único participante envolvido (o autor do artigo ou o pesquisador que estiver sendo mencionado).

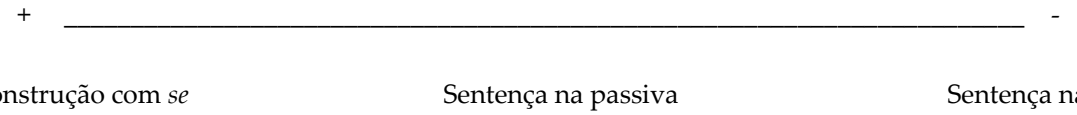
O presente estudo tem como foco o que Morais (2013) classifica como baixo grau de desfocamento do participante, construções nas quais é comum a presença de circunstâncias de lugar e a ocorrência de verbos no pretérito, como exemplifica a sentença em (2).

(2)
Nessa pesquisa verificou-se valor de produção de matéria seca…

[8] Os exemplos em (1) e (2) foram retirados de Morais (2016).

Como explica Morais (2016, p. 91), "o artigo de pesquisa é um texto em que se relata sobre uma pesquisa feita, por isso, os processos ligados ao processo de *fazer* pesquisa (*observar*, *verificar* e *analisar*) permitem pressupor um Agente, o pesquisador, responsável pelas etapas/ações do trabalho". Partindo desse pressuposto, propomos que as sentenças em (1) podem revelar um contínuo de desfocamento do agente, como mostra Fig. 1.

Figura 1 – Contínuo de desfocamento do agente.



Fonte: elaborado pelas autoras.

No polo à esquerda, estaria localizada a construção com *se*, que permite o grau máximo de desfocamento, como aquela em (1a). No polo à direita, estariam as construções ativas, em que não ocorre o desfocamento do agente, como aquela em (1c). Entre os polos, se encontrariam as construções na voz passiva, que permitiriam um grau intermediário de desfocamento, como aquela em (1b).

Tal contínuo baseia-se na análise de Shibatani (1985), segundo a qual a função primária das construções passivas é o desfocamento do agente (*agent defocusing*). Segundo o autor, tal função se evidencia pelo fato de que as construções passivas geralmente não expressam seus agentes de forma explícita, mesmo nas línguas que permitem tal estrutura. Camacho (2002), que analisou um *corpus* constituído por 916 ocorrências de estruturas sentenciais pertencentes ao NURC (Projeto da Norma Urbana Culta), mostra que este é o caso do português brasileiro. Segundo seus resultados, em 85,5% das ocorrências de passivas, não existe "a possibilidade de

recuperação, no contexto discursivo, de referência a uma entidade individuada que seja controladora da ação desenvolvida no predicado" (CAMACHO, 2002, p. 258).

Shibatani (1985), ao defender a sua proposta de análise das passivas como estruturas que têm por finalidade o desfocamento do agente, propõe a seguinte hierarquia para o foco no que diz respeito à estrutura de uma sentença: sujeito > objeto direto > objeto indireto > objetos oblíquos. Em línguas como o português, que permitem a expressão do agente em construções passivas, o agente é expresso por um objeto oblíquo, ou seja, ele assume "o grau mais baixo de foco entre os elementos sintaticamente codificados"[9] (SHIBATANI, 1985, p. 833).

No contínuo de desfocamento do agente proposto neste trabalho, situamos as construções com *se* no grau máximo de desfocamento, distinguindo-as das passivas, em função de que, como mostra Camacho (2002, p. 251), as primeiras "não autorizam a manifestação formal de um SN agentivo", como mostram os exemplos em (3)[10].

(3)
(a) João quebrou o vidro da janela.
(b) O vidro da janela foi quebrado (por João).
(c) O vidro da janela (se) quebrou (?por João).

Camacho (2002) explica que a diferença mostrada entre (3b) e (3c), que reside na possibilidade ou não de expressão de um agente, é tão relevante que isso se relaciona à possibilidade de haver uma expressão de instrumento nas passivas (4a), o que pressupõe a existência de uma entidade agentiva, enquanto as construções com *se* não a permitiriam (4b).

(4)
(a) O vidro da janela foi quebrado com uma pedrada.

[9] Texto original: "[...] the lowest degree of focus among the syntactically encoded elements".
[10] Os exemplos em (3) e (4) foram retirados de Camacho (2002).

(b) O vidro da janela (se) quebrou (?com uma pedrada).

Entre as características comuns à construção passiva e àquela com *se*, Camacho (2002) cita o fato de que ambas acarretam um argumento agentivo, ainda que na segunda tal argumento fique subentendido, uma vez que inexiste a possibilidade de sua manifestação formal. Ainda conforme o autor, "mesmo que, na passiva, o agente nem sempre se manifeste, enunciá-lo depende unicamente do ponto de vista do falante em relação ao evento e não de uma restrição sintático-semântica" (CAMACHO, 2002, p. 304), diferentemente do que ocorre nas construções com *se*, nas quais o agente não pode ser expresso.

É com base nestes fatores que propomos a diferença entre as construções com *se* e as passivas no que diz respeito ao seu lugar no contínuo de desfocamento do agente.

Destacamos que o presente trabalho é derivado de uma reflexão sobre a língua que foi possível a partir da análise dos resultados de uma pesquisa que se utilizou de ferramentas computacionais, o que permitiu o acesso a uma grande quantidade de textos (MORAIS, 2016), combinada com a análise de resultados de uma pesquisa que, apesar de não se valer de ferramentas computacionais, procedeu a um escrutínio criterioso dos dados (CAMACHO, 2002). Tal percurso teórico-metodológico nos permitiu apontar possibilidades de expansão do estudo do desfocamento do agente a partir da exploração de recursos computacionais, conforme mostra a seção 4 deste trabalho.

4 O desfocamento do agente em um *corpus* de escrita acadêmica

A seção anterior apresentou estudos que discutem as diferentes formas de desfocamento do agente e suas implicações. Embora os estudos apontem interessantes resultados, a pesquisa empírica com um alto volume de dados somada ao uso de

ferramentas computacionais adequadas pode enriquecer os resultados. É tal ideia que será exposta e defendida nas próximas subseções.

4.1 O *corpus*

Para este estudo, utilizamos o *Corpus* de Português Escrito em Periódicos (CoPEP) (KUHN; FERREIRA, 2020). O CoPEP é um *corpus* representativo da escrita acadêmica em língua portuguesa nas variedades brasileira e europeia (PE), composto de 9.900 textos e um total de 48.506.519 palavras. O *corpus* reúne textos publicados entre os anos 1992 e 2018 em revistas acadêmicas periódicas de seis áreas de conhecimento, agrupadas em três grandes colégios (Tabelas 1 e 2).

Além do texto *per se*, podemos também obter informações adicionais para cada unidade mínima, *token*, de um *corpus*. *Tokens* são, em sua maioria, palavras. Dígitos, siglas, acrônimos, pontuação são também considerados *tokens*. Essas informações adicionais, também conhecidas como *anotação*, são normalmente feitas de forma automática. Atualmente uma versão do CoPEP com anotações para classes gramaticais (POS - do inglês *parts-of-speech*) e lema (ou *tokens* em sua forma neutra) está disponível via SketchEngine (KILGARRIFF *et al*., 2014). O SketchEngine é uma poderosa plataforma *online* de busca em *corpus*, disponibilizada por meio de assinatura paga.

Para o presente trabalho, obtivemos dos autores o *corpus* bruto, i.e., sem anotações linguísticas, e permissão para distribuí-lo gratuitamente. Do *corpus* em seu estado bruto, preparamos e anotamos um *corpus* com dois sistemas diversos para permitir que estudos futuros trabalhem com eles e testem diferentes tipos de anotações e suas respectivas precisões. O primeiro sistema possui anotações para classes gramaticais e lema, implementadas a partir do TreeTagger (SCHMID, 1994). O segundo sistema foi implementado utilizando o Spacy[11], uma biblioteca para

[11] Disponível em: https://spacy.io

processamento de linguagem natural em Python. Neste segundo sistema, temos as seguintes anotações: lema, classe gramatical, classe gramatical simplificada, núcleo do sintagma e etiqueta sintática.

Além das informações atribuídas a cada *token* do *corpus*, é útil também obter informações a respeito de cada texto que compõe o *corpus*. No CoPEP, as informações preservadas em cada texto são as seguintes: variedade do português (europeu ou brasileiro); país sede da revista acadêmica; área de conhecimento; grande área de conhecimento; ISSN; ano de publicação; título do artigo. Tais informações são comumente conhecidas como *metadados*. Para melhor explorar o potencial das anotações e metadados, precisamos instalar ou carregar o *corpus* em uma ferramenta de busca. Para o presente trabalho, instalamos o CoPEP em um aplicativo *online* de busca em *corpus*, o CQPweb (HARDIE, 2012)[12].

Tabela 1 – Distribuição de *tokens* e textos no CoPEP (área de conhecimento).

	Europa		Brasil		CoPEP Total	
	tokens	textos	tokens	textos	tokens	textos
Ciências Humanas	12747013	2581	12751623	1219	25498636	3800
Ciências Sociais Aplicadas	2686764	517	2689639	319	5376403	836
Colégio de Humanidades	15433777	3098	15441262	1538	30875039	4636
Ciências da Saúde	6687507	2432	6695452	1564	13382959	3996
Ciências Agrícolas	1283054	385	1301763	522	2584817	907
Colégio de Ciências da Vida	7970561	2817	7997215	2086	15967776	4903
Ciências Exatas e da Terra	402028	67	401918	118	803946	185
Engenharia	422790	107	436968	69	859758	176
Colégio de Ciências Exatas, da Terra e Multidisciplinar	824818	174	838886	187	1663704	361

Fonte: elaborada pelas autoras.

Tabela 2 – Distribuição de *tokens* e textos no CoPEP (ano de publicação).

	Todos		europeu		brasileiro	
	tokens	textos	tokens	textos	tokens	textos
1992	12496	1	0	0	12496	1
1993	29810	3	0	0	29810	3
1994	29198	3	0	0	29198	3
1996	21240	2	0	0	21240	2

[12] Disponível em: https://ola.unito.it/CQPweb32/copep

1997	89312	15	9343	9	79969	6
1998	367897	91	5048	6	362849	85
1999	270118	63	26332	17	243786	46
2000	654587	97	241728	31	412859	66
2001	797574	173	232095	37	565479	136
2002	1267869	278	203339	37	1064530	241
2003	1917247	407	63109	22	1854138	385
2004	1673253	368	65742	38	1607511	330
2005	1731566	330	121155	87	1610411	243
2006	2252933	443	300393	122	1952540	321
2007	1684884	330	371905	168	1312979	162
2008	1973120	363	664464	209	1308656	154
2009	3660192	684	2334265	530	1325927	154
2010	4246448	871	2508551	604	1737897	267
2011	4326782	910	2541318	647	1785464	263
2012	5534955	1146	3811356	903	1723599	243
2013	5710364	1191	3711249	910	1999115	281
2014	6037394	1298	3972501	1023	2064893	275
2015	3899954	769	2943845	664	956109	105
2016	281453	54	65545	15	215908	39
2017	28534	8	28534	8	0	0
2018	7339	2	7339	2	0	0
Total	48506519	9900	24229156	6089	24277363	3811

Fonte: elaborada pelas autoras.

4.2 Ferramenta de busca

Escolhemos o CQPweb por ser uma ferramenta *online* e de código aberto. Ser uma ferramenta *online* auxilia o estudo colaborativo e elimina a necessidade de instalação de programas. Possuir o código aberto significa, neste caso, ter uma ferramenta gratuita para o consumidor final e ainda ter a possibilidade de adaptar ou criar novas funções no programa, caso haja necessidade.

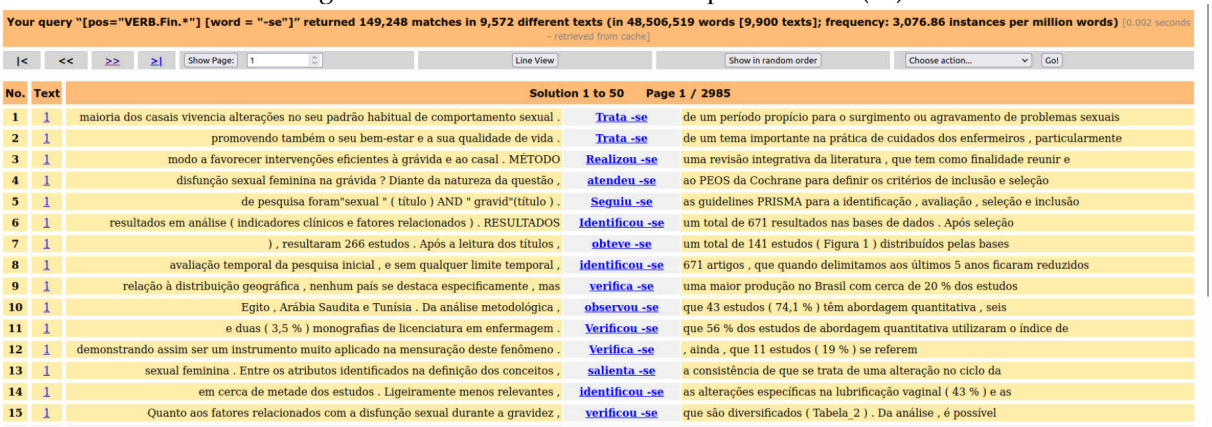
Com o CQPweb, é possível realizar buscas simples por palavras ou sequência de palavras e obter linhas de concordância, bem como a frequência relativa e absoluta dos elementos da busca e o número de textos que possuem o(s) item(ns) buscado(s). As buscas também podem ser restringidas (filtradas) de acordo com os metadados dos textos que compõem o *corpus*. Ou seja, é possível optar por fazer buscas apenas em textos com determinadas características.

Uma importante ferramenta do CQPweb é o poderoso sistema de busca *Corpus Query Processor* (CQP). Esse sistema permite que realizemos buscas refinadas por estruturas elaboradas. Por exemplo, para encontrarmos exemplos de desfocamento do agente, como descrição feita na seção três, podemos utilizar as formas de buscas em (5).

(5)
(a) [pos="VERB.Fin.*"] [word = "-se"]
(b) [(word = "foi") | (word = "foram")]{1,3} [(pos = "VERB.Part.*")]
(c) [(pos != "PRON.*") & (pos != "NOUN.*") & (word != "que")]{1,3} [(pos = "VERB.Fin.Plur") & (word = ".*mos")]

A busca (5a) procura por todas as ocorrências de um *token* com a etiqueta de verbo na forma finita seguido pelo *token* "-se" (Fig. 2). A estrutura em (5b) retorna ocorrências em que o *token* "foi" ou "foram" ocorre à esquerda de um verbo em sua forma participial (fig. 3). A notação {1,3} significa que o segundo *token* pode vir imediatamente após o primeiro (ex.: Foram analisados) ou até três casas à direita (ex.: Foi devidamente analisado).

Figura 2 – Linhas de concordância para a busca (5a).



Your query "[pos="VERB.Fin.*"] [word = "-se"]" returned 149,248 matches in 9,572 different texts (in 48,506,519 words [9,900 texts]; frequency: 3,076.86 instances per million words) [0.002 seconds - retrieved from cache]

|< << >> >| Show Page: 1 Line View Show in random order Choose action... Go!

No.	Text	Solution 1 to 50 Page 1 / 2985		
1	1	maioria dos casais vivencia alterações no seu padrão habitual de comportamento sexual .	Trata -se	de um período propício para o surgimento ou agravamento de problemas sexuais
2	1	promovendo também o seu bem-estar e a sua qualidade de vida .	Trata -se	de um tema importante na prática de cuidados dos enfermeiros , particularmente
3	1	modo a favorecer intervenções eficientes à grávida e ao casal . MÉTODO	Realizou -se	uma revisão integrativa da literatura , que tem como finalidade reunir e
4	1	disfunção sexual feminina na grávida ? Diante da natureza da questão ,	atendeu -se	ao PEOS da Cochrane para definir os critérios de inclusão e seleção
5	1	de pesquisa foram"sexual " (título) AND " gravid"(título) .	Seguiu -se	as guidelines PRISMA para a identificação , avaliação , seleção e inclusão
6	1	resultados em análise (indicadores clínicos e fatores relacionados) . RESULTADOS	Identificou -se	um total de 671 resultados nas bases de dados . Após seleção
7	1	) , resultaram 266 estudos . Após a leitura dos títulos ,	obteve -se	um total de 141 estudos (Figura 1) distribuídos pelas bases
8	1	avaliação temporal da pesquisa inicial , e sem qualquer limite temporal ,	identificou -se	671 artigos , que quando delimitamos aos últimos 5 anos ficaram reduzidos
9	1	relação à distribuição geográfica , nenhum país se destaca especificamente , mas	verifica -se	uma maior produção no Brasil com cerca de 20 % dos estudos
10	1	Egito , Arábia Saudita e Tunísia . Da análise metodológica ,	observou -se	que 43 estudos (74,1 %) têm abordagem quantitativa , seis
11	1	e duas (3,5 %) monografias de licenciatura em enfermagem .	Verificou -se	que 56 % dos estudos de abordagem quantitativa utilizaram o índice de
12	1	demonstrando assim ser um instrumento muito aplicado na mensuração deste fenômeno .	Verifica -se	, ainda , que 11 estudos (19 %) se referem
13	1	sexual feminina . Entre os atributos identificados na definição dos conceitos ,	salienta -se	a consistência de que se trata de uma alteração no ciclo da
14	1	em cerca de metade dos estudos . Ligeiramente menos relevantes ,	identificou -se	as alterações específicas na lubrificação vaginal (43 %) e as
15	1	Quanto aos fatores relacionados com a disfunção sexual durante a gravidez ,	verificou -se	que são diversificados (Tabela_2) . Da análise , é possível

Fonte: extraída do programa CQPweb.

Figura 3 – Linhas de concordância para a busca (5b).

Your query "[(word = "foi") | (word = "foram")]{1,3} [(pos = "VERB.Part.*")]" returned 95,129 matches in 8,799 different texts (in 48,506,519 words [9,900 texts]; frequency: 1,961.16 instances per million words) [2.523 seconds]

|< << >> >| Show Page: 1 Line View Show in random order New query Go!

No.	Text	Solution 1 to 50 Page 1 / 1903		
1	1	nos idiomas português , francês , espanhol e inglês . A pesquisa	foi realizada	nas bases de dados da EBSCHOhost (CINAHL , MEDLINE , Nursing
2	1	estudos na revisão . Por se tratar de uma revisão integrativa ,	foi realizada	pesquisa adicional nas referências dos artigos previamente selecionados e todos os resulta
3	1	pesquisa adicional nas referências dos artigos previamente selecionados e todos os resultados	foram integrados	, de modo a ter uma revisão mais ampla . A inclusão
4	1	modo a ter uma revisão mais ampla . A inclusão dos estudos	foi realizada	atendendo à concordância de dois investigadores de modo independente . Os dados
5	1	concordância de dois investigadores de modo independente . Os dados dos estudos	foram analisados	após preenchimento de um instrumento de colheita de dados com dados acerca
6	1	se encontravam em duplicado . Resultaram 68 estudos , dos quais 7	foram retirados	porque não cumpriam o critério de inclusão referente ao idioma e 3
7	1	47 estudos com texto completo (81 %) e os restantes	foram analisados	apenas a partir do resumo , uma vez que faziam parte de
8	1	uma vez que cerca de 40 % da produção sobre este tema	foi publicada	nos últimos 5 anos . Quanto à distribuição dos estudos neste período
9	1	anos . Quanto à distribuição dos estudos neste período , 24,1 %	foram publicados	em 2012 , seguindo -se 2014 com 22,4 % , 2011 e
10	1	com 5,1 % e Espanha e Israel com 3,5 % . Ainda	foram identificados	estudos oriundos de países como Canadá , Polônia , Austrália , Croácia
11	1	e na autoimagem (30 %) . Um pouco menos prevalentes	foram identificados	fatores religiosos , alterações na autoestima , presença de algumas vulnerabilidades (
12	2	aceroleira adubadas com fósforo e zinco . MATERIAL E MÉTODOS O experimento	foi instalado	em casa de vegetação situada no pomar do Departamento de Agricultura da
13	2	localizado na região . Após a retirada dos frutos , as sementes	foram lavadas	e semeadas em canteiros com areia lavada e , posteriormente , repicadas
14	2	cálcio e de magnésio p.a. , na relação 4:1 . A calagem	foi efetuada	20 dias antes da repicagem das mudas . O delineamento utilizado foi
15	2	misturado ao solo , até completa homogeneização . O sulfato de zinco	foi dissolvido	na água e aplicado via solução . Os demais nutrientes a as

Fonte: extraída do programa CQPweb.

O último caso em (5c) (Fig. 4) procura por ocorrências de verbos na forma finita na primeira pessoa do plural que não sejam precedidos por um pronome, por um substantivo ou pelo token "que".

Figura 4 – Linhas de concordância para a busca (5c).

Your query "[(pos != "PRON.*") & (pos != "NOUN.*") & (word != "que")]{1,3} [(pos = "VERB.Fin.Plur") & (word = ".*mos")]" returned 49,810 matches in 5,859 different texts (in 48,506,519 words [9,900 texts]; frequency: 1,026.87 instances per million words) [44.626 seconds]

|< << >> >| Show Page: 1 Line View Show in random order Choose action... Go!

No.	Text	Solution 1 to 50 Page 1 / 997		
1	1	e sem qualquer limite temporal , identificou -se 671 artigos , que	quando delimitamos	aos últimos 5 anos ficaram reduzidos a 266 artigos . Este fato
2	1	(3,5 %) artigos de opinião e dois (3,5 %	) editoriais . Destacamos	que 45 (77,5 %) são estudos originais , um (
3	1	) . Da análise comparativa com dos indicadores clínicos já existentes na	NANDA- I , verificamos	que cinco características definidoras já estão classificadas (alteração na atividade
4	3	como esta vinha se determinando até o presente momento . Por isso	, apresentaremos	, primeiramente , a constituição da liberdade de imprensa e o processo
5	3	da liberdade de imprensa e o processo de regulamentação em nível ocidental	. Depois , descreveremos	a constituição histórica da liberdade de imprensa no Brasil e o contraste
6	3	determinação arbitrária do editor do veículo.11 1.3 Liberdade de imprensa no Brasil	No Brasil , sabemos	que a Imprensa Nacional , órgão criado pelo Decreto de 13.05.1808 (
7	3	legislação específica , ou seja , um laissez-faire na atividade da imprensa	. Vejamos	como se articulam estas posições . 1.4 Lei de Imprensa ou ausência
8	3	no entanto , divergem quando se trata de encontrar uma solução	. Temos	duas posições : a) Necessidade de uma Lei de Imprensa :
9	3	da Filosofia do Direito . Porém , o que , de fato	, sabemos	é que o seu Prefácio expõe local e data , a saber
10	3	ou " caricatura insolente " do governo e de seus ministros .	Entretanto , não queremos	agora julgá -lo pelos padrões de hoje , e sim comparar suas
11	3	em algumas ocasiões , era até pretexto para perseguições e penalidades .	Além disso , sabemos	que Hegel vivenciou a experiência de censura , em 1808 , quando
12	3	imprensa e opinião pública a) Redator-chefe da Gazeta de Bamberg Hegel	, como já afirmamos	, trabalhou como diretor da Gazeta de Bamberg , de 1807 a
13	3	protagonistas , mostrar o que sabem fazer e expressar a sua opinião	. Temos	o MySpace , o YouTube , os blogs , lista de e-mails
14	4	de Pediatria da zona Norte do país e especialistas dedicados à Medicina	da Criança . Temos	ainda a colaboração de editores associados pertencentes a diversas áreas : Pediatria
15	4	pouco reconhecida , já que é habitualmente anônima . Neste último número	de 2012 agradecemos	, de modo particular , a todos os nossos revisores que contribuíram

Fonte: extraída do programa CQPweb.

Outras funcionalidades além das buscas por linhas de concordâncias são igualmente fáceis de se obter. Por exemplo, podemos identificar (sequências de) palavras que frequentemente coocorrem em um contexto, também conhecidos como

colocados. No CQPweb, podemos utilizar diferentes medidas estatísticas (ex.: MI, MI3, Z-score, T-score, log-likelihood, Coeficiente de Sorensen-Dice) para calcular e gerar uma lista de colocados. Por exemplo, a Fig. 5 mostra os 15 colocados mais fortes para a busca *(foram|foi)*, utilizando a medida estatística Log ratio (filtrada) e considerando apenas *tokens* ocorrendo à direita do nódulo e com a etiqueta gramatical de verbo no particípio. É importante notar que o etiquetador não tem 100% de precisão e algumas ocorrências (itens 2, 3 e 9, por exemplo) não estão no particípio.

Figura 5 – Lista de colocados para a busca (foram|foi).

Collocation based on:	Word form	Statistic:	Log Ratio (filtered)
Collocation window *from*:	1 to the Right	Collocation window *to*:	1 to the Right
Freq(node, collocate) at least:	5	Freq(collocate) at least:	5
Filter results by:	specific collocate:	and/or tag: VERB.Part	VERB.Part — Submit changed parameters — Go!

Extra information:

The **Log Ratio** statistic is a measurement of *how big the difference is* between the (relative) frequency of the collocate alongside the node, and its (relative) frequency in the rest of the corpus or subcorpus.

In **the current collocation analysis**, all collocates displayed have Log-likelihood of at least **14.57104**.

The use of a log-likelihood filter means that it is not necessary to set high minimum values for *Freq(node, collocate)* and *Freq(collocate)* when using Log Ratio.

There are 77,321 different words in the collocation database for this query (Query "(foram|foi)" returned 171,230 matches in 9,361 different texts) [5.435 seconds - retrieved from cache]

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log Ratio (filtered)
1	inaugurado	144	0.508	32	29	6.334
2	institucionalizando	25	0.088	5	5	6.141
3	evoluindo	159	0.561	29	28	5.977
4	decidido	334	1.179	56	50	5.829
5	verificado	1,605	5.666	263	218	5.79
6	desdobrado	32	0.113	5	5	5.708
7	admitido	215	0.759	33	31	5.678
8	excluído	300	1.059	46	43	5.676
9	surgindo	457	1.613	68	66	5.625
10	mantido	766	2.704	110	105	5.565
11	instalando	49	0.173	7	7	5.556
12	generalizando	36	0.127	5	4	5.509
13	revertido	52	0.184	7	7	5.456
14	respondido	204	0.72	25	24	5.301
15	forjando	41	0.145	5	5	5.293

Fonte: extraída do programa CQPweb.

Uma outra funcionalidade extremamente útil é a Distribuição. Com ela podemos ver, de forma clara, como os termos buscados se encontram distribuídos no *corpus* de acordo com os metadados. Tomamos como exemplo a busca descrita em (6).

(6)
[(word = "foi") | (word = "foram")] [(word = "observad.*") | (word = "verificad.*") | (word = "analisad.*")]

A primeira informação que obtemos é que, em todo o CoPEP, temos 3997 ocorrências do verbo "foi" ou "foram" seguido dos particípios passados de "observar", "verificar" e "analisar" em 2204 textos[13]. Com a função Distribuição, essa informação é calculada também para cada subseção do *corpus* (Tabelas 3, 4 e 5).

Tabela 3 – Distribuição de acordo com a variedade.

variedade	*tokens* na categoria	ocorrências na categoria	em n textos	de um total de n textos	Frequência por milhão de tokens na categoria
Br	24277363	2340	1212	3811	96,39
Eu	24229156	1657	992	6089	69,39

Fonte: elaborada pelas autoras.

Tabela 4 – Distribuição de acordo com a área de conhecimento.

área de conhecimento	*tokens* na categoria	ocorrências na categoria	em n textos	de um total de n textos	frequência por milhão de tokens na categoria
Ciências Agrícolas	2584817	946	442	907	365,98
Ciências Sociais Aplicadas	5376403	318	176	836	59,15
Engenharia	859758	85	47	176	98,87
Ciências Exatas e da Terra	803946	153	63	185	190.31
Ciências da Saúde	13382959	1958	1097	3996	146,31
Ciências Humanas	25498636	537	379	3800	21,06

Fonte: elaborada pelas autoras.

Tabela 5 – Distribuição de acordo com o ano de publicação.

anos	*tokens* na categoria	ocorrências na categoria	em n textos	de um total de n textos	frequência por milhão de tokens na categoria

[13] A justificativa para a seleção dos verbos "observar", "verificar" e "analisar" encontra-se na seção 3 deste trabalho.

1992-1997	162940	8	5	24	357,57
1998-2002	3358045	489	232	702	744,23
2003-2007	9259883	881	488	1878	480,97
2008-2012	19741497	1312	752	3974	332,86
2013-2018	15965038	13107	727	3322	521,55

Fonte: elaborada pelas autoras.

Os valores das tabelas acima nos permitem fazer comparações na frequência do uso da construção passiva nas diferentes categorias. Por exemplo, a frequência desse tipo de construção é bem maior nas publicações das Ciências Agrícolas do que nas Ciências Humanas, como mostra a Tabela 4.

4.3 Testes estatísticos

4.3.1 Teste de significância

A frequência por si só pode ser enganosa. Por isso, após obtermos os valores para as frequências, verificamos se a diferença entre duas observações é realmente significativa. Para essa verificação, aplicamos um teste de significância. Um teste de significância nos diz qual a probabilidade de se obter o resultado verificado em nossa amostragem, considerando que as variáveis sejam independentes (hipótese nula).

Assim, tomamos o exemplo em (6), na seção 4.2. A frequência normalizada do uso da passiva com os verbos "observar", "analisar" e "verificar" é maior em textos escritos em português brasileiro (96,39) do que em português europeu (68,39). Para verificar a chance da hipótese de que artigos em PB usem essa construção mais frequentemente que artigos escritos em PE, devemos aplicar um teste de significância. O teste de significância nos diz a probabilidade de termos o resultado que obtivemos com a nossa amostra no CoPEP se tivéssemos a certeza de que a variedade do português não está relacionada ao uso dessa construção.

Se a probabilidade for baixa, rejeitamos a hipótese nula e consideramos a nossa hipótese. Essa probabilidade é expressa como valor-p e varia entre 0 (não há chance de se obter os resultados observados) e 1 (certeza absoluta de se obter os resultados observados). Os valores de corte para o valor-p são arbitrários, mas geralmente considera-se que valores-p abaixo de 0.05 sejam significativos.

Para o caso supracitado, obtivemos um valor-p igual a $5{,}038995 \cdot 10^{-27}$ ao aplicar o teste exato de Fisher. Assim, podemos rejeitar a hipótese nula de que a variedade do português não influencia o uso da construção passiva e considerar nossa hipótese de que há diferença de uso entre as variedades europeia e brasileira. Caso o valor-p fosse acima de 0.05, aceitaríamos a hipótese nula e diríamos que a diferença na frequência observada não é significativa.

Existem diferentes testes de significância, sendo os três testes mencionados na sequência frequentemente utilizados na Linguística de *Corpus*. Chi quadrado é um teste amplamente utilizado em vários campos. Porém, apesar do seu alto uso também na Linguística de *Corpus*, este teste possui falhas. Entre essas falhas podemos ressaltar a baixa precisão e a baixa confiabilidade quando trabalhamos com números muito pequenos. O log-likelihood ou G2 teste é um teste amplamente utilizado na Linguística de Corpus (cf. DUNNING, 1993). O teste exato de Fisher, como o próprio nome diz, calcula a probabilidade exata.

4.3.2 Tamanho do efeito

Os testes de significância nos dizem a probabilidade de duas variáveis ou mais serem relacionadas. Para calcular a força com a qual elas estão associadas, utilizamos medidas de tamanho do efeito, também conhecidas como medidas de associação. Algumas medidas comumente usadas na Linguística de *Corpus* são o Coeficiente Phi e o V de Cramér. Para ambas as medidas, os valores vão de 0 (nenhuma correlação entre as variáveis) e 1 (as variáveis são completamente ligadas). O V de Cramér é

utilizado em situações que envolvem mais de duas variáveis. A interpretação usual para esses valores é a de que para valores próximos a 0,1 temos uma correlação baixa; uma correlação média para valores próximos a 0,3 e uma alta correlação caso este valor seja igual ou superior a 0,5. Para o nosso exemplo em (6), obtivemos um coeficiente Phi 0,0015, o que nos aponta a uma baixa correlação.

Os testes que apresentamos neste trabalho foram realizados no ambiente estatístico R, com o auxílio das bibliotecas *stats* (R CORE TEAM, 2018) e *rcompanion* (MANGIAFICO, 2016). Porém, há vários sites que oferecem serviço gratuito e simples de cálculo de testes estatísticos, como é o caso do Social Science Statistics[14].

5 Perspectivas e considerações finais

O presente trabalho apresentou recursos computacionais que permitem ampliar estudos sobre o fenômeno do desfocamento do agente, como os discutidos na seção três. O primeiro passo foi iniciar a preparação do *corpus* para estudar especificamente o contínuo de desfocamento do agente, como proposta apresentada na Fig. 1, bem como realizar algumas análises preliminares. A continuidade do trabalho tratará de, sem se limitar a, dois pontos principais.

Primeiro, faremos a identificação de cada seção (ex.: introdução, conclusão etc.) dos artigos presentes no *corpus*. Essa marcação nos permitirá restringir as buscas no *corpus* por seções do texto. Isso é particularmente útil no estudo do desfocamento do agente pois poderemos excluir seções relativas à revisão da literatura, por exemplo, concentrando-nos nas seções em que verbos como "observar", "analisar" e "verificar" tenham como agente os autores dos artigos, o que nos auxiliaria a localizar com mais precisão o que Morais (2013) identifica como baixo grau de desfocamento de participantes.

[14] Disponível em: https://www.socscistatistics.com/tests

O segundo ponto é implementar uma interface para realizar os testes de significância e tamanho do efeito diretamente na função distribuição do CQPweb. Um dos objetivos com essa nova funcionalidade é facilitar a validação das comparações feitas ao analisar o *corpus*.

Espera-se que o prosseguimento deste trabalho nos permita observar o contínuo de desfocamento do agente e testar a validade das propostas apresentadas neste trabalho, através da investigação de um *corpus* bem preparado através de uma ferramenta eficiente e confiável.

Esperamos também que os recursos aqui apresentados, seja o próprio *corpus* (CoPEP), seja as ferramentas utilizadas para anotação automática dos textos (Spacy), disponibilização e exploração do *corpus* (CQPweb), e aplicação de testes estatísticos (bibliotecas para R), possam servir como uma modesta exposição dos muitos recursos de que dispomos gratuitamente para pesquisas linguísticas.

Referências Bibliográficas

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

BERBER SARDINHA, T.; PINTO, M. V. **Multi-dimensional analysis, 25 years on a tribute to Douglas Biber**. Amsterdam/São Paulo: John Benjamins Publishing Company/Universidade Católica de São Paulo, 2014. (Studies in Corpus Linguistics, ISSN 1388-0373; v. 60). DOI https://doi.org/10.1075/scl.60

BIBER, D.; CONRAD, S.; REPPEN, R. **Corpus linguistics**: investigating language structure and use. Cambridge: Cambridge University Press, 1998. DOI https://doi.org/10.1017/CBO9780511804489

CAMACHO, R. G. Construções de voz. *In*: ABAURRE, M. B. M.; RODRIGUES, A. C. S. (org.). **Gramática do português falado**. Novos estudos descritivos. Campinas: UNICAMP, 2002. p. 227-316.

DUNNING, T. E. Accurate methods for the statistics of surprise and coincidence. **Computational linguistics**, v. 19, n. 1, p. 61-74, 1993.

HARDIE, A. CQPweb — combining power, flexibility and usability in a corpus analysis tool. **International Journal of Corpus Linguistics**, [S.L.], v. 17, n. 3, p. 380-409, 31 dez. 2012. DOI https://doi.org/10.1075/ijcl.17.3.04har

KILGARRIFF, A.; BAISA, V.; BUŁTA, J.; JAKUBÍčEK, M.; KOVÁř, V.; MICHELFEIT, J.; RYCHLÝ, P.; SUCHOMEL, V. The Sketch Engine: ten years on. **Lexicography**, [S.L.], v. 1, n. 1, p. 7-36, jul. 2014. DOI https://doi.org/10.1007/s40607-014-0009-9

KUHN, T. Z.; FERREIRA, J. P. O Corpus de Português Escrito em Periódicos - CoPEP. **Delta**: Documentação de Estudos em Lingüística Teórica e Aplicada, [S.L.], v. 36, n. 2, p. 1-42, fev. 2020. DOI https://doi.org/10.1590/1678-460x2020360209

LEECH, G. Corpora and theories of linguistic performance. *In*: **Directions in corpus Linguistics**. Proceedings of Nobel Symposium 82. Stockolm, 4-8 Aug 1991. Dan Svartvik (editor). Morvton de Guryter: Berlin, 1992. p.105-122.

LEECH, G. Review of Biber, Conrad, and Reppen. Corpus linguistics: Investigating language structure and use. **International Journal of Corpus Linguistics**, Amsterdã, John Benjamins, 4(1), p. 185-88, jun/1999. DOI https://doi.org/10.1075/ijcl.4.1.11lee

MANGIAFICO, S.S. **Summary and Analysis of Extension Program Evaluation in R**, version 1.18.8, 2016. Disponível em: https://rcompanion.org/handbook. Acesso em: 1 out. 2021.

MORAIS, F. B. C. de. **Entre alhos e bugalhos - os usos do clítico *se* na escrita acadêmica**. 2013. 183 f. Tese (Doutorado em Linguística Aplicada e Estudos da Linguagem) – Pontifícia Universidade Católica de São Paulo, São Paulo, 2013.

MORAIS, F. B. C. de. Variação de usos do clítico *se* na comunidade acadêmica: um estudo descritivo com base na linguística sistêmico-funcional. **Cadernos de Linguagem e Sociedade**, v. 17, n. 1, p. 70-100, 2016. Disponível em: https://periodicos.unb.br/index.php/les/article/view/4429. Acesso em: 22 jul. 2021. DOI https://doi.org/10.26512/les.v17i1.4429

R CORE TEAM. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2018. Disponível em: https://www.R-project.org. Acesso em: 1 out. 2021.

ROCHA, M. A. E. Introduction to the issue on corpus linguistics. **Ilha do Desterro**, n. 52, Florianópolis, p. 9-33, jan./jun. 2007.

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

SHIBATANI, M. Passives and related constructions: a prototype analysis. **Language**, v. 61, n. 4, p. 821-848, 1985. Disponível em: https://www.researchgate.net/publication/244437975_Passives_and_Related_Constructions_A_Prototype_Analysis. Acesso em: 9 mar. 2021. DOI https://doi.org/10.2307/414491

SINCLAIR, J. McH. From theory to practice. *In*: LEECH, G.; MYERS, G., THOMAS J. **Spoken English on computer**: transcription, mark-up and application. Londres: Longman, 1995. p 99-112.

Artigo recebido em: 15.12.2021 Artigo aprovado em: 21.02.2022

DOI: 10.14393/DL52-v16n4a2022-5

As contribuições da Linguística de *Corpus* e do Processamento de Linguagem Natural na elaboração do protótipo do Dicionário Ideológico de Locuções

The contributions of Corpus Linguistics and Natural Language Processing in the elaboration of the Ideological Dictionary of Locutions prototype

Thyago José DA CRUZ[*]

RESUMO: Neste trabalho, buscamos demonstrar como recursos e ferramentas da Linguística de *Corpus* e do Processamento da Linguagem Natural puderam ser empregados na elaboração do protótipo do Dicionário Ideológico de Locuções, de caráter monolíngue e, ao mesmo tempo, onomasiológico e semasiológico. Esse tipo de repertório fraseográfico compõe-se de três grandes seções no corpo do dicionário: a parte sinóptico-analógica, a analógica (correspondendo ambas à parte onomasiológica da obra) e a alfabética (de característica semasiológica). No desenvolver desse projeto, utilizamos como *corpora* o *Corpus* Brasileiro e a *Web*. Como ferramenta para a elaboração do corpo do dicionário, empregamos o software *FieldWorks Language Explore*, o FLEx. Ao final, foi possível verificar que esses instrumentos computacionais foram de fundamental relevância para a realização do propósito da pesquisa.

PALAVRAS-CHAVE: Fraseografia. Linguística Computacional. Linguística de *Corpus*. Dicionário Ideológico.

ABSTRACT: In this work, we seek to demonstrate how resources and tools of *Corpus* Linguistics and Natural Language Processing could be used in the elaboration of the Ideological Dictionary of Locutions prototype, monolingual and, at the same time, onomasiological and semasiological. This type of phraseographic repertoire is composed by three large sections in the dictionary body: the synoptic-analogical part, the analogical (both corresponding to the onomasiological part of the work) and the alphabetical (with a semasiological characteristic). In developing this project, we used the *Corpus* Brasileiro and the Web as *corpora*. As tool for the elaboration of the dictionary body, we used the FieldWorks Language Explore software, FLEx. As a result, it was possible to verify that these computational instruments were of fundamental relevance for the accomplishment of the research purpose.

KEYWORDS: Phraseography. Computational Linguistics. Corpus Linguistics. Natural Language Processing. Ideological Dictionary.

[*] Doutor em Letras pelo Programa de Pós-graduação em Letras (CPTL/UFMS), professor da Faculdade de Educação (FAED/ UFMS). ORCID: https://orcid.org/0000-0001-5562-8485. thyago.cruz@ufms.br

1 Introdução

Há tempos que se reconhece a relevância nas pesquisas lexicográficas do uso de dados informatizados em *corpora* para a elaboração de dicionários. Em meados da década de 60 do século XX, a recolha para a descrição e a análise das unidades do léxico eram realizadas por meio de coleta e registro em fichas – tarefa que demandava tempo e, muitas vezes, gastos elevados. Foi a partir do advento da computação e do acesso mais facilitado a suas máquinas que se possibilitou uma revolução nas pesquisas de caráter lexicográfico (BIDERMAN, 1984, p. 17), pois os dados poderiam ser armazenados, selecionados, analisados, corrigidos, recuperados e com um custo muito menor.

Neste trabalho, fruto de uma pesquisa de doutoramento, objetivamos demonstrar a aplicabilidade dos recursos informáticos nos estudos linguísticos, em especial os advindos da Linguística de *Corpus* (doravante LC) e do Processamento de Linguagem Natural (PLN), para a elaboração do protótipo de um dicionário ideológico de locuções.

Com relação à finalidade do dicionário ideológico de locuções, esse modelo de obra fraseográfica tem como propósito armazenar e descrever as locuções de um modo onomasiológico (na parte sinóptico-analógica[1] e na analógica[2]) e também de um modo

[1] O dicionário ideológico de locuções se divide em três seções principais do corpo do dicionário: a parte sinóptica, a analógica e a alfabética. Os quadros sinóptico-analógicos estão antecedidos pelo plano de classificação de mundo, elaborado por meio da análise semântica do material fraseológico selecionado e se divide em duas macrocategorias com suas ramificações: mundo externo (água, animal, ar, botânica, coisa material, espaço, fogo, quantidade e qualidade, tempo, terra) e ser humano (ação, aspecto, ciclos da vida, ciência, corpo humano, faculdades cognitivas, profissão, relações de influência e/ou posse, relações de preferência, religião e crenças, sentido). Sua elaboração foi facilitada graças aos recursos do software FieldWorks Language Explore, o FLEx, que, após comandos pré-definidos, organizaram os verbetes sob o campo lexical ao qual pertencem, como dissertamos mais adiante neste artigo.

[2] Já a parte analógica, também facilitada sua organização graças aos recursos do FLEx, possui a mesma rede de analogia dos quadros sinóptico-analógico, porém com um visual menos poluído, isto é, os verbetes não estão inseridos em quadros e se dispõem em ordem alfabética, partindo do conceito (ou do lexema que constitui o fraseologismo) para as locuções que lhes são análogas.

semasiológico (na parte alfabética[3]), tendo como público-alvo os potenciais interessados em estudos linguísticos e também tradutores.

Reconhecidas as características do dicionário ideológico de locuções, passamos a um segundo momento em que discorremos sobre o Processamento de Linguagem Natural) e Linguística de *Corpus* para, na sequência, indicar como essas metodologias científicas foram empregadas na nossa pesquisa.

2 Algumas considerações sobre o Processamento de Linguagem Natural e sobre a Linguística de *Corpus*

O uso da tecnologia nos estudos linguísticos vem, a cada década, incrementando-se e aperfeiçoando-se. Conforme Impacta (2019) e Souza (2019), o campo de Linguística Computacional[4] teve início em meados da década de 50, no contexto da Guerra Fria, quando se empregaram computadores a fim de traduzir para o inglês, de forma célere e automática, documentos redigidos em outras línguas. Embora essas máquinas nem se comparem com as atuais, no que diz respeito à qualidade e eficiência, já se era possível traduzir, por exemplo, textos do russo para o inglês de um modo bem satisfatório.

Vieira (2004), por sua vez, coaduna com Impacta (2019) e Souza (2019) sobre a origem da Linguística Computacional, pois, para a pesquisadora:

> a área possui aproximadamente meio século de existência, começou juntamente com a área de Inteligência Artificial (IA), que tem o

[3] Com relação à parte alfabética, também conhecida como índice remissivo nessa tipologia de dicionário, trata-se da apresentação dos verbetes com entradas dispostas em ordem alfabética e que, na sequência, se oferecem a definição e exemplo de uso, além de outras marcas, como função gramatical e indicação direta aos quadros sinóptico-analógicos (obrigatórias) ou contorno, marcas de uso, outras informações (de cunho gramatical, ortográfico, pragmático ou histórico-cultural) e relações semânticas de sinonímia ou antonímia (estas últimas não obrigatórias).

[4] Embora não seja um posicionamento unânime e haja discussão na área, consideramos, neste artigo, a Linguística Computacional e o Processamento de Linguagem Natural como termos sinônimos.

> objetivo de reproduzir comportamento inteligente em sistemas computacionais, como a solução de problemas e automatização do raciocínio (VIEIRA, 2004, p. 1).

Poersch (1987, p. 97) reconhece a Linguística Computacional como "sendo uma ciência interdisciplinar na qual o linguista se serve de, fornece subsídio a, e interage com a ciência da computação". Em outras palavras, esse domínio de estudo considera as máquinas computacionais como uma ferramenta de trabalho que possui a finalidade de processar editar, controlar e/ou analisar dados linguísticos. A Linguística Computacional possibilita o desenvolvimento de *softwares* básicos para pesquisas. Desse modo colabora, mutuamente com a Informática, com o desenvolvimento cada vez mais apurado da Inteligência Artificial.

Domínguez Burgos (2002, p. 104) elenca algumas funcionalidades que programas advindos das pesquisas em Linguística Computacional podem exercer: construir modelos de teorias linguísticas; auxiliar no ensino de língua estrangeira; apontar as correções ortográficas e normativas em textos de um dado idioma; reconhecer a voz humana e processar a mensagem contidas nas frases enunciadas por qualquer indivíduo; elaborar sistemas que facilitem o trabalho do pesquisador, que antes era realizado manualmente, como a construção de verbetes de dicionários; criar jogos virtuais que utilizem, de alguma forma, os comandos da linguagem natural; realizar traduções automáticas ou auxiliar os tradutores nesse processo; e inclusive produzir voz artificial cada vez mais próxima da humana, com transmissão de informação em alto grau de inteligibilidade.

A linha divisória entre esses dois âmbitos de estudos nem sempre está muito clara, haja vista que, no decorrer de várias pesquisas, a LC e o PLN se entrecruzam. Contudo, concordamos com os argumentos de Finatto, Lopes e Ciulla que consideram que o PLN: "denota especificamente o objeto da pesquisa de desenvolvimento de sistemas computacionais capazes de processar objetos de natureza linguística" (2015,

p. 43) e não se trata de um sinônimo de Linguística de *Corpus*, uma vez que esta é mais conhecida pela comunidade científica em geral da área da Linguagem. Tentaremos, brevemente, delimitar estes dois âmbitos de estudos nos parágrafos que se seguem.

A Linguística de *Corpus* busca armazenar amostras de linguagem natural, advindas de uma ou de variadas fontes, tanto na modalidade escrita como na oral. Direciona-se, portanto, a explorar a linguagem por meio de evidências empíricas, efetivadas com recursos da informática. Esse armazenamento denominado de *corpus* linguístico é tratado computacionalmente e se configura como:

> Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e a análise (BERBER SARDINHA, 2004, p. 18).

Ainda segundo Berber Sardinha (2004), há alguns pré-requisitos a serem cumpridos para a formação de um *corpus* linguístico, tais como a origem (deve ser de textos de linguagem natural), a autenticidade (escritos ou falados por nativos), o conteúdo (deve passar por critérios pré-estabelecidos pelo seu criador para que aquilo que foi coletado responda às características almejadas na investigação) e a representatividade (ser de uma extensão considerável e representativa).

Exemplos de *corpus* linguísticos podem ser encontrados no: *Corpus* Brasileiro[5], *Corpus* do Português[6]; *Corpus de Referencia del Español Actual* (CREA)[7]; *Corpus Diacrónico del Español* (CORDE)[8] e *Corpus of Contemporary American English (COCA)*[9].

Já o PLN direciona-se aos estudos da linguagem relacionados com a criação e desenvolvimento de *softwares*, aplicativos e sistemas computacionais. Para Othero (2006, p. 343), "cabe à área de PLN justamente a construção de programas capazes de interpretar e/ou gerar informações em linguagem natural".

Segundo Finatto, Lopes e Ciulla (2015), ainda que possa trabalhar com algum *corpus*, a finalidade do PLN não se reduz à descrição de elementos linguísticos, mas se propõe a oferecer soluções a problemas pontuais que se relacionam com o reconhecimento e a reprodução da linguagem natural em uma dada escala, priorizando a relação baixo custo e alto benefício.

Dias-da-Silva (1996) considerava o PLN como um "laboratório em ebulição", pelo fato de que, a cada ano, a indústria tecnológica crescia (e ainda cresce!) vertiginosamente. Com isso, os sistemas vinculados ao PLN se sofisticam em seus mais variados programas informáticos e buscam oferecer produtos tecnológicos de forma acessível e útil. Para o autor:

> a pesquisa [em PLN] reveste-se de um caráter tecnológico e transforma-se em um objeto cobiçado pela voraz indústria da informática que, cada vez mais, precisa tornar seus produtos menos 'enigmáticos' e mais adaptados às necessidades dos seus clientes" (DIAS-DA-SILVA, 1996, p. 66).

[5] Disponível em: https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS Acesso em 23 ago. 21.
[6] Disponível em: https://www.corpusdoportugues.org/ Acesso em 23 ago. 21.
[7] Disponível em: https://corpus.rae.es/creanet.html Acesso em 23 ago. 21.
[8] Disponível em: https://corpus.rae.es/cordenet.html Acesso em 23 ago. 21.
[9] Disponível em: https://www.english-corpora.org/coca/ Acesso em 23 ago. 21.

O pesquisador em PLN, portanto, empenha-se em tornar os sofisticados *softwares* em produtos de manuseio mais simples para o usuário, para que se diminua o risco de que este os abandone; ou prefira utilizar métodos mais exaustivos por não compreender o funcionamento do sistema informático criado ou por considerar que o tempo dispensado para entender seu funcionamento acabe sendo muito grande, equiparando-se ao período utilizado se adotassem os tradicionais procedimentos manuais.

Atualmente, temos como exemplos de programas de PLN: o Extrator Automático de Termos para Ontologias em Língua Portuguesa (ExATOlp), que emprega técnicas modernas de PLN em textos escritos em português, o *Neural Machine Translation* (NMT), um sistema de tradução automática sofisticado capaz de lidar com estruturas complexas da língua de modo satisfatório (como as unidades fraseológicas) e o *FieldWorks Language Explorer* (FLEx), a ser mais bem explorado neste artigo em tópico posterior.

Após essa contextualização sobre a Linguística de *Corpus* e o PLN, direcionamo-nos aos *corpora* empregados na elaboração do protótipo do dicionário ideológico de locuções e sobre o uso e a viabilidade da ferramenta computacional FLEx para a construção da parte sinóptico-analógica, da parte analógica e da parte alfabética do referido modelo de obra fraseográfica.

3 *Corpus* Brasileiro e *Web* como *corpus:* A importância do uso de *corpora* para a elaboração de um dicionário ideológico de locuções

Conforme Penadés Martínez (2015, p. 73), as locuções que compõem um dicionário de locuções não aparecem *ex nihilo* e nem devem partir exclusivamente da competência linguística do fraseógrafo – é aconselhável que sejam extraídas de fontes primárias e de secundárias. Neste contexto, é que surge a importância de *corpora* linguísticos que, ademais de possibilitarem a extração das unidades, permitem a

verificação da frequência de uso, a extração de exemplos, a identificação de significados ou de combinatórias que, muitas vezes, o redator jamais intuiria existir. Para a elaboração de um dicionário, portanto, é necessário que se parta de pelo menos um *corpus*.

Tendo em consideração essas informações sobre o labor lexicográfico aliado ao emprego de *corpus* (*corpora*) na redação de dicionários, escolhemos o *Corpus Brasileiro* e a *Web* como fontes primárias para a extração de algumas unidades fraseológicas para a elaboração do protótipo do dicionário ideológico de locuções, além da coleta de exemplos, de identificação dos significados e das estruturas argumentais, actanciais[10] ou informações pragmáticas (distintas das apresentadas nos dicionários analisados). Cabe-nos ressaltar que a grande maioria das locuções foi retirada de fontes secundárias – nas obras Ferreira (2010), Tesouro do Léxico Patrimonial Galego-Português[11] e o Dicionário de Expressões Idiomáticas[12] – uma vez que começamos a extração e a seleção das locuções por meio delas.

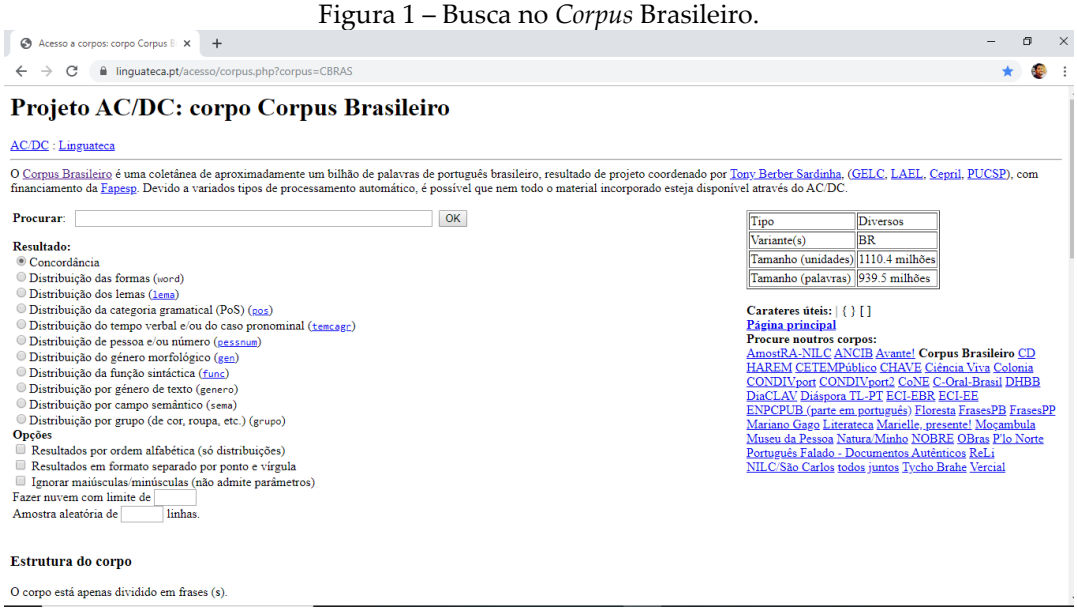
A escolha de *Corpus* Brasileiro deveu-se ao fato de que, atualmente, trata-se de um dos maiores *corpora* construído de palavras do português brasileiro, ao reunir mais de um bilhão delas. Idealizado e elaborado pelo Grupo de Estudo de Linguística de *Corpus* (GELC), coordenado pelo professor e pesquisador Tony Berber Sardinha, a base de dados online possibilita a busca de unidades lexicais (simples e compostas), terminológicas e fraseológicas; possui registros coletados dos mais variados gêneros textuais tanto em sua modalidade escrita como oral, além de apresentar a frequência de ocorrência da unidade pesquisada, no *corpus*. O *Corpus Brasileiro* pode ser acessado por meio do SketchEngine (plataforma paga) ou na Linguateca (acesso gratuito).

---

[10] As estruturas argumentais e actanciais referem-se à valência das locuções verbais e também de algumas nominais, adjetivas e adverbiais (PENÁDES MARTÍNEZ, 2015, p. 188)
[11] Disponível em: http://ilg.usc.es/Tesouro/pt. Acesso em: 20 jan. 2022.
[12] Disponível em: http://www.deipf.ibilce.unesp.br/pt/index.php Acesso em: 20 jan. 2022.

Para se localizar uma unidade fraseológica, deve-se desmembrar e separar por meio de aspas todos os seus elementos constituintes. Por exemplo, para a unidade "a cavalo" digita-se no campo *Procurar*: "a" e, em seguida, "cavalo". Deve estar selecionada a opção concordância para que ele faça a busca exata destas palavras. Vejamos o *layout* dessa página na *Web*:

Figura 1 – Busca no *Corpus* Brasileiro.



Fonte: elaborado pelo autor.

Com relação às locuções verbais, uma vez que as retiradas de fontes secundárias estão registradas em sua forma infinitiva, realizamos várias tentativas de buscas no *Corpus*, para cada unidade, conjugando-se o elemento verbal de sua estrutura em diferentes tempos e modos. Como, a princípio, para a elaboração deste protótipo de dicionário de locuções, preocupamo-nos somente se o fraseologismo em questão ainda está em uso e não registramos se seria uma forma pouco frequente, frequente ou muito frequente, este critério apresentado se tornou viável.

Ao todo eliminamos, até a elaboração da versão final do protótipo de dicionário, 76 locuções, extraídas das fontes secundárias, mas que não tinham nenhuma ocorrência no *Corpus* Brasileiro, nem na *Web*. Em contrapartida, incluímos mais 67

novas locuções, que correspondem a aproximadamente quinze por cento das locuções frequentes nos *corpora* analisados. A inclusão desses novos fraseologismos foi realizada por meio dos seguintes passos:

1- ou partiram de anotações de locuções ouvidas, por nós, em diálogos, em programas radiofônicos ou televisivos, que possuíssem o sema de /+ rural/ [13]e que houvesse ocorrência em um dos dois *corpora*;
2- ou por meio de digitação de partes dos fraseologismos já coletados nas fontes secundárias, em especial, os de maiores extensões, no *Corpus* Brasileiro (como em "botar o carro na frente dos bois", realizaram-se as seguintes buscas: "carro" "na" "frente" "do" "boi" > "frente" "do" "boi" > "do" "boi");
3- ou por meio da busca por unidade lexical que remetesse ao campo lexical do rural (por exemplo, digitávamos no campo de busca a unidade lexical "pato" e verificávamos, uma por uma, qual constituía um fraseologismo). Este método é mais moroso por apresentar, na maioria das vezes, muitas ocorrências de aparição do lexema buscado.

Verificada a importância do *Corpus* Brasileiro e o modo empregado de busca, seleção e extração de unidades, discutimos sobre a consideração e o uso da *Web* como *corpus*.

Para muitos pesquisadores, como Kilgarriff e Grefenstette (2003), a *Web* pode ser considerada como o maior *corpus* linguístico existente na atualidade. Valença e Sabino (2016) fazem uma relação entre a quantidade de páginas indexadas pelo

[13] O critério de selecionar somente as unidades fraseológicas que possuíssem o sema de /+rural/ se deve ao fato de que, por se tratar de uma pesquisa de doutorado e de que seria necessário, portanto, realizarmos recortes, elegemos, como ponto de partida, as locuções que continham um significado que as ligasse a esse conteúdo semântico (como, por exemplo, "aberto dos peitos", que se diz da cavalgadura ou do animal de sela que vive caindo devido a esforços extremos realizados) ou que um dos lexemas que compunha o fraseologismo se vinculasse ao referido sema (como na locução "amolar o boi", de significado "aborrecer", em que há um elemento em sua estrutura que se remete ao mundo rural, isto é, "boi").

buscador *Google* (mais de 60 bilhões) com o possível número de palavras existentes nelas, o que levaria, segundo as autoras, a uma quantidade quase imensurável desses lexemas.

Concordamos com Lüdeling, Evert e Baroni (2007) que assinalam que muitos *corpora* linguísticos tradicionais (armazenados em um banco de dados e sistematizados mediante uma programação) são adequados para determinados propósitos de pesquisa, como ao se desejar verificar quais são as unidades lexicais mais frequentes em um dado gênero textual. Todavia, percebemos, no decorrer desta pesquisa, que muitas unidades que não estavam presentes no *Corpus* Brasileiro, apareciam na *Web*, além de que o número de ocorrência dos fraseologismos naquele era, na grande maioria dos casos, inferior ao demonstrado neste. Sobre essa perspectiva, Lüdeling *et alii*, (2007, p. 7) acrescentam:

> [...] há casos em que os dados necessários para responder ou explorar uma pergunta não podem ser encontrados em um corpus padrão, porque o fenômeno em consideração é raro (dados escassos), pertence a um gênero ou registro não representado no corpus ou decorre de uma época em que os dados do corpus não cobrem (por exemplo, são novos demais) (LÜDELING *et alii*, 2007, p. 7) [14].

Essa afirmação vem ao encontro do assinalado por Valença e Sabino (2016, p. 482), isto é, "a frequência de fraseologismos presentes nesses *corpora* [tradicionais] é relativamente baixa, o que torna esses ambientes de pesquisa menos relevantes que a *web* para pesquisas fraseológicas". Concordamos com as autoras de que se torne menos relevante a busca em *corpora* tradicionais se partirmos somente do prisma da análise

[14] […] there are cases in which the data needed to answer or explore a question cannot be found in a standard corpus because the phenomenon under consideration is rare (sparse data) belongs to a genre or register not represented in the corpus, or stems from a time that the corpus, or stems from a time that the corpus data do not cover (for example, it is too new) (tradução nossa).

de frequência de usos dos elementos investigados. Já no que diz respeito ao controle e sistematização rigorosos do *corpus* pelo linguista, à dificuldade de um motor de busca eficiente, ao risco de encontrar uma linguagem de caráter artificial e à possível presença de erros ortográficos ou desvios gramaticais no material linguístico coletado, trazem, segundo Colson (2007), objeções a esse modo de pesquisa.

Ao levar esses fatores em consideração, tanto as vantagens como as desvantagens de empregar a *Web* como *corpus*, decidimos incluí-la, pela possibilidade de que ela pode trazer à vista do pesquisador unidades fraseológicas, significados novos e exemplos que não foram encontrados em um *corpus* tradicional, mas sempre com olhos atentos também para as dificuldades e objeções elencadas por Colson (2007).

Os passos assumidos para a busca das locuções, seus significados, seus contornos[15] e exemplos, foram:

1. acessamos o site *Google*, digitamos no campo de busca a locução entre aspas (para que fosse exatamente esta expressão). Em seguida, na opção de "Configurações", escolhemos "Pesquisas avançadas". Nela, selecionamos o idioma (Português) e o país (Brasil) – o site fez as buscas somente nesse idioma e espaço virtual;
2. para as locuções verbais, atuamos como fizemos nos passos metodológicos para a busca no *Corpus* Brasileiro, ou seja, conjugamos o lexema verbal que forma parte da estrutura do fraseologismo, em vários tempos e modos (por exemplo: cair do cavalo, caindo do cavalo, caído do cavalo, caio do cavalo, caiu do cavalo, caímos do cavalo, cairá do cavalo etc.);
3. para as locuções de extensões maiores, aos poucos, eliminávamos um de seus elementos constituintes para verificar a possibilidade de alguma variação

[15] Os contornos lexicográficos correspondem aos traços subcategoriais ou contextuais que acompanham a "definição propriamente dita" do verbete.

lexical dentro do fraseologismo. Há também a possibilidade de inserir, apenas uma vez por busca, um asterisco no lugar da palavra excluída – isso faz com que o buscador mostre outros elementos que podem compor as unidades pesquisadas (por exemplo, para a locução "caminho das cabras" é recomendável colocar como "caminho * cabras", pois, se não, há a possibilidade de não se encontrar nenhuma variação do fraseologismo);

4. os erros e desvios gramaticais que apareceram, em especial nos exemplos extraídos, foram corrigidos;

Discutidos os processos de utilização do *Corpus* Brasileiro e da *Web* como *Corpus*, passamos à seção que descreve como o software *FieldWorks Language Explorer* (FLEx) foi empregado na pesquisa, bem como seus benefícios e limitações.

3 Processamento de Linguagem Natural: O FLEx como ferramenta útil para a construção de verbetes

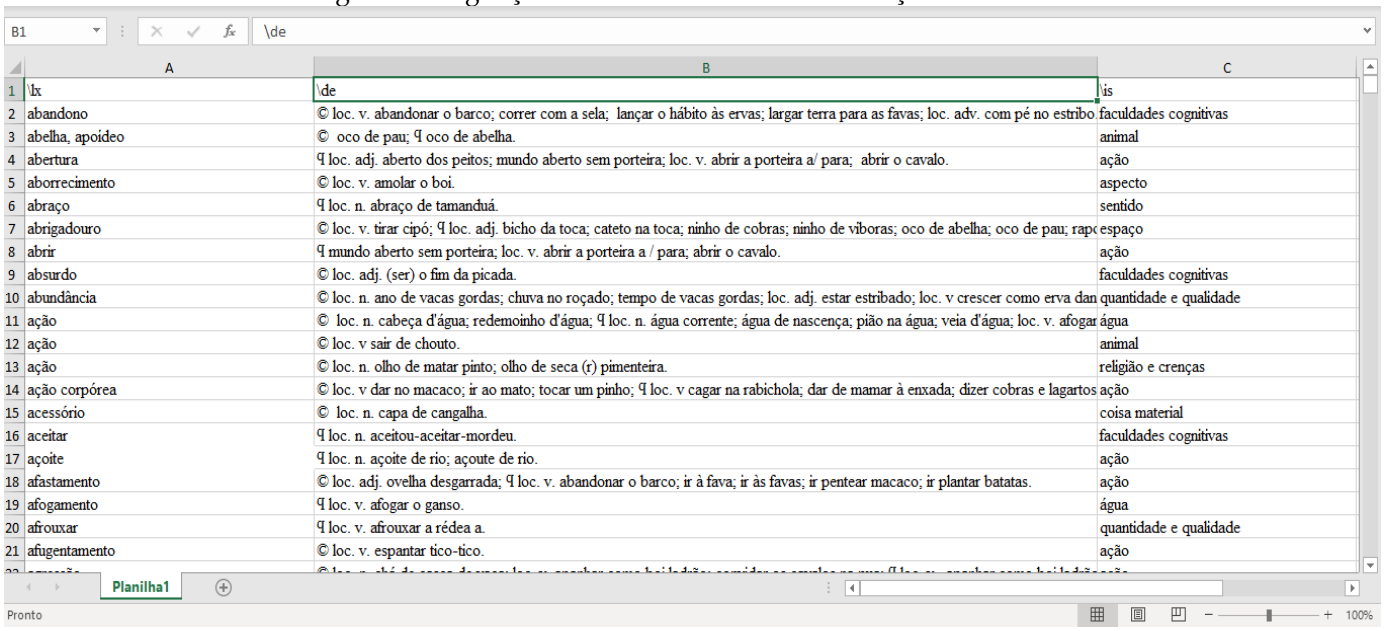
Para a elaboração dos verbetes (dos quadros sinóptico-analógicos, da parte analógica e da parte alfabética) empregamos o *software FieldWorks Language Explorer*, o FLEx. Trata-se de um programa criado e disponível para download[16], gratuitamente, pelo *Summer Institute of Linguistic* (*SIL International*). Essa ferramenta, que também pode ser usada nos estudos e pesquisas da Linguística de *Corpus*, organiza os dados e gera, de um modo automático, os verbetes para o usuário, mediante as configurações pré-estabelecidas.

Para a construção dos verbetes dos quadros sinóptico-analógicos e da parte analógica do dicionário ideológico de locuções, primeiramente, digitamos todos os dados em uma planilha de formato XLS. Haveria a possibilidade de colocá-los diretamente no programa, mas como ainda não tínhamos a segurança se este poderia

[16] Disponível em: https://software.sil.org/fieldworks/ Acesso em: 20 jan. 2022.

ser útil para a elaboração da microestrutura das diferentes seções, preferimos deixar gravado em um arquivo XLS, caso fosse necessário utilizar outro programa ou fazer manualmente em formato DOCS. Para tanto, a digitação de algumas codificações, no formato XLS, foi necessária para que o FLEx pudesse ler os dados e organizar os verbetes, como podemos ver na figura 2:

Figura 2 – Digitação dos comandos e das informações no Excel.



	A	B	C
1	\lx	\de	\is
2	abandono	© loc. v. abandonar o barco; correr com a sela; lançar o hábito às ervas; largar terra para as favas; loc. adv. com pé no estribo	faculdades cognitivas
3	abelha, apoideo	© oco de pau; ¶ oco de abelha.	animal
4	abertura	¶ loc. adj. aberto dos peitos; mundo aberto sem porteira; loc. v. abrir a porteira a/ para; abrir o cavalo.	ação
5	aborrecimento	© loc. v. amolar o boi.	aspecto
6	abraço	¶ loc. n. abraço de tamanduá.	sentido
7	abrigadouro	© loc. v. tirar cipó; ¶ loc. adj. bicho da toca; cateto na toca; ninho de cobras; ninho de viboras; oco de abelha; oco de pau; rapc	espaço
8	abrir	¶ mundo aberto sem porteira; loc. v. abrir a porteira a / para; abrir o cavalo.	ação
9	absurdo	© loc. adj. (ser) o fim da picada.	faculdades cognitivas
10	abundância	© loc. n. ano de vacas gordas; chuva no roçado; tempo de vacas gordas; loc. adj. estar estribado; loc. v crescer como erva dan	quantidade e qualidade
11	ação	© loc. n. cabeça d'água; redemoinho d'água; ¶ loc. n. água corrente; água de nascença; pião na água; veia d'água; loc. v. afogar	água
12	ação	© loc. v sair de chouto.	animal
13	ação	© loc. n. olho de matar pinto; olho de seca (r) pimenteira.	religião e crenças
14	ação corpórea	© loc. v dar no macaco; ir ao mato; tocar um pinho; ¶ loc. v cagar na rabichola; dar de mamar à enxada; dizer cobras e lagartos	ação
15	acessório	© loc. n. capa de cangalha.	coisa material
16	aceitar	¶ loc. n. aceitou-aceitar-mordeu.	faculdades cognitivas
17	açoite	¶ loc. n. açoite de rio; açoute de rio.	ação
18	afastamento	© loc. adj. ovelha desgarrada; ¶ loc. v. abandonar o barco; ir à fava; ir às favas; ir pentear macaco; ir plantar batatas.	ação
19	afogamento	¶ loc. v. afogar o ganso.	água
20	afrouxar	¶ loc. v. afrouxar a rédea a.	quantidade e qualidade
21	afugentamento	© loc. v. espantar tico-tico.	ação

Fonte: elaborado pelo autor.

O código "\lx" indica que os elementos daquela coluna devem ser lidos pelo programa como a entrada do verbete. Já o código "\de" corresponde à definição e "\is", ao domínio semântico (no caso, desta pesquisa, inserimos as macrocategorias).

Devido às características dos verbetes dos quadros e da parte analógica, tomamos alguns procedimentos para que o programa cumprisse com o objetivo que esperávamos dele. Como na definição necessitávamos da diferenciação das classes gramaticais das locuções e elas se encontravam dentro da planilha codificada por "\de", foi necessário que puséssemos manualmente, dentro de cada célula, as

referidas classificações, além também dos símbolos "©"[17] e "ɋ"[18]. Não nos pareceu viável abrir mais colunas com a codificação "\de", o que possibilitaria a inclusão do comando "\ng" (informação gramatical), pois, como decidimos que os verbetes destas duas partes se regeriam primordialmente pelos sentidos ou pela presença do lexema que constituiu o fraseologismo, o FLEx não conseguiria decodificar os dados para dispô-los como esta pesquisa almejava. Logo, elaboramos dois arquivos XLS, um com as locuções acompanhadas das classes gramaticais e outros sem essas informações.

Após todo esse processo, os dois arquivos XLS foram convertidos em formato *Simple File Manager* (SFM), por meio do programa SheetSwiper[19], pois se trata de uma extensão que pode ser lida pelo FLEx. Ao abrir o programa, escolhemos o arquivo que se desejava a conversão e em alguns segundos o novo formato já estava disponível. Salientamos que os dois arquivos SFM, no que diz respeito à inserção dos dados no FLEx, passaram pelo mesmo processo descrito a seguir.

O passo seguinte foi a inserção desses dados no programa FLEx. Para tanto, escolhemos a opção "Create a new Project", selecionamos o arquivo que se deseja abrir, nomeamos o projeto e seguimos alguns pequenos comandos que o programa solicita e o arquivo em seguida é aberto. A tela indicada para a elaboração dos quadros sinóptico-analógicos é a denominada "Dicionário Classificado", que organizará as etiquetas (isto é, as entradas dessa seção do dicionário ideológico) e as indicações lemáticas (isto é, as definições dessa seção do dicionário ideológico) a partir do

---

[17] O símbolo "©" presente na definição dos verbetes indica que, na sequência, apresentamos as locuções que se remetem conceitualmente ao lexema etiquetado na entrada. Por exemplo, na entrada "abundância", temos as unidades fraseológicas "ano das vacas gordas" e "chuva no roçado", dentre outras, que se relacionam conceitualmente ao elemento lematizado.

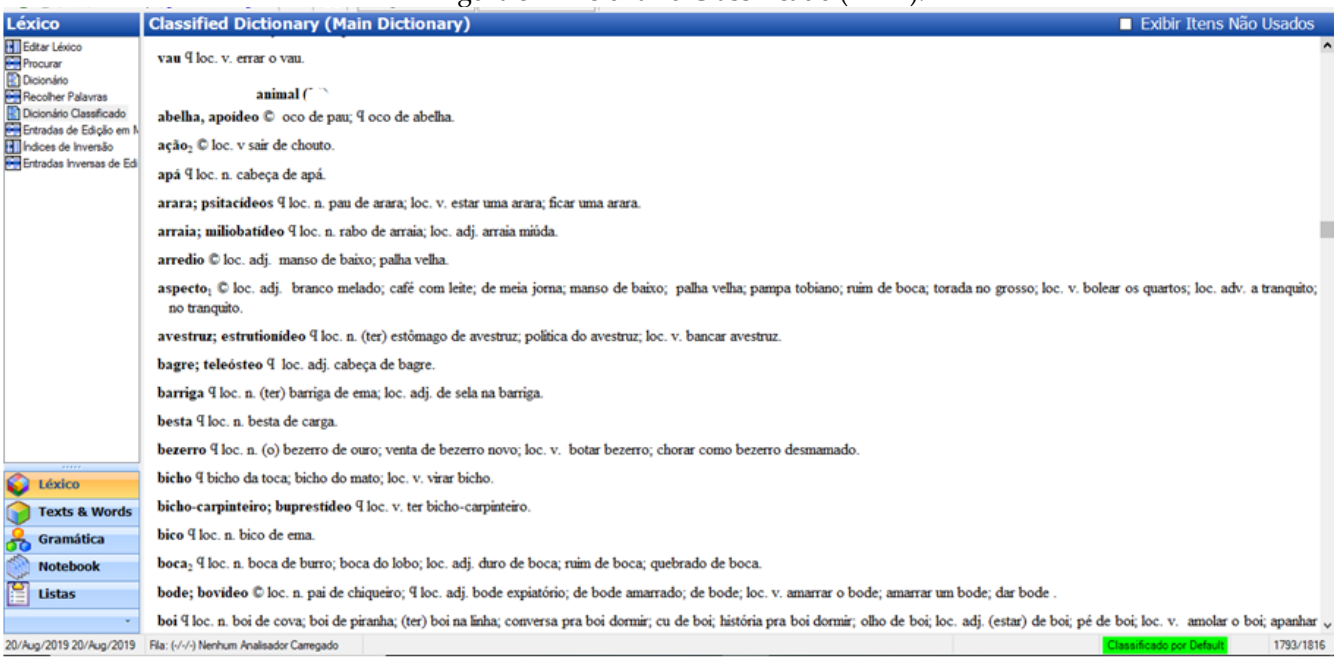
[18] O símbolo "ɋ" presente na definição dos verbetes sinaliza que o consulente encontrará, na sequência, locuções que, em sua estrutura, possuam o lexema (ou um radical em comum) do elemento lematizado. Por exemplo, na entrada "afrouxar", há na definição a locução "**afrouxar** a rédea"; ou em "assar", há "mão de mucura **assada**".

[19] Disponível gratuitamente para download em https://software.sil.org/sheetswiper/ Acesso em: 20 jan. 2022.

domínio semântico. No entanto, previamente, o usuário do programa deve direcionar-se à tela "Recolher Palavras" para configurar os domínios semânticos, pois, nesta seção, caso não se apague o que está identificado como "vernáculo", no momento em que os dados estiverem no "Dicionário Classificado," as macrocategorias aparecerão escritas repetidamente. Caso se julgue melhor, essa duplicação pode ser corrigida quando estiver em posse dos dados em formato DOC.

Após o processo de compilação de palavras, selecionamos a tela "Dicionário Classificado" e já aparecem os verbetes disponibilizados do modo almejado, como vemos na figura a seguir:

Figura 3 – Dicionário Classificado (FLEx).



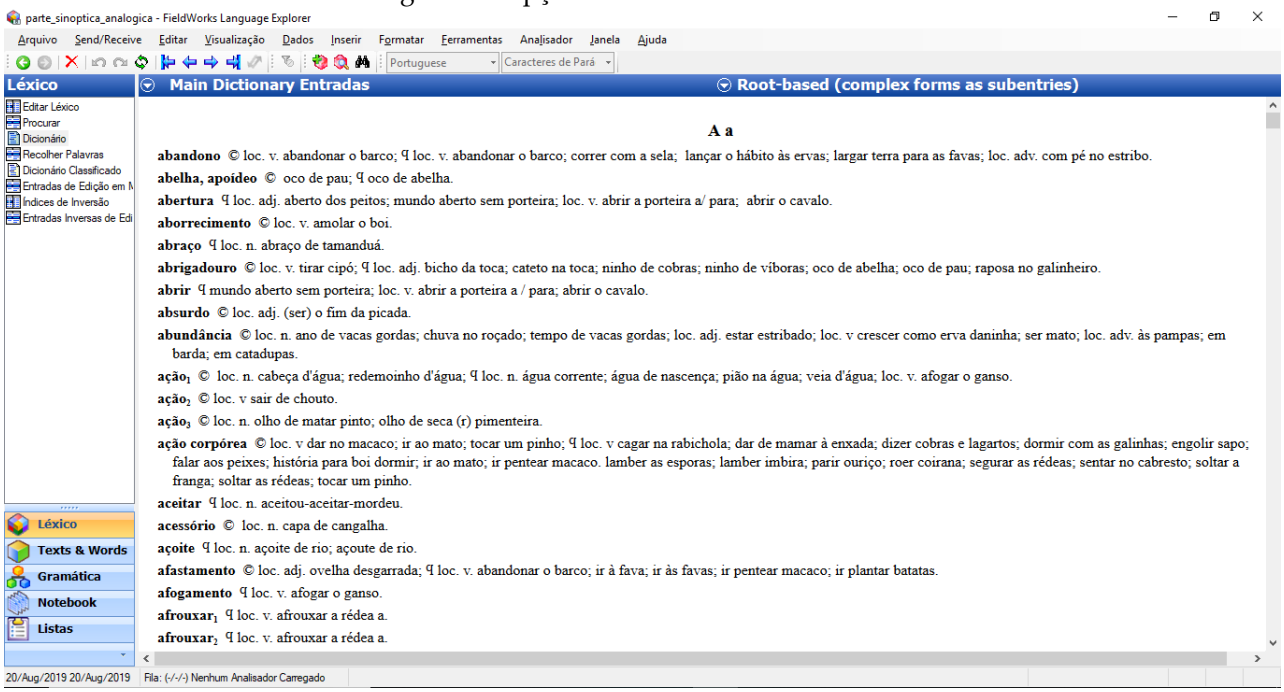
Fonte: elaborado pelo autor.

Para essa seção, o FLEx não possui a disponibilidade de conversão direta dos dados para o formato DOCS, somente para o XHTML. Por isso, foi necessário que os dados fossem convertidos e exportados para este último, posteriormente, aberto no navegador "Chrome", o qual possibilita salvá-lo em formato PDF para, em seguida,

ser convertido em DOCS. Para este trabalho, a conversão de PDF para DOCS foi realizada pelo site PDFCandy[20], uma vez que foi o que melhor realizou essa atividade, com perdas praticamente nulas do formato inicial. Realizado isso, cabe ao fraseógrafo definir a formatação do tamanho da fonte e outras que julgar necessário (como a inserção dos quadros e da numeração de cada verbete).

Para a parte analógica, com o arquivo que possui os verbetes sem as marcas de classificação gramatical, já convertido em formato SFM, seguimos os passos indicados até a abertura do arquivo no FLEx. Após isso, em vez de utilizar a tela "Dicionário Classificado", selecionamos a seção "Dicionário", que apresentará os verbetes dispostos em ordem alfabética sem as indicações das macrocategorias. É um processo semelhante a ser empregado na formação da parte alfabética, como perceberemos mais adiante. A seguir, a figura de alguns verbetes da parte analógica, no dicionário FLEx.

Figura 4 – Opção "Dicionário" no FLEx.



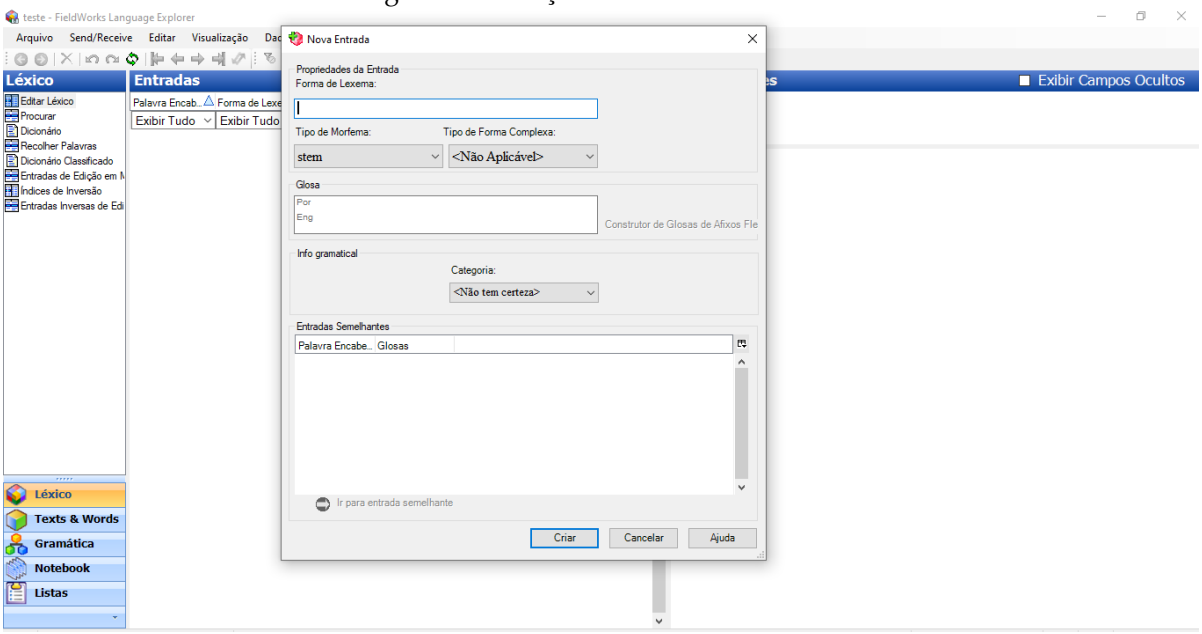
Fonte: elaborado pelo autor.

[20] Disponível em: https://pdfcandy.com/pt/result/ebde06fa.html Acesso em: 20 jan. 2022.

A exportação dos dados dessa tela para o formato DOCS é mais simples. Basta realizar o download, também gratuito, do programa Pathway[21]. Uma vez que esteja instalado no computador, devemos selecionar no FLEx a aba "Arquivo", logo após "Exportar" e escolher a opção "Dictionary, Reversal Index Pathway (various outputs)". Em seguida, elegemos a exportação para o programa Open Office/ Libre Office, que é compatível com o formato DOCS.

Com relação à parte alfabética, também foi realizada por meio do FLEx. No entanto, a modo de testar outras possibilidades, decidimos digitar os dados diretamente no programa. Para tanto, ao abri-lo, selecionamos a opção "Create a new Project". Escrevemos o nome do projeto e escolhemos o idioma. Selecionamos, na aba "Inserir" a opção "Entrada" e surgiu a seguinte janela:

Figura 5 – Inserção de entradas no FLEx.

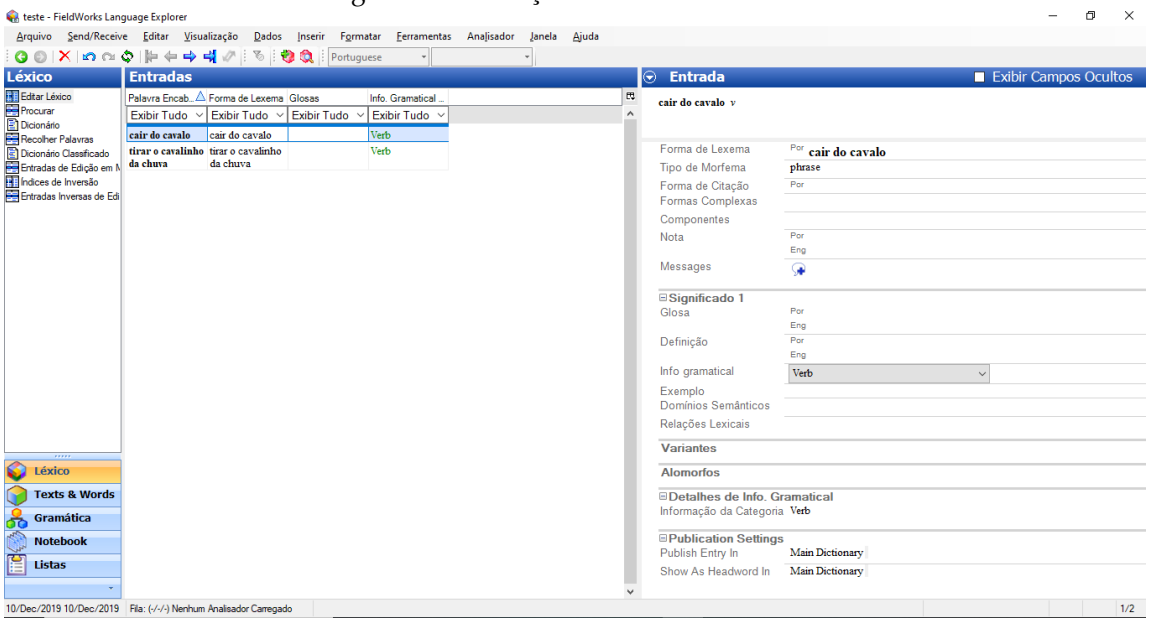


Fonte: elaborado pelo autor.

[21] Disponível em https://software.sil.org/products/ Acesso em: 20 jan. 2022.

No espaço "Forma de Lexema", dispomos a unidade fraseológica a ser lematizada. Automaticamente, a opção "Tipo de morfema" é preenchida como "phrase". Em "Tipo de forma complexa" recomendamos selecionar a opção "Não aplicável" para que no verbete não apareça a marca de elemento sem especificação ("unspecified complex form"). Pela característica assumida para o dicionário ideológico de locuções, não preenchemos a opção de glosa. Identificada a função gramatical que a locução pode assumir, selecionamos a opção que lhe é correspondente (nominal, adjetival, etc.) em "categoria". Caso houvesse mais de uma função, a depender do contexto de onde foram extraídas, decidimos dispô-las em acepções distintas. Por fim, selecionamos a opção "Criar". Realizado isso, o programa expõe a seguinte tela:

Figura 6 – Formação de verbetes no FLEx.



Fonte: elaborado pelo autor.

Com relação às marcas lexicográficas, devemos ir à aba "Configurar" e, em seguida, "Coluna" para adicionar a opção "Dialect Labels (Entry)"[22]. Já no espaço destinado ao significado, é possível preencher a definição e o exemplo, cada um em seu campo específico. Se há a necessidade de inserir mais exemplos ou mais acepções, basta clicar em "Inserir Exemplo" ou "Inserir Significado", respectivamente. No que diz respeito ao contorno lexicográfico, uma vez que decidimos colocá-lo entre parênteses antes das definições, realizamos isso manualmente, isto é, digitando-o diretamente no início de cada definição, no campo da "definição".

Para a inclusão das relações de sinonímia e/ ou antonímia entre as unidades, selecionamos, com o botão direito do *mouse*, a opção "Relações lexicais", e foi escolhida a remissão que almejávamos estabelecer. O *software* a faz, automaticamente, após o usuário determinar quais as unidades devem estar relacionadas.

Como a parte alfabética foi elaborada no FLEx em projetos diferentes das partes sinóptica-analógica e analógica e por não identificarmos um recurso que pudesse fazer a remissão dos verbetes aos quadros sinóptico-analógicos, realizamos essa inclusão de forma manual, após a construção de todos os verbetes.

