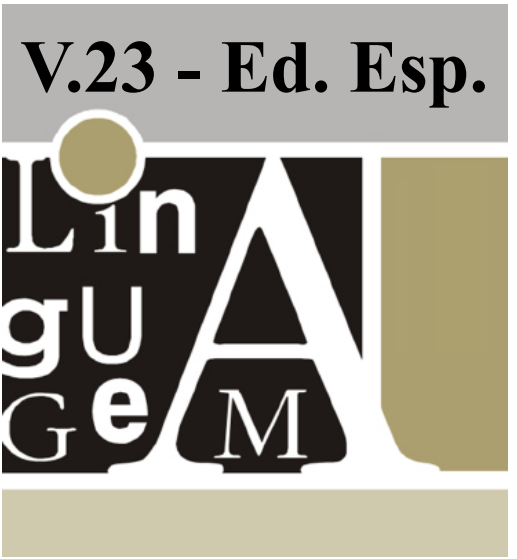
REVISTA DE ESTUDOS DA LINGUAGEM

Faculdade de Letras da UFMG



Revista de Estudos da Linguagem • Belo Horizonte • v. 23 • n. 3 • p. 581-922 • Edição Especial • 2015

REVISTA DE ESTUDOS DA LINGUAGEM

REVISTA DE ESTUDOS DA LINGUAGEM, v.1 - 1992 - Belo Horizonte, MG,
Faculdade de Letras da UFMG
Histórico:
1992 ano 1, n.1 (jul/dez)
1993 ano 2, n.2 (jan/jun)
1994 Publicação interrompida
1995 ano 4, n.3 (jan/jun); ano 4, n.3, v.2 (jul/dez)
1996 ano 5, n.4, v.1 (jan/jun); ano 5, n.4, v.2; ano 5, n. esp.
1997 ano 6, n.5, v.1 (jan/jun)
Nova Numeração:
1997 v.6, n.2 (jul/dez)
1998 v.7, n.1 (jan/jun)
1998 v.7, n.2 (jul/dez)

1. Linguagem - Periódicos I. Faculdade de Letras da UFMG, Ed.

CDD: 401.05

ISSN
On-line: 2237-2083

REVISTA DE ESTUDOS DA LINGUAGEM

V. 23 - Nº 3 - EDIÇÃO ESPECIAL - 2015

Indexadores:

CSA - Linguistics and Language Behavior Abstract
Directory of Open Access Journals
EBSCO
JournalSeek
Latindex
Matriu d'Informació per a l'Anàlisi de Revistes
Modern Language Association Bibliography
Portal Capes
WorldCat/Online Computer Library Center

REVISTA DE ESTUDOS DA LINGUAGEM

Sumário / Contents

DOI: 10.17851/2237-2083.23.3.589-597

Triangulando *corpus*, tecnologia e cultura: ELC e EBRALC na UFU

Este número da *Revista de Estudos da Linguagem* (RELIN) está voltado para a metodologia e a abordagem baseadas em *corpora*, e leva até o leitor textos oriundos de trabalhos advindos do XII Encontro de Linguística de Corpus (ELC) e da VII Escola Brasileira de Linguística Computacional (EBRALC). Os eventos foram realizados nas dependências do Instituto de Letras e Linguística (ILEEL) da Universidade Federal de Uberlândia (UFU), Minas Gerais, em novembro de 2014. Esses eventos contaram com o patrocínio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES –, da Fundação de Amparo à Pesquisa do Estado de Minas Gerais – FAPEMIG – e do Programa de Pós-Graduação em Estudos Linguísticos – PPGEL/ILEEL/UFU. A comissão organizadora foi composta por professores da própria instituição, da Universidade Federal de Minas Gerais – UFMG, da Universidade Federal do Rio Grande do Sul – UFRGS e da Universidade de São Paulo – USP, que contou com a participação ativa dos estudantes vinculados ao nosso Grupo de Pesquisas e Estudos em Linguística de *Corpus* – GPELC,[1] da UFU.

Nessa ocasião e com o tema *Corpus, tecnologia e cultura*, reunimos aproximadamente cento e dez participantes, que apresentaram seus trabalhos em diferentes modalidades, a saber: pôsteres, o já consagrado minuto de loucura e comunicações orais. Além de professores convidados de diferentes instituições nacionais, que plasmaram sua contribuição em palestras, mesa-redonda e oficinas, pudemos materializar a participação de dois professores de instituições internacionais, Carlos Ramisch (Universidade de Marseille, França) e Giovanni Parodi (PUC-Valparaíso, Chile), que compartilharam em plenária seus conhecimentos, pesquisas e recursos da área desenvolvidos no exterior.

[1] Disponível em: <http://dgp.cnpq.br/dgp/espelhogrupo/0919828121533301>. Acesso em: 19 nov. 2015.

No Brasil, apesar dos avanços consideráveis da última década, os trabalhos e resultados de pesquisa em Linguística de *Corpus* e Linguística Computacional ainda não contam com a merecida difusão. Um dos nossos objetivos foi o de mostrar os benefícios para a sociedade brasileira, pela promoção do diálogo entre as áreas da Linguística e da Computação tanto pela divulgação de produtos concretos desenvolvidos para o desempenho de tarefas imediatas como pelo subsídio para a realização de pesquisas e consequente descrição e análise linguísticas. Buscando oportunizar uma experiência nas mais diversas áreas envolvidas, durante os dias de encontro em Uberlândia, promovemos a familiarização com conceitos, abordagens e metodologias relacionadas à compilação, tratamento, etiquetagem e exploração de *corpora*, para as mais diferentes aplicações.

O ELC e a EBRALC são eventos itinerantes Brasil afora, que reúnem pesquisadores brasileiros e convidados internacionais, para a discussão de pesquisas de base empírica. Como bem pontuado por Novodvorski e Finatto (2014), o desenvolvimento da Linguística de *Corpus* (LC) no país já se consolida como uma aventura mais do que adequada. Desde a publicação do livro pioneiro de Berber Sardinha (2004), já são doze anos de um desenvolvimento contínuo de inúmeras pesquisas, que estão revolucionando os objetos de estudo das mais de quarenta subáreas da Linguística (FROMM; YAMAMOTO, 2013). Utilizando-se de dados concretos, baseados em *corpora* cuidadosamente compilados, preparados e, por vezes, também etiquetados, constatamos uma mudança de foco nas análises linguísticas brasileiras, especialmente naquelas que se baseiam na intuição dos pesquisadores como ponto de partida, levando-as para um patamar muito próximo das ciências exatas. Por outro lado, também cabe mencionar que inúmeros programas de análise lexical, como o *WordSmith Tools*, versão 6.0 (SCOTT, 2012)[2] e o *AntConc*, versão 3.4.4. (ANTHONY, 2014),[3] muitas das vezes porta de entrada para este contexto e vários outros que dão suporte a pesquisas

[2] Disponível em: <http://www.lexically.net/wordsmith/downloads/>. Acesso em: 10 nov. 2015.

[3] Disponível em: <http://www.laurenceanthony.net/software/antconc/>. Acesso em: 10 nov. 2015.

mais pontuais, como *UAM Corpus Tool, 3.2j* (O'DONNELL, 2015)[4] e *Sketch Engine* (KILGARRIFF *et al*., 2014),[5] estão sendo empregados em trabalhos dos mais diferentes campos do saber.

Os trabalhos na área da LC já foram tema de publicação em diferentes revistas dedicadas à Linguística: *Domínios de Lingu@gem*[6] e *Letras & Letras*[7] (UFU) e *Veredas*[8] (UFJF). Chegou o momento de a RELIN apresentar sua contribuição à área. Por meio de trabalhos baseados e / ou direcionados por *corpora* (TAGNIN, 2015), apresentamos aqui um pequeno panorama do que está sendo realizado em pesquisas linguísticas de ponta, de norte a sul do país.

Pablo Neves Machado e Vera Lúcia Strube de Lima, pesquisadores da PUC-RS, exploram em detalhe, no seu texto *Extração de relações hiponímicas em um corpus de língua portuguesa*, a importância das relações hiponímicas, na construção de estruturas de conhecimento (ontologias ou taxonomias), num *corpus* de língua portuguesa. Os autores objetivam o aprimoramento dos procedimentos de busca e extração de padrões, pela alimentação de um protótipo que gera as relações hiponímicas, e o processo de avaliação humana na produtividade dos padrões resultantes. As principais contribuições apontadas no trabalho dizem respeito à integração, em uma mesma pesquisa, de regras propostas por diferentes pesquisadores no assunto e pela análise detalhada dos resultados.

Com o texto *O léxico do corpo e anotação de sentidos em grandes corpora – o projeto Esqueleto*, Cláudia Freitas, Bruno Carriço (ambos da PUC-Rio), Diana Santos, Heidi Jansen (ambas da Universidade de Oslo) e Cristina Mota (Linguateca) mergulham num estudo sobre o léxico do corpo humano, motivado pelo sentido, com base

---

4 Disponível em: <http://www.wagsoft.com/CorpusTool/index.html>. Acesso em: 10 nov. 2015.

5 Disponível em: <https://www.sketchengine.co.uk/>. Acesso em: 15 nov. 2015.

6 Disponível em: <http://www.seer.ufu.br/index.php/dominiosdelinguagem/issue/view/615>. Acesso em: 19 nov. 2015.

7 Disponível em: <http://www.seer.ufu.br/index.php/letraseletras/issue/view/1217>. Acesso em: 19 nov. 2015.

8 Disponível em: <http://www.ufjf.br/revistaveredas/edicoes/2009-3/2009-2>. Acesso em: 19 nov. 2015.

em anotação e revisão de *corpora* de grande extensão em língua portuguesa. Os autores destacam a relevância do processo de anotação, caracterizado como "forma de estudo", que operaria "como uma lente de aumento", aliado ao processamento automático da língua e de sua descrição. Para além do detalhamento dos procedimentos metodológicos, os autores explicitam as diferentes informações pertinentes aos processos de anotação, às classes compreendidas e à exploração do *corpus* compilado de entrevistas e textos literários, que possui mais de 2,5 milhões de palavras. Uma informação relevante para o leitor é que todo o material está disponibilizado para consulta na página do projeto e, ainda, são oferecidas diversas sugestões para a realização de novas pesquisas, explorando os diversos sentidos do corpo na língua portuguesa.

Em *A tradução de suculentos jogos de palavras, sem perder o sabor*, Stella Tagnin (USP) concentra sua atenção em diferenças culturais entre a Inglaterra e o Brasil, pelo estudo da tradução para o português, no âmbito da culinária. A pesquisadora analisa, especificamente, a tradução dos títulos de receitas de um livro britânico de sucos para crianças. Segundo Tagnin, a maioria dos títulos analisados apresentam jogos de palavras e usos metafóricos, que exigem a busca de soluções por parte do tradutor, no sentido de preservar a riqueza semântica presente no apelo para o público infantil, realizado por meio de diferentes recursos como colocações, aliterações, rimas e referências culturais. Considerando as estratégias empregadas para a denominação dos sucos na língua de chegada e também observando a priorização do 'efeito' sobre a 'forma', em termos funcionalistas, a pesquisadora aponta que tais objetivos foram alcançados nas traduções e que, quando necessário, foram feitas adaptações culturais, deixando resguardado o caráter apelativo.

A proposta do texto *Topic Modeling for Keyword-Phrase Extraction: using Natural Language Processing methods for keyword extraction in Portal Min@s*, de Arnaldo Candido Junior (UTFPR), Célia Magalhães (UFMG), Helena Caseli (UFSCar) e Régis Zangirolami (USP - São Carlos), é apresentar uma comparação de métodos, baseados na Linguística de Corpus e no Processamento de Linguagem Natural, para extração de palavras-chave em textos ficcionais (romances, mais especificamente). Duas ferramentas, cujas técnicas de extração são

diferentes, foram analisadas no trabalho: uma mais clássica e comum aos linguistas, o *WordSmith Tools*, e outra com enfoque diferente, baseada em tópicos, o *Latent Dirichlet Allocation* (LDA), que pode ser acessada pelo *site Portal Min@s: Corpora de Fala e Escrita*. O *corpus* de análise consiste em três diferentes traduções para o português do romance *Heart of Darkness*, de Joseph Conrad. Os autores concluem que, no atual estágio de desenvolvimento de programas voltados para a análise linguística das palavras-chave, o uso do *WordSmith Tools* ainda seria o mais apropriado.

Comparação linguística e perfilação gramatical sistêmica em um corpus combinado, de Francieli Silvéria Oliveira (UFOP), analisa a organização gramatical e semântica, no par linguístico inglês / português brasileiro, de um *corpus* de manuais de instrução. Entre os objetivos principais do artigo, a autora aponta o estudo da variação linguística do registro, a comparação dos sistemas linguísticos envolvidos e a descrição da produção textual das traduções. Os padrões identificados por meio da análise, levaram a autora à afirmação de que o manual de instrução teria a macro função de 'capacitar' para o desenvolvimento de determinada atividade, segundo instruções, e não pela regulação de comportamentos. As funções semânticas de 'explicar', 'classificar', 'introduzir' e 'comandar' comporiam, desse modo, a macro função de 'capacitar'. Para além das semelhanças, a pesquisadora também observou diferenças no par linguístico estudado, no que diz respeito a funções semânticas como 'convidar', presente apenas nos manuais originais em língua inglesa, e 'introduzir' como Processo Verbal, somente observado no *subcorpus* de português original. Oliveira destaca que o método de pesquisa empregado, que integra a perfilação sistêmica e a comparação linguística, abre novas perspectivas na análise de *corpus*.