5 Considerações finais

Discorremos, neste artigo, como a Informática e a Linguística podem estar unidas em outros ramos científicos, como a Linguística de *Corpus* e o Processamento de Linguagem Natural. Com o desenvolver das investigações nesses dois âmbitos,

[22] Mediante as considerações de Cardoso (2010, p. 25), quem atribui a existência de dialetos a um dado espaço geográfico aliado aos traços de formação etnolinguística subjacente a um ato de fala, cabe ressaltar que as marcas de uso "vulgar", presente no modelo de dicionário, não são consideradas como formas de dialeto. Contudo, assim como ocorre no preenchimento das indicações gramaticais e pragmáticas no espaço destinado no FLEx para informações enciclopédicas, realizou-se deste modo pela disposição que o programa colocaria essas estruturas, aproximando o verbete construído à configuração do verbete prototípico elaborado para esta tese.

pesquisadores, cada vez mais, têm a oportunidade de ter às mãos bancos de armazenamento de material linguístico e/ou ferramentas computacionais que facilitem seu trabalho e que possam reduzir a chance de equívocos na pesquisa e de corrigi-los caso haja.

Demonstramos como a LC e o PLN foram de fundamental importância para a elaboração do protótipo do dicionário ideológico de locuções. Apresentamos, ainda, alguns pequenos obstáculos (como o caráter limitado de um *corpus* fechado, ou os riscos de não se trabalhar com uma linguagem natural ou as incorreções gramaticais que podem estar nas unidades léxicas da *Web*, ou as impossibilidades ou dificuldades de comandos no FLEx, dada à peculiaridade da obra fraseográfica delineada). Contudo, esses impasses puderam ser facilmente contornados, o que não desvalida a utilidade e o proveito que essas ferramentas computacionais podem proporcionar ao pesquisador.

Referências

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

BIDERMAN, M. T. C. A ciência da Lexicografia. **Alfa Revista de Linguística**. São Paulo, n. 28 (supl.), p. 1-26, 1984.

CARDOSO, S. A. **Geolinguística**: tradição e modernidade. São Paulo: Parábola Editorial, 2010.

COLSON, J-P. Corpus linguistics and phraseological statistics: a few hypotheses and examples. *In*: BURGER, H., HÄCHI BUHOFER, A., GRÉCIANO, G. (ed.). **Flut von texten – vielfalt der kulturen**. Ascona 2001 zu Methodologie und kulturspezifik der phraseologie. Baltmannsweiler: Schneider Verlag Hohengehren, p. 47-59, 2003.

DIAS-DA-SILVA, B. C. **A fase tecnológica dos estudos da linguagem:** o processamento automático das línguas naturais. Tese (Doutorado em Linguística e Língua Portuguesa) – Faculdade de Ciências e Letras, UNESP, Araraquara, 1996.

DOMÍNGUEZ BURGOS, A. Lingüística computacional: un esbozo. **Boletín de lingüística**, v. 18, p. 104-119, 2002. Disponível em: http://saber.ucv.ve/ojs/index.php/rev_bl/article/view/1437 Acesso em: 20 jan. 2022.

FERREIRA, A. B. H. **Dicionário Aurélio da Língua Portuguesa**. 5. ed. Rio de Janeiro: Positivo, 2010.

FINATTO, M. J. B.; LOPES, L.; CIULLA, A. Processamento de Linguagem Natural, Linguística de Corpus e Estudos Linguísticos: uma parceria bem-sucedida. **Domínios de Lingu@gem**. Uberlândia, MG. Vol. 9, n. 5 (dez. 2015), p. 41-59, 2015. Disponível em: https://seer.ufu.br/index.php/dominiosdelinguagem/article/view/28670/17075. Acesso em: 20 jan. 2022. DOI https://doi.org/10.14393/DLE-v9n5a2015-3

IMPACTA. **Conheça tudo sobre área de Linguística Computacional!** [S.l.], 2019. Disponível em: https://www.impacta.com.br/blog/conheca-tudo-area-linguistica-computacional/. Acesso em: 20 set. 2021.

KILGARRIFF, A.; GREFENSTETTE, G. Introduction to the special issue on the Web as Corpus. **Computational Linguistics**, 29(3), p. 333-347, 2003. DOI https://doi.org/10.1162/089120103322711569

LÜDELING, A; EVERT, S; BARONI, M. Using web data for linguistic purposes. *In*: HUNDT, M; NESSELHAUF, N; BIEWER, C (org.). **Corpus linguistics and the web**. Amsterdam: Rodopi, 2007.

OTHERO, G.A. Lingüística Computacional: uma breve introdução. **Letras de hoje**, v. 41, n. 2, 2006.

PENÁDEZ MARTÍNEZ, I. **Para un diccionario de locuciones:** de la lingüística teórica a la fraseografía práctica. Alcalá: Universidad de Alcalá, 2015. DOI https://doi.org/10.4067/S0718-93032016000100010

POERSCH, J. M. Linguística computacional: elaboração do diploma para a língua portuguesa. **Letras de Hoje**, v. 22, n. 1, 1987.

SOUZA, J. W. C. **Me vê um texto menor, por favor?** [S. l.], 2019. Disponível em: http://www.roseta.org.br/2019/07/31/me-ve-um-texto-menor-por-favor/ Acesso em: 20 set. 2021.

VALENÇA, E. M; SABINO, M. A. O uso da Web como corpus em pesquisas fraseológicas: Uma prática prejudicial ou um recurso valioso? **Calidoscopio**, v. 14, n. 3, p. 480-488, 2016. Disponível em: http://revistas.unisinos.br/index.php/calidoscopio/article/view/cld.2016.143.11 Acesso em: 20 jan. 2022. DOI https://doi.org/10.4013/cld.2016.143.11

VIEIRA, R. Lingüística Computacional: uma entrevista com Renata Vieira. **Revista Virtual de Estudos da Linguagem - ReVEL**. Vol. 2, n. 3, agosto de 2004. Disponível em: http://www.revel.inf.br/files/entrevistas/revel_3_entrevista_renata_vieira.pdf Acesso em: 20 jan. 2022.

Artigo recebido em: 27.08.2021 Artigo aprovado em: 21.01.2022

DOI: 10.14393/DL52-v16n4a2022-6

Modelação da valência verbal numa gramática computacional do português no formalismo HPSG

Modeling verb valency in a computational grammar for Portuguese in the HPSG formalism

Leonel Figueiredo de ALENCAR[*]
Alexandre RADEMAKER[**]

RESUMO: A HPSG é uma teoria gramatical lexicalista que propõe a formalização em paralelo das estruturas morfossintáticas e semânticas. Este trabalho descreve a implementação computacional de valências verbais numa nova gramática do português nesse formalismo. Essa gramática é relevante não somente para aplicações de compreensão textual, mas representa também uma contribuição à documentação formal das estruturas da língua, destacando-se pelo tratamento das construções de controle e alçamento. A gramática tem sido implementada incrementalmente por meio da sua aplicação a conjuntos de teste cada vez mais abrangentes. Com 278 entradas e um total de 215 lemas, o léxico verbal ainda é pequeno. No entanto, a hierarquia de tipos proposta modela as propriedades de 118 classes valenciais, das quais 57 são tipos que codificam classes de verbos. A gramática analisa 94% de um total de 581 sentenças gramaticais, apresentando, ao

ABSTRACT: HPSG is a lexicalist grammatical theory that proposes the parallel formalization of morphosyntactic and semantic structures. This work describes the computational implementation of verbal valences in a new Portuguese grammar in this formalism. This grammar is relevant not only for text understanding applications, but also represents a contribution to the formal documentation of the structures of the language, standing out for its treatment of control and raising constructions. The grammar has been incrementally implemented through its application to increasingly comprehensive test suites. With 278 entries and a total of 215 lemmas, the verb lexicon is still small. However, the proposed type hierarchy models the properties of 118 valence classes, of which 57 are types that encode verb classes. The grammar analyzes 94% of a total of 581 grammatical sentences, while showing low hypergeneration in a set of 167 ungrammatical examples.

[*] Doutor em Linguística, Professor Titular da Universidade Federal do Ceará e Professor Visitante da EMAp/FGV. ORCID: http://orcid.org/0000-0001-8148-6994. leonel.de.alencar@ufc.br
[**] Doutor em Informática, Professor Adjunto da EMAp/FGV e Pesquisador da IBM Research. ORCID: http://orcid.org/0000-0002-7583-0792. alexrad@br.ibm.com

mesmo tempo, baixa hipergeração em um conjunto de 167 exemplos agramaticais.

PALAVRAS-CHAVE: Linguística computacional. Engenharia da gramática. Análise sintática automática. Valência. Semântica computacional.

KEYWORDS: Computational linguistics. Grammar engineering. Syntactic parsing. Valence. Computational semantics.

1 Introdução

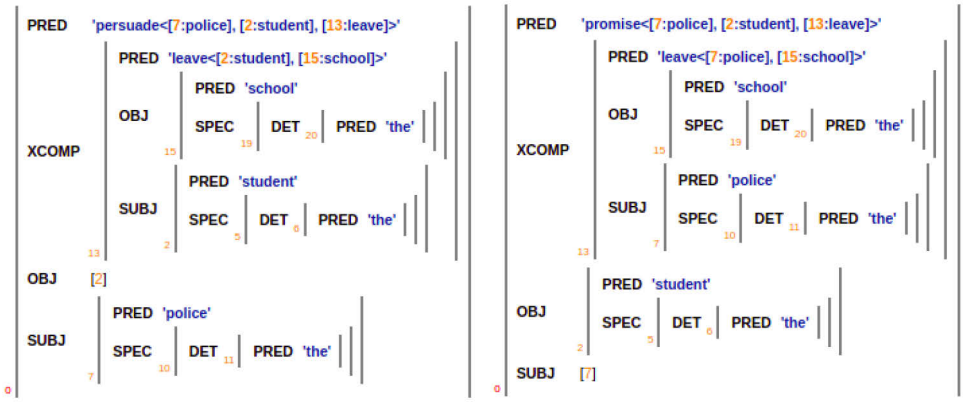
A análise sintática automática profunda é um componente de diversos sistemas comerciais de processamento de linguagem natural (doravante PLN). Essa tarefa consiste em construir, para as sentenças de um texto, representações suficientemente informativas para utilização por sistemas de compreensão textual. Um exemplo bastante característico é o Watson da IBM, sistema de resolução de perguntas representativo do estado da arte. Subjaz a esse sistema uma arquitetura híbrida, que conjuga abordagens estatísticas com simbólicas, da qual faz parte uma gramática computacional do inglês elaborada manualmente (FERRUCCI *et al.*, 2010; MCCORD; MURDOCK; BOGURAEV, 2012). Um dos principais marcos da inteligência artificial da última década, esse sistema, hoje amplamente empregado na área médica, venceu em 2011 dois campeões do *Jeopardy!*, um popular programa de perguntas e respostas da TV estadunidense.

É inegável que, em muitas aplicações de PLN, análises sintáticas rasas produzem resultados satisfatórios a um custo drasticamente menor do que o demandado pela implementação de analisadores profundos. No entanto, o maior esforço para construir esse tipo de componente é compensado pela maior riqueza de informações capaz de fornecer ao processamento semântico.

Para elucidar a distinção entre os dois tipos de análise sintática, comparem-se as diferentes representações de (1) e (2) nas Figuras 1 e 2.

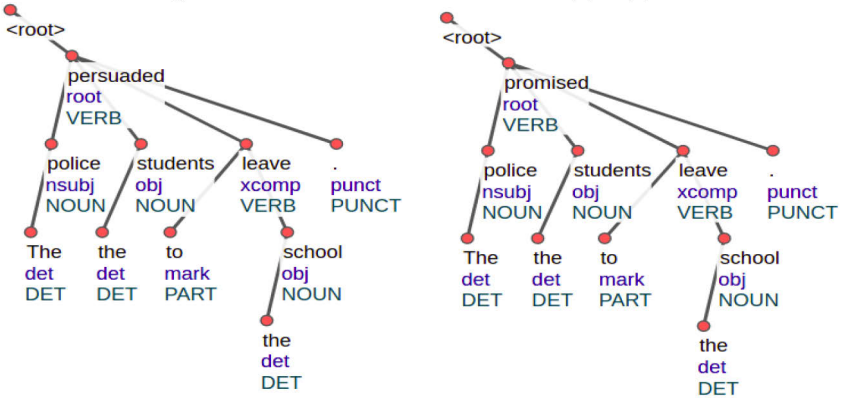
(1) The police persuaded the students to leave the school.
a polícia persuadir:PST;3SG os estudantes a deixar:INF a escola
A polícia convenceu os estudantes a deixarem a escola.'
(2) The police promised the students to leave the school.
a polícia prometer:PST;3SG os estudantes a deixar:INF a escola
'A polícia prometeu aos estudantes deixar a escola.'

Figura 1 – Análise sintática profunda de (1) e (2).



Fonte: gerada pelo programa XLE-Web (ROSÉN *et al.*, 2012) a partir da gramática do inglês no formalismo LFG.

Figura 2 – Análise sintática rasa de (1) e (2).



Fonte: gerada pelo analisador sintático estatístico UDPipe 2 com base no modelo *english-ewt-ud-2.6-200830* (STRAKA; STRAKOVÁ, 2017).

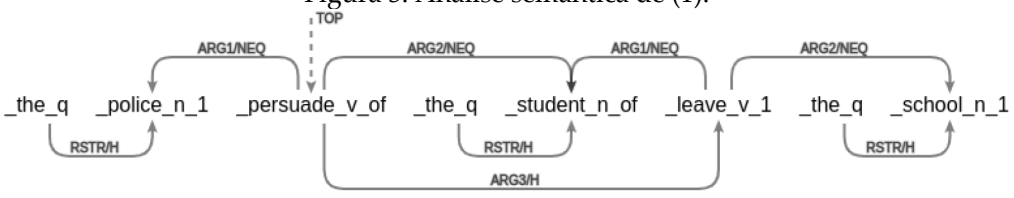
Nas duas representações da Figura 1, os verbos principais expressam uma relação de três lugares envolvendo duas entidades, expressas por *a polícia* e *os*

estudantes, e o estado de coisas referido pelo verbo encaixado. Esses três argumentos são realizados pelas funções sintáticas SUBJ (sujeito), OBJ (objeto) e XCOMP, denominado complemento predicativo ou complemento aberto, que corresponde ao complemento infinitivo. Este, por sua vez, possui dois argumentos, o segundo dos quais é a entidade referida por *a escola*. Em (1) e (2), o primeiro argumento do infinitivo, realizado em (3) na posição de sujeito, não é expresso nessa posição. Enquanto o primeiro argumento do verbo encaixado, na primeira análise da Figura 1, é a mesma entidade referida pelo objeto do verbo principal, na segunda análise é o sujeito do verbo principal que expressa esse argumento. O compartilhamento de informações entre o sujeito ou o objeto do verbo principal e o sujeito do verbo encaixado é expresso pelos índices entre colchetes. Por exemplo, o índice 2 em OBJ [2] da primeira análise da Figura 1 remete à estrutura do SUBJ do XCOMP, que recebe índice de igual valor.

(3) The police left the school.
'A polícia deixou a escola.'

A Figura 2 apresenta as análises de (1) e (2) conforme a teoria das Dependências Universais (doravante UD) (MARNEFFE *et al.*, 2021; NIVRE *et al.*, 2020). Salvo o lema verbal sob o nó <root>, os dois gráficos são idênticos, expressando as mesmas relações de dependência do verbo principal que nas análises da Figura 1, ou seja, sujeito (*nsubj*), objeto (*obj*) e complemento predicativo (*xcomp*). Uma diferença essencial entre as representações profundas no formalismo da LFG, na Figura 1, e suas contrapartes rasas em UD, na Figura 2, consiste em que estas últimas não especificam o sujeito do verbo encaixado.

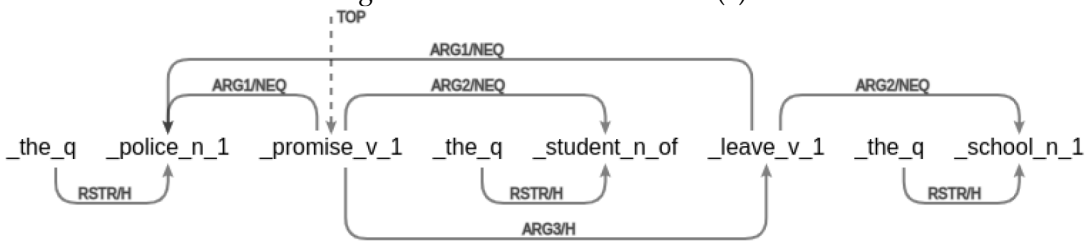
Figura 3: Análise semântica de (1).



Fonte: gerada pela ferramenta delphin-viz[1] a partir da gramática ERG 2018 (UW).

Além da LFG, outra teoria gramatical formal implementada computacionalmente e amplamente utilizada na construção de analisadores sintáticos profundos é a HPSG (SAG; WASOW; BENDER, 2003; MÜLLER, 2020). Uma vantagem importante de analisadores baseados nesse segundo modelo, sob a perspectiva da compreensão textual automática, é que integram a descrição sintática e a descrição semântica num único nível de representação. A Figura 3 e a Figura 4 são representações semânticas geradas pela *English Resource Grammar* (FLICKINGER, 2000) (doravante ERG), aparentemente a maior gramática computacional do inglês nesse formalismo. Conforme a Figura 3, o predicado *persuade* possui três argumentos, indicados por meio de setas rotuladas como ARG1, ARG2 e ARG3. Este último é o predicado *leave* que, por sua vez, tem dois argumentos, sendo o ARG1 destes o ARG2 de persuade. A Figura 4 é quase idêntica, exceto que o ARG1 de *leave* é o ARG1 do predicado *promise*.

Figura 4 – Análise semântica de (2).



Fonte: gerada pela ferramenta delphin-viz a partir da gramática ERG 2018 (UW).

Em (1) e (2), o primeiro argumento do verbo encaixado é determinado por uma função sintática do verbo matriz, numa relação denominada controle. Que função do verbo matriz controla o sujeito do verbo encaixado depende do tipo de verbo matriz. Para que um analisador sintático possa determinar esses sujeitos externos, é preciso

[1] http://delph-in.github.io/delphin-viz/demo/

que disponha de informações a respeito das relações de controle de cada verbo matriz. Isso, por sua vez, exige uma descrição exaustiva da valência verbal e sua codificação num formalismo computacional.

Observe que, em inglês, nenhuma característica formal distingue o controle do sujeito de (2) do controle do objeto de (1). Ambos os verbos governam um sujeito, um objeto direto e um complemento infinitivo. Desse modo, a mera anotação sintática de um corpus contendo sentenças com esses verbos aparentemente não seria suficiente para que um analisador estatístico determinasse o controlador no caso de verbos com essas propriedades de controle não presentes no corpus de treino.

O modelo Dependências Universais Estendidas (SCHUSTER; MANNING, 2016) propõe ampliar as anotações do projeto UD para uma representação sintática mais próxima da semântica. Para tanto, inclui, entre outras anotações adicionais, uma dependência rotulada *nsubj* entre o sujeito do verbo matriz e o verbo encaixado em sentenças como (4). No entanto, essa camada extra de anotação foi aplicada a poucas línguas, não incluindo o português (DROGANOVA; ZEMAN, 2019).

(4) Fred started to laugh.
'Fred começou a rir.'

Na release 2.8 da coleção *Deep Universal Dependencies* (DROGANOVA; ZEMAN, 2019), foram anotados no *UD_Portuguese-Bosque*, *treebank* do português do projeto UD (RADEMAKER *et al.*, 2017), mais de 2500 sujeitos externos (i.e., sujeitos de *xcomps*), dos quais a maioria é controlada pelo sujeito e o restante, pelo objeto do verbo matriz. Em português, contudo, não só o sujeito e o objeto do verbo matriz, mas também um objeto indireto podem controlar o sujeito de um verbo não finito encaixado, como evidencia (5).

(5) A mulher disse ao jardineiro para podar a árvore.

Outra limitação do esquema de anotação do projeto *Deep Universal Dependencies* é que não distingue entre os dois tipos de sujeitos externos exemplificados em (1), (2) e (5), por um lado, e (4), por outro. No primeiro grupo de exemplos, temos estruturas de controle, nas quais o sujeito externo é argumento semântico tanto do predicado expresso pelo verbo matriz quanto do predicado expresso pelo verbo encaixado. Pelo contrário, o sujeito externo de (4) é argumento semântico apenas do predicado do verbo encaixado, configurando uma construção de alçamento.

Em teorias lexicalistas como a LFG e a HPSG, controle e alçamento integram o domínio da valência, uma vez que constituem propriedades da estrutura argumental de determinados itens lexicais, notadamente verbos (POLINSKY, 2013; ABEILLÉ, 2021). Nas três últimas décadas, diversas iniciativas, sob diferentes perspectivas teóricas, voltaram-se ao levantamento das valência verbais do português, na esteira de Fernandes (1987), cuja primeira edição data de 1940. No entanto, esses trabalhos se furtam a uma modelação do controle e do alçamento, os projetos mais recentes se atendo à sentença simples.

Fernandes (1987) subdivide os seus mais de 12000 verbetes com base, primeiramente, em categorias de regência, que perfazem um total de oito, a saber, (i) intransitivo, (ii) transitivo, (iii) relativo, (iv) transitivo relativo (i.e., introduzido por preposição), (v) birrelativo, (vi) transitivo predicativo, (vii) predicativo e (viii) pronominal. A cada uma dessas subdivisões corresponde uma ou mais acepções, acompanhadas de abonações. Por exemplo, o verbo *convencer*, na acepção "obrigar com razões, argumentos (a reconhecer alguma coisa); persuadir" de (1), é transitivo-relativo. Não se indica como o sujeito do infinitivo que pode constituir o complemento relativo deve ser interpretado, trata-se de informação que precisa ser deduzida pelo usuário a partir das acepções dadas ou dos exemplos.

Borba (1991) oferece uma descrição bem mais detalhada de cerca de 6000 dos verbos do dicionário anterior, levando em conta não só aspectos morfossintáticos, mas também semânticos. Destaca-se pelo tratamento dos complementos oracionais, especificando não só a sua forma, mas também, ocasionalmente, relações de correferência entre o sujeito do verbo encaixado e um argumento do verbo matriz. No verbete de *impedir*, por exemplo, Borba (1991) afirma que (7) deriva de (6) pela transposição do sujeito da oração subordinada para complemento do verbo superior, repercutindo a análise transformacional do alçamento do sujeito (POLINSKY, 2013; ABEILLÉ, 2021). O verbete de *ouvir*, porém, carece de explicação dessa natureza, embora o verbo exiba comportamento análogo em (8) e (9). Analogamente, no verbete de *pedir*, consta que esse verbo significa "solicitar licença para" quando a oração principal e a infinitiva introduzida por *para* compartilham o mesmo sujeito, como em (10). No entanto, não explicitam qual seria o sujeito do infinitivo em (11) e (12).

(6) A polícia impediu que os ladrões assaltassem a joalheria.
(7) A polícia impediu os ladrões de assaltarem a joalheria.
(8) Bento ouviu bem que a velha gemia.
(9) [Pe. Lucas] ouvia o vento assobiar cada vez mais forte.
(10) [Telma] ligou direto, () pedindo para falar com Edith.
(11) Você pediu para não contar.
(12) Mamãe, peça ao Luiz para jantar conosco[2].

VerboWeb é um banco de dados sobre propriedades semânticas e sintáticas de verbos, com base na decomposição de predicados. Atualmente com 1418 lemas verbais e 1509 entradas, restringe-se a estruturas argumentais de até três argumentos realizados por sintagmas nominais e preposicionais (CANÇADO; AMARAL; MEIRELLES, 2017; CANÇADO *et al.*, 2021), limitação essa herdada das descrições de

[2] Exemplos de Borba (1991), dos quais (8)-(12) provêm de obras literárias.

diferentes classes verbais no mesmo quadro teórico que o antecederam e que incorporou, como Cançado, Godoy e Amaral (2013).

O projeto em andamento *Dicionário de valências verbais do português brasileiro*, tal como o VerboWeb, limita-se a sentenças simples, adiando para uma etapa futura fenômenos de valência envolvendo sentenças compostas, como os exemplificados em (1), (2), (4) e (5) (PERINI, 2016; PERINI *et al.*, 2019). Nesse dicionário, a valência de um verbo constitui-se de uma ou mais diáteses, concebidas como associações entre constituintes e papéis semânticos (PERINI, 2015). A versão atual abrange 635 verbos, aos quais se atribuem 326 diáteses ao todo. Controle e alçamento propriamente ditos não integram o quadro teórico subjacente, a noção que mais se aproxima desses construtos é a de emparelhamento de papéis semânticos, exemplificada em (13) (PERINI, 2015, p. 189–199). Na diátese de (14) associada a esse exemplo, considerada característica idiossincrática do verbo *achar*, emparelham-se os papéis dos sintagmas nominais complementos, porque entidade qualificada e qualidade implicam-se mutuamente, uma vez que esta predica sobre aquela. Uma desvantagem da representação de (14) em relação à entrada lexical desse verbo proposta por Abeillé (2021, p. 517–519) no quadro da HPSG é quebrar o paralelismo com o predicado de dois lugares de (15), paráfrase de (13), uma vez que implica que o verbo nesse exemplo é um predicado de três lugares. Conforme a abordagem de Abeillé (2021), pelo contrário, *achar* é um verbo de alçamento para objeto que expressa uma relação entre um experienciador (EXP) e um estado de coisas (SOA) tanto em (13) quanto (15), constituindo, portanto, em ambos os casos, um predicado de dois lugares, a diferença entre as duas variantes residindo na quantidade de complementos. Enquanto a segunda variante tem um único complemento, realizado como oração completiva, a primeira tem dois, um objeto direto e um complemento predicativo, cujo sujeito é compartilhado com o objeto direto.

(13) Marta acha Jim um idiota.
(14) VSubj>Agent V NP>*Qualified.thing* NP>*Quality*
(15) Marta acha que Jim é um idiota.

Visando permitir uma anotação sintaticamente profunda de corpora, entre outras aplicações que necessitem de uma compreensão textual mais refinada, iniciamos o desenvolvimento da PorGram, uma nova gramática computacional do português no formalismo HPSG. Cabe destacar que uma gramática computacional de uma língua não tem apenas utilidade prática na construção de softwares de processamento da linguagem. Cumpre, também, um importante papel na testagem de hipóteses linguísticas, constituindo uma forma de documentação de uma língua cuja consistência interna e validade empírica podem ser automaticamente verificadas (BENDER, 2010; BENDER; FLICKINGER; OEPEN, 2011). Diferentemente de sua congênere LXGram (COSTA; BRANCO, 2010), a PorGram é um projeto de software livre e de código aberto. Está ligada ao DELPH-IN Consortium, um movimento de implementação de gramáticas no formalismo da HPSG para várias línguas envolvendo grupos de pesquisa de vários países[3]. Código, documentação, conjuntos de teste, programas auxiliares desenvolvidos pela equipe da PorGram, tudo está disponível numa plataforma de software livre que integra diversas formas de interação social[4].

Neste trabalho, focamos a implementação da valência verbal, abrangendo a realização de argumentos tanto em sentenças simples, quanto em sentenças compostas, incluindo estruturas de controle e de alçamento. Na próxima seção, delineamos os princípios e conceitos fundamentais da teoria da HPSG relevantes para a compreensão do presente trabalho, o sistema *LinGO Grammar Matrix*, utilizado para

[3] https://github.com/delph-in/docs/wiki
[4] https://github.com/LR-POR/PorGram

construir parte substancial da PorGram, e a arquitetura da gramática. Na seção 3, descrevemos, em linhas gerais, o modelo de representação semântica gerada pela PorGram. Na seção 4, tratamos da implementação da valência, mostrando quais fenômenos puderam ser implementados por meio do questionário de customização e quais tiveram de ser codificados manualmente. Na seção 5, apresentamos os resultados da aplicação da PorGram sobre conjuntos de teste. Finalmente, na seção 6, sumarizamos os resultados alcançados e apontamos direções para os próximos passos a serem trilhados no desenvolvimento da gramática.

2 A teoria da HPSG, o sistema *Grammar Matrix* e a PorGram

A HPSG é uma teoria gramatical formal amplamente utilizada na implementação de analisadores sintáticos profundos. É uma teoria lexicalista, codificando parte substancial da informação gramatical no léxico, pelo que a formalização da valência assume enorme relevo. Diferentemente de teoria multiestratais como a LFG, a HPG é monoestratal. Em único nível de representação são codificadas informações sintáticas, fonológicas, semânticas, pragmático-discursivas etc.

A HPSG baseia-se na noção de signo, utilizando estruturas de traços providas de tipos para representar o significante e o significado tanto de palavras e sintagmas quanto de regras morfológicas e sintáticas. A unificação é a operação matemática fundamental desse formalismo. Essa operação permite construir representações mais complexas a partir de representações mais simples, assegurando que as informações sejam compatíveis entre si (FRANCEZ; WINTNER, 2012).

Estruturas de traços organizam as informações sobre as propriedades dos objetos linguísticos sob a forma de matrizes de atributos e valores. Por exemplo, podemos representar as propriedades de um item lexical como *amarela* por meio de

uma matriz com pares de atributos e valores como [CATEGORIA adjetivo], [GÊNERO feminino] e [NÚMERO singular]. Na HPSG, essas matrizes são providas de tipos, os quais desempenham duas funções. A primeira é restringir as estruturas de uma dada gramática, assegurando que contenham apenas atributos apropriados aos objetos que descrevem e restringindo os valores de cada atributo a determinados tipos. Por exemplo, o par [TEMPO presente] é apenas adequado na especificação de uma forma verbal, não de um adjetivo ou substantivo, ao passo que [TEMPO plural] não constitui uma especificação válida, pois *plural* não constitui subtipo do tipo tempo.

A segunda função dos tipos é formular generalizações sobre os objetos modelados, facilitando a descrição de classes, subclasses e instâncias específicas. Por exemplo, nas entradas lexicais dos verbos *convencer*, *forçar*, *proibir* e *impedir*, nas variantes dos exemplos abaixo, não precisamos descrever cada uma de suas propriedades individualmente:

(16) O diretor convenceu os estudantes a deixarem o local.
(17) Os seguranças forçaram os estudantes a deixar o local.
(18) A polícia proibiu os manifestantes de usarem capuzes.
(19) Os guardas impediram os manifestantes de invadir o prédio.

Observemos que (16)-(19) compartilham um conjunto de propriedades: (i) há um verbo principal numa forma finita e outro encaixado no infinitivo; (ii) o verbo principal expressa o tipo básico de evento de cada sentença (por exemplo, (18) descreve uma ação de proibição pela polícia e não o uso de capuzes por manifestantes); (iii) o objeto direto do verbo principal é o sujeito externo do verbo no infinitivo; (iv) o segundo complemento do verbo principal é um sintagma de complementador (CP, do inglês *complementizer phrase*). O tipo de complementador, porém, não é uniforme nas quatro sentenças, permitindo dividir os quadro verbos em dois subgrupos: enquanto convencer e forçar exigem como complementador a preposição *a*, *proibir* e *impedir*

exigem a preposição *de* (GABRIEL; MÜLLER, 2008, p. 39). Por outro lado, esses verbos se enquadram em duas classes diferentes relativamente a um outro critério, qual seja se o sujeito externo do infinitivo exerce papel temático do verbo matriz. Esse é o caso dos verbos de controle do objeto de (16)-(18), mas não de *impedir* em (19), classificado como verbo de alçamento para objeto. A propósito, o termo alçamento (*raising* em inglês) reflete a análise do fenômeno em modelos transformacionais da gramática gerativa, segundo a qual o objeto direto de (19) resulta do movimento de *os manifestantes* da posição de sujeito do verbo encaixado à posição mais elevada de objeto do verbo matriz. Utiliza-se esvaziado dessa acepção em modelos como a LFG e a HPSG, onde não há qualquer tipo de movimento de constituintes (FALK, 2001; SAG; WASOW; BENDER, 2003).

Uma maneira de modelar o comportamento dos verbos de (16)-(19) de forma elegante é codificar em tipos mais abstratos as propriedades gerais do grupo e as propriedades de subgrupos em tipos mais específicos, utilizando o mecanismo de múltipla herança na codificação dos tipos mais específicos. Por exemplo, podemos implementar um tipo *main-verb-lex* para dar conta da propriedade (ii) não somente desses quatro verbos, mas de qualquer outro verbo principal. A estrutura argumental dos verbos de (16)-(18) e o controle do sujeito externo do verbo encaixado pelo objeto direto do verbo principal podem ser codificados no tipo *ditrans-second-arg-control-verb-lex*, que, por sua vez, constitui subtipo de *ditrans-second-arg-control-lex-item*, que modela as propriedades análogas de itens lexicais de outras classes com a mesma aridade. A forma infinitiva do complemento oracional pode ser codificada no tipo *inf-ditrans-second-arg-control-verb-lex*, enquanto a exigência de uma preposição específica pode ser modelada pelos tipos *a-inf-ditrans-second-arg-control-verb-lex* e *de-inf-ditrans-second-arg-control-verb-lex*, exemplificados nas entradas de convencer e proibir da Figura 5. Analogamente, as propriedades de verbos de alçamento para objeto podem ser codificadas num tipo geral *ditrans-second-arg-raising-verb-lex*, propriedades essas

herdadas tanto pela variante de *impedir* em (19) quanto por *fazer* na variante causativa de (20). As particularidades de cada uma dessas variantes são modeladas em subtipos desse tipo geral, conforme as entradas da Figura 5, que determinam que o segundo complemento deve realizar-se como CP infinitivo encabeçado pelo complementador preposicional de, no caso de impedir, ou por um sintagma verbal nu, no caso de fazer.

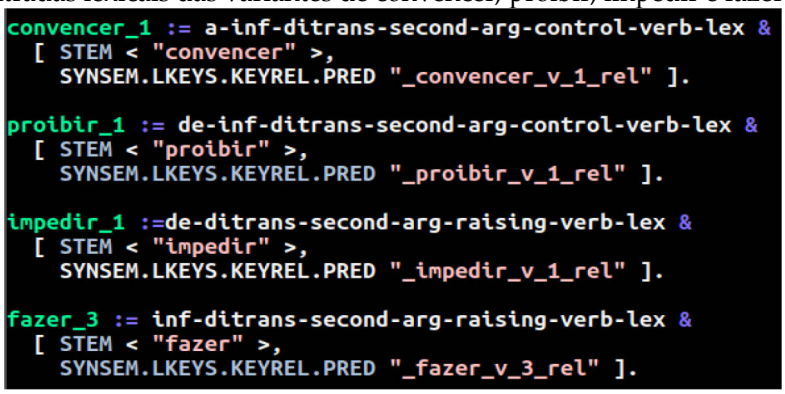
(20) a fumaça fez os agentes recuarem[5]

Como se pode constatar nas entradas lexicais da Figura 5, a organização dos tipos da gramática numa hierarquia com múltipla herança simplifica enormemente a codificação da valência. Na PorGram, como em gramáticas análogas, as entradas lexicais constituem definições de tipos, nas quais as informações se subdividem em quatro categorias. A primeira informação, antes do operador de definição ":=", é o tipo a ser definido, que funciona como identificador da entrada lexical. Por convenção, no caso de verbos, esse identificador constitui-se do lema com o sufixo _n, onde n indica a variante do verbo correspondente à definição do tipo à direita do operador ":=". Na Figura 5, a notação *impedir_1* indica que se trata de variante do verbo impedir distinta, por exemplo, de uma variante transitiva direta *impedir_2*, exemplificada em (21). A definição, por sua vez, constitui-se inicialmente de uma invocação de tipo, que codifica as propriedades da subclasse da qual o item em questão constitui instância, após o operador de unificação "&" seguem as propriedades idiossincráticas do item, que são o lema (STEM)[6] e o predicado semântico, que, convencionalmente, possui o mesmo índice do identificador.

(21) Os guardas impediram a invasão do prédio.

[5] Extraído de voto do TJ-SP de 15/06/2021 (https://tj-sp.jusbrasil.com.br/).
[6] O termo *stem* designa o radical de uma palavra, que, em inglês, corresponde ao lema. Em português, radical e lema não coincidem necessariamente.

Figura 5 – Entradas lexicais das variantes de convencer, proibir, impedir e fazer em (16)-(20).



Fonte: arquivo my-lexicon.tdl da PorGram.

Nessas duas entradas, temos uma amostra da linguagem TDL (do inglês *Type Description Language*), linguagem de descrição utilizada para elaborar a PorGram. Gramáticas especificadas nessa linguagem podem ser compiladas em analisadores sintáticos por meio de diversos sistemas, entre eles o LKB, que é o ambiente de desenvolvimento e testagem de gramáticas computacionais que utilizamos (COPESTAKE, 2002).

O desenvolvimento da PorGram desdobra-se em dois eixos. No primeiro, utilizamos o sistema de customização da *LinGO Grammar Matrix* (BENDER; FLICKINGER; OEPEN, 2002; BENDER *et al.*, 2010). Esse sistema gera o código TDL de um protótipo de gramática a partir de informações fornecidas pelo usuário sobre a estrutura da língua, sem necessidade de qualquer conhecimento da linguagem de descrição. Essas informações são elicitadas por meio de um questionário on-line dividido em 25 seções e submetidas a um sistema de validação, que emite avisos no caso de informações incompletas ou incompatíveis. Quatro seções visam especificar metadados, definir estratégias de toquenização, compilar conjuntos de teste etc. As demais cobrem diferentes aspectos da sintaxe e da morfologia e permitem especificar um número arbitrário de entradas lexicais para as principais classes de palavras. Cada seção cobre um amplo espectro de variação tipológica. Por exemplo, a seção sobre caso

permite especificar 9 padrões diferentes de marcação dos argumentos de verbos intransitivos e transitivos prototípicos, i.e., verbos monovalentes e verbos divalentes com estrutura argumental constituída de agente e paciente. Essa seção permite também especificar casos adicionais, não canônicos, para marcação desses argumentos.

A *Grammar Matrix* resulta de um esforço iniciado há duas décadas. Em sua primeira versão, teve como ponto de partida a ERG e a JACY (SIEGEL; BENDER; BOND, 2016), gramática de ampla cobertura do japonês. Esse núcleo inicial contemplava apenas fenômenos gramaticais universais (DRELLISHAK, 2009). Desde então, foram progressivamente adicionadas bibliotecas para lidar com fenômenos sujeitos a variação tipológica, por exemplo, caso e concordância (DRELLISHAK, 2009), tempo e aspecto (POULSON, 2011), regras morfológicas (GOODMAN, 2013), operações de mudança da valência (CURTIS, 2018), complementos oracionais (ZAMARAEVA; HOWELL; BENDER, 2019) e interrogativas parciais (ZAMARAEVA, 2021).

O segundo eixo de desenvolvimento da PorGram consiste na codificação manual na linguagem TDL, seja adaptando os tipos gerados automaticamente, seja criando novos tipos e entradas lexicais, de modo a analisar fenômenos gramaticais não abrangidos pelo sistema de customização. A Tabela 1 apresenta os componentes do código em TDL da gramática. Os 12 primeiros arquivos foram produzidos pelo sistema de customização, os três últimos foram codificados manualmente.

Os cinco primeiros arquivos não variam de uma língua para outra. Os dois primeiros constituem a espinha dorsal da gramática. Enquanto o segundo define 510 tipos disjuntivos para superclasses de palavras, como *+np*, que engloba substantivos e adposições, o primeiro fornece o arcabouço geral das estruturas gramaticais e semânticas, especificando as regras sintagmáticas e lexicais gerais, as classes de

valência etc. Como as especificações dos arquivos 3-5 não são relevantes no contexto do presente artigo, remetemos o leitor à documentação contida nos próprios arquivos.

Tabela 1 – Componentes em TDL da PorGram.7.

Número	Nome do arquivo	Tipos	Linhas	Palavras	Caracteres
	matrix.tdl	518	1957	8774	78278
	head-types.tdl	501	501	4947	24495
	labels.tdl	37	176	737	6188
	mtr.tdl	22	44	206	1113
	pet.tdl	0	21	48	415
	portuguese.tdl	363	1163	4960	60224
	lrules.tdl	8	8	24	268
	irules.tdl	90	270	540	5629
	lexicon.tdl	242	723	2894	24751
	rules.tdl	15	15	45	548
	roots.tdl	2	10	42	371
	portuguese-pet.tdl	0	22	50	438
	my-portuguese.tdl	179	487	2571	29023
	my-irules.tdl	135	424	1488	13754
	my-lexicon.tdl	198	606	2543	22736

Fonte: elaborada pelos autores.

Os arquivos 6-12 resultam do preenchimento do questionário de customização, refletindo particularidades gramaticais e lexicais do português. O arquivo 6 contém a maior parte das especificações gramaticais, os arquivos 7 e 8 se restringindo às regras lexicais. O arquivo 9 consiste de entradas lexicais.

Vejamos como se articulam as definições de tipos dos arquivos *martrix.tdl*, *portuguese.tdl* e *lexicon.tdl*. Por exemplo, a definição do tipo transitive-lex-item da Figura 6 abstrai da classe de palavra do item e da realização morfossintática dos argumentos, servindo como molde para qualquer item lexical biargumental. Desse

[7] Quantidades de linhas, palavras e caracteres computadas por meio da ferramenta wc do Unix, excluindo comentários e linhas em branco.

modo, pode ser utilizada como base para definir verbos transitivos canônicos tanto de línguas acusativas como o português, o alemão e o japonês quanto de línguas ergativas como o basco e o dirbal. Por outro lado, serve tanto para línguas que marcam os argumentos por meio de casos, como o alemão, quanto por meio de adposições, como o japonês.

Figura 6 – Definição do tipo *transitive-lex-item*.

```
transitive-lex-item := non-local-none-no-hcons & basic-icons-lex-item &
  [ ARG-ST < [ LOCAL [ CAT cat-sat,
                       CONT.HOOK [ INDEX ref-ind & #ind1,
                                   ICONS-KEY.IARG1 #clause ] ] ],
             [ LOCAL [ CAT cat-sat,
                       CONT.HOOK [ INDEX ref-ind & #ind2,
                                   ICONS-KEY.IARG1 #clause ] ] ]>,
    SYNSEM [ LKEYS.KEYREL [ ARG1 #ind1,
                            ARG2 #ind2 ],
             LOCAL.CONT.HOOK.CLAUSE-KEY #clause ] ].
```

Fonte: Imagem de parte do arquivo *matrix.tdl* no LKB.

Extrapolaria o âmbito deste artigo explicar detalhadamente toda a notação utilizada na definição do tipo da Figura 5 e de tipos análogos. É suficiente destacar que especifica a estrutura argumental (ARG-ST) do item como uma lista constituída de dois argumentos dotados de [INDEX ref-ind], ou seja, um índice referencial, próprio de signos nominais, por oposição a signos predicativos (BENDER; FLICKINGER; OEPEN, 2003). O valor do atributo HEAD, que especifica a classe de palavra de um dado núcleo, não é informado. Em vez disso, exige-se, por meio da especificação [LOCAL.CAT cat-sat] que o argumento constitua uma categoria saturada, ou seja, uma projeção máxima.

A partir das respostas que fornecemos no questionário de customização, o sistema gerou os tipos das Figuras 7-9. O primeiro é herdado tanto por verbos transitivos canônicos, como em (22), quanto não canônicos, como em (23), em que o objeto é marcado por uma preposição. Observe que esse tipo especifica a categoria de ambos os argumentos como *+np*, ou seja, preposição ou substantivo. O segundo tipo exige que o núcleo do primeiro argumento seja um substantivo no nominativo e o do

segundo, um substantivo no acusativo. Finalmente, a Figura 9 exibe a definição do tipo *trans-verb-lex*, herdado por verbos transitivos canônicos como *matar*, *amar* e *admirar*. Esses verbos herdam as especificações do tipo da Figura 8 e do tipo *noninh-refl-verb-lex*, que define a classe de verbos que não se submetem à regra lexical de afixação de um pronome reflexivo expletivo, de que trataremos mais adiante.

Figura 7 – Definição do tipo *transitive-verb-lex*.

```
transitive-verb-lex := main-verb-lex & transitive-lex-item &
  [ SYNSEM.LOCAL.CAT.VAL.COMPS < #comps >,
    ARG-ST < [ LOCAL.CAT.HEAD +np ],
             #comps &
             [ LOCAL.CAT cat-sat &
                         [ VAL [ SPR < >,
                                 COMPS < > ],
                           HEAD +np ] ] > ].
```

Fonte: Imagem de parte do arquivo *portuguese.tdl* no LKB.

Figura 8 – Definição do tipo *nom-acc-transitive-verb-lex*.

```
nom-acc-transitive-verb-lex := transitive-verb-lex &
  [ ARG-ST < [ LOCAL.CAT.HEAD noun &
                              [ CASE nom ] ],
             [ LOCAL.CAT.HEAD noun &
                              [ CASE acc ] ] >,
    SYNSEM.LOCAL.CAT.VAL [ SUBJ < [ LOCAL.CAT.HEAD.CASE-MARKED + ] >,
                           COMPS < [ LOCAL.CAT.HEAD.CASE-MARKED + ] > ] ].
```

Fonte: Imagem de parte do arquivo *portuguese.tdl* no LKB.

Figura 9 – Tipo *trans-verb-lex* e entradas lexicais com esse tipo.

```
trans-verb-lex := noninh-refl-verb-lex & nom-acc-transitive-verb-lex.
U:---  portuguese.tdl   17% L366   (TDL)

matar := trans-verb-lex &
  [ STEM < "matar" >,
    SYNSEM.LKEYS.KEYREL.PRED "_matar_v_rel" ].

amar := trans-verb-lex &
  [ STEM < "amar" >,
    SYNSEM.LKEYS.KEYREL.PRED "_amar_v_rel" ].

admirar := trans-verb-lex &
  [ STEM < "admirar" >,
    SYNSEM.LKEYS.KEYREL.PRED "_admirar_v_rel" ].

U:---  lexicon.tdl   36% L370   (TDL)
```

Fonte: Imagem de parte dos arquivos *portuguese.tdl* e *lexicon.tdl* no LKB.

(22) O gato matou uma ratazana.

(23) O cachorro obedece ao menino.

Os três últimos arquivos da Tabela 1 resultam de codificação manual. Alencar e Rademaker (2021, submetido à publicação) e Nunes, Rademaker e Alencar (2021) tratam das modificações manuais no terreno da morfologia flexional. Na seção 4, descrevemos as alterações em tipos gerados a partir do questionário de customização e os tipos completamente novos criados para suplantar as limitações do sistema no domínio da valência verbal.

3 O modelo da semântica de recursão mínima

Dentre os benefícios da análise sintática profunda, do tipo produzido por uma gramática no formalismo da HPSG, está a capacidade de produção de representações semânticas detalhadas, em paralelo e de forma composicional às análises sintáticas. Tais representações semânticas objetivam capturar o significado dos enunciados de forma canônica, abstraindo das diferentes variações sintáticas possíveis na linguagem natural. Considere as sentenças abaixo, todas expressam essencialmente o mesmo evento e os mesmos participantes, realizados sintaticamente de diferentes formas.

(24) Kim gave Sandy the green book.
Kim dar:PST;3SG Sandy o verde livro
'Kim deu o livro verde para Sandy.'
(25) Kim gave the green book to Sandy.
Kim dar:PST;3SG o verde livro para Sandy
'Kim deu o livro verde para Sandy.'
(26) The green book was given to Sandy by Kim.
o verde livro be:PST;3SG dar:PTCP para Sandy por Kim
'O livro verde foi dado para Sandy por Kim.'
(27) Sandy was given the green book by Kim.
Sandy be:PST;3SG dar:PTCP o verde livro por Kim

'Sandy ganhou o livro verde de Kim.'
(28) The green book, Kim gave Sandy.
o verde livro Kim dar:PST;3SG Sandy
'O livro verde Kim deu para Sandy.'

Para algumas aplicações de processamento automático de linguagem como extração de informações, inferência textual, resolução de perguntas, sumarização etc., é conveniente reconhecer que todas as sentenças expressam exatamente os mesmos eventos e que, a despeito de exercerem funções sintáticas distintas, o papel semântico das entidades envolvidas nos eventos é o mesmo. A lógica de primeira ordem (LPO) tem sido utilizada como linguagem de representação semântica, permitindo expressar a interpretação comum a diferentes construções sintáticas, como em (24)-(28). Por outro lado, permite representar as diferentes interpretações de uma mesma análise sintática de uma sentença ambígua como (29). Essa sentença possui duas possíveis interpretações em LPO, a saber (30a) e (30b). Note-se que a diferença de cada interpretação está principalmente na ordem dos quantificadores.

(29) All dogs chased a cat.
todos cães perseguir:PST;3PL um gato
'Todos os cães perseguiram um gato.'
(30) a. $\forall x\ (dog(x) \rightarrow \exists y(cat(y) \wedge chase(x, y)))$
b. $\exists y\ (cat(y) \wedge \forall x(dog(x) \rightarrow chase(x, y)))$
c. $\forall x\ dog(x) : \exists y\ cat(y) : chase(x, y)$
d. $\exists y\ cat(y) : \forall x\ dog(x) : chase(x, y)$

A representação em LPO, contudo, tem algumas limitações. A principal é não preservar a composicionalidade da estrutura sintática. Os dois sintagmas nominais quantificados de (29) não têm uma tradução para LPO independente, pois o sintagma nominal *all dogs* não pode ser associado a nenhuma subfórmula de (30a) ou (30b) isoladamente. Além disso, nas línguas naturais, uma variedade de outras expressões

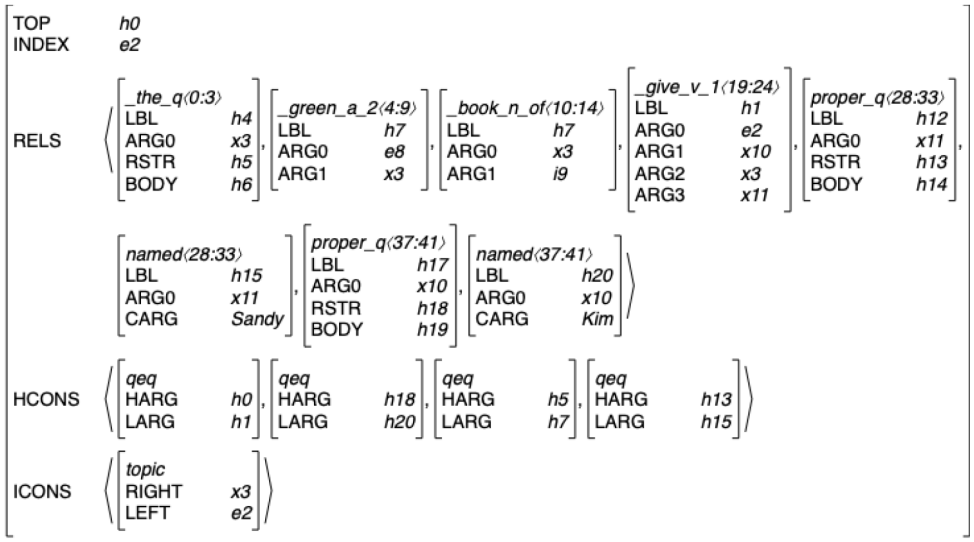
podem ser vistas como expressões quantificadoras, por exemplo, no inglês, *most* em (31), não representadas trivialmente com os quantificadores de LPO.

(31) Most dogs chased a cat.
a_maioria_dos cães perseguir:PST;3PL um gato
'A maioria dos cães perseguiu um gato.'

O que desejamos de uma representação semântica é um tratamento uniforme e composicional para expressões quantificadoras, incluindo as expressões não expressáveis com os quantificadores usuais de LPO. O uso dos quantificadores generalizados (WESTERSTÅHL, 2019), na forma de $Q\ x\ \alpha : \beta$, onde Q é um quantificador, x uma variável e α e β fórmulas em LPO, como em (30c,30d), constitui um passo nessa direção, oferecendo uma representação relacional dos quantificadores por meio de relações binárias entre subconjuntos do domínio. A expressão *all dogs*, por exemplo, denota o conjunto de todos os subconjuntos do domínio do qual todo cachorro é membro.

Finalmente, considere (32). A natureza binária da conjunção em LPO leva a uma ambiguidade espúria na representação, porque os possíveis agrupamentos (33a,33b) são irrelevantes para as condições de verdade do sintagma.

Figura 10 – MRS de (24)-(28).



Fonte: gerada pela ferramenta delphin-viz a partir da gramática ERG 2018 (UW).

(32) a furious white cat
um furioso branco gato
'um gato branco furioso'

(33) a. $\exists y(furious(x) \wedge (white(x) \wedge cat(y)))$
b. $\exists y((furious(x) \wedge white(x)) \wedge cat(y))$

As gramáticas produzidas pelo sistema *LinGO Grammar Matrix* adotam o formalismo da Semântica de Recursão Mínima (doravante MRS, do inglês *Minimal Recursion Semantics*) para representação semântica (COPESTAKE; LASCARIDES; FLICKINGER, 2001; BENDER; FLICKINGER; OEPEN, 2003; COPESTAKE *et al.*, 2005). A MRS não constitui uma teoria semântica, mas "uma linguagem de metanível para descrever estruturas semânticas em alguma linguagem de objeto subjacente" (COPESTAKE *et al.*, 2005, p. 282–283). Caracteriza-se por uma estrutura plana de predicações e por permitir subespecificação, isto é, que distinções semânticas permaneçam não resolvidas, de modo a permitir um raciocínio monotônico sobre tais representações semânticas parciais. A Figura 10 exibe a representação em MRS de (24)-(28).

A representação da Figura 10 é composta por 5 elementos principais: TOP, INDEX, RELS, HCONS e ICONS. No componente RELS temos a estrutura de predicados e argumentos expressa sob a forma de uma coleção de predicações (ou relações) n-árias vinculadas por variáveis (tipadas). Esta estrutura é caracterizada pela expressão "quem fez o que com quem" Apenas os signos linguísticos que contribuem com a semântica são representados, o verbo auxiliar da passiva, por exemplo, não contribuiu com nenhum predicado. Na figura, embora *book* (x3) seja o sujeito da passiva, corresponde ao ARG2 (elemento mais diretamente afetado pelo evento) do evento (e2) introduzido pelo predicado *_give_v_1*, cujo ARG1 (o causador volitivo de um evento) corresponde ao nome *Kim* (x9).

As predicações podem ser de superfície ou abstratas. Predicados de superfície seguem uma convenção de nomenclatura em que o símbolo é composto por três componentes separados pelo símbolo (_; U+005F): lema, POS (classe de palavra) e sentido. Por convenção, são marcados por um _ inicial e o campo de sentido é opcional. São introduzidos exclusivamente por entradas lexicais, cuja ortografia é uma forma (possivelmente flexionada) do campo lema no predicado. Na Figura 10, o predicado *_give_v_1* é um predicado de superfície.

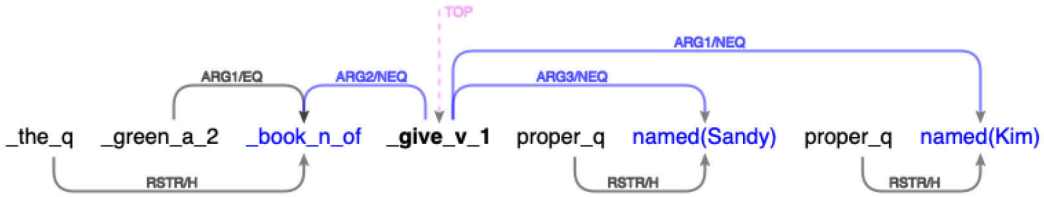
Predicados abstratos, por outro lado, são usados para representar a contribuição semântica de construções gramaticais ou entradas lexicais mais especializadas (como quantificadores implícitos, composição, nominalizações ou nomes próprios, no exemplo, *named*). Cada predicação é composta por: 1) uma relação ou predicado (por exemplo, _leave_v_1); 2) um rótulo LBL, cujo valor é sempre uma variável de escopo e servirá para agrupar predicações (*green* e *book* no exemplo); 3) um ARG0 com sua variável intrínseca; 3) ARG1. . . ARG4 correspondendo aos demais argumentos da relação, que podem ser preenchidos por variáveis sobre eventos (e), indivíduos (x), genéricas ou não realizadas (i ou u) ou de escopo (h, para predicados que recebem outras predicações como argumento).

Na representação da Figura 10, observamos três quantificadores generalizados que introduzem variáveis de indivíduos em seus ARG0 (instanciados pelo determinante *the* e dois determinantes implícitos para os nomes próprios). Estes quantificadores contêm dois argumentos RSTR (restrição) e BODY (escopo de aplicação).

Os componentes TOP e INDEX contêm como valores as variáveis h0 e e2, respectivamente. Em TOP temos a variável associada ao escopo mais externo ou abrangente da sentença e em INDEX o indivíduo principal da estrutura de predicados e argumentos, neste caso, o evento principal da sentença. Não iremos aqui detalhar o componente ICONS. Finalmente, o componente HCONS impõe restrições para os possíveis alinhamentos válidos para os quantificadores em uma possível resolução das subespecificações de escopo e ordem dos quantificadores. Por exemplo, o escopo h7, rótulo da predicação *book*, deve necessariamente estar embutido no escopo introduzido pela restrição (RSTR h5) do quantificador *the* que introduz a variável de indivíduo (x3), afinal, book predica sobre esta variável.

O formato de representação DMRS (COPESTAKE, 2009), exemplificado na Figura 11, constitui uma alternativa ao formato MRS na forma de um grafo livre de variáveis. De forma geral, as predicações são transformadas em nós e os argumentos, em arcos. As restrições de escopo (HCONS) são codificadas nos rótulos dos arcos, combinadas com os argumentos dos predicados (EQ para predicados com mesmo rótulo, no mesmo escopo, e NEQ para predicados em diferentes escopos) ou em arcos separados (RSTR/H).

Figura 11 – DMRS de (24)-(28).



Fonte: gerada pela ferramenta delphin-viz a partir da gramática ERG 2018 (UW).

4 Implementação da valência verbal na PorGram

A realização morfossintática de argumentos constitui um dos domínios gramaticais que mais variam de uma língua a outra. À diversidade tipológica soma-se uma grande variedade de padrões dentro de línguas específicas. O sistema de customização abrange um amplo leque de padrões em uma amostra bastante representativa de línguas das mais variadas famílias. Dada a complexidade dos fenômenos envolvidos, contudo, não poderia almejar a exaustividade de cobertura. Desse modo, a fim de tornar tratável, no âmbito do questionário de customização, a implementação das diferentes bibliotecas que interagem no tratamento computacional de sentenças com esses fenômenos, Drellishak (2009) limitou-se à realização dos argumentos de verbos intransitivos e de verbos transitivos como sintagmas nominais ou adposicionais. Verbos divalentes de alçamento para sujeito, cujo complemento é um verbo ou um sintagma verbal (34)-(36), classificados na *Grammar Matrix* como auxiliares, foram implementados posteriormente (POULSON, 2011).

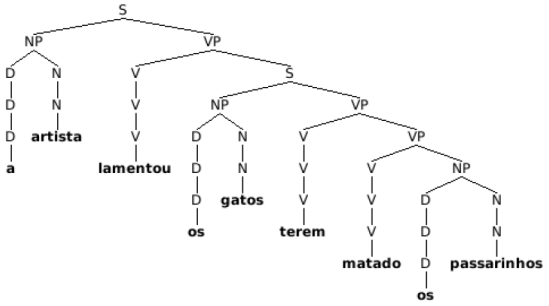
(34) Ele estava dormindo.
(35) Ele vai viajar.
(36) Ele tinha partido.

Analogamente, as biblioteca seguintes que permitiram a modelagem de complementos oracionais se restringiram a verbos divalentes (ZAMARAEVA; HOWELL; BENDER, 2019; ZAMARAEVA, 2021). Desse modo, pudemos, por meio do questionário, implementar sentenças como (37)-(42) com verbos regendo

complementos oracionais que expressam tanto proposições (37)-(39), quanto perguntas globais (41) e parciais (42). As Figuras 12 e 13 exemplificam os dois tipos de análise sintática gerados pela PorGram para as orações completivas de (37)-(41). Não obstante as discrepâncias estruturais entre as análises dessas duas figuras, ambos os exemplos compartilham a mesma representação semântica dependencial da Figura 14.

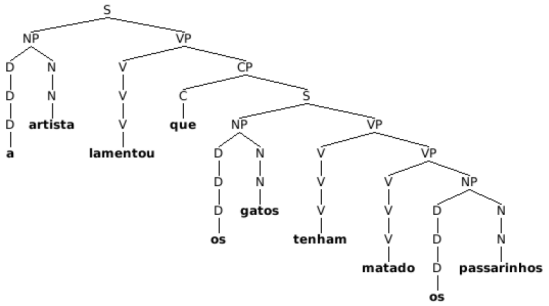
(37) Ela disse que ele estava dormindo.
(38) A Maria viu que as amigas choravam.
(39) A artista lamentou os gatos terem matado os passarinhos.
(40) A artista lamentou que os gatos tenham matado os passarinhos.
(41) Perguntei se ela estava triste.
(42) Ela perguntou quem tinha gritado.

Figura 12 – Árvore de (39).



Fonte: gerada pelo LKB a partir da PorGram.

Figura 13 – Árvore de (40).

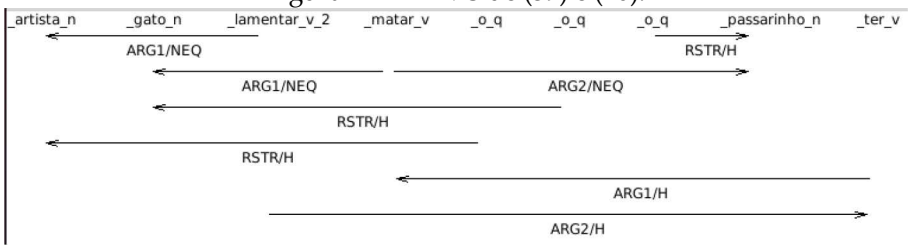


Fonte: gerada pelo LKB a partir da PorGram.

O questionário não contempla, portanto, sentenças com verbos trivalentes (43)-(47). Observe que, nesse último exemplo, temos um verbo de alçamento para objeto: o

sintagma os alunos ocupa a posição de objeto do verbo superior, mas constitui argumento semântico apenas do verbo encaixado (POLINSKY, 2013; ABEILLÉ, 2021).

Figura 14 – DRMS de (39) e (40).



Fonte: gerada pelo LKB a partir da PorGram.

(43) Ele doou uma bicicleta ao estudante.
(44) Ela contou ao estudante que o gato tinha morrido.
(45) Perguntei ao estudante se ele tinha dormido.
(46) Ela perguntou ao artista que gato o cachorro perseguiu.
(47) Ela fez os alunos chorar(em).

Uma outra limitação é que não permite a modelagem de estruturas de controle, como as exemplificadas abaixo:

(48) O artista tentou dormir.
(49) Ela prometeu ao pai estudar.
(50) Ela proibiu o artista de cantar.

Além disso, não oferece suporte para a implementação de verbos de alçamento para sujeito que regem complemento introduzido por complementador, exemplificados em (51)-(54), limitando-se àqueles que regem verbos ou sintagmas verbais nus como em (34)-(36).

(51) Ela tem que dormir.
(52) Temos de matar aquela ratazana.
(53) A estudante começou a sorrir.

(54) O cachorro tinha parado de latir.

Em que pesem essas limitações, o arquivo *matrix.tdl* inclui tipos abstratos para diversas outras classes valenciais, ainda não explorados pelo questionário de customização, que pudemos utilizar na codificação manual em TDL dos tipos necessários à implementação de (43)-(54), entre diversas outras construções.

A seguir, descrevemos, primeiro, as classes de valência implementadas por meio do questionário. Em seguida, tratamos das classes codificadas manualmente.

Na *LinGO Grammar Matrix*, os verbos se classificam em dois grandes grupos: verbos principais, subtipos de *main-verb-lex*, e verbos auxiliares, subtipos de *aux-lex*. Distinguem-se, primeiramente, pelos valores "+" e "-" do atributo AUX. Na PorGram, os auxiliares são verbos de alçamento para sujeito, constituindo subtipos de *subj-raise-aux*. Verbos de alçamento para sujeito, como em (55)-(59), têm como único argumento um VP não finito, não atribuindo, portanto, papel temático ao seu sujeito, que constitui argumento semântico do verbo encaixado.

(55) O gato tinha matado uma ratazana.
(56) O gato está perseguindo a ratazana.
(57) Ela precisa consertar o carro.
(58) Qualquer gato pode consumir esta ração.
(59) O artista começou a pintar o retrato.

Verbos de alçamento para sujeito distinguem-se de verbos de controle do sujeito, que ocorrem em construções superficialmente análogas, como em (60). Nesse caso, o sujeito da sentença é argumento semântico tanto do verbo matriz quanto do verbo encaixado. A diferença entre os dois tipos de verbos evidencia-se no contraste entre a gramaticalidade de (61) e (62), por um lado, e a agramaticalidade de (63), por outro (SAG; WASOW; BENDER, 2003, p. 376).

(60) O gato tentou agarrar a bola.
(61) O carro precisa ser consertado.
(62) O retrato começou a ser pintado pelo artista.
(63) *A bola tentou ser agarrada pelo gato.

Na *Grammar Matrix*, verbos auxiliares tanto podem ser dotados quanto desprovidos de um predicado próprio. Em português, integram o primeiro tipo, entre outros, verbos modais como *poder*, *dever* etc. e aspectuais como *começar* ou *parar* etc. No segundo tipo enquadram-se os auxiliares dos tempos compostos (55) e de locuções aspectuais (56). A existência ou não de dois predicados verbais em construções de verbo matriz e verbo encaixado reflete-se na possibilidade ou impossibilidade de modificação adverbial do verbo encaixado independentemente do verbo matriz, comparem-se os dois grupos de exemplos (64)-(69) e (70)-(73). No segundo grupo, o auxiliar não possui um predicado próprio que possa ser modificado pelo advérbio, contribuindo apenas com especificações de tempo, modo, pessoa e número para o significado da sentença, desempenhando, portanto, papel análogo ao da flexão verbal.

(64) Ele pode não ter chegado.
(65) Ele não pode ter chegado.
(66) O bebê parou de não querer comer.
(67) O bebê não parou de querer comer.
(68) Ela precisa muito trabalhar.
(69) Ela precisa trabalhar muito.
(70) Ele não tinha dormido.
(71) *Ele tinha não dormido.
(72) Ela não está trabalhando.
(73) *Ela está não trabalhando.

Numa versão anterior da PorGram, implementamos, por meio do questionário de customização, os dois tipos de auxiliares. No entanto, ao aplicar a gramática ao conjunto de teste, constatamos que sentenças como (74), com mais de um auxiliar,

sendo um deles desprovido de predicado, não eram analisadas, o que não ocorria com sentenças como (75), em que todos os auxiliares possuem um predicado.

(74) Ele está tentando trabalhar.
(75) Ele continua tentando trabalhar.

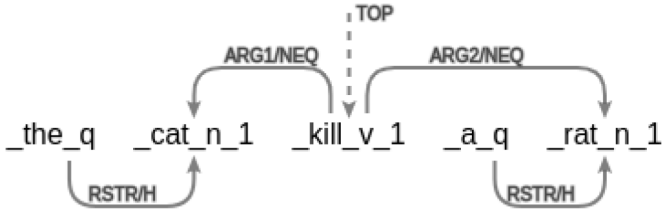
Essa assimetria decorre de uma limitação da *Grammar Matrix*, que não contempla a possibilidade de um auxiliar sem predicado tomar como complemento um VP nucleado por outro auxiliar (ZAMARAEVA, 2021, p. 320). Resolvemos provisoriamente esse problema, tal como Zamaraeva (2021) no seu fragmento de gramática da língua apinajé, atribuindo um predicado fictício (*dummy predicate*) a todos os auxiliares, indistintamente. Desse modo, o verbo auxiliar ter dos tempos compostos contribui com o predicado *_ter_v*, o verbo estar do progressivo com o predicado *_estar_v* etc. Reconhecemos que essa solução não é a ideal, uma vez que produz representações semânticas que discrepam estruturalmente das representações análogas produzidas pela ERG para exemplos do tipo de (76), sem que haja uma motivação linguística para tanto, como podemos constatar comparando as Figuras 15-18. No entanto, esses predicados fictícios podem ser assinalados de algum modo para que sejam ignorados no processamento semântico.

(77) the cat had killed a rat
o gato ter:PST;3SG matar:PTCP uma ratazana
o gato tinha matado uma ratazana
(78) the cat tried to catch the ball
o gato tentar:PST;3SG a agarrar:INF a bola
o gato tentou agarrar a bola

Na análise da Figura 15 para a tradução em inglês de (55), o auxiliar não contribui com nenhuma relação. No topo do gráfico dependencial, temos apenas o nó correspondente ao verbo principal, ou seja, *_kill_v_1*, de que constituem dependentes

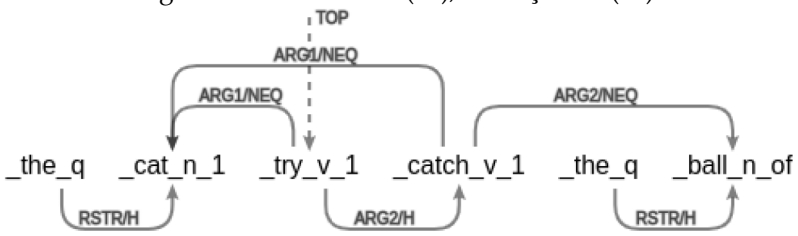
os nós correspondentes ao sujeito e objeto direto da sentença. Pelo contrário, o verbo matriz na análise da tradução de (60) na Figura 16 contribui com o predicado que ocupa o nó mais alto do gráfico dependencial, do qual o predicado correspondente ao verbo encaixado constitui o segundo argumento. Em ambas as análises da PorGram nas Figuras 17 e 18, o predicado do verbo matriz constitui o nó mais alto, quando somente deveria fazê-lo no segundo gráfico. Essa é uma deficiência que pretendemos sanar no futuro.

Figura 15 – Análise de (76), tradução de (55).



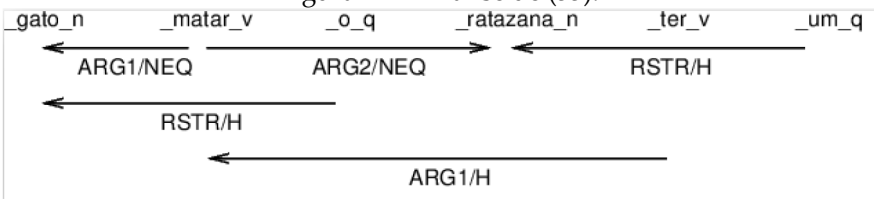
Fonte: gerada pela ferramenta delphin-viz a partir da gramática ERG 2018 (UW).

Figura 16 – Análise de (77), tradução de (60).



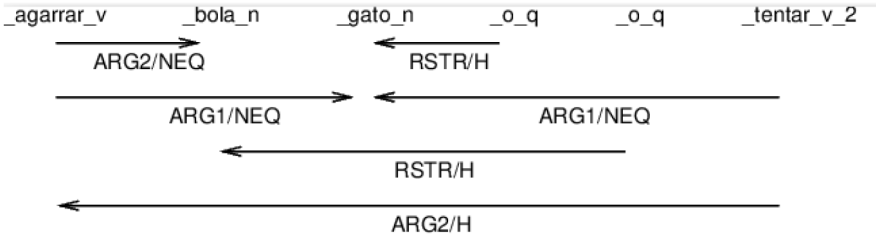
Fonte: gerada pela ferramenta delphin-viz a partir da gramática ERG 2018 (UW).

Figura 17 – Análise de (55).



Fonte: gerada pelo LKB a partir da PorGram.

Figura 18 – Análise de (60).



Fonte: gerada pelo LKB a partir da PorGram.

Os verbos principais implementados por meio do questionário dividem-se em dois grandes grupos: (i) verbos intransitivos e transitivos canônicos, i.e., verbos cujo segundo argumento realiza-se como um sintagma nominal ou clítico no acusativo; (ii) verbos com complemento oracional.

Utilizando o componente morfológico do sistema de customização, implementamos os verbos tradicionalmente chamados pronominais, incluindo tanto aqueles inerentemente reflexivos (78) quanto os facultativamente reflexivos (79). Por meio de uma regra lexical, esses verbos são sufixados com um reflexivo de forma obrigatória ou facultativa, dependendo do tipo do verbo. Essa regra não se aplica a verbos como *desaparecer* que não admitem o reflexivo expletivo (80).

(78) O artista queixou-se.
(79) A porta abriu(-se).
(80) *O gato desapareceu-se.

Quadro 1 – Estratégias de complementação oracional.

Número	Força iloc.	Compl.	Forma	Modo	Exemplos
1	proposição	que	finita	indicativo	(37)
2	proposição	(em|com|para) que	finita	subjuntivo	(84)-(88)
3	proposição	Ø	infinitivo flexionado	__	(39)
4	interrogação	se ou QU-	finita	indicativo	(41) (42)

Fonte: elaborado pelos autores.

O questionário permite especificar diferentes estratégias de complementação oracional verbal, área em que a língua portuguesa exibe grande variedade de padrões (MATEUS *et al.*, 1989, p. 265–279). O Quadro 1 sintetiza as estratégias implementadas por essa via. A segunda coluna especifica a força ilocucionária da oração completiva. A terceira especifica a forma do complementador, que pode ser foneticamente vazio ou consistir numa conjunção precedida ou não de uma preposição (*em*, *com* ou *para*) ou num elemento interrogativo QU-. A forma do verbo, na quarta coluna, pode ser finita ou o infinitivo flexionado. A estratégia 2 representa, na verdade, a consolidação de diversas estratégias, dependendo de cada verbo individual se o complementador pode ser precedido ou não por uma determinada preposição. Por meio da seção do léxico do questionário, utilizando essas estratégias, construímos tipos e entradas lexicais para todas as diáteses descritas por Mateus *et al.* (1989) envolvendo verbos divalentes cujo complemento é uma oração completiva, com as seguintes exceções: (i) verbos de controle; (ii) estruturas com inversão do sujeito (81), (iii) nominalizações (82) e (iv) construções com composição de predicados (83). Implementamos, também, a variante da construção causativa de (87) bem como a construção de (88), corriqueira em linguagem padrão, embora condenada por alguns puristas (MATOS, 2008; ARRAIS, 2017).

(81) Os professores acreditam terem os Centros recebido verba[8].
(82) As alunas lamentam o terem partido a jarra.
(83) Eu mandei escrever a carta aos alunos.
(84) Eu mandei que os alunos escrevessem a carta.
(85) O estudante insistiu em que a artista matasse a ratazana.
(86) O ruído fez que a criança chorasse.
(87) O estudante fez com que os cachorros perseguissem a ratazana.
(88) O estudante pediu para que a artista matasse a ratazana.

[8] Exemplos (81)-(84) extraídos de Mateus *et al*. (1989, p. 272-275).

Embora não descartemos que seja possível implementar as combinações de preposição e complementador de (85)-(88) de forma composicional por meio do questionário, optamos, na atual fase da PorGram, como mais simples, por codificar essas combinações sob a forma de locuções conjuncionais, ou seja, expressões de múltiplas palavras. No questionário, basta, para tanto, especificar *em que*, *com que*, *para que* etc. como lemas dos complementadores de diferentes estratégias de complementação oracional. Decerto se sacrifica, com isso, uma generalização linguística importante: verbos que licenciam uma oração completiva encabeçada por uma preposição via de regra também admitem um objeto preposicionado ou infinitivo com a mesma preposição, compare (89) e (90). No entanto, (91) e (92) mostram que há exceções. Por outro lado, (90) é uma estrutura de controle, não passível de implementação por meio do sistema de customização.

(89) Ela insistiu na viagem.
(90) Ela insistiu em viajar.
(91) Ela fez com que o filho viajasse.
(92) *Ela fez com a viagem do filho.

Por meio da codificação manual, recorrendo aos tipos abstratos definidos no arquivo *matrix.tdl*, implementamos entradas lexicais para todos os demais tipos de verbos que regem orações completivas descritos por Mateus *et al.* (1989), incluindo verbos de controle do sujeito (48) (49), do objeto direto (50) ou do objeto indireto (5) e verbos trivalentes com complementos realizados tanto por sintagmas nominais e preposicionais (43) quanto por orações (44), ressalvadas as exceções (ii)-(iv) mencionadas acima.

Também reformulamos manualmente as definições dos tipos *ind-cl-verb-lex* e *subj-cl-verb-lex*, referentes a verbos divalentes que regem orações completivas introduzidas pelo complementador que e exigem que o verbo encaixado esteja,

respectivamente, no modo indicativo e no subjuntivo. As definições geradas pelo sistema de customização estavam hipergerando, aceitando tanto sentenças gramaticais quanto suas contrapartes agramaticais, como em (93)-(96). A razão disso é que as estratégias de complementação oracional geradas automaticamente pelo sistema não permitem especificar o modo do verbo encaixado como uma propriedade lexical do verbo superior. Em vez disso, o verbo superior especifica um complementador que, por sua vez, determina o modo do seu complemento, ou seja, a oração encabeçada pelo verbo encaixado. O problema, nesse caso, é que um mesmo complementador licencia dois modos distintos, dependendo do verbo. Na reformulação manual, o verbo superior especifica o modo do verbo encaixado, permitindo que a gramática analise apenas os exemplos gramaticais de (93)-(96).

(93) Ela quer que ele durma.
(94) Ele viu que ela dormia.
(95) *Ela quer que ele dorme.
(96) *Ele viu que ela dormisse.

Implementamos, igualmente, de forma manual, verbos de alçamento para sujeito cujo complemento é encabeçado por um complementador, de modo a analisar sentenças com verbos modais e aspectuais como em (51)-(54). Para tanto, criamos tipos não só para esses verbos, mas também para os respectivos complementadores, os quais, com exceção de que com infinitivo do modal ter (variante mais informal da construção *ter de* + infinitivo), derivam de preposições. Como cada verbo exige um complementador específico, introduzimos o atributo COMP-FORM, cujos possíveis valores são tipos correspondentes às formas dessas preposições. Para analisar exemplos com o verbo ter modal de (51) e (52), criamos o tipo disjuntivo *que+de_comp* como supertipo de *que_comp* e *de_comp*. O verbo ter modal exige que a forma do

complementador do seu complemento seja esse supertipo, ao passo que um verbo como o aspectual parar *exige de_comp*.

O questionário possibilita incluir casos não canônicos para marcar argumentos nucleares, fenômeno comumente denominado *quirky* case na literatura anglofônica, como exemplificou Drellishak (2009) com um fragmento do alemão. Nessa língua, o objeto de verbos transitivos é canonicamente marcado com acusativo. No entanto, alguns verbos, como *helfen* 'ajudar', exigem um objeto no dativo. Drellishak (2009), porém, não implementou nenhum fragmento com a marcação não canônica por meio de adposições, também muito comum em alemão, em construções estruturalmente análogas a (97) e (98).

(97) Aquela artista depende de mim.
(98) O cachorro não obedece a mim.

Aparentemente, a implementação de sentenças como (97) e (98) por meio do questionário de customização é ou impossível ou excessivamente complexa. O sistema permite implementar tanto preposições marcadores de caso, i.e., categorias que constituem núcleos funcionais, semanticamente vazias, exercendo um papel meramente estrutural, quanto preposições que funcionam como núcleos lexicais, também chamadas preposições verdadeiras ou plenas (ZARING, 1991; MIOTO; SILVA; LOPES, 2005). Drellishak (2009) testou essa funcionalidade por meio de línguas em que todos os casos nucleares são realizados por adposições e línguas em que apenas um dos casos tem essa realização. No entanto, (97) e (98) não se enquadram em nenhuma dessas situações, uma vez que se trata de marcação não canônica.

A análise dos verbos (97) e (98) como verbos que regem caso conforma-se a análises de línguas como o francês, segundo as quais as preposições à e de marcam, respectivamente, os casos dativo e genitivo, exercendo função estruturalmente

análoga às flexões de caso correspondentes do latim (CARLIER; GOYENS; LAMIROY, 2013). A dificuldade da implementação desse tipo de análise via questionário decorre de que o caso expresso por uma adposição marcadora de caso é atribuído, também, ao seu complemento, como podemos constatar na Figura 19, onde o valor *#case* tanto do atributo *CASE* do núcleo adposicional quanto do núcleo nominal complemento indica identidade de valores. Com base nessa definição, portanto, se analisarmos a preposição *de* no exemplo (97) como marca de genitivo, o pronome *mim* será marcado igualmente com genitivo. Em (98), porém, esse mesmo pronome deverá estar no dativo. Essa duplicidade de marcação não nos parece razoável. Em vez disso, parece-nos mais sensato atribuir uniformemente o caso oblíquo à forma pronominal *mim* e às formas pronominais análogas em todas as construções em que funcionam como objeto de preposição (MIOTO; SILVA; LOPES, 2005), conforme o Quadro 2.

Quadro 2 – Pronomes pessoais no singular.

Pessoa	Nominativo	Acusativo	Dativo	Oblíquo
1	eu	me	me	mim
2	tu	te	te	ti
3	ele/ela	o/a	lhe	ele/ela

Fonte: elaborado pelos autores.

Desse modo, preferimos implementar manualmente as preposições marcadoras de caso e os verbos que regem complementos preposicionados. Para tanto, reformulamos a definição da Figura 19 como na Figura 20, onde o valor de CASE do núcleo adposicional e de seu complemento é o tipo case, pelo que não são necessariamente idênticos. Com base nesse tipo, criamos o subtipo *obl-case-marking-adp-lex*, cuja definição se encontra na parte inferior da Figura 20. As especificações desse subtipo, por sua vez, são herdadas pelas preposições marcadoras de caso, como veremos mais adiante.

Figura 19 – Tipo das adposições marcadoras de caso gerado pela *Grammar Matrix*.

```
case-marking-adp-lex := non-local-none-lex-item & raise-sem-lex-item &
  [ SYNSEM.LOCAL.CAT [ HEAD adp &
                             [ CASE #case,
                               MOD < > ],
                       VAL [ SPR < >,
                             SUBJ < >,
                             COMPS < #comps >,
                             SPEC < > ] ],
    ARG-ST < #comps &
             [ LOCAL.CAT [ VAL.SPR < >,
                           HEAD noun &
                                [ CASE #case,
                                  CASE-MARKED - ] ] ] > ].
```

Fonte: imagem de parte do arquivo *portuguese.tdl* no LKB.

Figura 20 – Codificação manual do tipo das adposições marcadoras de caso.

```
case-marking-adp-lex := non-local-none-lex-item & raise-sem-lex-item &
  [ SYNSEM.LOCAL.CAT [ HEAD adp &
                             [ CASE case,
                               MOD < > ],
                       VAL [ SPR < >,
                             SUBJ < >,
                             COMPS < #comps >,
                             SPEC < > ] ],
    ARG-ST < #comps &
             [ LOCAL.CAT [ VAL.SPR < >,
                           HEAD noun &
                                [ CASE case ] ] ] > ].

obl-case-marking-adp-lex := case-marking-adp-lex & [ ARG-ST.FIRST.LOCAL.CAT.HEAD.CASE obl ].
```

Fonte: imagem de parte do arquivo *my-portuguese.tdl* no LKB.

Verbos divalentes como obedecer e depender, que regem dativo e genitivo, respectivamente, herdam as propriedades dos tipos *dat-obj-verb-lex* e *gen-obj-verb-lex* da Figura 21. O tipo *prep-obj-verb-lex*, supertipo de ambos, constitui subtipo de *transitive-verb lex* da Figura 7, especificando adicionalmente que o objeto seja um sintagma adposicional. Enquanto *dat-obj-verb-lex* exige dativo, *gen-obj-verb-lex* requer genitivo. Ambos possuem subtipos com e sem um reflexivo expletivo, que se distinguem pelos prefixos *refl* e *nonrefl*.

Figura 21 – Alguns tipos de verbos de objeto preposicionado.

```
prep-obj-verb-lex := transitive-verb-lex &
                     [ SYNSEM.LOCAL.CAT.VAL.COMPS.FIRST.LOCAL.CAT.HEAD adp ].

dat-obj-verb-lex := prep-obj-verb-lex &
  [ SYNSEM.LOCAL.CAT.VAL.COMPS.FIRST.LOCAL.CAT.HEAD.CASE dat ].

gen-obj-verb-lex := prep-obj-verb-lex &
  [ SYNSEM.LOCAL.CAT.VAL.COMPS.FIRST.LOCAL.CAT.HEAD.CASE gen ].

nonrefl-gen-obj-verb-lex := noninh-refl-verb-lex & gen-obj-verb-lex.

refl-gen-obj-verb-lex := inh-refl-verb-lex & gen-obj-verb-lex.

nonrefl-dat-obj-verb-lex := noninh-refl-verb-lex & dat-obj-verb-lex.

refl-dat-obj-verb-lex := inh-refl-verb-lex & dat-obj-verb-lex.
```

Fonte: imagem de parte do arquivo *my-portuguese.tdl* no LKB.

Para verbos bitransitivos com complementos realizados por sintagmas nominais ou preposicionais, criamos vários subtipos do tipo abstrato *ditransitive-lex-item* do arquivo *matrix.tdl*. Verbos de transferência de posse (43), por exemplo, constituem instâncias do tipo *nom-acc-rec-ditrans*. Esse tipo especifica que os seus três argumentos, i.e., o sujeito e os complementos direto e indireto, sejam realizados por meio dos casos nominativo, acusativo e recipiente. Este último é uma espécie de supercaso, abarcando os casos dativo e alvo, realizados, respectivamente, pelas preposições *a* e *para*. Esse supercaso evita a necessidade de criar duas variantes verbais para cada verbo que participe da alternância exemplificada em (43), (44) e (49), por um lado, e (99)-(102), por outro. Verbos que não participam dessa alternância, como em (98) e (103), cujos complementos preposicionados devem ser encabeçado por *a* e *para*, instanciam tipos de verbos que exigem o caso dativo e o caso alvo, respectivamente.

(99) O artista doou uma bicicleta para o estudante.
(100) O artista contou para o estudante que o gato tinha matado uma ratazana.
(101) Perguntei para o estudante se ele tinha dormido.
(102) A artista prometeu para o estudante matar a ratazana.
(103) A artista desafiou o estudante para uma partida.

Também no campo dos verbos bitransitivos deparamos em português com uma grande diversidade de complementos oracionais, caracterizados por diferentes modos e formas verbais e uma ampla variedade de complementadores. Para analisar esses verbos, construímos diversos subtipos dos seguintes tipos abstratos do arquivo *matrix.tdl*: *ditransitive-lex-item*, *ditrans-first-arg-control-lex-item* e *ditrans-second-arg-control-lex-item*. Esses subtipos especificam as propriedades de verbos declarativos e de inquirição (44)-(46), de controle do sujeito (49), do objeto direto (16) e do objeto indireto (5).

O atributo COMP-FORM, criado para dar conta de exemplos como (51)-(54) com verbos de alçamento para sujeito, foi fundamental também para dar conta de verbos de controle cujo complemento oracional é encabeçado por complementador preposicional, como em (16) e (5), uma vez que cada um desses verbos exige um complementador específico. Utilizando esse atributo, implementamos entradas para outros complementadores preposicionais, como *para*.

Mateus *et al.* (1989) caracterizam como construções de controle apenas sentenças como (104)-(107) em que o verbo encaixado está no infinitivo não flexionado, cujo sujeito seria o pronome nulo anafórico PRO. Não haveria controle, contudo, dos sujeitos nulos de orações completivas no infinitivo flexionado ou numa forma finita, como em (108)-(115), não obstante serem correferentes do sujeito ou do objeto do verbo superior:

(104) Os artistas disseram ter ganho o festival.
(105) Nós acreditamos ter ganho o festival.
(106) As alunas lamentam ter partido a jarra.
(107) A Maria viu as amigas a chorar.
(108) As alunas lamentam terem partido a jarra.
(109) Os artistas disseram que ganharam o festival.
(110) Os artistas disseram terem ganho o festival.
(111) Nós acreditamos termos ganho o festival.

(112) Os alunos viram irem escorregar.
(113) Os alunos viram que iam escorregar.
(114) Eu autorizei os alunos a que escrevessem a carta.
(115) Eu autorizei os alunos a escreverem a carta.

Na PorGram, não restringimos a noção de controle a sentenças do tipo de (104)-(107), estendendo-a a alguns dos exemplos de (108)-(115). Esse segundo grupo de sentenças abriga estruturas superficialmente idênticas, porém heterogêneas do ponto de vista da valência verbal. Conforme Mateus *et al.* (1989), verbos declarativos, como *dizer*, e de atividade mental, como acreditar, ao contrário de verbos causativos, perceptivos e avaliativos de uso factivo, como *lamentar*, não licenciam orações completivas no infinitivo flexionado com sujeito na ordem canônica nos moldes de (39), exigindo a inversão exemplificada em (81). Esse contraste nos levou, por um lado, a atribuir a exemplos do tipo de (108) a mesma estrutura de (39), com a única diferença de que o sujeito da completiva é um pronome nulo *pro* (MIOTO; SILVA; LOPES, 2005, p. 245), não necessariamente correferente do sujeito do verbo matriz. Analogamente, os verbos superiores de (109) e (113) instanciam as mesmas variantes verbais de (37) e (38), respectivamente. Por outro lado, analisamos (110)-(112) como estruturas de controle, uma vez que não admitem a realização lexical do sujeito encaixado, necessariamente correferente do sujeito do verbo superior.