O texto *The relevance of the Sketch Engine software to build Field – Football Expressions Dictionary*, de Rove Luiza de Oliveira Chishman, Aline Nardes dos Santos, Diego Spader de Souza e João Gabriel Padilha (todos da UNISINOS), tece considerações acerca dos recursos que oferece a ferramenta *Sketch Engine* e de sua potencialidade para a análise de *Frames* Semânticos no desenvolvimento de um dicionário trilíngue de expressões de futebol, baseado em *corpora* linguísticos. Os pesquisadores destacam que um dos recursos do

programa, o *Word Sketch*, mostra o comportamento sintático-semântico do *frame*, por meio de evidências empíricas. Após a abrangência da fundamentação teórica e do potencial metodológico da Linguística de *Corpus*, os autores descrevem os procedimentos de análise, aliados empiricamente à identificação das unidades polissêmicas e das colocações presentes no *corpus*, fundamentalmente com o recurso *Word Sketch*, que possibilita a visualização detalhada dos usos produtivos da língua.

Marco A. Sobrevilla Cabezudo, Erick G. Maziero, Márcio S. Dias, Paula C. F. Cardoso, Pedro P. Balage Filho, Thiago A. S. Pardo, Verônica Agostini, Fernando A. A. Nobrega (todos do NILC/USP - S. Carlos), Jackson W. C. Souza e Ariani Di Felippo (ambos da UFSCar) e Cláudia D. Barros (IFET - São Paulo), em seu texto denominado *Anotação de Sentidos de Verbos no corpus CSTNew*, abordam um dos temas mais complexos para o Processamento de Linguagem Natural (PLN): a ambiguidade lexical. Considerando os ambientes de ocorrência, a polissemia das palavras impõe um desafio para o PLN, a determinação do sentido adequado de uma palavra em contexto. Conforme explicam os autores, a responsável por essa tarefa em PLN é a Desambiguação Lexical de Sentido (DLS). Para a realização de trabalhos dessa natureza, tanto o processo e desenvolvimento de métodos de anotação como o uso do *corpus* anotado é fundamental, para uma análise mais profunda da ambiguidade, isto é, a consulta a um *corpus* com anotação semântica dos verbos expande as possibilidades da DLS. Nesse sentido, o trabalho apresentado na publicação descreve os procedimentos pertinentes à anotação de um *corpus* jornalístico, por meio da *WordNet* de Princeton, como repositório de sentidos. Para além do detalhamento dos processos de etiquetagem e extração de dados e obtenção de resultados, uma das principais contribuições deste texto, resultante do projeto denominado SUCINTO, é a disponibilização para consulta do *corpus* jornalístico anotado e da própria ferramenta de edição, no intuito de poder subsidiar pesquisas futuras em DLS, tal como expressam seus autores.

No texto intitulado *Collocations workbook: um material de apoio pedagógico on-line e baseado em corpus para o ensino de colocações em inglês*, Adriane Orenha-Ottaiano (UNESP) justifica a elaboração de materiais de ensino que contemplem "uma seleção cuidadosa de

colocações", ajustada às principais dificuldades e necessidades dos aprendizes de inglês como língua estrangeira, principalmente dirigida a falantes de português brasileiro. Desse modo, a autora aponta ao tratamento das colocações nas aulas de inglês, buscando preencher a carência de propostas fraseológicas para o ensino, por meio da compilação de um *Workbook online* de colocações. Alimentam esse trabalho as produções de alunos universitários de tradução que, por sua vez, são parte integrante de um *corpus* paralelo de aprendizes, na direção português / inglês, além de redações que também compõem esse banco de dados. Por meio de uma combinação entre o programa *WordSmith Tools* e a consulta ao *Corpus of Contemporary American English*, é constatada a frequência dos padrões colocacionais. Com o objetivo de promover o desenvolvimento da fluência em língua inglesa, a pesquisadora destaca a importância de buscar conscientizar quanto ao fenômeno da colocação. Outro fator relevante da pesquisa é a disponibilização do recurso na *Web*.

Por último, Tommaso Raso (UFMG), Maryualê Malvessi Mittmann (FACVEST) e Anna Carolina Oliveira Mendes (UFMG) analisam a pausa silenciosa, como critério de segmentação da fala em unidades comunicativamente autônomas. Em seu texto intitulado *O papel da pausa na segmentação prosódica de corpora de fala*, os autores contextualizam e discutem as noções de unidade de referência da fala e de pausa. As fronteiras prosódicas delimitariam em unidades pragmáticas a segmentação da fala. Isto é, pelo estudo da fala espontânea com base no *corpus* C-ORAL-BRASIL e, neste caso particular, pelo estudo da duração das pausas, a investigação procurou estabelecer critérios confiáveis para a identificação de fronteiras de unidades de referência. Contudo, os pesquisadores chegam à conclusão acerca da impossibilidade de vincular genericamente as marcas físicas de pausas, o silêncio, no caso, às fronteiras de unidade de referência da fala. O estudo da quebra prosódica, em contrapartida, tem-se mostrado mais profícuo para a finalidade da segmentação da fala espontânea. Na expectativa de compreender melhor a percepção da natureza da fala, os autores adotam uma abordagem guiada pelo *corpus* e segundo os princípios de aprendizagem de máquina, reconhecendo a grandiosidade de desenvolver

um sistema (semi) automatizado para a segmentação de enunciados e de unidades tonais da fala.

Esperamos que todos tenham um bom proveito acadêmico com os textos aqui apresentados, resultantes do nosso encontro no triângulo mineiro, por ocasião do XII ELC e da VII EBRALC realizados na UFU, e que esses textos nos inspirem a continuar investindo nas pesquisas empíricas, para as quais a Linguística de *Corpus* vem demonstrando ser uma das principais aliadas no século 21.

Ariel Novodvorski (UFU)
Guilherme Fromm (UFU)

Referências

ANTHONY, L. *AntConc* (Version 3.4.4). Tokyo: Waseda University, 2014. Disponível em: <http://www.laurenceanthony.net/software.html>. Acesso em: 15 nov. 2015.

BERBER SARDINHA, T. *Lingüística de Corpus*. Barueri S.P.: Editora Manole, 2004. 410 p.

FROMM, G.; YAMAMOTTO, M. Terminologia, Terminografia, Tradução e Linguística de *Corpus*. In: TAGNIN, S. E. O.; BEVILACQUA, C. (Org.). *Corpora na Terminologia.* São Paulo: Hub Editorial, 2013. p.129-151.

NOVODVORSKI, A.; FINATTO, M. J. B. Linguística de corpus no Brasil: uma aventura mais do que adequada. *Letras & Letras*, v. 30, n. 2, Jul/dez 2014. Disponível em: <http://www.seer.ufu.br/index.php/letraseletras/article/view/28516/15799>. Acesso em: 14 nov. 2015.

SCOTT, M. *WordSmith Tools*. Versão 6, 2012. Disponível em: <http://www.lexically.net/wordsmith/downloads>. Acesso em: 15 nov. 2015.

TAGNIN, S. E. O. Glossário de Linguística de Corpus. In: VIANA, V.; TAGNIN, S. E. O. (Org.). *Corpora na tradução*. São Paulo: Hub Editorial, 2015. 331p.

DOI: 10.17851/2237-2083.23.3.599-640

Extração de relações hiponímicas em um *corpus* de língua portuguesa

Hyponym relation extraction in a Portuguese language corpus

Pablo Neves Machado
Pontifícia Universidade Católica (PUC-RS), Porto Alegre, RS, Brasil
machadoum@gmail.com

Vera Lúcia Strube de Lima
Pontifícia Universidade Católica (PUC-RS), Porto Alegre, RS, Brasil
vera.strube@pucrs.br

Resumo: As relações hiponímicas são importantes na construção de estruturas de conhecimento, tais como ontologias ou taxonomias, para melhorar o processo de busca. O presente trabalho estuda em detalhe padrões para extração de relações hiponímicas com base em um *corpus* de língua portuguesa. Para tanto, toma como base os padrões específicos propostos por Hearst (1992), Freitas e Quental (2007) e Taba e Caseli (2014). Constrói, a partir desses padrões, regras que alimentam um protótipo, o qual as aplica a um *corpus* e extrai, como resultado, relações hiponímicas. Avaliadores humanos avaliam as relações extraídas, utilizando a escala proposta por Freitas e Quental. A precisão das extrações é compatível com as da literatura. O trabalho ainda apresenta um minucioso estudo quanto à produtividade dos padrões e quanto à avaliação das extrações.

Palavras-chave: relações hiponímicas; extração de relações; padrões de extração de relações.

Abstract: Hyponym relations are important in the building of knowledge structures such as ontologies or taxonomies to enhance the search process. This work studies patterns for extraction of hyponym relations from a specific Portuguese language corpus. It starts from selected patterns proposed by Hearst (1992), Freitas and Quental (2007), and Taba and Caseli (2014). From these

patterns, it builds up rules that feed a prototype. This prototype applies them to the corpus and extracts, as a result, hyponym relations. Human evaluators assess the extracted relations, using the scale proposed by Freitas and Quental. Precision of the extractions is consistent with the literature. The paper also describes a detailed study on the productivity of the patterns and the assessment of the extractions.

Keywords: hyponym relations; relation extraction; relation extraction patterns

Recebido em: 29 de julho de 2015.
Aprovado em: 26 de outubro de 2015.

1 Introdução

O Processamento de Língua Natural (PLN) é área destacada por sua relevância no que tange ao processamento de grandes quantidades de documentos. O crescimento rápido da *World Wide Web* teve como consequência um desafio na compreensão do conteúdo das informações. Hoje, o acesso à informação na *web* é realizado, prioritariamente, por meio da busca de palavras-chave, e essa busca é realizada por mecanismos de comparação lexical. Devido ao gigantesco tamanho que a *web* apresenta atualmente e a sua contínua expansão, quando são realizadas buscas de palavras-chave, diversos conteúdos irrelevantes para o usuário são encontrados. Algumas fontes de dados manualmente estruturadas foram surgindo, mas, devido à grande quantidade de conteúdo existente na rede, fica evidente a importância de ferramentas para extração da informação disponível em língua natural.

Entre as informações a ser extraídas de textos, revestem-se de especial importância as relações hiponímicas, as quais podem ser usadas na construção de estruturas ontológicas e outras estruturas de representação do conhecimento.

O presente trabalho estuda a extração de relações com base em textos escritos em língua portuguesa. A abordagem inicial baseia-se na organização das contribuições presentes nos trabalhos de Marti Hearst (1992), Freitas e Quental (2007) e Taba e Caseli (2014), mas a arquitetura da solução e a prototipação seguem organização própria.

Também são aproveitados outros estudos, como o de Baségio (2007), que realizou a adaptação de padrões da língua inglesa para a língua portuguesa. Foi desenvolvido um protótipo de extração de relações hiponímicas de *corpus* em língua portuguesa, o qual permitiu a realização de experimentos com diferentes conjuntos de padrões e regras de extração. Os resultados obtidos com a execução do protótipo são analisados, avaliados e comparados com os registrados na literatura. Além disso, discute-se o processo de avaliação manual e analisa-se os erros frequentes.

O texto do trabalho está organizado em cinco seções, incluindo esta introdução. A seção 2 aborda, brevemente, a temática e os trabalhos correlatos ao presente estudo. A seção 3 descreve o trabalho realizado, com o detalhamento das regras de extração, as suas fontes e o processo de extração de relações. A seção 4 trata da avaliação e discussão dos resultados, incluindo uma análise dos erros encontrados. A seção 5 traz um fechamento do trabalho.

2 Extração de relações semânticas e trabalhos correlatos

2.1 A extração de relações

De acordo com o indicado no estudo de Jurafsky e Martin (2009), o significado de uma palavra pode ser expresso como sendo a sua relação com outras palavras. Tal como proposta por Gruber (1992), uma relação é um conjunto de n-uplas que representam um relacionamento entre objetos no universo do discurso. Cada n-upla é uma sequência finita e ordenada de objetos. Nesse contexto, a n-upla pode ser representada pela expressão (nome-da-relação, *arg*1, *arg*2,…, *arg*n), na qual *arg*1 é um objeto na n-upla. Um exemplo de formato de uma relação binária pode ser dado por (*arg*1, nome-da-relação, *arg*2), com pequena alteração na ordem dos componentes da n-upla.

Entre as relações semânticas, especialmente interessantes são as relações hierárquicas representadas pela hiperonímia e hiponímia. A hiperonímia expressa uma relação de significado geral, enquanto a hiponímia representa um significado hierárquico restrito. Alguns

exemplos de relações citadas nesta seção podem ser vistos no Quadro 1. No presente trabalho, focamos nas relações hiponímicas binárias, que representamos por "Hiponímia (*arg*1, *arg*2)".

Quadro 1 – Exemplos de relações semânticas

Nome da relação	**arg1**	**arg2**
Sinonímia	claro	alvo
Antonímia	claro	escuro
Hiperonímia	animal	cachorro
Hiponímia	cachorro	animal

Fonte: nossa autoria.

Existem outras relações semânticas que ligam argumentos no texto, verbais ou não verbais. Por exemplo, da oração "Alexandre adora fritas", pode ser extraída a n-upla (adora, Alexandre, fritas), na qual "Alexandre" e "fritas" são argumentos e "adora" representa a relação.

Os primeiros trabalhos relacionados à extração automática de relações abordam, principalmente, relações hiponímicas e meronímicas. Isso se deve ao fato de essas relações serem a base para a construção de ontologias. As relações hiponímicas são comumente representadas por "é um" e expressam relações entre instâncias e classes, como também entre classes. Mais recentemente as pesquisas nessa área têm incluído também as relações verbais, em geral representadas por verbos e seus argumentos.

Esses modelos de relação podem ser extraídos de textos em língua natural com base no processamento de *corpora*. Para Banko e coautores (2007), os sistemas de extração de relações normalmente buscam satisfazer determinadas demandas previamente especificadas, por exemplo, extrair o local e horário de um evento por meio de um conjunto de anúncios. Quando se tem de extrair relações de um novo domínio, costuma ser necessário um retrabalho. Pode ser preciso refazer algumas das tarefas, como o estabelecimento da heurística empregada na extração e a etiquetagem de um novo *corpus* de treino.

No atual estado da arte, existem trabalhos, principalmente para a língua inglesa, que abordam o tema da extração de relações. Há também ferramentas e recursos disponíveis que são interessantes. Introduzimos, a seguir, alguns desses trabalhos referentes ao tema. Entre as abordagens que estudam a extração de relações em *corpora* textuais, duas são as mais comuns: o aprendizado de máquina e a extração baseada em regras. Na exposição que segue, é dada ênfase maior à segunda abordagem.

2.2 Trabalhos com foco em língua estrangeira

Hearst (1992) propõe um método de aquisição de relações hiponímicas entre sintagmas nominais para a língua inglesa, com base em seis padrões simples que podem ser encontrados com frequência em textos. Esses padrões podem ser vistos no Quadro 2.

Quadro 2 – Padrões apresentados por Hearst

i.	NP such as {NP ,}* {(or | and)} NP
ii.	such NP as {NP ,}* {(or | and)} NP
iii.	NP {, NP}* {,} or other NP
iv.	NP {, NP}*{,} and other NP
v.	NP {,} including {NP ,}* {or | and} NP
vi.	NP {,} especially {NP ,}* {or | and} NP

Fonte: Hearst (1992).

Um dos objetivos que conduziu Hearst a essa abordagem foi criar um método aplicável a grandes quantidades de textos. A importância do trabalho de Hearst se deve ao fato de ser um dos primeiros a propor padrões lexicais na extração de relações semânticas. Um exemplo é aquele em que a autora mostra uma aplicação do padrão (vi) (HEARST, 1992):

> "... most European countries, especially France, England and Spain."

Aplicando o padrão "NP {,} especially {NP ,}* {or | and} NP", apresentado como (vi) no Quadro 2, no qual NP é um sintagma nominal (*noun phrase*), as seguintes relações são extraídas:

Hiponímia ("France", "European country")
Hiponímia ("England", "European country")
Hiponímia ("Spain", "European country")

Hearst aplicou seus padrões em *corpora* enciclopédicos e jornalísticos, avaliando que 63% das relações identificadas eram de boa qualidade.

Os padrões textuais criados por Hearst são utilizados em diversos trabalhos, por exemplo, os de Maedche e Staab (2002), Cederberg e Widdows (2003), Degeratu e Hatzivassiloglou (2004), Baségio (2007) e Freitas e Quental (2007).

Cederberg e Widdows (2003), por exemplo, utilizam os padrões propostos por Hearst com um método denominado *Latent Semantic Analysis* (LSA), para filtrar as relações incorretas, aumentando a precisão das extrações em 30%. Relações corretamente extraídas podem ser usadas como "sementes" para a extração de diversas outras relações, aumentando, assim, a cobertura.

Morin e Jacquemin (2003) apresentam padrões para a aquisição de relações hiponímicas em *corpora* de língua francesa, tal como pode ser visto no Quadro 3. O exemplo a seguir, dado pelos mesmos autores, mostra como esses padrões se comportam.

Se o padrão "{deux|trois...|2|3|4...} NP1 (LIST2)" é aplicado ao trecho:

"... analyse foliaire de quatre espèces ligneuses (chêne, frêne, lierre et cornouiller) dans...",

é possível identificar as seguintes relações:

Hiponímia ("chêne", "espèce ligneux")
Hiponímia ("frêne", "espèce ligneux")

Hiponímia (“lierre”, “espèce ligneux”)
Hiponímia (“cornouiller”, “espèce ligneux”)

Quadro 3 – Padrões para a língua francesa propostos por Morin e Jacquemin

i.	deux|trois...|2|3|4...} NP1 (LIST2)
ii.	{certain|quelque|de autre...} NP1 (LIST2)
iii.	{deux|trois...|2|3|4...} NP1: LIST2
iv.	{certain|quelque|de autre...} NP1: LIST2
v.	{de autre} NP1 tel que LIST2
vi.	NP1, particulièrement NP2
vii.	{de autre} NP1 comme LIST2
viii.	NP1 tel LIST2
ix.	NP2 {et|ou} de autre NP1
x.	NP1 et notamment NP2

Fonte: Morin e Jacquemin (2003).

Uma proposta mais recente para a extração de relações é a denominada *Open Information Extraction* (OpenIE), que visa à extração aberta e em grande escala, sem se preocupar em tipificar as relações extraídas. Nessa proposta, Corro e Gemulla (2013) apresentam o sistema ClausIE (*Clause-based Open Information Extraction*). Os experimentos de tais autores sugerem que o sistema obtenha os melhores resultados, tornando-se referência na OpenIE, dada a característica de utilizar uma abordagem baseada em cláusulas (orações), de mais forte cunho linguístico. O sistema ClausIE identifica conjuntos de orações e o seu tipo (de acordo com a função gramatical do conteúdo), sendo baseado em um *parser* (“analisador sintático”) de dependências e em um pequeno conjunto de léxicos independentes de domínio. Essa organização permite ao sistema o processamento em paralelo e, desse modo, o processamento de grandes coleções de conteúdo de maneira escalável. Assim como

ocorre no presente trabalho, ClausIE não necessita de pós-processamento e de dados de treino para sua execução. Contudo, os autores reportam que as principais incorreções nas relações extraídas são provenientes de erros de *parsing* ("análise sintática").

Em seu trabalho, Gamallo e coautores (2012) descrevem um método que igualmente utiliza o paradigma OpenIE para a extração de triplas baseadas em verbos de *corpora* multilíngues. O método extrai relações em *corpora* nos idiomas português, inglês, espanhol e galego. Segundo os autores, o método apresenta resultados superiores aos alcançados pelos trabalhos no estado da arte, devido principalmente ao fato de utilizar análise sintática profunda e um tokenizador[1] robusto e rápido.

2.3 Trabalhos com foco na língua portuguesa

O *software* PALAVRAS, proposto e desenvolvido por Bick (2000), reúne diversas ferramentas para o processamento de línguas naturais que aceitam como entrada textos em língua portuguesa, e pode ser utilizado para etiquetagem de *corpus*, processamento léxico-morfológico, geração de árvores sintáticas e reconhecimento de entidades nomeadas, entre outros. É relatada precisão maior que 97%, tanto em termos de morfologia quanto de sintaxe. O *parser* PALAVRAS é baseado em regras e está disponível por meio do projeto VISL.[2]

Freitas e Quental (2007) adaptam dois padrões de Hearst para a língua portuguesa ("such as" e "and/or others") e criam outros três padrões, com base em análise de ocorrências no texto, capazes de detectar relações hiponímicas. O trabalho utiliza o *parser* PALAVRAS para a identificação de sintagmas nominais. As regras foram aplicadas ao *corpus* CORSA (*CORpus* da SAúde Pública), que contém cerca de 2 milhões de palavras. Os resultados foram compatíveis com os de Hearst, mostrando um percentual de 73% de relações consideradas de boa qualidade. O Quadro 4 ilustra os dois padrões de Hearst adaptados por

---

1 Será usado o termo "tokenizador" em lugar de "*tokenizer*", por ser encontrado em outros trabalhos da área em língua portuguesa.

2 Sítio *web* do projeto VISL: <http://beta.visl.sdu.dk>. Acesso em: 26 jul. 2015.

Freitas e Quental (i(a), i(b) e ii), assim como os três padrões propostos pelas autoras (iii, iv e v).

Quadro 4 – Padrões extraídos de Freitas e Quental (2007)

i(a)	SN HHiper (tais como | como_PDEN) SN1 { , SN2 ... ,} (e | ou) Sni
i(b)	SN Hiper, (tais como | como_PDEN) SN1 { , SN2 ... ,} (e | ou) Sni
ii	SN HHipo { ,SN Hipoi } * { , } e|ou outros SN Hiper
iii	tipos de SN Hiper: SN1 { , SN2 ... ,} (e | ou) Sni
iv	SN HHiper chamado/s/a/as (de) SN Hipo
v	SN Hiper conhecido/s/a/as como SN Hipo

Fonte: Freitas e Quental (2007), adaptado de Hearst (1992).

O excerto de texto apresentado por Freitas (2007, p. 76) e reproduzido a seguir mostra a aplicação do padrão (iv):

> "e nele existe uma [substância] chamada [benzopireno]."

Com a aplicação, a seguinte relação deve ser extraída: "Hiponímia (benzopireno, substância)". Como exposto por Freitas, HHiper representa a configuração na qual o termo hiperônimo é o primeiro substantivo à esquerda.

O trabalho de Baségio (2007), da mesma época que o de Freitas e Quental, visualiza a construção semiautomática de ontologias com base em textos na língua portuguesa do Brasil. Para esse fim, o trabalho emprega uma abordagem que inclui extração de relações hiponímicas. O autor traduziu, para a língua portuguesa do Brasil, relações propostas em outros trabalhos consolidados, como, principalmente, o de Hearst (1992), o que pode ser visto no Quadro 5.

Quadro 5 – Padrões de Hearst adaptados por Baségio (2007)

NP such as {(NP,)*(or|and)} NP	SUB como {(SUB,)*(ou|e)} SUB SUB tal(is) como {(SUB,)*(ou|e)} SUB
NP such as {(NP,)*(or|and)} NP	SUB como {(SUB,)*(ou|e)} SUB
such NP as {(NP,)*(or|and)} NP	tal(is) SUB como {(SUB,)*(ou|e)} SUB
NP {, NP}* {,} or other NP	SUB {, SUB}* {,} ou outro(s) SUB
NP {, NP}* {,} and other NP	SUB {, SUB}* {,} e outro(s) SUB
NP {,} including {NP,}*{or|and} NP	SUB {,} incluindo {SUB,}*{ou|e} SUB
NP {,} especially {NP,}*{or|and} NP	SUB {,} especialmente {SUB,}*{ou|e} SUB SUB {,} principalmente {SUB,}*{ou|e} SUB SUB {,} particularmente {SUB,}*{ou|e} SUB SUB {,} em especial { SUB,}*{ou|e} SUB SUB {,} em particular { SUB,}*{ou|e} SUB SUB {,} de maneira especial { SUB,}*{ou|e} SUB SUB {,} sobretudo { SUB,}*{ou|e} SUB

Fonte: Baségio (2007).

Para obter resultados mais efetivos, Baségio implementou um processo de remoção de palavras entendidas como pouco relevantes para o domínio, fixando-se nos substantivos, como se observa nas regras propostas pelo autor. Esse processo removeu cerca de 70% das palavras analisadas. Observa-se, contudo, que a contribuição do trabalho de

Baségio reside mais no estudo realizado e na transposição de regras para o português do que na análise da produtividade das regras empregadas.

Batista e coautores (2013) trazem para a língua portuguesa uma proposta de uso de aprendizagem de máquina voltada à classificação de relações, mais do que à extração propriamente dita. O trabalho faz uso de métodos de aprendizagem semisupervisionada para identificar o tipo das relações, em vez de usar regras específicas e palavras-chave como ocorre nos estudos com regras de extração. Para tanto, emprega a similaridade entre elementos de um conjunto e, assim, identifica grupos de elementos similares entre exemplares de triplas representando relações. Nos experimentos relatados pelos autores, essas triplas correspondem a frases extraídas da Wikipedia que expressam relações entre pares de entidades extraídas da DBPedia. O algoritmo *k-nearest neighbors* (*k*-vizinhos mais próximos) é usado para construir os grupos, com base, principalmente, nas classes gramaticais das palavras que ocorrem antes, depois e entre duas entidades mencionadas. Embora o foco do trabalho e dos experimentos relatados esteja na classificação de relações entre entidades mencionadas, sem identificar relações hiponímicas, a abordagem descortina novas alternativas para estudos na área para a língua portuguesa, que efetivamente vêm a ser desenvolvidos por Taba e Caseli.

Em seus estudos, Taba e Caseli (2014) fazem uso de múltiplas abordagens de aprendizagem de máquina, bem como de regras, na extração automática de relações semânticas em textos em português. As abordagens são comparadas em diferentes experimentos, com utilização de dois *corpora* anotados pelo *parser* PALAVRAS. O primeiro, o CETENFolha, *corpus* de caráter jornalístico, é composto por 24 milhões de palavras de artigos do jornal Folha de São Paulo; o segundo, de caráter científico, conta com 870 mil palavras e é proveniente de textos de uma revista de divulgação científica (FAPESP). Os autores buscam extrair as seguintes relações:

- is-a
- part-of
- location-of
- effect-of
- property-of
- made-of

- used-for

No que tange à relação *is-a*, Taba e Caseli (2014) empregaram padrões provenientes de Hearst (1992) e de Freitas e Quental (2007), além de introduzir novos padrões manualmente definidos, que podem ser vistos no Quadro 6, no qual o termo T1 representa o hiperônimo de uma relação e os termos T2, T3 representam possíveis hipônimos.

Quadro 6 – Padrões para a relação *is-a* conforme Taba e Caseli (2014)

i	T1 (tais como|como) T2 {, T3}* (e|ou) TN
ii	T2 {, T3}* ,? (e|ou) outros T1
iii	tipos de T1: T2 {, T3}* (e|ou) TN
iv	T1 chamad(o|a|os|as) de? T2
v	T2 {, T3}* ,? (e|ou) (qualquer|quaisquer) outro{s} T1
vi	T2 é (o|a|um|uma) T1
vii	T2 são T1

Fonte: Taba e Caseli (2014).

Os autores relatam quatro experimentos: o primeiro com o método baseado em regras, e os três outros usando aprendizagem de máquina para a extração das relações. No primeiro caso, a precisão reportada para as relações *is-a* é de 61,1%, porém com muito baixa cobertura, de 1,2%. Os experimentos com aprendizagem de máquina visam, além dos resultados das extrações, estudar a influência das informações sintáticas, morfológicas ou superficiais no processo de extração. Os resultados apresentados mostram que o aprendizado de máquina pode trazer aumentos significativos na cobertura. A técnica conhecida como máquinas de vetores de suporte alcança a melhor precisão para as relações *is-a*, chegando a 78,2% nos *corpora* utilizados. A comparação de tais resultados com outros disponíveis na literatura não é possível devido às diferenças nos *corpora* e *parsers*, entre outras. Mas a técnica revela-se bastante promissora.