Também na análise de (114) e (115) divergimos de Mateus *et al.* (1989), que, embora reconheçam ser obrigatória a correferência nessa construção, não a tratam como estrutura de controle. Na PorGram, verbos como autorizar, em construções análogas a (114) e (115), instanciam os tipos *a-que-ditrans-second-arg-control-verb-lex* e *a-inf-ditrans-second-arg-control-verb-lex*. Ambos constituem subtipos do tipo *ditrans-second-arg-control-verb-lex*, que, como vimos na seção 2 a respeito de (16)-(18), codifica as propriedades comuns a todos os verbos trivalentes cujo segundo argumento

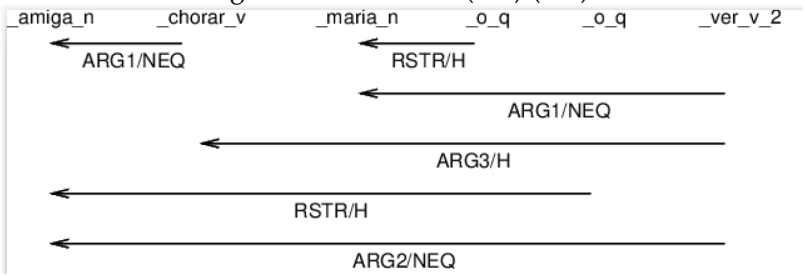
controla o sujeito da oração completiva que constitui o terceiro argumento desses verbos.

Para Mateus *et al.* (1989), subjaz a (116) estrutura paralela à de uma sentença como (39). Não se teria nesse exemplo, portanto, controle do objeto, uma vez que o sintagma as amigas não ocuparia a posição de objeto do verbo matriz, mas a de sujeito do verbo encaixado. Conforme essa abordagem, apenas haveria controle do objeto em (117), variante de (107) no português do Brasil. Por outro lado, (118), gramatical não só na variedade brasileira, mas também na europeia (GONÇALVES; CARRILHO; PEREIRA, 2016, p. 549), não é considerada por Mateus *et al.* (1989) plenamente aceitável no português europeu. Contrariamente a essa abordagem, analisamos *ver* em (116)-(118) como verbo de controle, dada a possibilidade de subida do clítico, ver (119)-(121)[9]. Na PorGram, a mesma variante de *ver* subjaz a (116)-(118), não obstante a variação de forma do verbo encaixado. Essa variante constitui instância do tipo *nonpast-nonfin-ditrans-second-arg-control-verb-lex*, que exige que o verbo encaixado tenha uma forma não finita e não passada, licenciando, portanto, tanto o gerúndio quanto o infinitivo flexionado ou não flexionado, mas excluindo o particípio passado. Essa análise, exemplificada nas Figuras 22 e 24, é adotada para todos os verbos perceptivos. Observe na Figura 22 que o predicado *ver_v_2*, atribuído à variante *ver_2*, possui três argumentos, correspondentes, respectivamente, aos sintagmas a *Maria* (ARG1), *as amigas* (ARG2) e *chorar* (ARG3) da Figura 24. Nas Figuras 23 e 25 temos as análises correspondentes da variante ver_1, que herda as propriedades do tipo *ind-cl-verb-lex*, próprio dos verbos que regem oração completiva no indicativo encabeçada por *que*. Como mostra a Figura 23, essa variante possui apenas dois argumentos. O sintagma *as amigas* constitui argumento apenas do verbo *chorar*.

[9] Para Abeillé (2021), a atribuição de caso em alemão permite enquadrar verbos perceptivos como de alçamento, questão que adiamos para uma próxima versão da gramática.

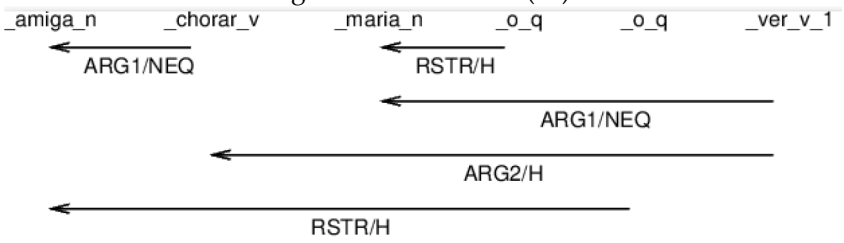
(116) A Maria viu as amigas chorarem.
(117) A Maria viu as amigas chorando.
(118) A Maria viu as amigas chorar.
(119) A Maria as viu chorarem.10
(120) A Maria as viu chorando.
(121) A Maria as viu chorar.

Figura 22 – DRMS de (116)-(118).



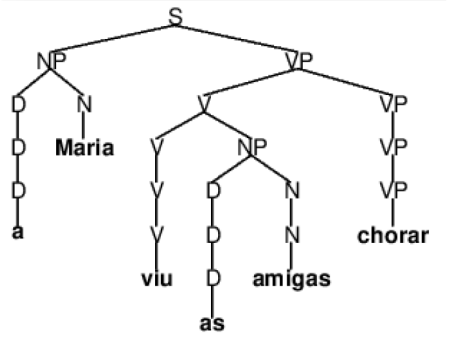
Fonte: gerada pelo LKB a partir da PorGram.

Figura 23 – DRMS de (38).



Fonte: gerada pelo LKB a partir da PorGram.

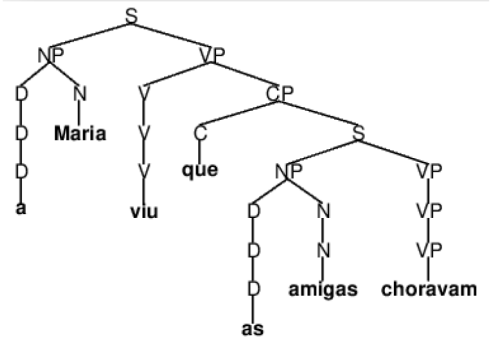
Figura 24 – Árvore de (118).



Fonte: gerada pelo LKB a partir da PorGram.

[10] Agramatical para Silveira (1997), essa construção ocorre na língua culta: "O intérprete da expedição os ouviu gritarem algo [...]." (SILVA; KOMISSAROV *et al.*, 1997)

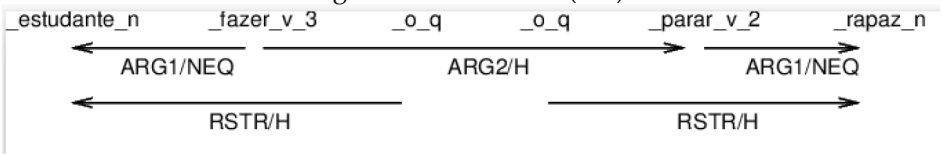
Figura 25 – Árvore de (38).



Fonte: gerada pelo LKB a partir da PorGram.

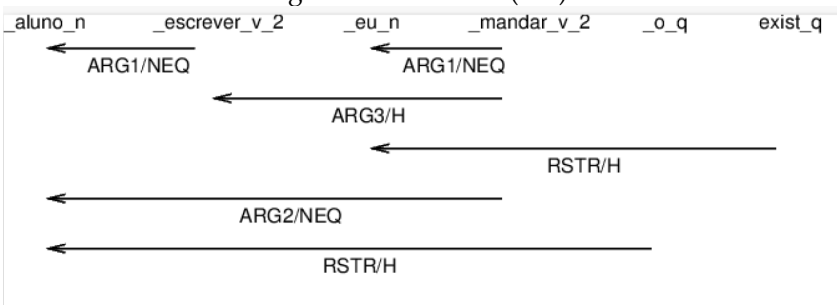
Outra divergência em relação a Mateus *et al.* (1989) consiste na análise dos verbos causativos, em exemplos como (122) e (123). Ambos os verbos licenciam a subida de clítico que realiza o agente do verbo encaixado em exemplos como (124). Ao que parece, esse exemplo seria agramatical para Mateus *et al.* (1989), que apresentam apenas o exemplo (131) com o agente do verbo encaixado no dativo. O caso acusativo do pronome em (124), ao nosso ver, evidencia que ocupa a posição de objeto do verbo superior. Não obstante essa semelhança superficial, as duas variantes verbais associam-se a tipos distintos na PorGram. Enquanto *mandar* no primeiro exemplo constitui um verbo de controle do objeto, *fazer* no segundo é um verbo de alçamento para objeto. Como podemos constatar nas Figuras 27 e 26, apenas no caso de *mandar* o objeto direto constitui argumento semântico do verbo superior. O contraste entre (126) e (127) mostra que *fazer*, ao contrário de *mandar*, não impõe restrições de seleção sobre o seu objeto (POLINSKY, 2013; ABEILLÉ, 2021).

Figura 26 – DRMS de (123).



Fonte: gerada pelo LKB a a partir da PorGram.

Figura 27: DRMS de (128).



Fonte: gerada pelo LKB a partir da PorGram.

Segundo Mateus *et al.* (1989, p. 275), verbos causativos como *mandar* e *fazer* "exprimem uma relação de causatividade entre um agente [...] e o estado de coisas descrito pela oração completiva", constituindo uma estrutura bioracional tanto em (84) e (86) quanto (122) e (123). Exemplos como (83), pelo contrário, teriam uma estrutura monooracional por conta da fusão do verbo superior e do encaixado em um predicado complexo. A flexão do infinitivo seria obrigatória nas variantes sem o complementador *que*, mas a sua ausência apenas reduziria a aceitabilidade da sentença, resultando em agramaticalidade apenas para alguns falantes (128). Ao que tudo indica, para Mateus *et al.* (1989), a subida de clítico somente é licenciada em estruturas monooracionais como (129)-(131). De fato, não apresentam exemplos análogos a (132) e (133), apesar de gramaticais não só no português brasileiro, mas também no europeu (GONÇALVES; CARRILHO; PEREIRA, 2016).

(122) Ela mandou os alunos escreverem uma carta.
(123) A estudante fez os rapazes parar.
(124) Ele a mandou comprar uma cerveja.
(125) O barulho os fez fugir.
(126) A cantora fez o chão tremer.
(127) *A cantora mandou o chão tremer.
(128) ?Eu mandei os alunos escrever.[11]

[11] Exemplos (128)-(131) extraídos de Mateus *et al.* (1989, p. 275-276), (132) e (133), de Gonçalves, Carrilho e Perira (2016, p. 551).

(129) Eu mandei-os escrever.
(130) Eu mandei-a escrever aos alunos.
(131) Eu mandei-lhes escrever a carta.
(132) A presidente da Assembleia mandou os deputados votar a lei.
(133) A presidente da Assembleia mandou-os votar a lei.

Na PorGram, tanto *fazer* quanto *mandar* aceitam as duas formas do infinitivo. Para dar conta dessa alternância, criamos o tipo disjuntivo *infl-or-not-inf*, supertipo dos tipos *inf* e *infl-inf*, correspondentes, respectivamente, ao infinitivo não flexionado e ao flexionado. Verbos como os modais e aspectuais constituem instâncias de tipos que especificam que a forma do complemento infinitivo é *inf*, ao passo que verbos como *lamentar* na variante de (39) herdam a especificação de que a forma da oração completiva é *infl-inf*.

Em suma, implementamos, por meio do questionário de customização da *Grammar Matrix*, não apenas os padrões canônicos de verbos monovalentes e divalentes, mas também verbos dessas duas superclasses com reflexivo expletivo e verbos que governam diferentes tipos de orações completivas. O sistema possibilitou, também, implementar diferentes tipos de verbos de alçamento para sujeito, classificados como auxiliares no quadro da *Grammar Matrix*, incluindo auxiliares de tempos compostos, modais e outros verbos que regem infinitivo, gerúndio e particípio. Por essa via, foram implementados, para verbos e auxiliares, 47 tipos, dos quais 27 são tipos com instâncias, que somam 140 entradas.

Por meio da exploração de tipos abstratos do arquivo *matrix.tdl*, superamos diversas limitações do questionário de customização. Em primeiro lugar, implementamos preposições como marcadores de caso, o que permitiu incluir no léxico verbos com objeto preposicionado. Em segundo lugar, implementamos verbos de alçamento para sujeito que regem complementos infinitivos encabeçados por complementadores, como os aspectuais *começar* e *parar*. Em terceiro lugar, ampliamos

a aridade dos verbos, incluindo verbos trivalentes com complementos realizados como sintagmas nominais, preposicionais ou oracionais, tanto finitos quanto não finitos, nus ou encabeçados por complementadores. Em quarto lugar, implementamos o controle não apenas do sujeito e do objeto direto, mas também do objeto indireto. Por último, implementamos verbos de alçamento para objeto. Esse esforço de codificação manual resultou em 71 novos tipos, dos quais 30 são tipos com instâncias. Ao todo, 138 entradas lexicais de verbos foram construídas por meio da manipulação direta do código em TDL, incluindo entradas cujos tipos foram codificados por meio do questionário. A maioria das classes valenciais implementadas possuem, além de uma ou mais entradas verbais, sentenças exemplificativas, constituindo, desse modo, um arcabouço para uma ampliação sistemática do léxico por meio da exploração de corpora e outros recursos disponíveis.

Os dicionários de valência mostram que a grande maioria dos verbos do português enquadra-se em múltiplas classes valenciais. No entanto, nossa preocupação principal na atual fase de desenvolvimento da PorGram não foi alcançar uma grande cobertura lexical. Desse modo, o léxico verbal da gramática contém tão somente 278 entradas para 215 lemas, a 80% dos quais foi atribuída uma única valência. Dos restantes, um total de 13% possui duas valências, e 6.5%, 3 ou 4.

O verbo *dizer* é o único com seis variantes, que instanciam seis tipos distintos, incluindo os exemplificadas em (5), (37), (110) e (134). Essa lista, contudo, não cobre toda a gama de construções documentadas em Fernandes (1987) e Borba (1991). Algumas dessas construções correspondem a tipos da PorGram, outras exigem a construção de novos tipos a partir dos existentes.

4 Avaliação

Nesta seção, apresentamos os resultados da avaliação da gramática, no que tange ao tratamento da valência verbal.

Na construção de gramáticas computacionais de línguas naturais, tipicamente se adota um ciclo de desenvolvimento com as seguintes etapas: (i) elaboração de um teste constituído de sentenças a serem analisadas pela gramática, (ii) implementação do código necessário para alcançar esse objetivo, (iii) aplicação da gramática ao teste, (iv) correção do código com base nos resultados, (v) reteste caso necessário, (vi) ampliação do teste e, finalmente, retorno à etapa (ii).

Como explicamos anteriormente, o código tem sido elaborado tanto automaticamente, por meio do sistema de customização, quanto manualmente por meio da edição em TDL. Em engenharia da gramática baseada no conhecimento, consolidou-se a utilização da técnica do fragmento (FRANCEZ; WINTNER, 2012; ZAMARAEVA, 2021). Em vez de se objetivar o esgotamento da análise de um dado fenômeno gramatical, define-se inicialmente um recorte que abarca apenas determinados aspectos. A implementação desse recorte visa a dois objetivos: (i) analisar sentenças que exemplificam os aspectos envolvidos, (ii) não analisar sentenças agramaticais que violam as restrições postuladas.

Para verificar em que medida os dois objetivos foram alcançados pela implementação, constroem-se dois conjuntos de teste: um conjunto de teste positivo, com sentenças gramaticais, e um conjunto de teste negativo, com sentenças agramaticais. As sentenças do conjunto de teste negativo derivam de sentenças do conjunto de teste positivo por meio da introdução de violações às restrições postuladas. Por exemplo, (139) contraria exigência relativa à forma do verbo encaixado na construção progressiva, observada em (138).

(138) Perguntei-me a quem o cachorro estava obedecendo.
(139) *Perguntei-me a quem o cachorro estava obedecido.

A PorGram, no estágio atual, resulta de dezenas de repetições do ciclo de desenvolvimento delineado. No LKB, não é necessário armazenar os dois conjuntos em arquivos separados, bastando prefixar as sentenças agramaticais com um asterisco. O *parser* ignora esse símbolo. Os atuais conjuntos de teste estão agrupados em dois arquivos, DEV-TEST e MAT-TEST. O primeiro arquivo foi construído incrementalmente, com o acréscimo de exemplos a cada repetição do ciclo. Constitui-se no momento de 631 exemplos, dos quais 139 são agramaticais.

O segundo arquivo contém 89 exemplos gramaticais e 28 agramaticais derivados das sentenças que exemplificam a descrição de Mateus *et al.* (1989) dos vários tipos de orações que funcionam como complementos verbais. Seguindo convenção de praxe, além do asterisco para marcar a agramaticalidade, Mateus *et al.* (1989) usam um ou dois sinais de interrogação para assinalar pouca aceitabilidade gramatical. Em gramáticas computacionais como a PorGram, porém, só é possível modelar julgamentos binários de gramaticalidade. Desse modo, convertemos todos os sinais de interrogação dos exemplos de Mateus *et al.* (1989) em asterisco, excetuando exemplos como (118) e (123), que, como vimos, são gramaticais.

Na adaptação das sentenças de Mateus *et al.* (1989), preservamos todos os aspectos estruturais relevantes para a avaliação da cobertura da PorGram em relação aos padrões de valência objeto da seção 4. No entanto, substituímos algumas construções, como, por exemplo, a voz passiva, e parte do vocabulário, dadas as presentes limitações da PorGram. Por outro lado, incluímos sentenças com a perífrase progressiva tanto na sua forma no português europeu padrão, quanto sua contraparte com gerúndio na variedade brasileira e em dialetos da europeia (GONÇALVES; CARRILHO; PEREIRA, 2016), comparem-se (107) e (117).

A Tabela 2 apresenta os resultados da aplicação da gramática sobre os dois arquivos de teste DEV-TEST e MAT-TEST, computando as quantidades de

verdadeiros negativos (TN), verdadeiros positivos (TP), falsos positivos (FP) e falsos negativos (FN). Os exemplos positivos consistem nas sentenças gramaticais, enquanto os negativos são as agramaticais. O índice R de cobertura (*recall*) da gramática em relação ao recorte gramatical e lexical pré-definido mede-se pela fórmula TP/(TP+FN), enquanto a precisão P é o resultado de TP/(TP+FP)). O índice FM, chamado Medida F (*F-Measure*), é a média harmônica de R e P (BIRD; KLEIN; LOPER, 2009). A hipergeração H da gramática, que consiste em analisar construções agramaticais como gramaticais, indicando que as restrições implementadas não são restritivas o suficiente, obtém-se pela fórmula FP/(TN+FP). Quanto menor esse índice, mais restritiva é a gramática.

Tabela 2 – Resultados da aplicação da gramática sobre os conjuntos de teste DEV-TEST e MAT-TEST.

	TN	TP	FP	FN	R	P	FM	H
DEV-TEST	130	480	9	12	0.98	0.98	0.98	0.06
MAT-TEST	21	66	7	23	0.74	0.90	0.81	0.25

Fonte: elaborada pelos autores.

Como podemos constatar na Tabela 2, a gramática alcançou uma alta cobertura em relação ao conjunto DEV-TEST, analisando 98% dos exemplos gramaticais, ao mesmo tempo que analisou apenas 6% dos exemplos agramaticais, resultando num índice FM de 0.98. A maior parte dos falsos positivos desse conjunto são sentenças declarativas com verbo não finito ou no subjuntivo. Esse problema se deve a que a *Grammar Matrix* não restringe o modo ou a forma verbal da sentença raiz. Quantos aos falsos negativos, (140) é o único exemplo referente à valência verbal. A regência de *insistir* nesse exemplo, consignada em Borba (1991), representa uma variação mínima da exemplificada em (85), codificada na PorGram como *insistir_3* com o tipo *em-que-cl-verb-lex*. Para que a gramática pudesse analisar (140) bastaria construir uma variante *insistir_5* com o tipo *para-que-cl-verb-lex*, analogamente a *pedir* em (88). De forma

alternativa, poderíamos implementar um tipo disjuntivo de forma de complementador que permitisse condensar as duas regências em uma única variante verbal, a exemplo do que fizemos para os verbos perceptivos.

(140) Ela insistiu para que o gato dormisse.

Do conjunto MAT-TEST, a PorGram analisou 74% dos exemplos gramaticais e 25% dos agramaticais, resultando num índice FM de 0.81. Ressalte-se que esse conjunto possui relativamente poucos exemplos agramaticais, uma vez que Mateus *et al.* (1989) não construíram sistematicamente, a partir das sentenças gramaticais, exemplos com desvios das regularidades modeladas, limitando-se a exemplos esparsos. Certamente, um maior número de exemplos agramaticais teria aumentado o índice FM e reduzido a hipergeração da gramática relativamente ao conjunto MAT-TEST.

Os falsos negativos do conjunto MAT-TEST exemplificam fenômenos ainda não implementados, incluindo particularidades do português europeu, como em (81)-(83) e (107).

As representações sintáticas e semânticas geradas pela PorGram para os falsos positivos do conjunto MAT-TEST revelam que se trata de sentenças talvez semanticamente anômalas, mas possíveis sintaticamente. A MRS de (141), por exemplo, apresenta a variante *dizer_v_2*, indicando que se trata do uso do verbo *dizer* como declarativo de ordem, ver (134).

(141) *Os críticos disseram que o filme ganhe o festival.

5 Considerações finais

Neste artigo, tratamos da implementação da valência na PorGram, uma nova gramática computacional do português no formalismo da HPSG. Ainda num estágio intermediário de desenvolvimento, a PorGram propõem-se a constituir, num futuro próximo, uma alternativa de software livre e de código aberto à LXGram. Também implementada em HPSG, mas com código proprietário, a LXGram é a única gramática do português de larga escala voltada à análise sintática profunda. Esse tipo de análise tem-se mostrado bastante relevante em tarefas de compreensão textual automática.

Tal como sua contraparte de código proprietário, a PorGram utilizou o questionário de customização da *LinGO Grammar Matrix* para construção automática do código em TDL de uma gramática inicial, modificado e expandido em etapas posteriores de desenvolvimento. Como outras gramáticas que compartilham a arquitetura da *LinGO Grammar Matrix*, a PorGram produz não apenas representações sintáticas, mas também semânticas, utilizando, para tanto, o formalismo da MRS, que supera diversas limitações da LPO.

Com esse sistema de customização, implementamos verbos intransitivos e transitivos prototípicos, ou seja, com sujeito no nominativo e complemento no acusativo, verbos com reflexivos inerentes bem como verbos divalentes que regem um amplo leque de tipos de orações completivas, caracterizados por diferentes combinações dos possíveis valores dos seguintes parâmetros: (i) força ilocucionária (proposição ou interrogação), (ii) forma do complementador (conjunção se ou pronome interrogativo, conjunção *que* precedida ou não de preposição ou foneticamente vazio), (iii) modo ou forma verbal (verbo no infinitivo flexionado, no modo indicativo ou no subjuntivo). O sistema permitiu igualmente implementar verbos de alçamento para sujeito que regem complementos no infinitivo, particípio ou gerúndio. Esses verbos incluem modais e auxiliares de tempos compostos e de perífrases aspectuais. Ao todo, 27 classes de verbos foram implementadas por essa via.

Por meio da manipulação direta do código em TDL, tomando como ponto de partida os tipos abstratos do arquivo *matrix.tdl* da *LinGO Grammar Matrix*, conseguimos superar diversas limitações do questionário de customização, que não contempla os seguintes fenômenos relacionados à valência verbal: (i) verbos com dois complementos (ou seja trivalentes), (ii) verbos de controle do sujeito e do objeto direto ou indireto, (iii) verbos de alçamento para objeto e (iv) verbos de alçamento para sujeito com complemento no infinitivo encabeçado por complementador. Para tanto, criamos diversos novos tipos de verbos e complementadores e redesenhamos a hierarquia de formas do verbo, de modo a dar conta da compatibilidades de determinados verbos, como os causativos e perceptivos, com mais de uma forma do verbo encaixado. Também implementamos manualmente preposições como marcadoras de caso e diversos tipos de verbos que exigem esse tipo de complemento, uma vez que a implementação desse fenômeno por meio do questionário se mostrou bastante difícil. Por outro lado, a maneira como o modo verbal das orações completivas introduzidas pelo complementador *que* é determinado na versão da gramática gerada pelo sistema revelou-se insuficientemente restritiva, deixando de bloquear exemplos agramaticais com o modo verbal incorreto. Isso nos levou a redefinir manualmente os tipos de verbos que exigem oração completiva encabeçada pelo complementador *que* num modo verbal específico. Ao todo, construímos de forma manual 71 tipos para codificação da valência verbal, dos quais 30 representam classes de verbos.

A PorGram foi testada em dois conjuntos de sentenças exemplificativas dos fenômenos abrangidos pelo recorte gramatical de que pretende constituir um modelo formal. Esses dois conjuntos de teste também contêm exemplos agramaticais que visam medir o quão restritivo é o modelo. O primeiro, com 492 sentenças gramaticais e 139 agramaticais, foi construído incrementalmente ao longo do desenvolvimento da gramática. No momento, são analisadas 480 das sentenças gramaticais e apenas 9 das construções agramaticais. Mostramos que o único falso negativo referente à valência

verbal poderia ser trivialmente corrigido incluindo uma variante do verbo *insistir* com um tipo já implementado na gramática.

Do segundo conjunto, derivado dos exemplos de Mateus *et al.* (1989), totalizando 117 exemplos, são analisados 74% das 89 sentenças gramaticais e 25% das consideradas agramaticais. A menor cobertura da PorGram em relação a esse conjunto resulta, sobretudo, dos diversos fenômenos que exemplifica não relacionados diretamente com a valência verbal, como, por exemplo, objetos clíticos, construções de clivagem e a perífrase progressiva do português europeu padrão, constituída de *estar a* e infinitivo. A formação de predicados complexos nessa variedade, porém, constitui fenômeno do domínio da valência não abarcado pela PorGram no momento, representando um dos desafios para uma versão futura da gramática. Quanto aos falsos positivos desse conjunto de teste, a um exame mais detalhado revelaram-se sintaticamente possíveis, embora semanticamente anômalos. Não indicam, portanto, hipergeração da gramática em sentido estrito.

A maior parte dos 57 tipos de verbos implementados, incluindo auxiliares, são representados por pelo menos uma entrada lexical e exemplificados com uma ou mais sentenças dos conjuntos de teste. Com 215 verbos codificados em 278 entradas lexicais, a PorGram cobre uma fração diminuta do léxico verbal português, do qual o número de lemas em circulação ativa na língua Borba (1991) estimaram em cerca de 6000, menos da metade dos verbetes de Fernandes (1987).

No entanto, a hierarquia de tipos proposta constitui uma infraestrutura para um léxico computacional de valências de uma profundidade gramatical que nos parece ímpar no domínio da língua portuguesa. De fato, modela a valência no domínio tanto da sentença simples quanto da sentença composta, distinguindo entre controle e alçamento. Pelo contrário, os dicionários de valência de grande cobertura disponíveis (FERNANDES, 1987; BORBA, 1991) não são formalizados, furtando-se a modelar as noções de controle e alçamento. Por outro lado, descrições mais aprofundadas de

grupos de verbos de determinados campos semânticos, como, por exemplo, verbos de mudança (CANÇADO; GODOY; AMARAL, 2013), restringem-se a sentenças simples, não abordando verbos com complementos oracionais. Analogamente, o projeto *Valências Verbais do Português Brasileiro*, ainda em andamento, não inclui esses padrões mais complexos de complementação (PERINI, 2016; PERINI *et al.*, 2019), controle e alçamento não integrando o quadro teórico subjacente a esse levantamento (PERINI, 2015).

Para concluir, destacamos os principais desafios que se apresentam para as próximas etapas de desenvolvimento da PorGram no terreno da valência verbal. Em primeiro lugar, falta construir tipos para verbos impessoais e verbos com sujeitos oracionais, adaptando às particularidades morfossintáticas do português os tipos abstratos correspondentes do arquivo nuclear *matrix.tdl* da *Grammar Matrix*. Em segundo lugar, é preciso ampliar o leque de descrições linguísticas dos padrões de complementação mais complexos a serem tomadas como base, visto que nos limitamos, no presente trabalho, à abordagem de Mateus *et al.* (1989), levando em conta julgamentos de aceitabilidade de Gonçalves, Carrilho e Pereira (2016) sobre algumas construções do português europeu. Em terceiro lugar, há que implementar fenômenos próprios do português europeu, como a formação de predicados complexos, a construção progressiva de *estar* com infinitivo encabeçado pelo complementador a ou a inversão do sujeito em orações completivas com infinitivo flexionado. Em quarto lugar, falta construir regras lexicais que modelem os processos sistemáticos de alteração da valência, de modo a poder analisar não somente sentenças na voz passiva, mas também simplificar a codificação do léxico. No momento, apenas uma regra lexical desse tipo está implementada, que é a adjunção de um reflexivo expletivo, aplicável de forma obrigatória ou facultativa sobre determinadas classes de verbos. Essa regra permite dar conta da variante incoativa de verbos como *abrir* por meio de uma única entrada lexical, ao invés das duas que seriam necessárias sem a

regra. Diversas variantes de verbos do léxico atual da PorGram poderiam ser geradas automaticamente por meio de regras análogas aplicadas sobre uma entrada lexical de base. Finalmente, é preciso povoar os tipos da gramática com instâncias, por meio da exploração sistemática dos recursos disponíveis, como *treebanks*, redes léxico-semânticas ou levantamentos das valências de classes semânticas individuais de verbos, como os mencionados.

Referências

ABEILLÉ, A. Control and raising. *In*: MÜLLER, S. *et al*. (ed.). **Head Driven Phrase Structure Grammar**: The handbook. Berlin: Language Science Press, 2021. p. 489–535.

ALENCAR, L. F. de; RADEMAKER, A. Cross-validating language resources for the development of a large-coverage computational grammar of Portuguese. **Language Resources and Evaluation**. Submetido à publicação.

ARRAIS, D. **Você sabe a diferença entre "pedir para" e "pedir que"?** 2017. Disponível em: https://exame.com/carreira/voce-sabe-a-diferenca-entre-pedir-para-e-pedir-que/. Acesso em: 21 nov. 2021.

BENDER, E. M. Reweaving a grammar for Wambaya: A case study in grammar engineering for linguistic hypothesis testing. **Linguistic Issues in Language Technology**, v. 3, p. 1–34, 2010. DOI https://doi.org/10.33011/lilt.v3i.1215

BENDER, E. M. et al. Grammar customization. **Research on Language & Computation**, v. 8, n. 1, p. 23–72, 2010. DOI https://doi.org/10.1007/s11168-010-9070-1

BENDER, E. M.; FLICKINGER, D.; OEPEN, S. The Grammar Matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. *In*: **COLING-GEE '02**: Proceedings of the 2002 Workshop on Grammar Engineering and Evaluation. [S.l.]: [s.n.], 2002. p. 8–14. DOI https://doi.org/10.3115/1118783.1118785

BENDER, E. M.; FLICKINGER, D.; OEPEN, S. **MRS in the LinGO Grammar Matrix**: A practical user's guide. [S.l.]: [s.n.], 2003. Disponível em: http://faculty.washington.edu/ebender/papers/userguide.pdf. Acesso em: 25 set. 2021.

BENDER, E. M.; FLICKINGER, D.; OEPEN, S. Grammar engineering and linguistic hypothesis testing: Computational support for complexity in syntactic analysis. *In*: BENDER, E. M.; ARNOLD, J. E. (ed.). **Language from a cognitive perspective**: Grammar, usage and processing. Stanford: CSLI, 2011. p. 5–29.

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python**: analyzing text with the Natural Language Toolkit. Sebastopol: O'Reilly, 2009.

BORBA, F. da S. (org.). **Dicionário gramatical de verbos do português contemporâneo do Brasil**. 2. ed. São Paulo: Editora da UNESP, 1991.

CANÇADO, M.; AMARAL, L.; MEIRELLES, L. **VerboWeb**: classificação sintático-semântica dos verbos do português brasileiro. Belo Horizonte: UFMG, 2017. Disponível em: http://www.letras.ufmg.br/verboweb. Acesso em: 13 dez. 2021.

CANÇADO, M. et al. Banco de dados VerboWeb: um panorama do léxico verbal do PB. *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 13 , 2021. Evento Online. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 372-380. DOI https://doi.org/10.5753/stil.2021.17817

CANÇADO, M.; GODOY, L.; AMARAL, L. **Catálogo de verbos do português brasileiro**: Classificação verbal segundo a decomposição de predicados. vol. 1: Verbos de mudança. Belo Horizonte: Editora da UFMG, 2013.

CARLIER, A.; GOYENS, M.; LAMIROY, B. De: A genitive marker in french? *In*: CARLIER, A.; VERSTRAETE, J.-C. (ed.). **The genitive**. Amsterdam: John Benjamins, 2013. p. 141–216. DOI https://doi.org/10.1075/cagral.5.07car

COPESTAKE, A. **Implementing typed feature structure grammars**. Stanford: CSLI, 2002.

COPESTAKE, A. Slacker semantics: Why superficiality, dependency and avoidance of commitment can be the right way to go. *In*: CONFERENCE OF THE EUROPEAN CHAPTER OF THE ACL, 12 , 2009, Athens. **Proceedings** [...]. Athens: Association for Computational Linguistics, 2009. p. 1–9. DOI https://doi.org/10.3115/1609067.1609167

COPESTAKE, A. et al. Minimal Recursion Semantics: An introduction. **Research on language and computation**, Springer, v. 3, n. 2, p. 281–332, 2005. DOI https://doi.org/10.1007/s11168-006-6327-9

COPESTAKE, A.; LASCARIDES, A.; FLICKINGER, D. An algebra for semantic construction in constraint-based grammars. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 39, Toulouse. **Proceedings** […]. Toulouse: Association for Computational Linguistics, 2001. p. 140–147. DOI https://doi.org/10.3115/1073012.1073031

COSTA, F.; BRANCO, A. LXGram: A deep linguistic processing grammar for Portuguese. *In*: PARDO, T. A. S. *et al*. (ed.). **Computational Processing of the Portuguese Language**. Berlin, Heidelberg: Springer, 2010. p. 86–89. DOI https://doi.org/10.1007/978-3-642-12320-7_11

CUNHA, C.; CINTRA, L. **Nova gramática do português contemporâneo**. Rio de Janeiro: Nova Fronteira, 1985.

CURTIS, C. **A Parametric implementation of valence-changing morphoplogy in the LinGO Grammar Matrix**. Dissertação (Mestrado) — University of Washington, Seattle, 2018. Disponível em: http://hdl.handle.net/1773/41814

DRELLISHAK, S. **Widespread but Not Universal**: Improving the typological coverage of the Grammar Matrix. Tese (Doutorado) — University of Washington, Seattle, 2009.

DROGANOVA, K.; ZEMAN, D. Towards deep Universal Dependencies. *In*: INTERNATIONAL CONFERENCE ON DEPENDENCY LINGUISTICS (DepLing), 5, 2019, Paris. **Proceedings** [...]. Paris: Association for Computational Linguistics, 2019. p. 144–152. DOI https://doi.org/10.18653/v1/W19-7717

FALK, Y. N. **Lexical-Functional Grammar**: An introduction to parallel constraint-based syntax. Stanford: CSLI, 2001.

FERNANDES, F. **Dicionário de verbos e regimes**. 35. ed. Rio de Janeiro: Globo, 1987.

FERRUCCI, D. et al. Building Watson: An overview of the DeepQA project. **AI Magazine**, v. 31, n. 3, p. 59–79, 2010. DOI https://doi.org/10.1609/aimag.v31i3.2303

FLICKINGER, D. On building a more efficient grammar by exploiting types. **Natural Language Engineering**, Cambridge University Press, v. 6, n. 1, p. 15–28, 2000. DOI https://doi.org/10.1017/S1351324900002370

FRANCEZ, N.; WINTNER, S. **Unification grammars**. Cambridge: Cambridge University Press, 2012.

GABRIEL, C.; MÜLLER, N. **Grundlagen der generativen Syntax**: Französisch, Italienisch, Spanisch. Tübingen: Niemeyer, 2008.

GONÇALVES, A.; CARRILHO, E.; PEREIRA, S. Predicados complexos numa perspetiva comparativa. *In*: MARTINS, A. M.; CARRILHO, E. (ed.). **Manual de linguística portuguesa**. Berlin: De Gruyter, 2016. p. 523–557. DOI https://doi.org/10.1515/9783110368840-022

GOODMAN, M. W. Generation of machine-readable morphological rules with human readable input. **University of Washington Working Papers in Linguistics**, v. 30, p. 1–34, 2013.

MARNEFFE, M.-C. de et al. Universal Dependencies. **Computational Linguistics**, v. 47, n. 2, p. 255–308, 2021.

MATEUS, M. H. M. *et al*. **Gramática da língua portuguesa**. Lisboa: Caminho, 1989.

MATOS, M. J. **"Pedir que" vs. "pedir para"**. 2008. Disponível em: https://ciberduvidas.iscte-iul.pt/consultorio/perguntas/pedir-que-vs-pedir-para24813. Acesso em: 11 nov. 2021.

MCCORD, M. C.; MURDOCK, J. W.; BOGURAEV, B. K. Deep parsing in Watson. **IBM Journal of research and development**, IBM, v. 56, n. 3.4, p. 3–1, 2012. DOI https://doi.org/10.1147/JRD.2012.2185409

MIOTO, C.; SILVA, M. C. F.; LOPES, R. E. V. **Novo manual de sintaxe**. 2. ed. Florianópolis: Insular, 2005.

MÜLLER, S. **Grammatical theory**: From transformational grammar to constraint-based approaches. 4. ed. Berlin: Language Science Press, 2020.

NIVRE, J. *et al*. Universal dependencies v2: An evergrowing multilingual treebank collection. *In*: LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 12,

2020, Marseille. **Proceedings** […]. Marseille: European Language Resources Association, 2020. p. 4034–4043. Disponível em: https://aclanthology.org/2020.lrec-1.497. Acesso em: 29 dez. 2021.

NUNES, A. L.; RADEMAKER, A.; ALENCAR, L. F. de: Utilizando um dicionário morfológico para expandir a cobertura lexical de uma gramática do português no formalismo HPSG. *In*: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 13 , 2021. Evento Online. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 11–18. DOI https://doi.org/10.5753/stil.2021.17779

PERINI, M. Construindo o Dicionário de Valências: problemas e resultados. **Scripta**, Belo Horizonte, v. 20, p. 148–167, 2016. DOI https://doi.org/10.5752/P.2358-3428.2016v20n38p148

PERINI, M. A. **Describing verb valency**: Practical and theoretical issues. Cham: Springer, 2015. DOI https://doi.org/10.1007/978-3-319-20985-2

PERINI, M. A. et al. **Valency dictionary of Brazilian Portuguese verbs**. Não publicado. 2019.

POLINSKY, M. Raising and control. *In*: DIKKEN, M. den (ed.). **The Cambridge handbook of generative syntax**: Grammar and syntax. Cambridge: Cambridge University Press, 2013. p. 577–606. DOI https://doi.org/10.1017/CBO9780511804571.021

POULSON, L. Meta-modeling of tense and aspect in a cross-linguistic grammar engineering platform. **University of Washington Working Papers in Linguistics**, v. 28, p. 1–67, 2011.

RADEMAKER, A. et al. Universal Dependencies for Portuguese. *In*: INTERNATIONAL CONFERENCE ON DEPENDENCY LINGUISTICS (DepLing), 4, 2017, Pisa. **Proceedings** […]. Pisa: Linköping University Electronic Press, 2017. p. 197–206.

ROSÉN, V. et al. An open infrastructure for advanced treebanking. *In*: HAJIČ, J. **META-RESEARCH Workshop on Advanced Treebanking at LREC2012**. [S.l.], 2012. p. 22–29.

SAG, I. A.; WASOW, T.; BENDER, E. M. **Syntactic theory**: A formal introduction. 2. ed. Stanford: CSLI, 2003.

SCHUSTER, S.; MANNING, C. D. Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 2016, 10, Portorož. **Proceedings** [...]. Portorož: European Language Resources Association, 2016.

SIEGEL, M.; BENDER, E. M.; BOND, F. **Jacy**: An implemented grammar of Japanese. Stanford: CSLI, 2016.

SILVA, D. G. B. da; KOMISSAROV, B. N. et al. (org.). **Os diários de Langsdorff**. Campinas: Associação Internacional de Estudos Langsdorff, 1997. DOI https://doi.org/10.7476/9788575412459

SILVEIRA, G. **O comportamento sintático dos clíticos no português brasileiro**. Dissertação (Mestrado) — Universidade Federal de Santa Catarina, Florianópolis, 1997. Disponível em: https://repositorio.ufsc.br/handle/123456789/112183

STRAKA, M.; STRAKOVÁ, J. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. *In*: **Proceedings of the CoNLL 2017 Shared Task**: Multilingual parsing from raw text to Universal Dependencies. Vancouver: Association for Computational Linguistics, 2017. p. 88–99. DOI https://doi.org/10.18653/v1/K17-3009

WESTERSTÅHL, D. Generalized quantifiers. *In*: ZALTA, E. N. (ed.). **The Stanford encyclopedia of philosophy**. Stanford: Stanford University, 2019.

ZAMARAEVA, O. **Assembling Syntax**: Modeling constituent questions in a grammar engineering framework. Tese (Doutorado) — University of Washington, Seattle, 2021. Disponível em: http://hdl.handle.net/1773/47087

ZAMARAEVA, O.; HOWELL, K.; BENDER, E. M. Modeling clausal complementation for a grammar engineering resource. *In*: SOCIETY FOR COMPUTATION IN LINGUISTICS (SciL), 2019, New York. **Proceedings** […]. [S.l.]: [s.n.], 2019. p. 39–49.

ZARING, L. On prepositions and case-marking in French. **Canadian Journal of Linguistics**, v. 36, p. 363–377, 1991. DOI https://doi.org/10.1017/S000841310001450X

Artigo recebido em: 29.12.2021 Artigo aprovado em: 28.02.2022

DOI: 10.14393/DL52-v16n4a2022-7

PFN-PT: A Framenet Annotator for Portuguese

Anotação semântica automática: um novo Framenet para o português

Eckhard BICK[*]

ABSTRACT: This article presents PFN-PT, a robust system for the automatic semantic annotation of Portuguese, consisting of a new, parsing-oriented framenet and a rule-based frame- and role-tagger. The framenet provides almost 13,000 valency frames covering 7,300 verb lemmas with 10,700 senses. Frame and role tagging is achieved by iterated matching of syntactic structures and semantic noun types with slot-filler conditions in the framenet. We discuss design principles and present frame and role statistics. In an evaluation run on news data, the system achieved an overall F-score of 92.2% for frame senses.

KEYWORDS: Portuguese Framenet (PFN-PT). Semantic Role Labeling (SRL). Semantic Parsing. Constraint Grammar (CG). PALAVRAS.

RESUMO: Este artigo apresenta o PFN-PT, um sistema robusto para a anotação semântica automática de Português, consistindo numa nova framenet com foco em *parsing*, e um tagger para frames e papéis semânticos baseado em regras. A framenet contêm cerca de 13.000 padrões sintáticos cobrindo 7.300 lemas verbais com 10.700 sentidos. A etiquetagem é realizada por meio de um alinhamento iterativo de estruturas sintáticas e classe semântica de substantivos com as condições listadas no framenet para argumentos sintáticos. Discutimos princípios de desenho e apresentamos estatísticas de distribuição de categorias. Numa avaliação realizada com base em textos jornalísticos, o sistema alcançou 92,2% sentidos/frames corretos para verbos.

PALAVRAS-CHAVE: FrameNet português (PFN-PT). Etiquetagem semântica. Papéis semânticos. Gramática de restrições (CG). PALAVRAS..

1 Introduction

In modern corpus linguistics, automatic grammatical annotation of free text is a central task, and the usefulness of a corpus depends on the linguistic complexity of

[*] PhD. University of Southern Denmark. ORCID: https://orcid.org/0000-0002-5505-4861. eckhard.bick@gmail.com

its annotation. However, the performance of annotation tools decreases with increasing complexity. Thus, while lower-level annotation such as lemmatization, morphology and part-of-speech (POS) work reasonably well, syntactic-structural annotation (treebanks) is more difficult to achieve, and for most languages other than English, robust semantic annotation, with the possible exception of named-entity-annotation (NER), remains an unsolved challenge. The task can be broken down into semantic classification and disambiguation on the one hand, and semantic function and relations on the other. For nouns, the classification task is addressed with ontologies like WordNet (FELLBAUM, 1998), that have good coverage and operate on an effective classification principle (hyponymy) but fail to provide the structural-relational information necessary to disambiguate senses. Verb classification is even more difficult, because though troponymy can be used instead of hyponymy, classification is less "local" than for nouns, and intertwined with the second semantic task, assigning semantic structure and argument relations to a clause. Thus, for English, VerbNet (KIPPER et al., 2006), Berkeley FrameNet (BAKER et al., 1998; JOHNSON; FILLMORE, 2000; RUPPENHOFER et al., 2010) and PropBank[1] (PALMER et al., 2005) classify entire predications, assigning a semantic class to the core lexeme (typically, but not necessarily a verb) and semantic roles (also called case/thematic roles, FILLMORE, 1968) to its arguments and possibly adjuncts. The combination is constructed as a lexical knowledge representation called a frame (FILLMORE; BAKER, 2001), that can be triggered (evoked) by the presence of certain frame elements (FE). For instance, in Portuguese, depending on concept granularity, the frame of selling can

[1] Methodologically, VerbNet and the English Berkeley FrameNet were conceived as lexicographical projects, one frame at a time, while PropBank and the German SALSA framenet depart from corpus data, one sentence at a time. Coverage problems are therefore different in nature: In the former, a common sense may be missing in a verb with several rare senses assigned. In the latter, common senses are registered first, but rare lemmas may be missing entirely.

involve verbs like vender (sell), *exportar* (export), *liquidar* (liquidate), *aleiloar* (auction off) as well as nouns like *venda* (sale) and *exportação* (export). The involved FE's are the core argument roles of a 'seller', 'buyer' and 'goods' and the peripheral adjunct role of 'price'. These could be specified as such, but can also seen as implied - in the context of a sell verb - by the more general 'donor' (or just 'agent'), 'receiver', 'theme' and 'value'. It is a matter of framenet design - and possibly language-dependent - where to use a separate frame where the perspective changes, i.e. *buy sth from sb instead of sell sth. to sb*., or where transitivity or agency changes, e.g. *change (self/subject) and alter (sth. else/object)*. The fact that the English Berkeley FrameNet has inspired framenets in several other languages, such as German (BURCHARDT et al., 2006), Spanish (CARLOS; PETRUCK, 2010), Portuguese (SALOMÃO, 2009; TORRENT; ELLSWORTH, 2013) and Japanese (OHARA et al., 2004) has shown that such distinctional choices, though inspired by English, have a level of abstraction and universal validity that allows them to be ported across languages.

In his comparison of the WordNet and FrameNet approaches, Boas (2005) stresses the added level of abstraction provided by the latter - e.g. that frames are independent of part-of-speech (POS) -, as well as the systematic link between semantic information and lexical-syntactic patterns. Thus, both FrameNet and PropBank provide morphosyntactic restrictions to frame realization. In addition, the former also specifies ontological slot filler information. Because it is anchored in lexico-syntax, frame-based semantics has the potential of providing a bridge between classical NLP (Natural Language Parsing) and real-world applications within AI (Artificial Intelligence). Machine translation (MT), for instance, can profit from word sense disambiguation (WSD), which is an inherent "by-product" of frame assignment.

The framenet resource presented here, PFN-PT (Parsing Framenet for Portuguese), is meant to provide, for Portuguese, such a syntactic-semantic bridge at a practical level. It therefore has a methodological focus, meaning to support robust

automatic frame and role annotation of running text. This methodological primacy, as well as the frame inventory, has been borrowed from the Danish Framenet (BICK, 2011), that also focuses on automatic annotation (BICK, 2017). The complete system consists of two complementary parts, the framenet and the frame annotator, linked by a semantically informed[2,] valency-driven dependency representation provided by a morphosyntactic parser. For such a resource to allow automatic annotation, good and, above all, evenly distributed lexical coverage is crucial, and granularity should be kept at a realistic level. This homogeneity in coverage and granularity is what sets PFN-PT apart from the concept/domain-based and example-based creation methods of traditional framenets and propbanks, respectively.

2 Building the framenet

In a first round of bootstrapping, using tailor-made software, we identified Danish-Portuguese verb sense matches by harvesting the machine-translation (MT) dictionary used in the Portuguese-Danish section of the GramTrans[3] system (BICK, 2014), where polysemy is handled by providing syntactic argument and semantic slot-filler information in much the same way a frame entry would. Rather than using MT to match existing framenets in two languages (GILARDI; BAKER, 2018), we use it to match valency patterns as an anchor for frame transfer. For instance, if a Portuguese verb is allowed four different translations depending on the semantics of its subject and object, the software would look up the translations in the Danish FrameNet and choose a frame with the same slot-filler conditions. Similarly, but more heuristically,

[2] To work optimally, the method requires semantic type tags on nouns and possibly adjectives or adverbs, such as 'human', 'tool'', 'unit', 'food', 'feature' etc. These need not necessarily be disambiguated, but should be of medium granularity, so as to have distinctional potential as slot-fillers without being too fine-grained to have abstractional value.

[3] https://gramtrans.com

prepositional complements were harvested and matched to a Danish frame asking for a corresponding preposition in Danish (as suggested by the MT dictionary). All in all, the method came up with frame suggestions and slot-filler semantics for 9,414 valency patterns covering 7245 Portuguese verb lemmas and 8816 verb senses (defined as different frame names for the same lemma). In a second round of manual revision, the frame candidates were checked and, if necessary, corrected. Particular care was taken to check prepositional arguments and reflexives, where different syntactic realizations in the two languages (e.g. intransitive as reflexive, or transitive as pp argument and vice versa) sometimes caused errors in the matches suggested by the harvesting program. In a third step, PALAVRAS' parsing lexicon was used to add missing lemmas and valency realizations. In these cases frames, argument roles and their semantic selection restrictions had to be written from scratch, or adapted/expanded from another frame for the same lemma. Finally, noun and adjective frames were specified manually for all cases where the parser lexicon specified a (prepositional) valency argument.

Using the parser lexicon - rather than e.g. a corpus-based frequency list - to decide which lemmas to include has three advantages. First, the method ensures broad coverage, also for rare words. But as important, the parser lexicon provides valency patterns that can be used as syntactic skeletons for the frames both by the lexicographer and by a live frame tagger adding frames and roles to syntactically annotated input.

3 Lexicon size, coverage and granularity

The verb part of PFN-PT currently contains 7,273 verb lemmas, with manually revised or assigned frames. All in all, these cover 12,835 valency frames (i.e. with differences in either valency, argument roles or semantic slot-filler conditions) and

10,612 different lemma+frame combinations (verb senses, sense frames). On average this amounts to 1.77 valency patterns and 1.46 frame senses per lemma, with a Zipfian distribution, where some frequent construction verbs and polysemic verbs needed 10 or more entries, while 60% had only one valency type and 70% only one sense frame. The top-scoring lemmas were *dar* (30 entries), *fazer* (27), *passar* (22), *levar* (19), *ficar* (19), *ter* (18), *estar* (18), *tornar* (15), *ir* (15), *sair* (14), *contar* (14), *chamar* (14), *trabalhar* (13), *pegar* (13), or - in terms of verb senses - *dar* (21), *fazer* (20), *passar* (14), *levar* (14), *estar* (14), *ter* (13), *ficar* (13), *ir* (11), *tomar* (10), *tirar* (10). As to coverage, it is important to stress that virtually all verb lemmas (over 7,000) in the parser lexicon (as well as all valency-marked nouns and adjectives) were assigned at least one frame and that corpus evaluation (chapter 5) indicates a raw lexical failure rate as low as 1-4% for lemma types and 0.3% for tokens[4]. By comparison, the above-mentioned SALSA resource for German, albeit more refined, more revised and less "bootstrapped", only contains about 1000 unique lemma-frame types (REHBEIN et al. 2014).

For Portuguese, the obvious comparison reference is FrameNet Brasil (FB, TORRENT; ELLSWORTH, 2013). Judging from its online database, the project pursues an example-based and domain-driven approach, where coverage progresses in an uneven fashion, good in one area (sports/tourism), but insufficient at a general level. Many verbs have no frame at all, or - worse for running-text applications - a specialized but rare frame rather than the most frequent one. Thus, a lookup of a randomized 1% sample of PFN-PT's verb list in FrameNet Brasil revealed that 61 out of 72 verbs (84.7%) had no frame at all. Detailed inspection of one common frame, 'adquirir' (fn:obtain in PFN-PT), found 4 verbs and 1 adjective in FB, but 25 verbs and 2 adjectives in PFN-PT.

[4] Lexical coverage is otherwise an issue even for English. Thus, Palmer & Sporleder (2010), comparing SemEval data with Framenet data, found that 3.4% of lexical units and 12.1% of frames from the former were not found in the latter. In terms of training data, the gaps were even more pronounced, 6.9% missing senses and 26.0% missing verbs.

The prototypical verb 'obter' (obtain) was not in FB's 'adquirir' frame set, but had only the less general 'obter_documento', as well as 'Posse', the latter being problematic because, according to the website, it asks for a POSSESSOR subject, whereas 'adquirir' needs a RECEIVER subject, the latter being a better match for the core meaning of *obter*. Another coverage difference between FB and PFN-PT concerns valency variation, i.e. meaning differences triggered by syntax. Thus, the verb *recuperar* (reclaim) did make it into the 'adquirir' set, but FB did not have an entry for the health domain meaning 'to recover' linked to the verb's reflexive form *recuperar-se*. Finally, the very fact that FB has a specialized coverage of the sports domain makes it less useful for ordinary text, because coverage is less balanced than in a full-lexicon parser extension like PFN-PT. The verb *marcar* (mark), for instance, has only domain-specific frames in FB, 'Ações do árbitro', 'Infrações', 'Jogadas_interativas' and 'Jogadas_pontuadas', all of which would count as errors if used to annotate marcar when used with non-sports meanings, such as 'mark' (a surface), 'decide on' (a meeting), 'identify/show' (time). PFN-PT, for its part, has to make do with the general frames in sport texts, too, e.g. fn:obtain for 'scoring goals' or fn:accompany for 'tagging a player', but in a general annotation context, a lack of precision is preferable to wrong or missing labels. Also, apart from the domain information in the frame name, precision is not necessarily better with the specialized frames. Thus, the frame 'Jogadas interativas' lumps *marcar* with other special meanings of e.g. *atacar* (attack), *bater* (beat), *combinar* (combine) and *cruzar* (cross) that arguably cover meaning distances as large as the one between the general *acompanhar* (accompany) and the specific *marcar* (tagging a player).

Out of the 494 different verb frame categories[5] available in the Danish framenet scheme, almost all (482) also ended up being used for Portuguese, too[6]. On top of these, we introduced 5 new frames: fn:repeat, fn:return, fn:path, fn:enough and fn:alter_soc. In order to capture further nuances, 230 frame senses were used as secondary frames senses in combination with a primary frame sense, yielding about 640 different combinations of these "atomic" frames. In some cases, the two frame senses have almost equal weight, either because each could be used on its own (insistir [to insist] fn:declare&demand), because they combine different aspects such as method and domain (empanar [to bread / coat with breadcrumbs] fn:cover_ize&prepare_food), or because they address two parts of a complex action (desenterrar [to unearth] fn:poke&take). Complex frames can also be used to capture lexicalized information such as aspect/aktionsart or polarity (more/less, better/worse), where there is no separate frame in the inventory to make the distinction. Sometimes Portuguese prefixes indicate a secondary frame in a systematic way (*re-* = &repeat, &return, *des-* = &fail)

estrear [to debut] - fn:start&perform
reabrir [to reopen] - fn:open&repeat
desgarrar-se [to stray] - fn:orient&fail
baratear [to become/make cheaper] - fn:cost&decrease

In principle, the same frame inventory can be used for both verbal and nominal predications, but obviously the distribution will be different. PFN-PL contains two types of nominal frames, static and dynamic. Static frames exist for both nouns and

[5] A smaller set of 200 frame senses exists with a hypernym-mapping from the more fine-grained set. This smaller set was meant to facilitate cross-language comparisons, syntactic-contextual uses and parser training.
[6] For a full overview, definitions and examples, see https://framenet.dk/verbal_prototypes.pdf

adjectives and consist of valency-based, manual entries in the framenet lexicon. There are currently 849 noun frames (for 682 lemmas) and 407 adjective frames (for 376 lemmas). Given the much smaller number of lemmas, the spread of frames is smaller than for verbs, with 296 different frame senses for nouns, and 178 for adjectives.

Dynamic frames are computed on the fly based on the morphological analysis of de-verbal nouns and participle-based adjectives in an actual annotation run. In this process, prepositional arguments are taken over as-is from the corresponding verb frame. For noun frames, the subject of intransitive verbs or the object of transitive verbs is turned into an argument with the preposition "de".

> *fertilizar* [to fertilize sth.] (transitive) --> *fertilização de* OBJECT [fertilization of]
> *capotar* [overturn] (intransitive) --> *capotamento de* SUBJECT [rollover]
> *atribuir X a Y* [to attribute X to Y] --> *atribuição de X a Y* [atribution of X to Y]
> *categorizar X como Y* [categorize X as Y]-> *categorizado como Y* [categorized as Y]

The derived frames inherit their semantic roles from the verb frame, and in all cases semantic slot-filler restrictions are maintained - for instance, 'vehicle' AG(ent) for the subject of *capotar/capotamento*, or 'human' REC(eiver)for the Y in *adjudicar/adjudicação*.

4 Frame role distinctors: Valency, syntax and semantic class

The distinctional backbone of all frames in PFN-PT's frame is provided by syntactic valency patterns, such as <vt> (monotransitive), <vdt> (ditransitive), <com^vrp> (reflexive verb with a prepositional argument headed by the preposition “com”, e.g. *aborrecer-se com* [to abhor]). For any given verb lemma, each of these

valency patterns is assigned at least one, and possibly more[7], verb senses, each corresponding to a separate semantic frame. There can be different surface realizations of a sense/frame, so more than one syntactic pattern or semantic slot order may trigger the same verb sense / frame name, but two different verb senses will almost always differ in at least one syntactic or semantic aspect of at least one of the arguments governed. Therefore, almost all senses can in principle be disambiguated exploiting the tags and dependency links provided for each argument by a deep syntactic parser like Portuguese PALAVRAS (BICK, 2014).

Though the frame inventory and role granularity of PFN-PT is modeled on the Danish framenet, we decided to make an important change in notational conventions, extending the shorthand system suggested by Bick (2017) for noun frames to cover the main, verbal lexicon, too. Thus, for each of the almost 13.000 verb sense frames, a list of arguments is provided in a single, composite tag ready to be used by CG rules. For instance, it is possible to differentiate at least seven different meanings[8] of the verb *apontar*, each with its only valency and frame patterns (<FN:...).

meaning: *point out (a person/thing)* valency: <vt> (monotransitive)
<FN:identify/S§AG'H/O§TH'H>
meaning: *point out (a fact)* valency: <vt> (monotransitive)
point out (that ...) valency: <vq> (que-clause)
<FN:emphasize/S§SP'H|sem/O§SOA'fact|ac|f|fcl>
meaning: *sharpen (a pencil)* valency: <vt> (monotransitive)
<FN:shape/S§AG'H/O§TH'"lápis"> # spidse
meaning: *appear (moon, sun, dolphin)* valency: <vi> (intransitive)
<FN:appear/S§TH'Lstar|A>

[7] In PFN-PL, 890 cases multiple verb senses share the same valency frame - in other words, in 7.6% of the verb entries, verb senses cannot be disambiguated on the grounds of syntactic function and form alone, but need help from semantic (noun) classes.

[8] For etymological reasons, the English verb point shares part of this polysemy (albeit with its own phrasal particles). In other languages, however, there may be no overlap at all.

meaning: *point at* valency: <para^vp> (PP-monotransitive)
<FN:gesture&identify/S§AG'H/P-para§DES'cc>
meaning: *point (a gun) at* valency: <a^vtp> (PP-ditransitive)
valency: <para^vtp> (PP-ditransitive)
<FN:orient/S§AG'H/O§TH'cc|tool|V/P-a|para§DES'H|L>
meaning: *appoint (a person) as* valency: <como^vtp> (PP-ditransitive)
<FN:appoint/S§AG'H/O§BEN'H/P-como§ROLE'H>

In this scheme, frames match valencies in terms of arity, so a monotransitive frame will have two role arguments. Arguments are slash-separated ('/') and contain themselves three information fields:

1. Syntactic function (**S** - subject, **O** - direct object, **D** - dative object etc.)
2. Thematic role (e.g. **§AG** - agent, **§TH** - theme, **§DES** - destination)
3. Semantic slot filler conditions (e.g. '**H**' - human, '**food**', '**act**') for argument nouns

For prepositional arguments, the preposition in field 1, for instance P-em for *insistir em* (to insist on). Syntactic form restrictions for material other than 'np' (noun phrase), field 3 is used (e.g. 'fcl' - finite clause, 'icl' - non-finite clause, 'num' - numeral).

In our evaluation of PFN-PL, 29% of verb lemma types were sense-ambiguous in terms of frame names. In 890 cases, valency patterns were ambiguous with regard to verb frames (senses + arguments), with a maximum of 8 frames for one valency pattern. However, in some of these, the difference only concerned roles and slot fillers, not the verb sense. Discounting these cases, 731 valency patterns (covering 661 verb lemmas) were verb-sense ambiguous, with only 133 valency patterns (128 verb lemmas) being affected by more than one such ambiguity. This means that almost 90% of all verbs could in theory be assigned a unique sense using syntactic argument structure alone (i.e. syntactic function, form and dependency), that is, with input from

a simple morphosyntactic parser without lexical semantics[9]. For the rest, semantic slot-filler clues are needed.

The most common valency patterns (table 1) were monotransitive with accusative (S/O), intransitive (S) and monotransitive with a PP argument.

Table 1 – Valency patterns.

Argument inventory	of all frames in lexicon
S/O	55.0 %
S	16.7 %
S/R	10.8 %
S/P	5.9 %
S/O/P	4.4 %
S/R/P	4.3 %
S/SA	0.6 %
S/D/O	0.5 %
S/O/OC	0.2 %
S/O/OA	0.2 %
S/D	0.2 %
S/SC	0.2 %

(S=subject, O=object, P=pp argument, REFL=reflexive, D=dative object, SA=subject-related adverbial argument, OA=object-related adverbial argument, A-INC=verb-incorporated adverbial

PFN-PT uses 44 "atomic" semantic roles (or case/thematic roles, FILLMORE, 1968). The role inventory closely resembles that of the original PALAVRAS role annotator (BICK, 2007), with a few additions, such as §COG (cognizer), §ASS (asset), §POSS possesor, as well as a few roles targeting clausal arguments in particular: §SOA (state-of-affairs), §ACT (action) and §EV (event). The number of roles is similar to the one used in the semantic level of treebanks such as the Prague Dependency Treebank (BÖMOVÁ et al., 2003) and the Spanish 3LBLEX/3LBSEM project (TAULÉ et al., 2005). This kind of medium category granularity is big enough to allow useful distinctions,

[9] Note that these numbers refer to types, not tokens. Obviously, in running text, the most frequent verbs are the most ambiguous ones.

but small enough for generalizations and to avoid simply duplicating lexical information. Also, a limited level of granularity is easier to disambiguate and easier to define for human consensus, thus lending itself particularly well to automatic corpus annotation.

Where necessary, these roles can be combine, e.g. §AG-EXP (agent & experiencer) for the subject of "zuhören" (listen). All in all, 61 such combinations occur in the lexicon, the most frequently used being §AG-EXP (e.g. watching), §TH-COM (e.g. associating) and §AG-PAT (especially reflexives wit agents causing changes in themselves). In addition, a number of adverbial roles is used that are never valency-bound and therefore do not occur in any frame, but can be used by the frame tagger rules to specify the function of free adverbials and adverbial subclauses (such as §COND for conditional subclauses). The 44 roles are far from evenly distributed in running text. Table 2 shows role frequencies at the token level, for News texts[10].

Table 2 – Semantic roles.

	Semantic Role	surface verb arguments %	secondary verb args %	free (adjunct) adverbials %	noun arguments %
§TH	Theme	24.2	23.6	0.4	28.8
§AG	Agent	12.7	26.4	0.6	6.0
§ATR	Attribute	12.2	10.4	0.5	1.7
§PAT	Patient	5.7	5.4	-	11.4
§ACT	Action	5.0	3.1	1.4	4.1
§SOA	State-of-affairs	3.3	1.6	-	1.4
§SP	Speaker	3.3	3.6	0.1	1.8
§MES	Message	3.2	2.3	-	1.4
§COG	Cognizer	3.1	5.1	-	3.7
§EV	Event	2.9	3.5	1.4	3.2
§RES	Result	2.1	2.4	0.1	2.7
§REFL	Reflexive	2.0	-	-	

[10] 24483 random sentences from the Leipzig corpus collection (cp. chapter 6, evaluation).

§LOC	Location	1.9	1.5	16.7	5.9
§BEN	Beneficiary	1.7	1.5	2.0	2.8
§MNR	Manner	1.6	-	8.8	
§DES	Destination	1.6	0.7	3.8	2.0
§REC	Receiver	1.5	1.0	0.9	1.3
§INS	Instrument	1.0	0.4	0.9	0.7
§EXP	Experiencer	0.9	1.7	0.1	0.4
§CAU	Cause	0.9	0.6	2.6	0.8
§LOC-TMP	Temporal Location	0.7	-	21.3	1.7
§ORI	Origin	0.7	0.6	2.1	0.7
§VAL	Value	0.4	0.7	0.7	0.7
§PART	Part	0.2	0.7	-	0.1
§DON	Donor	0.2	0.6	-	-
§FOC	Focus marker	-	-	8.6	(4.1)
§CIRC	Circumstance	-	-	7.0	0.4
§FIN	Purpose	-	-	5.8	1.7
§DES-TMP	Temp. Destination	-	-	1.7	-
§EXT	Extension	-	-	1.5	0.1
§TP	Topic	0.6	0.3	0.7	2.7
§ASS	Asset	0.3	0.2	-	1.8

Other roles: §STI - Stimulus; §REFL - Reflexive; §PATH - Path; §EXT-TMP - Duration; §ROLE - Role; §CONC - Concession; §COMP - Comparison; §HOL - Whole; §POSS Possessor; §CONT - Content; §ID - Identity; §ATR-RES - Attribute-result; §COM - Co-role; §INC - Incorporated

At the clause level, §TH (theme), §AG (agent) and §ATR are the most common argument roles, while §LOC (location), §LOC-TMP (temporal location) and §MNR (manner) dominate adverbial adjunct roles, joined by §PAT (patient) at the noun phrase level (last column). For primary dependencies, i.e. with verb and argument in the same clause (column 3), §ACT (action) and §SOA (state-of-affairs) are relatively common because they are typically describe the function of object clauses relative to the main clause verb. For secondary, cross clause, dependencies (4th column), typical subject roles, in particular §AG, is more frequent than otherwise because of the non-surface subjects of infinitives and relative clause.

For some roles, there is a very tight link to a syntactic function. Thus §ID (identity) is closely linked to the function of apposition (@APP), while subject and object complements (@SC and @OC) carry exclusively attributive roles, §ATR and §ATR-RES (attribute-result). The same is obviously true of syntactic "dummy" roles without independent semantic information: §PRED (predicator, for top verbs), §REFL (reflexive) and §INC (verb-incorporated). The latter is used for binding particles in verb chains and in support verb constructions, where the semantic weight and - to a certain degree - valency reside in a nominal element, typically a noun that syntactically fills a (direct or prepositional) object slot, but semantically orchestrates the other complements. While our input parser, PALAVRAS, already marks verb chain particles at the syntactic level (@PRT-AUX<), it assigns noun incorporates an ordinary syntactic tag (@ACC), which therefore needs lexical treatment in the frame lexicon to allow assignment of a special sense to the support verb.

ter ***que/de*** + INF (must)	<fn:must>
botar ***fogo*** *em* (set s.th. on fire)	<fn:burn&start>
dar ***certo*** (succeed)	<fn:succeed>
dar ***cabo*** *de* (put an end to)	<fn:destroy>

One could argue that the real frame arguments (like the §PAT - patients - of *botar fogo em* and *dar cabo de* should be dependency-linked to the §INC constituent, and the frame class marked on the latter, but for consistency and processing reasons we decided to mark the frame name on the verbal element of such support constructions, with a corresponding semantic dependency link from the verb.

PP incorporates are handled in a similar way, with both the preposition and its argument listed in the frame pattern. Because of the above-mentioned principle of semantic dependencies, the §INC role with fall on the nominal part of the PP, blocking assignment of other (adverbial) roles.

*ter **a ver** com* (have to do with) <fn:relate>
*por **em causa*** (jeopardize) <fn:risk>

5 Automatic frame annotation

In our scheme, frame annotation is achieved by checking the elements of a potential frame against the feature sets of syntactic arguments, exploiting morphosyntactic and semantic information already assigned and disambiguated by the PALAVRAS parser. For instance, the verb *contar* has a number of different meanings, each sense corresponding to one or more different valency frames:

<fn:tell> *contar que ... contar uma história* (to relate that ... a story)
<fn:math> *contar ovelhas* (count sheep)
<fn:contain_have> *contar 300,000 habitantes* (have 300,000 inhabitants)
<fn:assume&rely> *contar com* (count on somebody or something happening)
<fn:plan> *contar viajar para ..., contar em viajar para ...* (plan to travel to ...)
<fn:matter> *a sua ajuda conta muito para mim* (your help means a lot to me)

Since only one of these - a "tell" frame - has a slot for a finite clause complement (fcl), the presence of an object clause will trigger this frame, harvesting a speaker role (§SP) for the subject (S), a message role (§MES) for the subclause and a receiver role (§REC) for a possible dative object (D), as specified by the following frame template example:

<FN:**tell**/S§SP'H/D**§REC**'H/O**§MES**'fcl>

The technical implementation of this annotation mechanism uses the Constraint Grammar formalism (CG, BICK; DIDRIKSEN, 2015), also used by the frame annotator's input parser, PALAVRAS. Apart from the obvious advantage of notational compatibility (tags and dependency), CG also has the advantage of allowing very

complex contextual rules including the use of tag set operators, variable unification, named relations and regular expressions, to name just a few. This makes it possible to unify frame constraints with actual tags and relations, while at the same time performing additional checks on the context.

As a preparatory step, all possible frames for a given verb lemma are mapped as template tags on all main verbs (cp. the 'tell' frame above). Also, some minor dependency adaptations are performed to make the syntax tree more "semantic", such as making sure that the subject is linked to a main verb rather than an auxiliary and marking certain nouns as dependency-"transparent" (e.g. *a maioria de*, *um monte de*, *uma copa de*, *parte de*, *um dos/das*, *uma sorte de*). Also, we introduce secondary, semantic dependency links for implicit arguments in relative or non-finite clauses, and add shadow pronouns to subject-less but person-inflected verbs, where a semantic category (especially ±HUM) can be deduced.

In the annotator grammar itself, four main rule types are used:

1. frame template tag selection
2. frame template tag removal
3. role instantiation
4. mapping of free roles

Template removal is a simpler task than template selection, because a single mismatch is enough to trigger the former, while the latter optimally needs as many frame element matches as possible. Therefore, matching rules are ordered in heuristicity batches with high-arity valency matches and full semantic slot matches coming before partial valency matches and hypernym or semantics-free slot matches. Safest are lexical matches, where word forms (e.g. verb incorporates or prepositions) are mentioned as such in a frame template. Thus, in the *contar* example above, the

'assume&rely' frame can be safely discarded or selected based on the presence of a PP argument with th preposition *com*:

<FN:**tell**/S§SP'H/D**§REC**'H/O**§MES**'fcl>

Another relatively safe method are syntactic mismatches - at least as long as the parser gets syntactic functions right, with correct dependency links. Thus, subject and object complements are obligatory arguments and the corresponding frames can be safely discarded in their absence. For direct objects, the situation is more complicated because of passivization and ellipsis, so more contextual checks are needed. Also, there can be a fail-safe condition asking for the existence of a competing, similar frame with lower valency (e.g monotransitive instead of ditransitive). In Germanic languages, impersonal verb frames could be chosen or discarded based on formal subjects, but in Portuguese, there is ambiguity here, because the subject-less 3. person singular form used for impersonality can also just mean an omitted subject. Therefore, while weather verbs (snowing, raining etc.) are safe, because they do not have other meanings, the verb ter and haver (to have) are quite difficult for the frame annotator, because without a subject, a lot of context and semantics is needed to make the distinction between on the one hand <fn:exist> (*tem/há* X - there is X) and, on the other, ordinary having frames ('have', 'have part', 'have attribute'), not to mention auxiliary and support verb constructions.

The most important rules, however, are the ones capable of differentiating frame templates with identical syntactic/valency skeletons. In this scenario, semantic slot filler information is used, exploiting the so-called semantic prototype tags that PALAVRAS assigns to nouns and proper nouns. Although these are lexical tags, and only partially disambiguated by the parser, ambiguities rarely overlapped with frame ambiguities found in a given verb. The noun ontology has about 200 categories, e.g. <tool>, <food>, <act> (action) or <mon> (money). Categories are organized in a shallow

hierarchy, with lower-case or hyphenated subcategories. 'H', for instance, means +HUM (human) and occurs in tags such as <Hprof> (human professional), <Hfam> (family member), <Hideo> (ideological human) etc., while <sem-r> (readable), <sem-c> (concept), <sem-s> (sayable) etc. all belong to the class of semantic products, sharing the 'sem-' prefix.

The frame templates specify - for all non-trivial function/role pairs (e.g. 'S§AG' for subject agent) - at least one semantic slot filler, drawing on the categories discussed above. The frame tagger's matching mechanism proceeds from safe to unsafe by trying rules with many, or more specific, conditions first, followed by underspecified matches, using heuristic defaults as a last resort to decide unresolved ambiguities. In order to handle category fuzziness or overlaps, creative language use or just incomplete slot filler information, the frame tagger grammar uses "umbrella category" matches. Here 17 semantic hypernym sets, e.g. 'HUMAN', 'THING', 'PLACE' etc. At this intermediate level, two categories (from a frame template and a sentence token, respectively) will be considered a match if they are part of the same hypernym category. Applying these principles, the following (simplified) scheme for rule ordering/prioritization was applied:

- 2 or more syntactic slots with a full semantic match
- 1 slot with a full match, 1 with an "umbrella category" match
- 2 or more "umbrella" matches
- 1 slot with a full match
- 1 slot with an "umbrella" match
- syntactic match, slot(s) marked <all> or <cc>
- longest syntactic match (e.g. ditransitive match beats monotransitive match)

As long as roles manifest as surface constituents with a direct dependency link to the frame-evoking verb, the method is quite robust, since it can simply draw on the existing dependency parse. However, in many cases, the role-carrying constituent has no, or only pronominal, surface representation in the clause itself. In Portuguese, this

is the case for e.g. subjects of infinitive clauses and antecedents of relative clauses. Thus, in the example sentence, the word Eltern (parents) functions - in two different subclauses - as both an object-theme in an 'exist' frame and as a subject-cognizer in an 'allow' frame. Here, in order to be able to check slot-filler conditions and to assign roles, we first introduced secondary dependency relations using special, relation-mapping CG rules. These additional dependency relations can then be drawn upon by variants of the ordinary frame-matching rules.

In a vertical, one-word-per-line CG notation, the frame-tagger adds <fn:sense> tags on verbs (red), and §ROLE tags on arguments (blue). The example contains four ordinary main verb frames with one or more role dependents each, one auxiliary frame and two noun frames with a single dependent. Roles relate to the next-coming frame upward in the dependency tree, bypassing prepositions and transparent nouns[11]. Primary dependency arcs are shown as #n->m ID-links, secondary dependencies are marked as R:c- (child) and R:p (parent) relations (green). The latter are matched by a secondary role link on the dependent, marked as R:sd- (semantic dependency) relations. In the example, *unidades de infantaria* has such a secondary subject link to the verb of the relative clause, *faltam*, with the role of §TH (theme).

```
O               [o] <artd> DET M S @>N #1->2
Secretário-geral        [secretário-geral] <Hprof> N M S @SUBJ> §AG #2->10
de              [de] <sam-> <np-close> PRP @N< #3->2
a               [o] <-sam> <artd> DET F S @>N #4->5
ONU             [ONU] <org> PROP F S @P< #5->3
$,              [$,] PU @PU #6->0
Butros          [Butros] <hum> PROP M S @APP §ID #7->2
Butros-Ghali    [Butros-Ghali] <hum> PROP M S @N< #8->7
```

[11] If desired by a corpus user with other annotation conventions (e.g. Universal Dependencies), or if needed for the use with a specific corpus search tool, it would take only a couple of rules to change this dependency convention into direct semantic links, with prepositions as dependents of nouns, and the role-carrying noun getting a direct dependency arc to its frame-carrying semantic head.

$, [$,] PU @PU #9->0
deixou [deixar] <fn:leave> <mv> V PS 3S IND @FS-STA §PRED #10->0
ontem [ontem] <atemp> ADV @<ADVL §LOC-TMP #11->10
Luanda [Luanda] <civ> PROP F S @<ACC §ORI #12->10
com [com] PRP @<ADVL #13->10
a [o] <artd> DET F S @>N #14->15
promessa [promessa] <act-s> <fn:promise> N F S @P< §CIRC #15->13
de [de] PRP @N< #16->15
que [que] <clb> <clb-fs> KS @SUB #17->25
até [até] PRP @ADVL> #18->25
a [a] <sam-> PRP @P< #19->18
o [o] <-sam> <artd> DET M S @>N #20->21
fim [fim] <fn:end> <temp> N M S @P< #21->19 §TMP-DES
de [de] PRP @N< #22->21
Agosto [agosto] <month> N M S @P< §TMP-LOC #23->22
deverão [dever] <fn:must> <aux> V FUT 3P IND @FS-P< #24->16
chegar [chegar] <fn:reach> <mv> V INF 3P @ICL-AUX< §ACT #25->24
a [a] PRP @<SA #26->25
Angola [Angola] <civ> <*> PROP F S @P< §DES #27->26
as [o] <artd> DET F P @>N #28->29
unidades [unidade] <HH> N F P @<SUBJ §AG #29->25 ID:29 R:c-subj:33 R:sd-TH:33
de [de] PRP @N< #30->29
infantaria [infantaria] <HH> N F S @P< #31->30
que [que] <clb> <rel> <hum> SPEC M S @SUBJ> #32->33
faltam [faltar] <fn:lack_itr> <mv> V PR 3P IND @FS-N< §ATR #33->29 ID:33 R:p-subj:29
para [para] PRP @<ADVL #34->33
se [se] PERS F 3S/P ACC @ACC>-PASS §REFL #35->36
completar [completar] <fn:complete_finish> <mv> <passive> V INF 3S @ICL-P< §FIN #36->34
a [o] <artd> DET F S @>N #37->38
Missão=de=Verificação [Missão=de=Verificação] <project> PROP F S @<SUBJ §PAT #38->36
$. [.] PU @PU #39->0

6 Evaluation and discussion

For a quantitative evaluation of the system, we automatically annotated 24,483 random sentences of news text (610,620 tokens12) from the Leipzig corpus collection[13]. The sample contained 2,222 different verb types[14], or 64,313 tokens[15]. 213 tokens with verb tags had no frame. However, manual inspection revealed that among these there were 28 misanalysed foreign words, 22 spelling errors and 27 unrecognized correct spelling (mostly "frequentar", that had been listed with the older, pre-reform 'ü'). 25 were morphological analysis errors and 21 morphological disambiguation errors. Only 90 (44 types) were regular frame failures, amounting to 0.1% of verb tokens, or 2% of verb types. A recount for a larger sample (2 million words) yielded, as expected, the same 0.1% for tokens, but a higher percentage (4%) for verb types. This can be explained by Zipf curve characteristics - as the corpus grows, it covers logarithmically more verb types, but in this so-called "tail" of the Zipf curve rare is made up of rare words, a phenomenon that will affect types, but not tokens.

At the type level, the annotated sample corpus contained 1.38 senses per verb, close to the ratio of 1.46 found in the lexicon. 75.8% of verb types in the corpus occurred with only one frame sense. However, frame sense ambiguity is very unevenly distributed, with a lot of ambiguity at the token level. Thus, in terms of running verb tokens, the grammar had a substantial disambiguation load, with 4.94 frame senses per verb, and only 23.4% 1-sense verb tokens. For the assignment of complete frames, i.e. including argument roles, the disambiguation load is even higher, because not just

---

[12] Raw count with space-separated tokens, including punctuation. Due to fusion of multi-word-expressions (e.g. names), PALAVARAS annotated token count had 38,096 fewer tokens.

[13] https://wortschatz.uni-leipzig.de/en/download/Portuguese

[14] This is the corrected count, after subtraction of 82 types caused by spelling errors etc. The passive auxiliary use of ser and *ficar* was excluded, because by matter of design there is never assigned a frame in this case.

[15] This count is the annotation count, which is only an approximation of the true count, due to possible annotation errors. The type count, on the other hand, is reliable, because the verbs either received frames from the lexicon or were checked by hand. It is unlikely that a verb token should be consistently, in all instances/contexts, sim-disambiguated as a different part-of-speech.

senses, but specific valency frames need to be disambiguated for this task. Here, token ambiguity was 7.36 (valency) frames per verb, and only 12.3% of verb tokens were unambiguous.

For the 12 most frequent verbs, together accounting for 17.4% of all verb tokens in the corpus, frame sense ambiguity is shown in table 3. The numbers show that most lexicon senses were also present in the corpus, but for some verbs (*dar*, *fazer*, *ficar* ...), there is an extra disambiguation load for rare senses or support verb constructions.

Table 3 – Frame sense ambiguity.

Verb lemma	frame senses in corpus	frame senses to be disambiguated (lexicon)
passar	14	14
dar	14	21
levar	13	14
fazer	12	20
estar	12	14
ter	10	13
ir	9	11
ficar	9	13
trabalhar	7	7
tornar	7	8
tomar	7	10
tirar	7	10

Since matching the semantic categories of arguments with slot-filler conditions is very important for frame sense disambiguation, missing surface arguments can be a problem. Salomão (2009, p. 177), for FB, distinguished between three types: (a) definite (subject recoverable from matrix clause), (b) indefinite (e.g. object missing or unclear) and (c) constructional (e.g. passives). Accordingly, infinitives, passives, relative clauses, ellipsis and incomplete sentences all get special treatment in our frame

annotator to increase the system's robustness[16]. For instance, secondary dependencies are created for noun antecedents of relative clauses and subject antecedents of infinitives[17]. However, not every slot-filler can be recovered. Thus, passive participles (type c), though their object slot can be filled with a subject or noun antecedent, usually have an empty subject slot, as the role of agent-of-passive is rarely specified. In order to assess the size of the empty-slot problem, we ran a count of surface expression for the main argument types (table 4).

Table 4 – Surface expression of slot-filler arguments.

Argument type	filled slots (primary dep.)	filled slots (secondary dep.)	filled slots (all dep.)
SUBJ (subject)	51.3 %	17.4 %	62.1 %
SUBJ + 1./2. person inflection			66.2 %
ACC (object)	77.0 %	8.6 %	82.5 %
SC (subject complement)	86.1 %	1.3 %	86.1 %
PP (prepositional argument)	100 %	0 %	100 %

As can be seen, almost half of all verbs had no direct surface subject (2nd column), and over a third had neither a direct or indirect link to a subject (4th column), even if 1./2. person inflection is counted as a +HUM (human) slot-filler for an otherwise missing subject. For direct objects, there are more filled slots, but still a fair percentage of frames had to be chosen without slot fillers. The 100% expression of prepositional slots is due to the fact that these were never the shortest frames for a given lemma, and could be safely discarded if not matched. In other words, surviving PP-frames were

[16] One might want to add coordination to this list, but the situation is really rather the opposite: By attaching conjuncts in parallel to the same head, the latter's frame slots will actually get 2 or more chances of slot filling rather than just one.

[17] Portuguese infinitives can, in fact, carry some person-number inflection, or even govern surface subjects, but subject-less, uninflected infinitives are still the most common.

chosen precisely because the PP was matched. For all argument types it should be noted that the percentage of real semantic matches is somewhat lower, since on average 9.2 % of all verb arguments were pronouns - a part of speech with only limited semantic content (±HUM for some). Again, this problem is more pronounced for subjects (15.9%) than for direct objects (12.7%) and PP arguments (4%).

To judge the performance of verbal part of the frame tagger, we evaluated a random set of 300 sentences (about 6000 raw tokens) of mixed Portuguese news from the Leipzig corpus collection, harvested in 2019 from tweets from the minority-filtered subcorpus, with 9054 tokens. All in all, the parser tagged 739 words as main verbs and 165 as auxiliaries, only three of which (0.7%) were POS errors. Our frame tagger knew all verb lemmas in the sample and assigned frames to all of them, with the exception of the 47 *ser* passive auxiliaries, that PFN-PT by design doesn't consider as frame carriers. The other auxiliaries were spread over 13 types[18] and 16 different frames. Since auxiliaries are fairly easy to annotate (only 1 frame error), evaluation concentrated on main verbs. The frame tagger found a correct frame sense for 92.4% of the correctly tagged verbs (table 5), with 3 spurious frames (92.0% precision). Inspection of the 56 errors revealed 9 cases, where the reason was faulty input - 2 spelling errors and 7 parser errors (1 POS error and 6 syntactic). This indicates that in a perfect world with perfect input, the performance of the frame tagger on its own could be a couple of percentage points higher.

Table 5 – Verb frames: Recall and precision.

	Recall	Precision	F-score[19]

[18] The Portuguese definition of auxiliary is complex and includes more than tense and diathesis, using criteria like pronoun movement. There is no 100% consensus on the matter in the literature. In addition to auxiliaries proper, PALAVRAS annotates other verb chains with light verbs, all of which get individual frames in PFN-PT.

[19] This is the F1 measure, defined as 2 * R * P / (R + P).

main verb frames	92.4 %	92 %	92.2
auxiliary frames	99.2 %	100 %	99.6 %

Automatic frame tagging is not yet a standard part of NLP pipelines, and systems and approaches are difficult to compare due to different tag granularity and domain, not least across languages. However, even without another Portuguese frame tagger to compare with, these figures constitute an encouraging result. Thus, although the "weak", inspection-based evaluation method makes a direct comparison impossible, performance compares favorably with e.g. an early English baseline for rule-based frame tagging, Shi & Mihalcea (2004), with an F-score of 74.5%. The task of word sense disambiguation (WSD) can also be a relevant comparison, e.g. the German SHALMANESER tagger (BURCHARDT et al., 2009), with 79% accuracy. More recently, in the machine learning camp (ML), Hermann et al. (2014) report F=70.1% for predicate frame identification, and Cai & Lapata (2019), also for German, performed (in-domain) semantic role labeling (SRL) with an F-score of 82.7%, using neural networks. It is interesting that both of the latter exploited syntactic features, as does our own frame tagger. Without using linguistic features, Do et al. (2018) reached a somewhat lower F-Score (73.5) for SRL on the same German test data (CoNLL 2009).

As described above, the PFN-PT tagger also performs SRL. For valency-bound arguments this task is intertwined with frame disambiguation, because the same dependency links and semantic slot filler matches can be used to harvest the relevant roles from a matching frame and assign them to the argument in question. In fact, SRL is often more robust than frame identification. For instance, one verb may have several frames all sharing the same subject role (e.g. §AGent). Time constraints for this paper did not allow a proper SRL evaluation, but an inspection of 15% of the test corpus suggests a recall of 95.4% and a precision of 96.9% for argument roles. This is better than the F-score of 88.6% reported by (BICK, 2007) for a syntax-only approach to

Portuguese SRL, though a direct comparison is impossible because the latter also included adjunct roles. In any case, SRL results can be expected to closely correlate with frame identification, according to Hartmann et al. (2017, p. 475), who compared SRL with automatic frame identification to SRL based on correct "gold" frames, finding that gold-based SRL was stable across text types (domains), while automatic frame identification and SRL varied considerably, but with a stable accuracy ratio of about 3:2 (i.e. SRL accuracy 33% lower than frame accuracy). Given a stable ratio, frame identification results can be used as a rough predictor for SRL, and improvements in the former should manifest in the latter. That said, it is interesting that the SRL accuracy of our system is actually higher than its frame identification accuracy, rather than lower, as suggested by Hartmann et al.'s experiments. Possible reasons for this could be a difference in the size of either the overall role tag set or the number of roles per frame type. Also, quality differences in the underlying parser could mean a lower SRL performance because of wrong argument attachments despite the gold senses. More likely, however, the difference is simply methodological - in our setup there is no separate, statistical step of role assignment. Rather, given an established verb sense and a correct syntactic analysis, role assignment is a 1-on-1 lexical lookup. Even where the system, due to rule hierarchies, has decided on one or more roles first, and deduced a frame sense from these in a second step, wrong roles will lead to wrong frames, too. So it is all but impossible for frame accuracy to be higher than role accuracy. The inverse, however, is possible because several of a verb's frames may share, e.g., the same subject or object role. Here, SRL will be counted as correct even if a wrong frame sense was chosen.

Hartmann et al.'s comparison of in-domain and cross-domain performance also highlights an important difference between our rule- and lexicon-based frame-annotator and other current systems, most of which rely on machine learning (ML). An ML system needs training data and therefore heavily depends on the availability

of human gold annotations for a particular domain. Without (rule-based) matching of syntactic slot information in the lexicon on the one hand and a syntactic-structural annotation on the other, there is no direct way of exploiting a framenet for annotation (beyond lemma-filtering of frame candidates), since lexicon coverage does not automatically translate into training data coverage. In fact, Hartmann et al. claim that a lexicon-free setup, based only on a vector space model of the frame-word-bundle-pairs found in the training data can perform on par with the lexicon-informed standard setup. One interpretation of this is that the lexicon-free setup compensates for coverage problems and unevenly distributed granularity in the frame lexicon, something that is not relevant for our own system, where frame annotation is not based on training data, but on direct structural matches.