3 O trabalho realizado

Visando a extrair relações hiponímicas por meio de *corpora* de língua portuguesa, foram realizadas adaptações de padrões propostos pelos autores estudados, o que gerou regras. As regras foram inseridas em um programa de computador desenvolvido especialmente para esse fim.

O formato das regras segue o das expressões regulares, de uso corrente nas Ciências da Computação. As regras são escritas em linguagem formal e são interpretadas por um processador de expressões regulares, tal como ocorre no presente trabalho. O objetivo de uma regra é prover uma forma concisa e flexível de representar um padrão que identifica cadeias de caracteres, ou caracteres de interesse. A linguagem formal adotada representa sintagmas nominais por "SN" e utiliza os parênteses para agrupar as expressões; o símbolo "*" indica que uma expressão pode não ocorrer, pode ocorrer uma ou mais vezes. A interrogação significa nenhuma ou uma repetição. Outro símbolo comumente utilizado é a barra vertical "|", que representa um "ou exclusivo". Também foi utilizada a notação "<sn PALAVRA-CHAVE sn>" para identificar a ocorrência de uma palavra-chave que está contida dentro de um *chunk* ("porção de texto, sintaticamente analisada"). Como um Sintagma Nominal pode ser formado por outros SNs, o símbolo "sn" (minúsculo) foi empregado para identificar um Sintagma Nominal que é um dos elementos de um "SN". Caso a palavra-chave se encontre logo após o símbolo "<" ou antes do símbolo ">", significa que ela é, respectivamente, a primeira ou a última palavra do *chunk*, como em: "<outros sn>" (a palavra-chave é representada por "outros"). O Quadro 7 relaciona as regras definidas neste trabalho e empregadas no estudo realizado. Todas se originam de padrões específicos, compilados da literatura estudada, com breves adaptações e melhorias introduzidas. A numeração, de 1 a 11, visa a facilitar sua exposição; observe-se que a regra 10 subdivide-se em duas, 10A e 10B, levando a um total de doze regras implementadas.

Quadro 7 – Regras de extração empregadas no presente trabalho

Regra	Descrição
1	SN(,)? como (SN ,)*(SN (e|ou))*SN
2	SN(,)? ta(is|l) como (SN ,)*(SN (e|ou))*SN
3	SN(,)? incluindo (SN ,)*(SN (e|ou))*SN
4	SN(,)? especialmente (SN ,)*(SN (e|ou))*SN
5	(SN (ou|e|,))*<outr(a|o)(s)? sn>
6	<... tipo(s)? de sn> : (SN ,)*(SN (e|ou))*SN
7	SN(,| é| são| foram)? chamad(o|a|os|as)(de)? (SN ,)*(SN (e|ou))*SN
8	SN((,)? também)?(,|é|são|foram)? conhecid(o|a|os|as) como (SN ,)*SN (e|ou))*SN
9	(SN (ou|e|,))*< (qualquer|quaisquer) outr(a|o)(s)? sn>
10A	SN é < (o|a) sn>
10B	SN é < (um|uma) sn
11	SN são SN

Fonte: nossa autoria.

As seções 3.1 a 3.4 detalham essas regras, sua origem e sua aplicação.

3.1 Extração com padrões de Hearst

Os padrões propostos por Hearst (QUADRO 2) foram criados com o intuito de extrair relações hiponímicas em *corpus* de língua inglesa. Para a utilização desses padrões na língua portuguesa do Brasil, partiu-se de trabalhos prévios de tradução e contextualização de tais relações.

Por exemplo: dado o excerto de texto: “Países como o Brasil, Equador e os EUA”, o padrão SN(,)? como (SN ,)*(SN (e|ou))*SN pode extrair as seguintes relações: Hiponímia (Brasil, País), Hiponímia (Equador, País) e Hiponímia (EUA, País).

O padrão em questão é o referente ao “such as”, proveniente dos estudos de Hearst. Este corresponderia ao “como” em português, que pode exercer diversas funções sintáticas em uma sentença, o que causa dificuldade em obter altos níveis de precisão, como já indicado por Freitas e Quental (2007). Baségio (2007) também havia empreendido esforços para adaptar o padrão “como” para a língua portuguesa. Entretanto considerava apenas substantivos, simplificando a ideia de sintagma nominal presente nos padrões de Hearst. Em nosso trabalho, escolhemos utilizar SNs, empregando padrões mais complexos. Já na adaptação por Freitas e Quental, foram utilizadas regras levando em conta a existência de SNs, mas ocorreu, assim como no caso de Baségio, a flexibilização, nesse caso, visando ao uso apenas da palavra mais à direita dentro do sintagma nominal.

Uma melhoria introduzida em relação aos trabalhos de Freitas e Quental (2007), e detalhada mais adiante, foi o tratamento da vírgula, que pode ocorrer antes da palavra “como”, por exemplo, em:

“... [outras falhas], como [dois nomes para um mesmo fator]...”

Essa alteração aumentou em torno de 40% o número de relações extraídas com o padrão “como” em relação aos resultados anteriores. Utilizando uma abordagem semelhante, foi possível adaptar os padrões de Hearst identificados no Quadro 2, como (ii), (v) e (vi):

SN(,)? ta(is|l) como (SN ,)*(SN (e|ou))*SN
SN(,)? incluindo (SN ,)*(SN (e|ou))*SN
SN(,)? especialmente (SN ,)*(SN (e|ou))*SN

Já o padrão a seguir, inspirado nos padrões (iii) e (iv) de Hearst, precisou de uma implementação alternativa:

(SN (ou|e|,))*<outr(a|o)(s)? sn>

O *parser* PALAVRAS, ao processar um texto como "Brasil, Equador, EUA e outros países", identifica diversos SNs, um dos quais inclui o determinante "outros":

"[Brasil], [Equador], [EUA] e [outros países]"

O presente trabalho propõe uma adaptação para encontrar SNs nessa situação. Com tal adaptação, as relações que podem ser extraídas com o padrão para o texto do exemplo são: Hiponímia (Brasil, País), Hiponímia (Equador, País), Hiponímia (EUA, País). O Quadro 8 associa os padrões propostos por Hearst com as regras propostas neste trabalho e apresentadas no Quadro 7. Pode-se observar que a regra 5 foi utilizada para expressar dois padrões propostos.

Quadro 8 – Associação entre padrões de Hearst e regras propostas neste trabalho

Regra	Padrão de Hearst
1	NP such as {NP ,}* {(or | and)} NP
2	such NP as {NP ,}* {(or | and)} NP
3	NP {,} including {NP ,}* {or | and} NP
4	NP {,} especially {NP ,}* {or | and} NP
5	NP {, NP}* {,} or other NP
	NP {, NP}*{,} and other NP

Fonte: nossa autoria, com base em Hearst (1992).

3.2 Extração com padrões de Freitas e Quental

Os padrões (iii), (iv) e (v) do Quadro 4, provenientes dos estudos específicos de Freitas e Quental, dão origem a regras desenvolvidas neste trabalho e numeradas como 6 a 8, no Quadro 7.

O padrão (iii), denominado "tipos de", busca extrair relações com base nas palavras-chave que dão origem ao seu nome. O excerto de texto

a seguir, proveniente do *corpus* CORSA, visa a demonstrar as relações que a regra correspondente é capaz de extrair:

> "desenvolver [dois tipos de dengue] : [dengue clássica] e
> [dengue hemorrágica]".

Desse trecho, a regra deve ser capaz de extrair as relações: Hiponímia (dengue clássica, dengue), Hiponímia (dengue hemorrágica, dengue). O resultado da adaptação criada para realizar tal tarefa é descrito na regra 6 do Quadro 7, a saber:

> <... tipo(s)? de sn> : (SN ,)*(SN (e|ou))*SN

A semelhança com o padrão de Freitas e Quental se dá na utilização dos símbolos "<" e ">", para representar um sintagma nominal que contém em seu interior as palavras-chave da regra. Isso se deve ao fato de o *parser* PALAVRAS definir que a expressão "tipos de" faz parte de um *chunk* com outras palavras que podem vir antes ou depois do padrão, por exemplo, "[todos os tipos de cortes]" e "[os principais tipos de tifo]". Para maximizar o número de relações extraídas, a regra foi flexibilizada para aceitar a expressão "tipo de", sem a utilização do plural. Essa regra apresenta um alto grau de confiança, como pontua Freitas:

> "... o padrão 'tipos de' não apresenta problemas de ambiguidade relativos ao sintagma preposicionado, nem particularidades de natureza discursiva ou coesiva – o que significa que as relações identificadas são altamente confiáveis." (FREITAS, 2007, p. 75).

Outra adaptação, dos estudos de Freitas e Quental, foi a do padrão denominado "chamado/a/os/as", representado como (iv) no Quadro 4, que deve extrair relações de textos como:

> "... e nele existe uma [substância] chamada [benzopireno]."

Nesse caso, a relação extraída é Hiponímia (benzopireno, substância). A regra encarregada de tal extração é identificada como regra 7 no Quadro 7, a saber:

> SN(,| é| são| foram)? chamad(o|a|os|as)(de)? (SN ,)*(SN (e|ou))*SN

Para maximizar o número de relações extraídas, foram flexibilizados o uso do verbo "ser" em quatro formas – "é", "são", "foi", "foram"– e a utilização de vírgula. Foi também permitida a ocorrência de uma lista de sintagmas nominais após a palavra-chave "chamado". Esse formato de lista já está presente nas regras 1, 2, 3 e 4 do Quadro 7; e permite a extração de relações de excertos de textos como:

> "... vem estudando profundamente [o fenômeno] , chamado de [sinantropia] ou [domiciliação]..."

O padrão (v) do Quadro 4 é o último adaptado do trabalho de Freitas e Quental. As autoras o denominam "conhecido/a/os/as como", e deve extrair relações de excertos como:

"[vesículas esféricas de gordura] , conhecidas como [lipossomas]", obtendo a relação Hiponímia (lipossomas, vesículas esféricas de gordura).

Após o processo de adaptação, a regra criada, numerada como regra 8 no Quadro 7, ganhou a seguinte representação:

> SN((,)? também)?(,|é|são|foram)? conhecid(o|a|os|as) como (SN,)*SN (e|ou))*SN

Para maximizar o número de relações extraídas pela regra 8, foram realizadas alterações, permitindo a presença de vírgula e das formas verbais "é", "são", "foi" e "foram" antes da expressão "conhecido como", além de uma lista de sintagmas nominais ter sido acrescentada após a expressão. Ainda, a presença da palavra "também" após o primeiro sintagma nominal passou a ser prevista.

No Quadro 9, esses são associados ao presente trabalho.

Quadro 9 – Associação entre os padrões de Freitas e Quental e regras do presente trabalho

Regra	Padrão de Freitas e Quental
6	tipos de SN Hiper: SN 1 { , SN 2 ... ,} (e | ou) Sni
7	SN HHiper chamado/s/a/as (de) SN Hipo
8	SN Hiper conhecido/s/a/as como SN Hipo

Fonte: nossa autoria, baseado em Freitas e Quental (2007).

Analisando o trabalho de Freitas e Quental, nota-se um formato de sintagma nominal que está ausente nas regras do presente trabalho: "SN HHiper". As autoras utilizaram esse prefixo para os sintagmas nominais com o objetivo de melhorar a precisão das extrações. O SN HHiper é utilizado para identificar apenas a primeira palavra encontrada mais à direita de um sintagma nominal. Já os "SN Hipo" e "SN Hiper" são utilizados para representar um sintagma nominal como elemento hiponímico ou hiperonímico da relação.

3.3 Extração com padrões de Taba e Caseli (2014)

O trabalho de Taba e Caseli (2014) também estuda o emprego de padrões para extração automática de relações semânticas de *corpus* de língua portuguesa. Em sua pesquisa, os autores utilizam os padrões criados por Freitas e Quental, assim como outros de sua própria autoria. Destes, abordaremos apenas os padrões (v), (vi) e (vii) do Quadro 6, que realizam extração de relações hiponímicas (denominadas por Taba e Caseli de relações "is-a"). O primeiro padrão adaptado foi o padrão (v), que busca extrair relações de textos como:

> "... apresentar [febre] ou [qualquer outro sintoma da doença de Chagas]..."

Com base no padrão (v), obtém-se a relação Hiponímia (febre, sintoma da doença de Chagas). A sua adaptação, apresentada na regra de número 9 no Quadro 7, pode ser vista a seguir:

(SN (ou|e|,))*< (qualquer|quaisquer) outr(a|o)(s)? sn>

No padrão original (TABA; CASELI, 2014, p. 2741), são permitidas apenas as palavras "outro" ou "outros" antes do último SN. No corrente trabalho, esse modelo foi flexibilizado para que a palavra no gênero feminino também fosse válida ("outra", "outras"). Assim como em outras regras, foram utilizados os sinais ">" e "<" para indicar que as palavras-chave são encontradas dentro de um *chunk*, e uma parte desse *chunk*, que é representada por "sn", será considerada nas relações extraídas.

Já o padrão (vi) do Quadro 6 é capaz de extrair relações de sentenças como:

> "por [a agência local de a Fundação Instituto Brasileiro de Geografia e Estatística] , [Pelotas] é [uma cidade] [cuja zona urbana comporta 297.825 habitantes]"

No caso, é obtida a relação Hiponímia (Pelotas, cidade).

O padrão em questão gerou duas regras, respectivamente numeradas no Quadro 7 como 10A e 10B:

SN é < (o|a) sn>
SN é < (um|uma) sn>.

Como pode ser observado, as regras 10A e 10B são semelhantes, mas o motivo da criação de duas regras, nesse caso, é o fato de ambas serem generalistas. Como extraem um grande número de relações, essa separação visa a que futuras análises possam determinar a precisão das regras individualmente. Ambas as regras apresentam a estrutura que indica que as palavras-chave estão dentro do *chunk*.

O padrão (vii) de Taba e Caseli (QUADRO 6) objetiva extrair relações de textos como no exemplo a seguir:

> "[as hemoglobinopatias] são [doenças geneticamente determinadas] e apresentam [morbidade significativa] em todo o mundo",

obtendo a relação Hiponímia (as hemoglobinopatias, doenças geneticamente determinadas).

Por fim, podemos ver a última regra adaptada de Taba e Caseli (2014), referente ao padrão (vii) do Quadro 6 e descrita no Quadro 7 como regra 11. A construção dessa regra reflete basicamente a transcrição do padrão desses autores para a sintaxe utilizada neste trabalho:

SN são SN

No Quadro 10, podem ser vistos os padrões adaptados de Taba e Caseli (2014), com suas correspondências para quatro regras do presente trabalho.

Quadro 10 – Associação entre padrões de Taba e Caseli e as regras do presente trabalho

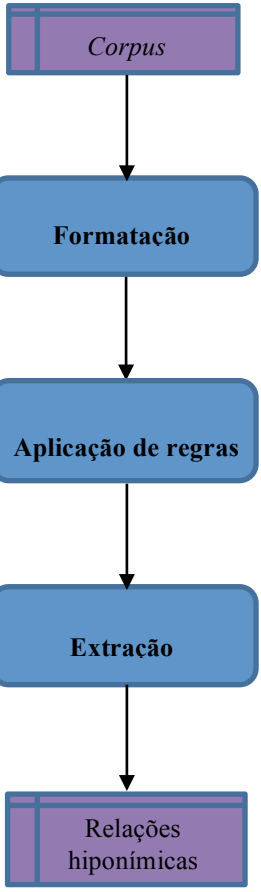
Regra	Padrão de Taba e Caseli
9	(SN (ou|e|,))*< (qualquer|quaisquer) outr(a|o)(s)? sn>
10A	SN é < (o|a) sn>
10B	SN é < (um|uma) sn>
11	SN são SN

Fonte: nossa autoria.

3.4 Protótipo e aplicação das regras

Para implementar e testar um extrator de relações hiponímicas de textos em português com base nos padrões trabalhados, foi desenvolvido um protótipo funcional, cuja arquitetura é descrita na Esquema 1.

Esquema 1 – Arquitetura utilizada na construção do protótipo



Fonte: nossa autoria.

Como mostra o Esquema 1, o fluxo inicia pela inserção do *corpus* como um parâmetro de entrada. O processo de formatação age sobre todo o *corpus* de entrada e produz, como parâmetro de saída, uma nova versão desse mesmo *corpus*, agora formatado adequadamente para a extração das relações. Então, o processo de aplicação de regras entra em ação, executando sobre cada sentença as regras do Quadro 7. Como

resultado, esse processo retorna todos os trechos de sentenças que foram identificados pelas regras. Na última etapa, esses trechos são inseridos como parâmetro de entrada para o processo de extração. Nesse processo, as relações resultantes são criadas e, então, é produzida uma lista com todas as extrações obtidas pela execução do protótipo.

As regras, na forma de expressões regulares, são escritas em linguagem formal e são interpretadas pelo processador de expressões regulares, que examina o texto e procura por trechos que atendam às regras determinadas pela expressão. A escolha desse método se deu por sua simplicidade e expressividade, assim como por estar disponível para uso em diversas linguagens de programação.

O *corpus* utilizado para experimentar o protótipo desenvolvido foi o CORSA, que é formado por 1.846.502 palavras armazenadas em um arquivo de 11Mb. O CORSA foi criado com base em textos da área de Saúde Pública, incluindo artigos acadêmicos, cartilhas, manuais, textos didáticos e também textos jornalísticos. A diversidade das fontes é proposital, com o objetivo de agregar variadas formas de escrita, assim como diferentes níveis de aprofundamento técnico. Esses textos foram analisados previamente pelo *parser* PALAVRAS. Em seguida, os sintagmas nominais foram etiquetados de acordo com as indicações propostas por Santos e Oliveira (2005). No *corpus*, cada linha apresenta uma palavra seguida de sua etiqueta POS. A palavra é separada de sua etiqueta pelo símbolo "_". Ainda, no final de cada linha é encontrada uma etiqueta do tipo "BIO": "I" para representar o início de um sintagma nominal, "O" para representar o fim, ou ainda "B", representando a ocorrência conjunta do fim do SN anterior e início de um novo. Segue um exemplo de trecho etiquetado do CORSA:

```
...
o_ART_I
Ministério=da=Saúde=do=Brasil_NPROP_I
(_(_O
MS_NPROP_I
)_)_O
lançou_V_O
uma_ART_I
campanha_N_I
nacional_ADJ_I
para_PREP_O
promover_V_O
o_ART_I
uso_N_I
```

```
de_PREP_I
preservativos_N_I
entre_PREP_I
meninas_N_I
adolescentes_ADJ_I
._._O
entre_PREP_I
13_NUM_I
e_KC_I
19_NUM_I
anos_N_I
,_,_O
um_ART_I
grupo_N_I
social_ADJ_I
que_PRO-KS_O
havia_VAUX_O
registrado_PCP_O
crescimento_N_I
em_PREP_O
o_ART_I
número_N_I
de_PREP_I
casos_N_I
de_PREP_I
AIDS_NPROP_I
e_KC_O
outras_PROADJ_I
doenças_N_I
sexualmente_ADV_I
transmissíveis_ADJ_I
._._
```

Nesse formato de *corpus*, não é possível que um sintagma nominal contenha outro, ou seja, não se pode representar aninhamentos de SNs nem, por consequência, empregar regras recursivas ou reentrantes. A escolha de um *corpus* já etiquetado foi realizada com o intuito de diminuir a influência do erro na fase de pré-processamento. Assim, possíveis erros nessa fase não são propagados para a fase de avaliação das extrações, evitando o prejuízo à análise dos resultados.

Com o objetivo de possibilitar o funcionamento com diferentes formatos de *corpus* e ainda facilitar a criação das regras, o *corpus* de entrada é convertido para um formato específico, o qual permite aplicar as expressões regulares, agregando a origem do padrão por autor de referência e número de relações geradas pelo padrão.

O formato adotado aceita sentenças descritas textualmente, com apenas um destaque para os sintagmas nominais. Esses estão entre colchetes, como pode ser visto a seguir:

> "... entre [os municípios maiores] , [Cáceres] e [Rondonópolis] são..."

Esse formato é aplicado a todo *corpus* e, na sequência, cada sentença é adicionada a uma lista para se dar início à próxima etapa, a de aplicação das regras. Nessa etapa a lista de sentenças é percorrida e, para cada sentença, todas as regras são aplicadas em forma de expressões regulares. Quando uma expressão "casa", ou seja, combina com uma sentença, dá-se início à etapa de identificação dos termos da relação.

Ao longo do trabalho de prototipação, foi preciso adicionar diversas regras e alterá-las. Percebeu-se que era preciso simplificar esse processo, já que, até então, era necessário escrever todo o código para a criação e aplicação de cada regra. Assim, foi adotado o conceito do armazenamento de regras em arquivo externo. As regras foram escritas em um arquivo externo, que foi usado como entrada na etapa de aplicação das regras. O arquivo de entrada consiste de um documento JSON (*Java Script Object Notation*), com todas as regras listadas e respectiva identificação da fonte. Esse formato de documento foi adotado por ser um padrão leve, de simples implementação e alta expressividade.

Nessa etapa, a relação já foi identificada na sentença, mas ainda se deve verificar quais dos SNs compõem cada relação extraída, posto que uma regra pode identificar mais de uma relação binária. Além disso, é necessário detectar qual sintagma nominal é o termo hiponímico e hiperonímico da relação. Por fim é gerada uma lista com todas as relações encontradas, no seguinte formato:

> Sentença: {Sentença analisada}
> Extrações: {Autor}-{Padrão} {Nome da Relação}({Argumento1}, {Argumento2}) ...

O campo {Autor} contém a identificação da fonte do padrão que deu origem à regra.

4 Avaliação e discussão dos resultados

4.1 Modelo de processo avaliativo

Durante a execução das etapas de avaliação, diversas dificuldades foram encontradas, entre as quais o grande número de relações extraídas pelo protótipo, ao todo 8.601, o que impossibilitou a análise manual de todas as extrações. Descartou-se a possibilidade de avaliação automática, devido à indisponibilidade de um *Gold Standard* na língua portuguesa com o qual os resultados pudessem ser comparados. Outro motivo que dificultou a execução de uma análise manual foi a falta de uma equipe com o número apropriado de avaliadores para realizar o processo avaliativo. Dois avaliadores realizaram a tarefa de avaliação, ambos com dedicação parcial.

Durante o processo de avaliação de resultados, torna-se necessário situar o trabalho em relação à bibliografia e comparar os resultados com os de outros autores. Na literatura, encontramos poucos trabalhos que realizam a extração de relações em *corpora* de língua portuguesa e, entre esses, não foi possível encontrar resultados que possam ser considerados um *baseline* (base para avaliação) para comparar a precisão e a cobertura. Mesmo assim, para situar o leitor, o texto apresenta na Seção 4.5 alguns dados de precisão e cobertura anunciados em pesquisas nessa área.

Para realizar os testes e a avaliação, o *corpus* CORSA serviu como entrada do protótipo construído. Essa escolha permitiu a comparação de resultados com os descritos por Freitas e Quental (2007), que realizaram a avaliação em dois formatos. No primeiro, as autoras analisaram os resultados dos padrões por elas propostos, individualmente, em busca de erros sintáticos. O objetivo era a eliminação dos erros mais frequentes para cada padrão. Já no segundo formato de avaliação, que se toma como principal referência, foi realizada uma validação humana, em que o foco era tornar os resultados "mais comparáveis" e "mais significativos". A avaliação se deu, no âmbito do trabalho de Freitas e Quental, após essa validação. As relações resultantes receberam notas de 0 a 3, com base nos critérios apresentados pelas autoras e indicados no Quadro 11.

Quadro 11 – Critérios de avaliação empregados por Freitas e Quental

Nota	Descrição
3	*A relação está correta da forma como foi extraída.*
2	*A relação está "um pouco" correta, isto é, o substantivo núcleo está correto, mas preposições, adjetivos etc. que o acompanham deixam a relação estranha.*
1	*A relação está correta em termos gerais; isto é, é muito geral ou muito específica para ser útil.*
0	*A relação está errada.*

Fonte: baseado em Freitas e Quental (2007).

No processo de avaliação desenvolvido por Freitas e Quental, três avaliadores realizaram em conjunto a análise de 436 relações, que representaram um terço das relações extraídas. O consenso entre os três avaliadores determinou a nota atribuída a cada relação. Esses avaliadores tinham formação em biologia, educação física e direito, ou seja, bastante diversificada.

Para avaliar as extrações, a análise foi realizada sobre um subconjunto do total de relações extraídas, composto por todas as 218 extrações obtidas pelas regras indicadas como 6, 7 e 8 no Quadro 7. Essas regras foram escolhidas por terem extraído uma quantidade aceitável de relações (considerando-se o processo avaliativo) e por pertencerem ao conjunto de regras adaptadas do trabalho de Freitas e Quental (2007), o que viabilizaria, em alguma medida, uma comparação. Para esse propósito, dois juízes humanos, sem treinamento prévio, analisaram as 218 relações sob os mesmos critérios utilizados no processo avaliativo de Freitas e Quental. Cada um dos dois avaliadores humanos realizou individualmente a avaliação e atribuiu notas de 0 a 3 às extrações.

As seções 4.2, 4.3 e 4.4 apresentam diferentes propostas de análise quanto à avaliação e às extrações. A Seção 4.5 se detém na

avaliação dos resultados em que houve concordância entre os avaliadores.

4.2 Distribuição das extrações

A aplicação das regras 1 a 11 (Quadro 7) sobre o *corpus* CORSA extraiu 8.601 relações, organizadas em três grupos. Cada grupo compõe-se das relações obtidas com a aplicação das regras sugeridas pelos autores de referência, a saber: grupo 1, regras 1 a 5, provenientes dos estudos de Hearst (1992); grupo 2, regras 6 a 8, exclusivamente de Freitas e Quental (2007); e grupo 3, regras 9 a 11, exclusivamente dos estudos de Taba e Caseli (2014). Nos grupos 2 e 3, tomou-se o cuidado de não se retomarem os padrões de Hearst, implementados pelos respectivos autores de referência. A constituição dos três grupos de regras visou a evidenciar a contribuição específica dos padrões trazidos pelos trabalhos de referência na adição de propostas voltadas à língua portuguesa. A Tabela 1 mostra o número de relações extraídas a partir de cada grupo de regras e o percentual com que cada grupo contribui para o total das relações extraídas. Conforme a tabela, o grupo das regras provenientes de Hearst trouxe a maior contribuição relativa para o resultado, gerando 69,02% das 8.601 relações obtidas, o que já era esperado. As regras propostas por Taba e Caseli constituíram o segundo grupo mais representativo, com 28,45% do total de relações obtidas. Por fim, o grupo das regras sugeridas por Freitas e Quental gerou 2,53% do total das 8.601 relações extraídas.

Tabela 1 – Número de relações extraídas do *corpus* CORSA por grupo de referência

Referência	Número de relações	Percentual
Grupo 1	5.936	69,02%
Grupo 2	218	2,53%
Grupo 3	2.447	28,45%
TOTAL	8.601	100%

Fonte: nossa autoria.[3]

[3] Todas as Tabelas neste trabalho são de nossa autoria.

Ressalta-se, contudo, que tal distribuição não pode ser entendida como uma medida da produtividade das regras, uma vez que se trata de uma ótica de extensão a padrões já consolidados e amplamente empregados, que são os de Hearst, com a capacidade de somar novas relações às já extraídas por aqueles padrões. As extrações produzidas, independentemente dos grupos de regras que as originam, constituem relações igualmente importantes às aplicações, por exemplo, de construção de estruturas de conhecimento, tais como ontologias, ou de extração de informações a partir de bases de dados textuais, tais como repositórios de artigos científicos. Ainda, a distribuição leva em conta exclusivamente o estudo sobre o *corpus* CORSA e pode sofrer influências tanto da adaptação das regras como das próprias ferramentas de processamento que promoveram a etiquetagem do CORSA.