7 Sense or metaphor?

Following the reductionist Constraint Grammar philosophy that category disambiguation gains robustness from removing readings incrementally as ambiguity decreases in the context, rather than selecting a category with one single rule, our frame tagger whittles down a lemmas list of potential senses (frames) step by step, and interactively with role assignment. The robustness of this method resides in the fact that it is not necessary to achieve a 100% match for all syntactic, lexical and semantic constraints of a given frame pattern. Instead, the most conflicting options will be removed first, and the least conflicting ones will survive longest. Ideally, the last surviving frame will be the correct one, even if an argument was missing or did not have the expected semantic type, or if it wasn't properly tagged or attached by the parser. Apart from being robust, the resulting "mismatching" frame annotation can be useful in its own right. First of all, mismatches can be semi-automatically exploited lexicographically - to hone the list of semantic selection restrictions for a given

argument or - if that is not enough - add a new sense (and frame) to the head lemma. But rather than adding more and more eclectic senses, there is also the alternative of keeping the sense and treating the mismatch as metaphor. Thus, the attribute '±metaphorical' foreseen by Torrent & Ellsworth (2013) in their discussion of frame annotation layers could be handled intra-sense rather than cross-sense, providing a mechanism for addressing the inherent fuzziness of word senses. For instance, the Portuguese verb *disparar* (generate) initially had three senses in PFN-PT - <fn:activate>, <fn:throw> and <fn:leave> with subject selection restrictions of 'human' for the first two and 'animal/human' for the third. Typical objects were listed as 'tools' and 'things' for the transitive senses, <fn:activate> and <fn:throw>, respectively. These are the senses found in GramTrans' MT lexicon and in the Portuguese online dictionary Dicio[20] and most lexicographers would probably not accept an etymologically motivated reduction to one core sense of 'letting loose' - but what about the opposite, increasing the number of senses? Should there be a fourth sense <fn:increase> for *preços disparando* (soaring prices) and a fifth sense <fn:start> or <fn:cause> for *disparar tensões* (trigger crises), because you can't literally turn on or throw a crisis, and because prices aren't entities that can 'run off'. And if not, which is the closest existing sense? Given the current granularity, our frame annotator will resolve the subject mismatch (a) as <fn:leave> and the object mismatch (b) as <fn:activate>, the former because of the intransitive use and lack of a +HUM subject, the latter heuristically by picking the top transitive meaning:

> (a) 'os preços dispararam' (prices soared)
> (b) 'A recessão internacional poderia disparar novas tensões e picos de crises financeiras' (The international recession could trigger new conflicts and peaks of financial crises)

[20] https://www.dicio.com.br (accessed 28. Oct. 2021).

The suggested new senses do exist in the frame inventory of PFN-PL, and obviously the frame annotator can make the distinction if given the relevant selection restrictions, i.e. <f-q> (quantifiable feature) for the subject in (a) and <event> for the object in (b). So once examples like the above are on the lexicographer's table, the framenet lexicon can be amended[21]. However, letting the system choose the "most matching" frame (rather than no frame) seems to be an acceptable solution where there is no precise frame entry (yet), lending robustness to the system and even preparing it for a future mark-up of potential productive metaphors. In this vein, we experimentally added rules to the annotator grammar that allow it to add a tag for "projected sense". The tag is assigned to mismatching arguments in the case of "fuzzy" or heuristic frame matches, i.e. if an argument matches syntactically, but not semantically, and it will contain the semantic class (or classes) specified for this argument in the last surviving frame. Thus, in example (a), preços will receive a tag <PROJ:A>, meaning the chosen fuzzy frame, <fn:leave>, projected the selection restriction of 'animal' (A), suggesting a metaphorical reading of prices running wild, like an unbridled horse[22].

8 Conclusion and future work

PFN-PT is meant to help closing the semantic gap in the rule-based parsing pipe of PALAVRAS, or - for that matter - any other morphosyntactic parser with capable of producing dependency trees. We have shown that this new framenet resource has the

[21] Some dictionaries, such as the online dictionary Priberam (https://dicionario.priberam.org, accessed 28. Oct. 2021), do list these senses. Still, the differences between dictionaries, as well as constantly evolving language usage, call for a robust backup solution, such as the one suggested here.

[22] The projection 'human' and 'tool' for the subject and object, respectively, in the recession/crisis example would be less illuminating. Though <fn:activate> can work as a hypernym for 'trigger', this is likely not a metaphor, but rather - if not a true sense - a case of to-be-amended slot fillers.

breadth and depth to provide good coverage of the Portuguese verb lexicon on running text in terms of both types and tokens. A preliminary evaluation of the system's frame tagger suggests a satisfactory performance for verbal frame sense disambiguation above the 90% correctness threshold, allowing meaningful automatic corpus annotation, as well as WSD for AI tasks such as machine translation.

However, in spite of its high lemma coverage for verbs, PFN-PT is by no means a finished task. Coverage and granularity should be improved, adding rarer senses and - not least - more verb constructions with incorporated nominal material. Also, the noun and adjective frame lexica have a far lower coverage, where performance could be improved by replacing ad-hoc frames with lexicon frames. Also, a clearer definition is needed of which nouns and adjectives can be frame carriers, e.g.:

- do you need surface dependents, or can a noun carry a frame on its own?
- how to treat participle-derived adjectives without dependents?

Last not least, a more thorough evaluation is needed, ideally after these improvements. This is relevant not least for semantic role labeling, where adjuncts/adverbials, as well as noun frames, need to be included in a future evaluation round. Finally, the theory-based claim of cross-domain robustness should be empirically corroborated.

Obviously, these tasks, as well as future maintenance of the resource, would greatly profit from a participation of the general research community, and hopefully PFN-PT will facilitate and inspire semantic annotation and research projects within the large and growing tapestry that is Portuguese language technology. To this end, the framenet will be open for collaboration and improvements, and the semantic

annotator version of PALAVRAS will be freely available[23] for academic research institutions making use of PFN-PT and its semantic annotation.

References

BAKER, C. F.; FILLMORE, J. C.; LOWE, J. B. The Berkeley FrameNet project. *In*: **Proceedings of the COLING-ACL** (Montreal, Canada). ACL, 1998. p. 86-90. DOI https://doi.org/10.3115/980451.980860

BICK, E. Automatic Semantic Role Annotation for Portuguese. *In*: **Proceedings of TIL 2007 - 5th Workshop on Information and Human Language Technology** / Anais do XXVII Congresso da SBC (Rio de Janeiro, July 5-6, 2007). Rio de Janeiro, 2007. p. 1713-1716.

BICK, E. A FrameNet for Danish. *In*: Proceedings of NODALIDA 2011 (May 11-13, Riga, Latvia). **NEALT Proceedings Series**, Vol 11. Tartu: Tartu University Library, 2011. p. 34-41.

BICK, E. PALAVRAS, a Constraint Grammar-based Parsing System for Portuguese. *In*: BERBER SARDINHA, T.; FERREIRA, T. de L. S. B. (ed.). **Working with Portuguese Corpora**. London/New York: Bloomsburry Academic, 2014. p 279-302.

BICK, E. Swedish-Danish Machine Translation in a Constraint Grammar Framework. *In*: PRZEPIÓRKOWSKI, A.; OGRODNICZUK, M. (ed.). Advances in Natural Language Processing, **Proceedings of 9th International Conference on NLP** (PolTAL 2014, Warsaw, Poland, September 17-19, 2014), Heidelberg: Springer, 2014. p. 216-227.

BICK, E.; DIDRIKSEN, T. CG-3 - Beyond Classical Constraint Grammar. *In*: MEGYESI, B. (ed.): **Proceedings of NODALIDA 2015** (May 11-13, 2015, Vilnius, Lithuania). Linköping: LiU Electronic Press, 2015. p. 31-39

BICK, E. From Treebank to Propbank: A Semantic-Role and VerbNet Corpus for Danish. *In*: TIEDEMANN, J. (ed.): **Proceedings of the 21st Nordic Conference on**

[23] If interested, please contact the author. Please note that unlike PFN-PT itself, the Constraint Grammar rules used by PALAVRAS and its semantic annotator module are not open for editing, and possible modifications should be made by output tag filtering or by adding a new, task-oriented rule set or machine learning module. For commercial use, a separate, paid license is required.

Computational Linguistics (NoDaLiDa 2017, Göteborg). NEALT Proceedings Series Vol. 29. Linköping University Electronic Press, 2017. p. 202-210.

BOAS, H. C. From theory to practice: Frame semantics and the design of FrameNet. *In*: LANGER, S.; SCHNORBUSCH, D. (ed.): **Semantik im Lexikon**. Tübingen: Gunter Narr Verlag, 2005. p. 129 - 159.

BÖHMOVÁ, A.; HAJIČ, J.; HAJIČOVÁ, E.; HLADKÁ, B. The Prague dependency treebank. *In*: **Treebanks**. Springer: Dordrecht, 2003. p. 103-127. DOI https://doi.org/10.1007/978-94-010-0201-1_7

BURCHARDT, A., ERK, K., FRANK, A., KOWALSKI, A; PADO, S.; PINKAL, M. Using FrameNet for the semantic analysis of German: Annotation, representation and automation. *In*: BOAS, H. C. (ed.): **Multilingual FrameNets in Computational Lexicography**: Methods and Applications. Mouton de Guyter, 2009. p. 209-244.

CAI, R.; LAPATA, M. Syntax-aware Semantic Role Labeling without Parsing. *In*: **Transactions of the Association for Computational Linguistics**, 7. ACL, 2019. p 343-356. DOI https://doi.org/10.1162/tacl_a_00272

DO, Q. N.; LEEUWENBERG, A.; HEYMAN, G.; MOENS, M. A Flexible and Easy-to-use Semantic Role Labeling Framework for Different Languages. *In*: **Proceedings of COLING 2018** (Demos). ACL, 2018. p. 161-165

FELLBAUM, C. (ed.). WordNet: An Eletronic Lexical Database. *In*: **Language, Speech and Communications**. MIT Press: Cambridge, Massachusetts, 1998. DOI https://doi.org/10.7551/mitpress/7287.001.0001

FILLMORE, C. J. The case for case. *In*: BACH; HARMS (ed.). **Universals in Linguistic Theory**. New York: Holt, Rinehart, and Winston, 1968. p. 1-88.

FILLMORE, C. J.; BAKER, C. F. Frame semantics for text understanding. *In*: **Proceedings of WordNet and Other Lexical Resources Workshop**. NAACL, 2001

GILARDI, L.; BAKER, C. F. Learning to Align across Languages: Toward Multilingual FrameNet. *In*: **International FrameNet Workshop 2018**: Multilingual FrameNets and Constructions (Miyaki, Japan). 2018. p. 13-22

HARTMANN, S.; KUZNETSOV, I.; MARTÍN-VALDIVIA, M. T.; GUREVYCH, I. Out-of-domain framenet semantic role labeling. *In*: **Proceedings of the 15th Conference of**

the European Chapter of the Association for Computational Linguistics. Vol. 1. Long Papers. ACL, 2017. p. 471-482. DOI https://doi.org/10.18653/v1/E17-1045

HERMANN, K.M; DAS, D.; WESTON, J.; GANCHEV, K. Semantic Frame Identification with distributed Word Representations. *In*: **Proceedings of the 52nd Annual Meeting of ACL** (Baltimore, Mariland). ACL, 2014. P. 1448-1458. DOI https://doi.org/10.3115/v1/P14-1136

JOHNSON, C. R.; FILLMORE, C. J. The FrameNet tagset for frame-semantic and syntactic coding of predicate-argument structure. *In*: **Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics** (ANLP-NAACL 2000, Seattle WA). ACL, 2000. p. 56-62.

KIPPER, K.; KORHONEN, A; RYANT, N.; PALMER, M. A Large-Scale Extension of VerbNet with Novel Verb Classes. *In*: **Proceedings of the 12th EURALEX International Congress** (Turin, Italy, September 2006). 2006. p. 173-184

OHARA, K., FUJII, S.; OHORI, T.; SUZUKI, R.; SAITO, H.; ISHIZAKI, S. The Japanese FrameNet project: an introduction. *In*: **Proceedings of the satellite workshop "Building lexical resources from semantically annotated corpora"** (LREC 2004). ELRA, 2004. p. 9–11.

PALMER, M; GILDEA, D.; KINGSBURY, P. The Proposition Bank: An Annotated Corpus of Semantic Roles. **Computational Linguistics**, 31:1., p. 71-105, 2005. DOI https://doi.org/10.1162/0891201053630264

PALMER, A.; SPORLEDER, C. Evaluating FrameNet-style semantic parsing: the role of coverage gaps in FrameNet. *In*: **Proceedings of the 23rd international conference on computational linguistics**: posters. ACL, 2010. p. 928-936.

REHBEIN, I; RUPPENHOFER, J.; SPORLEDER, C.; PINKAL, M. Adding nominal spice to SALSA - frame-semantic annotation of German nouns and verbs. In: **Proceedings of KONVENS 2012** (Vienna, Austria). 2014. p. 89-97.

RUPPENHOFER, J.; ELLSWORTH, M.; PETRUCK, M. R. L.; JOHNSON, C. R.; SCHEFFCZYK, J. **FrameNet II**: Extended Theory and Practice. 2010. Available at: http://framenet.icsi.berkeley.edu. (accessed 28. Oct. 2021).

SALOMÃO, M. FrameNet Brasil: um trabalho em progresso. **Calidoscopio** 7. p. 171-182, 2009. DOI https://doi.org/10.4013/cld.2009.73.01

SHI, L.; MIHALCEA, R. Open Text Semantic Parsing Using FrameNet and WordNet. *In*: **Proceedings of HLT-NAACL 2004**, Demonstration Papers. 2004. p. 19-22. DOI https://doi.org/10.3115/1614025.1614031

SUBIRATS, C.; PETRUCK, M. Surprise: Spanish FrameNet!. **Estudios de Lingüística del Español**, n. 31, 2010.

TAULÉ, M. et al. Mapping Syntactic Functions into Semantic Roles. *In*: **Proceedings of TLT2005**. 2005. p. 185-194.

TORRENT, T.T.; ELLSWORTH, M.J. Behind the Labels: Criteria for defining analytical categories in FrameNet Brasil. **Veredas** 17(1), p. 44-65, 2013.

Artigo recebido em: 29.10.2021 Artigo aprovado em: 21.01.2022

DOI: 10.14393/DL52-v16n4a2022-8

Modeling the prosodic forms of Discourse Markers

Para uma modelagem das formas prosódicas dos Marcadores Discursivos

Tommaso RASO[*]
Albert RILLIARD[**]
Saulo MENDES SANTOS[***]

ABSTRACT: This paper has a twofold goal: (i) to propose how and why to identify Discourse Markers (DM), showing that the formal features marking this category are of prosodic nature and can distinguish the six different functions of interactional nature performed by DMs. We describe both the prosodic characteristics responsible for a DM identification and the prosodic forms that convey each type of communicative function inside the more general category of DM; (ii) to show in detail the methodological steps adopted so far to allow the automatic extraction of different DMs from new data. The methodology is presented together with a statistical-computational discussion and explanation.

KEYWORDS: Discourse Markers. Prosody. Spontaneous Speech. Modeling.

RESUMO: Este artigo tem um objetivo duplo: (i) avançar uma proposta para a identificação da categoria de Marcador Discursivo (MD), mostrando que as marcas formais do MD são de natureza prosódica e também capazes de distinguir cerca de seis diferentes funções de natureza interacional veiculadas pelos MDs. Se descrevem tanto as características prosódicas responsáveis para a identificação de um MD quanto as formas prosódicas que veiculam cada tipo de função comunicativa dentro da categoria maior de MD; (ii) mostrar detalhes da metodologia que em maior medida será adotada para modelizar essas unidades e permitir uma extração automática a partir de novos dados. Ela é apresentada com uma reflexão estatístico computacional que a justifica.

PALAVRAS-CHAVE: Marcador Discursivo. Prosódia. Fala Espontânea. Modelização.

[*] Doutor em Linguística, Universidade Federal de Minas Gerais. ORCID: https://orcid.org/0000-0002-3446-313X. tommaso.raso@gmail.com

[**] Doutor em Ciência Cognitivas, Université Paris Saclay, CNRS. LISN. ORCID: https://orcid.org/0000-0001-6490-2386. albert.rilliard@limsi.fr

[***] Mestre em Linguística, Universidade Federal de Minas Gerais – Université Paris Saclay. ORCID: https://orcid.org/0000-0002-9399-9241. saulo.mendes@gmail.com

Within the Language into Act Theory (L-AcT; CRESTI, 2000; MONEGLIA; RASO, 2014; CAVALCANTE, 2020), Discourse Markers (DM) are considered interactional information units marked by specific prosodic devices with high flexibility as for their lexical fulfillment. The proposal of this paper stems from a systematic analysis of data from spontaneous speech corpora of Italian (CRESTI; MONEGLIA, 2005) and Brazilian Portuguese (BP) (RASO; MELLO, 2012), integrated by data from comparable corpora of American English (AE) and Spanish. This paper will explain this hypothesis and the methodology for its modeling.

1 L-AcT and the analysis of spontaneous speech

L-AcT is a corpus-driven theory that expands Austin's (1962) speech act theory, considering the illocution as the nucleus of the reference unit for speech and its informational patterns. The reference unit for speech is defined as the minimal stretch of speech which presents both pragmatical and prosodic autonomy. We can have two kinds of reference units: the utterance and the stanza. The utterance is made up of an information pattern around an illocutionary nucleus; the stanza (CRESTI, 2010) is made up of more subpatterns, each one with an illocutionary nucleus, juxtaposed by a prosodic continuity signal.

Therefore, the reference unit must end with a prosodic terminal boundary and is compound by one or more patterns. Each one performs an illocution and, many times, optional non-illocutionary units. Each information unit is enveloped in a prosodic unit, separated by the following one through a non-terminal prosodic boundary.

Example 1 and audio 1 show a sequence of simple utterances built up only by the illocutionary unit. Example 2 and audio 2 show a compound utterance built up by several intonation/information units, among which only the last one is the

illocutionary nucleus. Listening to audio 2a, one can perceive that the illocution is interpretable in isolation (the illocutionary unit does not need to be the last one); listening to audio 2b, one can perceive that the rest of the utterance, without the illocution, cannot be interpreted in isolation. Example 3 shows a stanza with two subpatterns, each one with its illocution (in bold). The first subpattern is linked to the following subpattern by a prosodic continuity signal (on the word *cemitério*). Both subpatterns, especially the first one, feature other non-illocutionary units that build a pattern with the illocution of their specific subpattern. Stanzas can easily present a more significant number of subpatterns.

Example 1
bpubdl03[118-128)] (áudio 1)
*GUI: faz força // mais // beleza // contrai o abdômen // joga o tronco só um pouquinho pra frente // aí // beleza // descansou // vou baixar um pouquinho mais // vai //
stronger // more // great // contract your abdomen // push the trunk just a little bit forward // ok // great // did you rest // I'm gonna slower it a little bit // go //
The utterances in example 1 are performed by a personal trainer to his client.

Example 2
bfamdl02[101] (áudio 2)
*BAO: porque / se eu for empregado / por exemplo / alguém vê que eu sou muito foda / medo de perder / o posto deles / es vão [/2] es vão me dizar né //
because / if I am hired / for instance / they see I am very good / fear of losing / their job / they will sack me //

Example 3
bfammn03[28] (áudio 3)
*ALO: aí / determinada hora lá / tava na hora de sair o [/1] o [/1] o velório / de ir po cemitério / o filho [/2] o filho mais velho vai lá dento / porque a dona Elvira até então nũ tinha aparecido cá na [/2] cá fora /=COM= né //
so / at a certain moment / it was time the wake to go out / to reach the cemetery /

the older son goes inside / become madame Elvira so far hadn't shown up outside here / did you get this //

The information units are of two types: textual and dialogic. The textual units build the semantic content of the utterance. Besides the illocutionary unit, called Comment, the textual units are Topic, Appendix of Comment, and Appendix of Topic, Parenthetical, and Locutive Introducer. Their analysis is not within the scope of this work. Dialogic units correspond to what in other frameworks are called Discourse Markers (or Pragmatic markers, Discourse particles, or other similar names).

In the last 30 years, a growing amount of literature has been published about DMs[1]. Nevertheless, it is still not clear what a DM is. More precisely, we do not have a clear answer for the following questions:

(a) How can we predict when a lexical item is a DM?

(b) Once we conclude that a lexical item (or a small sequence) behaves as a DM, how do we identify its specific function?

In our opinion, there is a fundamental problem (with few and partial exceptions) that biases the studies on DMs: the point of departure is usually the lexicon. This is very evident from the title of most of the works on DMs. The following made-up titles could exemplify the reality very closely: *The DM* well *in American conversation*; *El marcador* o sea *en el discurso académico*; *Il segnale discorsivo* cioè *nel parlato giovanile*; etc. We will try to answer questions (a) and (b) arguing that what defines both a DM and its specific function is prosody. Then, we will discuss a methodology

[1] See, among others, Bolden (2015), Degand (2014), Aijmer (2013), Traugott (2012), Fischer (2006a; 2006b), Romero Trillo (2006), Frank-Job (2006), Aijmer & Simon-Vandenbergen (2006), Schourup (1999), Brinton (1996), Bazzanella (1990), Schiffrin (1987).

for modeling different DMs according to their prosodic form, as the main feature responsible for their particular function.

2 Discussing some crucial characteristics of DMs according to the literature

Usually, in the literature, some properties are considered typical of DMs. We will discuss some of them:

(i) DMs are lexical units, or small locutions, that do not play a role at the semantic and syntactic level of the utterance since they do not partake in its propositional content; therefore, they are non-compositional items;

(ii) DMs are lexical units, or small locutions, that partially or totally lost their semantic meaning, acquiring a pragmatic function;

(iii) DMs are polyfunctional; this statement may be used in two different senses: in one sense, it means that one DM may cover different functions at the same time in the same specific occurrence; in another sense, it means that a specific lexeme may cover different functions in different occurrences;

(iv) among scholars, different functions are reported. We can divide them into cohesive and interactional functions. Cohesive functions are outside the scope of this work. As for interactional functions, the literature mentions modal functions, illocutive functions, conative function, turn-taking function, and often a strong role in politeness functions (but other functions are mentioned too).

First of all, we can say that modal functions, if we should take it as a lexeme that modifies the modality of the utterance (in the sense of BALLY, 1950), are in contradiction with the fact that there is no compositionality between the DM and the utterance. In fact, something that works as a modal operator needs to be compositional

to its semantic scope. In our view, this should be enough to say that DM are not involved in any kind of modalization. Secondly, we agree with (i) and think that DMs do not partake in the propositional content of the utterance and therefore are non-compositional with the rest of the utterance. But what marks the fact that DMs are not compositional? What makes the difference between:

(4a) *God save the queen!*
and
(4b) *God, save the queen!* (Where God is an exclamation and the rest of the utterance performs an order)

In (4b), functionally, *God* could be substituted by *Jesus* or by semantically very different exclamations, even imprecations. The function would not be affected. The choice of a specific lexeme for this function could depend on diastratic (age, cultural level, gender, specific group) or diaphasic reasons (formal or informal context, specific situation – a party with friends, a soldier of the monarchy that sees the queen is in danger, someone who is watching a chess game, etc.). In any case, we know that in (4a) *God* is the subject of *save*, while in (4b) there is no compositionality between *God* and the rest of the utterance. In (4a), we see the same thing we can also find in

(4c) See *the house of my friend!*
while in (4b) we have the same phenomenon of
(4d) See, *the house of my friend!*

Also, *see* in (4d) could be substituted by many different lexemes without losing its pragmatic function.

But the main point that these examples bring out is that the lexical form cannot be responsible for marking the loss of compositionality. How can we distinguish

between the compositional and the non-compositional status of the same lexical sequences? We can look for an answer only in prosody. The interruption of compositionality is marked by a prosodic non-terminal boundary (that is frequently transcribed with a comma). We will observe better how it is performed in spontaneous speech. The prosodic boundary may or may not coincide with a silent pause, but most of the time, in spontaneous speech, it is marked by prosodic cues other than pause.

We also agree that DMs are polyfunctional, but only if this means that a given lexeme, due to the loss of the semantic content, may perform different functions in different occurrences. In this case, what marks the function since the lexeme is the same? Again, the answer can be encountered in prosody. In fact, what marks the function is not the choice of the specific lexeme but the prosodic form of its concrete realization. Many lexemes can fulfill each function, but what conveys the functional attribution is the specific prosodic form with which the item is performed. Of course, it does exist some correlation between lexicon and function, but (i) any lexeme can perform more than one function, and in each function, it shows different prosodic characteristics; (ii) it is well known that the lexicon changes a lot diachronically, diatopically, diastratically and diaphasically. The importance of the lexicon is, in fact, much greater if we look at the effects for politeness. But this is not the main functional aspect of the DM concerning other linguistic categories. Politeness is involved in all linguistic choices, from the kind of illocution we strategically chose for our goals to any lexical choice, as well to the prosodic contour chosen to express a certain attitude. Politeness is a social category. We cannot say the same thing for the specific prosodic cues that convey grammatical functions; and more importantly, (iii) if what marks the function of a DM should be found in the lexicon, how many functions would we have, and how would we group different lexemes in the same function? We will show that, by paying attention to prosody, we can find probably six forms with interactional

functions and one form with a cohesive function. We will not discuss the latter. We refer the reader to Cresti & Moneglia (2019) for a discussion on this unit.

We still need to discuss an important aspect that we mentioned in (iv). DMs should not be confused with illocutions. Illocutions are textual units; they build the text of the utterance. This is clear in most cases since the illocutionary units are formed by more words, often many, that together build the semantic content of the most important part of the utterance. This might be not so clear in the case of illocutions expressed by exclamations, as we will see. But even in this case, since they perform an action, they can be pragmatically interpreted in isolation. They can even be the only item of a whole turn. On the contrary, DMs can never be interpreted in isolation and always depend on the illocution expressed by a different intonation unit in the utterance.

Examples (5-7) show respectively the same lexeme in a compositional context, with an illocutionary value (this means that the whole utterance is built up by just this lexeme) and as a DM. The examples show this difference using Italian, BP, and AE.

The three examples (5) show the lexemes *vedi*, *não,* and *well* in a compositional context.

> Example 5a: [audio 5a] [audio 5a1] ifamcv15[40]
> SAB*: poi / in piedi / hai visto / anche se il palco è un po' rialzato / però / se ti viene uno davanti alto / non vedi nulla //
> *then / standing up / you see / even if the stage is a little bit raised / however / if someone tall is in front of you / you don't see anything //*
>
> Example 5b: [audio es 5b] [audio es 5b1] bpubdl01[116]
> PAU*: ah / não acaba não / acaba //
> *ah / it does not end / does it //*
>
> Example 5c: [audio es 5c] [audio es 5c1] afamcv03[179]
> TOC*: &he / I didn't do terribly well there //

The three examples in (6) show the same three lexemes with an illocutionary function. In 6b, it is possible to observe that the compositional use of *não* would lead to the opposite meaning; in 6c, the lexeme is the whole content of the speaker's turn.

Example 6a: [audio es_6a] [audio es_6a1] ifamcv15 [42-45]
FER*: vedi // la metti dentro / fa finta di pigliarla / e poi la ributta fuori // vedi // non la vuole //
see // you put it inside / it seems it takes it / and then it throws it out again // see // it doesn't want it

Example 6b: [audio 6b] [audio 6b1] bpubdl01[14-15]
PAU*: não // tá dando a altura daquele que a Isa marcou lá / né //
no // it is the same height of that one that Isa marked there / isn't it //

Example 6c: [audio es_6c] [audio es_6c1] afamdl02[33-35]
*PAM: and where do you get those thoughts //
*DAR: processing what goes on around me //
*PAM: well //

Finally, the three examples in (7) show the same lexemes as DMs. The reader can listen to the audios and verify that the lexemes, which are evidently not compositional, cannot be interpreted in isolation.

Example 7a: [audio es_7a] [audio es_7a1] ifamdl02[611]
LID*: no / poi / vedi / succede questo //
no / then / see / that's what happens //

Example 7b: [audio es_7b] [audio es_7b1] bpubdl01[194]
PAU*: ah / não / ela disse que é pa ficar / por algum tempo //
ah / no / she said it must stay / for some more time //

Example 7c: [audio es_7c] [audio es_7c1] afamcv02[214]
SHR*: well / &e / you've really become &f / good friends with //

Examples (5-7) show how it is possible to answer question (a). DMs are lexemes or small locutions isolated in an intonation unit but non-interpretable in isolation. Their non-interpretability and their non-compositionality are both due to prosodic features.

3 Methodological aspects related to the prosodic modeling

3.1. Methodological premise

Throughout this work, we refer to the prosodic characteristics of DMs. For the sake of interpretability, we do not provide acoustic measurement (the statistical description and the modeling will be the object of a future work). Instead, we refer to high or low intensity, high or low f0, and long or short duration. Since spontaneous spoken data deal with different speakers and different articulation rates or intensities even in the speech of the same speaker, we need to establish a term of comparison with respect to which the prosodic parameters can be considered high, low, long, or short.

This term of comparison must be found within the utterance embedding the DM we want to describe, given the high variability of the information structure and the syntactic and lexical composition of different utterances, and given other factors that influence the speaker voice in different moments of an interaction. The only possible term of comparison is the illocutionary unit, the *Comment*. This is due to two reasons: the first one is that Comment is the only mandatory unit in order to perform an utterance (or a subpattern of the *stanza*); therefore, it is the only term of comparison we can always find. But there is also a theoretical reason: in the L-AcT framework, the illocution is the nucleus of the utterance, while all the other units build a pattern with the Comment and are informationally and prosodically subordinated to it.

Therefore, whenever we say that a DM is marked by a high or low intensity, high or low f0, and long or short duration, we mean that it has these characteristics

with respect to the syllabic mean of the Comment. Of course, the Comment can express different illocutions and therefore can have different prosodic forms; however, its nucleus (the chunk responsible for prosodically marking the illocutionary value in a Comment unit) is composed of a few syllables, normally just one or two, and this strongly reduces the impact of this variability on the mean syllabic measurements.

3.2 Acoustical analysis of prosodic correlates

The voice fundamental frequency (F0), its intensity, and the rhythm of enunciation form the most important correlates of the cognitive processing of the prosody main functions (HIRST; Di CRISTO, 1998). Their estimation and interpretation from the recording are the subjects of a large bibliography that shows the complexity of the task (for a review, see, e.g., COLE, 2015). If pitch estimation is now routinely proposed in many software suites (PRAAT, WINPITCH, SpTK etc.), the output is not systematically reliable. Particularly, non-modal phonations (GERRATT; KREIMAN, 2001) raise significant difficulties regarding the evaluation of a frequency (i.e., a regular phenomenon), whose definition is problematic when vocal folds vibration mechanisms are anything but regular. In Brazilian Portuguese, notably, post-stress syllables generally receive much lower energy, which tends to lead to complete or partial devoicing, or to the apparition of non-modal mechanisms. In English, many studies have shown utterance-final creak functions as a boundary marker (e.g., DAVIDSON, 2019). Martin (2012) shows a sample of detection errors that links particular deteriorations of the speech signal quality to various types of pitch detection algorithms: each pitch detection algorithm (PDA) appears to show specific robustness and weaknesses to different alterations of the input signal. To that aim, it may be interesting to compare the results of several PDAs on a given target signal so as to be able to detect potential problems and zones of high reliability of the F0 detection. The

current ongoing work proposed, for example, an estimation of F0 on the target corpus using the following PDA: Praat's ac, cc, and shs algorithms (BOERSMA; WEENINK, 2020), yin (DE CHEVEIGNÉ; KAWAHARA, 2002), swipe (CAMACHO; HARRIS, 2008), rapt (TALKIN, 1995), and openSMILE:) (EYBEN; SCHÜLLER, 2015). Once candidates from each algorithm are obtained, the output F0 vectors are time-aligned so as to deal with variation in the sampling frequencies and time alignment of each PDA, using typically a linear interpolation. Interpolated F0 candidates are then compared between PDA so as to detect if and where some may show differences in the estimated F0: a "gross error rate" variation above 20% between two candidates is generally considered problematic, as most errors are linked to octave jump (SIGNOL; BARRAS; LIENARD, 2008; CAMACHO; HARRIS, 2008). Note that in our case, unlike the situation in a typical evaluation process, there is no ground truth or reference value: all the candidate PDA algorithms potentially output reliable or erroneous estimations. In a first approximation, the value which is agreed by the majority of the tested algorithms was selected here; in the future, a dynamic programming algorithm will be set up so as to select the candidate (among the proposals of all algorithms) that satisfies criteria of continuity (smooth curve) and majority decision (see, e.g., VINCENT; ROSEC; CHONAVEL, 2006, for an example of application).

Of a different nature than F0, the signal's intensity is relatively straightforward to estimate, but its uses as a reliable correlate of perceived loudness are also problematic: first, its absolute level (sound pressure level) is generally lost during the recording procedure due to many factors (for details and solutions, see ŠVEC; GRANQVIST, 2018). For the intensity value to reflect the perceived loudness, it shall also be submitted to a specific weighting of its frequencies: typically, the A-weighting was designed to reflect human ear perception. Loudness also has complex relations with pitch and the spectral characteristics of the sound (MEUNIER et al., 2018), thus one may use dedicated models of loudness to express the perceived strength of speech

sounds, despite the complexity of such tools (MOORE et al., 2016). Another avenue to estimate information linked to the energy of the signal is to estimate changes in the signal linked to vocal effort (TITZE; SUNDBERG, 1992; LIÉNARD; DI BENEDETTO, 1999; TRAUNMÜLLER; ERIKSSON, 2000); this is basically done via estimations of the energy decrease in the spectrum since an important effect of higher effort is to raise the spectral slope (NORDENBERG; SUNDBERG, 2004; SUNDBERG; NORDENBERG, 2006). Meanwhile, vowel articulation has a major impact on the spectral slope when estimated on the speech signal: the proposed measurements of spectral emphasis thus generally rely on the long-term average spectrum (LTAS) with speech sequences of about 20 seconds so as to stabilize the spectrum (about 20 seconds of speech, from which about 10s. of voiced framed are extracted; see LÖFQVIST; MANDERSSON, 1987). From such LTAS, based on voiced frames only, the following indexes are often used in the literature: the so-called "Hammarberg index", which is the difference between the peak amplitudes of the 0-2 and 2-5kHz bands of the power spectrogram (HAMMARBERG et al., 1980); the alpha index, which is the ratio of sound energy above and below 1kHz (SUNDBERG; NORDENBERG, 2006), or the "spectral emphasis" as defined by Traunmüller & Eriksson (2000), which is the spectral energy above 1.5 the mean F0 on the concatenated voiced segments. These indexes are all derived ways of estimating the increase of spectral slope produced by vocal effort, typically to estimate changes linked to arousal (BANSE; SCHERER, 1996; GOUDBEEK; SCHERER, 2010). Liénard (2019) shows that it is possible from LTAS to estimate the original "voice strength" thanks to the characteristics of the low and mid-range of the spectrum, for gender-specific data, with an accuracy of 3dB. The main limitation of these measurements, with regard to an approach that is looking for short-term measurements, is the fact they need long-term speech data (as defined above) to be reliably evaluated (reliably meaning without the effect of the segmental changes produced by articulation). Another possibility would be to estimate parameters of the

source component related to loudness (see table 2 in D'ALESSANDRO, 2006, for propositions) thanks to an inverse filtering approach; unfortunately, the inverse filtering is not a reliable process on sustained vowels from many different speakers (KREIMAN; GERRATT; ANTOÑANZAS-BARROSO, 2007), hence shall give even worst results from spontaneous speech recorded during field works.

Considering the limitations of both raw intensity estimation from uncalibrated signals and the fact that reliable vocal effort estimations (in dB) can only be done on long-term signals, a potential approach may consist in estimating the mean effort of a speaker from the longest possible connected utterance at hand, containing the targeted unit from which intensity is to be extracted, and using this value to "calibrate" the mean intensity of the complete utterance. The intensity of the targeted unit would then be normalized with respect to the mean calibrated intensity of the sentence, following equation 1, where the calibrated intensity I_c estimated at time t equals the mean vocal effort E estimated (in dB) on the complete utterance plus the difference between the observed raw intensity at time i, $I_r(t)$ minus the mean of the raw intensities estimated on the N samples of the utterance.

Eq. 1: $I_c(t) = E + \left(I_r(t) - \frac{1}{N}\sum_{x=1}^{N} I_r(x)\right)$

The problem is, thus, the estimation of the mean effort, or voice strength in Liénard's terms (2019), expressed in dB: in the absence of a reliable model to estimate such value (there are none currently available to our knowledge), the mean intensity of the sentence, or of a reference part of the sentence (typically, in this context, the Comment), may be used as a reference so as to express the intensity measurement differentially regarding the reference illocutionary nucleus.

Manual speech segmentation is a relatively simple task, if heavily time-consuming. Meanwhile, deriving the perceived rhythm from the raw duration of segments is also non-trivial due to phoneme-specific intrinsic and co-intrinsic durational constraints. Normalization procedures have been developed, based on the standardization of raw durations with regard to the observed distribution of duration for similar segments in similar contexts (CAMPBELL; ISARD, 1991): this statistical approach allows an efficient estimation of segmental lengthening, expressed as the deviation from the expected duration for a phoneme with a set of characteristics. Building on this approach at the segmental level, Barbosa (2007) developed a model of speech rhythm that estimates the lengthening for syllable-like units (the so-called vowel-to-vowel unit) according to the duration characteristics of its phonemes and the contextual lengthening. This model (and other information) was then implemented into a semi-automatic tool available to the research community so as to derive reliable measures of rhythm from segmented speech (BARBOSA, 2013).

3.3 Stylization & normalization of measurements

Once reliable measurements of acoustic parameters are available, one may try to extract information from this dataset so as to describe what in the acoustic characteristic of prosody is related to the targeted functional aspects. Not all changes in the measurements are relevant to prosodic functions. For example, articulation has a dramatic effect on the measured intensity, typically with consonantal articulation. Intrinsic phonemic differences are also observed in the F0 and intensity values, similar to what was described earlier for duration. Such changes in parameters are regrouped under the umbrella term of microprosodic effects (DI CRISTO; HIRST, 1986; DUBĚDA; KELLER, 2005), where the articulation constraints produce systematic and predictable changes (e.g., higher F0 on close vowels) on both measurements. These changes may

be relatively large, but on vowels, they generally do not exceed perception thresholds; the changes are larger for co-intrinsic factors, with consonantal articulation having important effects on vowel's F0 – a fact that helps phonemic identification (HONDA, 2004).

If important for segmental perception, microprosody introduces changes that are not of interest for the macroprosodic aspects of speech and are more a nuisance to prosodic analysis. F0 is also a measurement that changes continuously throughout an utterance, while the pitch perceived on a given syllable may often be represented by a single note by trained musicians (see musical performances of Hermeto Pascoal, for example). This is partly due to the integration of the F0 value as a single note by our perceptual system for variations below a given threshold of perception (ROSSI, 1971). This threshold has a (negative) linear relation with the duration of the tone, on logarithmic scales ('T HART; COLLIER; COHEN, 1990) – i.e., the longer the sound, the smaller the threshold of perception for pitch differences. These facts lead the researchers of the IPO school to propose a straight-line stylization of the original F0 curve ('T HART et al., 1990), with line segments fitting the raw measurements, so the resynthesis of the stylized version is indistinguishable from the original (their so-called "close-copy stylization"). The stylized pitch curve has (at least) two main advantages over the raw measurements: (i) it proposes an economical description of F0 changes in terms of quantity of information required to describe the curve; (ii) it removes changes that do not participate to the perceptual processing of prosody, so changes that have no interest for prosodic models. Several variants of IPO close-copy stylization have been proposed in the literature that departs from 't Hart et al. (1990) proposal on several aspects but kept the perceptual equivalence principle. Two renowned implementations may be cited here: Hirst & Espesser (1993) MOMEL algorithm, based on quadratic spline functions producing a continuous smooth curve stylization of a sentence melody, and d'Alessandro & Mertens (1995) model of tonal perception that

stylizes F0 variation on each vocalic segments by straight lines, and is implemented as a Praat script in the Prosogram (MERTENS, 2004).

3.4 Parametric description of prosodic variations

F0 stylization is an important step in simplifying the F0 curve so as to remove some of its undesirable characteristics (microprosody, declination line), keeping the meaningful features that are thus described via an economical set of parameters. Phonological approaches to the prosodic description would take a further step after close-copy stylization, reducing even more F0 changes so as to keep only variations that are relevant to phonological functions. This led 't Hart et al. (1990) to propose their standardized stylization – a straight line curve that is no more perceptually equivalent to the original (in a psycholinguistic approach) but carries the same (phonological) functions as the original; they propose that the rising and falling movements of their stylization are the basic elements of their proposed grammar of intonation. Other implementations of phonological models have started with similar principles (close-copy stylization, based on various processes), but then proposed the target levels of the stylized pitch movements as the basic elements descriptive of the phonological models – e.g., the INSINT alphabet (HIRST; DI CRISTO; ESPESSER, 2000), Pierrehumbert (1980, 1981) High and Low targets, or the Polytonia system based on the Prosogram (MERTENS, 2014). Note that these systems of phonological description (without tackling here position on their theoretical differences, see the discussions in the referred literature for details) may be used for the latter processing, especially in cases where phonological functions are targeted, using their respective descriptive alphabets as feature set – see Rilliard (2019) for an example of application. Meanwhile, the functions that are targeted here are implemented at a pragmatic level; hence we will prefer keeping it at the phonetic level for the description of prosodic variation.

From a close-copy stylization, it is possible to extract a set of parameters that will describe the features of the prosodic variation. This may be done in several ways. 't Hart et al. (1990), for example, propose a feature matrix (rising, falling, etc.) that describes the changes in the stylized prosody, along the same lines as a proposition by Martin (1973, 1987) or Contini & Profili (1989). One may also describe the prosodic changes in terms of the fit of the estimated F0 points (raw or stylized ones) by polynomial functions; this is particularly efficient in cases where the prosodic units that are compared have a comparable duration. This can be done by fitting orthogonal polynomials to the measured F0 points and using the coefficients of these polynomials as descriptors of the underlying intonation shapes (LEVITT; RABINER, 1971; KOCHANSKI; SHIH, 2003; GRABE; KOCHANSKI; COLEMAN, 2007; LAI, 2014). Note the MOMEL approach (HIRST; ESPESSER1993) may fall in the same category, if not based on polynomials but on splines. Similar curve-fitting approaches are found with the application of the Fujisaki model (FUJISAKI, 1983, 1988) to prosodic description (MIXDORFF, 2000): the parameters of the fitting model are here motivated by the F0 production mechanisms, and they allow the description and comparison of prosodic performances (MIXDORFF; PFITZINGER, 2005; GURLEKIAN et al., 2010). Other model-based approaches have been followed, but without the motivation that characterizes Fujisaki's model, using a functional data analysis framework (FDA, RAMSAY; SILVERMAN, 2005): these works take advantage of the descriptive power of this statistical framework so as to compare the prosodic characteristics of the investigated functions (EVANS et al., 2010; HADJIPANTELIS; ASTON; EVANS, 2012; CAVALCANTE, 2020). The main point of all these approaches is their ability to describe and compare the shapes of F0 contours – thus potentially catching an important aspect of intonational variation, as in the case of tones (EVANS et al., 2010) or nuclear accents (GRABE et al., 2007).

The question of the comparability of the prosodic units, in terms of duration, is a complex one: there always is some durational variations between performances of similar speech "items"; this is easily overcome by introducing some temporal normalization – e.g., by considering a fixed number of points along the targeted contours (XU, 2005) – but such a solution is valid only if the structures of the compared objects are similar, typically in terms of the number of syllables. Not taking care of this may lead to misalignment of the different structures – while alignment of F0 with the underlying syllabic and lexical structure is a crucial prosodic characteristic (KOHLER, 2006).

Another way to describe prosodic variation and to compare its performances is to extract sets of summary statistics from each item to be characterized. This typical machine-learning approach will extract a set of low levels descriptors (e.g., F0, intensity, rhythmic measurements, several voice quality-related parameters, and often spectral information) and propose statistical descriptors of them across the complete item or relevant parts of the item (the nucleus, for example). Such statistical descriptors are typically the mean, standard deviation, first and second derivative, etc. Barbosa (2013) proposes a Praat script that allows the extraction of these kinds of descriptors; software like openSMILE (EYBEN; SCHULLER, 2015) was built so as to allow the extraction of feature-rich vectors then used for the categorization of, e.g., affective variations (EYBEN et al., 2016; MACARY et al., 2021). A frequent limitation of these approaches (for linguistic studies) is the predominance of black-box algorithms used for the classification tasks, which prevents an interpretation of the results, together with the large and mostly uninterpretable feature set. It is nonetheless possible to follow this path using smaller (and motivated) feature sets and to use white-box statistical algorithms.

3.2 Classification or clustering of observations

One aim of this methodological paper is to describe possible paths in the description of prosodic variations related to the description of communicative functions. Once the steps described in the previous parts have been done, the researcher has at her/his disposal a set of prosodic features on the one hand and communicative functions related to the items observed in a given corpus on the other; the problem being thus to describe/find the relations between these two groups.

There are two possibilities here (for details and more in-depth descriptions of possible options, the reader is referred to, e.g., VENABLES; RIPLEY, 2002; RENCHER; CHRISTENSEN, 2012): or the theoretical categories (communicative functions, or other – here the functions of the MDs) are already known for each item, or they are not. In the first case, classification algorithms, or tasks of supervised learning (in the machine learning world and speak), is the most obvious path; a classical one is the use of a discriminant analysis approach in the search for the best combination of the available prosodic features able to *predict* the theoretical categories. The aim is to build a predictive model able to attribute one of the theoretical functions (closed set) to any new observation (i.e., one not part of the training set used to build the model); a derived use is to describe how the prosodic features are combined to allow this predictive capacity, so as to better understand possible relationships between the two levels – i.e., one may use a white-box algorithm so as to be able to observe the specific feature combination build by the model, the aim not being the best possible predictive capacity, at the expense of descriptive ability. Note that any type of prosodic features (curve fitting parameters or feature vectors, or also characteristics extracted from phonological models) may be used in this process. In the second case, the theoretical categories are not previously known, and the aim is more descriptive, unsupervised, or exploratory (in the sense of VENABLES; RIPLEY, 2002): how the measured prosodic

features do separate the collected items into groups (or "clusters") that share similar prosodic characteristics within a cluster and maximal differences across clusters. These methods help observe patterns in the data that are able to show similarities within a subset and differences between subsets. The interpretative work – how the obtained groups or dimensions do relate to the theoretical communication functions – is left, of course, to the researcher; note also this process is potentially limitless, as the potential combinations are open.

One of the first steps in an exploratory process is the visualization of the dataset, generally using techniques linked to Principal Component Analysis (with or without rotations). This allows the reduction of the complexity of the prosodic features by finding the correlations between them and potentially allows observing potential patterns in the cloud of points formed by the observations. An approach of widespread use in the linguistic literature (e.g., NERBONNE et al., 2011) is the Multidimensional Scaling algorithm (MDS), which seeks a projection of the data set onto a reduced (2 or 3) number of dimensions, a dimensionality that best fit the geographic spread of dialectological data (but not necessarily the complexity of pragmatic functions). The method is based on a dissimilarity matrix that allows the comparison of a "distance" between each pair of items. The computation of such a (dis)similarity amounts to being able to calculate the prosodic "distance" between items. This prosodic difference, if one's goal is to link prosodic parameters to communicative functions, has to reflect the perceptual difference listeners would make between these functions. There is no perfect match between a set of prosodic features and their combination to propose an objective dissimilarity and the perceptual grouping obtained in different perceptual evaluation, but it is exactly this goal such techniques would ultimately try to produce (HERMES, 1998a, 1998b; HIRST; RILLIARD; AUBERGÉ, 1998; DE CASTRO MOUTINHO et al., 2011; D'ALESSANDRO et al., 2014).

4 Results: description of DMs' prosody

4.1 Raw characteristics of main DM categories

4.1.1 Previous studies

The proposal that Dialogic Information Units could explain DMs was firstly advanced by Cresti (2000), where an initial distinction was made. Cresti, studying Italian, proposed that there were four DMs (Dialogic Units in the L-AcT terminology), which she called Incipit (INP), Conative (CNT), Allocutive (ALL), and Phatic (PHA).

Incipit is described as having the function of taking the turn or beginning the utterance, expressing affective contrast with the previous utterance. Its distribution is always the initial one (in the utterance or in the stanza's subpattern), and its prosody shows very high intensity and f0, and a very short duration. Cresti says that INP can have different forms: rising, falling, and rising-falling. This poses one problem: why the same function should be conveyed by different forms? We will answer this question below.

Conative is described as having the function to push the listener to do something or to stop doing something. It has a free distribution. Prosodically, it is described with a falling profile, short duration, and high intensity. Also, Allocutive is described as distributionally free, but we will be back to this point later. Functionally, it serves to establish social cohesion among the interlocutors or to disambiguate who is the addressee of the utterance, using titles, epithets, and proper names. ALL is the DM where the lexical category is more defined. Cresti (2000) attributes to ALL a form similar to CNT: falling movement, short duration but low intensity. The description of these two units generates confusion when CNT is fulfilled by the lexicon typical for ALL, which is not rare. In fact, this has led to overestimating the number of ALL and underestimating that of CNT. As we will see below, a better prosodic description of these two DMs allows a clear distinction.

Phatic is described functionally as serving to maintain the channel open. This function seems too vague. It could be attributed to any kind of filled pause or even to textual informational units scanned in more intonation units. Besides, its prosodic description emphasizes the very short duration and very low intensity of the unit but does not individualize a specific form, which would be explained by its very low duration. Distributionally, it would be free too. This unit, therefore, poses some problems: we do not have a clear function for it, its prosodic form is not defined, nor its position could help its recognition. We will pick up on these difficulties below.

Frosali (2008), using the same theoretical framework, proposed two more DMs: Expressive (EXP) and Discourse Connector (DCT). Since DCT (CRESTI; MONEGLIA, 2019) has a cohesive function, it will remain outside of the scope of this work. As for EXP, it is described as follows: functionally, it would express emotional support for the illocution; distributionally, it would be free; and prosodically, it would present medium intensity and duration, and a form defined as "modulated", meaning that it is possible to see more than one small f0 movement in it. Evidently, also EXP poses some problems: what would it really mean to have the function of emotionally supporting the illocution? Besides, we need to have a clear prosodic form in order to say that DMs could be grouped in a limited set of prosodic cues mainly responsible for conveying their function. Also, this case will be better explored.

Raso (2014), using Cresti's and Frosali's categories and descriptions, tries a first systematization of the proposal studying two comparable sub-corpora of C-ORAL-ROM Italian and C-ORAL-BRASIL. At this point, it became clear that a better prosodic analysis was needed. This was the goal of Raso & Vieira (2016). They found an answer for the different profiles of INP and turned clear the prosodic distinction between ALL and CNT. But they leave the other DMs still in need of a better understanding.

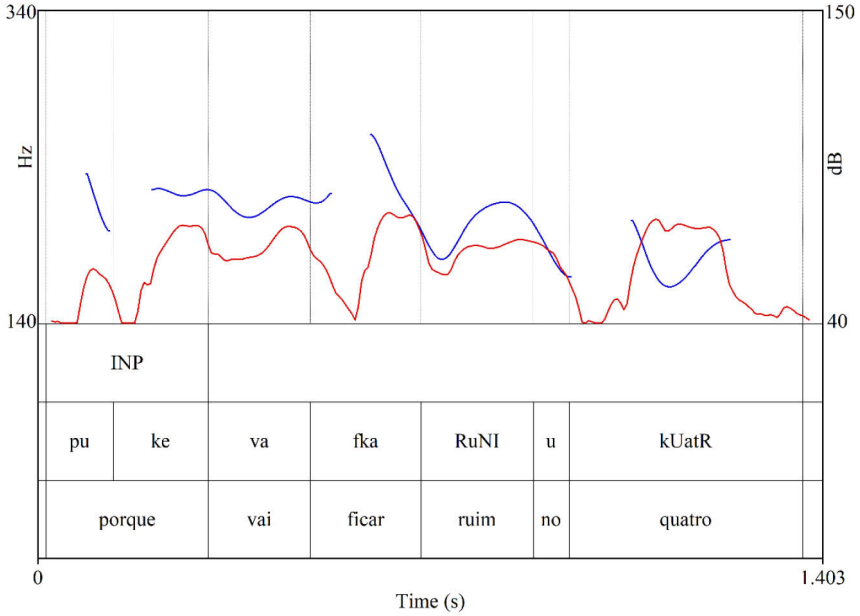
4.1.2 Incipit

As the examples below show, the different forms of INP are due to microprosodic effects. In this specific case, these effects are important because they are magnified by the very high f0 and the very short duration of this unit. Its form, in fact, should be described as a flat stressed vowel with a very high range, higher than the Comment mean, and very high intensity, higher than the Comment mean. Nevertheless, if the stressed syllable is preceded or followed by consonants or another syllable, its form is affected by the microprosodic segmental material; a non-stressed syllable before (or a single voiced consonant) produces a rising movement, in order to reach the high pitch; a non-stressed syllable (or the semivowel of a diphthong) after the stressed vowel produces a falling movement. These movements have a very high f0 variation rate.

Example 8 [audio 8] bfamcv03[257]
*CEL: porque /=INP= vai ficar ruim no quatro //
because / it will be bad for the four //

Example 8 (fig. 1) shows how the prestressed syllable is influenced by the micromelodic effect of the unvoiced bilabial, while the stressed one is influenced by the unvoiced velar, whose f0 range is maintained by the stressed vowel.

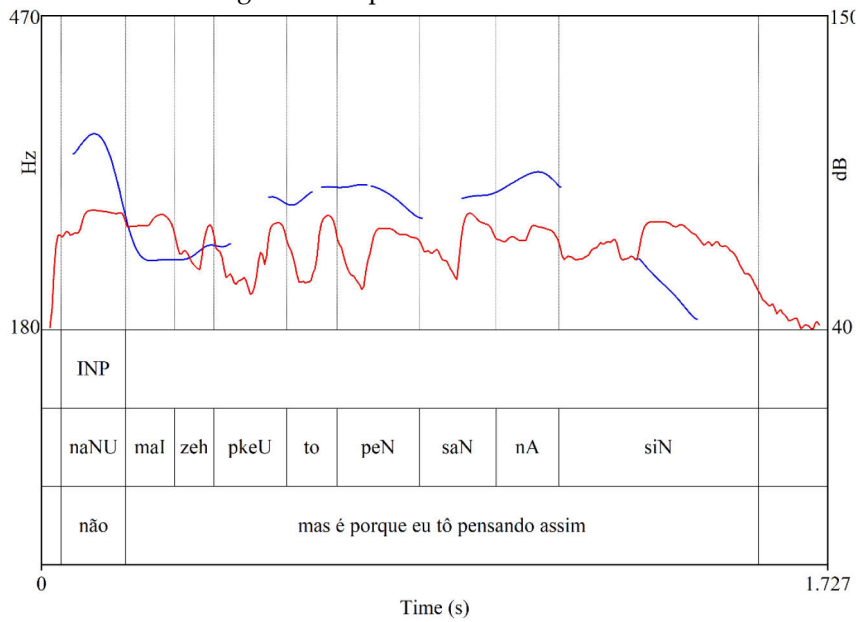
Figure 1 – Representation of audio 8: f0 (in blue), intensity (in red), tagging of the target unit, broad syllabic transcription, and orthographic transcription of the utterance[2].



Source: Raso & Vieira, 2016.

Example 9: [audio 9] bfamdl02[195]
*BAO: não /=INP= mas é porque eu tô pensando assim //
no / but it is because I'm thinking this way //

Figure 2 – Representation of audio 9.



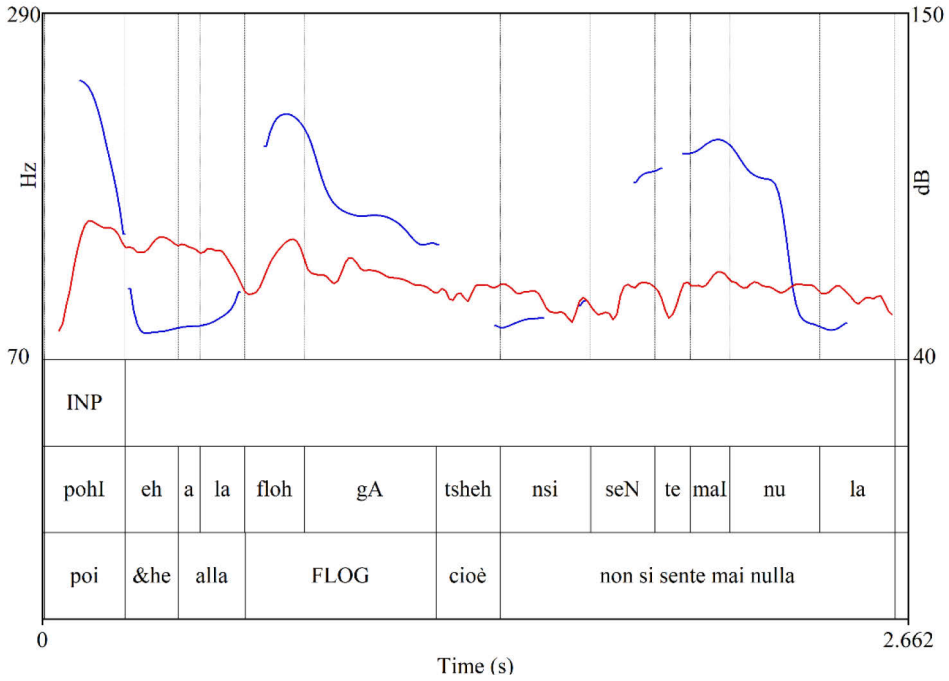
Source: Raso & Vieira, 2016.

[2] All the figures showing the different units have the same structure.

Figure 2 shows how the form of the INP is affected by an initial voiced consonant that causes a rising movement until the vowel of the diphthong, which falls on the semivowel.

Example 10: [audio 10] ifamcv06[32]
*ILA: poi /=INP= alla FLOG / cioè / non si sente mai nulla //
The / at the FLOG / I mean / we never listen to anything //

Figure 3 – Representation of audio 10.

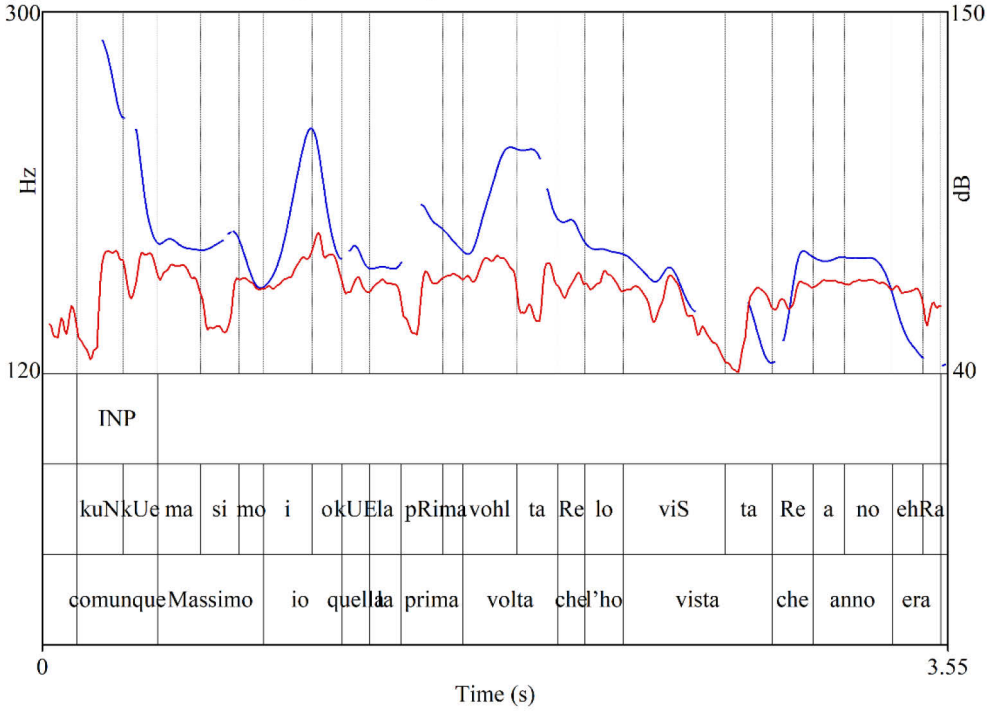


Source: Raso & Vieira, 2016.

Figure 3 shows the effect of the unvoiced consonant, while, like in figure 2, the semivowel of the diphthong causes a falling movement.

Example 11: [audio 11] ifamcv01[871]
*ELA: comunque /=INP= Massimo / io / la prima volta che l'ho vista / che anno era //
anyway / Massimo / I / the first time I saw her / what year was it //

Figure 4 – Representation of audio 11.



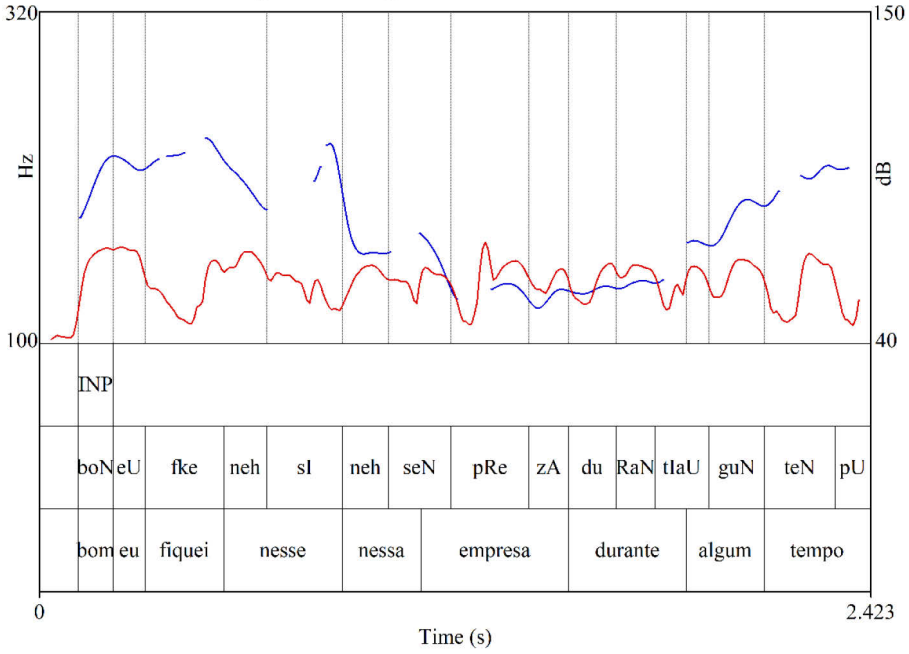
Source: Raso & Vieira, 2016.

Figure 4 shows the high range of the beginning of the INP due to an unvoiced consonant and the falling profile of the unstressed syllable. Note that the Italian word *comunque* [ko'mũkwe] is pronounced [kũkwe].

Example 12 [audio 12] bfammn06[49]
*JOR: bom / eu fiquei nesse [/1] nessa empresa durante algum tempo /
well / I remained in this firm for a while /

In figure 5, it is possible to see the rising profile due to a voiced consonant that needs to reach the high range of the stressed vowel, not followed by any segmental material.

Figure 5 – Representation of audio 13.



Source: Raso & Vieira, 2016.

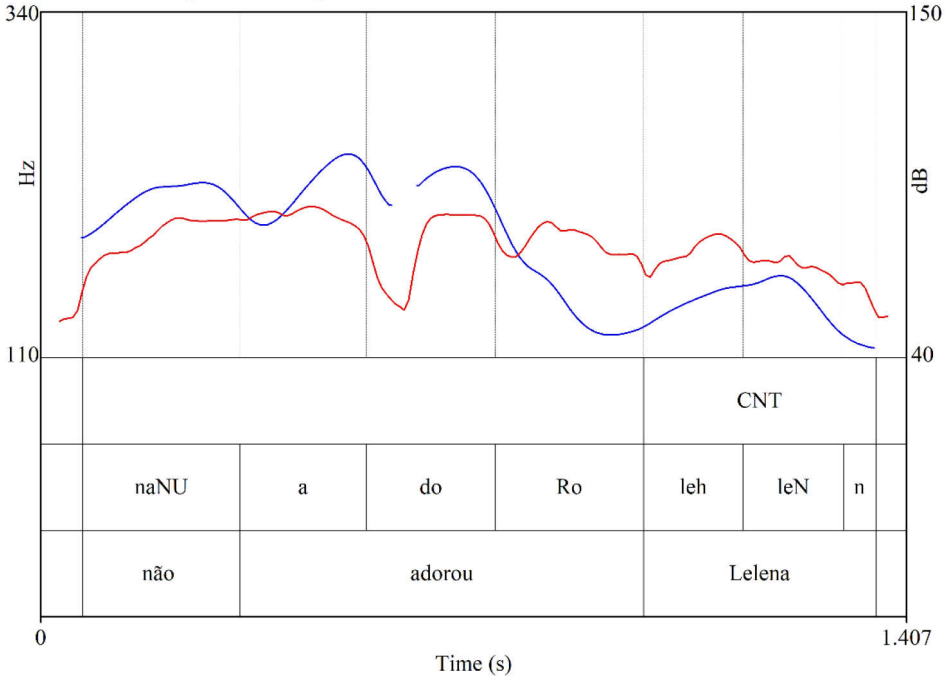
4.1.3 Conative and Allocutive

The examples below show the differences between CNT and ALL. CNT has a falling movement, usually with a high f0 variation rate (even if not so high as in INP and not higher than the Comment) and a high intensity (but not so high as in INP and not higher than the Comment). But what is more interesting is that in CNT the falling movement is aligned with stress syllable, while ALL falls since the beginning. Later (RASO; FERRARI, 2020), it was also observed that in CNT, before the falling movement, there is a slightly rising movement. This is more evident when the stressed syllable is not the first one, but it can be observed even if the stressed syllable is the first one and there is enough segmental material before the vowel; of course, it cannot be seen if this material is a non-voiced consonant. Raso & Ferrari (2020) propose a probably clearer functional definition of CNT. It would signal the illocutionary solution of the utterance.

Looking more in-depth at the behavior of ALL, Raso & Ferrari (2020) noted that it cannot appear in the initial position and prefers the final one. However, the few cases of ALL in medial positions help to better understand its form. ALL has a falling movement that becomes flat in its second part, regardless of the position of the stress. Most of the ALLs are in the final position, which partially masks their real form. In fact, due to the end of the utterance and to the fact that ALL features a low intensity, the falling part is more evident, and the flat part is frequently non-articulated or with such a low intensity that its f0 cannot be tracked.

Example 13 : [áudio 13]
*LUR: não / adorou / Lelena //=CNT=
no / he liked it // Lelena //

Figure 6 – Representation of audio 13. ifamdl12[238].

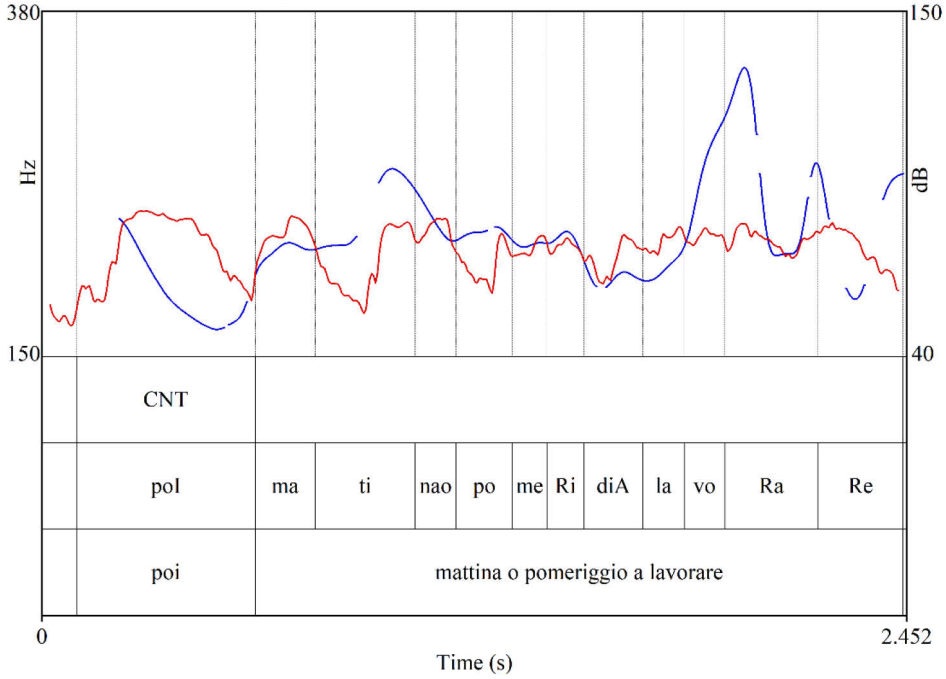


Source: Raso & Vieira, 2016.

Figure 6 allows us to observe the alignment of the falling f0 movement on the stressed syllable and the rising movement in the prestressed one.

Example 14: [audio 14] ifamdl12[235]
FRA*: poi /=CNT= vabbé / mattina o pomeriggio a lavorare /
then / okay / morning and afternoon working /

Figure 7 – Representation of audio 14.



Source: Raso & Vieira, 2016.

This figure can be compared with example 10 (fig. 3), where the same lexeme *poi* has the function (and the prosodic form) of INP. In figure 7, the CNT does not have the possibility to feature the initial rising movement since the stressed vowel is preceded by an unvoiced consonant.

The next four figures aim at showing once more how the lexicon cannot be considered a strong functional vehicle. In examples 15, 17, and 18 (fig. 8, 10, and 11), we have the same lexical item *Rena* (from the name *Renata*) working respectively as a calling illocution, as an ALL, and as a CNT. In 16 (fig. 9), we have another proper name, with the same accentual structure; also, in this last case, as in 15 (fig. 8), the name conveys an illocution, but it is not a calling; it is a warning. The prosodic form of calling in 15 (fig 8) is clearly different from that of warning in 16 (fig.9). Likewise, the prosodic

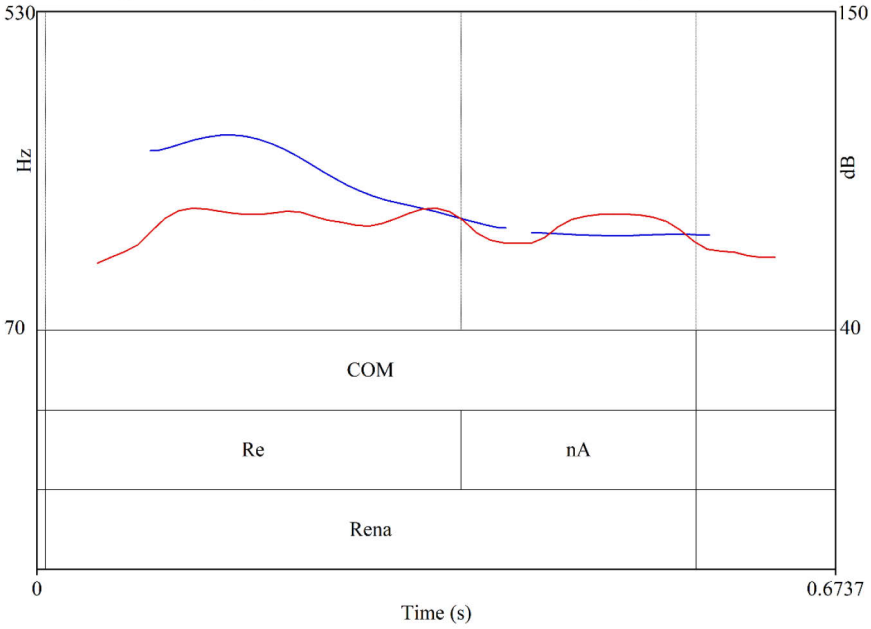
forms in 17 (fig. 10) and 18 (fig. 11) are different both with respect to each other and with respect to the two illocutionary forms.

In the calling illocution, the name *Rena* features a duration of 476 ms, an f0 mean of 248 Hz, reaching its maximum at 355 Hz, and mean intensity of 78 dB. In the warning illocution, the name Bruno features a duration of 272 ms, an f0 mean of 210 Hz, reaching its maximum at 263 Hz, and mean intensity of 84.6 dB. These parameters are much different from those of example 17 (ALL) and 18 (CNT), uttered by the same speaker of 15. In example 17 (fig. 10), Rena has a duration of 203 ms, an f0 mean of 335 Hz (due to the very high pitch of the whole utterance), and a mean intensity of 72 dB. In 18, Rena features a duration of 371 ms, a mean f0 of 232,5 Hz, with its maximum at 258 Hz, and a mean intensity of 70,5.

Some of these prosodic aspects can be evaluated only with respect to the specific illocution of the utterance, but what is interesting is to observe (i) that ALL and CNT are much shorter than the illocution of calling (as explained in note 6, the warning needed to be very fast due to a clear attitude of urgency); (ii) that, in both illocutions, a prominence with its specific illocutionary form is clearly recognizable; (iii) that, in the ALL, the f0 is falling from the beginning, but in the CNT it starts at 230 Hz, reaches 258 Hz and then falls to 204 Hz; the falling part starts three or four pulses after the beginning of the vowel.

Example 15: [audio 15] bfamdl01[255]
FLA*: Rena //=COM=

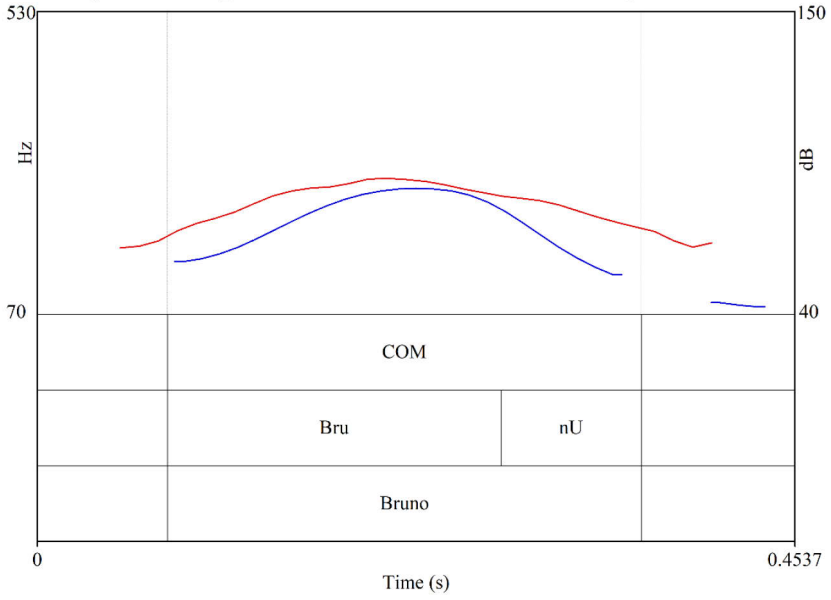
Figure 8 – Representation of audio 15 – Calling illocution.



Source: Raso & Ferrari, 2020.

Example 16[3]: [audio 16]
DEI*: Bruno //=COM=

Figure 9 – Representation of audio 16 – Warning illocution.

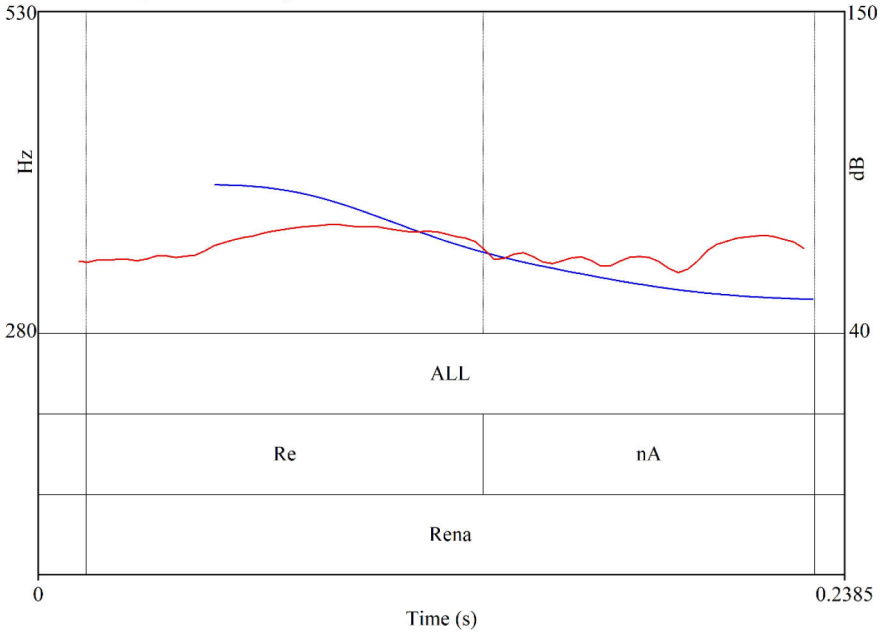


Source: Raso & Ferrari, 2020.

[3] This example is extracted from a corpus of Angolan Portuguese not yet published (ROCHA et al. 2018). In this context, *Bruno* is walking together with an Angolan guide while a truck is dangerously approximating. The "vocative" by the guide is a warning illocution, not a calling.

Example 17: [audio 17] bfamdl01[194]
FLA*: (&va [/1] vai esse / né /) Rena //=ALL=

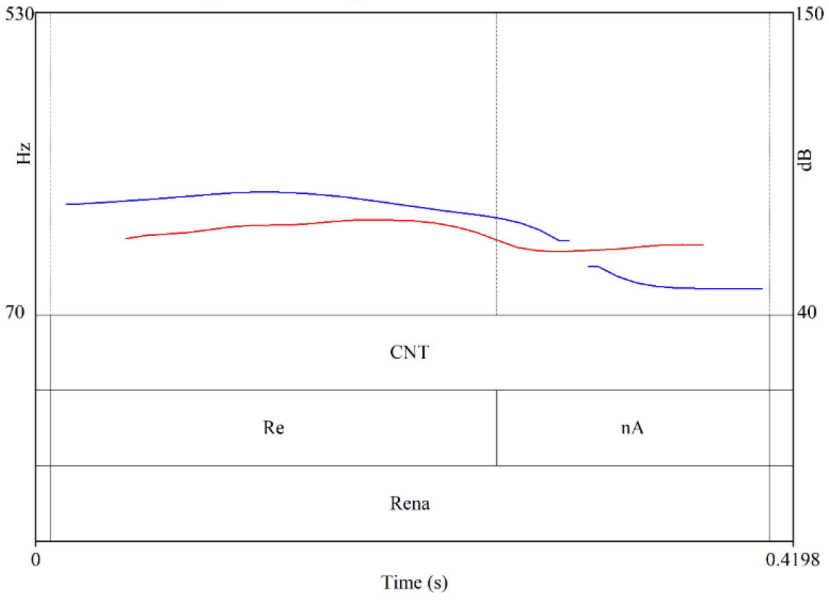
Figure 10 – Representation of the ALL unit of audio 17.



Source: Raso & Ferrari, 2020.

Example 18: [audio 18] bfamdl01[27]
FLA*: (<cê vai embora que> dia /) Rena //=CNT=

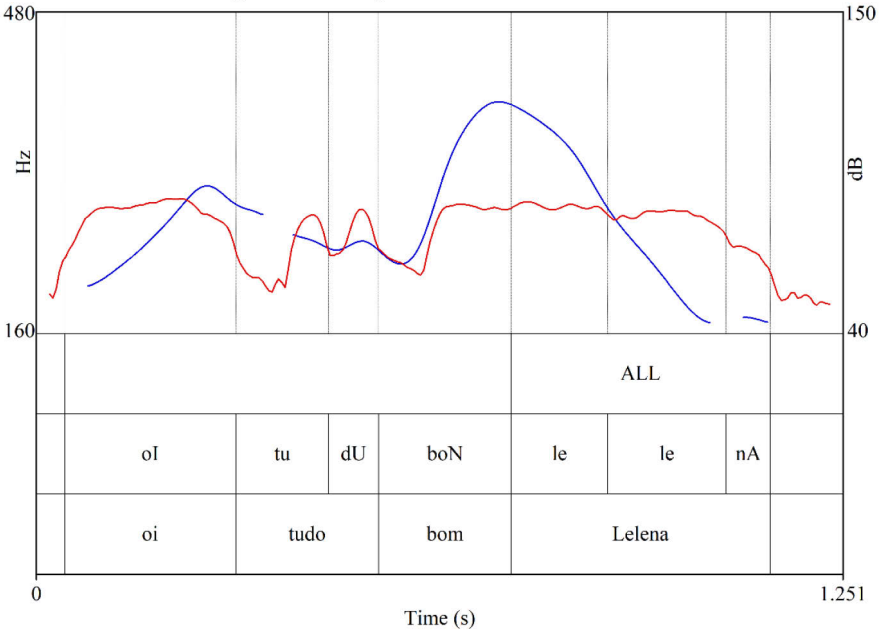
Figure 11 – Representation of audio 18.



Source: Raso & Ferrari, 2020.

Example 19: [audio 20]
*LUR: oi / tudo bom / Lelena //=ALL=
hello / everything OK / Lelena //

Figure 12 – Representation of audio 19.

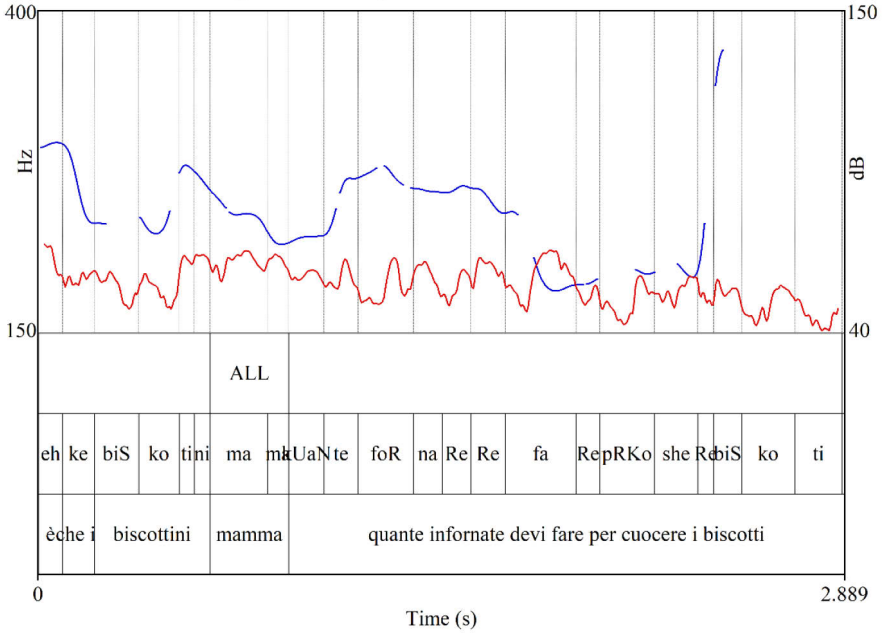


Fonte: Vieira & Raso, 2016.

Figure 12 can be compared with figure 6 since the same lexeme (Lelena), stressed on the second syllable, is performed as a CNT in example 13 (fig. 6) and as an ALL here.

Example 20: [audio 20] ifamdl14[134]
*VAL: è che i biscottini / mamma /=ALL= quante infornate devi fare per cuocere i biscotti //
the fact is that the cookies / mom / how many times should you put them in the oven to cook them //

Figure 13 – Representation of audio 20.



Source: Raso & Vieira, 2016.

Figure 13 shows the form of ALL when its position is not the final one, but it is inside the utterance. The unit has a falling movement that becomes flat and then falls at the end. This is even more evident in figure 4 (ex. 11), where the second unit (*Massimo*) is an ALL. This shows that ALL's form is influenced by its position in the utterance, while CNT is not.

4.2 Parameter extraction

To describe these changes in the shapes of DM contours, as described in the preceding section, we propose to compare two approaches. One, deriving prosodic descriptors from the raw measurements, has been tested by Gobbo (2019) for the classification of DMs. The other one, still untested on DMs, consists in extracting parameters describing the shapes of a curve fitting the raw F0 measurement; it has been tested, e.g., by Cavalcante (2020) for describing the topic unit.

4.2.1 Prosodic descriptors of prosody

The work presented in Gobbo (2019) is a good example of a prosodic feature set applied to the DM case. The author extracted from a corpus the following set of prosodic descriptors (details on their measurement in the original work):

- Five duration descriptors: the raw and normalized duration (in milliseconds) of the DM and of its syllables (mean of the raw and normalized duration), the number of syllables of the DM (normalized durations are obtained using BARBOSA, 2013, algorithm).
- Four intensity descriptors: the mean, standard deviation, minimum and maximum of intensity (in dB) level over the DM (calculated with reference to the comment unit's level).
- Eight F0 descriptors: the mean, standard deviation, minimum and maximum of F0 level (in semitones) over the DM, plus the F0 range, and F0 slopes (over the complete DM, until the final stress, from the final stress).
- Eight "alignment" parameters, indicating where stands the intensity or F0 maximum or minimum, in relation to the DM duration or the stress: relative position of the minimum or maximum of intensity or F0 within the DM

These 25 descriptors were then fed in multivariate data analysis procedures, which will be described in 4.3.

4.2.1 Parameters of shape description

On the basis of the raw F0 measurements, a curve fitting algorithm can be applied on the part of the DM that has to be taken into account. Within the FDA framework (RAMSAY; SILVERMAN, 2005), there is a need to find a single timeframe that fits all the units to be compared, so the timing of the items has to be comparable;

as the events of interests do not necessarily appear at the same position across items, a registration procedure allows the definition of anchor points between which comparable variations do occur. This process can, for instance, be used to match stress syllables across items of different metric structures (see CAVALCANTE, 2020). The curve fitting is then done minimizing a cost function (the roughness of the curve) that deals with the smoothness of the final curve – this process is similar to the MOMEL approach (HIRST; ESPESSER 1993) that fits a continuous spline across a set of anchor points and proved an adequate manner to obtain a close-copy stylization of the raw F0 points.

By doing so, we only use the intonation curve, and we consider F0 points as having an identical weight on the perception of speech melody. It may be possible to introduce weighting factors to put more importance on the passages of pitch that are produced with a stronger voice and thus shall bear a more important perceptual role. Building on Hermes (1998a, b), one may use the maximum amplitude of the subharmonic subspectrum as a weighting factor or another measure of loudness related to voicing (see, e.g., D'ALESSANDRO; RILLIARD; LE BEUX, 2011). Duration is also certainly an important factor to take into account so as to have a reliable prosodic representation. One problem with the duration being its measurement at a syllabic (or phonemic) level – but local speech rate estimations may be derived as a continuous curve (MIXDORFF; PFITZINGER, 2005): one may use it as a weighting factor or as a second dimension of the curve fitting (fitting then a surface; the interpretation may be more complex). The obtained models can be summarized by their parameters, which are used in the following steps.

4.3 Multivariate analyses

4.3.1 Classification

On the basis of his 25 prosodic descriptors, Gobbo (2019) applied multivariate methods so as to find the best combination of them able to separate the three labeled DMs categories (ALL, CNT, and INP). He found that a subset of three descriptors gave a global 74% accuracy for classifying these three categories. These descriptors were the mean intensity and F0, and the F0 slope up to the stressed syllable. Adding more descriptors led to higher accuracies (up to about 80%). However, by changing the set of selected parameters as we increase the number of parameters, one may get less reliable and interpretable results on unseen data.

Classification of DM's prosodic shapes has still not been done using FDA data (but is an ongoing process), but the feasibility of this approach has been demonstrated by Cavalcante (2020) for the forms of the Topic unit.

4.3.2 Qualitative analysis

On the same 25 prosodic descriptors (or potentially on the FDA parameters), one may pursue a descriptive analysis of the DMs that were not attributed a given function so as to regroup them on the basis of the prosodic characteristics to check if this leads to clusters that can then be interpreted at a functional level. Gobbo (2019) proposed a rapid overview of this possibility, sorting the AUX units[4] of his corpus into categories of shapes. A qualitative analysis of this data[5] reveals that, besides the forms

[4] Out of 564 tokens analyzed by Gobbo (2020) 414 tokens received the tag AUX so as to indicate that they were DMs (or Auxiliary Units) but that they could not, at that moment, be classified into a well-known category.

[5] Qualitative reanalysis of the 414 tokens left out of the main model proposed by Gobbo (2020). This work resulted in the regrouping of data in three main categories and in the exclusion of items due to hypersegmentation and quality issues (RASO; SANTOS, in preparation).

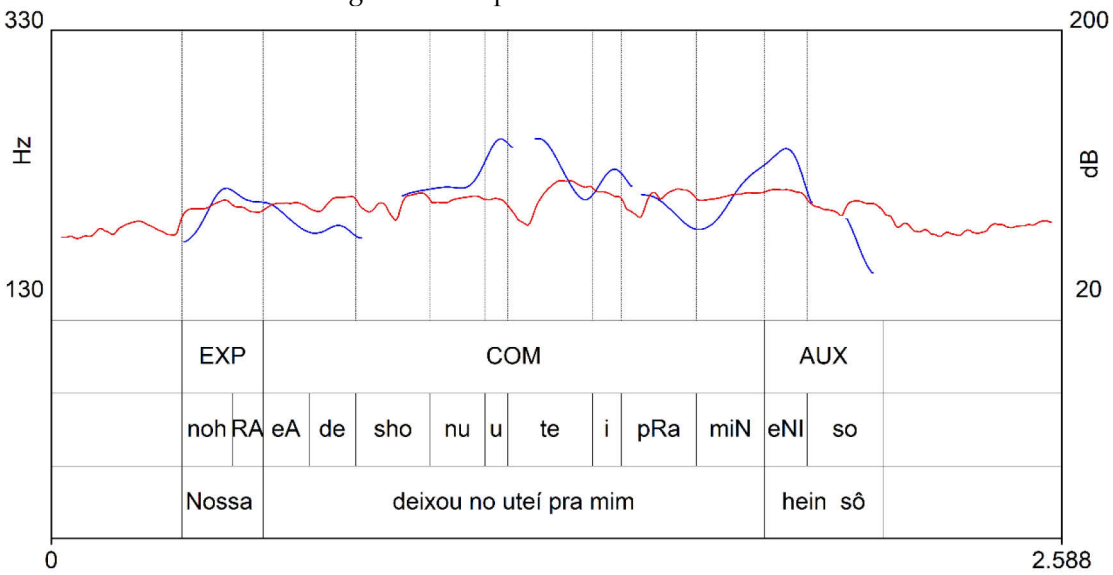
described for ALL, CNP, and INP, other three forms can be put forth that seem to bear prosodic and functional coherences. We give a brief overview of these forms, showing some prototypical examples and giving their functions. The reader is warned that this is an ongoing work and that a prosodic description (with acoustic measurements) is still needed in order to confirm the forms.