A Tabela 2 exibe o número de relações obtidas com as regras que adaptam os padrões de Hearst e o percentual que essas relações representam em relação ao total de 5.936 extrações que se encontram nesse grupo (cf. TABELA 1). A Seção 2.2 apresenta os padrões originais. A regra 1, que busca extrair relações por meio da palavra-chave "como", gerou um número grande de relações, representando 76,9% das extrações. O resultado já era esperado, pois a palavra-chave em questão é comum na língua portuguesa. Esse grande número de relações extraídas influenciou no resultado obtido com base em padrões de Hearst.

Tabela 2 – Distribuição das relações extraídas por regras adaptadas de Hearst

Regra	Número de relações	Percentual
1	4.565	76,90 %
2	351	5,91 %
3	578	9,74 %
4	376	6,33 %
5	63	1,06 %
TOTAL	5.936	100 %

As regras adaptadas de Freitas e Quental (2007) são baseadas em termos com menor frequência em textos em língua portuguesa e extraem 218 relações. A Tabela 3 mostra, por exemplo, que a regra 6, representada como "<... tipo(s)? de sn> : (SN ,)*(SN (e|ou))*SN", extrai 44,95% das relações.

Tabela 3 – Distribuição das relações extraídas por regras adaptadas de Freitas e Quental

Regra	Número de relações	Percentual
6	98	44,95%
7	75	34,40%
8	45	20,64%
TOTAL	218	100%

A Tabela 4 mostra a distribuição das 2.447 extrações obtidas com a adaptação dos padrões de Taba e Caseli (2014). Algumas regras são abrangentes, por exemplo as que se baseiam em expressões como "é um" e "são", obtendo alto número de relações.

Tabela 4 – Distribuição das relações extraídas por regras adaptadas de Taba e Caseli (2014)

Regra	Número de relações	Percentual
9	23	1,00%
10A	920	37,59%
10B	694	28,36%
11	810	33,10%
TOTAL	2.447	100 %

Ainda analisando as relações extraídas com base em Taba e Caseli (2014), observa-se que a regra 9 apresenta uma quantidade de extrações muito inferior às demais. Essa se baseia na combinação das palavras "qualquer" e "outros", que é menos comum na língua portuguesa, tornando-se uma regra de aplicação menos frequente.

4.3 Avaliação dos resultados

Esta seção apresenta o resultado da avaliação de um subconjunto das extrações obtidas pela aplicação das regras. Embora empregando uma metodologia de avaliação já aplicada, o sentido, mais do que buscar comparações, é prover análises dos resultados obtidos, sob diferentes prismas.

Foi utilizado o processo de avaliação manual dos resultados, também relatado na literatura. Devido ao fato de o total de resultados em foco ser superior a 8 mil extrações, a análise manual tornou-se inviável no tempo disponível, sendo estabelecido um subconjunto de relações para análise. Foram escolhidas as 218 relações extraídas com base nas regras adaptadas de Freitas e Quental (2007) e, com essas, foi possível realizar a avaliação manual. Outro motivo importante para a escolha das relações utilizadas nessa etapa foi a possível comparação de resultados com os relatados em Freitas e Quental (2007).

O Avaliador 1 classificou cada resultado em um de quatro grupos que são representados por notas de 0 a 3, gerando os dados presentes na Tabela 5.

Tabela 5 – Resultado da Avaliação 1: total de relações por nota de avaliação

Nota	Número de relações	Percentual
0	29	13,3%
1	41	18,8%
2	46	21,1%
3	102	46,8%
TOTAL	218	100 %

Analisando a Tabela 5, observa-se que um total de 46,8% de relações extraídas com 100% de correção (nota 3) não é um valor alto. Por outro lado, apenas 13,3% das relações foram consideradas totalmente erradas, o que é um resultado promissor.

Na avaliação feita pelo Avaliador 2 (TABELA 6), os resultados se assemelham com os obtidos na primeira avaliação, com um leve desvio nas relações classificadas com nota 1 e 2, o que pode demonstrar

alguma dificuldade em trabalhar com a escala proposta por Freitas e Quental.

Tabela 6 – Resultado da Avaliação 2: total de relações por nota de avaliação

Nota	Número de Relações	Percentual
0	26	11,9%
1	53	24,3%
2	41	18,8%
3	98	45,0%
TOTAL	218	100, %

Na busca de melhor entendimento do processo avaliativo, a Tabela 7 apresenta uma distribuição das notas atribuídas, considerando o total de 436 avaliações (218 de cada avaliador). Por exemplo, observa-se que, das notas atribuídas, 21,6% foram notas 1, valor que representa a proporção de notas 1 atribuídas no contexto das 436 avaliações.

Tabela 7 – Notas atribuídas pelos avaliadores

Nota	Percentual
0	12,6%
1	21,6%
2	19,9%
3	45,9%
TOTAL	100,%

Outro ponto interessante é a diferença entre os julgamentos atribuídos pelos avaliadores, como mostra a Tabela 8. Ou seja, das 218 avaliações em tela, 109 são coincidentes. É sobre elas que se desenvolve, mais adiante, a avaliação consolidada do trabalho.

Tabela 8 – Comparação entre resultados de julgamento pelos avaliadores

Nota	Avaliações idênticas	Percentual
0	13	11,9%
1	14	12,8%
2	13	11,9%
3	69	63,3%
TOTAL	109	100%

Constata-se que o número de relações que receberam a mesma nota pelos avaliadores é consideravelmente baixo. Esse resultado demonstra uma diferença nos critérios de cada avaliador ao determinar se uma relação está correta, fator que tem um significado importante na avaliação. A Tabela 9 apresenta um exemplo dessa diferença, para as extrações A, B, C e D a seguir:

A Hiponímia (transtorno de a compulsão alimentar periódica; transtorno alimentar)
B Hiponímia (questionário individual de homens; questionários)
C Hiponímia (questionário individual de mulheres; questionários)
D Hiponímia (colinesterase verdadeira; colinesterases)

Tabela 9 – Comparação entre julgamentos para cinco relações específicas

Relação	Avaliador 1	Avaliador 2
A	3	1
B	2	3
C	2	3
D	3	1

Todas as relações de A até E foram avaliadas com nota 3 no processo avaliativo conduzido por Freitas e Quental. Já no processo de avaliação realizado no presente trabalho, essas relações receberam notas distintas. Na Tabela 9, pode-se notar que apenas a relação A obteve o mesmo resultado nas três avaliações (a saber, Avaliador 1, Avaliador 2 e

avaliação relatada por Freitas e Quental). A discordância entre as avaliações sugere que os critérios de julgamento possam ser ambíguos.

Outra forma utilizada para analisar os resultados, provavelmente a mais expressiva, é a que leva em consideração apenas os resultados em que há concordância entre as avaliações providas pelos avaliadores humanos. A título de exemplo, a Tabela 10 traça uma distribuição das extrações conforme notas atribuídas pelos avaliadores exclusivamente nos casos em que houve concordância. Nas colunas, encontram-se as notas de 0 a 3. As linhas agrupam as extrações por regra (6, 7 ou 8) cuja aplicação as originou.

Tabela 10 – Percentual médio de relações encontradas por avaliação e por regra, segundo critério de concordância entre avaliadores

Regra/Nota	0	1	2	3
6	11,1%	9,3%	14,8%	64,8%
7	8,1%	16,2%	10,8%	64,9%
8	22,2%	16,7%	6,25%	55,6%

Considerando os dados na Tabela 10, as regras 6 e 7 apresentam resultados muito semelhantes no que se refere à nota 3. Já a regra 8 apresenta percentuais um pouco inferiores nas notas mais altas, o que pode indicar que ela apresente uma precisão também inferior se comparada com as regras 6 e 7.

É importante ressaltar que, para além dos resultados da avaliação humana (TABELAS 5, 6), optou-se por examinar aqui, cuidadosamente, o comportamento das notas atribuídas pelos avaliadores, tanto no geral (TABELAS 7, 8) quanto em extrações a partir de regras específicas (TABELA 9). A Tabela 10 mostra o comportamento dos dados nos casos em que há concordância entre os avaliadores, situação que será retomada na seção 4.5.

4.4 Uma análise de erros encontrados

Analisando as relações que obtiveram nota 0 de ambos os avaliadores, podemos destacar alguns motivos de erros mais frequentes. Um desses é o erro de *chunking* quando o *parser* realiza uma identificação incorreta. Esse mesmo tipo de erro foi apontado em Corro e Gemulla (2013) como uma das principais fontes de incorreções naquele trabalho. O erro ocorre após a etapa de tokenização, quando o *chunker* identifica os sintagmas nominais. A seguir consta de um exemplo em que o *parser* identificou, incorretamente, a letra "o" como sendo um sintagma nominal:

> "... [dois tipos de modelos] : [o] logístico e [o] hierárquico..."

Em alguns casos, um sintagma nominal pode ser subdividido em SNs menores, sem de fato gerar um erro sintático. Esse comportamento não pode ser considerado uma falha no *chunker.* Tecnicamente, tanto a identificação de um Sintagma Nominal composto (formado por um grupo de SNs) quanto a identificação de apenas um elemento desse conjunto estão corretas, mas esse comportamento gera resultados incoerentes. O trecho a seguir provê alguns exemplos:

> "[o aparecimento de anticorpos] em [o sangue] , chamado de [janela imunológica]".

O *parser* identificou "[o aparecimento de anticorpos]" e "[o sangue]" como dois SNs distintos, gerando uma possível extração errada: Hiponímia (o sangue, janela imunológica). Caso o *parser* identificasse os SNs como um só, uma relação mais precisa poderia ser extraída: Hiponímia (o aparecimento de anticorpos em o sangue, janela imunológica). Para corrigir essa falha, seria preciso de um *chunker* que agrupasse os SNs nesses casos. Outra solução seria prover uma etapa de pré-processamento que unisse *chunks* que podem dar origem a esses casos.

Outro erro encontrado é o de correferência, que ocorre quando o sintagma nominal faz referência a outro SN citado anteriormente na

sentença. Um exemplo pode ser visto no trecho a seguir, no qual o SN faz referência a "corpo":

> "tornar dócil [um corpo] não é [coisa simples], pois ele, normalmente , está submetido a [seu chefe natural], chamado [personalidade]".

Uma extração adequada para essa sentença seria Hiponímia (personalidade, chefe natural do corpo). Uma abordagem para solucionar esse problema seria a utilização de métodos de resolução de correferência.

Outro erro encontrado se refere à falta de contexto. Ele ocorre quando o termo é extraído corretamente, mas ele só faz sentido quando está inserido em um determinado contexto. Segue um exemplo:

> "... [a segunda fase] , chamada de [análise]..."

A regra está correta ao extrair a relação Hiponímia (a segunda fase, análise), mas como não sabemos a que entidade a palavra "fase" faz referência, a extração perde o significado se analisada fora do seu contexto.

Outro erro encontrado está presente na expressão que explora relações formadas por listas de SNs. Tal expressão considera que todos os SNs seguidos por "e", "ou" e "," fazem parte da mesma lista, mas em determinados casos esses conectores podem apenas ligar duas sentenças, não tendo a função de criar lista de sintagmas nominais. Seguem alguns exemplos:

> "[um gênero de vírus] conhecido como [flavivírus] , [a enfermidade] apresenta ..." "[a bactéria] chamada [*Rickettsia mooseri*] e [os sintomas] são praticamente..."

A relação Hiponímia (a enfermidade, um gênero de vírus) é extraída indevidamente, assim como Hiponímia (os sintomas, a bactéria). Apesar de, em ambas as sentenças, o padrão ser aplicado corretamente ao

primeiro SN, o segundo sintagma nominal é considerado indevidamente como parte da lista.

Já quando analisamos as relações apontadas por ambos os avaliadores como pertencendo ao grupo da nota 1, o erro mais comum encontrado é a ocorrência de palavras desnecessárias para o significado da relação dentro de um dos sintagmas nominais. A seguir podem ser vistos exemplos desse fenômeno:

> "[a ação de os vírus] conhecidos como [Influenza A]" "[essas lesões] , chamadas de [isquemia]".

As relações extraídas nesse caso são Hiponímia (a ação de os vírus, Influenza A) e Hiponímia (essas lesões, isquemia). Caso as relações extraídas fossem respectivamente Hiponímia (influenza A, vírus) e Hiponímia (isquemia, lesões), as relações obteriam, acredita-se, uma classificação melhor. Para solucionar esse tipo de problema, Freitas e Quental criaram uma etapa de pós-processamento que aplica filtros para remover palavras dos sintagmas nominais que não agreguem significado à relação. Uma etapa semelhante poderia ser utilizada no trabalho aqui apresentado, com o objetivo de melhorar a precisão. Para isso, no entanto, é essencial dispor de uma lista de palavras que frequentemente não agregam valor semântico, por exemplo, preposições e pronomes.

4.5 Discussão dos resultados

Ao longo das seções 4.2 a 4.4 foram relatados os resultados dos testes realizados. A presente seção discute esses resultados no cenário dos trabalhos correlatos. A precisão das extrações, neste estudo, será tomada como 63%, percentual representado pelas 69 extrações às quais foi atribuída nota máxima por ambos os avaliadores, entre as 218 extrações avaliadas (regras 6, 7 e 8), mostrado na Tabela 8.

A dificuldade de avaliar os resultados por comparação está no uso de regras diferentes pelos diferentes autores, assim como na escolha de *corpora* distintos, além de etapas distintas de pré-processamento ou pós-

processamento. A avaliação humana também tem caráter subjetivo, o que repercute nas avaliações.

A primeira comparação é realizada em relação ao publicado em Freitas e Quental (2007), utilizando, em uma das etapas, o *corpus* CORSA. Em uma das etapas avaliativas, as autoras afirmam obter 73,4% quando aplicam as regras "como / tais como", "e outros", "tipos de", "chamado" e "conhecido como" sobre o *corpus* CORSA.

Já de início destaca-se a inclusão do primeiro padrão, proveniente de Hearst, que seguramente pode trazer um viés de ampliação aos resultados exclusivos das regras 6, 7 e 8 analisadas. A precisão de 73%, expressivamente superior aos 63% do presente trabalho, deve ser relativizada devido à introdução de uma primeira etapa de avaliação, conduzida por Freitas e Quental. Nessa etapa, foi realizada uma análise manual sobre o resultado e foram removidas 726 relações consideradas sintaticamente erradas. Uma tal remoção configura uma filtragem que visa a melhorar a precisão. O resultado de 73,4% obtido por Freitas e Quental considera a etapa de avaliação realizada pela autora sobre extrações no *corpus* CORSA, após a primeira etapa mencionada.

Na conclusão de seu trabalho, Freitas e Quental consideram o resultado final de 75%, esse resultado foi calculado utilizando o *corpus* CETEN-Folha, sem a realização da primeira etapa, em que são removidas manualmente relações sintaticamente errôneas, mas já utilizando filtros propostos com base no estudo sobre o primeiro *corpus*. O primeiro filtro remove relações cujo argumento hiperonímico é substantivo com alto grau de generalidade ou falta de especificidade. Outros dois filtros buscam remover palavras que não agregam valor semântico. Com esse objetivo o primeiro filtro remove pronomes dêiticos, e o segundo remove alguns adjetivos. Observa-se, aqui, um grau de subjetividade na definição dos filtros, que pode ser tema de estudos futuros. Feitas essas observações, a diferença de 73,4% para 75% de precisão necessitaria um aprofundamento, seja com a ampliação de experimentos com diferentes *corpora*, seja com uma sistematização dos filtros empregados.

A título de informação, é interessante trazerem-se resultados obtidos por outros autores como Hearst (1998) e Morin e Jacquemin (2003), sempre levando em conta as diferenças entre *corpora*, processo

avaliativo, regras e também idioma. Analisando a Tabela 12, é possível constatar que, embora todas as diferenças do escopo de avaliação, idioma e *corpora*, que impedem qualquer comparação, a precisão de 63%, obtida no presente estudo, não se distancia da reportada nos estudos de Hearst (1998) nem do trabalho de Cederberg e Widdows (2003). Por outro lado, a precisão alcançada por Morin e Jacquemin (2003), bem superior às demais, reflete uma cuidadosa base de aplicação dos estudos de Hearst para a língua francesa, ampliada com relações específicas associadas ao idioma francês. A precisão relatada também foi produzida com teste sobre um só *corpus*.

Tabela 12 – Precisão reportada em estudos da área

	Corpus em Língua Portuguesa		*Corpus* em Língua Estrangeira		
	Nosso trabalho	Freitas e Quental (2007)	Morin e Jacquemin (2003)	Cederberg e Widdows (2003)	Hearst (1998)
Precisão	63%	73,4%	81%	64%	63%

Ainda assim, existem técnicas, algumas já mencionadas, que poderiam melhorar essa precisão, para chegar aos patamares dos demais trabalhos. Por exemplo, a filtragem de palavras que não agregam valor semântico, ou a inclusão de uma etapa de pré-processamento que una *chunks* em situações específicas (como já exposto). Todas essas técnicas de filtragem, entretanto, baseiam-se em observação e, por consequência, sofrem limitações.

5 Considerações finais

Uma das principais contribuições do presente trabalho é a agregação, num único estudo, de regras elencadas por diferentes autores, como as propostas por Freitas e Quental (2007), Hearst (1992) e Taba e Caseli (2014). Não menos importante é a contribuição da análise minuciosa dos resultados obtidos. Esses foram estudados segundo diferentes critérios, tais como: por regras, por autor, por nota e por avaliador. Ainda foram discutidos fatores que tornam subjetivo o processo de avaliação manual.

Durante o desenvolvimento deste trabalho ficou evidente a necessidade de criação de uma *Gold Standard* para extração de relações hiponímicas na língua portuguesa. Esse artefato contribuiria para o desenvolvimento das pesquisas na área, permitindo o cálculo de precisão e cobertura e comparações mais efetivas. A tarefa, entretanto, teria de contar com a condução de especialistas trabalhando também questões de escopo, contexto e referência, bem além da etiquetagem de relações, esforço que teria de ser amplamente registrado, para formalizar critérios e condutas a ser adotadas.

6 Referências

BANKO, M. *et al*. Open information extraction from the web. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE – IJCAI, 20, 2007, Hyderabad, India. *Proceedings*... Editor: Manuela M. Veloso. Hyderabad, India, January 6-12, 2007. p. 2670-2676.

BASÉGIO, T. L. *Uma abordagem semiautomática para identificação de estruturas ontológicas a partir de textos na língua portuguesa do Brasil*. 2007. 124 p. Dissertação (Mestrado em Ciência da Computação) – Programa de Pós-Graduação em Ciência da Computação, Pontifícia Universidade Católica, RS, 2007.

BATISTA, D. S. *et al*. Extração de relações semânticas de textos em português explorando a DBpédia e a Wikipédia. *Linguamática*, Braga, v. 5, n.1, p. 41-57, 2013.

BICK, E. *The parsing system PALAVRAS* – automatic grammatical analysis of Portuguese in a constraint grammar framework. Aarhus: Aarhus University Press, 2000.

CEDERBERG, S.; WIDDOWS, D. Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In: CONFERENCE ON COMPUTATIONAL NATURAL LANGUAGE LEARNING, 7, 2003, Edmonton. *Proceedings*...

Edmonton: Association for Computational Linguistics, 2003. CONLL, v. 4 p. 111-118. DOI: <http://dx.doi.org/10.3115/1119176.1119191>

CORRO, L.; GEMULLA, R. ClausIE: clause-based open information extraction. In: INTERNATIONAL WORLD WIDE WEB CONFERENCE, 22, 2013, Rio de Janeiro. *Proceedings*… Rio de Janeiro, 2013. p. 355-366.

DEGERATU, M.; HATZIVASSILOGLOU, V. An automatic method for constructing domain-specific ontology resources. In: LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 4, 2004, Lisboa. *Proceedings*... Lisboa, Porutgal, 2004. p. 2001-2004.

FREITAS, M. C. *Elaboração automática de ontologias de domínio*. 2007. 142 p. Tese (Doutorado em Linguística) – Programa de Pós-Graduação em Letras, Pontifícia Universidade Católica, Rio de Janeiro, RJ, 2007.

FREITAS, M. C.; QUENTAL, V. Subsídios para a elaboração automática de taxonomias. In: WORKSHOP DE TECNOLOGIAS DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 5, 2007, Rio de Janeiro. *Anais*... 2007. p. 1585-1594.

GAMALLO, P.; GARCIA, M.; FERNÁNDEZ-LANZA, S. Dependency-based open information extraction. In: JOINT WORKSHOP ON UNSUPERVISED AND SEMI-SUPERVISED LEARNING IN NLP, 9, 2012, Avignon. *Proceedings*... Avignon, France: Association for Computational Linguistics, 2012. p. 10-18.

GRUBER, T. *Ontolingua: a mechanism to support portable ontologies*. 1992. 61p. Technical Report - Knowledge Systems Laboratory, Stanford University, 1992.

HEARST, M. Automatic acquisition of hyponyms from large text *corpora*. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 14, 1992, Nantes. *Proceedings*... v. 2. Nantes, France: Association for Computational Linguistics. p. 539-545. DOI: <http://dx.doi.org/10.3115/992133.992154>

HEARST, M. Automated discovery of WordNet relations. In: FELLBAUM, C. (Org.) *WordNet*: an electronic lexical database. Cambridge: MIT Press, 1998. p. 131-153.

JURAFSKY, D.; MARTIN, J. *Speech and language processing*: an introduction to natural language processing, computational linguistics, and speech recognition. 2. ed. Upper Saddle River: Prentice Hall, 2009.

MAEDCHE, A.; STAAB, S. *Ontology learning for the semantic web*. Massachusetts: Kluwer Academic Publishers, 2002. DOI: <http://dx.doi.org/10.1007/978-1-4615-0925-7>

MORIN, E.; JACQUEMIN, C. Automatic acquisition and expansion of hypernym links. *Computer and the humanities*, Dordrecht, v. 38, n. 4, p. 363-396, 2003.

SANTOS, N.; OLIVEIRA, C. Aplicação de aprendizado baseado em transformações na identificação de sintagmas nominais. In: Congresso da Sociedade Brasileira de Computação, 25, 2005, São Leopoldo. *Anais...* Workshop de Tecnologias da Informação e da Linguagem Humana, 3, São Leopoldo, RS, 2005. p. 2138-2147.

TABA, L.; CASELI, H. Automatic semantic relation extraction from Portuguese texts. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 9, 2014, Reykjavic. *Proceedings...* Reykjavic, 2014. p. 2739-2746.

DOI: 10.17851/2237-2083.23.3.641-680

O léxico do corpo e anotação de sentidos em grandes *corpora*: o projeto *Esqueleto*

The lexicon of the human body and sense annotation: a corpus based study

Cláudia Freitas
Pontifícia Universidade Católica do Rio de Janeiro, RJ; Linguateca
claudiafreitas@puc-rio.br

Diana Santos
Universidade de Oslo, Noruega; Linguateca
d.s.m.santos@ilos.uio.no

Bruno Carriço
Pontifícia Universidade Católica, Rio de Janeiro, RJ[1]
carrico85@gmail.com

Cristina Mota
Linguateca
cmota21@gmail.com

Heidi Jansen
Universidade de Oslo, Noruega
heidi.lanorvegese@gmail.com

Resumo: Apresentamos aqui os resultados iniciais de um amplo estudo sobre o léxico do corpo humano e os seus sentidos, realizado por meio da anotação e revisão de *corpora* de grandes dimensões. Ao longo do artigo explicitamos as decisões linguísticas subjacentes à anotação, relatamos o resultado de um estudo sobre as classes de anotação e exploramos o vasto material criado: um *corpus* de entrevistas (1,4 milhão de palavras) e um *corpus* literário (1,2 milhão de palavras) anotados e integralmente revistos, e demais *corpora* do projeto,

[1] Bolsista de Iniciação Científica da FAPERJ/ E-26/200.812/2015.

parcialmente revistos. Todo o material está publicamente disponível para a comunidade.

Palavras-chave: corpo humano; léxico; anotação semântica; *corpus*; descrição do português.

Abstract: This paper presents the first results of a broad study regarding the lexicon of the human body. The study was based on the annotation of large corpora of Portuguese language. We explain the linguistic annotation choices, present the results of an agreement study and explore the material made available: a corpus of interviews (1.4 million words) and a literary corpus (1.2 million words) full annotated and revised, and the remained corpora partially revised. The whole material is publicly available.

Keywords: human body; lexicon; sense annotation; corpus linguistics; Portuguese.

Recebido em: 30 de junho de 2015.
Aprovado em: 20 de novembro de 2015.

1 Apresentação: motivação e interesses

Cada vez mais é reconhecida a importância da informação semântica associada aos *corpora*, ao mesmo tempo em que é indiscutível a dificuldade envolvida no processo de anotação, não apenas do ponto de vista da anotação automática mas também da anotação manual ou semiautomática.

A importância – e interesse – nesse tipo de informação semântica deve-se à sua vasta aplicação: tarefas do processamento automático de linguagem (PLN) se beneficiam de *corpora* semanticamente anotados e pesquisas linguísticas (voltadas para tradução, ensino de línguas e estudos contrastivos) também.

Em geral, a motivação para a anotação costuma vir de aplicações em PLN: a análise de sentimento / opinião motivou a elaboração de *corpora* anotados quanto a opinião e sentimentos (WIEBE *et al*., 2005; CARVALHO *et al.*, 2011; FREITAS *et al*., 2014; MOTA; SANTOS,

2015, VILLENA ROMÁN *et al*., 2015); tarefas de reconhecimento de entidades mencionadas motivaram a anotação semântica de nomes próprios e relações entre eles, como ilustram iniciativas, como o ACE (DODDINGTON *et al*. 2004) e o HAREM (MOTA; SANTOS, 2008), e o interesse na identificação de relações semânticas entre elementos do texto motivou a criação do PropBank (KINGSBURY *et al*., 2002; PALMER *et al*., 2005) e sua versão brasileira, o PropBank-BR (DURAN; ALUÍSIO, 2012), para citar apenas alguns.

Quando a motivação para a anotação vem, sobretudo, do estudo da língua, a inclusão de informação semântica em *corpus* tem sido mais escassa. Com relação à língua portuguesa, temos conhecimento apenas do projeto C-ORAL Brasil (RASO; MELLO, 2012) e do material do AC/DC (COSTA *et al*., 2009), que hoje contém informação relativa ao campo semântico das cores (FREITAS *et al*., 2012) e do vestuário, e está em andamento a anotação de sentimento (SANTOS; MOTA, 2015). A anotação é uma atividade linguística que poderia ser mais explorada, uma vez que propicia o contato intenso com a língua em uso ao mesmo tempo em que obriga o pesquisador_linguista_anotador a sistematizar (enquadrar nas categorias de anotação) porções dessa língua. A anotação linguística nunca é neutra e, portanto, também interessa teoricamente, pois, (a) ao anotar conforme um dado modelo teórico os pesquisadores_anotadores têm a chance de pôr o modelo à prova, reforçando-o, enriquecendo-o ou refutando-o; ou (b) ao anotar conforme uma dada perspectiva ou motivação (que não necessariamente precisa estar vinculada a uma teoria linguística), os pesquisadores_anotadores têm a chance de testar empiricamente suas hipóteses de categorização acerca do fenômeno investigado.

Se a anotação semântica já é uma tarefa morosa, no viés (b) a morosidade é ainda mais evidente: o processo envolve não apenas a anotação propriamente e toda a discussão inerente a esse processo; envolve uma etapa anterior, de modelagem do fenômeno que se deseja anotar / investigar, ausente quando se parte de categorias pré-estabelecidas por uma teoria específica. Envolve, portanto, idas e vindas, à medida que os dados do *corpus* estão sempre prontos a confrontar nossas pré-concepções, nos fazendo rearranjar e reajustar as categorias-hipótese iniciais, em um processo empírico valoroso à prática linguística.