4.3.2.1 EXP

The first form is labeled under the tag EXP (Expressive), which is generally used to convey some surprise (but not as an illocution) or to give emotional support to the act being performed. It has a rising f0 shape on the stressed syllable, which can shortly fall in the presence of post-stress segmental material. There seems to be some lengthening on the stressed syllable (at least) with respect to the mean syllabic duration of COM (Comment unit), although this aspect needs to be confirmed. Its intensity seems to be on the same level as that of COM or a bit lower. It is frequently found at the very beginning of the terminated sequence (utterance/stanza) or at the beginning of reported speech.

Example 21: [audio 21] bfamcv03[138]
*TON: Nossa /=EXP= ea deixou no uteí pra mim /=COM= hein sô //=CNT=
Holy /=EXP= she left it in the UCI for me / you saw //

Figure 14 – Representation of audio 21.



Source: the authors.

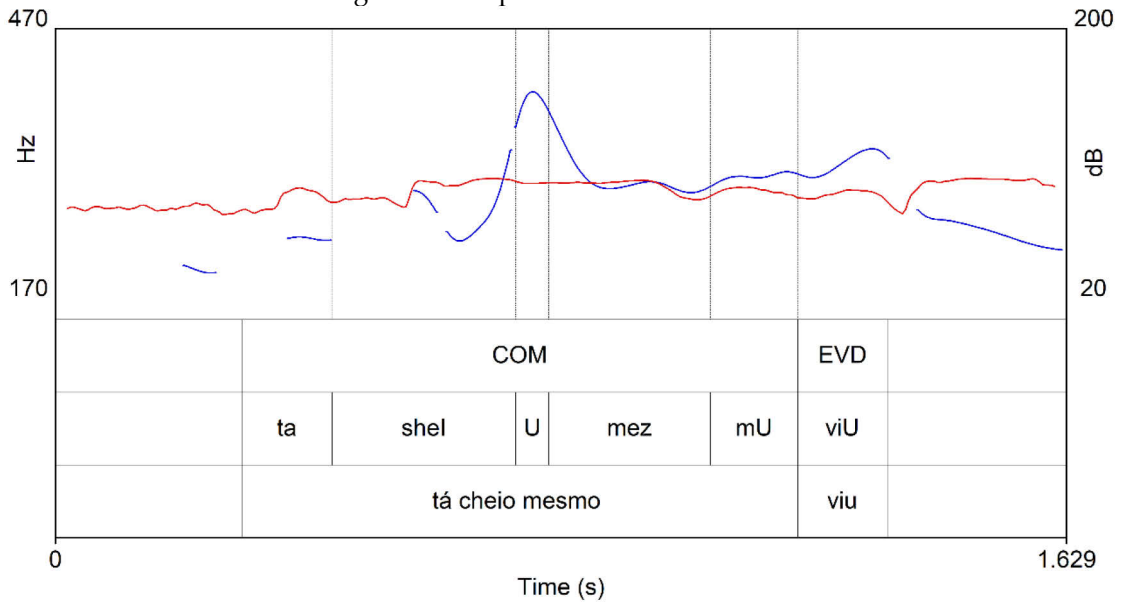
Frequent lexical fillers are *Nossa/No'* (Holy), *ah* (oh), and *não* (no).

4.3.2.2 EVD

The second form is tentatively tagged as EVD (Evidentiator - not to be confounded with grammatical markers of evidentiality). It can highlight what was said and secure, at the same time, the addressee's attention. It is characterized by a (generally) slightly rising f0 shape, lower intensity with respect to COM, and syllabic duration that tends to be shorter than that of COM. The slope of the rising movement can vary from almost flat to markedly slopy. All the same, it suffices to create a contrast with a falling f0 movement with the same f0 mean. This effect seems to be in connection with the very low intensity it is often produced with.

Example 22: [audio 22] bfamdl01[199]
*REN: tá cheio mesmo /=COM= viu //=EVD=
it's really crowded /=COM= huh //=EVD=

Figure 15 – Representation of audio 22.



Source: the authors.

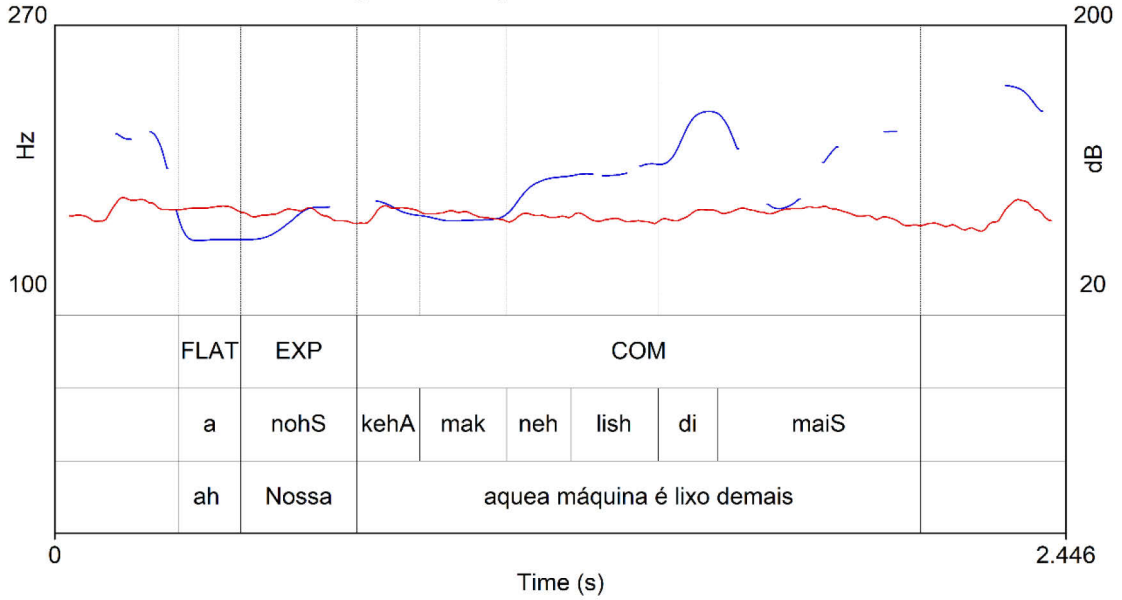
Frequent lexical items fulfilling this form are *né* (isn't it), *hein* (huh), *viu* (you see), and *sabe* (you know).

4.3.2.3 Initial flat form

The last form seems to convey a coherent function, but more data are needed for a better functional definition. We hypothesize that it may be used as a sign that the speaker has just realized something about what s/he is about to say. As to its prosodic form, this group has, as indicated by the tentative tag, a flat f0 profile, syllabic mean duration shorter than that of COM, and generally mean to high intensity levels. We found it exclusively in the initial position.

Example 23: [audio 23] bfammn05[102]
*JUN: ah /=FLAT= Nossa /=EXP= aquea máquina é lixo demais //=COM=
oh /=FLAT= Holy / that camera is complete trash //

Figure 16 – Representation of audio 23.



Source: the authors.

Frequent lexical fillers found for this form are *ah* (oh), *é* (yes), and *não* (no). It is noteworthy that the degree of pragmaticalization of these items can be shown by the fact that *yes* and *no* can be interchanged without detriment to the meaning of the utterance.

5 Conclusions & perspectives

In this paper, we propose a new way to identify DMs and their specific functions based on prosodic parameters after discussing different aspects of the mainstream literature on this topic. Prosody allows accounting for the main formal features that convey both the function of being a DM and the specific functions performed by different kinds of DMs. We showed how and why it is possible to establish when a lexical item behaves as DM, consequently giving a clear functional and formal definition of what should be considered a DM. We also described the specific formal prosodic cues that allow us to recognize the specific function of the

different types of DMs: at this point, we think that there are probably six types of DMs with interactional functions, for five of which we presented a more reliable picture.

We presented the methodology applied to the description of the different prosodic cues and the steps we have followed so far and will follow in order to automatically extract from new data the different kinds of DMs, discussing different strategies and the pros and contras of each of them.

The contribution of the paper is, therefore, twofold: it proposes a new linguistic solution to the problem of DM identification, and it shows how to model them in order to turn possible an automatic extraction of each specific functional item.

References

AIJMER, K. **Understanding pragmatic markers**: a variational pragmatic approach. Edinburgh: Edinburgh Univ. Press, 2013. DOI https://doi.org/10.1515/9780748635511

AIJMER, K.; SIMON-VANDENBERGEN, A.-M. (ed.). **Pragmatic markers in contrast**. Amsterdam: Elsevier, 2006. DOI https://doi.org/10.1163/9780080480299

D'ALESSANDRO, C. Voice source parameters and prosodic analysis. *In*: SUDHOFF, S.; LENERTOVA, D.; MEYER, R.; PAPPERT, S.; AUGURZKY, P.; MLEINEK, I.; RICHTER, N.; SCHLIESSER, J. (ed.). **Methods in empirical prosody research**. Berlin: Walter de Gruyter, 2006. p. 63–87.

D'ALESSANDRO, C.; MERTENS, P. Automatic pitch contour stylization using a model of tonal perception. **Computer Speech & Language**, vol. 9, no. 3, p. 257–288, 1995. DOI https://doi.org/10.1006/csla.1995.0013

D'ALESSANDRO, C.; RILLIARD, A.; LE BEUX, S. Chironomic stylization of intonation. **Journal of the Acoustical Society of America**, vol. 129, no. 3, p. 1594–1604, 2011. DOI https://doi.org/10.1121/1.3531802

D'ALESSANDRO, C.; FEUGÈRE, L.; LE BEUX, S.; PERROTIN, O.; RILLIARD, A. Drawing melodies: Evaluation of chironomic singing synthesis. **The Journal of the Acoustical Society of America**, vol. 135, no. 6, p. 3601–3612, 2014. DOI https://doi.org/10.1121/1.4875718

AUSTIN, J. L. **How to do things with words**. Oxford: Clarendon Press, 1962.

BALLY, C. **Linguistique générale et linguistique française**. 3rd ed. Berne: A. Francke, 1950.

BANSE, R.; SCHERER, K. R. Acoustic profiles in vocal emotion expression. **Journal of Personality and Social Psychology**, vol. 70, p. 614–636, 1996. DOI https://doi.org/10.1037/0022-3514.70.3.614

BARBOSA, P. A. From syntax to acoustic duration: A dynamical model of speech rhythm production. **Speech Communication**, vol. 49, no. 9, p. 725–742, 2007. DOI https://doi.org/10.1016/j.specom.2007.04.013

BARBOSA, P. A. Semi-automatic and automatic tools for generating prosodic descriptors for prosody research. **Proceedings from TRASP**, p. 86–90, 2013.

BAZZANELLA, C. Phatic connectives as interactional cues in contemporary spoken Italian. **Journal of Pragmatics**, vol. 14, no. 4, p. 629–647, Aug. 1990. DOI https://doi.org/10.1016/0378-2166(90)90034-B

BOERSMA, P.; WEENINK, D. **Praat**: doing phonetics by computer [Computer program]. Version 6.1.16, 2020. Available at: http://www.praat.org/.

BOLDEN, G. B. Discourse Markers. *In*: TRACY, K.; SANDEL, T.; ILIE, C. (ed.). **The International Encyclopedia of Language and Social Interaction**. 1st ed. Wiley, 2015. p. 1–7.

BRINTON, L. J. **Pragmatic Markers in English**: Grammaticalization and Discourse Functions. De Gruyter Mouton, 1996. DOI https://doi.org/10.1515/9783110907582

CAMACHO, A.; HARRIS, J. G. A sawtooth waveform inspired pitch estimator for speech and music. **The Journal of the Acoustical Society of America**, vol. 124, no. 3, p. 1638–1652, Sep. 2008. DOI https://doi.org/10.1121/1.2951592

CAMPBELL, W.N.; ISARD, S.D. Segment durations in a syllable frame. **Journal of Phonetics**, vol. 19, no. 1, p. 37–47, Jan. 1991. DOI https://doi.org/10.1016/S0095-4470(19)30315-8

CAVALCANTE, F. A. **The information unit of topic**: a crosslinguistic, statistical study based on spontaneous speech corpora. 2020. PhD Thesis – Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, 2020.

COLE, J. Prosody in context: a review. **Language, Cognition and Neuroscience**, vol. 30, no. 1–2, p. 1–31, 7 Feb. 2015. DOI https://doi.org/10.1080/23273798.2014.963130

CONTINI, M.; PROFILI, O. L'intonation de l'italien régional - un modèle de description par traits. 1989. **Mélanges de phonétique générale et expérimentale offerts à Péla Simon**. Strasbourg: Publications de l'Institut de phonétique de Strasbourg, 1989. p. 855–870.

CRESTI, E. **Corpus di italiano parlato**. Firenze: presso l'Accademia della Crusca, 2000.

CRESTI, E. La stanza: Un'unità di costruzione testuale del parlato. *In*: SILFI 2008, 2010. **Atti del X Congresso della Società Internazionale di Linguistica e Filologia Italiana**. Basel: Firenze: Cesati, 2010. p. 713–732.

CRESTI, E.; MONEGLIA, M. (ed.). **C-ORAL-ROM**: Integrated Reference Corpora for Spoken Romance Languages. vol. 15 (Studies in Corpus Linguistics). Amsterdam: John Benjamins Publishing Company, 2005. DOI https://doi.org/10.1075/scl.15

CRESTI, E.; MONEGLIA, M. The Discourse Connector according to the Language into Act Theory: data from IPIC Italiano. *In*: BIDESE, E.; CASALICCHIO, J.; MORONI, M. C. (ed.). **La linguistica vista dalle Alpi Linguistic views from the Alps**. Berlin: Peter Lang Verlag, 2020.

DAVIDSON, L. The Effects of Pitch, Gender, and Prosodic Context on the Identification of Creaky Voice. **Phonetica**, vol. 76, no. 4, p. 235–262, 1 Jul. 2019. DOI https://doi.org/10.1159/000490948

DE CASTRO MOUTINHO, L.; COIMBRA, R. L.; RILLIARD, A.; ROMANO, A. Mesure de la variation prosodique diatopique en portugais européen. **Estudios de fonética experimental**, vol. 20, p. 33–55, 2011.

DE CHEVEIGNÉ, A.; KAWAHARA, H. YIN, a fundamental frequency estimator for speech and music. **The Journal of the Acoustical Society of America**, vol. 111, no. 4, p. 1917–1930, 2002. DOI https://doi.org/10.1121/1.1458024

DEGAND, L. 7 'So very fast then' Discourse Markers at Left and Right Periphery in Spoken French. *In*: BEECHING, K.; DETGES, U. (ed.). **Discourse Functions at the Left and Right Periphery**, BRILL, 2014. p. 151–178. DOI https://doi.org/10.1163/9789004274822_008

DI CRISTO, A.; HIRST, DJ Modelling French Micromelody: Analysis and Synthesis. **Phonetica**, vol. 43, no. 1–3, p. 11–30, 1 Jan. 1986. DOI https://doi.org/10.1159/000261758

DUBĚDA, T.; KELLER, E. Microprosodic aspects of vowel dynamics—an acoustic study of French, English and Czech. **Journal of Phonetics**, vol. 33, no. 4, p. 447–464, Oct. 2005. DOI https://doi.org/10.1016/j.wocn.2005.02.003

EVANS, J.; CHU, M.; ASTON, J. A. D.; SU, C. Linguistic and human effects on F0 in a tonal dialect of Qiang. **Phonetica**, vol. 67, no. 1–2, p. 82–99, 2010. DOI https://doi.org/10.1159/000319380

EYBEN, F; SCHULLER, B. openSMILE:): the Munich open-source large-scale multimedia feature extractor. **ACM SIGMultimedia Records**, vol. 6, no. 4, p. 4–13, 28 Jan. 2015. DOI https://doi.org/10.1145/2729095.2729097

EYBEN, F.; SCHERER, K. R.; SCHULLER, B. W.; SUNDBERG, J.; ANDRE, E.; BUSSO, C.; DEVILLERS, L. Y.; EPPS, J.; LAUKKA, P.; NARAYANAN, S. S.; TRUONG, K. P. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. **IEEE Transactions on Affective Computing**, vol. 7, no. 2, p. 190–202, 1 Apr. 2016. DOI https://doi.org/10.1109/TAFFC.2015.2457417

FISCHER, K. (ed.). **Approaches to discourse particles**. Studies in pragmatics, 1. Oxford: Elsevier, 2006a. DOI https://doi.org/10.1163/9780080461588

FISCHER, K. Towards an understanding of the spectrum of approaches to discourse particles: introduction to the volume. *In*: FISCHER, K. (ed.). **Approaches to discourse particles**. Studies in pragmatics. Oxford: Elsevier, 2006b. p. 1–20. DOI https://doi.org/10.1163/9780080461588_002

FRANK-JOB, B. A dynamic-interactional approach to discourse markers. In: FISCHER, Kerstin (ed.). **Approaches to discourse particles**. Studies in pragmatics. Oxford: Elsevier, 2006. p. 395–413.

FROSALI, F. Il lessico degli Ausili Dialogici. *In*: CRESTI, E. (ed.). Prospettive nello studio del lessico italiano: **Atti del IX Congresso SILFI**. Proceedings e report. 1st ed. Florence: Firenze University Press, 2008. vol. 40, p. 417–424.

FUJISAKI, H. Dynamic characteristics of voice fundamental frequency in speech and singing. In: MACNEILAGE, P. (ed.). **The production of speech**. New York, NY: Springer, 1983. p. 39–55. DOI https://doi.org/10.1007/978-1-4613-8202-7_3

FUJISAKI, H. A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. *In*: FUJIMURA, O. (ed.). **Vocal fold physiology**: voice production, mechanisms and functions. New York, NY: Raven, 1988. p. 347–355.

GERRATT, B. R.; KREIMAN, J. Toward a taxonomy of non-modal phonation. **Journal of Phonetics**, vol. 29, no. 4, p. 365–381, Oct. 2001. DOI https://doi.org/10.1006/jpho.2001.0149

GOBBO, O. **Marcadores discursivos em uma perspectiva informacional**: análise prosódica e estatística. 2019. Master Thesis – Federal University of Minas Gerais, Belo Horizonte, Brazil, 2019.

GOUDBEEK, M.; SCHERER, K. Beyond arousal: Valence and potency control cues in the vocal expression of emotion. **The Journal of the Acoustical Society of America**, vol. 128, no. 3, p. 1322–1336, 2010. DOI https://doi.org/10.1121/1.3466853

GRABE, E.; KOCHANSKI, G.; COLEMAN, J. Connecting intonation labels to mathematical descriptions of fundamental frequency. **Language and speech**, vol. 50, no. 3, p. 281–310, 2007. DOI https://doi.org/10.1177/00238309070500030101

GURLEKIAN, J.; MIXDORFF, H.; EVIN, D.; TORRES, H.; PFITZINGER, H. R. Alignment of F0 model parameters with final and non-final accents in Argentinean Spanish. 2010. **Proceedings**. Speech Prosody 2010. Chicago, USA: 2010. p. paper 131.

HADJIPANTELIS, P. Z; ASTON, J. A. D.; EVANS, J. P. Characterizing fundamental frequency in Mandarin: A functional principal component approach utilizing mixed effect models. **The Journal of the Acoustical Society of America**, vol. 131, no. 6, p. 4651–4664, 2012. DOI https://doi.org/10.1121/1.4714345

HAMMARBERG, B.; FRITZELL, B.; GAUFIN, J.; SUNDBERG, J.; WEDIN, L. Perceptual and Acoustic Correlates of Abnormal Voice Qualities. **Acta Oto-**

Laryngologica, vol. 90, no. 1–6, p. 441–451, Jan. 1980. DOI https://doi.org/10.3109/00016488009131746

HERMES, D. J. Auditory and Visual Similarity of Pitch Contours. **Journal of Speech**, Language, and Hearing Research, vol. 41, no. 1, p. 63–72, 1998a. DOI https://doi.org/10.1044/jslhr.4101.63

HERMES, D. J. Measuring the Perceptual Similarity of Pitch Contours. **Journal of Speech, Language, and Hearing Research**, vol. 41, no. 1, p. 73–82, 1998b. DOI https://doi.org/10.1044/jslhr.4101.73

HIRST, D.; DI CRISTO, A. A survey of intonation systems. *In*: HIRST, D.; DI CRISTO, A. (ed.). **Intonation systems**: a survey of twenty languages. Cambridge, U.K.: Cambridge University Press, 1998. p. 1–44.

HIRST, D; DI CRISTO, A.; ESPESSER, R. Levels of Representation and Levels of Analysis for the Description of Intonation Systems. *In*: HORNE, M. (ed.). **Prosody**: Theory and Experiment. Text, Speech and Language Technology. vol. 14. Dordrecht: Springer Netherlands, 2000. p. 51–87. DOI https://doi.org/10.1007/978-94-015-9413-4_4

HIRST, D.; ESPESSER, R. Automatic Modelling of Fundamental Frequency Using a Quadratic Spline Function. **Travaux de l'Institut de Phonétique d'Aix**, vol. 15, p. 75–85, 1993.

HIRST, D.; RILLIARD, A.; AUBERGÉ, V. Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis. **The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis**. 1998.

HONDA, K. Physiological factors causing tonal characteristics of speech: from global to local prosody. 2004. **International Conference on Speech Prosody 2004**. Nara, Japan: 2004. p. 739–744.

KOCHANSKI, G.; SHIH, C. Prosody modeling with soft templates. **Speech Communication**, vol. 39, no. 3–4, p. 311–352, Feb. 2003. DOI https://doi.org/10.1016/S0167-6393(02)00047-X

KOHLER, K. J. What is emphasis and how is it coded? 2006. **Proceedings**. Speech Prosody 2006. Dresden, Germany: 2006. paper 015.

KREIMAN, J.; GERRATT, B. R.; ANTOÑANZAS-BARROSO, N. Measures of the Glottal Source Spectrum. **Journal of Speech, Language, and Hearing Research**, vol. 50, no. 3, p. 595–610, Jun. 2007. https://doi.org/10.1044/1092-4388(2007/042)

LAI, C. Interpreting final rises: task and role factors. **7th International Conference on Speech Prosody**, Dublin, Ireland, 2014. p. 520–524.

LEVITT, H.; RABINER, L. R. Analysis of fundamental frequency contours in speech. **The Journal of the Acoustical Society of America**, vol. 49, p. 569, 1971. DOI https://doi.org/10.1121/1.1912388

LIÉNARD, J.-S.; DI BENEDETTO, M.-G. Effect of vocal effort on spectral properties of vowels. **The Journal of the Acoustical Society of America**, vol. 106, no. 1, p. 411–422, Jul. 1999. DOI https://doi.org/10.1121/1.428140

LIÉNARD, J.-S. Quantifying vocal effort from the shape of the one-third octave long-term-average spectrum of speech. **The Journal of the Acoustical Society of America**, vol. 146, no. 4, p. EL369–EL375, Oct. 2019. DOI https://doi.org/10.1121/1.5129677

LÖFQVIST, A.; MANDERSSON, B. Long-Time Average Spectrum of Speech and Voice Analysis. **Folia Phoniatrica et Logopaedica**, vol. 39, no. 5, p. 221–229, 1987. DOI https://doi.org/10.1159/000265863

MACARY, M.; TAHON, M.; ESTEVE, Y.; ROUSSEAU, A. On the Use of Self-Supervised Pre-Trained Acoustic and Linguistic Features for Continuous Speech Emotion Recognition. *In*: **IEEE SPOKEN LANGUAGE TECHNOLOGY WORKSHOP (SLT)**, Shenzhen, China, 2021. p. 373–380. DOI https://doi.org/10.1109/SLT48900.2021.9383456

MAIA ROCHA, B.; RASO, T. A unidade informacional de Introdutor Locutivo no português do Brasil: uma primeira descrição baseada em corpus. **Domínios de Lingu@gem**, vol. 5, no. 1, p. 327–343, Jul. 2011. Available at: https://seer.ufu.br/index.php/dominiosdelinguagem/article/view/12479

MARTIN, Ph. Les problèmes de l'intonation: recherches et applications. **Langue française**, vol. 19, no. 1, p. 4–32, 1973. DOI https://doi.org/10.3406/lfr.1973.5638

MARTIN, Ph. Prosodic and rhythmic structures in French. **Linguistics**, vol. 25, no. 5, p. 925–950, 1987.

MARTIN, Ph. Multi methods pitch tracking. *In*: **Speech Prosody**, 2012. Shanghai, China, 2012, p. 47–50.

MERTENS, P. The prosogram: semi-automatic transcription of prosody based on a tonal perception model. **International Conference on Speech Prosody 2004**, Nara, Japan, 2004, p. 549–552. DOI https://doi.org/10.20396/joss.v4i2.15053

MERTENS, P. Polytonia: a system for the automatic transcription of tonal aspects in speech corpora. **Journal of Speech Sciences**, vol. 4, no. 2, p. 17–57, 5 Feb. 2014.

MEUNIER, S.; CHATRON, J.; ABS, B.; PONSOT, E.; SUSINI, P. Effect of Pitch on the Asymmetry in Global Loudness Between Rising- and Falling-Intensity Sounds. **Acta Acustica united with Acustica**, vol. 104, no. 5, p. 770–773, 1 Sep. 2018. DOI https://doi.org/10.3813/AAA.919220

MITTMANN, M. M. **O C-ORAL-BRASIL e o estudo da fala informal**: um novo olhar sobre o tópico no português brasileiro. 248 f. PhD Thesis – Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, 2012.

MIXDORFF, H. A novel approach to the fully automatic extraction of Fujisaki model parameters. Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. **Proceedings**. 2000 IEEE International Conference on Speech and Signal Processing, IEEE, 2000. vol. 3. p. 1281–1284.

MIXDORFF, H.; PFITZINGER, H. R. Analysing fundamental frequency contours and local speech rate in map task dialogs. **Speech Communication**, vol. 46, no. 3–4, p. 310–325, Jul. 2005. DOI https://doi.org/10.1016/j.specom.2005.02.019

MONEGLIA, M.; RASO, T. Appendix: Notes on the Language into Act Theory. *In*: RASO, T.; MELLO, H. (eds.). **Studies in Corpus Linguistics**. Amsterdam: John Benjamins Publishing Company, 2014. vol. 61, p. 468–495. DOI https://doi.org/10.1075/scl.61.15mon

MOORE, B. C. J.; GLASBERG, B. R.; VARATHANATHAN, A.; SCHLITTENLACHER, J. A Loudness Model for Time-Varying Sounds Incorporating Binaural Inhibition. **Trends in Hearing**, vol. 20, Jan. 2016. DOI https://doi.org/10.1177/2331216516682698

NERBONNE, J.; COLEN, R.; GOOSKENS, C. S.; LEINONEN, T.; KLEIWEG, P. Gabmap – A web application for dialectology. **Dialectologia**, vol. SI II, p. 65–89, 2011.

NORDENBERG, M.; SUNDBERG, J. Effect on LTAS of vocal loudness variation. **Logopedics Phoniatrics Vocology**, vol. 29, no. 4, p. 183–191, Dec. 2004. DOI https://doi.org/10.1080/14015430410004689

PIERREHUMBERT, J. B. **The phonology and phonetics of English intonation**. PhD Thesis. Massachusetts Institute of Technology, 1980.

PIERREHUMBERT, J. Synthesizing intonation. **The Journal of the Acoustical Society of America**, vol. 70, no. 4, p. 985–995, Oct. 1981. DOI https://doi.org/10.1121/1.387033

RAMSAY, J. O.; SILVERMAN, B. W. **Functional data analysis**. 2nd ed. New York: Springer, 2005. DOI https://doi.org/10.1007/b98888

RASO, T. Prosodic constraints for discourse markers. *In*: RASO, T.; MELLO, H. (ed.). **Studies in Corpus Linguistics**. vol. 61. Amsterdam: John Benjamins Publishing Company, 2014. p. 411–467. DOI https://doi.org/10.1075/scl.61.14ras

RASO, T.; CAVALCANTE, F. A.; MITTMANN, M. M. Prosodic forms of the Topic information unit in a cross–linguistic perspective. A first survey. *In*: DE MEO, A.; DOVETTO, F. M. (ed.). **La comunicazione parlata / Spoken Communication**. Rome: Aracne, 2017. p. 473–498.

RASO, T.; FERRARI, L. A. Uso dei Segnali Discorsivi in corpora di parlato spontaneo italiano e brasiliano. *In*: FERRONI, R.; BIRELLO, M. (ed.). **La competenza discorsiva e interazionale: a lezione di lingua straniera**. Canterano (Roma): Aracne, 2020. p. 61–107.

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

RASO, T.; VIEIRA, M. A. A description of Dialogic Units/Discourse Markers in spontaneous speech corpora based on phonetic parameters. **CHIMERA**: Romance Corpora and Linguistic Studies, vol. 3, no. 2, p. 221–249, 2016.

RENCHER, A. C.; CHRISTENSEN, W. F. **Methods of multivariate analysis**. Third Edition. Hoboken, New Jersey: Wiley, 2012. DOI https://doi.org/10.1002/9781118391686

RILLIARD, A. Geoprosody – Quantitative approaches of prosodic variation across dialects. *In*: VIEIRA, M. S. M.; WIEDEMER, M. L. (ed.). **Dimensões e Experiências em**

Sociolinguística. São Paulo: Editora Blucher, 2019. p. 55–83. DOI https://doi.org/10.5151/9788521218746-02

ROCHA, B.; MELLO, H.; RASO, T. Para a compilação do C-ORAL-ANGOLA. **Filologia e Linguística Portuguesa**, vol. 20, no. Especial, p. 139–157, 30 Dec. 2018. DOI https://doi.org/10.11606/issn.2176-9419.v20iEspecialp139-157

ROMERO-TRILLO, J. Discourse Markers. *In*: MEY, J. (ed.). **Concise encyclopedia of pragmatics**. Amsterdam; New York: Elsevier, 1998. p. 191–194.

ROSSI, M. Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. **Phonetica**, vol. 23, no. 1, p. 1–33, 1971. DOI https://doi.org/10.1159/000259328

SCHIFFRIN, D. **Discourse Markers**. Cambridge University Press, 1987. DOI https://doi.org/10.1017/CBO9780511611841

SCHOURUP, L. Discourse markers. **Lingua**, vol. 107, no. 3–4, p. 227–265, Apr. 1999. DOI https://doi.org/10.1016/S0024-3841(96)90026-1

SIGNOL, F.; BARRAS, C.; LIENARD, J.-S. Evaluation of the Pitch Estimation Algorithms in the monopitch and multipitch cases. *In*: **Proceedings of Acoustics'08**. Paris, France: May 2008. p. 675–680.

SIGNORINI, S. **Topic e soggetto in corpora di italiano parlato spontaneo**. 2005. PhD Thesis. Università di Firenze, Firenze, Italy, 2005.

SUNDBERG, J.; NORDENBERG, M. Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. **The Journal of the Acoustical Society of America**, vol. 120, no. 1, p. 453–457, Jul. 2006. DOI https://doi.org/10.1121/1.2208451

ŠVEC, J. G.; GRANQVIST, S. Tutorial and Guidelines on Measurement of Sound Pressure Level in Voice and Speech. **Journal of Speech, Language, and Hearing Research**, vol. 61, no. 3, p. 441–461, 15 Mar. 2018. DOI https://doi.org/10.1044/2017_JSLHR-S-17-0095

'T HART, J.; COLLIER, R.; COHEN, A. **A perceptual study of intonation**: an experimental-phonetic approach to speech melody. Cambridge: Cambridge University Press, 1990. DOI https://doi.org/10.1017/CBO9780511627743

TALKIN, D. A robust algorithm for pitch tracking (RAPT). *In*: KLEIJN, W. Bastiaan; PALIWAL, K. K. (ed.). **Speech coding and synthesis**. Amsterdam: Elsevier, 1995. p. 495–518.

TITZE, I. R.; SUNDBERG, J. Vocal intensity in speakers and singers. **The Journal of the Acoustical Society of America**, vol. 91, no. 5, p. 2936–2946, May 1992. DOI https://doi.org/10.1121/1.402929

TRAUGOTT, E. C. Intersubjectification and clause periphery. **English Text Construction**, vol. 5, no. 1, p. 7–28, 4 May 2012. DOI https://doi.org/10.1075/etc.5.1.02trau

TRAUNMÜLLER, H.; ERIKSSON, A. Acoustic effects of variation in vocal effort by men, women, and children. **The Journal of the Acoustical Society of America**, vol. 107, no. 6, p. 3438–3451, Jun. 2000. DOI https://doi.org/10.1121/1.429414

TUCCI, I. L'inciso: caratteristiche morfosintattiche e intonative in un corpus di riferimento. *In*: ALBANO LEONI, F.; SENZA PELUSO, M. (ed.). **Il parlato italiano**: atti del convegno nazionale (Napoli, 13-15 febbraio 2003). Napoli: M. D'Auria editore, 2004. p. 1–14.

TUCCI, I. «Obiter dictum» La funzione informativa delle unità parentetiche. *In*: PETTORINO, M.; GIANNINI, A.; DOVETTO, F. M. (eds.). **La comunicazione parlata 3**. (Napoli, 23-25 febbraio 2009). Napoli: Liguori, 2010. p. 635–654.

VENABLES, W N; RIPLEY, B D. **Modern Applied Statistics with S**. Fourth edition. Springer, 2002. DOI https://doi.org/10.1007/978-0-387-21706-2

VINCENT, D.; ROSEC, O.; CHONAVEL, T. Glottal Closure Instant Estimation using an Appropriateness Measure of the Source and Continuity Constraints. *In*: **2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings**. vol. 1. Toulouse, France: IEEE, 2006. p. I-381-I–384.

XU, Y. Speech melody as articulatorily implemented communicative functions. **Speech Communication**, vol. 46, no. 3–4, p. 220–251, Jul. 2005. DOI https://doi.org/10.1016/j.specom.2005.02.014

Article received in: 12.21.2021 Article approved in: 21.03.2022

DOI: 10.14393/DL52-v16n4a2022-9

Bases lexicais verbais do português brasileiro

Brazilian Portuguese verbal databases

Roana RODRIGUES[*]
Marcella LEMOS-COUTO[**]
Francimeire Leme COELHO[***]
Isaac Souza de MIRANDA JUNIOR[****]
Oto VALE[*****]

RESUMO: Este artigo apresenta o levantamento e a análise de bases lexicais verbais do português brasileiro (PB), que podem ser utilizadas em tarefas de Processamento de Língua Natural (PLN). Para tanto, considerou-se para análise apenas bases de dados com extensão superior a 1.000 lexemas verbais, disponíveis de maneira on-line e gratuita e atualizadas nos últimos 10 anos. Sendo assim, o estudo recaiu sob a análise crítica e comparada de três bases lexicais: VerbNet.Br (SCARTON, 2013), Verbo-Brasil (DURAN; ALUÍSIO, 2015) e VerboWeb (CANÇADO *et al.*, 2018), destacando-se seus pontos comuns e divergentes. Acredita-se que esta pesquisa contribui com a atualização do estado da arte, no que se refere às bases lexicais verbais do PB da última década,

ABSTRACT: This paper presents a survey and an analysis of Brazilian Portuguese lexical verbal databases, which are possible to use in Natural Language Processing (NLP) tasks. For this purpose, we considered for analysis only databases with a dimension greater than 1.000 verbal lexemes, free online access, and updated in the last 10 years. Therefore, the study fell on the critical and comparative analysis of three lexical databases: VerbNet.Br (SCARTON, 2013), Verbo-Brasil (DURAN; ALUÍSIO, 2015) and VerboWeb (CANÇADO *et. al.*, 2018), highlighting their commonalities and divergences. It is believed that this research contributes to updating the state-of-the-art, regarding the BP lexical verbal databases of the last decade, in addition to listing future investigations to create,

[*] Doutora em Linguística pela UFSCar. Professora do Departamento de Letras Estrangeiras da UFS. ORCID: https://orcid.org/0000-0002-7748-8716. roana@academico.ufs.br.

[**] Doutoranda em Linguística pela UFSCar. ORCID: https://orcid.org/0000-0001-9669-9458. marcella.couto@estudante.ufscar.br.

[***] Mestranda em Linguística pela UFSCar. ORCID: https://orcid.org/0000-0002-9333-9534. flcoelho@estudante.ufscar.br.

[****] Mestrando em Linguística pela UFSCar. ORCID: https://orcid.org/0000-0002-4004-3182. isaacmiranda@estudante.ufscar.br.

[*****] Doutor em Linguística e Língua Portuguesa pela UNESP. Professor do Departamento de Letras da UFSCar. ORCID: https://orcid.org/0000-0002-0091-8079. otovale@ufscar.br.

além de elencar ações investigativas futuras para criação, revisão e/ou ampliação de recursos descritivos linguísticos do PB.

revise and/or extend BP descriptive linguistic resources.

PALAVRAS-CHAVE: Processamento de Língua Natural. Lexicologia. Bases de dados verbais.

KEYWORDS: Natural Language Processing. Lexicology. Verbal databases.

1 Introdução

O interesse pela descrição dos verbos é uma constante nos estudos descritivos do português brasileiro, doravante PB. Pelo fato de o verbo ser o tipo de palavra com maior produtividade morfológica, ser, *grosso modo*, o elemento central nas frases, articulando os demais elementos, e ser o portador de informações como tempo, modo, aspecto, o verbo é sempre alvo de estudos nas mais variadas abordagens.

Nas gramáticas tradicionais, como as de Cunha e Cintra (2008 [1984]) e Azeredo (2008), o verbo ocupa um capítulo importante, sendo classificado por suas propriedades morfossintáticas, podendo ser regular, irregular, defectivo, abundante, auxiliar, principal/pleno, impessoal, pronominal, de ligação/copulativo, transitivo e intransitivo. Nas mais distintas abordagens teórico-metodológicas, há estudos linguísticos que visam à descrição e categorização dos verbos, segundo não só suas propriedades morfossintáticas, como também por suas características semânticas (CANÇADO; GODOY, 2012; RASSI; VALE, 2013; PERINI, 2016; CANÇADO; AMARAL, 2016).

O interesse descritivo sobre as particularidades dos verbos verifica-se também em trabalhos lexicográficos especializados. Sobre PB, pode-se mencionar, como uma obra de referência, o *Dicionário de verbos e regimes*, de Francisco Fernandes, publicado em 1940. De inquestionável valor, o dicionário abrange a definição de mais de 12 mil verbos, com abonações de textos literários. Em 1990, outra obra de igual repercussão

foi publicada, o *Dicionário Gramatical de Verbos do Português Contemporâneo do Brasil*, elaborado sob coordenação de Francisco da Silva Borba. Nessa obra, aplicam-se conceitos da Linguística de Corpus para a sua produção, culminando em uma seleção mais criteriosa dos verbos descritos. O dicionário apresenta cerca de 6 mil lexemas, com a proposta de uma classificação semântica para os verbos, categorizando-os como verbos de ação, processo, ação-processo e estado. Mais recentemente, em 2013, tendo como base os dados do dicionário de Borba (1990), foi publicado o *Catálogo de Verbos do Português Brasileiro*, coordenado por Márcia Cançado. O catálogo, que contém mais de 800 verbos de mudança do PB, organiza os lexemas verbais em 4 classes, de acordo com as suas propriedades sintático-semânticas. Essas obras lexicográficas compartilham tanto o empenho pela catalogação dos verbos do PB, quanto o público-alvo, que é o usuário comum.

Em consonância com as obras citadas, a presente pesquisa se preocupa com a descrição, sistematização e classificação dos verbos do PB, no entanto, direciona sua atenção a recursos descritivos que podem ser utilizados em tarefas de Processamento de Língua Natural (PLN). Segundo Greghi (2002), as aplicações em PLN se realizam a partir de recursos linguísticos descritivos, como analisadores sintáticos (*parsers*), *corpora*, dicionários e bases lexicais. Para a autora,

> "[as bases lexicais são] volumosas e abrangentes, compreendendo vários atributos linguísticos para cada item lexical, e não necessariamente servindo a uma aplicação específica, mas à centralização e organização das informações lexicais, a fim de apoiar a pesquisa e o desenvolvimento de aplicações de PLN para uma dada língua" (GREGHI, 2002, p. 2).

Portanto, as bases lexicais atuam como recurso basilar para a construção de ferramentas computacionais como revisores ortográficos e gramaticais, além de sumarizadores e tradutores automáticos, por exemplo. Salienta-se que essas bases

podem ser úteis também para os usuários comuns, diante da possibilidade de realização de consultas, daí a importância da construção de repositórios com uma interface de fácil interação.

Esta pesquisa visa, portanto, contribuir com a atualização do estado da arte, no que se refere às bases lexicais verbais do PB desta década. Para tanto, realizou-se o levantamento das bases, a partir de buscas na web, de acordo com os seguintes critérios: (i) extensão: a base deve conter pelo menos 1.000 lexemas verbais descritos; (ii) acesso: a base deve estar disponível para consulta on-line, de maneira gratuita; e (iii) atualização: a base deve ter sofrido atualizações nos últimos 10 anos. Ao todo, foram elencadas três bases lexicais, as quais serão analisadas nas próximas seções deste artigo, a saber: VerbNet.Br, Verbo-Brasil e VerboWeb.

De acordo com as particularidades de cada base, e considerando os seus arcabouços teórico-metodológicos, esta pesquisa apresenta também uma análise crítica sobre as bases de dados verbais do PB, com o intuito de discutir os seus pontos comuns e divergentes, assim como explorar as informações sintático-semânticas disponibilizadas por cada uma, com vistas a traçar planos para melhorias que possam ocorrer na área.

Sendo assim, este artigo organiza-se da seguinte maneira: na seção dois, são apresentadas as bases lexicais verbais selecionadas; em seguida, realiza-se uma análise qualitativa e comparativa sobre esses trabalhos, considerando os seus objetivos e delimitações. Por fim, são descritas as considerações finais desta pesquisa, assim como os encaminhamentos para investigações e ações futuras.

2 Bases de dados verbais do PB

A construção de bases lexicais é uma tarefa primorosa, que demanda tempo e mão de obra especializada. A participação ativa de linguistas se faz necessária, diante

da necessidade de uma descrição e/ou revisão pormenorizada de diferentes fenômenos da língua. Posto isso, elucida-se que este trabalho prestigia as atividades realizadas nos projetos descritos, que viabilizam reflexões sobre o comportamento morfossintático-semântico dos verbos do PB, além de atuarem como ricos repositórios descritivos que podem ser utilizados em diferentes empreendimentos na área de PLN.

As próximas seções apresentam as bases lexicais verbais selecionadas, com informações detalhadas sobre: (i) desenvolvedores; (ii) fundamentação teórico-metodológica; (iii) exemplo de descrição linguística; (iv) classificações e dados quantitativos; e (v) interface.

2.1 VerbNet.Br

A VerbNet (SCHULER, 2005) é um recurso léxico-computacional (RLC) que agrupa verbos do inglês em diferentes classes semânticas inspiradas nas classes propostas por Levin (1993). Em seu trabalho, Levin (1993) agrupou os verbos do inglês em 51 grandes classes (com diversas subclasses) de acordo com as alternâncias sintáticas e características semânticas que esses verbos compartilham. A hipótese é de que há uma relação entre os sentidos do verbo e seu comportamento sintático. A partir do RLC do inglês foram criados outros para várias línguas como espanhol, francês, alemão etc.

Para o PB, a pesquisadora Carolina Evaristo Scarton, durante seu Mestrado em Ciências da Computação e Matemática Computacional da Universidade de São Paulo, criou a VerbNet.Br[1], que também parte da VerbNet do inglês e que conta com 1.766 unidades lexicais verbais e 4.333 sentidos de verbos, distribuídos em 202 classes[2].

[1] VerbNet.Br. Disponível em: http://143.107.183.175:21380/verbnetbr/index.html. Acesso em: 29 jun. 2021.
[2] Salienta-se que o RLC do inglês possui 3.769 unidades lexicais verbais e 274 classes.

A metodologia proposta por Scarton (2013) foi fundamentada em quatro etapas: uma etapa manual e três, realizadas de maneira automática. A primeira etapa, única realizada de maneira manual, consistiu na definição das alternâncias sintáticas do português para cada classe presente na VerbNet do inglês. A segunda, propôs a busca automática das alternâncias sintáticas em um *corpus* etiquetado. A terceira etapa serviu para definir automaticamente os candidatos a membros das classes por meio de alinhamentos entre VerbNet, WordNet e WordNet.Br[3]. Por fim, a quarta e última etapa consistiu em escolher, de fato, os membros das classes. Por conta dessa abordagem *cross-linguística* que foi utilizada na criação do recurso, as classes da VerbNet.Br apresentam-se em inglês. Para uma melhor exemplificação das informações presentes nas classes verbais, no Quadro 1, são apresentados os componentes presentes na classe *disassemble-23.3*.

Quadro 2 – Informações sintático-semânticas da classe disassemble-23.3 da VerbNet.

disassemble-23.3	
Members: 39, Frames: 3	
Papéis temáticos e [restrições seletivas]	
Agent [+animate | +machine]	
Patient [+concrete]	
Co- Patient [+concrete]	
Frames	
VerbNet.Br	**VerbNet**
V_NP	NP V NP - Basic Transitive
V_NP_PP[de]	NP V NP PP.patient2 - NP-PPSource-PP

Fonte: VerbNet.Br. Disponível em: http://verbs.colorado.edu/verb-index/vn/disassemble-23.3.php. Acesso em: 29 jun. 2021.

A classe reproduzida acima, com informações disponíveis na versão do PB, é composta por 39 membros e 3 frames. Os papéis temáticos e as restrições de seleção da

[3] WordNets são bases de dados que visam sistematizar conjuntos de substantivos, adjetivos, advérbios e verbos em quatro relações: sinonímia, antonímia, hiponímia/hiperonímia e meronímia/holonímia. (SCARTON, 2013, p. 50).

classe são, respectivamente, Agente, com restrição seletiva "animado" e "máquina"; e Paciente e Co-Paciente, com a restrição que seleciona, para esses papéis temáticos, uma entidade "concreta". Ainda no Quadro 1, são apresentados os frames sintáticos e alternâncias de transitividade tanto do português (VerbNet.Br), quanto do inglês (VerbNet).

A VerbNet do inglês contém uma lista de 23 papéis temáticos para dar conta dos argumentos selecionados pelos verbos; todos esses papéis foram incorporados na criação da VerbNet.Br. As restrições seletivas que, por sua vez, são impostas aos papéis temáticos, também não sofreram alterações na versão do PB. Os frames sintáticos, que descrevem, além da transitividade verbal, outros itens lexicais selecionados em alternâncias em particular (como a exigência de preposições), sofrem alterações em suas traduções. Por fim, é possível consultar os predicados semânticos, que fornecem informações sobre as relações entre os participantes e o evento da ação verbal. Os predicados semânticos são divididos em quatro classes:

> **Predicados gerais:** inclui predicados como *motion* (movimento) e *cause* (causa) e são genéricos em uma grande quantidade de classes e, também, em várias línguas.
> **Predicados variáveis:** predicados cujo significado admite uma relação um-por-um com um conjunto de palavras em uma língua.
> **Predicados específicos:** carregam um sentido verbal específico.
> **Predicados para múltiplos eventos:** são predicados usados para expressar relações entre eventos. (SCARTON, 2013, p. 65-66).

No Quadro 2 abaixo, estão reproduzidos os *frames* da classe *disassemble-23.3* e, em seguida, são discutidas as relações semânticas para melhor entendimento da metalinguagem semântica empregada.

Quadro 3 – Exemplo de *frames* da classe *disassemble-23.3* na VerbNet.

FRAMES	
NP V NP	
example	*"I unscrewed the handle."*
syntax	*Agent V Patient*
semantics	*cause(Agent, E) together(start(E), physical, Patient, ?Co-Patient) apart(result(E), physical, Patient, ?Co-Patient)*
NP V NP PP.co-patient	
example	*"I unscrewed the handle from the box."*
syntax	*Agent V Patient {from} Co-Patient*
semantics	*cause(Agent, E) together(start(E), physical, Patient, Co-Patient) apart(result(E), physical, Patient, Co-Patient)*
NP V ADV-Middle	
example	*"That new handle unscrews easily."*
syntax	*Patient <+plural> V ADV*
semantics	*property(Patient, Prop) Adv(Prop)*

Fonte: VerbNet. Disponível em: http://verbs.colorado.edu/verb-index/vn/disassemble-23.3.php. Acesso em: 29 jun. 2021.

Como é possível observar, a VerbNet utiliza uma metalinguagem de representação semântica, a qual inclui informação temporal que desempenha a função de denotar se "o predicado é verdadeiro no início do evento (*start*(E)), na preparação do evento (*during*(E)), no final do evento (*end*(E)) ou no resultado do evento (*result*(E))" (SCARTON, 2013, p. 65). Sendo assim, é retomado, em (1a), seguido de sua tradução livre ao PB (1b), o primeiro *frame* (NP V NP):

(1) a. I unscrewed the handle.
b. Eu desparafusei a maçaneta.

Para uma interpretação da metalinguagem *cause(Agent,E) together(start(E), physical, Patient, ?Co-Patient) apart(result(E), physical, Patient, ?Co-Patient)*, em uma aplicação para (1b), leia-se: o Agente causa *(cause)* um evento que, em seu início (*start),*

é verdadeiro que Paciente e Co-paciente estavam unidos (*together*) fisicamente (*physical*); e, como resultado do evento verbal (*result*), é verdadeira a separação (*apart*) física (*physical*) de Paciente e Co-paciente, aqui, a "maçaneta desparafusada". É possível notar que há uma interrogação antes de *Co-patient* na representação da metalinguagem, justamente porque o argumento do Co-paciente não é expresso nesse *frame* sintático. Ao passo que, no segundo *frame* da classe *disassemble-23.3* (**NP V NP PP.co-patient**), em sua equivalência do *frame* em português (**V_NP_PP[de]**), o Co-paciente é um argumento obrigatório e introduzido por preposição "de", como em (1c):

(1) c. Eu desparafusei a maçaneta da porta.

Em Scarton (2013), foram avaliadas 16 classes de verbos com base em um *golden standard* produzido manualmente, tendo sido realizados 5 experimentos diferentes. Em seus resultados, todas as classes apresentaram uma boa cobertura (de até 96,17% em um dos experimentos), mas a maioria foi avaliada com uma precisão muito baixa (de, no máximo, 44,99% de precisão entre as 16 classes). A maior parte das falhas e erros de classificação da VerbNet.Br são decorrentes do caráter predominantemente automático de sua construção. É o que se verifica, por exemplo, na polissemia relacionada ao verbo *trabalhar*, que aparece como pertencente a quatro classes verbais distintas na VerbNet.Br, conforme é reproduzido no Quadro 3.

Quadro 4 – Verbo *trabalhar* na VerbNet.Br.

TRABALHAR				
Classes	build-26.1	carve-21.2	disassemble-23.3	other_cos-45.4
Papéis temáticos	**Agent Asset Beneficiary Material Product**	Agent Instrument Patient	Agent Patient Patient1 Patient2	**Agent Instrument Patient**
Restrições seletivas	Agent [+animate | +machine] Material [+concrete] Product Beneficiary [+animate | +organization] Asset [+currency]	Agent [+int_control] Patient [+concrete] Instrument [+concrete]	Agent [+animate | +machine] Patient [+concrete] Co-Patient [+concrete]	Agent [+int_control] Patient Instrument Result
Alternâncias	V_SN, V, V_SN_SP[em], V_SN_SP[de], V_SN_SP[com], V_SN_SP[em], V_SN_SP[para], V_SN_SP[com]_SP[para], V_SN_SP[de]_SP[para], V_SN_SP[em]_SP[para], V_SN_SP[para], V_SN_SP[em], V_SN_SP[em]_SP[por], V_SN_SP[com]_SP[por], V_SN_SP[de]_SP[por].	all.	V_SN, V_SN_SP[de].	V_SN, V_SN_SP[com], V, V_SN, V_SN_SP[em], V_SN_SP[em]_SP [com].

Fonte: VerbNet.Br. Disponível em: http://www.nilc.icmc.usp.br/verbnetbr/glossario.php?letra=t. Acesso em: 29 jun. 2021.

O principal critério para a classificação dos verbos na VerbNet e na VerbNet.Br é o conjunto de alternâncias sintáticas que um determinado grupo de verbos admite. Na VerbNet.Br, esse conjunto de alternâncias é definido para cada classe. Existem, no entanto, 33 classes cujas alternâncias são definidas como "all", ou seja, admitem todas as alternâncias sintáticas possíveis, é o caso da classe *carve 21.2*. Por outro lado, há também classes muito genéricas, nas quais é difícil identificar a relação semântica a que está submetida. É o caso, por exemplo, da classe *other_cos-45.4*, que possui 368 membros diversos, tais como *adormecer*, *apertar*, *oxidar*, *estrangular*, *explodir*, *moer* etc. Tais classes demonstram que a base apresenta a necessidade de refinamento da classificação.

A interface dessa base de dados verbais é simples e intuitiva. As buscas são organizadas por ordem alfabética e podem ser realizadas tanto por verbos, quanto por classes. Informações como restrições seletivas, predicados semânticos e subclasses estão disponíveis somente na versão do inglês, no entanto, existe o direcionamento às páginas por meio de *hiperlinks* nos títulos das classes da VerbNet.Br. Algo importante a ressaltar é que, para os *frames* sintáticos presentes em cada uma dessas classes, não há exemplos reais do PB, diferente do RLC do inglês, que apresenta uma frase de exemplo para cada alternância sintática representada por novo *frame* sintático.

2.2 Verbo-Brasil

A Verbo-Brasil (DURAN; ALUÍSIO, 2015)[4] é uma base de dados composta de 1.093 verbos com frequência acima de 1.000 no *corpus* PLN-Br, com a finalidade de apoiar a tarefa de anotação de papéis semânticos[5] nos projetos PropBank-Br v1 e PropBank-Br v2 (DURAN; ALUÍSIO, 2012)[6].

A Verbo-Brasil foi concebida pela pesquisadora Magali Sanches Duran em uma equipe supervisionada pela pesquisadora Sandra Maria Aluísio, ambas integrantes do Núcleo Interinstitucional de Linguística Computacional (NILC), da Universidade de São Paulo (USP). O repositório verbal fez parte do Pros@ (Processamento Semântico de Textos em Português Brasileiro)[7].

---

[4] Verbo-Brasil. Disponível em: http://143.107.183.175:21380/verbobrasil/index.php?lang=pt-br. Acesso em: 29 jun. 2021.

[5] Os termos "papel temático" e "papel semântico" foram apresentados, ao longo deste trabalho, mantendo o emprego dado na fundamentação teórica de cada base verbal.

[6] Projeto PropBank-Br. Disponível em: http://143.107.183.175:21380/portlex/index.php/pt/projetos/propbankbr. Acesso em: 29 jun. 2021.

[7] Projeto Pros@. Disponível em: http://www.nilc.icmc.usp.br/semanticnlp/index.php?id=principal&dir=includes&lang=pt-b. Acesso em: 29 jun. 2021.

A base verbal surge como apoio ao projeto PropBank-Br (DURAN; ALUÍSIO, 2012), com o objetivo de construir uma camada de anotação de papéis semânticos em um *corpus* do PB, aproveitando a metodologia existente em língua inglesa para a tarefa: o repositório de informação semântica *Proposition Bank* (PALMER *et al.*, 2005)[8]. O Propbank-Br v1 contém anotado uma porção brasileira do *corpus* Bosque (AFONSO *et al.*, 2002); já o PropBank-Br v2 contém 8.350 instâncias anotadas do *corpus* PLN-Br (BRUCKSCHEN *et al.*, 2008) e uma amostra de 840 instâncias do *corpus* Buscapé (HARTMANN *et al.*, 2014). Ambos os projetos mencionados (Propbank-Br v1 e Propbank-Br v2) possuem anotação com rótulos de papéis semânticos realizada sobre as árvores sintáticas geradas pelo *parser* Palavras (BICK, 2000), seguindo os parâmetros do PropBank (PALMER *et al.*, 2005), a fim de identificar a estrutura argumental dos verbos de língua inglesa equivalentes aos verbos anotados no PB. No entanto, devido às particularidades do PB, foram realizadas decisões complementares (DURAN; ALUÍSIO, 2011).

Sendo assim, o desenvolvimento da Verbo-Brasil fundamentou-se nos projetos mencionados: PropBank (PALMER *et al.*, 2005) e PropBank-Br (DURAN; ALUÍSIO, 2012), com destaque para o primeiro. Além disso, para determinar a relação semântica, o projeto também contou com as classes da VerbNet de língua inglesa (SCHULER, 2005). Segundo Palmer *et al.* (2005, p. 73-74), os verbos do PropBank são classificados de acordo com os estudos de Levin (1993), em que são organizados em função de seus aspectos semânticos e propriedades sintáticas compartilhados.

Ainda de acordo com Palmer *et al.* (2005), os argumentos semânticos de um verbo são numerados, começando com zero. Para um verbo específico, *Arg0* é geralmente o argumento que possui características de um Agente *Prototípico*, enquanto

[8] Base de dados da língua inglesa Propbank. Disponível em: http://verbs.colorado.edu/~mpalmer/projects/ace.html. Acesso em: 02 jun. 2021.

Arg1 é um Paciente ou *Tema Prototípico*. Um verbo polissêmico pode ter mais de um *frameset* ou *conjunto de quadros* (sintático-semânticos) quando possui significados diferentes, apresentando um *frameset* para cada conjunto de papéis semânticos. Para exemplificar, no Quadro 4, são apresentados os papéis semânticos do verbo *trabalhar* na base Verbo-Brasil.

Quadro 5 – Verbo *trabalhar* na Verbo-Brasil.

Verbo: Trabalhar **Exemplo:** ***A República Movimento de Emaús trabalha há 20 anos com adolescentes carentes.***
Frameset **trabalhar.01.** *Exercer atividade*. Mapeamento para o inglês: **work.01** **Arg0**: *trabalhador* (vnrole: -agent) **Arg1**: *trabalho (trabalhar com, trabalhar em)* (vnrole: -theme) **Arg2**: *empregador (trabalhar para)* (vnrole: -beneficiary) **Arg3**: *co-trabalhadores (trabalhar com alguém)* **Arg4:** ***instrumento (trabalhar com)*** **(vnrole: -instrument)**

Fonte: Verbo-Brasil. Disponível em: http://143.107.183.175:21380/verbobrasil/textoFrames/trabalhar-v.html. Acesso em: 29 jun. 2021.

Os ArgNs (Arg0, Arg1, Arg2, Arg3, Arg4) são previstos pela semântica dos verbos e cada sentido compreende os papéis numerados em palavras, segundo o Guia de Anotação do Propbank-Br (DURAN, 2014). O *frameset* do verbo *trabalhar* inclui os sentidos identificados a partir do Propbank e os papéis e as classes conforme a VerbNet (SCHULER, 2005). Segundo o Quadro 4, o conjunto de papéis semânticos do verbo *trabalhar*, ou *trabalhar.01*, prevê cinco ArgNs: o Arg0 é *trabalhador* ou Agente, o Arg1 é *trabalho - trabalhar com, trabalhar em* ou Theme[9], o Arg2 é *empregador - trabalhar para* ou Beneficiário, o Arg3 é *co-trabalhadores - trabalhar com alguém* e o Arg4 é *trabalhar com* ou Instrumento.

[9] O papel semântico Theme utilizado na base de dados Verbo-Brasil parece referir-se ao "assunto do trabalho", distanciando-se da definição comumente utilizada na literatura, em que *Theme* se refere a uma entidade que se desloca (ou sofre deslocamento), como definido em Palmer *et al.* (2010, p. 4).

O repositório da Verbo-Brasil apresenta exemplos anotados para ilustrar a atribuição dos papéis. Embora previstos, os ArgNs não precisam ocorrer todos ao mesmo tempo, como se verifica no exemplo do Quadro 4, retomado em (2):

(2) A República Movimento de Emaús trabalha há 20 anos com adolescentes carentes.

Em (2), tem-se: *Arg0* (Agente), que corresponde a "*A República Movimento de Emaús*" e o verbo em terceira pessoa do singular (*trabalha*), que expressa a relação entre constituintes sintáticos e os papéis semânticos na árvore sintática. O sintagma "*há 20 anos*" é etiquetado como *Argm-tmp* (modificadores de tempo) e o *Arg1* (*com adolescentes carentes*) possui a anotação de Paciente ou Tema. Cada *frameset* apresenta um conjunto de exemplos. Nesse caso, *trabalhar* possui 23 exemplos em um *frameset* na Verbo-Brasil. Os exemplos descritos no repositório foram extraídos de *corpora* utilizados no projeto Propbank-Br.

A interface da base de dados apresenta as entradas verbais em ordem alfabética, sendo a busca realizada de forma fácil e intuitiva. Em cada entrada verbal consta uma lista dos sentidos de cada verbo num arquivo (*framefile*) e, para cada sentido, um conjunto de papéis semânticos previstos (*roleset*) e um conjunto de exemplos para cada quadro sintático-semântico. As *tags* foram anotadas de acordo com o projeto Propbank-Br. Além disso, cada verbo possui um mapeamento para o Propbank de língua inglesa. Na interface também consta informações do projeto, dados da equipe e downloads dos *framefiles* do repositório.

2.3 VerboWeb

A VerboWeb é uma base de dados lexicais com a classificação sintático-semântica dos verbos do PB, criada pelas pesquisadoras Márcia Cançado, Luana Amaral e Letícia Meirelles, da Universidade Federal de Minas Gerais (UFMG). Trata-

se de uma base de dados em desenvolvimento, que se fundamenta nos dados do *Catálogo de verbos do português brasileiro*, publicado pelas criadoras em 2013 e nos trabalhos que são desenvolvidos pelo Núcleo de Pesquisa em Semântica Lexical da Faculdade de Letras da UFMG (NuPes)[10].

Essa iniciativa pretende realizar a análise de cerca de 3.000 verbos e, até o momento, já divulga em seu site a descrição de 1.482 construções. As entradas lexicais verbais, disponíveis na base, foram retiradas do Dicionário do Borba (1990) e do Dicionário eletrônico Houaiss (2009) e as frases criadas, atestadas em *corpora* da Linguateca e na *web*.

Propondo uma descrição e classificação dos verbos do PB, a VerboWeb se pauta nos princípios da Semântica Lexical, mais especificamente da chamada *Interface Sintaxe-Semântica Lexical*, considerando, para a proposta de tipologia, nos moldes expostos em Cançado, Godoy e Amaral (2013): (i) *os papéis temáticos*, que consistem nas relações estabelecidas entre um predicado e seus argumentos; (ii) *o aspecto lexical*, que remete a como um evento verbal se desenrola no decorrer do tempo; e (iii) *a decomposição de predicados*, que se refere à metalinguagem formal e sistemática, que lida com os sentidos dos verbos, decompondo o sentido dos itens lexicais em um sistema de predicados primitivos.

As frases (3a) e (4a) exemplificam os comportamentos do verbo *partir*, seguidas pela representação lógica em (b) e sua paráfrase em (c). Os dados foram retirados da VerboWeb:

(3) a. O peso do machado / O lenhador partiu a tora de madeira para fazer uma fogueira
b. [[X ACT (volition)] CAUSE [BECOME [Y <STATE>]]]
c. X (volicionalmente ou não) age causando Y ficar em determinado estado

[10] VerboWeb. Disponível em: http://www.letras.ufmg.br/verboweb/. Acesso em: 29 jun. 2021.

(4) a. As tropas partiram para o campo de batalha
b. [BECOME<event> [X LOC Y]]
c. X passa a ficar em Y através de um evento

Segundo Cançado *et al.* (2013, p. 107), "classificar verbos implica agrupá-los em classes que partilham certas propriedades não só semânticas, mas também sintáticas, ou, ainda, implica agrupá-los por propriedades semânticas que tenham impacto no seu comportamento gramatical". Sendo assim, o lexema *partir*, conforme exemplificado acima, é duplicado por apresentar diferenças semânticas e sintáticas acentuadas em sua constituição. O Quadro 5 apresenta um exemplo detalhado da descrição sintático-semântica do verbo *trabalhar* na VerboWeb.

Quadro 6 – Verbo *trabalhar* na VerboWeb.

Verbo: Trabalhar
Exemplo: O rapaz trabalhava demais.

Classe: Verbos de atividade: internamente causados (inergativos)
Propriedades da Classe:
- Conteúdo semântico recorrente na classe: x faz/produz um evento em si mesmo
- Estrutura sintática básica: [SN V] (verbo intransitivo)
- Estrutura de papéis temáticos: {Agente}
- Estrutura de decomposição de predicados: [X DO <EVENT>]
- Aspecto lexical básico: atividade

- Licencia um objeto cognato: O rapaz trabalhava, trabalhos voluntários.
- Licencia um adjunto equivalente ao objeto cognato: O rapaz trabalhava voluntariamente

Fonte: VerboWeb. Disponível em: http://www.letras.ufmg.br/sistemas/verboweb_cliente/ver_verbo.php?id=1579. Acesso em: 29 jun. 21.

Conforme se verifica no Quadro 5, a partir da busca pelo lexema verbal *trabalhar*, obtém-se informações sobre a classe e as propriedades do verbo: conteúdo semântico, papéis temáticos, aspecto lexical, estrutura sintática, estrutura de decomposição de predicados e seus licenciamentos sintáticos, que, no caso, se referem à atuação com um *objeto cognato* e um *adjunto que equivale a esse objeto*.

Na VerboWeb, propõe-se uma classificação que consiste em 4 grandes categorias, 17 classes e 7 subclasses, a saber[11]:

(i) *Verbos de atividade* (*de ação*): inergativos, como *correr* (subclasses: sujeito recíproco e expressão); inergativos com instrumento incorporado, como *esquiar*; contato com instrumento incorporado, como *martelar* (subclasse: instrumentais com dois instrumentos); contato mediado por instrumento, como *afiar* (subclasse: remoção); e contato mediado por corpo, como *abraçar*.

(ii) *Verbos de causação*: de mudança de estado opcionalmente agentivo, como *quebrar* (subclasses: objeto recíproco e contato); mudança de estado não agentivo, como *cansar*; mudança de estado locativo, como *dependurar* (subclasse: criação de imagem); mudança de estado de posse, como *encher*; mudança de lugar, como *engarrafar*; mudança de posse, como *acorrentar*; transferência do tipo locatum, como *abençoar*; e estado psicológico, como *preocupar*.

(iii) *Verbos de culminação* (*processo*): de mudança de estado, como *amadurecer*; e mudança de lugar, como *entrar*.

(iv) *Verbos de estado*: psicológicos, como *amar*; e existenciais, como *haver*.

Salienta-se que a página da VerboWeb é de fácil acesso e está atualizada. As buscas podem ser feitas por *verbo, classe, subclasse* ou a partir de alguma *propriedade não classificatória*. Além disso, é possível estabelecer relações de busca entre classe e subclasse ou cruzamento de propriedades, por exemplo. No site, encontram-se

[11] Dados baseados em Cançado, Amaral e Meirelles (2018) e na palestra *Um raio x dos verbos do PB: o projeto VerboWeb*, disponível em: https://youtu.be/VnzrqNfW2QI. Acesso em: 29 jun. 2021.

também informações sobre os pesquisadores envolvidos no projeto, referências bibliográficas e um tutorial sobre como fazer as buscas na base de dados. Pode-se afirmar, portanto, a existência de um recurso descritivo rico sobre os verbos da língua portuguesa, disponibilizado de maneira gratuita e acessível, o que a instaura como uma ferramenta interessante não só para linguistas e profissionais da computação, como também para usuários comuns.

3 As descrições sintático-semânticas verbais nas bases de dados do PB: reflexões e análise

Ao realizar a comparação entre as três bases de dados lexicais verbais do PB (VerbNet.Br, Verbo-Brasil e VerboWeb), é possível observar seus pontos comuns e divergentes, os quais serão descritos nesta seção.

Todas as bases analisadas consideram as classes semânticas propostas por Levin (1993) para a categorização dos verbos e apresentam, como objeto de descrição, o comportamento sintático e semântico de verbos *plenos* do PB. A VerbNet.Br e a VerboWeb se assemelham ainda na utilização de metalinguagens formais para a descrição dos fenômenos verbais.

Cada um dos recursos possui um número específico de entradas anotadas e categorizadas: VerbNet.Br, 1.766; Verbo-Brasil, 1.093; e VerboWeb, 1.486[12]. No entanto, apresentam a descrição de apenas 298 comportamentos verbais em comum, incluindo verbos muito frequentes (*ir*, *mandar*, *pedir*, *voltar*) e verbos pouco frequentes (*afiar*, *congestionar*, *fascinar*, *rachar*) da língua.

[12] É importante mencionar que tanto a VerbNet.Br, quanto a VerboWeb, duplicam as entradas verbais quando o lema apresenta diferentes comportamentos sintático-semânticos (*partir 1; partir 2*). Já na Verbo-Brasil, não há casos de entradas duplicadas, pois é apenas no interior de cada lema verbal que se tem acesso às suas diferentes construções sintático-semânticas (*framesets*).

Para além das diferenças concernentes aos arcabouços teórico-metodológicos de cada recurso, verifica-se também critérios distintos para a criação de diferentes entradas com a mesma forma verbal. Enquanto a VerbNet.Br é uma base mais granular e relaciona um lema verbal a diferentes classes, a depender de suas alternâncias sintáticas e características semânticas, a Verbo-Brasil e a VerboWeb não apresentam muitos casos de duplicação, salvo quando os verbos possuem comportamentos muito destoantes (como o exemplo de *partir* na seção 2.3).

Conforme se verifica nos exemplos dados com *trabalhar*, na VerbNet.Br, o verbo está relacionado a quatro classes distintas (*build-26.1, carve-21.2, disassemble-23.3, other_cos-45.4*), ao passo que, na Verbo-Brasil e na VerboWeb, esse mesmo lexema possui apenas uma entrada: na Verbo-Brasil, embora esteja descrito em apenas um *frameset*, o verbo *trabalhar* apresenta ao todo 23 exemplos e propõe a informação sintática e semântica de quatro argumentos distintos: Agente, na posição de sujeito, e Tema, Beneficiário e/ou Instrumento, na posição de objeto; já na VerboWeb, *trabalhar* está descrito com o comportamento de uma construção intransitiva da classe de *verbos inergativos*, em que se descreve apenas o papel semântico do argumento que ocupa a posição de sujeito (Agente). A abrangência de informações sintático-semânticas da Verbo-Brasil também se destaca quando se observa a preocupação em anotar os sentidos não só de verbos plenos (e verbos pronominais), como também os usos de verbos auxiliares, de construções multipalavras e, ainda, de expressões idiomáticas verbais do PB.

Com base na análise dos dados, é possível contrapor dois projetos de investigação distintos, em que um se dedica à descrição linguística de fenômenos verbais (VerboWeb), e o outro, elabora propostas descritivas para fins computacionais, inserindo-se na área de PLN (VerbNet.Br e Verbo-Brasil).

Sendo assim, a VerboWeb, projeto ainda em desenvolvimento, ilustra o primeiro tipo de empreendimento descritivo, o qual tem como objetivo a realização de

descrições linguísticas sobre diferentes fenômenos verbais da língua portuguesa. São, portanto, dados produzidos de maneira manual, atividade que demanda tempo e mão de obra especializada, com a aplicação de resultados de pesquisas realizadas há mais de 20 anos na área da Linguística. O fato de o banco de dados não ser uma tradução ou adaptação de trabalhos anteriores e/ou realizados para outras línguas naturais garante uma maior uniformidade dos dados e o estabelecimento de um recurso coerente com a sua base teórico-metodológica.

Por outro lado, têm-se a VerbNet.Br e a Verbo-Brasil como exemplos de recursos elaborados para fins computacionais, os quais unem as competências profissionais de linguistas e cientistas da computação, privilegiando-se, desse modo, a otimização dos sistemas e visando aplicações na área de PLN. A Verbo-Brasil, que é um apêndice verbal do projeto Propbank-Br, partiu de um mapeamento verbal existente para a língua inglesa, assim como ocorreu com a elaboração da VerbNet.Br, criada a partir de dados do inglês e de forma semiautomática. Apesar de essa metodologia otimizar o trabalho desenvolvido, apresenta determinadas lacunas para o PB, como a ausência de verbos frequentes da língua (*atualizar*, *adiar*, *emprestar*[13]) e, no caso da VerbNet.Br, a proposta de classes muito genéricas, nas quais é difícil identificar a relação semântica e sintática de sua composição. Sobre isso, Scarton (2013, p. 98-99) enfatiza a importância da VerbNet.Br como um repositório de dados verbais do PB, mas reconhece a necessidade de validação linguística, principalmente por se tratar de um recurso desenvolvido com base na tradução. A anotação realizada por um linguista permitiria a revisão dos dados e a inserção de alternâncias específicas para a língua portuguesa.

[13] Os exemplos mencionados se justificam pela frequência superior a 2.500 ocorrências no *corpus* Folha Kaggle, disponível em: https://www.kaggle.com/marlesson/news-of-the-site-folhauol. Acesso em: 29 jun. 2021.

A distinção entre esses "dois projetos de investigação" parece refletir também na disponibilização dos dados para consulta, já que a ferramenta de caráter mais linguístico (VerboWeb) possui uma interface mais acessível a especialistas e usuários comuns se comparada aos recursos elaborados com o propósito de servirem a aplicações de PLN (VerbNet.Br e Verbo-Brasil). Como já mencionado, a VerboWeb possui alternativas de buscas, para além da entrada verbal, e apresenta definições de informações sintático-semânticas anotadas nos itens lexicais, além de sugestões de referências acadêmicas sobre os temas. A Verbo-Brasil, por sua vez, embora não apresente atualizações nos últimos cinco anos, permite buscas simples com a entrada lexical verbal ou a listagem de verbos disponíveis, organizados em ordem alfabética, com a disposição das informações na página, que contribuem para o seu uso de maneira intuitiva. Já a VerbNet.Br parece ser o recurso com menor facilidade de acesso aos dados: apesar de ser possível realizar buscas por classes ou pelo lexema verbal e ainda consultar listas de verbos em ordem alfabética, as informações detalhadas sobre os predicados semânticos das classes estão em inglês. Portanto, para se ter acesso a essas informações, a página redireciona o usuário aos dados da VerbNet de língua inglesa.

No Quadro 6, são resumidas as características das bases comparadas neste trabalho.

Quadro 7 – Aspectos comuns e divergentes entre as bases lexicais verbais do PB.

Aspecto	**Convergência**	**Divergência**

Quantidade de lexemas	Todas as bases possuem mais de 1.000 entradas verbais descritas.	Cada base apresenta uma quantidade distinta de entradas lexicais, assim como diferentes critérios para duplicação e classificação dos lemas: VerbNet.Br (1.766 entradas), Verbo-Brasil (1.093 entradas) e VerboWeb (1.486 entradas).
Base teórico-metodológica	Todas as bases se inspiram nas classes da semântica lexical de Levin (1993); VerbNet.Br e VerboWeb apresentam uma metalinguagem formal de representação sintático-semântica.	Embora se inspirem nas classes de Levin (1993), cada base é guiada ainda por outros aspectos sintático-semânticos.
Objetivo	Todas as bases objetivam propor uma descrição sintático-semântica dos comportamentos verbais do PB; VerbNet.Br e Verbo-Brasil são recursos criados para fins computacionais.	Apesar de também poder ser utilizada para fins computacionais, a VerboWeb não se construiu especificamente para essa função.
Organização dos dados	VerbNet.Br e VerboWeb, embora partam de critérios distintos, organizam os verbos em classes sintático-semânticas: VerbNet.Br, com 202 classes semânticas; VerboWeb, com 4 grandes categorias sintático-semânticas.	Verbo-Brasil apresenta a descrição sintática e semântica dos verbos, em *framesets*, mas não apresenta uma proposta de tipologia para a sua organização.
Origem dos dados	Verbo-Brasil e VerboWeb são recursos feitos sobre o PB e validados por linguistas brasileiras.	VerbNet.Br e Verbo-Brasil se apoiam em descrições linguísticas do inglês; VerbNet.Br e Verbo-Brasil possuem dados decorrentes de processos automatizados; VerbNet.Br apresenta dados que carecem de revisão realizada por linguistas; VerboWeb é um recurso feito inteiramente por linguistas.
Interface	Todas as bases de dados possuem páginas gratuitas de acesso aos dados, com possibilidade de buscas. VerbNet.Br e Verbo-Brasil não apresentam atualizações recentes. No entanto, as bases disponibilizam a opção de *download* dos dados.	VerbNet.Br redireciona o usuário para os dados da VerbNet do inglês; VerboWeb apresenta dados atualizados e possibilita diferentes opções de busca, sendo uma ferramenta interessante para usuários comuns. No entanto, a base não disponibiliza o *download* dos dados.

Fonte: autoria própria.

4 Considerações finais

Neste trabalho, foram descritas e comparadas três bases de dados verbais do PB, disponíveis de maneira gratuita atualmente e comandadas por pesquisadoras brasileiras nos últimos 10 anos: VerbNet.Br, VerboWeb e Verbo-Brasil. Os dados revelam a existência de trabalhos com descrições robustas com informações relevantes sobre a sintaxe e a semântica dos verbos do PB.

A análise comparada entre essas bases enaltece o papel imprescindível do linguista na elaboração e revisão de dados lexicais para fins computacionais: a base realizada de maneira inteiramente manual por linguistas – VerboWeb – demonstrou ser a que apresenta maior coerência com a sua fundamentação teórico-metodológica, enquanto que a VerbNet.Br, feita de forma automática e com base em traduções do inglês, é a base que necessita de maior revisão dos dados, conforme sugere sua própria idealizadora, Scarton (2013). Por sua vez, a Verbo-Brasil, embora tenha sido criada para cumprir funções de anotação fundamentalmente computacionais, parece ser a base que contém mais informações sintático-semânticas das construções verbais em seu acervo, já que inclui, em seu repositório, além do comportamento de verbos plenos, verbos auxiliares, construções multipalavras e expressões idiomáticas.

Nesse sentido, ressalta-se a necessidade de adaptação, ampliação e/ou criação de repositórios verbais que considerem a polissemia de determinados verbos e, ainda, o seu comportamento quando não atuam como verbos *plenos*. Sobre isso, pode-se citar trabalhos lexicogramaticais sobre as construções com nomes predicativos e verbos-suporte de Barros (2014), Santos (2015) e Rassi (2015), que descrevem os comportamentos dos verbos *fazer* (*fazer exercício*), *ter* (*ter apoio*) e *dar* (*dar trabalho*), respectivamente, com foco em sua atuação como verbo-suporte. Acredita-se que estas descrições precisam constar em bases de dados disponíveis para consulta por usuários comuns e também para aplicações linguístico-computacionais na área de PLN.

É importante salientar ainda que há, obviamente, para a língua portuguesa, outras bases de dados verbais disponíveis de igual - ou maior - impacto que as

selecionadas. Esses casos não constam em nosso artigo por não preencherem os critérios estipulados nesta pesquisa. A título de exemplo, menciona-se o projeto ViPEr, uma base de dados de construções léxico-sintáticas dos verbos do português europeu, que culminou na publicação do *Dicionário gramatical dos verbos do português* (BAPTISTA, MAMEDE, 2020), o qual consta de 6.000 entradas verbais, com a apresentação de suas propriedades sintático-semânticas.

Outro trabalho de descrição dos verbos do PB importante de ser mencionado é o Projeto *Valências Verbais do Português* (VVP), sob a coordenação de Mário A. Perini. O objetivo do projeto é listar, em um dicionário, todos os conjuntos de construções (diáteses) que ocorrem no PB, assim como os verbos associados a cada um deles. A orientação do trabalho é fundamentalmente descritiva, procurando levantar e sistematizar dados, com base, tanto quanto possível, em pressupostos teóricos de ampla aceitação. Para isso, não há endosso de nenhuma corrente teórica, porque, segundo Perini (2008), as teorias devem basear-se no maior número de dados possível. Até o momento, o VVP contém cerca de 700 verbetes verbais descritos por linguistas associados ao projeto. Alguns resultados prévios podem ser conferidos em Perini (2015, 2016, 2019).

Como mencionado, o verbo é um elemento nuclear para as línguas naturais, daí a necessidade de elaboração de trabalhos acadêmicos e bases de dados lexicais dedicados especificamente ao comportamento sintático-semântico dessa classe de palavras. Por isso, como ações de pesquisas futuras, espera-se ampliar o reconhecimento e análise de bases de dados verbais disponíveis, principalmente de maneira gratuita, para o português, incluindo outras variantes da língua. Além disso, almeja-se verificar também como o verbo é tratado em diferentes recursos linguístico-computacionais, como a FrameNet.Br (SALOMÃO *et al.* 2013) e a anotação em Universal Dependencies (MCDONALD *et al*., 2013) do *corpus* Bosque (AFONSO *et al*., 2002), por exemplo.

Portanto, acredita-se que esta pesquisa contribui com a atualização do estado da arte sobre as bases de dados lexicais verbais do PB, dá visibilidade às obras analisadas e propõe interessantes desdobramentos de pesquisas e ações futuras para o incremento de recursos linguístico-computacionais do português, sobretudo da variante brasileira.

Agradecimentos

Os autores deste trabalho agradecem ao Centro de Inteligência Artificial (C4AI-USP) e o apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM Corporation.

Referências

AFONSO, S. BICK, E.; HABER, R.; SANTOS, D. Floresta sintá(c)tica: a treebank for Portuguese. *In*: RODRIGUES, M. G.; ARAUJO, C. P. S. (org.). **Proceedings of the Third International Conference on Language Resources and Evaluation** (LREC 2002). Paris: ELRA, 2002.

AZEREDO, J. C. **Gramática Houaiss da Língua Portuguesa**. São Paulo: Publifolha, 2008.

BAPTISTA, J.; MAMEDE, N. **Dicionário gramatical de verbos do português**. Faro: Editora UAlg, 2020.

BARROS, C. D. **Descrição e classificação dos predicados nominais com o verbo-suporte *fazer* em Português do Brasil**. Tese (Doutorado) - Universidade Federal de São Carlos (UFSCar), São Carlos, 2014.

BICK, E. **The parsing system palavras**: Automatic grammatical analysis of Portuguese in a constraint grammar famework. Aarhus Universitetsforlag, 2000.

BORBA, F. S. (coord.). **Dicionário gramatical de verbos do português contemporâneo do Brasil**. São Paulo: Editora UNESP, 1990.

CANÇADO, M.; GODOY, L. Representação lexical de classes verbais do PB. **Alfa**, São Paulo, 56 (1), 2012. DOI https://doi.org/10.1590/S1981-57942012000100006

CANÇADO, M.; GODOY, L.; AMARAL, L. **Catálogo de verbos do português brasileiro**: classificação verbal segundo a decomposição de predicados. v. 1. Verbo de mudança. Belo Horizonte: Editora UFMG, 2013.

CANÇADO, M.; AMARAL, L. **Introdução à Semântica Lexical**. Petrópolis: Vozes, 2016.

CANÇADO, M.; AMARAL, L.; MEIRELLES, L. VerboWeb: uma proposta de classificação verbal. **Revista da Anpoll**, v. 1, 2018. DOI https://doi.org/10.18309/anp.v1i46.1077

CANÇADO, M.; AMARAL, L.; MEIRELLES, L. **Banco de Dados Lexicais VerboWeb**: classificação sintático-semântica dos verbos do português brasileiro, UFMG. Disponível em: http://www.letras.ufmg.br/verboweb/. Acesso em: 29 jun. 2021.

CUNHA, C.; CINTRA, L. **Nova gramática do português contemporâneo**. 5 ed. Rio de Janeiro: Lexikon, 2008 [1984].

DURAN, M. S.; ALUÍSIO, S. M. Propbank-Br: A Brazilian Portuguese corpus annotated with semantic role labels. *In*: **Proceedings of the 8th Symposium in Information and Human Language Technology**. Cuiabá, Brazil, 2011.

DURAN, M. S., ALUÍSIO, S. M. Propbank-Br: A Brazilian Treebank annotated with semantic role labels. *In*: **Proceedings of the 8th International Conference on Language Resources and Evaluation**. Istanbul, Turkey, 2012.

DURAN, M. S.; MARTINS, J. P.; ALUÍSIO, S. M. Um repositório de verbos para a anotação de papéis semânticos disponível na web. *In*: **Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology**. Fortaleza, Brazil, 2013. Disponível em: https://www.aclweb.org/anthology/W13-4820.pdf. Acesso em: 29 jun. 2021.

DURAN, M. S. **Guia de Anotação**: Propbank-Br. 2014. Disponível em: https://docplayer.com.br/81173801-Guia-de-anotacao-propbank-br.html. Acesso em: 29 jun. 2021.

DURAN, M. S.; ALUÍSIO, S. M. Automatic Generation of a Lexical Resource to support Semantic Role Labeling in Portuguese. *In*: **Proceedings of SEM 2015**: The Fourth Joint

Conference on Lexical and Computational Semantics. Colorado, US, 2015. Disponível em: https://www.aclweb.org/anthology/S15-1026.pdf. Acesso em: 29 jun. 2021.

FERNANDES, F. **Dicionário de verbos e regimes**. 45 ed. Porto Alegre: Globo, 2005 [1940]. DOI https://doi.org/10.22456/2177-0018.7658

GREGHI, J. G. **Projeto e desenvolvimento de uma base de dados lexicais do português**. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional). – Universidade de São Paulo (USP), São Carlos, 2002.

LEVIN, B. **English Verb Classes and Alternations**: A Preliminary Investigation. Chicago: University of Chicago Press, 1993.

MCDONALD, R.; NIVRE, J.; QUIRMBACH-BRUNDAGE, Y.; GOLDBERG; Y.; DAS, D.; GANCHEV, K.; HALL, K.; PETROV, S.; ZHANG, H.; TÄCKSTRÖM, O.; BEDINI, C.; CASTELLÓ, N. B.; LEE, J. Universal dependency annotation for multilingual parsing. *In*: **Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics**. Sofia, Bulgaria, 2013. Disponível em: https://aclanthology.org/P13-2017.pdf. Acesso em: 18 set. 2021

PALMER, M.; GILDEA, D.; KINGSBURY, P. The Proposition Bank: An annotated corpus of semantic roles. **Computational Linguistics**, 31(1), 71–106, 2005. DOI https://doi.org/10.1162/0891201053630264

PALMER, M.; GILDEA, D.; XUE, N. Semantic role labeling. **Synthesis Lectures on Human Language Technologies**, v. 3, n. 1, p. 1-103, 2010. DOI https://doi.org/10.2200/S00239ED1V01Y200912HLT006

PERINI, M. A. **Estudos de gramática descritiva**. São Paulo: Parábola Editorial, 2008.

PERINI, M. A. **Describing verb valency**: practical and theoretical issues. Switzerland: Springer, 2015. DOI https://doi.org/10.1007/978-3-319-20985-2

PERINI, M. A. Construindo o dicionário de valências: problemas e resultados. **Scripta** (PUCMG), v. 20, 2016. DOI https://doi.org/10.5752/P.2358-3428.2016v20n38p148

PERINI, M. A. **Gramática descritiva do português brasileiro**. Petrópolis: Vozes, 2016.

PERINI, M. A. **Thematic Relations**. Switzerland: Springer, 2019. DOI https://doi.org/10.1007/978-3-030-28538-8

RASSI, A. P.; VALE, O. A. Tipologia das construções verbais em PB: uma proposta de classificação do verbo dar. **Caligrama**, Belo Horizonte, v. 18, n. 2, 2013. DOI https://doi.org/10.17851/2238-3824.18.2.105-130

RASSI, A. P. **Descrição, classificação e processamento automático das construções com o verbo *dar* em português brasileiro**. Tese (Doutorado) - Universidade Federal de São Carlos (UFSCar), São Carlos, 2015.

SALOMÃO, M. M. M.; TORRENT, T. T.; SAMPAIO, T. F. A Linguística Cognitiva encontra a Linguística Computacional: Notícias do Projeto FrameNet Brasil. **Cadernos de Estudos Linguísticos**, 55(1), 7-34, 2013. DOI https://doi.org/10.20396/cel.v55i1.8636592

SANTOS, M. C. A. **Descrição e classificação dos predicados nominais com o verbo-suporte *ter* em Português do Brasil**. Tese (Doutorado) - Universidade Federal de São Carlos (UFSCar), São Carlos, 2015.

SCARTON, C. E. **VerbNet.Br**: construção semiautomática de um léxico verbal online e independente de domínio para o português do Brasil. 2013. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Universidade de São Paulo (USP), São Carlos, 2013.

SCHULER, K. K. **Verbnet**: A broad coverage, comprehensive verb lexicon. Ph.D. Thesis (Philosophy) - University of Pennsylvania, 2005.

Artigo recebido em: 04.10.2021 Artigo aprovado em: 10.03.2022

DOI: 10.14393/DL52-v16n4a2022-10

Evaluating a typology of signals for automatic detection of complementarity

Avaliação de uma tipologia de sinais para a detecção automática da complementaridade

Jackson Wilke da Cruz SOUZA[*]
Ariani DI FELIPPO[**]

ABSTRACT: In a cluster of news texts on the same event, two sentences from different documents might express different multi-document phenomena (redundancy, complementarity, and contradiction). Cross-Document Structure Theory (CST) provides labels to explicitly represent these phenomena. The automatic identification of the multi-document phenomena and their correspondent CST relations is definitely handy for Automatic Multi-Document Summarization since it helps computers understand text meaning. In this paper, we evaluated a typology of (textual) signals for the automatic detection of the CST relations of complementarity (i.e., *Historical background*, *Follow-up* and *Elaboration*) in a multi-document corpus of news texts in Brazilian Portuguese. Using algorithms from different machine-learning paradigms, we obtained classifiers that achieved high general accuracy (higher than 90%), indicating the potential of the signals.

RESUMO: Em uma coleção de notícias sobre um mesmo evento, duas sentenças de textos distintos podem expressar diferentes fenômenos multidocumento (redundância, complementaridade e contradição). A *Cross-Document Structure Theory* (CST) provê rótulos para representar esses fenômenos. A identificação automática dos fenômenos multidocumento e das relações CST correspondentes é central à Sumarização Automática Multidocumento, pois ajuda a máquina a entender o conteúdo textual. Neste artigo, avaliou-se uma tipologia de sinais (textuais) para a detecção automática das relações CST de complementaridade (*Historical background*, *Follow-up* e *Elaboration*) em um *corpus* multidocumento de notícias em Português do Brasil. Utilizando algoritmos de diferentes paradigmas de Aprendizado de Máquina, obtiveram-se classificadores que atingiram alto índice de acurácia geral (superior a 90%), indicando o potencial dos sinais.

[*] PhD in Linguistics (UFSCar), professor in Instituto de Ciências Sociais Aplicadas at Universidade Federal de Alfenas (UNIFAL-MG). ORCID: https://orcid.org/0000-0003-1881-6780. jackcruzsouza@gmail.com

[**] PhD in Linguistics (UNESP), professor in Departamento de Letras at Universidade Federal de São Carlos (UFSCar). ORCID: https://orcid.org/0000-0002-4566-9352. arianidf@gmail.com

KEYWORDS: *Cross-Document Structure Theory.* Automatic summarization. Multi-document Corpus. Complementarity. Textual signal.

PALAVRAS-CHAVE: *Cross-Document Structure Theory.* Sumarização automática. Complementaridade. *Corpus* multidocumento. Sinal textual.

1 Introduction

Since the estimated size of the indexed Web is around 3.7 billion pages[1], its amount of textual information has already exceeded human limits of manageability. Given this scenario, subareas of Natural Language Processing (NLP) can produce computational solutions to deal with this large amount of data available to the user. Sub-areas dealing with content production and selection are those that have gained prominence in recent years, producing, for example, sentiment analysis, question and answer systems and automatic summarizers.

Specifically Automatic Multi-Document Summarization (MDS) NLP application that may assist users in acquiring relevant information in a short time. MDS aims at identifying the main information in a cluster of texts and presenting it as a summary (MANI, 2001). Much of the work to date has focused on extracts, i.e., summaries produced by concatenating sentences taken exactly as they appear in the source-documents (NENKOVA; MACKEOWN, 2011).

One important theory that guides extractive methods is Cross-Document Structure Theory (CST) (RADEV, 2000). It proposes relations to connect sentences from topically related texts. Such relations can be grouped into two categories (MAZIERO; JORGE; PARDO, 2010). Content relation indicates similarities and differences between sentences (*Identity*, *Equivalence*, *Summary*, *Overlap*, and *Subsumption*), complementarity (*Historical background*, *Follow-up*, and *Elaboration*), and contradiction (*Contradiction*). The form category conveys relations (*Indirect-speech*, *Modality*, *Attribution*, *Citation*, and *Translation*) that deal with shallow aspects of texts.

[1] Available from: https://www.worldwidewebsize.com. Access in 01/09/2021.

In MDS, there are very important challenges such as capturing the most important information of a topic within a generic perspective or prioritizing information preferences specified by the user (such as context information or the evolution of an event in time). CST annotation provides the means to deal with that, since the relations allow for detecting the multi-document phenomena (redundancy, complementarity and contradiction). Relations as Equivalence and Identity, for example, help to exclude repeated information, since a coherent extract should not have redundancy. Otherwise, if a user requires more context information about an event in the summary, the Historical background relation is helpful.

There have been many efforts to automatically detect the CST relations (e.g., ZHANG; BLAIR-GOLDENSHON; RADEV, 2002; ZHANG; OTTERBACHER; RADEV, 2003; MAYABE; TAKAMURA; OKUMURA, 2008; ZAHRI;, FUKUMOTO, 2011; KUMAR; SALIM; RAZA, 2012). One of them is CSTParser (MAZIERO; PARDO, 2012), an online multi-document parser based on CST for Brazilian Portuguese (BP). Using machine learning (ML) techniques, the system detects the relations with a general accuracy of 68,13%. Except for *Contradiction* and *Identity* (which are detected by rules), the parser uses similarity features to decide which CST relation (including those of complementarity) is held between sentences, since this type of relation only occurs between semantically related sentences.

Since complementary content might be important to build a multi-document extract and similarity is not sufficient for recognizing the different types (or CST relations) of complementarity, some efforts have been made in the last years to provide more extensive descriptions about the phenomenon and more efficient automatic methods for detecting it (SOUZA; DI-FELIPPO, 2018; SOUZA, 2015, 2019, 2021).

In Souza (2019), a corpus annotation of temporal markers and a wide variety of other (textual) signals of complementarity was carried out that resulted in a typology. It contributes to better understanding how complementarity is marked in the text, and

also provides attributes that can be used by automatic classifiers to recognize the different types (and CST relations) of complementarity. Recently, Souza (2021) refined the typology by exploring other aspects that seem to guide the readers to recognize complementary relations that hold between sentences.

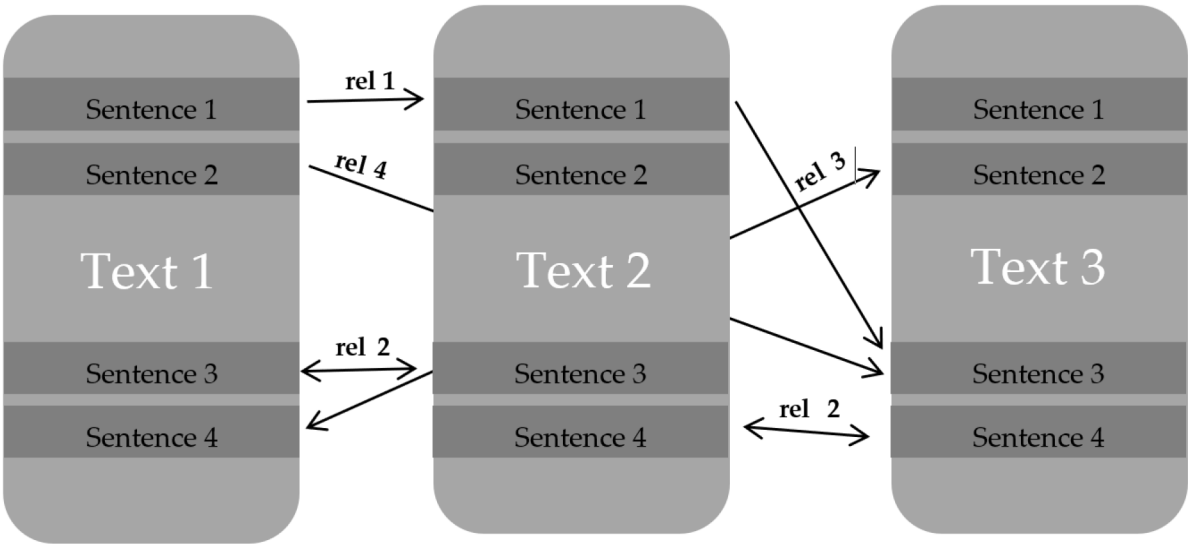
In this paper, we focused on evaluating the typology of Souza (2019) for the automatic detection of the CST relations of complementarity (i.e., *Historical background, Follow-up* and *Elaboration*) in a multi-document corpus of news in BP. The evaluation was performed using algorithms from different ML paradigms, and the results are promising.

Following the Introduction, Section 2 provides a brief introduction to CST, the main framework for the analysis, and an overview of the notion of complementarity. Section 3 presents the CSTNews corpus and the typology of signals. Section 4 investigates, through ML algorithms, the use of signals to distinguish the CST relations of complementarity. Finally, Section 5 summarizes the paper, and presents a few directions for future research.

2 Related works

Two sentences from topically related texts can be similar and different in several ways. One of the most relevant models to represent multi-document relations is CST, which was inspired by Rhetorical Structure Theory (RST) (MANN; THOMPSON, 1987). The difference between theories is that RST is aimed at capturing the rhetorical relation between adjacent text units while CST goes across topically related texts. Figure 1 illustrates a generic multi-document analysis at the sentence level.

Figure 1 – Generic scheme of multi-document analysis.



Source: Maziero (2012).

The original version of CST includes a set of 24 relations (RADEV, 2000) (Table 1).

Table 1 – Original set of CST relations.

Identity	Modality	Judgment
Equivalence	Attribution	Fulfillment
Translation	Summary	Description
Subsumption	Follow-up	Reader profile
Contradiction	Elaboration	Contrast
Historical background	Indirect speech	Parallel
Cross-reference	Refinement	Generalization
Citation	Agreement	Change of perspective

Source: based on Radev (2000).

In the last decade, interest in CST applications began to arise, especially in MDS, but also in other areas such as Query Reformulation, Learning Support, and Opinion Mining in the web (e.g., BELTRAME; CURY; MENEZES, 2012, INAM *et al.*, 2012; MURAKAMI *et al.*, 2010).

In order to obtain a better formalization and improving annotation concordance (by reducing ambiguity), Maziero, Jorge and Pardo (2010) proposed, based on the annotation of CSTNews, a typology for 14 CST relations according to their semantic

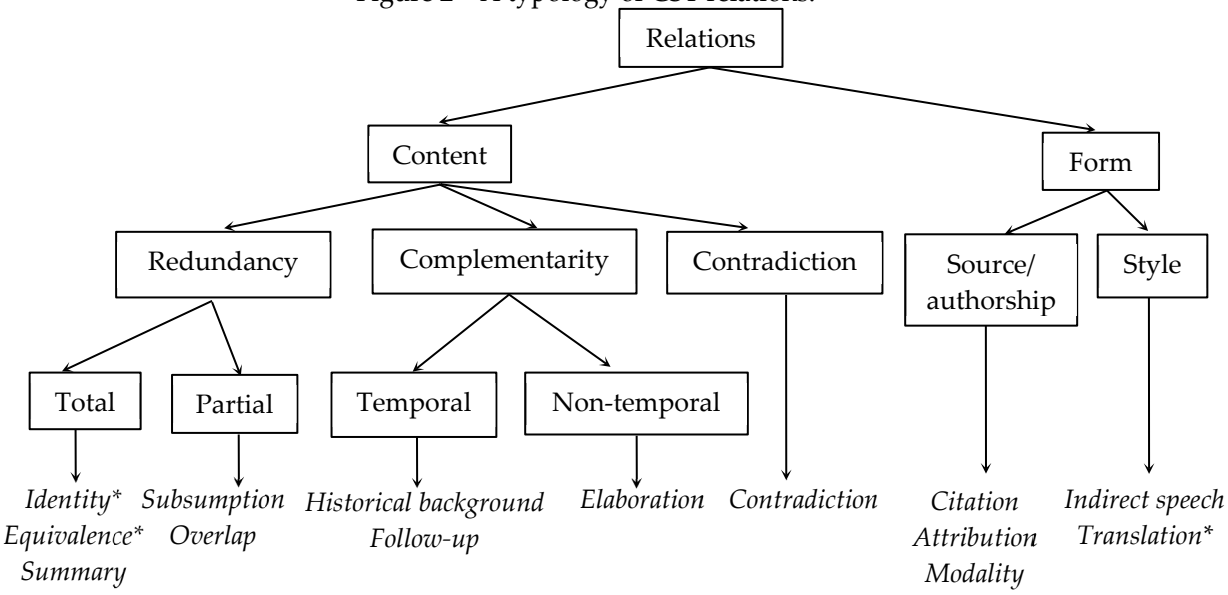
nature. This typology is illustrated in Figure 2, where the CST relations are at the lowest level of the hierarchy. One may observe in Figure 2 that there are two main categories of relations in the typology: content and form.

The content category refers to relations that indicate or capture (informational) similarities and differences among sentences. This category is divided into 3 subcategories: redundancy, complementarity, and contradiction. Redundancy includes relations that express total (i.e., *Identity*, *Equivalence* and *Summary*) or partial (i.e., *Overlap* and *Subsumption*) similarity among sentences. Complementarity relations link segments that elaborate, give continuity or background to some other information. *Historical background* and *Follow-up* are considered temporal, while *Elaboration* is non-temporal. The last subcategory only includes *Contradiction*.

The form category includes all the relations that deal with superficial aspects of information, for example, writing styles (*Indirect-speech*, *Modality*), citations (*Attribution*, *Citation*) or language (*Translation*). CST relations may also have directionality, being classified as symmetric or non-symmetric. In Figure 2, the asterisk indicates the symmetric relations, since one may read them in any direction.

To illustrate how complementarity relations occur among texts, Table 2 shows examples extracted from CSTNews. Each pair in Table 2 was selected from distinct news on the same event (i.e. from distinct news of the same cluster). The pair of sentences (S1 and S2) in (1) illustrates *Historical background*. In this case, the sentences were extracted from cluster C1, which comprises news reporting "a plane crash in Congo". S1 informs the place and the number of victims, while S2 provides a historical setting about S1 (i.e., air accidents are routine in the history of Congo).

Figure 2 – A typology of CST relations.



Source: Maziero, Jorge and Pardo (2010).

Table 2 – Example of types and CST relations of complementary from CSTNews[2].

Type	Relation	Pair of sentences
Temporal	Historical background	(1) S1: A plane crash in Bukavu, in the Eastern Democratic Republic of the Congo, killed 17 people on Thursday, said a United Nations spokesman on Friday. *(Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira, informou nesta sexta-feira um porta-voz das Nações Unidas)* S2: Air accidents are frequent in Congo, where 51 private companies operate elderly planes built in the former Soviet Union. *(Acidentes aéreos são frequentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética)*
	Follow-up	(2) S1: In his speech, Lula put emphasis on the end of agricultural protectionism. *(O discurso de Lula na ONU deu grande ênfase ao fim do protecionismo agrícola.)* S2: After Lula, it was the turn of the U.S. President George W. Bush to address the United Nations 62nd General Assembly. *(Depois de Lula, foi a vez do presidente americano George W. Bush discursar na Assembleia Geral da ONU)*

[2] In this paper, we first present an English translated version of the example, followed by the original one in Portuguese language.

Non-Temporal	Elaboration	(3) S1: The victims of the accident were 14 passengers and three crew members. *(As vítimas do acidente foram 14 passageiros e três membros da tripulação.)* S2: According to air traffic control, all crew members were Russian nationals. *(Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.)*

Example (2) illustrates the *Follow-up* relation, since S2 presents additional information which has happened since S1. The sentences were compiled from news on "the speech of the former Brazilian president, Luiz Inácio Lula da Silva, at the 62nd session of the United Nations General Assembly". In this case, *Follow-up* is signaled by the temporal expression "after Lula" in S2.

The sentences in example (3) were also compiled from the cluster C1, and they illustrate *Elaboration*. More specifically, S1 presents the number and profile of the victims, while S2 details an element present in S1 without redundancy. In this case, S2 provides the nationality of the crew members who died in the crash.

Given its usefulness, especially for MDS applications, some efforts have been made in the last years to provide extensive descriptions about the multi-document phenomena covered by the CST relations. The first work on MDS in PB, by Maziero, Jorge and Pardo (2014), addressed the automatic detection of CST relations, including those of complementarity. Following the literature, the authors only used features that track some form of repetition or redundancy (such as the difference in number of words, percentage of words in common, number of words in the longest common substring, difference in the number of nouns, etc.) to predict the relations between two sentences. The ML techniques explored for the prediction of the content relations achieved a general accuracy of 70,51%, which is considered a good result given the subjectivity of the CST analysis.

With the common goal of automatically identifying and characterizing CST relations in texts, a few computational-linguistic studies have explored a wide array of linguistic or textual features of complementarity (SOUZA; DI-FELIPPO, 2018; SOUZA,

2015, 2019, 2021). More specifically, Souza (2015) and Souza and Di-Felippo (2018) developed the first corpus description of complementarity using 90 sentence pairs compiled from the CSTNews corpus. In this work, a set of 7 potential attributes was explored for temporal complementarity detection. The attributes are: noun overlap, sentence distance[3], subtopic overlap, temporal expression in S1, temporal expression in S2, adverb in S1, and adverb in S2. With the exception of temporal expressions and adverbs, the majority of the attributes is based on lexical similarity because it is known that the CST relations only occur between semantically related sentences.

Using some ML algorithms from Weka (Waikato Environment for Knowledge Analysis) (WITTEN; FRANK 2005), the potential of the attributes to discriminate between *Historical background* and *Follow-up* were evaluated. The JRip classifier learned the smallest set of rules with the highest general accuracy (80%). Among the 5 rules of JRip, three of them are based on "temporal expression in S2" to classify the *Follow-up* pairs. For attribute selection[4], the InfoGainAttributeEval algorithm was applied, and it also indicated the relevance of this feature (i.e., temporal expression in S2) in the task.

Later, Souza (2019) expanded the previous work by investigating temporal and non-temporal complementarities and a wide variety of signals (morphological, syntactic, semantic, and pragmatic). As a result, the signals were organized into a typology, yielding a hierarchical structure of textual cues.

More recently, in a study with no computational motivations, Souza (2021) refined the typology by exploring other aspects (not expressed in the text) on how the

[3] The relative distance between complementary sentences according to the position of the sentences in their correspondent source text. For example, given a sentence pair from a cluster x (S1 and S2), where S1 is Sentence 6 from Text 1 and S2 is Sentence 4 from Text 2, the distance value between them is equal to 2 (positions). The authors normalized this value by dividing it by the longest distance between two sentences identified in the subcorpus of complementarity relations.

[4] Attribute selection aims at improving the performance of the ML algorithms by removing irrelevant attributes, which reduces the processing time and generates simpler models.

complementarity between sentences is recognized by readers, such as the reading of their source texts.

In the next section, we briefly present the CSTNews corpus. We also present the mentioned typology (or taxonomy) of signals in detail, as it is the focus of the evaluation.

3 The CSTNews corpus and the typology of signals of complementarity

The study of Souza (2019) was conducted over CSTNews, a multi-document corpus contained 50 clusters of news in BP, totaling 140 texts, 2,088 sentences and 47,240 words (CARDOSO *et al.* 2011). The clusters are organized into 6 categories: world, politics, daily news, science, money, and sports. The source texts were compiled from the online news agencies *Folha de São Paulo*, *Estadão*, *O Globo*, *Jornal do Brasil*, and *Gazeta do Povo*.

Each cluster contains: (i) 2 or 3 texts; (ii) mono-document abstracts (produced by a human summarizer); (iii) 12 manual multi-document summaries (6 abstracts and 6 extracts); (iv) 1 automatic multi-document extract, and several types of annotation, corresponding to different levels of linguistic analysis.

One of the annotations is the manual identification of the CST relations across the input documents within the clusters. This annotation was performed by 4 computational linguists using the CSTTool editor (ALEIXO; PARDO, 2008) and the typology shown in Figure 2. From the total of 1,641 sentence pairs annotated in CSTNews, there are 713 pairs of complementarity, distributed in 370 of temporal (i.e., 77 of *Historical Background* and 293 of *Follow-up*), and 343 of non-temporal relations (i.e., *Elaboration*). Thus, complementarity corresponds to 43.44% of the total relations in the corpus.

For the study of (temporal and non-temporal) complementarity, Souza (2019) only used 655 pairs (i.e., 76 of *Historical Background*, 260 of *Follow-up* and 319 of

Elaboration), since he excluded 58 pairs for disagreeing with the original annotation. The main goal of this work was to explore how many, and what types of cues could be found if the signaling information was studied beyond temporal expressions and lexical similarity. The most important aspect of the description was to select and classify the types of cues.

For the description of these relations, a sequence of three main tasks was performed: (i) analysis of each sentence pair in the subcorpus of complementarity from CSTNews, (ii) delimitation (with brackets) of the signals involved in complementarity to indicate the specific CST relation, and, finally, (iii) documentation on how the relation is signaled.

We illustrate these tasks with a *Historical background* pair from the CSTNews corpus (4). This type of complementarity is codified by a CST relation with directionality (from S2 to S1), which means that S2 provides the contextual information about S1. According to the author´s analysis, the previous annotation of the sentence pair with *Historical background* is strongly related to the textual segment "foi o pior do país desde 1995" ("it was the country´s worst (earthquake)") (in parentheses) that occurs in S2. This segment contains 3 different textual signals (in bold), which are delimitated by square brackets and indexed according to the position of occurrence in the sentence: (i) superlative expression, (ii) preposition "desde", and (iii) named entity ("1995").

In the coding or delimitation task (ii), the author added signaling information to the existing relations from the CSTNews corpus. Then, the signals identified were extracted, and documented along with relevant information about their function. For example, the superlative expression ("foi o pior (do país)") directly refers back to the earthquake magnitude mentioned in S1, and provides a relative description about it, i.e., very high size or amplitude in comparison to the last strongest one. The remaining two signals, "since" and "1995", are used (together) to specify the particular time in

the past when the last strongest earthquake has happened. The author documented or described the signaling information in a separate Excel file.

A detailed description of the annotation for the sentence pair in (4) is provided in Table 3.

(4) S1: In the case of Japan, the mentioned magnitude of 6.8 is considered "strong".
("*No caso do Japão, a magnitude apontada de 6,8 é considerada "forte".*")

S2: ([**It was the country´s worst** (earthquake)]$_1$ **[since]**$_2$ [**1995**]$_3$), when a 7.3 magnitude earthquake killed more than 6,400 people in the city of Kobe.
(*"Foi o pior do país desde 1995, quando um tremor de magnitude 7,3 matou mais de 6.400 pessoas na cidade de Kobe.").*

Table 3 – Example of signals description.

Cluster/ Pair	Relation name	Signal type	Specific signal	Explanation – how the relation is signalled
32/52	Historical background	Main clause	Superlative	The expression "the worst" in main clause of S2 is used to compare the magnitude mentioned in S1 to the last strongest earthquake
		Word class	Preposition	"Since" is used to situate the event in time, specifically in relation to the year of the last strongest earthquake
		Time	Named entity	The NE "1995" functions as a signal because it refers back to a specific point in time.

After describing the 655 pairs, all signals were hierarchically organized in 3 levels (i.e., signal class, signal type and specific signal), following Taboada and Das (2013) and Das and Taboada (2018). The signal class is the top-level classification, and it has 5 tags representing the major classes of signals (i.e., referential, morphological, syntactic, semantic, and pragmatic), as described in Table 4. For each class, a second level is defined; for example, the class referential has only one type (anaphor), while

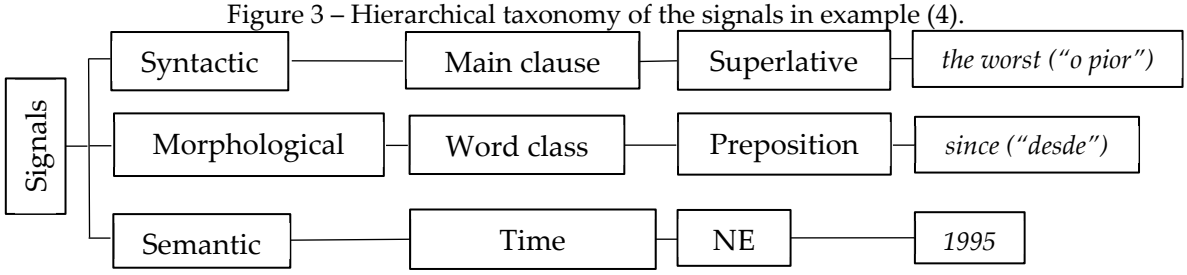
the semantic class is divided into 4 types (i.e., semantic field, semantic relation, addition, and time). Finally, the third level in the hierarchy refers to specific signals. The anaphoric type, for example, has two specific signals: associative and nominal.

Table 4 – Description of the different classes of complementarity signals[5].

Class	Description
Referential	Features include links where entities, similar or dissimilar, help interpret the relations.
Morphological	Among morphological features, tense is a very prominent feature, indicating temporal relations between the sentences.
Syntactic	At the syntactic level, there are a host of constructions that help identify relation. For example, subornative clauses tend to express details of some information given more generally in the other sentence of a pair.
Semantic	A semantic feature has two components, each belonging to one of the sentences. The components are in a semantic relationship with each other, such as *meronymy*, and *semantic field*.
Pragmatic	At the pragmatic level, there are several signals to guide the interpretation of relations. In the case of the news genre (which all the sentences/texts in the corpus belong to), the complementary information might be a list of items that details an event, a similar fact, a posterior or future event, etc.

The taxonomy of the signals in example (4) is provided in Figure 3.

Figure 3 – Hierarchical taxonomy of the signals in example (4).



We show the hierarchical organization of the signaling taxonomy in Table 4 as well as descriptive statistics of the frequency of signals in our corpus of

[5] For a more detailed description about the classes, types, and specific signals of the typology, along with example from the corpus, see Souza (2019, 2021).

complementarity. According to Table 5, the corpus includes 2,022 signals, distributed in 30 different specific signals. Referential and morphological classes are the less frequent, with a very similar distribution (i.e., 15.9% and 15.6%, respectively); pragmatic is the most frequent (25.4%), closely followed by syntactic (22.2%) and semantic classes (21%).

In Table 6, there is a summary of the relationship between CST relations and signals. The summary provides descriptive statistics of the frequency of the relations and how often each of them is signaled by specific cues.

We would like to point out that what Souza (2019) has found are positive signals, and this does not mean that the signals are used exclusively to indicate the relation. In other words, this means that relation identification, by humans and machines, relies or can rely on signals as indicators that a relation is present, but there are many-to-many correspondences between relations and signals, as we can see in Table 6. Although, the occurrence of certain signals seems to be typical of a relation. This is the case for the *superlative expression* in sentence 2 of a pair, which is a syntactic signal that only indicates *Historical background* (see example (4)). Also, the concept of a signaling information as an indicator of a relation also means that the signals, as textual devices, are not exclusively used to mark a CST relation; they may well have other purposes in the document or text. For instance, an associative anaphora, as a type of the referential class, also contributes to cohesion, in addition to signaling a relation.

In other words, we can say that textual signals are compatible with a CST relation, not necessarily indicators of the relation exclusively. However, the results seem to provide evidence that relation signaling is widespread and has potential for computational applications.

Table 5 – Typology of complementarity signals and its statistics.

Class	Signal type	Specific signal	Total		%	
Referential (322 – 15,9%)	Anaphor	Associative	81	322	4,0	15.9
		Nominal	241		11.9	
Morphological (315 – 15,6%)	Word class	Numeral	48	315	2.4	15.6
		Noun	17		0.8	
		Preposition	24		1.2	
		Verbal tense	149		7.4	
		Elocution verb	77		3.8	
Syntactic (450 – 22,2%)	Simple period (or main clause)	Adverbial phrase	73	99	3.6	4.9
		Superlative expression	26		1.3	
	Compound period	Reported speech	119	240	5.9	11.9
		Additional clause	28		1.4	
		Explanation clause	49		2.4	
		Direct object clause	29		1.4	
		Reduced relative clause	15		0.7	
	Displacement	Theme-Rheme	111	111	5.5	5.4
Semantic (422 – 20,9%)	Semantic field	Related words	63	63	3.1	3.1
	Semantic relation	Cause-effect	35	112	1.7	5,5
		Hyponymy (*is-a*)	20		1.0	
		Meronymy (*part-whole*)	57		2.8	
	Addition	Indicative word/phrase	170	170	8.4	8.4
	Time	Named entity (NE)	77	77	3.8	3.8
Pragmatic (495 – 25,4%)	Genre (about the event)	Detailing (*list of items*)	162	368	8.0	18,2
		Posteriority (*later event*)	92		4.5	
		Futurity (*future event*)	57		2.8	
		Continuity (*continuous event*)	18		0.9	
		Similarity	39		1.9	
	Argumentation	Focus	17	17	0.8	0.8
	Addition	Related information	52	52	2.6	2.6
	Aspectuality	Punctual event	38	76	1.9	3.8
		Durative event	38		1.9	

Source: Souza (2019).

Table 6 – Distribution of CST relations by signaling devices. Source: Souza (2019).

Tipology			**CST relations of complementarity**			**Total**
Class	**Signal type**	**Specific signal**	***Elab.***	***Follow-up***	***Historical background***	
Referential	Anaphor	Associative	50	31	0	81
		Nominal	132	89	20	241
Morph.	Word class	Numeral	11	35	2	48
		Noun	2	15	0	17
		Preposition	7	0	17	24
		Verbal tense	12	134	3	149
		Elocution verb	26	51	0	77
Syntactic	Simple period (or main clause)	Adverbial phrase	31	40	2	73
		Superlative expression	0	0	26	26
	Compound period	Reported speech	67	52	0	119
		Additional clause	26	2	0	28
		Explanation clause	37	5	7	49
		Direct object clause	22	7	0	29
		Reduced relative clause	12	3	0	15
	Displacement	Theme-Rheme	108	1	2	111
Semantic	Semantic field	Related words	29	34	0	63
	Semantic relation	Cause-effect	12	23	0	35
		Hyponymy (*is-a*)	16	4	0	20
		Meronymy (*part-whole*)	42	15	0	57
	Addition	Indicative word/phrase	4	109	57	170
	Time	Named entity (NE)	27	42	8	77
Pragmatic	Genre (about the event)	Detailing (*list of items*)	103	59	0	162
		Posteriority (*later event*)	0	92	0	92
		Futurity (*future event*)	0	57	0	57
		Continuity (*continuous event*)	0	18	0	18
		Similarity	0	0	39	39
	Argumentation	Focus	17	0	0	17
	Addition	Related information	52	0	0	52
	Aspectuality	Punctual event	0	0	38	38
		Durative event	17	0	0	17

Since Souza (2019) has proposed the typology or taxonomy of complementarity signals, we were interested in automatic detecting the CST relations of complementarity based on the signals. Thus, we investigated the discriminative power of the annotated signals.

4 Automatic validation of the signals

In order to determine whether certain signals (or combinations of them) predicate a CST relation of complementarity, we used ML algorithms available in Weka, which is a state-of-art facility for developing ML techniques and their application to real-world data mining tasks (WITTEN; FRANK, 2005). We conducted such automatic study because ML techniques consider every combination of signals to predict the classes (i.e., the CST relations of complementarity).

In this work, we used supervised ML algorithms, which basically map a function from known input-output pairs to estimate relationships between them. Most fundamentally, supervised learning utilizes a data set which includes both input features as well as the output class (or target) which are labeled at the start of training. In our case, each instance of the training data consisting of a sentence pair and its signals (i.e., the features) and the desired CST relation of complementarity (i.e., the output classes). Thus, the algorithms train on the input data set to produce a model which will differentiate among the output labels based on the most relevant attributes. In other words, the algorithms analyze the training data and produce a classifier that should be able to predict the correct classes (MITCHELL, 1997).

For using Weka, we converted the descriptive information illustrated in Table 3 into an ARFF (Attribute-Relation File Format) file, which is the most common format for data used in Weka (Figure 4). According to Figure 4, an ARFF file has two parts. The first one is a Header describing what each data instance should be like, and the second part is the Data (entry).

Figure 4 – Illustration of the ARFF file format for ML.

@ATTRIBUTE PAIR NUMERIC
@ATTRIBUTE RELATION {ELABORATION, FOLLOW_UP, HISTORICAL_BACKGROUND}
@ATTRIBUTE ASSOCIATIVE_ANAPHOR {YES, NO}
@ATTRIBUTE NOMINAL_ANAPHOR {YES, NO}
@ATTRIBUTE NUMERAL {YES, NO}
@ATTRIBUTE NOUN {YES, NO}
@ATTRIBUTE PREPOSITION {YES, NO}
@ATTRIBUTE VERBAL_TENSE {YES, NO}
@ATTRIBUTE ELOCUTION_VERB {YES, NO}
@ATTRIBUTE ADVERBAL_PHRASE {YES, NO}
@ATTRIBUTE SUPERLATIVE {YES, NO}
@ATTRIBUTE REPORTED_SPEECH {YES, NO}
@ATTRIBUTE ADDITIONAL_CLAUSE {YES, NO}
@ATTRIBUTE EXPLANATION_CLAUSE {YES, NO}
@ATTRIBUTE DIRECT_OBJECT_CLAUSE {YES, NO}
@ATTRIBUTE REDUCED_RELATIVE_CLAUSE {YES, NO}
@ATTRIBUTE THEME-RHEME {YES, NO}
@ATTRIBUTE RELATED_WORDS {YES, NO}
@ATTRIBUTE CAUSE-EFFECT {YES, NO}
@ATTRIBUTE HYPONYMY {YES, NO}
@ATTRIBUTE MERONYMY {YES, NO}
@ATTRIBUTE INDICATIVE_WORD {YES, NO}
@ATTRIBUTE NAMED_ENTITY {YES, NO}
@ATTRIBUTE DETAILING {YES, NO}
@ATTRIBUTE POSTERIORITY {YES, NO}
@ATTRIBUTE FUTURITY {YES, NO}
@ATTRIBUTE CONTINUITY {YES, NO}
@ATTRIBUTE SIMILARITY {YES, NO}
@ATTRIBUTE FOCUS {YES, NO}
@ATTRIBUTE RELATED_INFORMATION {YES, NO}
@ATTRIBUTE PUNCTUAL_EVENT {YES, NO}
@ATTRIBUTE DURATIVE_EVENT {YES, NO}

@DATA

01, HISTORICAL_BACKGROUND, NO, NO, NO, NO, **YES**, NO, NO, NO, **YES**, NO, NO, NO, NO, NO, NO, NO, NO, NO, NO, NO, **YES**, NO, NO, NO, NO, NO, NO, NO, NO, NO

More precisely, the Header describes the list of attributes. The format of @attribute is "@attribute [attribute-name] [values]". In our case, we have 31 features or attributes. The first one (@attribute PAIR) codifies the numeric id of each sentence pair of the corpus. The second attribute (@attribute RELATION) describes the CST relations of complementarity, which means that it has 3 possible nominal values: elaboration, historical_background, and follow-up. Then, there is an attribute for each of the 29 specific signals of the typology. Each of them was codified as a binary feature, admitting 2 possible values: "yes" for the presence and "no" for the absence of a signal.

The order the attributes are declared indicates the column position in the Data section, which describes the corpus examples (or instances) for training. If an attribute is the third one declared, then Weka expects that all those attribute values will be found in the third comma delimited column. As an illustration, consider the instance in the Data section of Figure 4, which corresponds to example (4). According to the order of the attributes in the Header, the 3 signals of complementarity that occur in this example (i.e., morphological=preposition, syntactic=superlative, and semantic=named entity) (see Figure 3), for instance, are respectively found in the 7th, 11th and 23rd comma delimited columns, which is indicated by the value "yes". The null occurrence of any referential or even pragmatic signal in example (4) is indicated by the value "no".

To perform the ML over the complementarity subcorpus from CSTNews, we applied the *10-fold cross-validation* technique. In the basic *10-fold cross-validation*, the corpus is randomly partitioned into *k* equal sized subsamples. Of the *k* subsamples, a single one is retained for test, and the remaining (*k – 1*) subsamples are used as training data. The process is repeated *k* times, with each of the k subsamples used once as the test data. The results are averaged over all the runs. We selected the *10-fold cross-validation* method because it gets more realistic estimates of the error rates for

classification, since our dataset is relatively small (i.e., 655 pairs of sentences) and unbalanced (i.e., 76 of *Historical background*, 260 of *Follow-up*, and 319 of *Elaboration*).

Additionally, we also performed the attribute selection process using Weka. Even with a relatively small set of attributes from the typology (i.e., 29 features, precisely), this selection is an ML technique that reveals the importance of the features, reducing processing time as well as increasing the performance of mining task.

Although there are different ML paradigms available in Weka, i.e., connectionist, mathematical (or probabilistic) and symbolic, we focused on symbolic algorithms, because they produce rules that can be easily interpreted and verified by human experts. Nonetheless, we have also tested other machine algorithms from other Artificial Intelligence paradigms, for comparison purposes only (Table 7).

To evaluate the results, we have used the following metrics: accuracy, precision (P), recall (R), and *f*-measure (f-m). Accuracy indicates an overall performance of the model or classifier; such metric determines how far the output can be from the optimal one). Precision is the percentage value indicating how many of the instances returned by the algorithm are correctly classified. Recall quantifies the number of positive class predictions made out of all positive examples in the dataset. F-Measure provides a single score that balances both the concerns of precision and recall in one number, being a unique indicator of the algorithm performance (SHALEV-SHWARTZ, BEN-DAVID, 2014; JURAFSKY, MARTIN, 2021).

Table 7 – Automatic validation of the signal typology.

Paradigms / **CST relations**	**Symbolic**						**Mathematical**			**Connectionist**		
	JRip (96.6%)			**J48 (95.8%)**			**NaïveBayes (96.7%)**			**MLP (96.1%)**		
	P	**R**	***f-m***	**P**	**R**	***f-m***	**P**	**R**	***f-m***	**P**	**R**	***f-m***
HB	1.00	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.99	1.00	1.00	1.00
Follow-up	0.96	0.92	0.94	0.98	0.93	0.95	0.98	0.94	0.96	0.95	0.94	0.95
Elaboration	0.94	0.97	0.95	0.94	0.96	0.97	0.95	0.98	0.96	0.95	0.96	0.96

According to Table 7, the algorithms from different paradigms present very similar general accuracy. The well-known connectionist Multi-Layer Perception (MLP) algorithm achieved 96.1% of general accuracy with the default Weka configurations. Among the several mathematical or probabilistic methods in Weka, we ran Naïve-Bayes, which presents the highest general accuracy among all the algorithms, achieving 96.7%. More specifically, we tried JRip and J48 algorithms from the symbolic paradigm. JRip and J48 generated sets of rules with 96.6% and 95.8% of accuracy, respectively. The decision tree produced by J48, however, has more rules than the JRip classifier (22 and 9 rules, respectively).

More than a good classification accuracy, we wanted to be able to make the characterization of the different CST relations (of complementarity) explicit. Thus, we explored the results achieved by the symbolic algorithms in more detail. We chose to use the JRip classifier based on the combination of two factors: (i) manageable rule set and (ii) highest accuracy among the symbolic algorithms. We believe that combination is a good scenario for our purposes in this work. In other words, we chose JRip because its classifier learned a small set of rules with the best general accuracy

We present in Table 8 the 9 rules of the JRip algorithm, which are followed by the number of instances (sentence pairs) correctly (C) classified and incorrectly (I) classified, and the precision of the rule. For example, the 1^{rst} rule of Table 7 predicated a total of 38 pairs, and all of them were correctly classified as *Historical background,* thus it has 100% precision. In the rules, one can say that aspectuality is a signal type (from the pragmatic class) that characterizes well the *Historical Background* relation, since the 1^{rst} and 2nd rules are based on durative and punctual events, respectively. This means that the historical context is commonly an event/fact that occurs frequently or a specific

event back in time. By the way, the attribute selection using the InfoGainAttributeEval[6] algorithm (at Weka) also indicated the relevance of the aspectuality feature.

Table 8 – JRip logic rules.

Rule	C	I	P (%)
1. If DurativeEvent=*yes* then *Historical background*	38	0	100
2. Else-if PunctualEvent=*yes* then *Historical background*	38	0	100
3. Else-if Tense=*yes* and Theme-Rheme=*no* and RelatedInfo=*no* then *Follow-up*	136	2	98.5
4. Else-if NamedEntity(Time)=*yes* then *Follow-up*	58	3	95
5. Else-if Posteriority=*yes* then *Follow-up*	39	0	100
6. Else-if Continuity=*yes* then *Follow-up*	11	0	100
7. Else-if RelatedWords=*yes* and Numeral=*yes* then *Follow-up*	2	0	100
8. Else-if Detailing=*yes* and ElocVerb=*yes* and NomAnaphor=*yes* then *Follow-up*	8	2	80
9. Else-if *Elaboration*	325	13	96.6

The *Follow-up* relation is characterized by signals from all classes (referential, morphological, syntactic, semantic and pragmatic), since all rules, from the 3rd to the 8th, are based on them. Rules four, five and six are based on individual signals, while rules three, seven and eight are based on combinations of specific signals. The 3rd rule, for example, combines 3 features from different signal types/classes to classify the higher number of *Follow-up* instances: (i) verbal tense (word class→morphological), (ii) theme-rheme (displacement→syntactic class) and (iii) related information (addition→pragmatic class). In rule five, we see that posteriority is the signal that characterizes individually the higher number of *Follow-up* instances (39 sentences pairs) with 100% precision.

[6] The algorithm evaluates the worth of a feature by measuring the information gain with respect to the class.

If none of the eight first rules are applied, the default class is *Elaboration*, which is given by the 9th rule. This could indicate that *Elaboration* is not characterized by particular signals present on this set of attributes, being a very generic CST relation.

5 Final remarks and future works

We have presented a validation task of the signaling taxonomy of complementarity proposed by Souza (2019). The purpose of this work was to determine to what extent complementarity carry textual signals that may help NPL applications identify the correspondent CST relations. NLP research so far has focused mainly on temporal complementarity (i.e., *Historical background* and *Follow-up* relations) (SOUZA; DI-FELIPPO, 2018), but complementary information not related to temporal attributes is very frequent in language. This can be seen in CSTNews, where *Elaboration* is the second most frequent relation (20.90%) in the corpus, corresponding to 48.10% of the total CST relations of complementarity. Thus, it is essential to explore automatic ways for identifying both.

In the process of annotating or describing the sentence pairs in the subcorpus of temporal and non-temporal complementarity, Souza and Di-Felippo (2018) and Souza (2019) have noticed that delimiting and classifying the signals are not easy tasks, given the subjectivity involved in the task. Additionally, the ML study reveals the relevance of the signals for distinguishing the different CST relations in question. In this respect, we can confidently say that signals can potentially support the automatic detection of the relations because the JRip´s classifier had 96.9% of accuracy with a small set of rules.

However, it is important to note that not all signals from the typology are machine-tractable attributes, mainly those from the pragmatic class. This means that there are no NLP tools for automatically annotating them in corpora. Thus, one future goal is to investigate only machine-treatable signals using the whole subcorpus of

complementarity (i.e., 655 pairs). This is the case of the morphological and syntactic signals, for example, which can be identified by taggers and parsers, respectively.

We emphasize the contributions of this work from two natural instances to this field of study: (i) *Descriptive Linguistics* and (ii) *NLP*. In (i), when we systematize a broad set of linguistic signals of complementarity that expand the linguistic knowledge that we had until then about the phenomenon; and in (ii) by providing subsidies (linguistic signals) for the automatic identification of complementarity, one of the most frequent linguistic phenomena in multi-document journalistic corpora, and whose identification can help in the task of automatic summarization.

Finally, our work can help to enrich CSTNews, which is the reference corpus for MDS. The delimitation and description tasks of signals (see example (4)) can be used to insert a new type or layer of annotation to the corpus. We believe that this type of linguistic annotation may be used in future research on multi-document analysis.

Acknowledgment

We thank the Coordination for the Improvement of Higher Education Personnel for the financial support and the Interinstitutional Center of Computational Linguistics for the research support.

References

ALEIXO, P.; PARDO, T.A.S. Finding Related Sentences in Multiple Documents for Multidocument Discourse Parsing of Brazilian Portuguese Texts. *In*: **Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web**. 2008. p. 298-303. DOI https://doi.org/10.1145/1809980.1810055

BELTRAME, W.; CURY, D.; MENEZES, C. S. Fique Sabendo: um Sistema de Disseminação Seletiva da Informação para Apoio à Aprendizagem. *In*: **Brazilian symposium on Computers in Education**. Rio de Janeiro – Brazil. 2012. 10p.

CARDOSO, P. C. F.; MAZIERO, E. G.; JORGE, M. L. C.; SENO, E. M. R.; DI-FELIPPO, A.; RINO, L. H. M.; NUNES, M. G. V.; PARDO, T. A. S. CSTNews: a discourse-annotated corpus for Single and Multi-Document Summarization of news texts in Brazilian Portuguese. *In*: **Proceedings of the 3rd RST Brazilian Meeting**. Cuiabá – Brazil. 2011. p. 88-105.

DAS, D.; TABOADA, M. RST signalling corpus: a corpus of signals of coherence relations. **Language Resources and Evaluation**, v. 52, n. 1, p. 149–184, 2018. DOI https://doi.org/10.1007/s10579-017-9383-x

INAM, S.; SHOAIB, M.; MAJEED, F.; SHAERJEEL, M. I. Ontology based query reformulation using rhetorical relations. **International Journal of Computer Sciences IJCS**, Vol 9, Issue 4. p. 261-268, 2012.

JURAFSKY, D; MARTIN, J. **Speech and Language Processing**: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3ª Edition (Draft), 2021. Available at: https://web.stanford.edu/~jurafsky/slp3/. Access in: 08 Sept. 2021.

KUMAR, Y. J.; SALIM, N.; RAZA, B. Cross-document structural relationship identification using supervised machine learning. **Applied Soft Computing**, v. 12, n. 10, p. 3124-3131, 2012. DOI https://doi.org/10.1016/j.asoc.2012.06.017

MANI, I. **Automatic summarization**. Vol. 3. John Benjamins Publishing. 2001. DOI https://doi.org/10.1075/nlp.3

MANN, W. C.; THOMPSON, S. A. **Rhetorical structure theory**: A theory of text organization. University of Southern California, Information Sciences Institute, 1987. DOI https://doi.org/10.1515/text.1.1988.8.3.243

MAZIERO, E. G.; JORGE, M. L. C.; PARDO, T. A. S. Identifying multi-document relations. *In*: **Proceedings of the International Workshop on Natural Language Processing and Cognitive Science**. Funchal, Madeira/Funchal. 2010. p. 60-69.

MAZIERO, E.; PARDO, T. A. CSTParser–a multi-document discourse parser. *In*: **Proceedings of the PROPOR**. Coimbra – Portugal. 2012. p. 1-3.

MAZIERO, E. G. **Identificação automática de relações multidocumento**. Master's dissertation (Masters in Computer Science and Computational Mathematics) - Institute of Mathematical and Computer Sciences, University of São Paulo, São Carlos, 2012.

MAZIERO, E. G.; JORGE, M. L. R. C.; PARDO, T. A. S. Revisiting Cross-document Structure Theory for multi-document discourse parsing. **Information Processing & Management**, v. 50, n. 2. p. 297-314, 2014. DOI https://doi.org/10.1016/j.ipm.2013.12.003

MITCHELL, T. M. Does machine learning really work? **AI magazine**, v. 18, n. 3, p. 11. 1997.

MURAKAMI, K.; NICHOLS, E.; MIZUNO, J.; WATANABE, Y.; GOTO, H.; OHKI, M. Automatic classification of semantic relations between facts and opinions. *In*: **Proceedings of 2nd workshop on NLP challenges in the information explosion Era NLPIX**. Beijing – China. 2010. p. 21–30.

NENKOVA, A.; MCKEOWN, K. Automatic summarization. **Foundations and Trends in Information Retrieval**, 5(2-3), p. 103–233, 2011. DOI https://doi.org/10.1561/1500000015

RADEV, D. R. A. Common theory of information fusion from multiple text sources step one: cross-document structure. *In*: **Proceedings of the 1st SIGdial Workshop on Discourse and Dialogue**. Volume 10. 2000. p. 74-83. DOI https://doi.org/10.3115/1117736.1117745

SHALEV-SHWARTZ, S.; BEN-DAVID, S. **Understanding Machine Learning**: From Theory to Algorithms. New York: Cambridge University Press, 2014. DOI https://doi.org/10.1017/CBO9781107298019

SOUZA, J. W. C. **Descrição linguística da complementaridade para a sumarização automática multidocumento**. *Dissertação* (Mestrado em Linguística) – Universidade Federal de São Carlos. 2015. p. 102.

SOUZA, J. W. C.; DI-FELIPPO, A. Caracterização da complementaridade temporal: subsídios para sumarização automática multidocumento. **Alfa: Revista de Linguística** (São José do Rio Preto), v. 62, p. 125-150, 2018. DOI https://doi.org/10.1590/1981-5794-1804-6

SOUZA, J. W. C. **Aprofundamento da caracterização linguístico-computacional da complementaridade em um corpus jornalístico multidocumento**. 2019. *Tese* (Doutorado em Linguística) – Universidade Federal de São Carlos, São Carlos, p. 117. 2019.

SOUZA, J. W. C. O papel do corpus de estudo no aprimoramento descritivo da complementaridade informacional multidocumento. **Revista de Estudos da Linguagem**, v. 29, n. 2, 2021. DOI https://doi.org/10.17851/2237-2083.29.2.1059-1087

TABOADA, M.; DAS, D. Annotation upon annotation: adding signalling information to a corpus of discourse relations. **Dialogue and Discourse**. v. 4, n. 2, p. 249-281, 2013. DOI https://doi.org/10.5087/dad.2013.211

WITTEN, I. H.; FRANK, E. **Data Mining**: Practical machine learning tools and techniques. 2nd edition. Morgan Kaufmann, San Francisco. 2005.

ZAHRI, N. A. H. B.; FUKUMOTO, F. Multi-document Summarization using link analysis based on rhetorical relations between sentences. *In*: **CICling Lectures Notes in Computer Science**. 2011. p. 328-338. DOI https://doi.org/10.1007/978-3-642-19437-5_27

ZHANG, Z.; BLAIR-GOLDENSOHN, S.; RADEV, D. R. Towards CST-enhanced summarization. *In*: **Proceedings of the 18th National Conference on Artificial Intelligence** (AAAI-2002), Edmonton – Canada. 2002. p. 439-446.

ZHANG, Z.; OTTERBACHER, J.; RADEV, D. R. Learning cross-document structural relationships using boosting. *In*: **Proceedings of 12th ICIKM**. New Orleans, USA. 2003. p. 124–130. DOI https://doi.org/10.1145/956863.956887

Article received in: 10.26.2021 Article approved in: 04.28.2022

DOI: 10.14393/DL52-v16n4a2022-11

A construção de um banco de dados lexicográficos em XML a partir de dados dialetais: o Processamento Automático de Linguagem Natural (PLN)

The construction of a lexicographic database in XML from dialectal data: the Natural Language Processing (NLP)

Aparecida Negri ISQUERDO[*]
Jorge Luiz Nunes dos SANTOS JUNIOR[**]

RESUMO: Este artigo situa-se na interface entre a Lexicografia (PORTO DAPENA, 2002; HARTMANN, 2016), a Dialetologia (CARDOSO, 2010; CHAMBERS; THUDGILL, 1994) e a Linguística Computacional (HABERT, 2004; PÉREZ HERNÁNDEZ; MORENO ORTIZ, 2009; HAUSSER, 2014; KURDI, 2016). Objetiva-se discutir a proposta de construção de um banco de dados em *XML (Extensible Markup Language)*, explorando os resultados obtidos com o PLN (Processamento Automático de Linguagem Natural). O arquivo *XML* também se fundamenta em parâmetros da Lexicografia Dialetal (EZQUERRA, 1997; NAVARRO CARRASCO, 1993) e está sendo alimentado com dados dialetais oriundos do Projeto Atlas Linguístico do Brasil (ALiB) documentados na região Norte do país. Para tanto, utilizou-se como editor de

ABSTRACT: This paper is situed at the interface between Lexicography (PORTO DAPENA, 2002; HARTMANN, 2016), Dialectology (CARDOSO, 2010; CHAMBERS; THUDGILL, 1994) and Computational Linguistics (HABERT, 2004; PÉREZ HERNÁNDEZ; MORENO ORTIZ, 2009; HAUSSER, 2014; KURDI, 2016). The objective is to discuss the proposal of building a database in XML (Extensible Markup Language), exploring the results obtained with NLP (Natural Language Processing). The XML file is also based on parameters of Dialectal Lexicography (ESQUERRA, 1997; NAVARRO CARRASCO, 1993) and is being fed with dialectal data from the project Atlas Linguístico do Brasil (ALiB) documented in the country's Northern region. Therefore, the jEdit software was used as a text editor and, to manage the database, the BaseX program. The

[*]Doutora em Letras (Linguística e Língua Portuguesa) pela UNESP/Araraquara. Pesquisadora Sênior/UFMS. Docente da UFMS – Estudos de Linguagens/FAALC e Letras/CPTL. ORCID: https://orcid.org/0000-0003-1129-5775. aparecida.isquerdo@gmail.com.

[**]Doutorando do Programa de Pós-Graduação em Letras da Universidade Federal de Mato Grosso do Sul, *campus* de Três Lagoas (UFMS/CPTL). Bolsista CAPES. ORCID: https://orcid.org/0000-0002-1111-4148. jorgesantosjunior@gmail.com.

texto o software *jEdit* e, para gerenciar o banco de dados, o programa *BaseX*. A extração das informações linguísticas foi realizada, no *BaseX*, a partir de uma amostra de dados e com o auxílio de expressões *X-Query*. Assim, foram executadas as seguintes manipulações de dados: i) localização de uma unidade lexical específica; ii) visualização de qualquer dado da microestrutura filtrada pelas variáveis sexo, idade, escolaridade e localidade; iii) seleção de informações a partir de uma das 14 áreas semânticas em que as questões do questionário semântico-lexical do ALiB foram organizadas. Em síntese, entende-se que a construção do banco de dados em *XML* confere agilidade em relação à extração de informações e compatibilidade dos dados para executar interfaces com outras aplicações como, por exemplo, a elaboração de um produto lexicográfico a ser publicado em suporte *on-line*.

PALAVRAS-CHAVE: Lexicografia Dialetal. Linguística Computacional. Banco de dados em XML. PLN.

linguistic information extraction was performed in the BaseX, from a sample of data and with the X-Query expressions support. Thus, the following data manipulations were performed: i) location of a specific lexical unit; ii) visualization of any microstructure data filtered by variables gender, age, education and location; iii) selection of information from one of the 14 semantic areas in which the questions of the ALiB semantic-lexical questionnaire were organized. In summary, it is understands that the construction of a XML database provides agility in concerning the information extraction and data compatibility to implement interfaces with another applications, for example, the development of a lexicographic product to be published in online support.

KEYWORDS: Dialectal Lexicography. Computational Linguistics. XML database. NLP.

1 Introdução[1]

A Tecnologia da Informação (TI) está cada vez mais presente no dia a dia das pessoas. A sua aplicação pode ser observada nos diversos segmentos da sociedade como, por exemplo, nos dispositivos que estabelecem uma comunicação virtual entre as pessoas ao redor do mundo, tecnologias que auxiliam o diagnóstico e o tratamento de doenças, *softwares* que realizam tarefas hercúleas praticamente impossíveis de se executar por mãos humanas, enfim, as aplicações da TI são múltiplas. Trata-se de uma

[1] Este artigo é resultado da metodologia que está sendo aplicada à pesquisa de doutorado do segundo autor.

área interdisciplinar que está em constante evolução e qualquer área do conhecimento pode se beneficiar de suas contribuições.

No caso das pesquisas realizadas no âmbito da Linguística, os recursos informáticos desenvolvidos pela área da TI permitem automatizar diversas etapas metodológicas, contribuindo para a elaboração de novos produtos e de novas perspectivas de acesso e análise de dados. Nesse sentido, por meio da Linguística Computacional o estudioso do ramo dos Estudos da Linguagem pode realizar o chamado Processamento Automático de Linguagem Natural (PLN), acrônimo do termo em inglês *Natural Language Processing (NLP)*, ampliando os horizontes da pesquisa científica.

Nesse cenário situa-se este artigo, que tem como objetivo discutir e refletir sobre o uso de ferramentas informáticas destinadas a criar e gerenciar *corpora* com a finalidade de executar a extração automática de dados. Para tanto, um banco de dados em *XML (Extensible Markup Language)* está sendo construído a partir de informações lexicais de cunho dialetal, com vistas a realizar o PLN.

As informações que estão sendo inseridas nesse banco de dados pertencem ao *corpus* do Projeto Atlas Linguístico do Brasil (ALiB), mais especificadamente a parcela documentada nos pontos de inquérito da rede de pontos do ALiB circunscritos à região Norte do país em 24 localidades. Esses dados são de natureza oral e precisam ser transcritos em um editor de texto para que seja possível a manipulação eletrônica.

O ALiB é um Projeto interinstitucional iniciado em 1996 com sede na Universidade Federal da Bahia[2], que tem como objetivo principal documentar e descrever as variedades linguísticas do português do Brasil representando-as por meio de cartas linguísticas. Para tanto, a equipe iniciou a coleta de dados em 2001 e percorreu 250 localidades brasileiras em busca de informantes que atendessem ao

[2] Informações mais amplas sobre o Projeto ALiB podem ser obtidas por meio de consulta ao site do projeto: www.alib.ufba.br.

perfil estabelecido pelo projeto com base nos critérios definidos pela Geolinguística[3]. Essa coleta, levada a cabo até 2013, resultou num extenso *corpus* oral que tem sido utilizado como fonte para a elaboração do *Atlas Linguístico do Brasil* que, em 2014, teve os seus dois primeiros volumes publicados. Além disso, o *corpus* do Projeto ALiB tem subsidiado diversos estudos em nível de pós-graduação que versam sobre a variação linguística da língua portuguesa.

Atualmente, há três trabalhos de caráter lexicográfico concluídos com base em dados do Projeto ALiB, a saber: O *Vocabulário dialetal da região Norte do Brasil: um estudo das capitais com base nos dados do Projeto ALiB* (CORREIA DE SOUZA, 2019)[4]; o *Vocabulário Dialetal do Centro-Oeste* (COSTA, 2018) e o *Vocabulário Dialetal Baiano* (NEIVA, 2017). Esses estudos estão filiados ao projeto do *Dicionário Dialetal Brasileiro (DDB)* (MACHADO FILHO, 2011) de cunho interinstitucional, em andamento, que tem como objetivo dar tratamento lexicográfico ao *corpus* nacional do Projeto ALiB no nível lexical. Todavia, esses três trabalhos, produzidos no âmbito da pós-graduação, não contemplaram a transformação dos dados do ALiB em *XML*, a exemplo do realizado com o projeto de tese a que se associa este estudo[5].

Em síntese, este artigo focaliza parâmetros para a construção de uma base de dados em *XML* a partir do *corpus* dialetal do Projeto ALiB relativo à região Norte do Brasil, atentando para uma organização lexicográfica e eletrônica que permita, num primeiro momento, realizar a extração automática de informações linguísticas e, futuramente, desenvolver uma aplicação web que servirá de suporte para a elaboração de um vocabulário dialetal *on-line*.

---

[3] No item destinado à metodologia são apresentadas informações mais específicas quanto à metodologia do Projeto ALiB.

[4] Tendo em vista que Correia de Souza (2019) trabalhou com dados das capitais da região Norte e a presente pesquisa está trabalhando com os dados do interior, toda a região Norte do Brasil ficará coberta em relação ao tratamento lexicográfico dos dados dialetais do ALiB.

[5] Outros estudos com base em dados do Projeto ALiB e filiados ao *Dicionário Dialetal Brasileiro* estão em andamento, sob a orientação do prof. Américo Venâncio Machado Filho, na Universidade Federal da Bahia.

2 Pressupostos teóricos

O desenvolvimento do banco de dados já referenciado fundamenta-se na Lexicografia (PORTO DAPENA, 2002; HARTMANN, 2016), mais especificamente na Lexicografia Dialetal (EZQUERRA, 1997; NAVARRO CARRASCO, 1993), na Dialetologia (CARDOSO, 2010; CHAMBERS; THUDGILL, 1994) e na Linguística Computacional (HABERT, 2004; PÉREZ HERNÁNDEZ; MORENO ORTIZ, 2009; HAUSSER, 2014; KURDI, 2016).

2.1 Lexicografia

Um dos legados que a Lexicografia deixa para as gerações futuras é o registro do léxico de uma língua numa determinada sincronia. Assim, a missão dos lexicógrafos é a de registrar o repertório lexical de uma língua natural que, por sua vez, envolve dois aspectos intrínsecos: o linguístico e o extralinguístico. O aspecto linguístico abarca todas as regras convencionadas pelas normas gramaticais, subsidiando a comunicação humana por meio do *sistema de possibilidades* (COSERIU, 1980, p. 122). Por sua vez, o aspecto extralinguístico é resultado do uso do *sistema de possibilidades* por indivíduos plurais do ponto de vista social, econômico, cultural e geográfico. Vale destacar que a carga extralinguística das línguas naturais exerce influências no falar de um grupo linguístico capaz de (re)configurar a norma lexical de tempos em tempos.

Dessa maneira, para que a norma lexical de um povo não se perca com o passar dos anos é preciso que seja registrada por obras lexicográficas. Na arte de fazer dicionários (PORTO DAPENA, 2002, p. 20) uma fotografia do falar contemporâneo é construída a cada dicionário elaborado.

Tendo em vista a plasticidade das línguas, o registro do léxico por meio dos dicionários tem sua relevância no transcorrer da história. É impossível saber quais

unidades lexicais entrarão em desuso no futuro e quais estarão vivas na fala do povo daqui a 100 anos, dado o caráter dinâmico do léxico (BIDERMAN, 2001, p. 197).

Nesse sentido, as marcas de uso registradas nos dicionários gerais configuram-se como formas de repertoriar os múltiplos falares, incluindo a perspectiva espacial. Além disso, constituem-se em formas de consulta lexicográfica no que diz respeito ao léxico de uma dada região. No entanto, em razão de essa tipologia de obras ter o objetivo de abarcar a língua como um todo, o registro da variação diatópica não é o principal objetivo dos dicionários gerais, o que pode acarretar, em alguns casos, uma diminuição do rigor metodológico na identificação e descrição dos falares regionais.

Diferentemente do que ocorre na Lexicografia Geral, os dicionários especializados têm como objetivo focar em uma determinada parcela do conhecimento ou uma área de uso como ocorre, por exemplo, na Lexicografia Dialetal que tem como foco repertoriar os falares regionais, partindo de dados empíricos consistentes e registrados, por exemplo, por pesquisas orientadas pela Dialetologia.

Ressalta-se, dessa forma, que a Lexicografia Dialetal compartilha dos mesmos elementos que estruturam uma obra lexicográfica de cunho geral como, por exemplo, a *front matter*, a *middle matter* e a *back matter* que constituem a macroestrutura de um dicionário, bem como o conjunto das acepções de cada verbete que representa a microestrutura dessas obras (HARTMANN, 2016, p. 59).

2.2 Lexicografia Dialetal

Para que uma obra lexicográfica possa ser classificada como dialetal é preciso que represente, de fato, a variação lexical evidenciada por falantes pertencentes a uma dada área geográfica. Nesse particular, os atlas linguísticos configuram-se como fontes confiáveis da norma lexical, conferindo rigor metodológico à obra lexicográfica de cunho regional (NAVARRO CARRASCO, 1993, p. 93).

A Lexicografia Geral, por sua vez, nem sempre utiliza os atlas linguísticos publicados como fonte de atestação do uso e, por extensão, do sentido assumido por determinados itens lexicais no âmbito de dialetos circunscritos a um determinado território. Desse modo, não é raro o registro de marcas de uso em definições fornecidas por dicionários gerais que não representam com fidedignidade a perspectiva diatópica, como mostra Sá (2021, p. 224) que identificou divergências entre as acepções lexicográficas de *igarapé* nos dicionários Houaiss (2009), Ferreira (2010) e Michaelis (2015), ao compará-las com os dados do *Atlas Linguístico do Amazonas (ALAM)* (CRUZ, 2004) e do *Atlas Linguístico do Sul Amazonense (ALSAM)* (MAIA, 2018).

Desse modo, além de o estudo de Sá (2021) indicar a necessidade da atualização dos dicionários mencionados em relação à marca de uso e à definição da unidade lexical *igarapé,* também atesta que a Lexicografia e a Dialetologia podem se relacionar de modo a estabelecerem uma interdisciplinaridade que beneficie ambas as disciplinas, pois a Lexicografia recorre aos dados dialetais como fonte para registro das marcas de uso nos dicionários, enquanto a Dialetologia vale-se, frequentemente, de informações fornecidas pela Lexicografia para comprovar seus dados (EZQUERRA, 1997, p. 79). Vale destacar, ainda, que a natureza interdisciplinar da Lexicografia permite intersecções com várias áreas do conhecimento, dentre as quais as disciplinas teóricas e as orientações metodológicas que embasam a pesquisa que deu origem a este artigo.

2.3 Dialetologia

Para que a Dialetologia possa identificar a variação linguística em um determinado espaço geográfico é preciso construir um *corpus,* documentado *in loco,* que registre a realidade de fala das comunidades que habitam esse espaço. A coleta do falar regional pautada em parâmetros da Dialetologia e da Geolinguística é realizada por meio de entrevistas, geralmente gravadas em áudio, com informantes naturais da

região investigada. A documentação de dados geolinguísticos é orientada por um questionário linguístico de natureza onomasiológica e a seleção dos informantes considera o perfil previamente definido que considera variáveis geográficas e sociais, de maneira a haver o controle necessário para garantir a comparabilidade dos dados.

Como mencionado anteriormente, o *corpus* do Projeto ALiB tem sido estudado no âmbito de diversos trabalhos acadêmicos e a análise desse material tem revelado fenômenos linguísticos significativos no falar dos brasileiros. É preciso considerar também que a metodologia do Projeto ALiB tem inspirado a realização de outros estudos acerca da norma lexical de diferentes regiões brasileiras.

Além disso, os dados coletados a partir dos critérios dialetais podem servir como uma fonte empírica para pesquisas em outras disciplinas linguísticas (CHAMBERS; TRUDGILL, 1994, p. 45), atestando a dimensão interdisciplinar da Dialetologia.

Destaca-se, ainda, que a variação linguística estudada pela Dialetologia auxilia a descrição da norma lexical de uma comunidade de falantes, levando em consideração fatores extralinguísticos como, por exemplo, a riqueza cultural, as influências de outras línguas, além de reflexos da formação demográfica no decorrer da história de um povo(ado) (CARDOSO, 2010, p. 15).

Além da Lexicografia e da Dialetologia, outra disciplina liga-se ao arcabouço teórico-metodológico deste estudo aumentando intersecções que se entremeiam, harmoniosamente, a partir das contribuições da Linguística Computacional.

2.4 Linguística Computacional

Hausser (2014, p. xix) defende que o objetivo básico da Linguística Computacional é "Transmitting information by means of a natural language like

Chinese, English, or German is a real and well-structured procedure.[6]" A ponderação desse autor remete à comparação entre as línguas naturais e as linguagens de programação que podem ser vistas, desde uma perspectiva linguística, como linguagens estrangeiras para os seres humanos.

Tendo em vista que os computadores ainda não entendem as línguas naturais, na realização de procedimentos que envolvem o processamento de *corpora* por meio de ferramentas de Processamento Automático de Linguagem Natural, também chamado de Processamento de Linguagem Natural (PLN) (PÉREZ HERNÁNDEZ; MORENO ORTIZ, 2009, p. 68), são utilizadas linguagens que o computador compreende. Desse modo, a tradução para o computador das manipulações textuais que um pesquisador deseja executar favorece a apropriação dos benefícios que as ferramentas computacionais oferecem no campo da pesquisa científica.

Em outras palavras, a manipulação de dados textuais, escritos em linguagem humana, por meio de softwares que possibilitam a recuperação de informações de modo automatizado tem permitido o desenvolvimento de "um conjunto de técnicas e ferramentas capazes de realizar tarefas que vão da identificação de estruturas morfossintáticas até a atribuição de informação semântica a porções de texto, como o reconhecimento de entidades nomeadas" (HIGUCHI; FREITAS, 2017, p. 1-2).

Desse modo, a Linguística Computacional "[…] may be defined as an application of computer science to modeling natural language communication as a software system. This includes a linguistic analysis of natural language using computers.7" (HAUSSER, 2014, p. 3).

---

[6] "Transmitir informações por meio de um idioma natural como chinês, inglês ou alemão é um procedimento real e bem estruturado." - (T.N.).

[7] "[...] pode der definida como uma aplicação da Ciência da Computação para modelagem de comunicação em linguagem natural por um sistema informatizado. Isso inclui uma análise da linguagem natural utilizando computadores." – (T.N.).

No cenário brasileiro, o Núcleo Interinstitucional de Linguística Computacional (NILC)[8], da Universidade de São Paulo, em São Carlos, se destaca pelo pioneirismo em termos de análise da linguagem natural por meio de computadores. O grupo nasceu com o "objetivo de formar recursos humanos e desenvolver pesquisa e sistemas de PLN especialmente para o português do Brasil (PB)" (NUNES; ALUÍSIO; PARDO, 2010, p. 13).

Para tanto, o NILC tem focado esforços na criação de *corpora* em língua portuguesa, além de investir na elaboração de ferramentas de PLN que realizam tarefas como, por exemplo, a marcação morfossintática; a simplificação ou sumarização de um texto; assistentes de escrita e de fala destinado a estudantes de inglês como língua estrangeira; analisador sintático-semântico; analisador de discurso; tradutores automáticos, entre outras ferramentas que podem ser acessadas no portal do grupo.

Vale destacar que, mesmo havendo no mercado uma infinidade de *softwares* criados para o desenvolvimento de trabalhos voltados para o PLN, muitas vezes, o linguista precisará desenvolver suas próprias aplicações informáticas para que os objetivos de uma pesquisa sejam, satisfatoriamente, alcançados. Ocorre que, em determinadas situações, os *softwares* utilizados não oferecem soluções para a execução de determinadas tarefas, forçando o pesquisador a realizar o trabalho manualmente ou até mesmo desistir de seus propósitos por falta de suporte eletrônico adequado.

Nessas ocasiões, o estudioso poderá recorrer à programação. A primeira opção é terceirizar a missão de criar soluções informáticas que contemplem as necessidades da pesquisa a um profissional da área de TI. A segunda opção, mais vantajosa, é o linguista se apropriar do funcionamento básico de algumas linguagens de marcação e de programação para definir os caminhos que melhor atendam ao objeto de seu estudo. No entanto, ainda será preciso a ajuda de um profissional de TI para definir os

[8] Para maiores informações acesse: <http://www.nilc.icmc.usp.br/nilc/index.php>.

passos a serem seguidos nessa investida, ou seja, elencar quais linguagens e temas da área da Computação o linguista deverá estudar para atender os requisitos metodológicos de sua pesquisa. É preciso observar que o campo da Informática é extremamente vasto e as possibilidades de se executar uma tarefa de maneira automática são múltiplas. Dessa maneira, não há a necessidade, por exemplo, de se dominar uma gama de conteúdos da TI para construir um arquivo *XML*. Nesse caso, basta identificar quais conhecimentos são basilares para a execução da tarefa e lançar mão deles.

Vale destacar que, ao investir na construção de aplicações informáticas, o linguista amplia as possibilidades de visualização dos dados, abrindo novos horizontes, no que diz respeito à extração de informações linguísticas e à compreensão dos fenômenos observados (HABERT, 2005, sem página). Além disso, o uso de ferramentas que executam o PLN oferece a possibilidade de testar empiricamente modelos teóricos (KURDI, 2016, p. x-ix).

3 Metodologia

Os dados disponíveis no *corpus* construído pelo Projeto ALiB estão em formato de áudio e organizados por regiões, estados e cidades. Nesse banco de dados oral há informações importantes que serão recuperadas de maneira eletrônica como, por exemplo, as perguntas e as respostas de cada entrevista, além das variantes lexicais que permitem identificar a localidade, a idade, o sexo e a escolaridade de cada entrevistado.

A metodologia de coleta de dados do ALiB contempla um amplo questionário estruturado para registrar a variação lexical, fonológica e morfossintática, por meio de três grupos de perguntas, a saber: i) Questionário Fonético-fonológico (QFF) com 159 perguntas; ii) Questionário Semântico-lexical (QSL) com 202 perguntas; iii) Questionário Morfossintático (QMS) com 49 perguntas. A seleção dos informantes

para as entrevistas, por sua vez, foi pautada nos seguintes critérios: i) oito informantes nas capitais e quatro informantes nas regiões do interior, distribuídos equitativamenta por faixas etárias (18 a 30 anos e 50 a 65 anos); por sexo (masculino e feminino); ii) por escolaridade (informantes das cidades do interior devem possuir o ensino fundamental incompleto enquanto nas capitais são controlados dois níveis de escolaridade: ensino fundamental incompleto e curso superior completo (COMITÊ NACIONAL..., 2001, p. viii).

O banco de dados focalizado neste estudo e ainda em construção incorpora as respostas fornecidas pelos 120 informantes do ALiB, naturais das 24 localidades da rede de pontos relativas à região Norte do Brasil, para as 202 perguntas do QSL/ALiB. Para tanto, a criação do arquivo com extensão *.xml* foi realizada de modo a organizar as informações de maneira lexicográfica, estabelecendo-se, dessa forma, um modelo de microestrutura descrito no quadro a seguir:

Quadro 1 – Organização lexicográfica dos dados do ALiB – Região Norte do Brasil.

Informações lexicográficas	**Descrição**
1) lema	Denominação fornecida como resposta pelo informante para a pergunta formulada pelo entrevistador.
2) pergunta	Número e pergunta do QSL.
3) abonação	Contexto de fala do informante sobre o emprego de determinada denominação.
4) observação	Informações adicionais identificadas durante a audição do inquérito.
5) fonética	Registro da variação fonética do referente do item lexical documentado.
6) áudio	Execução de áudio e tempo de gravação.
7) remissiva	Variação de respostas para a mesma pergunta do QSL.
8) informante	Identificação conforme a idade, o sexo e a escolaridade.
9) legenda dialetal	Dados relativos à localidade em que o dado foi registrado e indicação do informante que o mencionou.
10) classe gramatical	Classe gramatical da unidade lexical registrada.
11) definição	Texto da definição.

Fonte: elaboração dos autores.

O banco de dados em *XML*, construído no *jEdit*[9], foi planejado para estruturar os dados dialetais em 11 categorias de informações lexicográficas que foram transformadas em *tags*[10]. É importante destacar que essa fase inicial da construção do arquivo *.xml* exige um planejamento prévio de como as informações serão armazenadas e de quantas *tags* serão necessárias para que se possa realizar a extração das informações automaticamente. Isso significa que, se o pesquisador sentir a necessidade de realizar modificações nas *tags* ou até mesmo inserir uma nova *tag*, deverá realizar esse procedimento manualmente.

Como os dados de entrada estão em formato de áudio e, nesse momento, a pesquisa concentra-se na tarefa de transcrição dos áudios em formato de texto, ainda não é possível realizar o pré-processamento automático do *corpus*. Porém, assim que todas as transcrições estiverem finalizadas, os dados poderão ser submetidos à etiquetação morfossintática, gerando a classificação gramatical[11] de cada unidade lexical armazenada no banco de dados.

Vale destacar que as informações estão armazenadas em dois tipos de *tags*, *a saber: i) tag* do tipo *elemento:* destinadas às informações textuais e que não possuem qualquer tipo de restrição quanto à quantidade de caracteres[12]; ii) *tag* do tipo *atributo*: utilizadas para armazenar informações curtas e específicas como, por exemplo, as variantes sexo, idade, localidade e escolaridade dentre outras.

Desse modo, os dados estão sendo estruturados da seguinte forma:

---

[9] Editor de texto próprio para programação.

[10] As *tags* armazenam os dados e são escritas de modo a poder identificar seu conteúdo genérico. Por exemplo, em <lema>igarapé</lema>, *igarapé* está etiquetado pelo elemento<lema> de abertura e </ lema> de fechamento.

[11] A classificação gramatical de um *corpus* é realizada por meio de *taggers*, ou seja, *softwares* de *Part of Speeck* (POS). Para maiores informações sobre esse assunto acessar a página do NILC: http://nilc.icmc.usp.br/nilc/tools/nilctaggers.html.

[12] Essa nomenclatura é utilizada para distinguir caracteres de letras, já que o computador, em um primeiro momento, só reconhece caracteres.

Quadro 2 – Estrutura *XML* do banco de dados da pesquisa.

```
<?xml version="1.0" encoding="utf8" ?>
<!DOCTYPE dicio SYSTEM "corpus-1.dtd">

<dicio>
    <entrada id="acid.geo.água.1" abc="i">
        <lema>igarapé</lema>
        <perg campo="Acidentes geográficos" ref="QSL-1">Como chama um rio pequeno,
de uns dois metros de largura?</perg>
        <abo>Aqui é garapé, né...(Tem outros nomes?) Não. Aqui é garapé só.</abo>
        <obs></obs>
        <fone>garapé</fone>
        <aud src="nome-do-arquivo" type="mp3">001_01_QFF01_QSL051_A
48:56</aud>
        <ver name="rio" ref="acid.geo.água.230"/>
        <info sexo="M" escolaridade="F" idade="J" >Wilson, 28 anos</info>
        <lg ponto="1" cidade="Oiapoque" estado="AP" />
        <gram></gram>
        <def></def>
    </entrada>
</dicio>
```

Fonte: elaboração dos autores.

Observa-se, a partir dos dados registrados no quadro 2, que as duas primeiras linhas do *XML* são destinadas à descrição do documento, indicação da versão e à codificação de caracteres (utf8) e faz referência ao *Document Type Definition* (*DTD)*[13] que é um conjunto de diretrizes que estabelecem a arquitetura do arquivo *XML.*

Todos os dados do *XML* estão inseridos dentro da *tag* <dicio></dicio> e cada conjunto de dados, que representa cada questão do QSL/ALiB, está armazenado na tag <entrada></entrada>. Por sua vez, dentro de cada *tag* <entrada></entrada>, nomeada por uma identificação única no banco de dados (id), encontram-se as informações lexicográficas, mencionadas no quadro 1, que ao seu turno, receberam

[13] O *DTD* é um arquivo onde estão escritas as regras que estruturam o arquivo *XML.* Sua elaboração é importante, pois garante a validação do *XML* e sua compatibilidade com outros *softwares* e/ou linguagens de programação.

individualmente uma uma *tag* que armazena dados de acordo com a descrição apresentada no quadro que segue:

Quadro 3 – Descrição das *tags* que armazenam as informações lexicográficas no interior de cada *tag* <entrada></entrada>.

Tag	**Descrição**
<lema></lema>	Armazena os lemas em formato de texto. É uma *tag* do tipo elemento.
<abo></abo>	Abreviação de *abonação* e armazena a fala de cada informante em formato de texto. É uma *tag* do tipo elemento.
<obs></obs>	Abreviação de *observação* que armazena as informações que o pesquisador julgar importantes na etapa da transcrição dos áudios.
<fone></fone>	Abreviação de *fonética* e armazena a ocorrência de variação fonética para determinado lema.
<aud src="nome-do-arquivo" type="mp3"> 001_01_QFF01_QSL051_A 48:56</aud>	Abreviação de *áudio*. Essa *tag* contém o atributo *src* que indica o nome do arquivo de áudio a ser selecionado pela ferramenta de áudio no verbete da aplicação web, além do atributo *type* que especifica o formato do áudio (mp3) selecionado para a aplicação web. Há também uma informação textual que indica o nome do arquivo da entrevista, bem como a indicação dos minutos e segundos em que a fala do informante foi registrada.
<ver name="rio" ref="acid.geo.água.230"/>	Essa *tag* contém somente atributos e é responsável pela execução do sistema de remissivas. Desse modo, o lema em questão recebe um link que remete para outro lema, nomeado pelo atributo *name* e ligado pelo atributo *ref.*
<info sexo="M" escolaridade="F" idade="J" > 28 anos</info>	Abreviação de *informante* e armazena as variáveis sexo, escolaridade e idade em formato de atributos. Há também informações textuais sobre o informante que podem ser adicionadas.
<lg ponto="1" cidade="Oiapoque" estado="AP" />	Abreviação da *legenda dialetal* que indica, em forma de atributos, o número da rede de pontos do ALiB, a cidade e o estado.
<gram></gram>	Abreviação de *gramática* e armazena a classe gramatical do lema em uma *tag* do tipo elemento.
<def></def>	Abreviação de *definição* que armazena o texto definitório do verbete em uma *tag* do tipo elemento.

Fonte: elaboração do autores.

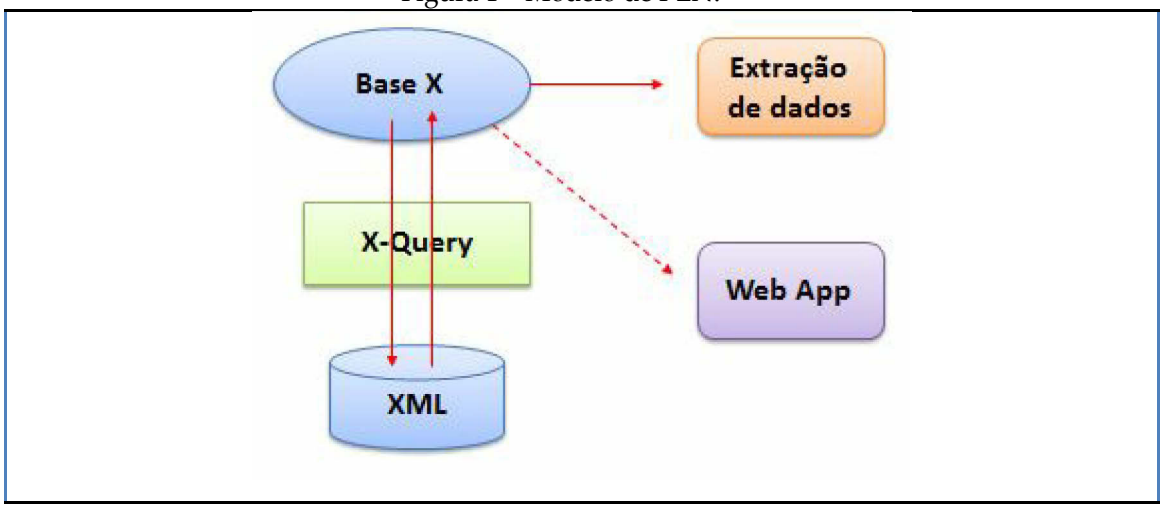
Essa estrutura *XML* permite recuperar as informações de acordo com o aspecto a ser observado. Assim, se o pesquisador deseja, por exemplo, visualizar a resposta para a questão número 1 do QSL/ALiB fornecida por um informante jovem, do sexo masculino e morador do município de Oiapoque/AP, deverá orientar o computador para que realize essa filtragem. Essas especificações precisam ser traduzidas para uma linguagem compreensível pelo computador. Assim, ao escrever essa filtragem de dados em forma de linhas de código, o resultado será o conteúdo que se encontra na *tag* <abo>Aqui é garapé, né...(Tem outros nomes?) Não. Aqui é garapé só.</abo>, por exemplo.

Nesse sentido, observa-se que essa é uma das maiores contribuições que o *XML* pode oferecer ao pesquisador, no que diz respeito à manipulação eletrônica de um *corpus*, ou seja, a possibilidade de acessar dados de múltiplas formas, de acordo com o objetivo de cada pesquisa.

Ressalte-se, ainda, que a construção e a alimentação do banco de dados em *XML* configuram-se como os primeiros passos para que se possa realizar o PLN. Isso significa que a base de dados por si só não faz muita coisa. A extração de informações, como já mencionado, é realizada por meio de linhas de código e esses códigos devem ser escritos em um *software* específico, chamado *BaseX*, que realiza o gerenciamento e o processamento de bancos de dados. Dentre suas utilidades, permite realizar a recuperação de informações por meio de expressões *X-Query*[14]. A figura a seguir ilustra os processos realizados na pesquisa, até o estágio atual, a partir do arquivo *XML* e do gerenciador de banco de dados:

[14] As expressões *X-Query* podem ser entendidas, de uma maneira ampla, como linhas de código escritas no editor do *BaseX* e que são responsáveis por filtrar e extrair as informações que estão armazenadas no banco de dados.

Figura 1 – Modelo de PLN.



Fonte: elaboração dos autores.

A Figura 1 traz, pois, um exemplo de PLN a partir do software *BaseX*. A extração de dados ocorre quando o usuário solicita um pedido por meio de uma expressão *X-Query*, que especifica o dado a ser retirado do arquivo *XML*, mostrando o resultado em uma janela do *software*. Esse processo depende da escrita correta das expressões *X-Query* e esse modelo de PLN pode ser ampliado, por exemplo, na construção de uma aplicação web (Web App) na qual um conjunto de dados é selecionado para ser exibido em uma página na Internet.

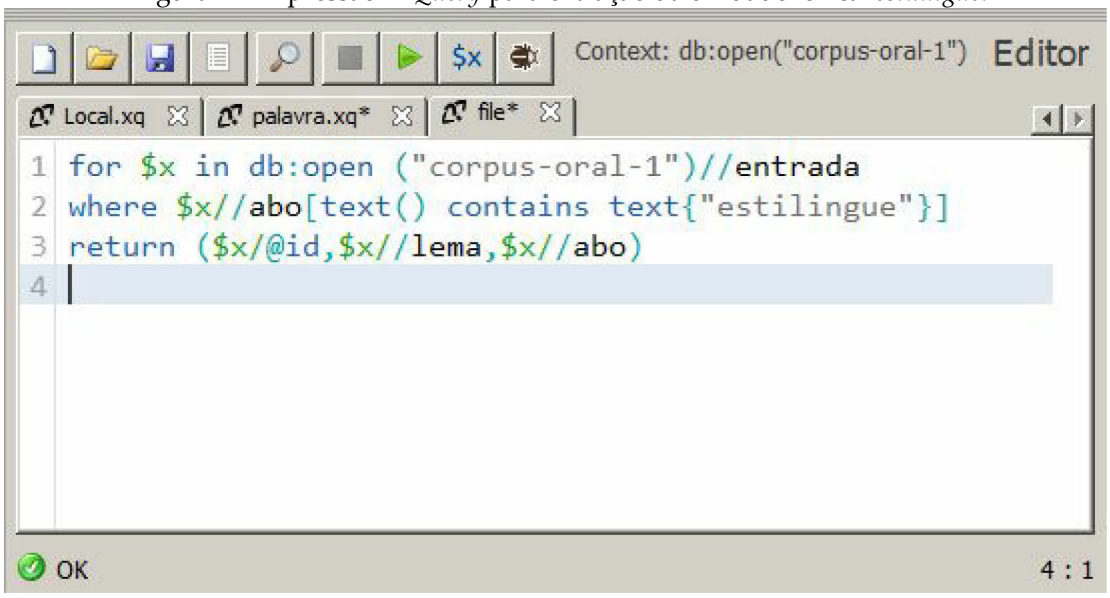
A partir de uma amostra dos dados digitalizados é possível realizar a extração de informações relevantes como, por exemplo: i) a localização de uma unidade lexical específica; ii) a visualização de qualquer dado da microestrutura filtrada pelas variáveis sexo, idade, escolaridade e localidade; iii) a seleção de informações a partir de uma das 14 áreas semânticas a que as questões do QSL/ALiB se vinculam. Essas manipulações são detalhadas a seguir.

3.1 A localização de uma unidade lexical específica

Toda extração de informações no banco de dados deve ser solicitada por meio de uma expressão *X-Query*, escrita no editor do software *BaseX*. A função da *X-Query* é localizar o dado a partir dos parâmetros de filtragem escritos nas linhas de código

do editor. O resultado da busca é exibido em uma janela ao lado do editor. A imagem a seguir ilustra como solicitar a exibição de unidades lexicais:

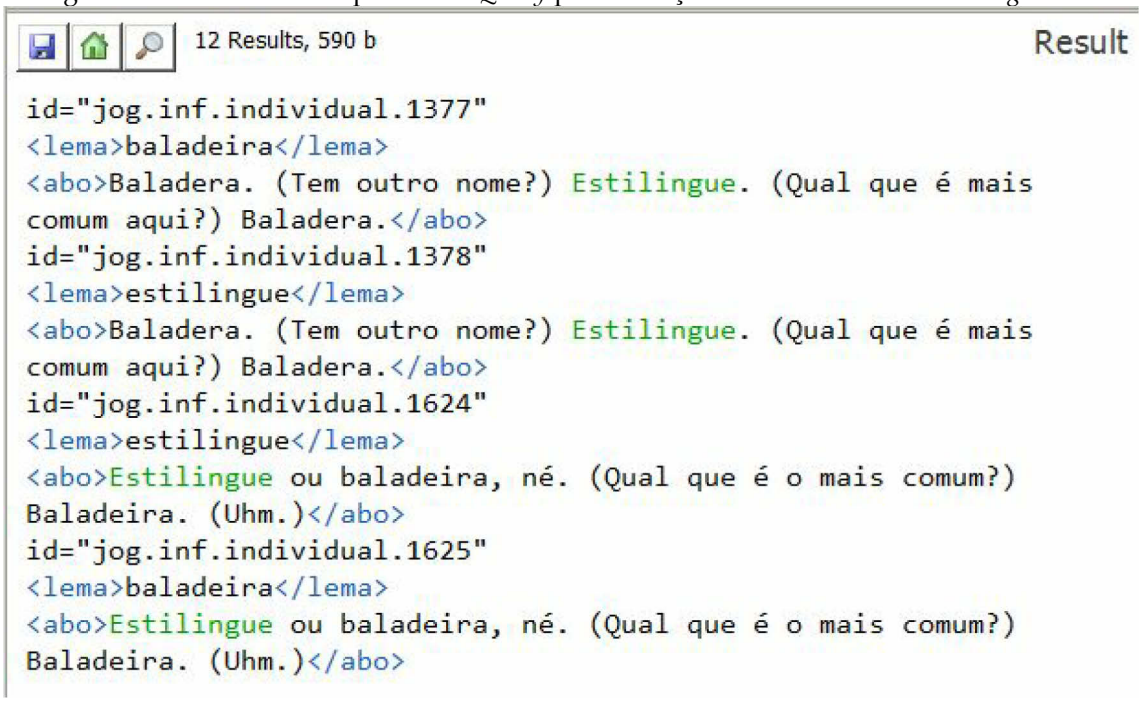
Figura 2 – Expressão *X-Query* para extração da unidade lexical *estilingue*.



Fonte: software *BaseX*.

Na Figura 2, foram escritas quatro linhas de código. Sem entrar em detalhes técnicos, é possível resumir a função de cada linha, a fim de explicar o funcionamento básico dessa expressão *X-Query*. O PLN ocorre a partir desses comandos e, ao alterá-los adequadamente, uma extração de dados diferente é processada. Desse modo, foi criada na linha 1 uma variável *$x* para armazenar a informação que será extraída do banco de dados *corpus-oral-1*. A busca abrange todas as entradas do banco de dados a partir do comando *//entrada*. Na linha 2, o comando especifica que o dado requerido é do tipo texto e que deverá ser extraído das falas dos informantes, ou seja, dos dados armazenados nas *tags* <abo></abo> (abreviação de abonação). A unidade lexical *estilingue* foi pesquisada e o resultado da busca, configurado na linha 3, exibirá a identificação da entrada, ou seja, a id, o lema e o contexto. O resultado pode ser visto na imagem a seguir:

Figura 3 – Resultado da expressão *X-Query* para extração de unidade lexical *estilingue.*



Fonte: software BaseX.

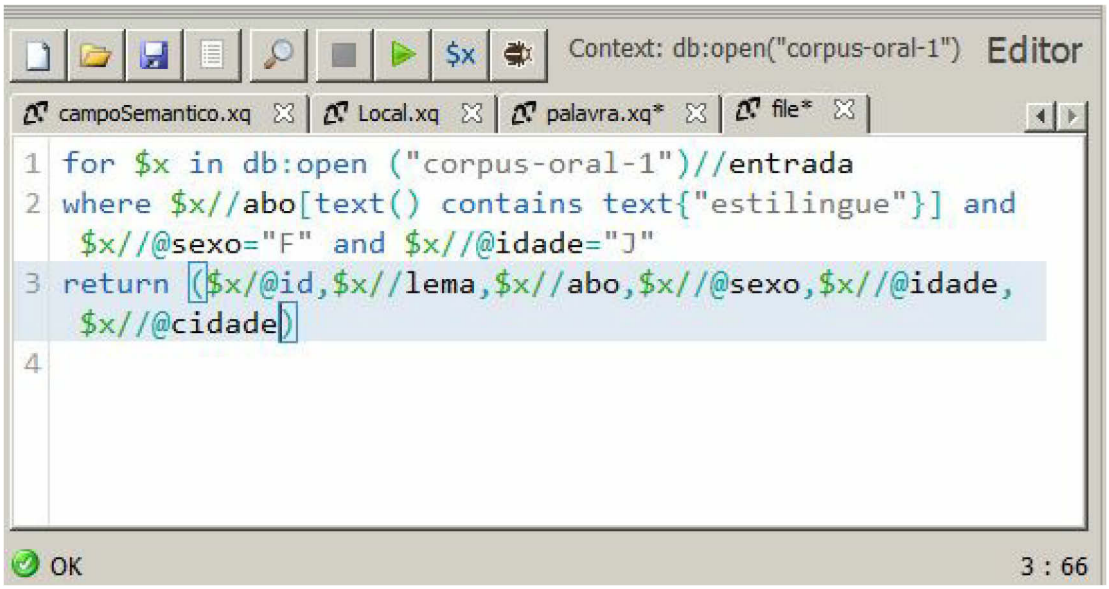
Os dados exibidos na Figura 3 seguem a ordem do comando escrito na linha 3 da *X-Query* apresentada na Figura 2. Desse modo, nota-se que há um conjunto de quatro entradas que são identificadas pelas quatro linhas que contêm o atributo id. Após cada id, há uma sequência de caracteres e números entre aspas que identificam a entrada. Logo abaixo da id situa-se o lema e, em seguida, a abonação, que contém a unidade lexical pesquisada destacada na cor verde.

Para realizar buscas de outras informações no banco de dados é preciso compreender minimamente o funcionamento das expressões *X-Query*, já que é por meio dessa linguagem que o editor irá entender e processar a busca. Vale destacar que a linha 1 da *X-Query* permanecerá a mesma, pois as buscas se referem ao mesmo banco de dados. Desse modo, apenas as linhas 2 e 3 devem ser editadas de acordo com o tipo de dado a ser extraído.

3.2 A visualização de dados da microestrutura filtrada pelas variáveis sexo, idade, escolaridade e localidade

Para realizar extrações de dados a partir do controle das variáveis sexo, idade, escolaridade e localidade é preciso informar, na linha 2 do editor, o atributo correspondente à variável que se deseja pesquisar e, na sequência, indicar na linha 3 quais elementos da microestrutura deverão figurar no resultado dessa busca, como é possível constatar na figura a seguir:

Figura 4 – Expressão *X-Query* para extração de unidades lexicais com controle das variáveis *sexo* e *idade*.

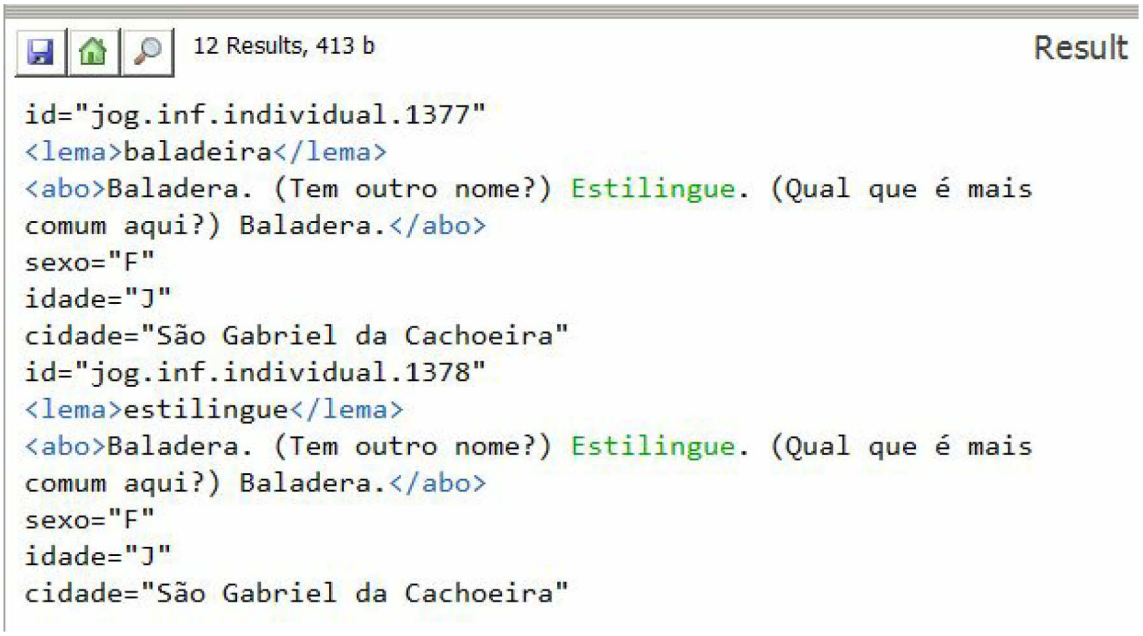


Fonte: software *BaseX*.

Observa-se, na Figura 4, que foram acrescidas na linha 2 do editor mais duas condições a serem calculadas pela *X-Query*, especificadas pelos atributos *@sexo* e *@idade*. Desse modo, solicita-se, por exemplo, que o software mostre a unidade lexical *estilingue* mencionada por informantes jovens e do sexo feminino. O resultado está configurado para exibir a id, o lema, a abonação, o sexo, a idade e a cidade. Vale lembrar que esse resultado pode ser configurado para obter qualquer dado da microestrutura relativo a uma extração. Desse modo, repetiu-se a exibição dos dados

referentes ao sexo e a idade, na linha 3 do editor, apenas para confirmar o dado solicitado. O resultado pode ser observado na figura a seguir:

Figura 5 – Resultado da expressão *X-Query* para extração de unidades lexicais com controle das variáveis *sexo* e *idade.*



Fonte: software BaseX.

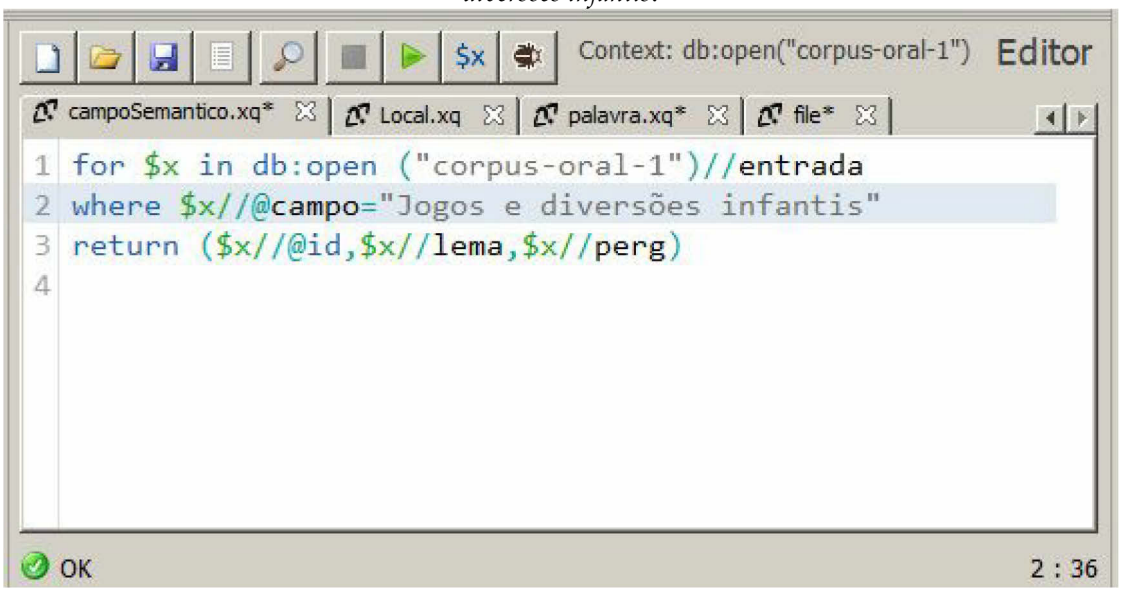
Os dados da Figura 5 demonstram que o resultado da busca trouxe os lemas *baladeira* e *estilingue*, mencionados por uma informante jovem, do sexo feminino e moradora da cidade de São Gabriel da Cachoeira/AM. Essa pesquisa no banco de dados também foi feita alternando-se as variáveis sexo (masculino e feminino) e idade (jovem e idoso). No entanto, não houve resultados com os demais perfis de informantes, o que se justifica pelo estágio inicial de alimentação do banco de dados.

Vale destacar que todos os informantes das localidades do interior da rede de pontos do Projeto ALiB possuem o nível fundamental incompleto de escolaridade e, por essa razão, não há a necessidade de filtrar os dados a partir da variável escolaridade. Porém, se necessária a extração desse tipo de dado, uma condição na linha 2 do editor deverá ser inserida para que a expressão *X-Query* processe os resultados levando em consideração o nível de escolaridade.

3.3 A seleção de informações a partir da área semântica *jogos e diversões infantis* (QSL/ALiB)[15]

A estrutura dos elementos e dos atributos que formam o banco de dados *XML* foi desenhada com o propósito de permitir a extração de informações a partir de uma das 14 áreas semânticas em que estão distribuídas as 202 perguntas do QSL. Para executar esse processamento, é preciso escrever a expressão *X-Query* no editor do *software Base-X*:

Figura 6 – Expressão *X-Query* para extração de unidades lexicais a partir da área semântica *Jogos e diversões infantis*.

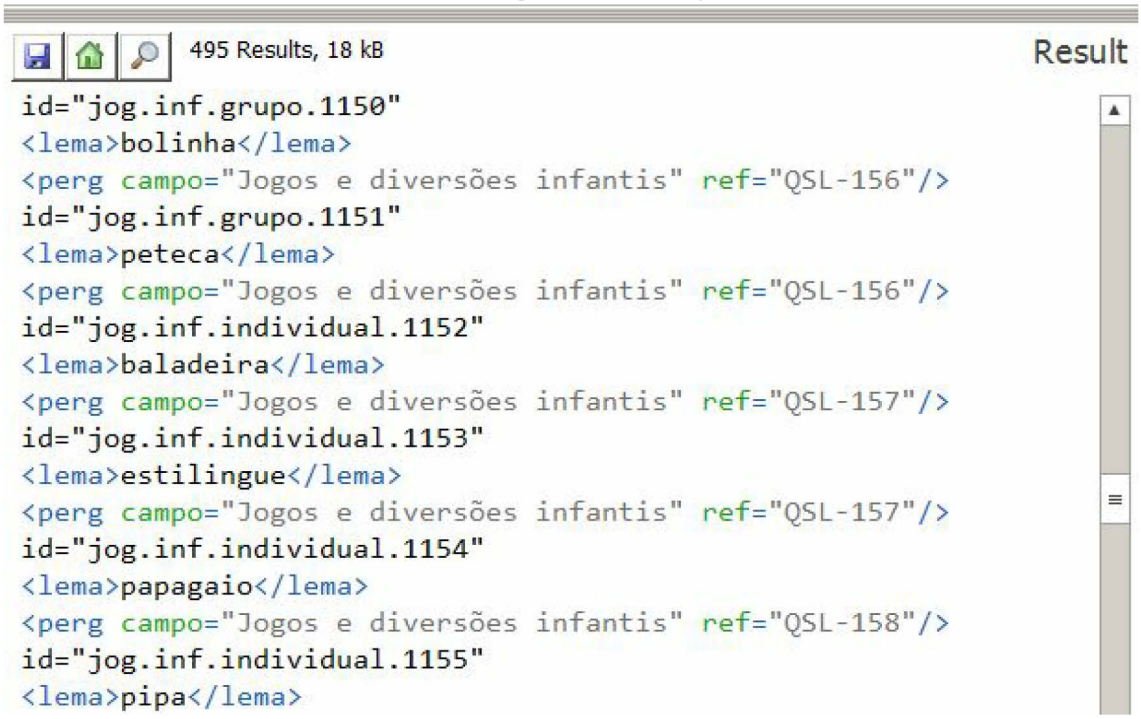


Fonte: software *BaseX*.

Na Figura 6, observa-se que os dados requeridos são aqueles que apresentam o atributo *@campo= Jogos e diversões infantis"*, que é especificado na linha 2 do editor. O resultado dessa extração está configurada, na linha 3 da expressão *X-Query*, para exibir a id, o lema e a pergunta. Esses elementos da microestrutura podem ser visualizados na figura a seguir:

[15] Áreas semânticas contempladas pelo Questionário Semântico-lexical do Projeto ALiB: Acidentes geográficos, Fenômenos atmosféricos, Astros e tempo, Atividades agropastoris, Fauna, Corpo humano, Ciclos da vida, Convívio e comportamento social, Religião e crenças, Jogos e diversões infantis, Habitação, Alimentação e cozinha, Vestuário e acessórios, Vida urbana.

Figura 7 – Resultado da expressão *X-Query* para extração de unidades lexicais a partir da área semântica *Jogos e diversões infantis*.



Fonte: software BaseX.

Nota-se, na Figura 7, uma pequena parcela dos resultados obtidos com a operacionalização da proposta, contendo seis *tags* com suas respectivas informações, a saber: a id, o lema e a *tag* que armazena os dados da pergunta do QSL em questão, isto é, a área semântica e o número da pergunta. Como anteriormente pontuado, cada extração de dados pode exibir qualquer informação presente na microestrutura, bastando adicioná-la na linha 3 da expressão *X-Query*.

Vale frisar que essa manipulação de dados pode ser alterada, na linha de código 2, para que se mostrem os resultados de outras áreas semânticas do *corpus* bastando, para isso, editar a especificação "Jogos e diversões infantis" por uma das outras 13 áreas semânticas que formam o QSL do Projeto ALiB.

4 Resultados preliminares

A experiência obtida por meio da construção do banco de dados aqui focalizado, cuja alimentação continua em andamento, e com os resultados alcançados em cada extração de dados, dá mostras da importância do planejamento da arquitetura dos elementos e dos atributos em um documento *XML*, assim como a definição das regras dentro do *DTD* de maneira adequada aos propósitos do estudo. Nesse sentido, o linguista precisa ter em mente, antes de iniciar a escrita de um documento *XML*, o que exatamente pretende obter com a investigação. Ou seja, listar as tarefas que deverão ser automatizadas segundo os propósitos da pesquisa para, num segundo momento, estudar como proceder no campo da Informática para que as ações se concretizem.

A extração de dados a partir de áreas semânticas, por exemplo, só foi possível porque a organização dos dados foi pensada, previamente, para atender essa demanda. Isso significa que, uma vez iniciado o processo de alimentação do banco de dados, editar sua estrutura e as diretrizes do *DTD* para adicionar uma nova funcionalidade significa investir mais tempo para a reescrita do arquivo XML, já que esse processo é manual.

Vale observar que o procedimento de construção do banco de dados é lento e manual, ou seja, não é possível automatizar essa etapa. Desse modo, é preciso prever todas as funções que deverão ser executadas a partir do banco de dados e testá-las num protótipo para, só então, iniciar a alimentação dos dados em definitivo.

5 Considerações finais

O *BaseX* é um *software* que permite a visualização dos dados de diferentes formas. Nesse sentido, ainda não foi possível explorar e compreender todas as funcionalidades oferecidas pelo programa. Porém, o que mais importa no momento é que foi possível realizar extrações de informações relevantes à pesquisa satisfazendo,

por hora, os objetivos do estudo. No entanto, sabe-se que é possível habilitar a porta *localhost:8984* para realizar simulações em um navegador de internet, isto é, utilizar os dados em *XML* para desenvolver uma aplicação web, tarefa que se buscará compreender e realizar na próxima etapa da pesquisa.

Os benefícios da digitalização dos dados do Projeto ALiB em formato *XML* vão além das extrações de informações pertinentes à pesquisa científica, que subsidia reflexões e discussões acerca dos fenômenos linguísticos presentes na fala dos habitantes da região Norte do Brasil. Dessa maneira, uma vez completo, os dados poderão ser compartilhados com outras bases de dados e utilizados para serem processados por outros softwares e/ou linguagens de programação, subsidiando, dessa forma, outros projetos. Vale destacar que essa "[...] relação entre as práticas tradicionais de registro do conhecimento e as novas tecnologias é a marca indelével do movimento das Humanidades Digitais" (HIGUICHI; FREITAS, 2017, p. 1).

O uso das expressões *X-Query* são fundamentais para a extração de informações linguísticas no software *BaseX*. Além do mais, essas expressões e outras linguagens de programação serão requeridas para a construção da aplicação web, que tem como objetivo a produção e publicação, como produto final, do *Vocabulário Dialetal da região Norte do Brasil.*

Referências

BIDERMAN, M. T. C. **Teoria linguística:** Teoria lexical e linguística computacional. 2ª ed. São Paulo: Martins Fontes, 2001.

CARDOSO, S. A. **Geolinguística: tradição e modernidade.** São Paulo: São Paulo, 2010.

CHAMBERS, J.; TRUDGILL, P. **La dialectología**. Madrid: Visor Libros, S. L., 1994. p. 35-61.

COMITÊ NACIONAL DO PROJETO ALIB. **Atlas Lingüístico do Brasil**: questionário 2001. Londrina: EDUEL, 2001.

COSERIU, E. **Lições de Linguística Geral**; tradução do Prof. Evanildo Bechara. Rio de Janeiro: Ao Livro Técnico, 1980.

COSTA, D. de S. S. **Vocabulário Dialetal do Centro-Oeste**: interfaces entre a Lexicografia e a Dialetologia. 2018. 353 f. Tese (Doutorado em Estudos da Linguagem) – Universidade Estadual de Londrina, Londrina/PR, 2018.

CORREIA DE SOUZA, C. **Vocabulário Dialetal da região Norte do Brasil:** um estudo das capitais com base nos dados do Projeto ALiB. 2019. 134 f. Dissertação (Mestrado em Língua e Cultura) - Universidade Federal da Bahia, 2019.

EZQUERRA, M. A. Lexicografía dialectal. **ELUA**, Estudios de Lingüística, [*S.l.*] nº 11, p.79-109, (1996-1997). Disponível em: https://scholar.google.es/citations?user=mEEtglQAAAAJ&hl=es. Acessso em: 23 nov. 2020. DOI https://doi.org/10.14198/ELUA1996-1997.11.03

GRÜN, C. **BaseX.** Versão 9.4.3, [*S.l.*], 2020. Software de computador. Disponível em: https://basex.org/. Acesso em: 23 set. 2021.

HABERT, B. Portrait de linguiste(s) à l'instrument. **Texto!** [*S.l.*], vol. X, n°4, 2005. Disponível em: http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html. Acesso em: 14 dez. 2020.

HARTMANN, R. R. K. Estructural and typological perspectives. *In*: **Teaching and Researching Lexicograph**. New York: Routledge, 2016, p. 57-65. Disponível em: https://books.google.com.br/books?id=duzeCwAAQBAJ&pg=PA59&hl=pt-BR&source=gbs_selected_pages#v=onepage&q&f=false. Acesso em: 30 set. 2019.

HAUSSER, R. **Fondations of Computacional Linguistics:** Human-Computer Communication in Natural Language. 3. ed. Heidelberg: Springer, 2014. DOI https://doi.org/10.1007/978-3-642-41431-2

HIGUCHI, S.; FREITAS, C. Linguística computacional, humanidades digitais e os desafios na mineração de um dicionário histórico-biográfico. In: X Congresso Internacional da Abralin, Niterói, 2017. **Anais**. X Congresso Internacional da Abralin, 2017. Disponível em: https://bibliotecadigital.fgv.br/dspace/handle/10438/29142. Acesso em: 13 mar. 2022.

KURDI, M. Za. **Natural Language Processing and Computacional Linguistics 1:** Speech, Morphology and Syntax. London: ISTE, 2016. DOI https://doi.org/10.1002/9781119145554

MACHADO FILHO, A. V. L. Um ponto de interseção para a dialectologia e a lexicografia: a proposição de um dicionário dialetal brasileiro com base nos dados do ALiB. **Estudos** (UFBA), v. 41, p. 49-70, 2010.

NAVARRO CARRASCO, A. I. Geografía lingüística y diccionarios. **ELUA,** Estudios de Lingüística. [*S.l.*], nº 9, p. 73-96, 1993. Disponível em: http://rua.ua.es/dspace/handle/10045/6467. Acesso em: 23 nov. 2020. DOI https://doi.org/10.14198/ELUA1993.9.05

NEIVA, I. **Vocabulário Dialetal Baiano.** 2017. v. 1, 270 f. Tese (Doutorado em Língua e Cultura). Universidade Federal da Bahia, Salvador/BA, 2017.

NUNES, M. das G. V.; ALUÍSIO, S. M.; PARDO, T. A. S. Um panorama do Núcleo Interinstitucional de Linguística Computacional às vésperas de sua maioridade. **Linguamática**, v. 2, n. 2, p. 13-27, 29 mai. 2010. Disponível em: https://www.linguamatica.com/index.php/linguamatica/article/view/66/75> Acesso em: 13 mar. 2022.

PÉREZ HERNÁNDEZ, C.; MORENO ORTIZ, A. **Lingüística computacional y lingüística de corpus.** Potencialidades para la investigación textual. 2009. Disponível em: http://tecnolengua.uma.es/doc2/trea2009.pdf. Acesso em: 16 jan. 2021.

PESTOV, S. *et al.* **jEdit.** Versão 5.4.0. [*S.I.*], [2017?]. Software de computador. Disponível em: https://sourceforge.net/projects/jedit/files/jedit/5.4.0/. Acesso em: 06 set. 2020.

PORTO DAPENA, J.-Á. **Manual de técnica lexicográfica**. Madrid: ARCO/LIBROS, S.A., 2002.

SÁ, E. J. de. Variação lexical no falar amazonense: um estudo dialetal e metalexicográfico das denominações para riacho/córrego. **Entrepalavras**, [S.l.], v. 11, n. 10esp, p. 213-226, jun. 2021. ISSN 2237-6321. Disponível em: http://www.entrepalavras.ufc.br/revista/index.php/Revista/article/view/2088. Acesso em: 14 fev. 2022. DOI https://doi.org/10.22168/2237-6321-10esp2088

Artigo recebido em: 30.09.2021 Artigo aprovado em: 10.03.2022

DOI: 10.14393/DL52-v16n4a2022-12

Reflexões metodológicas sobre *datasets* e Linguística de *Corpus*: uma análise preliminar de dados legislativos

Methodological reflections on datasets and Corpus Linguistics: a preliminary analysis of legislative data

Lúcia de Almeida FERRARI[*]
Evandro Landulfo Teixeira Paradela CUNHA[**]

RESUMO: Ferramentas e métodos computacionais são, cada vez mais, importantes aliados para a realização de pesquisas no âmbito das humanidades. Em particular, o uso dessas ferramentas é relevante para a análise linguística diacrônica. Neste estudo, é apresentada uma discussão sobre o uso de *corpora* e *datasets* na linguística, destacando algumas potencialidades e limitações desses recursos. Para ilustrar as possibilidades de uso de um *dataset* para pesquisa linguística, apresenta-se, também, uma análise preliminar da Base de Normas Jurídicas Brasileiras.

PALAVRAS-CHAVE: Processamento de texto. *Dataset* de normas jurídicas. Análise diacrônica. Linguagem e direito.

ABSTRACT: Computational tools and methods are increasingly important for conducting research in the humanities. In particular, these tools are relevant for diachronic linguistic analysis. In this study, we present a discussion about the use of corpora and datasets in linguistics, highlighting some strengths and limitations of these resources. To illustrate the possibilities of using a dataset for linguistic research, a preliminary study employing a dataset of Brazilian legal norms is also presented.

KEYWORDS: Text processing. Legal norms dataset. Diachronic analysis. Language and law.

[*] Doutora em Estudos Linguísticos pela Universidade Federal de Minas Gerais (UFMG). Professora na Faculdade de Letras da UFMG. ORCID: https://orcid.org/0000-0002-9855-0646. ferrari.lu@gmail.com.
[**] Doutor em Linguística pela Universitaet Leiden e em Ciência da Computação pela Universidade Federal de Minas Gerais (UFMG). Professor na Faculdade de Letras da UFMG. ORCID: https://orcid.org/0000-0002-5302-2946. cunhae@ufmg.br.

1 Introdução

A relação entre linguagem e direito é profunda: o direito, enquanto regulamentação do comportamento humano, se manifesta por meio da língua, oral ou escrita; ao mesmo tempo, é por meio da linguagem que se discutem as normas de convivência na sociedade civil[1]. No passado distante, os usos e costumes aceitos e compartilhados por uma sociedade eram transmitidos oralmente. No entanto, já na antiguidade surgiu a demanda pela codificação escrita das noções primordiais dos direitos e deveres dos indivíduos. Prescrições sobre o comportamento e a moral, escritas em tábuas de pedra, são aquelas do decálogo bíblico (Êxodo, 20:1-17; 34:1-5). Historicamente, é à Lei das XII Tábuas da República Romana (*Lex Duodecim Tabularum*, 451 a.C.) que se deve a primeira elaboração articulada de uma legislação escrita à disposição de toda a população, para que não fosse manipulada, e que ainda hoje serve de inspiração, em muitos países, para o Direito Público e o Direito Civil. O sistema legislativo de um Estado é, portanto, uma expressão tangível de fenômenos sociais distintos (direito e língua), mas intimamente ligados entre si (MACIEL, 2001, p. 56).

Uma pesquisa bibliográfica sobre o estudo da linguagem jurídica[2] no Brasil aponta para duas grandes áreas: uma ligada ao mundo do Direito propriamente dito e outra, ainda em expansão, mais próxima aos estudos da linguagem. Além de dicionários de terminologia da área (DINIZ, 1998; SANTOS, 2001; GUIMARÃES, 2013; entre outros), há livros didáticos para o ensino de redação e argumentação em textos jurídicos, direcionados especialmente a estudantes de Direito, nos quais a referência à língua é de tipo instrumental ou sobre seu uso retórico (p. ex., DAMIÃO; HENRIQUES, 2020; AQUINO; DOUGLAS, 2017; PETRI, 2017). Volumes mais

[1] Para uma discussão sobre direito e linguagem, ver Warat (1995) e Ivo (2020).

[2] Empregamos, aqui, o termo "linguagem jurídica" por ser ele o mais utilizado e por possuir um significado mais amplo, apesar de concordarmos com Ramos (2017, *apud* SVOBODOVÁ, 2017) sobre o fato de que a linguagem empregada na área do Direito possa ser dividida em dois tipos: "linguagem legal", utilizada nas normas jurídicas, e "linguagem forense", utilizada no foro, ou seja, em tribunal.

específicos abordam as dificuldades encontradas por possíveis redatores de normas ou atos administrativos sobre o uso do vernáculo, em oposição à terminologia e usos da linguagem do domínio jurídico, não somente do ponto de vista da correção, mas também da compreensão por parte de não especialistas[3]. Reflexões epistemológicas ou semióticas, assim como ponderações linguístico-pragmáticas, são encontradas em obras de juristas que discorrem sobre lógica ou ontologia jurídica (p. ex., WARAT, 1995; BITTAR, 2009; IVO, 2020).

Os trabalhos de cunho linguístico em interface com a área jurídica podem estar, por exemplo, no âmbito da análise do discurso jurídico, que tem tradição no Brasil sobretudo graças ao Grupo de Pesquisa Linguagem e Direito e à Associação de Linguagem & Direito (ALIDI). O projeto TermiSul[4], por sua vez, se ocupa de estudos terminológicos e tradutórios específicos em diversas áreas do conhecimento. O grupo tem publicado, ao longo dos anos, desde dicionários e glossários multilíngues de gestão e direito ambiental (KRIEGER *et al.*, 1998, 2006, 2008) até estudos de léxico especializado da linguagem legal. A partir do início dos anos 2000, o grupo se vale de metodologias de linguística de *corpus* em suas pesquisas. O Projeto CoMET – Corpus Multilíngue para Ensino e Tradução[5] compila e disponibiliza *corpora* técnico-científicos desde 2003. Entre os trabalhos da equipe, se destacam o CorTec – Corpus Técnico-

[3] "Juridiquês" é o termo pejorativo que indica o uso de uma linguagem jurídica especialmente arcaica e repleta de latinismos, expressões consideradas pedantes e desnecessariamente rebuscadas e de raciocínio intricado. A Campanha pela Simplificação da Linguagem Jurídica (https://www.amb.com.br/campanha-pela-simplificacao-da-linguagem-juridica-sera-lancada-as-11-horas/. Acesso em: 20 dez. 2021) tem como objetivo "incentivar os magistrados e operadores do Direito a simplificar a linguagem jurídica, valorizando o uso de um vocabulário mais objetivo, simples e direto para aproximar os cidadãos do Poder Judiciário brasileiro". Os termos técnicos deveriam ser mantidos, mas utilizados de forma adequada e destinados a facilitar a compreensão pelo cidadão comum. As ações vão desde campanhas por meio de concursos para estudantes e magistrados promovidas pela Associação de Magistrados Brasileiros (AMB) a projetos de lei para a elaboração de sentenças em linguagem simples (ver, a esse respeito: https://www12.senado.leg.br/noticias/materias/2012/06/27/guerra-contra-o-2018juridiques2019-pode-levar-a-mudancas-em-projetos-de-lei. Acesso em: 18 dez. 2021).

[4] http://www.ufrgs.br/termisul/. Acesso em: 20 dez. 2021.

[5] https://comet.fflch.usp.br/. Acesso em: 20 dez. 2021.

Científico[6], que contempla uma seção de textos da área do Direito sobre instrumentos contratuais, além de artigos e publicações sobre linguagem jurídica voltados ao público de profissionais, como tradutores, advogados e especialistas (p. ex., CARVALHO, 2006). Além disso, nos últimos anos, pesquisadores brasileiros vêm se inserindo no panorama mundial dos eventos e discussões que reúnem legislação, linguagem e computação: um exemplo é a 18th International Conference on Artificial Intelligence and Law (ICAIL 2021)[7], organizada de forma virtual em São Paulo e que possibilitou grande interação entre estudiosos, analistas, juristas, linguistas e programadores do Brasil e do mundo no debate sobre as possibilidades de análises automáticas e semiautomáticas de textos legislativos e jurídicos.

A proposta do presente artigo é realizar um estudo linguístico exploratório de um conjunto de leis brasileiras, tecendo reflexões de caráter metodológico sobre a utilização desse tipo de dado sob o viés da linguística de *corpus* e sobre a diferença entre a utilização de *corpora* e de *datasets*. O intuito do artigo é discutir as possibilidades investigativas dos conjuntos de leis e realizar uma primeira análise de tipo diacrônica. Os dados utilizados são formados por um *dataset* contendo normas jurídicas federais coletado e disponibilizado por Martim, Lima e Araujo (2018), que foi subdividido em períodos e explorado utilizando *scripts* desenvolvidos em linguagem de programação Python.

O artigo está assim dividido. Na primeira parte, se discute a diferença entre *corpora* e *datasets*. Em seguida, antes de introduzir o *dataset* utilizado no presente trabalho, expomos brevemente como ocorre o processo de elaboração de leis no Brasil. A seção seguinte apresenta as etapas e resultados da análise exploratória. As conclusões discutem as observações efetuadas, inclusive de cunho metodológico, e possibilidades de pesquisas futuras.

[6] https://cortec.fflch.usp.br/. Acesso em: 20 dez. 2021.
[7] https://icail.lawgorithm.com.br/. Acesso em: 21 dez. 2021.

2 *Datasets* e *corpora*

Nesta seção, são apresentados alguns conceitos fundamentais relacionados ao que, em linguística, recebe o nome de *corpus* (plural *corpora*) e sua diferença ao que se pode chamar de *datasets*, ponderando sobre as diferentes metodologias e resultados a que esses diferentes tipos de dados podem levar no contexto dos estudos linguísticos.

Datasets, ou conjuntos de dados, podem ser definidos como coleções de dados agrupados por possuírem características em comum[8]. Frequentemente, esses dados são tabulados, com colunas correspondendo às diferentes variáveis, mas podem ser organizados também em forma de diretórios com arquivos do mesmo tipo e, por vezes, até mesmo contendo interfaces gráficas para a visualização dos dados. Tais dados, geralmente empíricos, podem ser de vários tipos: desde observações espaciais a informações meteorológicas, desde prontuários médicos a informações sobre flutuações monetárias, e são a base para investigações de cientistas e analistas de diversas áreas. No caso de *datasets* para estudos linguísticos e/ou processamento de língua/linguagem natural (PLN), seus conteúdos são igualmente diversificados – como, por exemplo, áudios de fala, arquivos de texto, tabulações com dados linguísticos e/ou extralinguísticos diversos etc. A coleta desses dados pode ser realizada com recursos muito diferentes, como gravações de conversas inteiras ou de trechos curtos (palavras ou frases) e textos dos mais variados gêneros (que podem provir, por exemplo, de páginas da internet ou da digitalização de obras de bibliotecas). Geralmente, grandes *datasets* são a base de aplicações em PLN, as quais têm necessitado de extensos volumes de dados para pesquisas ou para o desenvolvimento de produtos como filtros de spam para e-mails, assistentes virtuais inteligentes, corretores automáticos, tradutores automáticos, entre outros.

Datasets não devem ser confundidos com *databases*, ou bancos de dados, que

[8] Para uma discussão mais ampla sobre como o termo *dataset* é utilizado na literatura científica e técnica, ver Renear, Sacchi e Wickett (2010).

armazenam dados usando uma estrutura mais rígida. O termo remete às fichas catalográficas, por muito tempo utilizadas pelos linguistas, mas também, e especialmente, à importância dada à estrutura em que são arquivados e sucessivamente apresentados os dados (DIMITRIADIS; MUSGRAVE, 2009). Hoje em dia, os próprios catálogos dos sistemas de bibliotecas se comunicam com bancos de dados, na busca por obras e autores, por meio digital. O mesmo ocorre com os bancos de dados linguísticos, que frequentemente utilizam um DBMS (*database management system*). Os dados e os motores de busca ficam armazenados em servidores que podem ser acessados, por meio da internet, pelo usuário final, sem a necessidade de se sobrecarregar o equipamento de cada pesquisador ou organização. Muitas bases de dados não possuem uma interface gráfica amigável para usuários não programadores (ou GUI, *graphical user interface*), mas podem ser acessadas por meio de consultas utilizando linguagens de pesquisa específicas (como a SQL, *Structured Query Language*) ou de aplicativos instalados no computador do usuário que permitem tais buscas. Alguns bancos de dados funcionam com uma interface própria e permitem uma série de pesquisas pela própria página web do serviço[9].

Corpora também são considerados conjuntos de dados, mas, no contexto da linguística de *corpus*, sua arquitetura se relaciona a três conceitos básicos: representatividade, amostragem e balanceamento. Um *corpus* deve ser representativo

[9] Bancos de dados linguísticos para PLN são frequentemente organizados e utilizados por grandes empresas e seu acesso, quando possível, costuma ser pago. Para o português brasileiro, os bancos de dados disponíveis gratuitamente entram na categoria dos *corpora*, pois são compilados por grupos de pesquisa ligados a universidades e seguem uma arquitetura precisa. O portal https://www.linguateca.pt/ (acesso em: 10 dez. 2021) é um site que tem por objetivo catalogar, reunir e disponibilizar (no próprio site ou por meio de links) o maior número possível de recursos disponíveis ao público para o processamento computacional do português. O projeto teve seu financiamento suspenso em 2011 e a própria equipe manteve a página até a última atualização, em 2015. Apesar de não haver nele informações mais recentes, trata-se de um ponto de partida valioso pela riqueza de informações e links para pesquisadores do português em geral e do português brasileiro em específico, indicando e disponibilizando desde *corpora*, ontologias, dicionários, até recursos de buscas e ferramentas, além de informações sobre numerosos projetos e pesquisas.

de uma língua ou variedade linguística: isso envolve conhecimentos sobre a população a ser estudada e decisões sobre o tipo de amostragem estatística que será utilizada, ponto de partida para a coleta dos dados linguísticos. A noção de balanceamento se refere ao tamanho de cada amostra e seu peso relativo com relação ao conjunto dos dados do *corpus*. Outro conceito basilar na arquitetura de um *corpus* é sua finalidade, ou seja, o porquê de sua compilação. Um *corpus* de referência é pensado para ser representativo de dada língua ou variedade em sua totalidade: caso seja escrito, levará em conta as diferentes tipologias textuais e sua distribuição no panorama geral das publicações; caso seja oral, serão as variáveis sociolinguísticas que servirão como ponto de partida para a amostragem da população. *Corpora* de referência são pensados como projetos de muitos anos, que demandam equipes treinadas e coesas para sua compilação, além dos financiamentos necessários para sua implementação. Tais *corpora* permitem pesquisas além daquelas para as quais foram inicialmente planejados e são geralmente disponibilizados à comunidade acadêmica, de forma gratuita ou por meio de acordos entre instituições.

Um *corpus* pode, também, ser compilado para uma pesquisa relativamente restrita e pequena, como em trabalhos de mestrado e doutoramento. *Corpora* para finalidades específicas tendem a ser menores em tamanho, pela demora no trabalho de coleta e preparação dos dados, que muitas vezes recai sobre somente uma pessoa, e nem sempre são adequados ao tipo de pesquisa que se pretende realizar. Muitas vezes, após a conclusão do trabalho, esses *corpora* são abandonados por seus autores, perdendo-se todo o trabalho de compilação. É frequente, portanto, que estudiosos utilizem *corpora* já prontos em suas pesquisas, extraindo amostras ou adaptando-os a suas necessidades.

A utilização de *corpora* subdimensionados ou cuja arquitetura não foi devidamente pensada pode se tornar um problema, pois cada tipo de pesquisa linguística requer uma amostragem não somente da população a ser estudada, mas

também dos textos, além da definição do tamanho de cada um deles – ou seja, seu balanceamento[10]. É de se sublinhar que conseguir atingir a devida representatividade de um *corpus* não é somente uma questão de aumentar seu tamanho ou o tamanho de suas amostras: é importante, ainda, que a arquitetura se preocupe em prever as características necessárias e adequadas à pesquisa que se pretende empreender.

Biber (1988, 1990, 1992, 1993) analisa a distribuição de traços linguísticos[11] frequentes em amostras de registros escritos diferentes. Sua conclusão é de que há uma grande diversificação nos itens linguísticos encontrados à medida que há variação de textos, e estabilização dessa variação quando os textos são muito longos e pouco variados. Componentes linguísticos comuns (como nomes e preposições) apresentam uma distribuição estável entre textos diferentes e podem ser encontrados em amostras relativamente pequenas (com cerca de 2000 palavras). Componentes linguísticos menos frequentes (como verbos conjugados no passado ou no futuro) necessitam de amostras maiores pois sua distribuição varia muito mais dentro dos mesmos textos.

Estudos em *corpora* sobre dispersão lexical (BIBER *et al.*, 2016; EGBERT; BURCH; BIBER, 2020) concluem que subdividir dados de *corpora* em trechos aleatórios de mesmo tamanho para análises pode gerar resultados distorcidos, pois perde-se o contexto específico em que os itens comparecem. Novamente, essas pesquisas apontam para a necessidade de muita cautela na amostragem dos dados. A dispersão lexical resulta ser muito maior pelo efeito do contexto (e dos próprios textos) quando são extraídos de um *corpus* trechos de forma arbitrária do que quando esses fazem parte de agrupamentos linguisticamente relevantes. Em outras palavras, a amostragem deve se preocupar com as tipologias textuais, enquanto a taxa de

[10] Egbert, Larsson e Biber (2020) discutem e exemplificam em detalhes tais problemas. Aqui, nos limitaremos a citar alguns pontos que reputamos importantes sem nos adentrarmos na ampla discussão metodológica dos autores.

[11] Biber (1993) utiliza o termo *features*, traduzido em português como *traços* por Resende e Maverick (2016), para indicar os elementos analisáveis e quantificáveis nas investigações em *corpora*.

dispersão lexical varia muito de texto para texto, sendo necessário levar em conta seus contextos específicos. Análises desse tipo devem se atentar, portanto, a uma cuidadosa seleção dos registros que comporão o *corpus*.

Em *corpora* orais, a representatividade dos textos é obtida geralmente por meio da seleção dos informantes. A tradição sociolinguística (LABOV, 1966, 1972a; BAKER, 2010) fornece domínios específicos para as interações formais: há, portanto, situações como a entrevista estruturada ou semiestruturada, mas também palestras, contextos de sala de aula, interações profissionais, religiosas, políticas, entre outras. Já no caso de interações espontâneas, o domínio é totalmente aberto e a coleta de gravações com ampla variação diafásica tem se mostrado muito produtiva para a obtenção de dados linguísticos extremamente variados e com excelente balanceamento entre sexo e idade, mantendo-se um grau de instrução médio-alto (RASO, 2012; MELLO, 2014). Partir de situações comunicativas heterogêneas e totalmente espontâneas permite: (a) atingir diferenças de natureza diastrática, pois variam muito os informantes; e (b) determinar variações de natureza informacional e ilocucionária, ou seja, atos linguísticos diferentes (SEARLE, 1969), enquanto situações diversas acarretam estruturas organizacionais variadas entre os textos e, como consequência, léxico, morfologia e sintaxe diversificados.

A disponibilização de *corpora* compilados seja por projetos maiores, seja por iniciativas individuais é outro ponto relevante para a pesquisa empírica. Um *corpus*, especialmente de fala ou multimodal, mas mesmo escrito, é um recurso informático que ocupa um espaço virtual grande em termos de armazenamento. Seu compartilhamento, até entre membros da própria equipe, requer que os dados sejam carregados em algum site ou "em nuvem". Não é raro que tais dados sejam retirados do ar após a finalização do projeto ou quando os financiamentos são suspensos, pois manter espaço de armazenamento e páginas na web pode ser relativamente dispendioso.

Em vista de tais problemas, o CQPweb[12] tem se tornado um recurso interessante para vários projetos: trata-se de uma interface gráfica online de elementos de processamento de pesquisa, especificamente do CQP (*Corpus Query Processor*). Sua base é o CWB (*Corpus Workbench*), uma coletânea de ferramentas de código aberto para o processamento e consultas de grandes *corpora* com anotações linguísticas. O Linguateca AC/DC[13] utiliza a interface do CQPweb para a publicação e consultas a seus recursos, que incluem *corpora* do português (inclusive português brasileiro), assim como o CLUL (Centro de Linguística da Universidade de Lisboa)[14], que disponibiliza *corpora* abarcando variedades africanas e asiáticas do português.

A diferença entre se disponibilizar um *dataset* ou *corpus* (com e/ou sem anotações) em uma plataforma para que seja baixado pelo usuário ou em outra que possibilite consultas online se relaciona com a capacidade de processamento e com as competências computacionais do pesquisador[15]. Dados volumosos requerem competências técnicas nem sempre viáveis para todos os linguistas. Como afirma Hardie (2012),

> Alguns pesquisadores em linguística de corpus com maior conhecimento técnico recomendam uma abordagem 'faça você mesmo' para ferramentas de análise de *corpora*, em que, em vez de utilizar um concordanciador pronto, o pesquisador escreve seus próprios programas de computador para processar seus dados. [...] Claramente, essa abordagem 'faça você mesmo' para softwares de análise de *corpora* é a mais poderosa imaginável se a capacidade [de processamento] for equiparada à adaptabilidade – embora os programas 'faça você mesmo' possam ser executados lentamente caso não sejam

[12] Para as especificações da plataforma e uma descrição detalhada de seus recursos, ver Hardie (2012).
[13] Disponível em: https://www.linguateca.pt/ACDC/. Acesso em: 18 dez. 2021.
[14] As buscas nos *corpora* disponibilizados pelo CLUL podem ser realizadas na página: http://gamma.clul.ul.pt/CQPweb/. Acesso em: 5 mar. 2022.
[15] A título de exemplo sobre tais possibilidades, citamos o projeto C-ORAL-BRASIL, que disponibiliza seus *corpora* de fala para download no site http://www.c-oral-brasil.org/ (acesso em: 15 dez. 2021), mas também possui uma base de dados própria, o DB-CoM (*Database for Corpora Multimidia*), que possibilita a consulta online a alguns de seus mini *corpora* etiquetados informacionalmente em http://www.c-oral-brasil.org/db-com (acesso em: 15 dez. 2021).

> acrescentados os 'truques', como indexação, necessários para a alta velocidade em grandes conjuntos de dados. No entanto, essa abordagem é insuficiente em termos de usabilidade para a maioria dos linguistas de *corpus* em potencial, que não podem ou não desejam aprender programação de computadores [...]. Para esses pesquisadores, a usabilidade sempre será mais relevante que a capacidade de processamento" (tradução nossa) (HARDIE, 2012, p. 383)[16].

Ferramentas gratuitas como o AntConc (disponível em https://www.laurenceanthony.net/software/antconc/, acesso em: 10 dez. 2021) conseguem processar *corpora* de tamanhos médios de forma geralmente satisfatória e possibilitam análises cada vez mais sofisticadas. Todavia, quando os dados provêm de *corpora* ou de *datasets* de tamanho muito grande, tais programas podem não suportar o volume de dados. Nesses casos, acaba por se fazer necessária a participação de alguém que saiba produzir seus próprios códigos. Essa tarefa é ainda mais importante quando se pensa que *datasets* não apresentam necessariamente uma arquitetura como, em teoria, os *corpora*, e devem, por sua vez, ser ainda mais cuidadosamente amostrados e selecionados. O trabalho aqui apresentado mostrará, justamente, parte da metodologia e das dificuldades encontradas no processamento de dados previamente coletados e disponibilizados no formato de *dataset*.

Passaremos agora a um breve panorama sobre o processo legislativo brasileiro para, em seguida, seguirmos à apresentação dos dados inicialmente investigados neste artigo e a suas análises.

---

[16] [s]ome technically sophisticated corpus researchers recommend a "do-it-yourself" approach to corpus analysis tools, where rather than exploiting an off-the shelf concordancer, the analyst instead writes their own computer programs to process their data. [...] Clearly this "do-it-yourself" approach to corpus analysis software is the most powerful imaginable if power is equated to maximum scope for adaptability – although "do-it-yourself" programs may run slowly if they do not incorporate the "tricks", such as indexing, needed for high speed on large datasets. However, this approach falls short in terms of usability for the majority of potential corpus analysts, who either cannot or do not wish to learn computer programming [...]. For such researchers, usability will always be more critical than power.

3 O processo legislativo no Brasil e o *dataset* de normas jurídicas brasileiras

O Brasil é uma república federativa presidencialista: a União (governo federal), os 26 estados e o Distrito Federal, além de seus mais de 5000 municípios, formam um estado democrático com divisão entre três poderes. No âmbito federal, o Poder Legislativo é exercido pelo Congresso Nacional, o Executivo pela Presidência da República e o Poder Judiciário pelos Supremo Tribunal Federal, Conselho Nacional de Justiça, Superior Tribunal de Justiça, Tribunal Superior do Trabalho, tribunais regionais federais, tribunais eleitorais, tribunais militares, tribunais de justiça dos estados e do Distrito Federal e territórios.

A Constituição Federal é a lei superior que orienta sobre as competências de cada poder no nível federal, detalhando sobre quais são seus campos de atuação e determinando que um poder fiscalize o outro. O Congresso Nacional, por exemplo, além de legislar, anualmente deve julgar as contas prestadas pela Presidência da República e examinar os relatórios sobre a execução dos planos de governo. A Presidência da República organiza e atua no funcionamento da administração federal, além de se ocupar das relações com os outros países e nomear ministros ou vetar projetos de lei. Ao Poder Judiciário compete zelar pelo cumprimento da Constituição nas diferentes instâncias.

O processo legislativo brasileiro, a partir da Constituição Federal de 1988, prevê que o conjunto de atos para a produção de normas jurídicas (ou leis) passe pelas duas casas do Congresso Nacional: a Câmara dos Deputados e o Senado Federal. O processo bicameral das leis federais estabelece que um projeto de lei, que pode ser proposto por um deputado ou por um senador, após ser aprovado por uma comissão que o avaliará e pela casa proponente, deve passar pela outra casa – a casa revisora –, podendo ser ali modificado. Caso isso ocorra, o projeto retorna à casa proponente, que pode decidir por acatar ou não as mudanças no texto. O quórum necessário para a aprovação do projeto de lei depende do tipo de lei em questão.

Em seguida, o projeto de lei passa para a sanção da Presidência da República: em caso de desacordo, o projeto pode ser vetado em sua integralidade ou em partes. Todavia, caso deputados e senadores não estejam de acordo com o veto, eles podem invalidar, isto é, rejeitar, o veto da Presidência. Em caso de sanção, o projeto é promulgado e, a partir desse ato, se torna lei, dependendo, contudo, de publicação para que tenha validade.

As emendas à Constituição contam com regras próprias, seja em sua proposta, que prevê um quórum mínimo de 1/3 de deputados ou senadores, seja pelos turnos e quóruns necessários à sua aprovação. Há, contudo, algumas cláusulas da Constituição (as cláusulas pétreas) que não podem ser modificadas ou abolidas, sendo elas: o Estado federal; o voto como direito, secreto, universal e periódico; a separação dos poderes; e os direitos e garantias individuais.

O Poder Legislativo, por meio de suas casas, produz e coleta uma quantidade significativa de materiais, pois, além de todos os atos administrativos e de fiscalização, cada projeto de lei é registrado em suas sucessivas modificações até sua eventual sanção.

3.1 O *dataset* de normas jurídicas brasileiras de Martim, Lima e Araujo (2018)

O *Open Government Data* (OGD), em português Dados Governamentais Abertos (DGA), é uma iniciativa internacional que promove a transparência e responsabilidade ética das informações por parte dos governos (cf. https://publicadministration.un.org/en/ogd, acesso em: 10 dez. 2021). Os países que aderem ao DGA se empenham em promover iniciativas e políticas que possibilitem a disponibilização de seus dados a todos os cidadãos, de maneira a tornar as instituições públicas mais transparentes e direcionadas à colaboração democrática de seus habitantes.

Tal abertura se iniciou nos anos 2000 (GRAY, 2014), tendo o Brasil aderido ao

DGA em setembro de 2011, com uma série de ações políticas e técnicas para a publicação na web de dados oficiais, que culminaram no lançamento do Portal Brasileiro de Dados Abertos (https://dados.gov.br, acesso em: 20 dez. 2021). A Controladoria-Geral da União (CGU) se ocupa da gestão e monitoramento da Política de Dados Abertos, através da Infraestrutura Nacional de Dados Abertos (INDA), a qual atua por meio de padrões, tecnologias e orientações para a disseminação e o compartilhamento de dados e informações públicas.

Martim, Lima e Araujo (2018) coletaram e publicaram uma base de normas jurídicas federais acessível online[17]. Trata-se de um conjunto de oito *datasets* que contêm, em princípio, todas as normas legislativas federais desde 4 de outubro de 1946 até 12 de abril de 2017. Segundo os autores, o recorte metodológico selecionou "uma quantidade razoável de normas com maior probabilidade de estarem vigentes. Isso não exclui normas expressamente revogadas, pois o histórico dos textos é importante para a obtenção do texto vigente em uma determinada data" (MARTIM; LIMA; ARAUJO, 2018, p. 137). Os autores do *dataset* optaram por não incluir, na captura de dados, os Decretos, Decretos-Leis, Emendas Constitucionais e outros atos ou proposições normativas.

Os *datasets* que compõem a base, intitulada Base de Normas Jurídicas Brasileiras, estão assim divididos:

(a) *dataset* 1: contém os textos articulados das normas – para cada lei há um arquivo em formato .rtf com o conteúdo legislativo completo;

(b) *dataset* 2: contém as representações LexML dos textos articulados das normas – cada artigo é estruturado com o esquema XML em padrão LexML;

(c) *dataset* 3: nele foram processados em formato .txt os dados do *dataset* 2 e foram subdivididas as sentenças da epígrafe, ementa, preâmbulo, dispositivos e fecho das normas, de maneira a possibilitar pesquisas mais pontuais. Contudo, essa divisão

[17] Disponível em: https://doi.org/10.6084/m9.figshare.c.4029253.v1. Acesso em: 12 nov. 2021.

acabou gerando uma série de sentenças incompletas, fazendo-se necessária a criação do *dataset* 4;

(d) *dataset* 4: é formado por um conjunto de arquivos contendo somente os dispositivos agregadores, ou seja, com acréscimo nas sentenças de incisos e alíneas, apresentados em .txt, como dispositivos distintos;

(e) *dataset* 5: é formado por arquivos em formato .txt de cada dispositivo, com anotação gramatical em formato CoNLL-U;

(f) *dataset* 6: contém, para cada dispositivo, arquivos em formato .json que foram submetidos a análise sintática e identificação de entidades pela API Google Natural Language Processing;

(g) *dataset* 7: são apresentadas, em formato .txt, as sentenças dos dispositivos das normas de maneira completa e, em arquivos separados, suas respectivas ementas;

(h) *dataset* 8: são apresentados os arquivos .json das normas e de suas respectivas ementas, também processados pela API Google Natural Language Processing.

Teixeira *et al.* (2019) utilizaram a Base de Normas Jurídicas Brasileiras para testar um processo de extração de definições utilizando o *dataset* 8. O processo de extração empregou filtros heurísticos, que utilizam um conjunto de regras diversas, para: (a) filtrar sentenças candidatas utilizando expressões em que comparece o verbo *ser* + artigo; (b) isolar as leis que não foram corretamente parseadas pelo esquema LexML; (c) filtrar as sentenças com o verbo *ser* etiquetadas morfossintaticamente; (d) retirar as leis autorizativas, as quais possuem estrutura VSO; (e) extrair do conjunto de dados as entidades nomeadas, ou seja, itens utilizados como designadores específicos como nomes próprios de pessoas, locais ou organizações, datas, valores monetários, entre outros.

O trabalho apresentado na próxima seção utiliza os dados do *dataset* 7 para realizar uma série de análises preliminares e avaliar possíveis caminhos de investigação para pesquisas linguísticas utilizando os dados em questão.

4 Exploração e análise da Base de Normas Jurídicas Brasileiras

Nesta seção, são apresentados resultados preliminares referentes à exploração da Base de Normas Jurídicas Brasileiras disponibilizada por Martim, Lima e Araujo (2018). As análises implementadas aqui utilizaram *scripts* desenvolvidos na linguagem de programação Python[18]. Os principais objetivos foram realizar uma caracterização geral dos dados a partir de uma perspectiva quantitativa e efetuar análises lexicais que levassem em consideração mudanças ao longo do tempo no *dataset* estudado, demonstrando algumas das possibilidades do uso de ferramentas de programação relativamente simples para análise diacrônica. Todas as explorações apresentadas aqui foram realizadas utilizando o *dataset* 7 da Base de Normas Jurídicas Brasileiras – isto é, aquele contendo os textos das normas legislativas completos (em teoria) e, em arquivos separados, suas respectivas ementas, todos em formato .txt, totalizando 25.944 arquivos.

Embora a Base de Normas Jurídicas Brasileiras tenha sido previamente processada pelos seus autores, verificou-se a necessidade de realizar uma etapa adicional de pré-processamento no *dataset* analisado para uma melhor compreensão dos dados. Nessa etapa, notaram-se dois erros que merecem destaque:

(1) no *dataset* original, a Lei 13.407/2016 está duplicada, constando seja no diretório 'LEI-2016-13407' (correto), seja no diretório 'LEI-2016-13047' (com inversão dos dígitos '0' e '4'). Isso acabou gerando conflito com a Lei 13.047/2014 (a 'verdadeira' lei de número 13.047). Para resolver essa questão, optou-se por excluir o diretório 'LEI-2016-13047' e, consequentemente, os dois arquivos que dele faziam parte ('LEI-2016-13047-dispositivos.txt' e 'LEI-2016-13047-ementa.txt');

(2) alguns arquivos estão em branco, inconsistentes ou incompletos. Isso foi constatado

[18] Para fins de reprodutibilidade e para oferecer suporte a pesquisas relacionadas, os *scripts* utilizados durante a realização deste trabalho estão disponíveis em https://github.com/evandrocunha/dados-legislativos.

apenas em arquivos referentes a dispositivos das normas, nunca naqueles referentes às ementas. Por exemplo, o arquivo 'LEI-1979-06765-dispositivos.txt' está em branco, apesar de a ementa correspondente ('LEI-1979-06765-ementa.txt') estar completa. Foram identificados sete arquivos em branco no *dataset*. Detectaram-se, ainda, arquivos contendo dispositivos de normas com um número muito baixo de palavras, o que nos pareceu atípico. Ao analisar os arquivos contendo menos de dez palavras, encontraram-se casos como os dos arquivos 'LEI-1951-01326-dispositivos.txt' (onde se lê apenas "O Quadro de Oficiais Farmacêuticos da Aeronáutica compor-se-á:"), 'LEI-1994-08883-dispositivos.txt' (onde se lê apenas "(VETADO). (VETADO).") e 'LEI-2004-11003-dispositivos.txt' (onde se lê apenas "2.2 - ....."). Nesses e em outros casos, o que consta nos arquivos não corresponde ao conteúdo real das respectivas normas[19]. Ocorrências similares foram identificadas em arquivos contendo entre dez e vinte palavras, mas, nesses casos, havia, também, normas contendo textos efetivamente curtos: é o caso, por exemplo, do arquivo 'LEI-2011-12449-dispositivos.txt', que contém apenas o seguinte texto: "O ator Paulo Autran é declarado Patrono do Teatro Brasileiro. Esta Lei entra em vigor na data de sua publicação.", que corresponde, de fato, ao conteúdo da Lei 12.449/2011[20]. Apesar dessas falhas nos dados, realizou-se a opção metodológica de manter, para a execução das análises aqui apresentadas, todos esses arquivos. Duas razões nos levaram a essa

[19] Para efeito de comparação, os textos integrais das normas mencionadas estão disponíveis em: Lei 1.326/1951: http://www.planalto.gov.br/ccivil_03/leis/1950-1969/l1326.htm; Lei 8.883/1994: http://www.planalto.gov.br/ccivil_03/leis/l8883.htm; Lei 11.003/2004: http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l11003.htm. Acesso em: 2 nov. 2021.

[20] Ainda assim, observa-se uma inconsistência na forma como os textos das normas são disponibilizados no *dataset*. Em alguns arquivos, está presente o fechamento do texto; em outros, não. Por exemplo: o arquivo 'LEI-1954-02158-dispositivos.txt' é finalizado com "Rio de Janeiro, em 2 de janeiro de 1954; 133º da Independência e 66º República. GEtúlio VARGAS João Goulart"; já o arquivo contendo os dispositivos da norma que declara o ator Paulo Autran Patrono do Teatro Brasileiro, conforme já indicado, não inclui os dizeres correspondentes (que seriam, nesse caso: "Brasília, 15 de julho de 2011; 190º da Independência e 123º da República. DILMA ROUSSEFF Anna Maria Buarque de Hollanda").

decisão: em primeiro lugar, as análises preliminares realizadas aqui não exigem, necessariamente, acesso ao conteúdo completo das normas; em segundo lugar, a maioria dos mais de vinte mil arquivos parecem estar satisfatórios – portanto, a tendência é que os poucos arquivos problemáticos se percam em meio à grande massa de dados.

A identificação desses erros no *dataset* se deu durante a etapa de caracterização quantitativa dos dados. Ainda que o impacto nos dados analisados não tenha sido grande (apenas dois arquivos foram removidos), isso demonstra a importância dessa etapa, frequentemente negligenciada. Para os estudiosos da linguagem, sua importância é ainda maior quando se sabe que os dados disponibilizados para a pesquisa não foram previamente preparados por uma equipe contendo um/a linguista, como é o caso da Base de Normas Jurídicas Brasileiras utilizada aqui.

Assim, apesar de Martim, Lima e Araujo (2018, p. 143) informarem que "[h]á 25.944 arquivos nesse dataset, sendo metade com arquivos das articulações das normas, metade com arquivos das ementas das normas", o número utilizado neste trabalho, após a remoção da norma duplicada, é 25.942 arquivos (sete deles em branco e muitos deles incompletos) – correspondendo, portanto, a 12.971 normas, cada uma com um arquivo referente à sua ementa e outro aos seus dispositivos.

Para a realização das análises a partir de uma perspectiva diacrônica, o *dataset* foi dividido em três períodos: I (de 04/10/1946 a 31/03/1964), II (de 01/04/1964 a 04/10/1988) e III (de 05/10/1988 a 12/04/2017). Essa periodização leva em consideração importantes mudanças no cenário legislativo brasileiro. O ano de 1946 inaugura o período I por ser a data de promulgação da quinta Constituição brasileira. O próprio *dataset* utilizado aqui tem início com a Lei 1/1946[21], de 04 de outubro daquele ano. O

[21] Conforme informado pelo Centro de Estudos Jurídicos da Secretaria-Geral da Presidência da República, "a partir de 04.10.1946, teve início a numeração de leis ordinárias que vigora até hoje" (informação disponível em: http://www4.planalto.gov.br/centrodeestudos/assuntos/legislacao/reflegis. Acesso em: 20 out. 2021).

ano de 1964, mais precisamente no dia 1 de abril, marca o fim desse período democrático, com a instauração da ditadura militar – ainda que a substituição da Constituição tenha ocorrido apenas em 1967. Finalmente, o período III tem início com a promulgação da Constituição de 1988, a chamada "Constituição Cidadã", no dia 5 de outubro, e vai até o final do período incluído no *dataset*, em 2017. Mesmo que a ditadura militar tenha terminado em 1985, considera-se que apenas com a nova Constituição os direitos sociais tenham sido, de fato, ampliados e universalizados. Adota-se aqui, portanto, a periodização empregada por autores como Santos (1997), cuja análise considera "[d]e um lado, [o período] que se estende de 1946 a 1964, regulado pela Constituição de 1946; de outro, o período atual, marcado pela Constituição de 1988". No caso do nosso estudo, acrescenta-se, ainda, o período intermediário (1964-1988), orientado, inicialmente (1964-1967), por normas emanadas fora do Estado Democrático de Direito; e, posteriormente (1967-1988), pela própria Constituição que institucionalizou e legalizou o regime militar.

O *dataset* analisado contém 6.212.514 palavras, sendo 371.526 nas ementas e 5.840.988 nos dispositivos[22]. Como informado anteriormente, é importante salientar que, em geral, softwares amplamente utilizados em linguística de *corpus* (como o AntConc, o WordSmith Tools, entre outros) podem não ser capazes de processar adequadamente quantidades de dados dessa magnitude, motivo pelo qual a programação de computadores pode ser útil para essas tarefas. Informações gerais sobre o *dataset*, inclusive considerando sua divisão em períodos, são mostradas na Tabela 1. É interessante observar como o valor da média de palavras nos dispositivos difere do valor da mediana para esse mesmo tipo de texto (450,3 *vs*. 120 no *dataset* completo; 246,2 *vs*. 107 no período I; 428,5 *vs*. 130 no período II; 624,4 *vs*. 131 no período III). Isso revela uma grande variabilidade na quantidade de palavras por texto, com

[22] Neste trabalho, o número de palavras contabilizado considera todos os *tokens* exclusivamente alfanuméricos.

certos textos contendo valores atípicos (*outliers*), isto é, apresentando grandes afastamentos dos demais valores presentes no conjunto de dados. O mesmo fenômeno não é observado nas ementas: nesses textos, as médias de palavras se aproximam das medianas, sugerindo uma maior uniformidade na distribuição dos valores ao longo do *dataset*.

Tabela 1 – Caracterização quantitativa do *dataset* 7 da Base de Normas Jurídicas Brasileiras após pré-processamento.

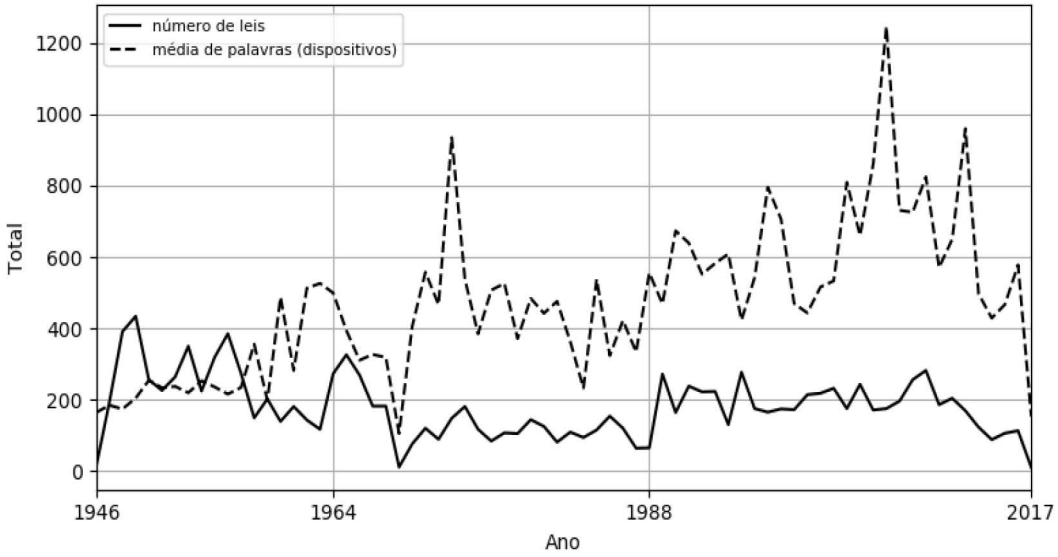
Período	Número de normas	Palavras			
		Tipo dos textos	Total de palavras	Média por texto	Mediana
Dataset completo (1946-2017) [Lei 1/1946 a 13.435/2017]	12.971	Ementas	371.526	28,6	24
		Dispositivos	5.840.988	450,3	120
Período I (1946-1964) [Lei 1/1946 a 4.320/1964]	4.260	Ementas	107.100	25,1	22
		Dispositivos	1.048.748	246,2	107
Período II (1964-1988) [Lei 4.321/1964 a 7.675/1988]	3.301	Ementas	76.838	23,3	21
		Dispositivos	1.414.444	428,5	130
Período III (1988-2017) [Lei 7.676/1988 a 13.435/2017]	5.410	Ementas	187.588	34,7	28
		Dispositivos	3.377.796	624,4	131

Fonte: elaborada pelos autores.

A Figura 1 complementa a Tabela 1 ao mostrar o número de leis e a média de palavras por lei (considerando apenas os dispositivos, não as ementas) em cada ano incluído no *dataset*. Observa-se uma variação natural nos valores; entretanto, é interessante notar alguns padrões. No período II, o número de leis por ano é consistentemente mais baixo que nos períodos I e III. Com relação à média de palavras por lei, convém observar que esse valor mostra uma tendência geral e gradual de crescimento desde 1970, pouco após a emissão do Ato Institucional Número Cinco (AI-

5), em 1968, e da Lei de Segurança Nacional de 1969, que marcaram o início do período que ficou conhecido como "anos de chumbo" na historiografia brasileira (CORDEIRO, 2009). Esse crescimento continua, também, no período III. Os números de leis nos extremos do dataset (anos de 1946 e 2017) são naturalmente mais baixos pois o *dataset* inclui apenas alguns meses desses anos.

Figura 1 – Número de leis e média de palavras por lei (dispositivo) em cada ano incluído no *dataset*.



Fonte: elaborada pelos autores.

O objetivo das análises apresentadas a seguir é comparar alguns aspectos linguísticos observáveis quantitativamente nas normas jurídicas brasileiras nos três períodos considerados. Por se tratar de uma análise preliminar, o *dataset* não passou por nenhum processo de anotação (por exemplo, anotação sintática), nem mesmo por um processo de lematização (convém recordar, no entanto, que versões da Base de Normas Jurídicas Brasileiras lematizadas e anotadas sintaticamente estão disponíveis nos *datasets* 5 e 6, não examinados neste estudo). Os únicos aspectos modificados nos textos foram a padronização ortográfica em caixa baixa e a remoção de pontuação.

4.1 Razão forma/item (type/token ratio)

A razão forma/item – ou, como é mais conhecida, razão *type*/*token* – indica a variabilidade, diversidade ou riqueza lexical do texto. Esse valor é obtido dividindo-se o número de palavras únicas (isto é, sem contar suas repetições – os *types*) pelo total absoluto de palavras (incluindo repetições – os *tokens*). Conforme explica Berber Sardinha (2004, p. 94), quanto maior for seu resultado, mais palavras diferentes (proporcionalmente) o texto conterá: uma razão forma/item próxima ao valor 1 indica um altíssimo grau de variabilidade lexical (pois o número de *types* se aproxima ao de *tokens*), ao passo que uma razão forma/item mais baixa sugere uma maior repetição dos itens lexicais ao longo do texto.

Aqui, comparou-se a razão forma/item entre os três períodos analisados. Os resultados obtidos estão indicados na Tabela 2. Observa-se uma tendência de diminuição na variabilidade lexical ao longo do tempo, passando de 0,0179 (1,79%) no período I a 0,0151 (1,51%) no período II e, posteriormente, a 0,0088 (0,88%) no período III.

Tabela 2 – Razão forma/item (*type/token ratio*) nos três períodos analisados no *dataset*.

Período	Razão forma/item
Período I (1946-1964)	0,0179
Período II (1964-1988)	0,0151
Período III (1988-2017)	0,0088

Fonte: elaborada pelos autores.

Alguns fatores podem explicar, ao menos parcialmente, essa tendência de diminuição. Em primeiro lugar, é preciso considerar o fenômeno conhecido como Lei de Herdan ou Lei de Heaps (HERDAN, 1964; HEAPS, 1978), segundo a qual a probabilidade de uma palavra nova (isto é, que não apareceu ainda) entrar em um

texto/*corpus* é menor à medida que o texto/*corpus* cresce. O motivo para isso é trivial: como o léxico é limitado, sempre que o texto/*corpus* cresce há menos palavras novas disponíveis para ingressar. Já que o *dataset* é menor (em número de palavras) no período I e, progressivamente, cresce nos períodos II e III, essa poderia ser uma explicação para o fenômeno.

Uma outra possível explicação para o fenômeno é que nas últimas décadas foram desenvolvidos manuais de padronização de textos oficiais, inclusive de normas jurídicas. Com isso, a tendência é que haja mais uniformização no léxico presente nesses textos, com uma maior repetição de estruturas fixas e uma maior reincidência de padrões linguísticos.

Por fim, é necessário considerar as discussões recentes sobre a necessidade de simplificação da linguagem jurídica (vide discussão anterior sobre o "juridiquês"). É possível que a diminuição da variabilidade lexical ao longo do tempo esteja relacionada, de fato, a esforços para a simplificação dos textos das normas jurídicas. Uma análise mais aprofundada da diversidade lexical poderá elucidar quais desses fatores possuem, de fato, maior influência na tendência de diminuição da variabilidade lexical observada ao longo do tempo.

4.2 *N*-gramas frequentes

A observação das palavras e sequências de palavras mais frequentes em diferentes trechos do *dataset* pode fornecer sugestões a respeito de elementos relevantes em cada período analisado. Nesta seção, são investigados os *n*-gramas mais frequentes em cada período, com *n* igual a 1 e a 2 – isto é, são analisadas as palavras isoladamente (1-gramas, ou unigramas) e pares de palavras contíguas (2-gramas, ou bigramas).

Para esta análise, foram removidas as *stop words* ("palavras vazias"), ou seja, palavras consideradas irrelevantes para as análises. A lista inicial de palavras vazias

utilizada foi aquela disponibilizada para a língua portuguesa pelo Natural Language Toolkit – NLTK (BIRD; KLEIN; LOPER, 2009) (https://www.nltk.org/, acesso em: 30 nov. 2021), biblioteca para processamento de texto disponível para Python. Essa lista inclui 204 itens, dentre os quais se destacam palavras funcionais/gramaticais (de classe fechada, como artigos, conjunções, preposições, pronomes) e verbos de alta frequência ("estar", "ir", "haver", "ser", "ter").

A primeira análise realizada foi a comparação entre as vinte palavras lexicais (substantivos, verbos, adjetivos e advérbios) mais frequentes em cada período, considerando apenas aquelas compostas exclusivamente por caracteres alfabéticos. Como esperado, em todos os períodos, tais buscas retornaram, sobretudo, termos ligados ao gênero textual legislativo, como "lei", "artigo", "publicação" e "vigor". No período I (1946-1964), o verbo "entrar" utilizado performativamente ("entrará em vigor") é encontrado no futuro, assim como no período II, sendo substituído pelo presente no período III ("entra em vigor"). No período II, destaque para o termo "militar", que também comparece entre os mais frequentes.

Para a análise de bigramas por período, utilizou-se a função *collocations* disponibilizada pelo NLTK, que permite a identificação de bigramas cuja frequência de combinação entre seus elementos é mais alta que o estatisticamente esperado. Os vinte bigramas com associação mais relevante (de acordo com o NLTK) em cada período são apresentados no Quadro 1. Novamente, se destaca a presença de itens próprios do gênero textual legislativo ("sua publicação", "esta lei", "lei entrará", "lei entra", "caput deste", entre tantos outros) – mas, também, de colocações referentes a questões orçamentárias (por exemplo, "mil cruzeiros", "por cento", "crédito suplementar", "orçamento fiscal", "recursos necessários"), instituições e serviços públicos (como "previdência social" no período I, "tribunal regional" no período II, "congresso nacional" no período III), e até mesmo um antropônimo ("juscelino kubitschek", no período I).

Quadro 1 – Bigramas com associação mais relevante de acordo com a função *collocations* do NLTK.

Período	**Bigramas**
Período I (1946-1964)	poder executivo; sua publicação; crédito especial; esta lei; lei entrará; executivo autorizado; pelo ministério; desta lei; juscelino kubitschek; mil cruzeiros; distrito federal; presente lei; para atender; que trata; por cento; outras providências; obras públicas; acôrdo com; nos têrmos; previdência social
Período II (1964-1988)	esta lei; desta lei; sua publicação; poder executivo; distrito federal; que trata; outras providências; bem como; neste artigo; lei entra; lei entrará; por cento; crédito especial; executivo autorizado; presente lei; nesta lei; parágrafo único; tribunal regional; seguinte redação; poderá ser
Período III (1988-2017)	desta lei; que trata; deste artigo; poder executivo; esta lei; sua publicação; lei entra; por cento; nos termos; bem como; caput deste; crédito suplementar; distrito federal; seguridade social; orçamento fiscal; metros até; congresso nacional; neste artigo; recursos necessários; para atender

Fonte: elaborado pelos autores.

Uma forma interessante de visualização geral de *n*-gramas são as nuvens de palavras (*word clouds*), que permitem observar aqueles mais frequentes no conjunto de textos. Para efeito de comparação, a Figura 2 mostra as nuvens de palavras referentes aos três períodos analisados neste estudo. Os resultados encontrados, à primeira vista, não indicam uma mudança tão significativa no padrão de construção dos bigramas frequentes nos três períodos analisados do *dataset*, conforme já havia sido antecipado pelas colocações identificadas no Quadro 1. Uma pesquisa que remova mais elementos irrelevantes para a investigação (por exemplo, que inclua palavras como "lei", "artigo", "publicação", "disposições" e "vigor" na lista de *stop words*) poderá, possivelmente, promover análises mais interessantes da mudança lexical nesse *dataset* ao longo do tempo.

Figura 2 – Nuvens de palavras referentes, respectivamente, aos períodos I, II e III.

Fonte: elaborada pelos autores.

4.3 *Hapax legomena*

Nesta análise preliminar, também foram considerados os *hapax legomena*, isto é, ocorrências que só aparecem uma vez no *dataset*. Em todos os períodos foram encontrados erros de digitação (ou de digitalização), mas foi possível também perceber ocorrências que podem ser úteis como indicadores de mudanças diacrônicas a serem verificadas.

No período I (1946-1964), encontram-se erros de digitação (ou de digitalização) como: "atiântica" por "atlântica", "ciqüenta" por "cinqüenta", "dafender" por "defender", "federai" por "federal", "fiacalização" por "fiscalização", "jalgamento" por "julgamento", entre outros, mas também palavras com grafias diferentes ("datilógrafo" e "dactilógrafo"). Na grande maioria dos casos, os *hapax legomena* encontrados são nomes próprios de pessoa ("thyco", "teodoro", "lucélia", "izidoro") ou topônimos ("uaupés", "urucará", "turiaçú"). Foram observados vários termos estrangeiros, seja antropônimos ("truman", referente a Sally Truman; "kirkons", de Kirkons Nodhjalp), seja nomes de empresas e instituições ("thomas", da firma Thomas de la Rue & Company Limited; "telefonaktisebolaget", da firma Telefonaktisebolaget - l.m. Ericsson de Stokolmo), seja utilizados como terminologia de campos específicos ("truks", em "caixas de graxas para truks de carros e vagões"; "royalty", também grafado como "rorgarties").

No Quadro 1, observou-se que entre as colocações mais relevantes desse período está o nome do presidente Juscelino Kubitschek. Curiosamente, a grafia de seu nome está presente de formas muito variadas, tendo sido encontrada como "juscelimo" e "juscellino", e o sobrenome com nove variações de ocorrência única: "kubigtschek", "kubitscbek", "kubitschet", "kubitsckek", "kubitshek", "kubltsckek", "kubsitschek", "kubstischek", "kubstschek".

Para finalizar, no período I foram encontrados três interessantes *hapax legomena* não presentes nos períodos II e III: "chefatura" (com significado de "chefia", no caso em

questão da polícia), "locupletamento" (no período II há uma ocorrência de "locupletaram", do verbo "locupletar", com o significado de "enriquecer, não necessariamente de forma lícita") e "leprólogos" (peritos em hanseníase, antigamente/popularmente conhecida como "lepra").

A análise dos *hapax legomena* do período II (1964-1988) inclui erros de digitação (ou de digitalização) como "acôdo" por "acordo", "administraç/ão" por "administração", "ainciso" por "inciso", "aiterações" por "alterações", "anônino" por "anônimo", "biihetes" por "bilhetes", "canpos" por "campos", "exêcito" por "exército", "imcompatibilidade" por "incompatibilidade", além de grafias diferentes ("balisamento" e "balizamento", "autárquias" uma única vez frente a "autarquias"). Nesse período também há muitos nomes próprios de pessoa ("adolfo", "agripino", "doris") e topônimos ("afuá", "alpercata", "botuverá", "cajamar", "erechim", "iepê", "pojuca"). Comparecem, aparentemente em número menor, termos estrangeiros: nomes próprios ("charles", de Charles de Gaulle), de empresas e instituições (como da empresa Deutsche Ibero-Amerika Stiftung, ou da Standard Elektrik Arktiengeselischaft), ou utilizados como terminologia de campos específicos ("broadcastinge", em "estação de broadcastinge de televisão"). Assinalamos, ainda, uma presença considerável de siglas, muitas delas referentes a órgãos, instituições ou programas estatais, como: "capes" (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), "cbdf" (Corpo de Bombeiros do Distrito Federal), "cdfa" (Comissão Desportiva das Forças Armadas), "cbee" (Companhia Brasileira de Energia Elétrica), "celf" (Centrais Elétricas Fluminenses), "codeplan" (Companhia de Desenvolvimento do Planalto Central), "dag" (Departamento de Administração Geral), "dnff" (Departamento Nacional de Estradas de Ferro), "iee" (Instituto de Educação do Excepcional), "pebe" (Programa Especial de Bôlsas de Estudo), "pgpm" (Política de Garantia de Preços Mínimos), entre muitíssimas outras. O termo "nosológico" comparece com uma única ocorrência e não se encontra nos outros dois períodos.

A varredura dos *hapax legomena* do período III (1988-2017) mais uma vez evidenciou erros de digitação (ou de digitalização), sempre numerosos (entre os vários, citamos: "vidaou" por "vida ou", "uperávir" por "superávit", "títulares" por "titulares", "trêsvaras" por "três varas", "trêsreais" por "três reais", "trubutária" por "tributária", "tratameto" por "tratamento", "prsidente" por "presidente", "cenros" por "centros"). Enquanto os topônimos de ocorrência única também se mantiveram numerosos (como "xexéu", "urupá", "uruarama", "uarini", "sooretama", "mozarlândia", "iraci", "cujubim", "aracoiaba"), os nomes próprios de pessoa foram relativamente poucos ("thaumaturgo", "teotônio", "sidney", "januário", "ezequiel", "euclydes"), assim como os nomes de instituições em língua estrangeira (como Morgan Guaranty Trust Company of New York). Foram novamente encontrados termos ingleses utilizados em campos específicos, como "sweepstakes", "pellets", "loft", "lockout". Mais uma vez houve uma presença considerável de siglas e acrônimos, como: "zees" (zoneamentos ecológico-econômicos), "utn" (usinas termonucleares), "tsb" (técnico em saúde bucal), "transpetro" (Petrobras Transporte S.A.), "telpe" (Telecomunicações de Pernambuco S.A.), "ted" (transferência eletrônica de dados), "srbf" (Secretaria da Receita Federal do Brasil), "snsm" (Sistema Nacional de Sementes e Mudas), "serse" (Secretaria Especial da Região Sudeste), "renavam" (Registro Nacional de Veículos Automotores), "qofarm" (Quadro de Oficiais Farmacêuticos), "cgee" (Centro de Gestão e Estudos Estratégicos), "boh" (Boletim de Ocupação Hoteleira), entre muitos outros. Destaque para o número crescente de siglas referentes a instituições de ensino e de pesquisa, tais quais: "usp" (Universidade de São Paulo), "unirio" (Universidade do Rio de Janeiro), "unifesp" (Universidade Federal de São Paulo), "ufla" (Universidade Federal de Lavras), "ufg" (Universidade Federal de Goiás), "inpa" (Instituto Nacional de Pesquisas da Amazônia), "fua" (Fundação Universidade do Amazonas), entre tantas outras.

5 Considerações finais

Neste artigo, discutiu-se a relação entre linguagem e direito, tecendo um pequeno panorama sobre os estudos de interface entre as duas áreas no Brasil. Em seguida, após uma reflexão sobre a utilização de *corpora* e *datasets* em pesquisas linguísticas, é apresentada uma análise lexical preliminar da Base de Normas Jurídicas Brasileiras, que contém o texto de quase treze mil leis promulgadas no Brasil entre 1946 e 2017. Apesar de possuir inconsistências e imperfeições, o *dataset* disponibilizado por Martim, Lima e Araujo (2018) se mostra uma fonte de dados interessante para estudos que possuam como objetivo investigar aspectos (sejam eles linguísticos e/ou sociais) da legislação brasileira, em particular sob uma perspectiva diacrônica, como a que é apresentada aqui.

Labov (1972b, p. 100) afirma que "A grande arte do linguista histórico é tirar o melhor proveito desses dados ruins – 'ruins' no sentido de que podem ser fragmentários, corrompidos ou muitas vezes removidos das produções reais dos falantes nativos" (tradução nossa)[23]. Em um certo sentido, o trabalho do linguista que lida com grandes *datasets* não é muito diferente. Como pode ser observado neste trabalho, dados disponibilizados ao pesquisador, sobretudo quando foram obtidos sem a devida curadoria linguística, podem se revelar falhos, incompletos e com os mais variados tipos de erros. Ainda assim, a viabilização desse tipo de dado se mostra importante por permitir que sejam realizadas análises em grande escala e a um custo relativamente baixo – seja de tempo, seja propriamente financeiro.

Conforme se discutiu anteriormente, *corpora* compilados para estudos linguísticos preveem que seja pensada uma arquitetura específica para as finalidades da pesquisa que se pretende empreender, por meio de uma amostragem adequada dos dados e seu devido balanceamento, a fim de se obter a representatividade de uma

[23] [t]he great art of the historical linguist is to make the best of this bad data – 'bad' in the sense that it may be fragmentary, corrupted, or many times removed from the actual productions of native speakers.

determinada variedade linguística. Tal processo implica que haja conhecimento, por parte da equipe de compilação, das características específicas de uma linguagem setorial (neste caso, das leis), assim como do contexto de sua produção. A compilação de um *corpus* de referência da linguagem jurídica do português brasileiro, recentemente empreendida por uma das autoras, tenta cobrir justamente esses pontos. O LEX-BR-Ius (FERRARI; MARQUES, em preparação) será a seção legislativa de tal *corpus* e incluirá, também, textos completos de normas legislativas federais brasileiras[24]. Além disso, (a) as normas que farão parte do corpus deverão estar em vigor no momento da coleta; e (b) o uso efetivo das leis está sendo verificado através de uma análise prévia com base no critério de relevância, o que será fator decisivo na escolha por aquelas que irão entrar ou não no corpus. O balanceamento entre textos que possuem, por sua própria natureza, dimensões muito variáveis se dará pela criação de subseções relativas às diferentes tipologias legislativas. As subseções terão números de palavras similares, mantendo a integridade dos textos, mas garantindo, assim, sua comparabilidade interna em termos quantitativos. A extração dos dados será seguida por diferentes tipos de tratamento computacional, o que permitirá que o *corpus* esteja disponível em sua versão bruta (*raw text*), completo com anotação textual (*markup* textual) e anotação POS (classe gramatical). O texto bruto passará por uma limpeza prévia que eliminará expressões formulaicas típicas do gênero, assim como os campos de assinaturas finais de cada norma. Tais informações, contudo, serão mantidas nos metadados, que incluirão data de extração, pessoas responsáveis por sua promulgação, número de palavras, artigos em que é subdividida cada norma, ementa, assunto e alterações. De tal forma, a varredura dos metadados possibilitará ao pesquisador uma definição prévia do tipo de lei que será interessante para sua

[24] A decisão por não extrair excertos das normas de tamanho igual (o que facilitaria o balanceamento) respeita a integridade do texto, por entender que a língua necessita de seu contexto próprio de uso para que as regularidades e padrões não fiquem enviesados (BIBER, 1998; BIBER; CONRAD, 2009; BERBER SARDINHA, 2010; BIBER *et al*., 2016; EGBERT; LARSSON; BIBER, 2020).

pesquisa (por legislatura, por período, por assunto etc.).

Embora não tenha sido compilado utilizando critérios tão rigorosos quanto esses, o *dataset* disponibilizado por Martim, Lima e Araujo (2018) e aqui apresentado nos permitiu vislumbrar padrões interessantes que podem constituir o ponto de partida para pesquisas futuras. Salientamos a necessidade de cautela nas generalizações dos resultados, mas acreditamos que, com os devidos cuidados, esta pesquisa possa abrir uma série de oportunidades para análises futuras utilizando esses mesmos dados. Do ponto de vista lexical, pretende-se refinar, em um trabalho posterior, a análise da variação da frequência de determinadas palavras e expressões ao longo do tempo – observando, inclusive, a conceitualização em torno desses itens lexicais, conforme apresentado, por exemplo, por Cunha *et al*. (2018). Para isso, poderá ser importante adicionar procedimentos como a lematização do texto, além da definição de periodizações mais finas (por exemplo, pode-se trabalhar com períodos de dez ou cinco anos, ou ainda menos, em vez de apenas três períodos, como foi realizado aqui). Uma possibilidade promissora de investigação é a utilização do método apresentado por Cunha e Wichmann (2021) para a identificação das datas de fixação e de obsolescência de palavras e expressões no conjunto de dados analisado: assim, será possível identificar, por exemplo, em que momento determinados *n*-gramas passaram a ser mencionados nas leis e quando outros deixaram de aparecer nas normas brasileiras. Por fim, uma melhor compreensão dos fenômenos será beneficiada por uma pesquisa mais aprofundada acerca dos períodos históricos aqui apresentados e seus desdobramentos nas leis; e, ao mesmo tempo, por um aprofundamento na relevância de determinadas leis em cada período, sua validade e sua efetiva utilização na jurisprudência.

Referências

AQUINO, R.; DOUGLAS, W. **Manual de português e redação jurídica**. 6. ed. Niterói: Impetus, 2017.

BAKER, P. **Sociolinguistics and Corpus Linguistics**. Edinburgh: Edinburgh University Press, 2010.

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

BERBER SARDINHA, T. A abordagem metodológica da análise multidimensional. **Gragoatá**, v. 15, n. 29, p. 107-125, 2010. DOI https://doi.org/10.22409/gragoata.v15i29.33077

BIBER, D. **Variation across speech and writing**. Cambridge: Cambridge University Press, 1988. DOI https://doi.org/10.1017/CBO9780511621024

BIBER, D. Methodological issues regarding corpus-based analyses of linguistic variation. **Literary and Linguistic Computing**, v. 5, n. 4, p. 257-269, 1990. DOI https://doi.org/10.1093/llc/5.4.257

BIBER, D. On the complexity of discourse complexity: A multidimensional analysis. **Discourse Processes**, v. 15, n. 2, p. 133-163, 1992. DOI https://doi.org/10.1080/01638539209544806

BIBER, D. Representativeness in Corpus Design. Literary and Linguistic Computing, v. 8, n. 4, p. 243-257, 1993. DOI https://doi.org/10.1093/llc/8.4.243

BIBER, D; CONRAD, S. **Register, genre, and style**. Cambridge: CUP, 2009. DOI https://doi.org/10.1017/CBO9780511814358

BIBER, D.; REPPEN, R.; SCHNUR, E.; GHANEM, R. On the (non)utility of Juilland's *D* to measure lexical dispersion in large corpora. **International Journal of Corpus Linguistics**, v. 21, n. 4, p. 439-464, 2016. DOI https://doi.org/10.1075/ijcl.21.4.01bib

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python**: analyzing text with the Natural Language Toolkit. Sebastopol: O'Reilly, 2009.

BITTAR, E. C. B. **Linguagem jurídica**. 4. ed. São Paulo: Saraiva, 2009.

CARVALHO, L. Os dicionários jurídicos bilíngües e o tradutor - dois binômios em Direito Contratual. **TradTerm**, v. 12, p. 309-347, 2006. DOI https://doi.org/10.11606/issn.2317-9511.tradterm.2006.46903

CORDEIRO, J. M. Anos de chumbo ou anos de ouro? A memória social sobre o governo Médici. **Estudos Históricos**, v. 22, n. 43, p. 85-104, 2009. DOI https://doi.org/10.1590/S0103-21862009000100005

CUNHA, E.; MAGNO, G.; CAETANO, J.; TEIXEIRA, D.; ALMEIDA, V. Fake news as we feel it: perception and conceptualization of the term "fake news" in the media. *In*: STAAB, S.; KOLTSOVA, O.; IGNATOV, D. I. (ed.). **Social informatics** [Lecture Notes in Computer Science, n. 11185]. Cham: Springer, 2018. p. 151-166. DOI https://doi.org/10.1007/978-3-030-01129-1_10

CUNHA, E. L. T. P.; WICHMANN, S. An algorithm to identify periods of establishment and obsolescence of linguistic items in a diachronic corpus. **Corpora**, Edinburgh, v. 16, n. 2, p. 205-236, 2021. DOI https://doi.org/10.3366/cor.2021.0218

DAMIÃO, R. T.; HENRIQUES, A. **Curso de português jurídico**. 14. ed. São Paulo: Atlas, 2020.

DIMITRIADIS, A.; MUSGRAVE, S. Designing linguistic databases: a primer for linguists. *In*: EVERAERT, M.; MUSGRAVE, S.; DIMITRIADIS, A. (ed.). **The use of databases in cross-linguistic studies**. Berlin/New York: Mouton de Gruyter, 2009. p. 13-75.

DINIZ, M. H. **Dicionário jurídico**. São Paulo: Saraiva, 1998.

EGBERT, J.; BURCH, B.; BIBER, D. Lexical dispersion and corpus design. **International Journal of Corpus Linguistics**, v. 25, n. 1, p. 89-115, 2020. DOI https://doi.org/10.1075/ijcl.18010.egb

EGBERT, J.; LARSSON, T.; BIBER, D. **Doing Linguistics with a Corpus**. Methodological Considerations for the Everyday User. Cambridge: Cambridge University Press, 2020. DOI https://doi.org/10.1017/9781108888790

FERRARI, L. A; MARQUES, C. G. F. **O corpus LEX-BR-Ius, seção legislativa das leis federais brasileiras**: arquitetura e primeiras análises. Em preparação.

GRAY, J. Towards a Genealogy of Open Data. *In*: **GENERAL CONFERENCE OF THE EUROPEAN CONSORTIUM FOR POLITICAL RESEARCH**. Glasgow, 2014. DOI https://dx.doi.org/10.2139/ssrn.2605828

GUIMARÃES, D. T. **Dicionário técnico jurídico**. São Paulo: Rideel, 2013.

HARDIE, A. CQPweb — combining power, flexibility and usability in a corpus analysis tool. **International Journal of Corpus Linguistics**, v. 17, n. 3, p. 380-409, 2012. DOI https://doi.org/10.1075/ijcl.17.3.04har

HEAPS, H. S. **Information retrieval**: computational and theoretical aspects. New York: Academic Press, 1978.

HERDAN, G. **Quantitative linguistics**. London: Butterworth, 1964.

IVO, G. O direito e a inevitabilidade do cerco da linguagem. *In*: CARVALHO, P. de B.; CARVALHO, A. T. de (org.). **Constructivismo lógico-semântico**. 2. ed. revista v. 1. São Paulo: Noeses, 2020. p. 65-91.

KRIEGER, M. G.; MACIEL, A. M. B.; BEVILACQUA, C. R.; ROCHA, J. C.; FINATTO, M. J. B. **Dicionário de Direito Ambiental**. Porto Alegre: Editora da Universidade, 1998.

KRIEGER, M. G.; MACIEL, A. M. B.; BEVILACQUA, C. R.; FINATTO, M. J. B.; REUILLARD, P. C. R. **Glossário de Gestão Ambiental**. São Paulo: Disal Editora, 2006.

KRIEGER, M. G.; MACIEL, A. M. B.; BEVILACQUA, C. R.; FINATTO, M. J. B. **Dicionário de Direito Ambiental**. 2. ed. Rio de Janeiro: Lexikon, 2008.

LABOV, W. **The social stratification of English in New York City**. Washington: Center for Applied Linguistics, 1966.

LABOV, W. **Sociolinguistic patterns**. Philadelphia: University of Pennsylvania Press, 1972a.

LABOV, W. Some principles of linguistic methodology. **Language in Society**, v. 1, n. 1, p. 97-120, 1972b. Disponível em: https://www.jstor.org/stable/4166672. Acesso em: 12 nov. 2021. DOI https://doi.org/10.1017/S0047404500006576

MACIEL, A. M. B. **Para o reconhecimento da especificidade do termo jurídico**. 2001. 291 f. Tese. (Doutorado em Estudos da Linguagem) – Programa de Pós-Graduação em Letras. Universidade Federal do Rio Grande do Sul, 2001.

MARTIM, H.; LIMA, J. A. O.; ARAUJO, L. C. Base de Normas Jurídicas Brasileiras: uma iniciativa de Open Government Data. **Perspectivas em Ciência da Informação**, v. 23, n. 4, p. 133, 2018. DOI https://doi.org/10.1590/1981-5344/3567

MELLO, H. Methodological issues for spontaneous speech corpora compilation: the case of C-ORAL-BRASIL. *In*: RASO, T.; MELLO, H. (org.). **Spoken Corpora and Linguistic Studies**. Amsterdam: John Benjamins, 2014. v. 1, p. 27-68. DOI https://doi.org/10.1075/scl.61.01mel

PETRI, M. J. C. **Manual de linguagem jurídica**. 3. ed. São Paulo: Saraiva, 2017.

RAMOS, J. J. S. C. **Ocorrência e interpretação dos verbos modais 'dever' e 'poder' em contexto jurídico**: contributos para uma análise juslinguística. 207 f. Tese (Doutorado) – Filozofická Fackultá, Univerzita Karlova, Praha, Rep. Tcheca, 2017 *apud* SVOBODOVÁ (2017).

RASO, T. O corpus C-ORAL-BRASIL. *In*: RASO, T.; MELLO, H. (org.). **C-ORAL-BRASIL I**. Corpus de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012. p. 55-90.

RENEAR, A. H.; SACCHI, S.; WICKETT, K. M. Definitions of *dataset* in the scientific and technical literature. **Proceedings of the American Society for Information Science and Technology**, v. 47, n. 1, 2010. DOI https://doi.org/10.1002/meet.14504701240

RESENDE, S. V.; MAVERICK, R. Planejamento, compilação e organização de corpora. *In*: **Anais do EBRALC 2015 & ELC 2015** [Blucher Social Science Proceedings, n. 3, v. 2]. São Paulo: Blucher, 2016. p. 27-35. DOI https://doi.org/10.5151/sosci-viiieblc-xiii-elc-06_artigo_03

SANTOS, F. Patronagem e Poder de Agenda na Política Brasileira. **Dados**: Revista de Ciências Sociais, v. 40, n. 3, 1997. DOI https://doi.org/10.1590/S0011-52581997000300007

SANTOS, W. **Dicionário jurídico brasileiro**. Belo Horizonte: Del Rey, 2001.

SEARLE, J. R. **Speech Acts**. An Essay in the Philosophy of Language. Cambridge: Cambridge University Press, 1969. DOI https://doi.org/10.1017/CBO9781139173438

SVOBODOVÁ, I. Modalidade não epistêmica na linguagem jurídica: um estudo contrastivo. **Caligrama**, Belo Horizonte, v. 22, n. 2, p. 103-133, 2017. DOI http://dx.doi.org/10.17851/2238-3824.22.2.103-133

TEIXEIRA, W. R.; LIMA, J. A. O.; ARAUJO, L. C.; VIERO, D. M.; SANTANA, F. F.; HERINGER, F. R. A.; MARTIM, H.; VIEIRA FILHO, J. J. Exemplo de extração de definições em textos articulados de normas jurídicas com o apoio do processamento de linguagem natural. **Cadernos de Informação Jurídica**, v. 6, n. 1, p. 49-64, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/119039. Acesso em: 20 dez. 2021.

WARAT, L. A. **O direito e sua linguagem**. Porto Alegre: Sergio Antonio Fabris Editor, 1995.

Artigo recebido em: 30.12.2021 Artigo aprovado em: 30.05.2022

DOI: 10.14393/DL52-v16n4a2022-13

Manual de anotação como recurso de Processamento de Linguagem Natural: o modelo *Universal Dependencies* em língua portuguesa

An annotation manual as a Natural Language Processing resource: the Universal Dependencies model in Portuguese

Magali Sanches DURAN[*]
Maria das Graças Volpe NUNES[**]
Lucelene LOPES[***]
Thiago Alexandre Salgueiro PARDO[****]

RESUMO: Com o avanço da área de Processamento de Linguagem Natural (PLN), *corpora* são recursos que têm tido um lugar de destaque. Mais do que subsidiar estudos linguísticos, eles constituem as bases para o treinamento de modelos de Aprendizagem de Máquina e para o desenvolvimento de aplicações computacionais de ponta. Particularmente, há grande necessidade de *corpora* anotados, porém sua geração requer outro recurso essencial, o manual de anotação, que instancia o modelo de anotação de interesse para a língua em questão e delineia as decisões de anotação que devem ser adotadas. Neste artigo,

ABSTRACT: With the advances of the Natural Language Processing area, corpora are resources that have had a prominent place. More than subsidizing linguistic studies, they constitute the basis for training Machine Learning models and developing cutting-edge computational applications. In particular, there is a great need for annotated corpora, but their production requires another essential resource, the annotation manual, which instantiates the annotation model of interest for the language in question and outlines the annotation decisions that should be adopted. In this paper, we explore issues related to the

[*] Doutora em Estudos Linguísticos pela UNESP de São José do Rio Preto e pesquisadora de pós-doutorado no NILC. ORCID: https://orcid.org/0000-0002-3843-4600. magali.duran@uol.com.br

[**] Professora Doutora do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, no campus de São Carlos. ORCID: https://orcid.org/0000-0002-2776-6140. gracan@icmc.usp.br

[***] Doutora em Ciência da Computação pela PUC Rio Grande do Sul e pesquisadora de pós-doutorado no NILC. ORCID: https://orcid.org/0000-0003-0314-140X. lucelene@gmail.com

[****] Professor Doutor do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, no campus de São Carlos. ORCID: https://orcid.org/0000-0003-2111-1319. taspardo@icmc.usp.br

exploramos questões relacionadas ao desenvolvimento de manuais para a anotação de *corpus* em português brasileiro segundo o modelo internacional *Universal Dependencies*, amplamente adotado na área. Partimos da discussão da evolução do PLN e o uso de *corpora*, passamos pelas questões, recursos e ferramentas fundamentais relacionados à representação sintática, discutimos o modelo *Universal Dependencies* e apresentamos as principais decisões tomadas na instanciação de suas diretrizes no português brasileiro. Por questões práticas e de didática, dividimos o manual em duas partes: o Manual de Anotação de *PoS tags* (anotação morfossintática) e o Manual de Anotação Relações de Dependência. Ambos foram resultado do processo relatado neste artigo e estão disponíveis para livre acesso no site do projeto POeTiSA na Web.

PALAVRAS-CHAVE: *Corpora* anotados. Manual de anotação. *Universal Dependencies*. Árvores de dependência. Português brasileiro.

development of manuals for the annotation of Brazilian Portuguese corpora according to the Universal Dependencies model, widely adopted in the field. We discuss the evolution of NLP and the use of corpora, the fundamental issues, resources and tools related to syntactic representation, the Universal Dependencies model, and the main decisions made in the instantiation of UD guidelines in Brazilian Portuguese. For practical and didactic reasons, we divided the manual into two parts: the PoS Tag Annotation Manual (morphosyntactic annotation) and the Dependency Relations Annotation Manual. Both resulted from the process reported in this paper and are available for free access on the POeTiSA project's Web site.

KEYWORDS: Annotated corpora. Annotation manual. Universal Dependencies. Dependency trees. Brazilian Portuguese.

1 Introdução

Neste artigo, apresentamos o manual de anotação como recurso de Processamento de Linguagem Natural (PLN), o que não é uma ideia muito difundida nem nos círculos de linguistas nem nos círculos de cientistas da computação. Para fundamentar tal afirmação, dividimos esta introdução em três subseções, iniciando com uma retrospectiva do papel do linguista ao longo da evolução do PLN (1.1), passando pela ascensão dos *corpora* anotados como recurso valioso para os novos métodos de PLN (1.2) e terminando por discorrer sobre a importância dos manuais de anotação como parte indissociável de um esquema de anotação (1.3).

1.1 Processamento de Linguagem Natural: a evolução rumo aos *corpora*

O PLN desenvolveu-se muito nas últimas décadas e o papel do linguista nos projetos de PLN também se modificou. Até a década de 1990, acreditava-se que a representação do conhecimento linguístico na forma de léxicos, ontologias, gramáticas, regras e outros formalismos de representação de conhecimento seria capaz de fornecer todo o conhecimento necessário para o PLN.

A partir dos anos 1990, ocorreram alguns avanços tecnológicos e mudanças de comportamento social que revolucionaram toda essa área. O advento da internet e de sua interface fez aumentar exponencialmente a quantidade de documentos em línguas naturais na Web e, consequentemente, aumentou a demanda por tarefas de processamento: de busca, de classificação, de auxílio à produção, de análise e correção, de consulta e de sumarização, entre muitas outras. Os avanços em matéria de *hardware* aumentaram de forma gigantesca a capacidade de processamento e de armazenagem de dados, permitindo o emprego de métodos sofisticados, antes proibitivos por consumirem muito tempo e espaço de memória. E, por fim, mais recentemente, o surgimento das redes sociais aumentou o conteúdo gerado por usuários, exigindo que o PLN passasse a lidar com textos que fogem em muito do registro formal da língua.

Essas mudanças de cenário, de uma forma ou de outra, cada uma com sua influência, trouxeram o PLN para a era da Aprendizagem de Máquina (AM) (MITCHELL, 1997). Com a AM, é possível que a máquina aprenda tarefas baseadas em conhecimento humano sem que seja necessário que o humano represente e estruture formal e explicitamente esse conhecimento. Isso é vantajoso principalmente no caso de problemas cuja solução requer conhecimento cuja formalização é difícil ou altamente custosa. O importante para a AM é ter um conjunto suficientemente grande e representativo de exemplos sobre o qual os algoritmos possam aprender. No caso de PLN, esses exemplos constituem os *corpora* de texto ou de fala transcrita, que,

alinhando-se à tradição empirista de pesquisa, representam um recorte de interesse da língua (MANNING; SCHÜTZE, 1999; SARDINHA, 2000).

A introdução da AM no PLN foi revolucionária e inaugurou um ritmo acelerado de novas conquistas. Um dos mais modernos modelos de AM, as Redes Neurais Profundas (*Deep Learning*) (GOODFELLOW; BENGIO; COURVILLE, 2016) alavancaram o desempenho dos tradutores automáticos - a primeira grande tarefa de PLN - e têm sido usadas, com excelentes resultados, nos mais variados cenários: *chatbots*, reconhecimento e geração de fala, geração de documentos, análise de sentimentos, detecção de discurso de ódio e *fake news*, entre muitos outros.

Nesse atual modo de fazer PLN, o papel do linguista passa a incluir o de subsidiar o projeto do *corpus*, o qual fornecerá o conhecimento aos algoritmos de aprendizagem. Há um compromisso entre o *corpus* fornecido, o algoritmo de AM utilizado e os resultados obtidos. Assim, quanto mais "informados" forem os exemplos fornecidos, melhores serão os resultados alcançados. Diversos atributos podem ser anotados em textos, tanto associados a análises linguísticas (como os atributos morfológicos, sintáticos e semânticos), quanto associados a outros tipos de análise, como análise de sentimentos e de opiniões.

Se, pelo lado da computação, o PLN evolui por meio da melhoria dos métodos de aprendizagem automática, do lado da linguística o PLN evolui por meio da produção de *corpora* anotados de forma cada vez mais lógica, completa e consistente, guiando e solidificando modelos e teorias linguísticas.

1.2 *Corpus* no coração do PLN

A anotação de *corpus* é uma ciência (IDE, 2017). Ela começa pela modelagem do fenômeno a ser anotado, que culmina na definição de um esquema de anotação, com um conjunto de etiquetas a serem atribuídas aos segmentos de texto. A modelagem de

fenômenos para anotação é um trabalho árduo e, por isso, é comum haver reaproveitamento de esquemas de anotação bem sucedidos na comunidade de PLN.

As línguas que possuem recursos mais avançados de PLN criam modelos de anotação e esses costumam ser replicados e adaptados por outras línguas que têm menos recursos humanos dedicados à criação de modelos. Foi assim que proliferaram, por exemplo, clones e adaptações da Framenet (BAKER; FILLMORE; LOWE, 1998) e do Propbank (PALMER; GILDEA; KINGSBURY, 2005), que modelaram, de formas diferentes, a teoria de papéis semânticos de Fillmore (1968), e da Verbnet (KIPPER-SCHULER, 2005), que modelou a teoria de classes verbais de Levin (1993).

Embora em um dado momento da história o reaproveitamento de esquemas tinha só o objetivo de poupar esforço de modelagem, não se demorou a perceber o grande potencial que a utilização de um mesmo esquema de anotação poderia fornecer para o PLN, no sentido de facilitar a comparação entre as línguas e permitir o desenvolvimento de ferramentas e aplicações de PLN multilíngues.

Em 2006, um desafio de PLN para criar um *parser* de dependências multilíngue (BUCHHOLZ; MARSI, 2006) exigiu que os organizadores adotassem um esquema de anotação de *corpus* razoavelmente flexível para anotar as treze línguas envolvidas da competição. Devido a sua simplicidade, o esquema escolhido foi o apresentado em Nivre *et al*. (2006). Os resultados no treinamento de *parsers* multilíngues a partir desses *corpora* foram promissores e, em 2015, um esquema de anotação "universal" de dependências sintáticas já estava sendo utilizado por várias línguas. Esse esquema, chamado de *Universal Dependencies* (UD) (NIVRE *et al*., 2015, 2020; MARNEFFE *et al*., 2021) pretende ser independente de língua e tem ganhado adeptos em diversas partes do mundo. Mais de cem línguas, das mais diferentes famílias, já possuem *corpora* anotados no esquema UD.

A disponibilidade de *corpora* anotados em diversas línguas seguindo um mesmo esquema de anotação abriu oportunidade para o desenvolvimento de vários

parsers[1] cujos métodos são independentes de língua. Como exemplos, podemos citar o UDPipe (STRAKA; STRAKOVA, 2017), o UDify (KONDRATYUK; STRAKA, 2019), o Stanza (QI et al., 2020) e o *parser* do spaCy (HONNIBAL; JOHNSON, 2015). Esses *parsers*, por terem uma capacidade muito boa de aprendizado[2], mesmo a partir de poucos dados, muitas vezes são usados para pré-anotar novos *corpora* no formato UD, que são revisados por humanos. Assim, num ciclo virtuoso, com mais dados para retreinamento, os *parsers* podem ser continuamente aprimorados.

O reaproveitamento de um esquema de anotação não se dá, no entanto, sem esforço. Pustejovsky, Bunt e Zaene (2017) distinguem o que chamam de *markables*[3] dos *non-markables* em um esquema de anotação. Os *markables* estão diretamente associados aos conjuntos de etiquetas (elementos explícitos na anotação) e os *non-markables* são as partes de um esquema de anotação expressas por meio de diretrizes que devem ser seguidas a fim de se alcançar consistência na anotação (elementos implícitos na anotação). Os *markables* não podem ser alterados, pois garantem a comparabilidade física entre as línguas que o utilizam. Sem os *non-markables*, contudo, corre-se o risco de atribuir etiquetas iguais a fenômenos diferentes e isso também compromete a comparabilidade entre as línguas.

Os *non-markables* são expressos por meio de diretrizes de anotação e, como as diretrizes fazem uso de muitos exemplos, elas normalmente são dependentes de língua. Assim, seja para reaproveitar um esquema de anotação feito para uma outra língua, seja para utilizar um esquema "universal", é essencial instanciar as instruções de anotação na língua em que se vai aplicar o esquema.

---

[1] Anotadores automáticos de funções sintáticas.

[2] O desempenho de um *parser* é, em grande parte, função da qualidade e quantidade de dados usados para seu treinamento. Por isso, um mesmo *parser* multilíngue pode apresentar desempenhos diferentes em línguas diferentes, sendo relativamente melhor em línguas que disponibilizam *corpora* maiores e mais consistentemente anotados.

[3] Pustejovsky et al. (2017) dizem que os *markables* são a parte do modelo que consome etiquetas (*consuming tags*) e os *non-markables* são a parte do modelo que não consome etiquetas (*non-consuming tags*).

1.3 Recursos essenciais aos *corpora*: diretrizes e manuais

As diretrizes em uma nova língua têm por objetivo tornar claro para os anotadores como o conjunto de etiquetas deve ser utilizado, com rica exemplificação que contemple desde casos comuns e frequentes até casos mais raros e difíceis de anotar.

As diretrizes da UD[4], por não serem específicas de língua, deixam grandes lacunas a serem preenchidas pelos linguistas que as instanciam numa língua. Um fórum mantido pela UD e disponível no *github*[5] mostra o quanto essa instanciação é desafiadora, pois centenas de tópicos são discutidos e diversas opiniões concorrem para solucionar as dúvidas.

Para o português, já foram disponibilizados no site da UD três *corpora*[6]: o PUD, o GSD e o Bosque-UD, este último descrito em Rademaker *et al*. (2017). É desejável que novos *corpora*, maiores e de novos domínios, sejam anotados seguindo o mesmo modelo, pois isso vai contribuir para a melhoria do desempenho dos *parsers* de português. Contudo, a falta de um manual de anotação dificulta as iniciativas de produção de novos *corpora* de UD em português, exigindo que cada novo projeto tenha que enfrentar o desafio de instanciar as diretrizes da UD para a língua portuguesa. Além disso, se cada projeto adota decisões diferentes para lidar com desafios iguais na língua, os *corpora* anotados deixam de ser comparáveis e isso pode impedir que sejam "somados" no esforço de treinar *parsers* mais robustos.

Foi essa dificuldade que enfrentamos ao iniciarmos um projeto de anotação de *corpus* seguindo o modelo UD: para guiar a anotação, precisávamos de um manual que contivesse explicações e exemplos de cada uma das etiquetas dos conjuntos de

[4] https://universaldependencies.org/guidelines.html
[5] https://github.com/universaldependencies/docs/issues
[6] https://universaldependencies.org/#download

etiquetas da UD, bem como discussões sobre como resolver ambiguidades que frequentemente geram dúvidas.

O relatório produzido após a anotação do *corpus* Bosque-UD (SOUZA *et al*., 2020) foi de grande auxílio para iniciarmos nosso trabalho, pois reporta diversos dos problemas enfrentados durante a anotação do esquema UD em língua portuguesa. Contudo, tal relatório não parece ter a pretensão de servir como um manual de anotação UD, pois não é organizado de forma a contemplar de forma sistemática cada uma das etiquetas dos conjuntos de etiquetas da UD.

Diante dessa lacuna, empreendemos a construção de um manual de anotação de *corpus* em UD, no qual estão instanciadas diretrizes da UD, englobando decisões de projeto e estudos acerca de palavras altamente ambíguas que, no decorrer do processo de anotação, se mostraram foco de dúvidas e discordâncias entre anotadores. Neste artigo, exploramos detalhadamente as questões relacionadas ao desenvolvimento desse manual.

O artigo está organizado em seis seções além desta introdução. Na Seção 2, fazemos uma breve revisão sobre a anotação sintática de *corpus* de língua portuguesa e sua relação com o desenvolvimento de *parsers* do português. Dedicamos a Seção 3 à discussão sobre os motivos que levaram a comunidade de PLN a consagrar a teoria de Tesnière (1959, 2015) como vantajosa para a anotação sintática. Na Seção 4, apresentamos o esquema de anotação da UD. Na Seção 5, apresentamos os conjuntos de etiquetas morfossintáticas (5.1) e de relações sintáticas (5.2) da UD, relacionando-as às classes morfossintáticas e funções sintáticas das gramáticas normativas do português. Na Seção 6, discutimos quais das diretrizes da UD são facilmente adotáveis no português, quais são problemáticas e exigiram tomada de decisão no projeto com base em fundamentos linguísticos e computacionais, e quais exigiram tomada de decisões arbitrárias, apenas para manter consistência na anotação, mas que poderão

ser alteradas no futuro caso uma solução melhor se apresente. Na Seção 7, tecemos considerações finais e discutimos trabalhos futuros.

2 *Parsers* e *treebanks* de português

Durante bastante tempo foi possível construir sistemas de PLN que prescindiam da análise sintática. Quase todo conhecimento era provido pelos itens lexicais e suas posições relativas (mesmo que, indiretamente, isso espelhasse a sintaxe). Para muitas aplicações mais simples e limitadas, como os primeiros corretores ortográficos, isso foi suficiente. Para outras aplicações mais avançadas, especialmente as que envolvem qualquer nível de interpretação do conteúdo, como, por exemplo, sistemas de pergunta e resposta, a análise sintática mostrou-se fundamental. À medida que se avançava nas tarefas de PLN, a análise sintática foi se tornando um ponto crítico, pois de sua qualidade dependia toda uma série de processamentos subsequentes.

Nos anos de 1990, os *parsers* estatísticos marcaram o início de uma nova era: a da geração de *parsers* por meio de aprendizagem automática. Para treinar esses *parsers*, eram necessários *corpora* anotados sintaticamente, de preferência revisados por humanos, pois isso garantiria que o aprendizado não se daria sobre desvios de anotação. A fim de facilitar o aprendizado automático, praticamente toda anotação adotava o formato de árvores, seja de constituintes, seja de dependências, e é por isso que os *corpora* anotados sintaticamente são chamados *treebanks*.

A estratégia era a seguinte: os *corpora* eram anotados com os *parsers* existentes, baseados em regras; em seguida, eram revisados por humanos e disponibilizados como exemplo para o aprendizado automático.

O primeiro *parser* para português largamente conhecido e utilizado em PLN tanto no Brasil quanto em Portugal foi o Palavras (BICK, 2000), e sua primeira versão comercial é de meados dos anos 1990. Ele é constituído por centenas de regras escritas

pelo autor e que seguem o paradigma de *Constraint Grammar* (KARLSSON, 1990). Inicialmente, o Palavras só gerava saída de árvores de dependências, mas a partir dos anos 2000 passou também a gerar uma saída de árvores de constituintes. E foi essa saída de constituintes a escolhida para pré-anotar o *corpus* Floresta Sintá(c)tica (AFONSO *et al*., 2002), primeira iniciativa bem-sucedida de gerar um *treebank* de língua portuguesa.

Na ocasião de seu lançamento, apenas 10% do *corpus* Floresta Sintá(c)tica havia sido revisado, porém todos os desvios encontrados na porção revisada serviram para melhorar as regras do *parser* Palavras. Como a parte não revisada do *corpus* Floresta foi reanotada com a versão aperfeiçoada do Palavras, pode-se dizer que a revisão beneficiou todo o *treebank*. Uma vez que o Floresta Sintá(c)tica possui partes totalmente revisadas, partes parcialmente revisadas e partes não revisadas, essas partes passaram a ter nomes distintos[7].

O Bosque é a parte do *corpus* Floresta Sintá(c)tica que foi totalmente revisada. Ele possui 9.368 sentenças, extraídas de *corpora* jornalísticos, das quais cerca de metade são de português brasileiro (CetenFolha) e metade do português europeu (Cetempúblico) e até hoje é o único *treebank* revisado que contém português brasileiro. O Selva, que teve suas árvores parcialmente revisadas, possui cerca de 30.000 sentenças e contém amostras de língua falada, bem como de textos literários e científicos, separadamente. O Floresta Virgem possui cerca de 96 mil sentenças não revisadas. E por fim, o Amazônia, anexado posteriormente, possui cerca de 275 mil sentenças não revisadas extraídas de *blogs* e, portanto, é uma porção do *treebank* que destoa das demais por ser constituída de conteúdos produzidos por usuários da Web e apresentar um registro de língua coloquial (FREITAS; ROCHA; BICK, 2008).

Em 2016, o Palavras passou a gerar também uma saída na forma de dependências em formato UD (BICK, 2016). Esse conversor de formato foi usado sobre

[7] https://www.linguateca.pt/Floresta/*corpus*.html

o *corpus* Bosque e o resultado foi a base para a revisão manual que culminou na produção do *corpus* Bosque-UD (RADEMAKER *et al*., 2017). O *corpus* Bosque-UD, disponibilizado no site da UD[8], passou a ser utilizado para treinar *parsers* de dependência que utilizam técnicas independentes de língua[9], que já vinham apresentando alta capacidade de aprendizado em outras línguas (como os já citados UDPipe, UDify, SpaCy e Stanza). A vantagem desses *parsers* é que são livres (enquanto o Palavras não é) e, por isso, podem ser retreinados por outras equipes, com novos *corpora* e em diferentes gêneros.

Outros *parsers* e *treebanks* surgiram ao longo dos últimos anos, principalmente por iniciativa da comunidade linguística portuguesa, como o LX-Parser (SILVA *et al*., 2010) e o treebank Cintil, mas o Palavras continuou sendo muito utilizado. Seu papel no processamento do português é inegável e, graças a ele, pôde-se fazer a transição entre um *parser* baseado em regras para os *parsers* baseados em AM. Da mesma forma, é inegável o papel do *corpus* Bosque, primeiro *treebank* do português cujas árvores sintáticas foram totalmente revisadas por humanos e, por isso, ideal para ser usado para aprendizado automático.

Embora no site da UD hoje existam três *corpora* já disponibilizados em formato UD, apenas o Bosque é constituído de sentenças originalmente produzidas em português. O *corpus* PUD tem apenas 1.000 sentenças e é uma versão em português do *corpus* paralelo produzido para a competição CoNLL 2017 de *parsers* multilíngues[10]. O *corpus* GSD tem 12.078 sentenças da variante brasileira do português, resultado da conversão para UD da anotação do *corpus* do Google, um *corpus* paralelo de traduções para várias línguas.

---

[8] https://lindat.mff.cuni.cz/services/udpipe/

[9] Esses *parsers* são dependentes de modelo: eles só aprendem a partir de *corpus* anotado no modelo UD, e desde que uma língua tenha um *corpus* anotado no modelo UD, eles podem ser treinados para se tornarem um parser para essa língua.

[10] *CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies* (http://universaldependencies.org/conll17/).

Quando comparamos o tamanho e a variedade de gêneros dos *treebanks* que temos com os dos *treebanks* disponibilizados nas línguas mais bem providas de recursos de PLN, vemos que há muito a ser feito no português e, especialmente, no português do Brasil.

3 O PLN e a estrutura sintática de dependências

As árvores sintáticas de dependência concorreram alternativamente com as árvores de constituintes durante muito tempo no PLN. No entanto, quando passou-se a utilizar árvores sintáticas para comparar línguas automaticamente, percebeu-se que as árvores de dependência eram mais adequadas. Isso se deve principalmente ao fato de que a estrutura de dependências permite anotar relações sintáticas entre as palavras independentemente da posição em que elas ocorrem na oração, ao passo que a estrutura de constituintes é mais rígida quanto à ordem de realização de seus constituintes.

Essa característica da estrutura de dependências é fundamental quando o objetivo é relacionar línguas posicionais (línguas como o Português, em que a ordem dos constituintes é relevante, pois o papel sintático está em grande parte atrelado à posição das palavras na oração) a línguas desinenciais (línguas como o Latim, em que a ordem dos constituintes da oração é livre, pois o papel sintático das palavras é marcado por desinências).

Além disso, em termos de técnicas de AM, a estrutura de dependências permite um aprendizado mais lexicalizado (por ser apoiado em tokens e não em sintagmas), o que tem se mostrado vantajoso em várias aplicações de PLN.

Por esses motivos, as gramáticas de dependência (inspiradas na proposta de Tesnière (1959)), passaram a ser revalorizadas como um modelo promissor para o PLN.

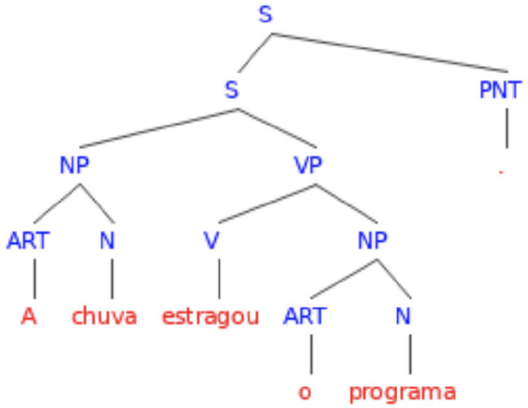
Tesnière (1959) rejeitava a divisão dualista da estrutura sintática em sujeito e predicado. Para ele, o verbo é a raiz de toda estrutura sintática e o sujeito e o objeto são seus subordinados. Ele percebeu que uma sentença não era composta apenas das palavras que a integravam (elementos explícitos), mas também de relações implícitas entre essas palavras, relações que tinham uma hierarquia (um governante e um subordinado) e uma direção (que marca qual elemento está governando a relação). Para explicitar essas relações de dependência, Tesnière (1959) introduziu a ideia de representar a sintaxe por meio de uma árvore, que ele chamava de *stemma*, na qual estão representados todos os elementos de uma sentença e todas as dependências que ligam esses elementos em uma única estrutura.

As árvores de dependência são o cerne do esquema de anotação da UD, cujo objetivo é, em última instância, subsidiar a anotação de *corpora* que serão insumo para desenvolver aplicações de PLN multilíngues. A iniciativa UD motivou, assim, a disseminação tardia do trabalho de Tesnière em língua inglesa (TESNIÈRE[11], 2015).

Para fins de ilustração, apresentamos um exemplo de representação da sintaxe por meio de árvore de constituintes (Figura 1) e um exemplo de representação por árvores de dependências (Figura 2). Contrastando as duas figuras, é possível perceber que a segunda é mais simples, pois contém menos nós no total, relacionando mais diretamente os tokens.

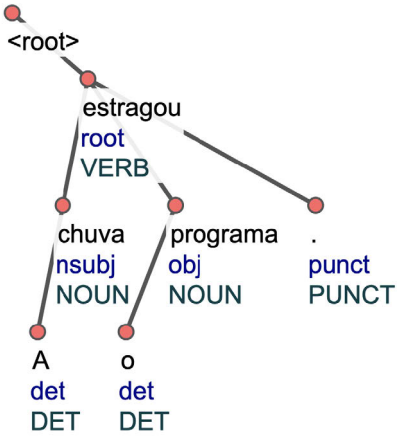
[11] A tradução de 2015 para o inglês é comentada à luz do emprego atual da teoria.

Figura 1 — Análise de "A chuva estragou o programa" sob forma de árvore de constituintes.



Fonte: árvore sintática produzida pelo parser LX-Parser[12].

Figura 2 — Análise de "A chuva estragou o programa" sob forma de árvore de dependências.



Fonte: árvore sintática em formato UD produzida no software UDPipe LINDAT/CLARIN[13].

4 O modelo UD

O modelo de anotação UD é lexicalista, o que significa que todas as etiquetas são associadas diretamente aos *tokens*. Em uma relação de dependência, apenas um *token* é o *head* (governante) e apenas um *token* é o dependente. Todos os *tokens*, sem exceção, recebem uma etiqueta morfossintática e participam de pelo menos uma relação de dependência. A UD define que apenas as palavras de conteúdo (substantivos, pronomes substantivos, verbos, adjetivos, advérbios e numerais)

[12] https://portulanclarin.net/workbench/lx-parser/
[13] http://lindat.mff.cuni.cz/services/udpipe/run.php

podem ser *head* de relações (e podem ser dependentes de relações também). Já as palavras funcionais (como auxiliares, preposições, determinantes e conjunções), símbolos e sinais de pontuação só podem ser dependentes de relações.

O *corpus* a ser anotado deve ser previamente processado para: 1) fazer a separação do texto em sentenças (pois cada sentença é anotada individualmente) e 2) fazer a tokenização, que inclui separar os sinais de pontuação colados às palavras (exceto: pontos das abreviações, pontos separadores das milhares, vírgulas das casas decimais dos numerais e hifens das palavras compostas), separar os verbos dos pronomes enclíticos e mesoclíticos, eliminando os hifens que os juntavam e separar os *tokens* que constituem contrações na língua (ex: deste = de este, comigo = com mim, naquela = em aquela). As palavras compostas (unidas por hífen) são aceitas normalmente como *tokens* da língua, mas *tokens* que integram multipalavras, como "panela de pressão" e "energia solar", são anotados como *tokens* independentes. A UD não trabalha com o conceito de multipalavras no nível morfossintático e, no nível sintático, tem três relações de dependência para anotar apenas as multipalavras que não apresentam relações sintáticas entre os *tokens* que as integram.

A anotação no esquema UD segue o formato definido na conferência Computational Natural Language Learning (CoNLL) e por isso é chamado CoNLL-U (U para "universal"). Várias ferramentas de anotação de interface gráfica e amigável estão disponíveis para anotar nesse formato[14], não representando, portanto, uma dificuldade para os projetos. É importante, contudo, entender as partes que constituem esse formato. Apesar de o objetivo ser a anotação de relações de dependência sintática, o CoNLL-U permite que se registre muito mais informações além dessas relações. Isso porque, quanto mais rica a anotação, mais atributos podem contribuir para a distinção entre as etiquetas no aprendizado automático.

[14] https://universaldependencies.org/tools.html

O formato CoNLL-U é constituído por uma tabela de 10 colunas com informações associadas aos tokens nas linhas. As colunas mais importantes para o anotador são a quarta, que se chama UPOS (preenchida com uma das 17 etiquetas morfossintáticas da UD), e a oitava, que se chama DepRel (preenchida com uma das 37 etiquetas de relações de dependência sintática da UD). As demais colunas contêm: o identificador de posição de cada token, a forma, o lema, os atributos morfológicos (flexões, por exemplo), o *head* da relação de dependência e três colunas de preenchimento opcional. Esse formato, ao mesmo tempo que restringe o conteúdo da maioria das colunas, caracterizando a anotação segundo a abordagem UD, prevê algum espaço para que também sejam anotadas características específicas de cada língua (a quinta coluna, XPOS, é reservada para etiquetas morfossintáticas específicas da língua e a décima coluna, MISC, é de uso livre). A Tabela 1 mostra um exemplo de sentença no formato CoNLL-U.

Tabela 1 - Exemplo de sentença anotada no formato CoNLL-U.

# text = **O estúdio também divulgou novo poster de o filme .**									
id.	*forma*	*lema*	*UPOS*	*XPOS*	*Atributos morfológicos*	*head*	*DelRel*	*Deps*	*MISC*
1	**O**	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	2	det	_	_
2	**estúdio**	estúdio	NOUN	_	Gender=Masc|Number=Sing	4	nsubj	_	_
3	**também**	também	ADV	_	_	4	advmod	_	_
4	**divulgou**	divulgar	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	0	root	_	_
5	**novo**	novo	ADJ	_	Gender=Masc|Number=Sing	6	amod	_	_
6	**poster**	poster	NOUN	_	Foreign=Yes|Gender=Masc|Number=Sing	4	obj	_	_
7	**de**	de	ADP	_	_	9	case	_	_
8	**o**	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	9	det	_	_
9	**filme**	filme	NOUN	_	Gender=Masc|Number=Sing	6	nmod	_	_
10	**.**	.	PUNCT	_	_	4	punct	_	SpacesAfter=\s\s\n

5 A relação entre o esquema UD e a Gramática Normativa

Olhando apenas para o conjunto de 17 etiquetas morfossintáticas (*part-of-speech tags* ou simplesmente *PoS tags*) e de 37 relações sintáticas da UD, poderíamos nos guiar pelos conhecimentos adquiridos com o estudo das Gramáticas Normativas e arriscar-nos a fazer a atribuição de grande parte delas às palavras do texto, pois os nomes das etiquetas muitas vezes coincidem com termos utilizados na Nomenclatura Gramatical Brasileira (HENRIQUES, 2009). No entanto, as diretrizes da UD nos mostram que a

interpretação dessas etiquetas nem sempre é simples de ser feita a partir apenas de conhecimentos prévios de gramáticas. Para exemplificar como os conceitos das etiquetas da UD diferem dos conceitos largamente difundidos pelas gramáticas normativas, abordaremos as etiquetas morfossintáticas e as etiquetas de relações de dependência em duas seções diferentes (da mesma forma como abordamos em dois manuais separados as diretrizes para atribuição desses dois conjuntos de etiquetas).

Embora apresentadas em seções diferentes, é importante destacar que na UD há alta correlação entre as etiquetas morfossintáticas e as etiquetas de relações sintáticas. Na Seção 5.2 mostraremos como essa correlação se traduz num conjunto de restrições que as relações sintáticas impõem sobre as etiquetas morfossintáticas.

Ressaltamos que este artigo exemplifica apenas algumas das questões contempladas nas diretrizes que integram nossos manuais de anotação. Os manuais foram elaborados a partir de experiências iniciais de anotação e, como são geridos pelas mesmas pessoas que anotam o *corpus*, sofrem manutenções periódicas com vistas a refletir novos fatos observados no *corpus* e que ainda não tenham sido contemplados nas diretrizes[15].

5.1 Etiquetas morfossintáticas da UD

As etiquetas morfossintáticas padrões da UD são:

- Para classes abertas: ADJ (adjetivos), ADV (advérbios), INTJ (interjeições), NOUN (substantivos), PROPN (nomes próprios) e VERB (verbos);
- Para classes fechadas: ADP (preposições), AUX (verbos auxiliares), CCONJ (conjunções coordenativas), DET (determinantes), NUM (numerais cardinais), PART (partículas), PRON (pronomes) e SCONJ (conjunções subordinativas);

[15] A versão do manual divulgada neste artigo baseia-se nas diretrizes da UD vigentes em abril de 2022. Novas diretrizes da UD que venham a ser divulgadas serão in*corpora*das em futuras versões do manual.

- Outros tokens que não se enquadram nas classes abertas e fechadas: PUNCT (pontuações), SYM (símbolos) e X (demais casos, incluindo onomatopeias e palavras em língua estrangeira).

As principais diferenças entre as etiquetas morfossintáticas da UD e as classes morfossintáticas da gramática normativa do português brasileiro são:

- DET: a UD trabalha com o conceito de determinante e sob essa etiqueta reúne os artigos e os pronomes não nominais (demonstrativos e possessivos);
- NUM: apenas os numerais cardinais devem ser anotados como NUM, enquanto os numerais ordinais devem ser anotados como ADJ;
- PRON: apenas os pronomes nominais são anotados com essa etiqueta, e os demais pronomes, desde que estejam modificando um nominal, devem ser anotados como DET;
- NOUN: apenas os substantivos comuns são anotados sob essa etiqueta, pois os nomes próprios são anotados como PROPN;
- VERB: apenas os verbos considerados plenos e passíveis de serem classificados como predicados verbais devem ser anotados com essa etiqueta.
- AUX: verbos auxiliares e verbos de cópula "altamente gramaticalizados", ou seja, sem carga semântica significativa, devem ser anotados com essa etiqueta;
- PROPN: etiqueta destinada a anotar nomes próprios, desde que não coincidam com palavras comuns da língua.

Em nosso manual de anotação de *PoS tags* (etiquetas morfossintáticas) cada etiqueta é tratada em uma seção, com definição, exemplos, léxico das palavras (no caso

de classes fechadas) e diretrizes de desambiguação em relação a outras etiquetas que costumam ser confundidas com ela em determinados contextos. A ambiguidade entre etiquetas morfossintáticas é, em grande parte, dependente de língua, e por isso não é contemplada nas diretrizes da UD. Por exemplo, a palavra "menos", antes de um substantivo, com o sentido de "menor quantidade de", é um ADJ, pois modifica um NOUN. Quando, porém, a palavra "menos" modifica um VERB, um ADJ ou um ADV, com o sentido de "com menor intensidade" ou "em menor grau" ela é anotada como ADV. Há contextos, porém, em que "menos" ocorre entre um VERB e um NOUN e o anotador desavisado pode se confundir na atribuição. Exemplo:

O plano favoreceu **menos** classes D e E do que classe média. (aqui "menos" é um ADV que modifica o verbo "favorecer": "favoreceu em menor grau as classes D e E")

O plano atual favoreceu **menos** pessoas que o plano anterior. (aqui "menos" é um ADJ e modifica o substantivo "pessoas": "uma menor quantidade de pessoas").

5.2 Relações de dependência da UD

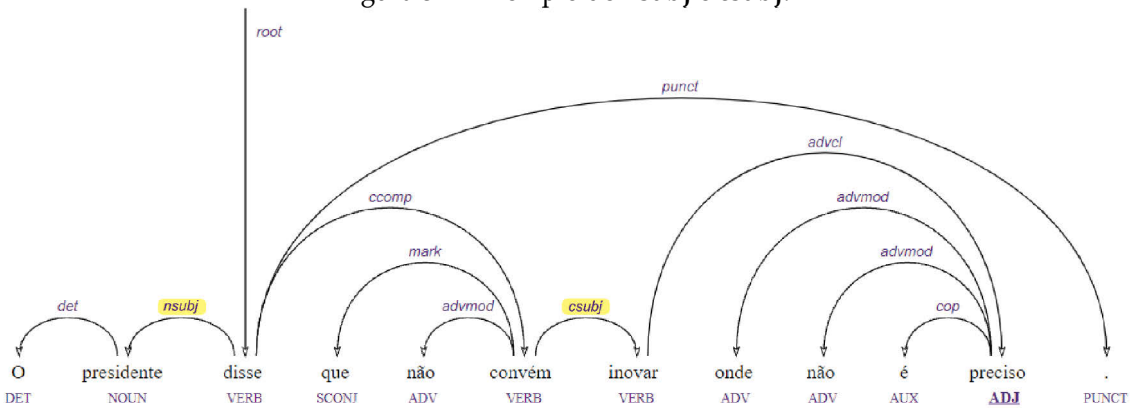
A UD possui um conjunto de 37 etiquetas para marcar as relações de dependência[16] e o descasamento entre essas relações e os componentes da análise sintática descritos nas gramáticas normativas é ainda maior do que o observado no

[16] Um quadro completo das relações de dependência do UD está disponível em https://universaldependencies.org/u/dep/index.html

conjunto de etiquetas morfossintáticas. A seguir serão discutidas as principais diferenças entre o esquema de anotação sintática da UD e as funções sintáticas que constam das gramáticas normativas.

A UD tem relações separadas para as funções sintáticas preenchidas por orações (tanto predicados verbais quanto nominais) e as mesmas funções sintáticas preenchidas por palavras núcleos de sintagmas. Assim, por exemplo, um sujeito pode ser **nsubj**, se for um nominal, ou **csubj**, se for constituído por uma oração, conforme pode ser observado na Figura 3.

Figura 3 — Exemplo de **nsubj** e **csubj**.

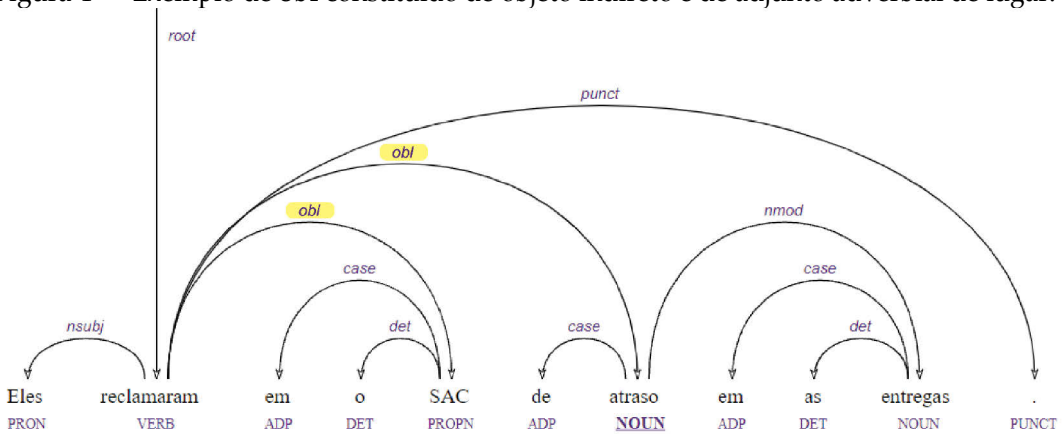


Fonte: árvore sintática em formato UD produzida no software Arborator-Grew[17].

Na UD, os adjuntos adverbiais cujo núcleo é um NOUN (em sua maioria introduzidos por preposição) não são discriminados dos objetos indiretos introduzidos por preposição, pois ambos são considerados modificadores verbais do tipo nominal e anotados com a relação **obl** (ilustrados na Figura 4).

[17] https://arborator.icmc.usp.br/#/

Figura 4 — Exemplo de **obl** constituído de objeto indireto e de adjunto adverbial de lugar.

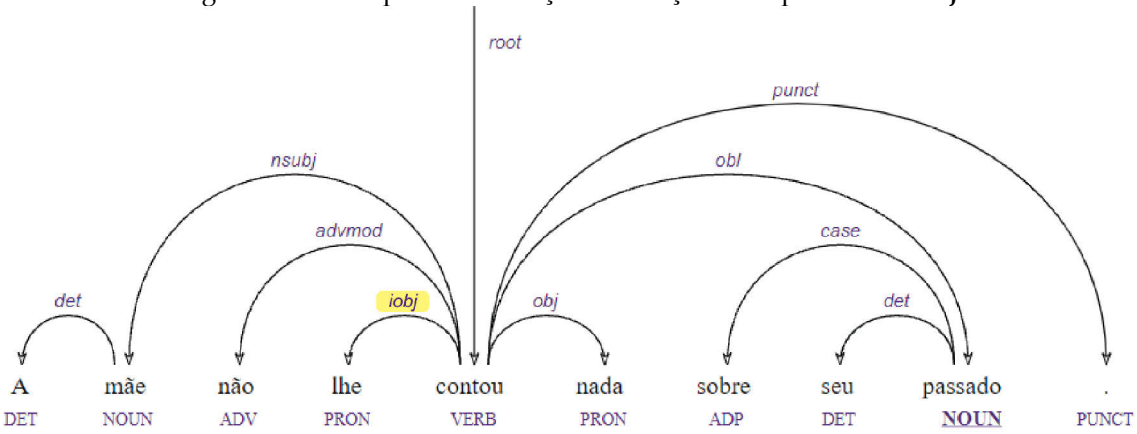


Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

Talvez uma das relações da UD mais fáceis de serem mal interpretadas seja **iobj** (objeto indireto). Essa relação só deve ser atribuída a objetos indiretos não preposicionados, o que é um contrassenso se pensarmos no que significa "indireto" no nome dessa função sintática. Esse é um resquício das origens das relações na anotação em língua inglesa, pois o inglês tem casos de dois objetos não preposicionados[18] e ambos podem ser alçados à posição de sujeito na diátese de voz passiva. No português, só devemos utilizar a relação **iobj** para anotar objetos indiretos na forma de pronomes oblíquos (*me, te, se, lhe, nos, vos, lhes*), como mostrado na Figura 5.

[18] Dois objetos não preposicionados podem ocorrer nos verbos dativos do inglês. Em "John sent Mary a message", ambos os objetos podem ser alçados a sujeito da passiva: "A message was sent to Mary" e "Mary was sent a message".

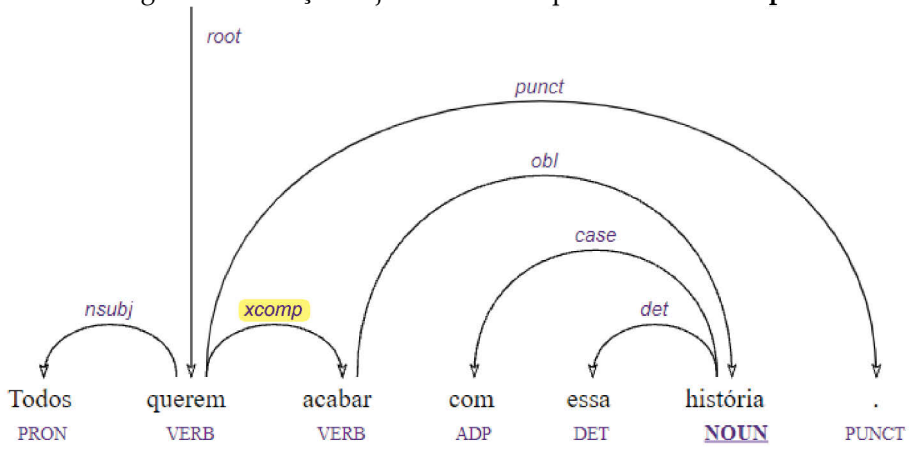
Figura 5 — Exemplo de atribuição da relação de dependência **iobj**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

Na UD, objetos diretos são anotados com a relação **obj**, mas, se forem realizados por meio de oração, recebem dois tipos de anotação diferentes. As orações objetivas diretas que têm sujeito NULL[19] governado pelo sujeito ou pelo objeto da oração matriz, são anotadas como dependentes da relação **xcomp** (Figura 6), e as que têm sujeito realizável (embora possa estar elíptico), são anotadas com a relação **ccomp** (Figura 7).

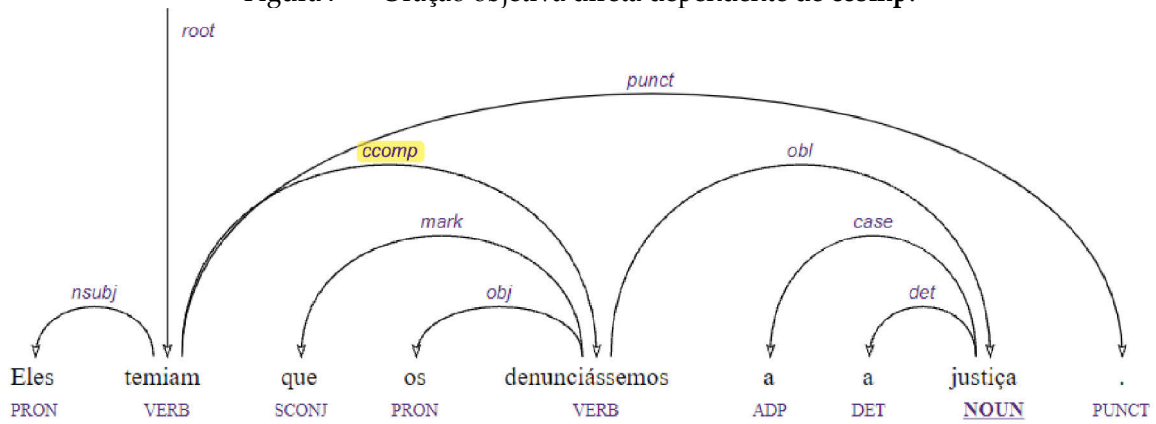
Figura 6 — Oração objetiva direta dependente de **xcomp**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

[19] Para um aprofundamento sobre o conceito de sujeito NULL e xcomp, recomendamos a leitura de Bresnan (1982).

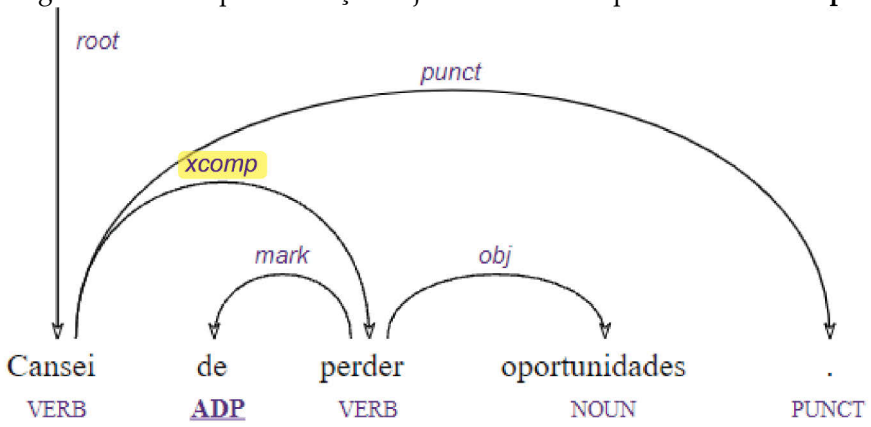
Figura 7 — Oração objetiva direta dependente de **ccomp**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew

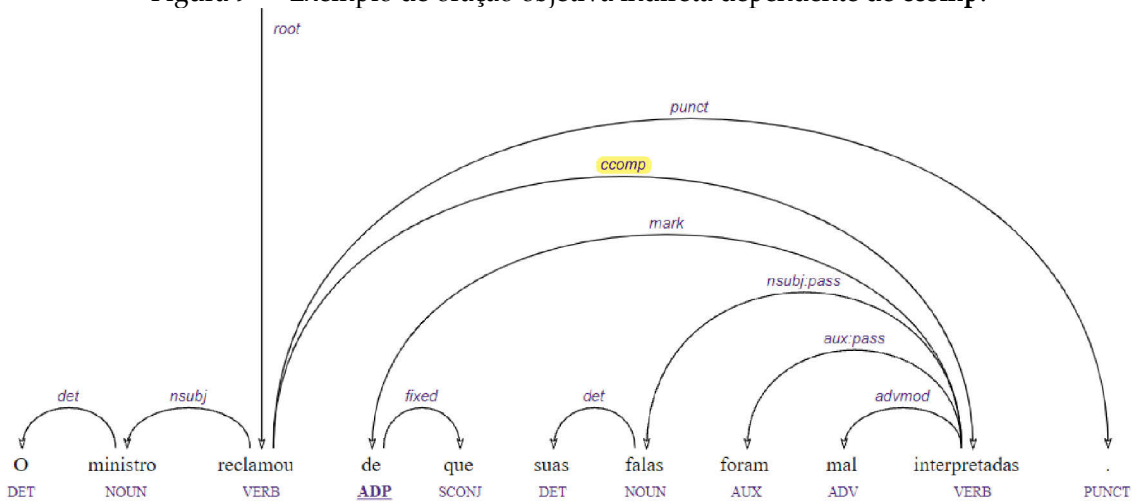
O mesmo ocorre com os objetos indiretos realizados sob forma de oração (subordinada substantiva objetiva indireta): se a oração subordinada tiver sujeito NULL governado pelo sujeito ou pelo objeto da oração matriz, será dependente de **xcomp** (Figura 8) e, se tiver sujeito realizável, será dependente de **ccomp** (Figura 9).

Figura 8 — Exemplo de oração objetiva indireta dependente de **xcomp**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

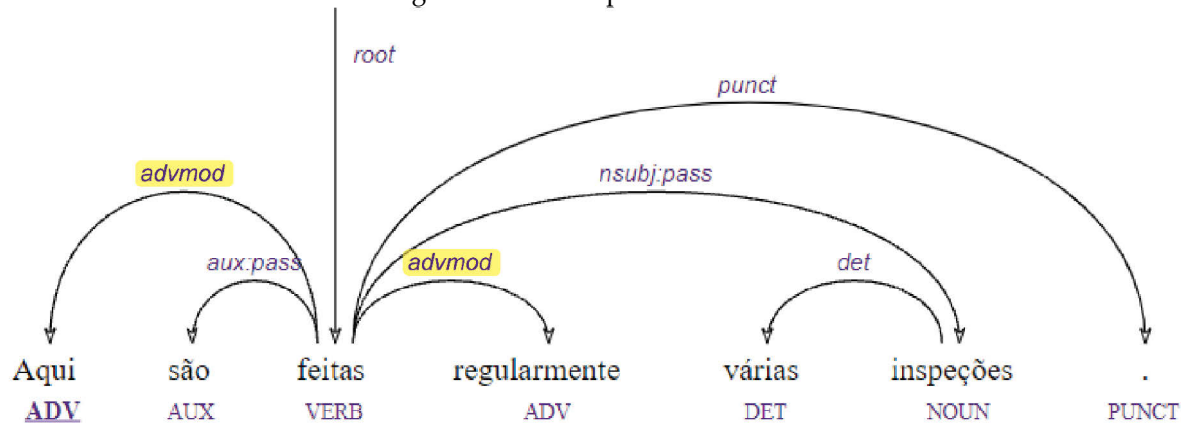
Figura 9 — Exemplo de oração objetiva indireta dependente de **ccomp**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew

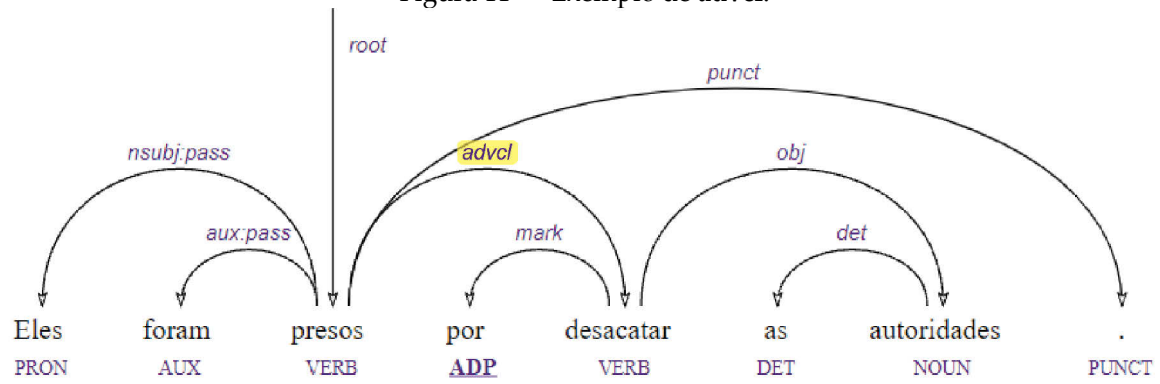
Os modificadores constituídos por advérbios (ADV) são anotados com a relação **advmod** (Figura 10). Porém, em se tratando de uma oração adverbial, a relação será **advcl** (Figura 11).

Figura 10 — Exemplo de **advmod**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew

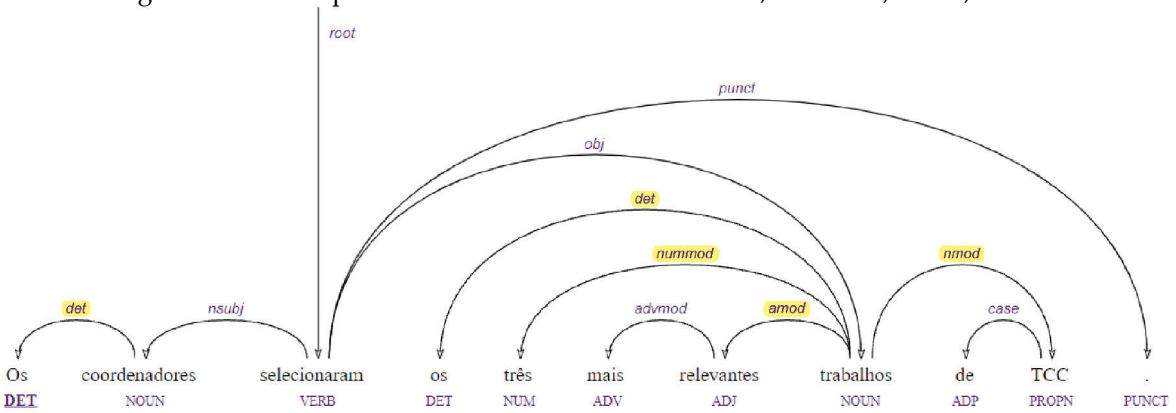
Figura 11 — Exemplo de **advcl**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew

A UD também tem uma granularidade maior para os modificadores nominais do que as gramáticas normativas. Os modificadores nominais podem ser: **amod** (para ADJ), **nmod** (para NOUN e PROPN) e **nummod** (para NUM). As preposições (ADP) que introduzem um **nmod** são ligadas pela relação **case** ao nominal *head* da relação. Os determinantes (DET) são ligados aos nominais que determinam por meio da relação **det**, como pode ser observado na Figura 12.

Figura 12 — Exemplos de modificadores nominais: det, nummod, amod, nmod.

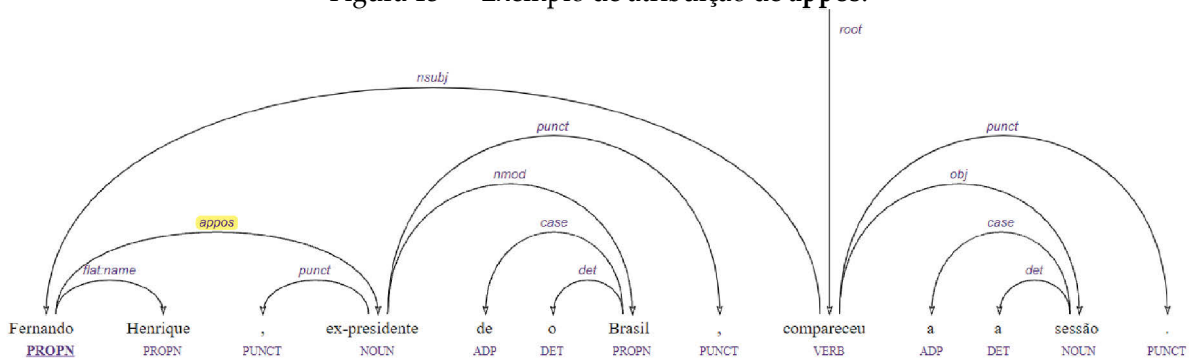


Fonte: árvore sintática em formato UD produzida no software Arborator-Grew

Se, contudo, um modificador nominal estiver sob forma de oração, a relação de dependência será sempre **acl**, pois a UD não distingue orações completivas nominais de orações adjetivas.

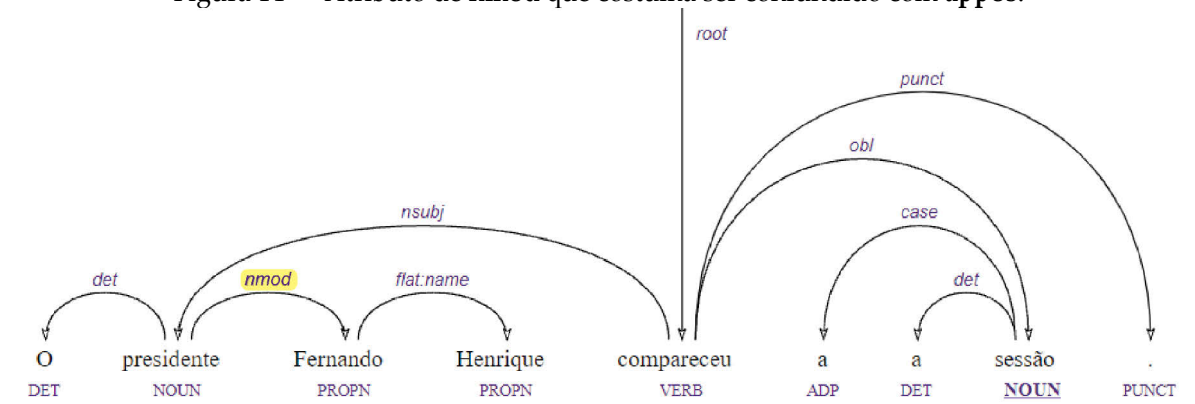
Outra diferença que merece ser destacada é o conceito de aposto na UD. Segundo as instruções da UD, só deve ser anotada como **appos** a relação entre dois elementos intercambiáveis, que sejam independentes e que possam ser trocados de ordem sem prejuízo para a gramaticalidade da sentença, como na Figura 13. Casos como o mostrado na Figura 14, contudo, tradicionalmente tratados como aposto, são anotados como um atributo na UD, **nmod**.

Figura 13 — Exemplo de atribuição de **appos**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

Figura 14 — Atributo de **nmod** que costuma ser confundido com **appos**.

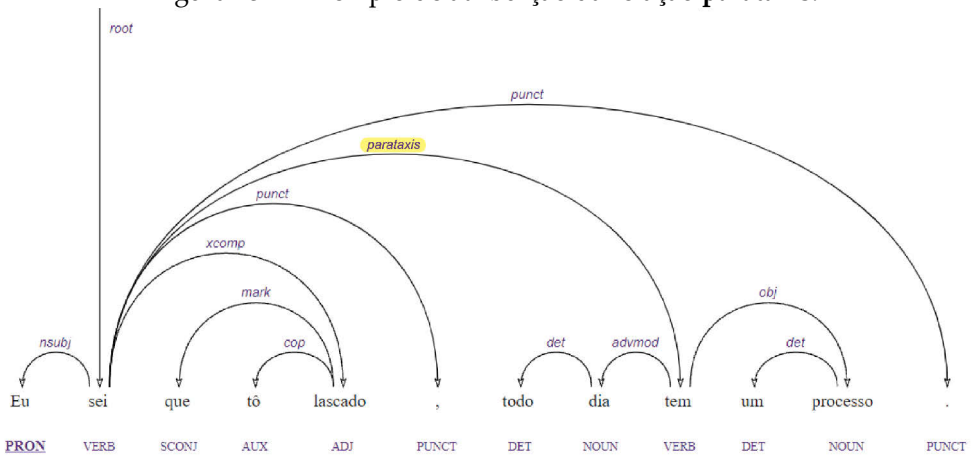


Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

Além das relações de dependência associadas a funções sintáticas da gramática normativa, a UD prevê uma série de relações para ligar os elementos da sentença que não têm relação sintática com outros elementos, como as relações **parataxis**, **discourse**, **reparandum**, **dislocated** e **intj**.

A relação de **parataxis** (ilustrada na Figura 15), por exemplo, serve para ligar duas orações sem coesão lógica explícita dentro da mesma sentença.

Figura 15 — Exemplo de atribuição da relação **parataxis**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

As diretrizes da UD apresentam também três relações de dependência para anotar multipalavras que não possuem relação sintática entre si (**compound**, **flat** e **fixed**). O uso dessas etiquetas deve ser parcimonioso, pois elas simplesmente atestam que as palavras ligadas por elas não possuem nenhuma relação sintática entre si, o que as torna opacas para que o *parser* aprenda sintaxe com elas. Por exemplo:

João Paulo - PROPN PROPN ligados pela relação **flat**
além de - ADV ADP ligados pela relação **fixed**
ir embora - VERB ADV ligados pela relação **compound**.

Não se pode dizer, portanto, que as diretrizes da UD determinem a anotação de todas as multipalavras, pois aquelas que possuem relações sintáticas entre os tokens que as compõem são anotadas com relações comuns. Por exemplo:

efeito estufa é um NOUN NOUN ligado pela relação **nmod**, cujo *head* é *efeito* e o dependente (modificador) é *estufa*
protetor solar é um NOUN ADJ ligado pela relação **amod**, cujo *head* é *protetor* e o dependente (modificador) é *solar*
casa de câmbio é um NOUN ADP NOUN ligado pelas relações **nmod** (entre *casa* e *câmbio*) e **case** (entre *câmbio* e *de*)

Algo que deve ser observado nos exemplos apresentados nesta seção é o fato de que algumas relações sintáticas só admitem dependentes que tenham uma determinada etiqueta morfossintática: **det** só admite DET, **case** só admite ADP, **advmod** só admite ADV, **nmod** só admite NOUN ou PROPN, **amod** só admite ADJ, **nummod** só admite NUM, **aux** só admite AUX, **cop** só admite AUX, **punct** só admite PUNCT, **intj** só admite INTJ e **mark** só admite SCONJ e ADP. Isso ilustra como a UD restringiu automaticamente a anotação, impedindo a ocorrência de grande parte dos erros de atribuição de etiquetas.

6 Os desafios da instanciação das diretrizes UD e da construção do manual de anotação

A tarefa de produzir os dois manuais de anotação (de etiquetas morfossintáticas e de etiquetas de relações dependência) iniciou-se com a leitura minuciosa das diretrizes da UD, complementada pela leitura de outras fontes de consulta sobre a atribuição das etiquetas do esquema UD, como artigos e fórum de discussão disponível no *github* da UD. Realizamos o trabalho em duas etapas , uma vez que a anotação das etiquetas morfossintáticas seria realizada em uma fase anterior à fase de anotação de relações de dependência (por isso organizamos as diretrizes em dois manuais).

Antes de iniciarmos a anotação de cada fase, já tínhamos uma versão preliminar dos respectivos manuais, os quais foram utilizados no treinamento dos anotadores. Essas versões preliminares dos manuais foram elaboradas por uma linguista e revisadas por uma pesquisadora de PLN (não linguista) que participa do processo de anotação. A equipe de anotadores foi ampliada na sequência, e as pessoas responsáveis pela versão preliminar do manual continuaram na equipe, participando da anotação.

Durante cada fase do processo de anotação, as necessidades de melhoria das definições e de maior detalhamento das instruções tornaram-se evidentes. Utilizamos as divergências entre os anotadores como subsídio para aperfeiçoar os manuais, assim como utilizamos exemplos do *corpus* para ilustrar questões que se mostraram desafiadoras para a anotação.

Outra fonte de consulta foram os *corpora* já anotados em português e em outras línguas com o modelo UD. Para isso, utilizamos o buscador Grew-Match[20], que permite buscas por palavras, por lemas, por etiquetas morfossintáticas e por relações de dependência. Essa prática foi importante tanto para descobrirmos soluções adotadas em outros *corpora* que poderiam ser aplicáveis no nosso *corpus* quanto para percebermos que 1) diferentes *corpora* de uma mesma língua às vezes adotam formas diferentes de atribuir uma mesma etiqueta ou relação; 2) *corpora* de diferentes línguas adotam, eventualmente, etiquetas e relações diferentes para anotar fenômenos semelhantes e 3) muitas vezes um padrão não é reconhecido pelos anotadores e ele acaba sendo anotado de diferentes maneiras ao longo do mesmo *corpus*, gerando inconsistência.

Durante a confecção dos manuais, nos deparamos com três tipos de situações: 1) aquelas em que a orientação da UD era clara e facilmente aplicável no português; 2) aquelas em que havia lacunas na orientação da UD, mas que, seguindo princípios da linguística, conseguimos preencher com decisões claras; e 3) aquelas em que as

[20] http://match.grew.fr/

diretrizes da UD não eram claras ou apresentavam lacunas e a fundamentação linguística não era suficiente para garantir uma decisão que pudesse ser considerada "correta", situações em que tivemos que tomar decisões arbitrárias a fim de garantir consistência na anotação.

Por exemplo, os *corpora* de UD de português anotam quantidades diferentes de verbos com AUX: o PUD anota 13, o GSD anota 59 e o Bosque anota 6 verbos[21]. Nas outras línguas, há grande variação na atribuição dessa etiqueta também. O inglês, por exemplo, anota modais como *would, should, can, could, may, might* e *must*, enquanto o francês anota também o causativo *faire* como AUX.

A questão dos verbos auxiliares e de cópula foi uma das que exigiram que tomássemos uma decisão arbitrária (e provisória até que solução melhor se apresente), já que as diretrizes da UD deixam a cargo de cada língua fazer essa decisão. A única recomendação da UD é a de que só devem ser anotados como AUX os verbos auxiliares altamente gramaticalizados e os verbos de cópula esvaziados de semântica. Atualmente, estamos anotando como AUX apenas os auxiliares de tempo e de voz passiva e os verbos de cópula *ser* e *estar*.

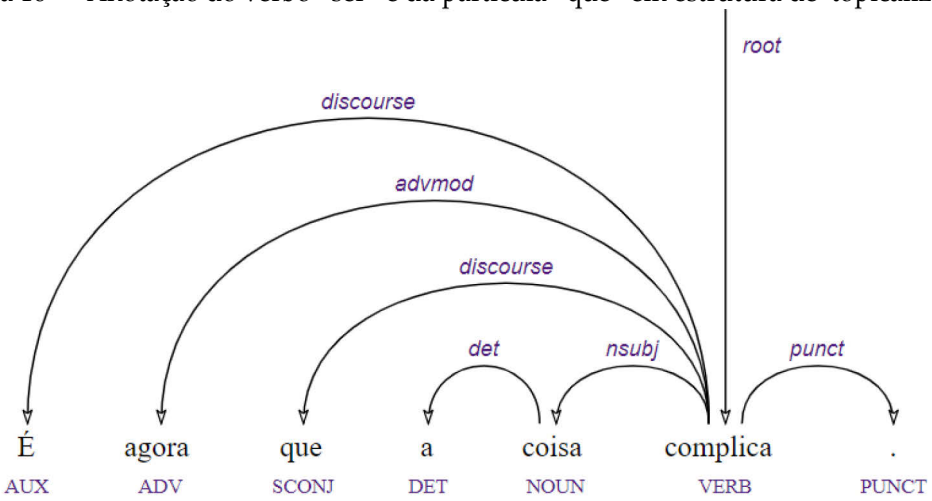
Já uma questão que exigiu uma decisão baseada em critérios linguísticos foi a anotação das preposições que introduzem orações reduzidas de infinitivo como ADP e não como SCONJ. A UD só diz que SCONJ é usada para *complementizers* (no português, *que* e *se*) e palavras que introduzem orações adverbiais (como *quando*, nas temporais, e *como*, nas causais e conformativas). Segundo Azeredo (2013, item 291) as conjunções subordinativas são utilizadas para introduzir orações desenvolvidas e não orações reduzidas, o que nos permitiu inferir, portanto, que as preposições que introduzem orações subordinadas reduzidas de infinitivo devem continuar sendo anotadas como ADP, embora a respectiva relação sintática seja **mark**. Essa anotação pode beneficiar futuramente tarefas de anotação de papéis semânticos, pois as

[21] Para uma comparação detalhada: https://universaldependencies.org/treebanks/pt-comparison.html

preposições que introduzem nominais ou orações no infinitivo são importantes pistas para discriminar os papéis semânticos na estrutura argumental de um verbo.

A anotação do verbo *ser* e da conjunção *que* com função de topicalização foi um caso em que reproduzimos a decisão arbitrária tomada no *corpus* Bosque-UD: anotamos o verbo como AUX, a conjunção como SCONJ e a relação de ambos com o predicado como **discourse**.

Figura 16 — Anotação do verbo "ser" e da partícula "que" em estrutura de topicalização.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

Um caso em que não seguimos cegamente as diretrizes da UD foi o da anotação dos nomes próprios. A UD prescreve que os nomes próprios que constituem substantivos e adjetivos comuns da língua sejam anotados com as etiquetas próprias de suas categorias de origem e sua relação sintática seja anotada com as relações sintáticas usadas para palavras comuns. Assim, *Maria das Dores* seria composto por PROPN, ADP, DET e NOUN, e *Companhia de Água e Esgoto* por NOUN, ADP, NOUN, CCONJ e NOUN. No entanto, fomos conservadores e mantivemos como PROPN todas as palavras grafadas em letra maiúscula e as ligamos com a relação **flat:name**, pois não queríamos limitar tarefas de PLN que dependem do resultado do *parser*, como o reconhecimento de entidades nomeadas.

7 Considerações finais e trabalhos futuros

Conforme exposto, o trabalho de construir o manual de anotação do modelo UD em *corpus* do português exigiu bastante esforço e constantes aperfeiçoamentos. Por razões práticas e didáticas, separamos as diretrizes em dois manuais, um para anotação de etiquetas morfossintáticas (Manual de Anotação de *PoS tags*) e outro para relações de dependência (Manual de Anotação de Relações de Dependência). Esses dois manuais, que se completam, estão disponíveis no site do projeto POeTiSA[22], e poderão abreviar o esforço necessário para novos empreendimentos de anotação sintática de *corpus* seguindo o modelo UD. No mesmo site serão publicadas novas versões desses manuais, sempre que manutenções no conteúdo se façam necessárias.

Como nosso projeto prevê a anotação de outros *corpora*, em diferentes domínios, é de se esperar que novas diretrizes venham a ser somadas ao material existente, na forma, por exemplo, de apêndices dedicados a domínios específicos.

Agradecimentos

Os autores agradecem o apoio do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI-http://c4ai.inova.usp.br/), financiado pela IBM e pela FAPESP (processo#2019/07665-4).

Referências Bibliográficas

AFONSO, S.; BICK, E.; HABER, R.; SANTOS, D. Floresta sintá(c)tica: a treebank for Portuguese. *In*: RODRÍGUEZ, M. G.; ARAUJO, C. P. S. (ed.), **Proceedings of the Third International Conference on Language Resources and Evaluatio**n (LREC 2002). European Language Resources Association, 2002. p. 1698-1703.

AZEREDO, J. C. S. **Fundamentos de Gramática do Português**. Rio de Janeiro: Jorge Zahar Editores, 2013. E-book.

[22] https://sites.google.com/icmc.usp.br/poetisa

BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. The Berkeley Framenet Project. *In*: **36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics**, Vol. 1. Quebec: Association for Computational Linguistics, 1998. p. 86-90. DOI https://doi.org/10.3115/980845.980860

BICK, E. **The Parsing System PALAVRAS**: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press, 2000.

BICK, E. Constraint grammar-based conversion of dependency treebanks. *In*: **Proceedings of the 13th International Conference on Natural Language Processing** (ICON). Varanasi: NLP Association of India, 2016, p. 109–114.

BRESNAN, J. Control and Complementation. **Linguistic Inquiry**, Vol. 13, n. 3, p. 343-434, 1982.

BUCHHOLZ, S.; MARSI, E. CoNLL-X Shared Task on Multilingual Dependency Parsing. **Proceedings of the Tenth Conference on Computational Natural Language Learning**. New York: Association for Computational Linguistics, 2006. p. 149–164. DOI https://doi.org/10.3115/1596276.1596305

FILLMORE, C. J. The Case for Case. *In*: BACH, E.; HARMS R. T. (ed.) **Universals in Linguistic Theory**. London: Holt, Rinehart and Winston. p. 1-88, 1968.

FREITAS, C.; ROCHA, P.; BICK, E. "Floresta Sintá(c)tica: Bigger, Thicker and Easier". *In*: TEIXEIRA, A.; LIMA, V. L. St. de; OLIVEIRA, L. C. de; QUARESMA, P. (ed.), **Proceedings of Computational Processing of the Portuguese Language, 8th International Conference,** (PROPOR 2008), vol. 5190. Springer Verlag, 2008. p. 216-219. DOI https://doi.org/10.1007/978-3-540-85980-2_23

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.. **Deep Learning**. Cambridge MA: MIT Press, 2016.

HENRIQUES, C. C. **Nomenclatura Gramatical Brasileira**: cinquenta anos depois. São Paulo: Parábola, 2009.

HONNIBAL, M.; JOHNSON, M. An Improved Non-monotonic Transition System for Dependency Parsing. *In*: **Proceedings of the Conference on Empirical Methods in Natural Language Processing** (EMNLP). Lisboa: Association for Computational Linguistics, 2015. p. 1373-1378. DOI https://doi.org/10.18653/v1/D15-1162

IDE, N. Introduction: The Handbook of Linguistic Annotation. In: Ide, Nancy; Pustejovsky, James (ed). **Handbook of Linguistic Annotation**. Springer, 2017. DOI https://doi.org/10.1007/978-94-024-0881-2

KARLSSON, F. Constraint Grammar as a Framework for Parsing Unrestricted Text. *In*: KARLGREN, H. (ed.), **Proceedings of the 13th International Conference of Computational Linguistics**, Vol. 3. ACM Digital Library, 1990. p. 168-173. DOI https://doi.org/10.3115/991146.991176

KIPPER-SCHULER, K. **VerbNet:** A Broad-Coverage, Comprehensive Verb Lexicon. Tese de Doutorado (Ciência da Computação). University of Pennsylvania, 2005.

KONDRATYUK, D.; STRAKA, M. 75 Languages, 1 Model: Parsing Universal Dependencies Universally. *In*: INUI, K.; JIANG, J.; NG, V.; WAN, X. (ed.) **Proceedings of the 2019 Conference on Empirical Metho1982)ds in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing** (EMNLP), Hong Kong, China. Association for Computational Linguistics, 2019. p. 2779–2795. DOI https://doi.org/10.18653/v1/D19-1279

LEVIN, B. **English Verb Classes and Alternations:** A Preliminary Investigation. University of Chicago Press, 1993.

MANNING, C.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge, MA: MIT Press, 1999.

MARNEFFE, M.-C. de; MANNING, C. D.; NIVRE, J.; ZEMAN, D.. Universal Dependencies. **Computational Linguistics** 47 (2), p. 255–308, 2021.

MITCHELL, T. Machine Learning. New York: McGraw Hill, 1997.

NIVRE, J. Towards a Universal Grammar for Natural Language Processing. *In*: **Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics** (CICLing), 2015. p. 3-16. DOI https://doi.org/10.1007/978-3-319-18111-0_1

NIVRE, J.; HALL, J.; NILSSON, J.; ERYIĞIT, G.; MARINOV, S. Labeled Pseudo-Projective Dependency Parsing with Support Vector Machines. *In*: **Proceedings of the Tenth Conference on Computational Natural Language Learning** (CoNLL-X). New York: Association for Computational Linguistics, 2006. p. 221–225. DOI https://doi.org/10.3115/1596276.1596318

NIVRE, J.; MARNEFFE, M.-C. de; GINTER, F.; HAJIČ, J.; MANNING, C. D.; PYYSALO, S.; SCHUSTER, S.; TYERS, F.; ZEMAN, D. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. *In*: **Proceedings of the 12nd International Conference on Language Resources and Evaluation** (LREC 2020). Marseille: European Language Resources Association, 2020. p. 4034-4043.

PALMER, M.; GILDEA, D.; KINGSBURY, P. The Proposition Bank: An Annotated *Corpus* of Semantic Roles. **Computational Linguistics**, 31:1., p. 71-105, March, 2005. DOI https://doi.org/10.1162/0891201053630264

PUSTEJOVSKY, J.; BUNT, H.; ZAENE, A. Designing Annotation Schemes: From Theory to Model. *In*: IDE, N.; PUSTEJOVSKY, J. (ed.). **Handbook of Linguistic Annotation**. Springer, 2017. DOI https://doi.org/10.1007/978-94-024-0881-2_2

QI, P.; ZHANG, Y.; ZHANG, Y.; BOLTON, J.; MANNING, C. D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *In*: JURAFSKY, D.; CHAI, J.; SCHLUTER, N.; TETRAULT, J. R. (ed.). **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.** Association for Computational Linguistics, 2020. p. 101-108. DOI https://doi.org/10.18653/v1/2020.acl-demos.14

RADEMAKER, A.; CHALUB, F.; REAL, L.; FREITAS, C.; BICK, E.; PAIVA, V. de. Universal Dependencies for Portuguese. *In*: **Proceedings of the Fourth International Conference on Dependency Linguistics.** Linköping University Electronic Press, 2017. p. 197-206.

SARDINHA, T. B. Linguística de *corpus*: Histórico e Problemática. **Documentação e Estudos em Linguística Teórica e Aplicada** (DELTA), 16:2, 2000. p. 323-367. DOI https://doi.org/10.1590/S0102-44502000000200005

SILVA, J.; BRANCO, A.; CASTRO, S.; REIS, R.. Out-of-the-Box Robust Parsing of Portuguese. *In*: **Proceedings of the 9th International Conference on the Computational Processing of Portuguese**. Springer, 2010. p. 75–85. DOI https://doi.org/10.1007/978-3-642-12320-7_10

SOUZA, E. de; CAVALCANTI, T.; SILVEIRA, A.; EVELYN, W.; FREITAS, C. **Diretivas e documentação de anotação UD em português (e para língua portuguesa)**. Rio de Janeiro: PUC-RIO, 2020. Disponível em: http://comcorhd.letras.puc-rio.br/recursos/.

STRAKA, M.; STRAKOVA, J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *In*: **Proceedings of the CoNLL 2017 Shared Task**: Multilingual

Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics, 2017. p. 88-99. DOI https://doi.org/10.18653/v1/K17-3009

TESNIÈRE, L. **Éléments de Syntaxe Structurale**. Paris: Librarie C. Klincksieck, 1959.

TESNIÈRE, L. **Elements of Structural Syntax.** Tradução de OSBORNE, T.; KAHANE, S. Amsterdam: John Benjamins, 2015. DOI https://doi.org/10.1075/z.185

Artigo recebido em: 19.10.2021 Artigo aprovado em: 25.04.2022