Este trabalho apresenta os resultados do projeto *Esqueleto*: um esforço conjunto de anotação semântica das palavras do corpo humano, que teve como objetivo geral investigar a estruturação do léxico do corpo na língua portuguesa a partir de sua ocorrência em *corpora*, que vem sendo executado na Linguateca por meio de uma colaboração entre a PUC-Rio e a Universidade de Oslo.

A motivação inicial para esse projeto surgiu de resultados anteriores relacionados à identificação de opinião em resenhas (FREITAS *et al*. 2012), quando foi constatada a grande presença de expressões de opinião associadas a palavras do corpo humano. Investigar as palavras do corpo humano, portanto, contribui também para a descrição de como expressamos opinião em português, ao mesmo tempo em que fornece subsídios para o levantamento de pistas lexicais que devem ser consideradas por sistemas interessados em detectar opinião em textos. Qualquer pessoa interessada na expressão de sentimentos ou opiniões tem de se debruçar, quer queira, quer não, sobre (algumas das) expressões associadas ao corpo humano.

Surgiu também da consciência de que o corpo humano era muito frequente e especial na anotação semântica em português da cor (SILVA; SANTOS, 2012), e das diferenças contrastivas conhecidas entre o português e as línguas germânicas em relação a partes do corpo. Assim, um estudo sobre o léxico do corpo também contribui para questões vinculadas a especificidades culturais. Shigehisa Kuriyama (1999 *apud* GREINER, 2005), em um estudo comparativo entre as diferentes concepções do corpo na China e no Ocidente, explica que a noção de corpo na China

> nunca foi um substantivo (um corpo com nome), e aparece descrita de forma mais próxima de adjetivos (...) [caracterizados] pela descrição de posturas, de atitudes, de gestos, como [...] corpo sentado, corpo em pé, corpo andando, corpo risonho, corpo que chora(...) (GREINER, 2005, p. 22.)

Ainda quanto a especificidades culturais, a indicação de traços de caráter, como alguém "bundão" ou "linguarudo" e expressão de sentimentos, como "nó na garganta" ou "dor de cotovelo", podem trazer

dificuldades para sistemas de apoio à tradução ou para aprendizes de uma segunda língua.

Outra característica relevante do léxico do corpo é sua alta frequência na língua, o que o torna especialmente interessante para estudos baseados em *corpora*. Adicionalmente à ampla ocorrência, uma busca superficial por palavras do corpo humano (e, como amplamente relatado nas abordagens de viés cognitivista (LAKOFF; JOHNSON, 1980) nos revela que boa parte das ocorrências refere-se a usos "não físicos". O interesse na identificação das palavras do corpo, portanto, reside não apenas na exploração de como, em português, falamos sobre o nosso corpo, mas também na investigação sobre a que tipos de outras coisas nos referimos quando usamos palavras do corpo.

Nos estudos pós-estruturalistas vinculados à desconstrução, reconhecendo-se o amplo papel da linguagem na constituição das identificações, o corpo, e especificamente as maneiras de falar do corpo, de referi-lo, são parte dos processos de identificação dos sujeitos, isto é, de marcação social (BUTLER, 2000; PINTO, 2007). Um levantamento amplo das maneiras de caracterizar o corpo – belo, feio, saudável, dócil, áspero, perfumado, alegre – também oferece insumo para explorações nesta perspectiva.

Especificamente, o *Esqueleto* busca responder às seguintes perguntas: (i) quando usamos palavras do léxico do corpo humano, que outros sentidos – que não os do corpo humano – estão em jogo?; (ii) como falamos do corpo humano em português?

Para tanto, conduzimos um processo de anotação semiautomática, que parte de um léxico de palavras do corpo humano para detectar automaticamente as palavras que serão anotadas. A anotação consiste em, considerando as 15 etiquetas (classes semânticas) propostas, atribuir uma ou mais a uma palavra ou expressão, conforme o contexto:

- Sofreu múltiplas fraturas nas duas pernas[sema="corpo"]
- Aí eu fui trabalhando a cabeça[sema="corpo:faculdade"] dele e também a minha pra gente mudar a situação.

Neste artigo, relatamos as opções linguísticas subjacentes à anotação do *Esqueleto*, bem como os resultados deste processo de

anotação, com base na revisão de dois *corpora*: o *corpus* Museu da Pessoa,[2] com 1,4 milhão de palavras, distribuídos em 200 entrevistas que tematizam a vida pessoal dos entrevistados; o *corpus* Obras,[3] atualmente com 1,2 milhão de palavras, composto por 25 obras da literatura brasileira já no domínio público.

Todo o material é público, e está disponível para consulta por meio da interface do serviço AC/DC.[4] Com isso, não apenas apresentamos os resultados de uma pesquisa sobre o léxico do corpo humano baseada em grandes *corpora*, mas também fornecemos à comunidade interessada um rico material para pesquisas futuras.

2 Outros estudos sobre o léxico do corpo humano e o enquadramento do *Esqueleto*

Muitas das pesquisas recentes que se interessam pelo léxico do corpo humano o fazem segundo o viés cognitivista, já que o corpo ocupa um espaço central nesse modelo, estando na base de processos metafóricos (LAKOFF; JOHNSON, 1980). Maalej e Yu (2011) tratam dos usos do léxico do corpo humano em diferentes línguas, em uma perspectiva também contrastiva.

Na perspetiva cognitivista, com relação a estudos envolvendo o corpo e a língua portuguesa, destacamos Soares da Silva (1992) e Leitão de Almeida *et al*. (2009).

Em outras perspectivas, e tratando também apenas da língua portuguesa, Baptista (2000) apresenta o comportamento de descrições de partes do corpo segundo o formalismo da léxico-gramática. Também de um ponto de vista da léxico-gramática, Vale (2013) trata de expressões verbais do tipo [N V (C de N_{hum})[5]], em que C corresponde a uma parte do

---

2 http://www.linguateca.pt/acesso/corpus.php?corpus=MUSEUDAPESSOA

3 http://www.linguateca.pt/OBras/OBras.html

4 http://www.linguateca.pt/ACDC. Esse serviço não só dá acesso a muitos *corpora* cujos donos autorizaram mas também adiciona muita informação linguística, e neste caso semântica, a esse material.

5 Na expressão, C é um substantivo que denomina parte do corpo ou um "elemento inalienável" do N humano. Por exemplo, "(o filme (N_0) encheu(V) o (saco(C) de

corpo humano, levantando a hipótese, a ser posteriormente confirmada em *corpus*, de que a presença de tais expressões fixas seria indicativa da presença de opinião em textos.

Orsi e Zavaglia (2010), agora de um ponto de vista lexicográfico, tratam de expressões idiomáticas tabu que envolvem unidades lexicais obscenas, especificamente referentes à genitália feminina, nas línguas portuguesa e italiana.

Do nosso ponto de vista, o *Esqueleto* é acima de tudo um projeto de anotação de *corpus*. Um projeto de anotação pode estar associado a uma dada teoria linguística, mas também pode estar comprometido mais diretamente com uma motivação empírica, e esse é o nosso caso. Considerando uma das perguntas norteadoras iniciais – quando usamos o léxico do corpo humano, que outros campos do sentido estão em jogo?–, nossa estratégia foi anotar / indicar a presença ou não de um uso associado ao corpo.

Dessa forma evitamos indicar se estamos diante de usos literais ou metafóricos / metonímicos, categorizando as palavras do corpo segundo sua distribuição pelos diferentes campos em que aparecem. Com essa decisão, evitamos, ao longo do processo de anotação, nos posicionar explicitamente no amplo debate sobre a metáfora e a busca de um sentido literal (veja-se ECO, 1991; ARROJO; RAJAGOPALAN, 1992, MARTINS, 2010; para ampla problematização sobre o literal e o metafórico na linguagem).

Mas não negamos os diferentes sentidos que podem ser atribuídos às palavras do corpo como as opções de anotação indicam. Nossa intenção foi agrupar as palavras do corpo segundo seus usos mais gerais. Aos interessados apenas em usos não corporais (metafóricos, segundo a metalinguagem tradicional), é possível uma busca em que se eliminam os demais usos, como será explorado na seção 3.

3 Mão na massa: o processo de anotação

3.1 A escolha do *corpus*

Ana(N_{hum}))", ou "a notícia(N_0) envenenou(V) a (alma(C) de Rui(N_{hum}))". Exemplos extraídos de Vale (2013).

Toda a anotação e estudo conduzidos com o *Esqueleto* tomaram por base os *corpora* do AC/DC. A opção por este material justifica-se, sobretudo, por 3 motivos:

i. livre acesso para pesquisas linguísticas. Dessa maneira, os resultados, bem com toda a documentação das opções de anotação, encontram-se públicos e disponíveis para aqueles interessados em investigar a distribuição e ocorrências em contexto das palavras do léxico do corpo na língua portuguesa;

ii. todo o material já foi anotado morfossintaticamente pelo *parser* PALAVRAS (BICK, 2000), o que se reflete tanto em facilidades no processo de anotação, uma vez que podemos dispor de informações, como lema e classe gramatical, na criação de regras, como em mais possibilidades de exploração – por exemplo, pode-se procurar por verbos que têm como complemento / argumento palavras do léxico do corpo; e

iii. o tamanho e variedade do AC/DC, que conta hoje com mais de 20 *corpora* diferentes, distribuídos em diferentes tipos de texto, como texto jornalístico, acadêmico, entrevistas e obras literárias, totalizando cerca de 1 bilhão de palavras.

Atualmente, já temos dois *corpora* integralmente revistos: o *corpus* OBras e o *corpus* Museu da Pessoa. No entanto, a ênfase na revisão desse material não significa que os demais estejam intocados em relação ao léxico do corpo humano. Devido à maneira como é realizada a anotação, que parte de regras bastante gerais que vão sendo refinadas conforme as ocorrências, todo o material está parcialmente revisto.

3.2 A anotação

O processo de anotação é semiautomático, utilizando uma ferramenta desenvolvida para esse tipo de atividade (SANTOS; MOTA,

2010). As regras são linguisticamente motivadas, e tiramos proveito da informação semântica e morfossintática previamente existente, já incluída no *corpus* pelo PALAVRAS. Em termos gerais, o processo parte de um léxico inicial (no nosso caso, uma lista com palavras do corpo humano, que pode conter palavras simples como "pé" ou expressões como "batata da perna" e "céu da boca") que é aplicado às palavras do *corpus*, anotando-as como palavras relativas ao corpo humano. Em seguida, por meio da análise das palavras inicialmente anotadas, são criadas regras de especialização ou de eliminação, para corrigir casos como "umbigo do mundo", que receberá uma etiqueta semântica específica, e "coluna social", em que "coluna" será desconsiderada como palavra do corpo.

A definição das categorias de anotação consiste no principal desafio do *Esqueleto*. Como mencionamos, optamos por não partir de categorias determinadas aprioristicamente. A estratégia utilizada consistiu em, considerando a observação das ocorrências em *corpus*:

i. estabilizar a primeira grande classificação relevante para as motivações do *Esqueleto*: palavras do corpo humano que se referem ao corpo humano *vs.* palavras e expressões do corpo humano que se distribuem por outros campos semânticos.

ii. a partir da análise das ocorrências, criar subclasses que organizassem as palavras e expressões do corpo humano por outros campos semânticos.

Quando a palavra do corpo é usada para fazer referência ao corpo humano, recebe a etiqueta corpo (procurável por meio da pesquisa [sema="corpo"]). Nos outros casos, receberá semas específicos, na forma[6] [sema="corpo:xxx"], conforme o seu uso.

Considerando a análise integral de dois *corpora* e a análise parcial dos demais *corpora* do AC/DC, temos hoje 15 classes, ou semas, estáveis. O Quadro 1 apresenta os semas, com exemplos de uso.

[6] Daqui em diante usamos a expressão da pesquisa para referir o nome da categoria.

Nesta etapa, cuidamos apenas dos substantivos e adjetivos do corpo.

Quadro1 – Classes semânticas do corpo humano no *Esqueleto*[7]

Sema	Exemplos
corpo	torceu o *pé* na corrida; ter *olhos* azuis
corpo:animal	*orelha* de porco;
corpo:centralidade	Seu departamento é o *cérebro* da operação; Sem revelar o *coração* do plano, Itamar rebatizou o conjunto de medidas(...)
corpo:doenca	não tenho medo do *pé* de atleta; esta medonha epidemia de *bexiga*
corpo:faculdade	uma provocação plástica para *olhos* e *ouvidos* livres
corpo:grupo	*corpo* docente; *coluna* do exército
corpo:lugar	no *coração* da floresta amazônica
corpo:medida	dois *dedos* de pinga; onda de 3 *pés*
corpo:movimento	ir a *pé*; assim que pôs os *pés* na cidade
corpo:opiniao	ele é um *bundão*; tem sempre um *orelhudo* na conversa
corpo:parte	*boca* do fogão; *membro* do Parlamento; *braço* da máfia
corpo:posicao	suplicou de *joelhos*; dormiu em *pé*
corpo:sentimento	com o *coração* apertado; o meu *sangue* ferve por vocês
corpo:vegetal	*dente* de alho; *pé* de jabuticaba
corpo:outros	*boca* da noite; uma *veia* pop

Uma vez que não partimos de categorias familiares a uma teoria específica, explicitamos a seguir as motivações e princípios que nortearam a estabilização de uma classe semântica.

A partir de uma exploração inicial do *corpus*, iniciamos o processo de anotação com cinco classes candidatas: corpo, opinião, sentimento, lugar e outros. À medida que a anotação e revisão avançaram, fomos criando novas classes.[8] Nesse processo, para que um sema se estabilizasse como categoria de anotação, ele deveria englobar (i) diferentes palavras do corpo que compartilhassem o mesmo tipo de sentido; ou (ii) poucas palavras do corpo, mas com um uso muitíssimo frequente e sistemático.

[7] Todos os quadros, tabelas, gráficos e figuras neste trabalho são de nossa autoria.

[8] Uma nova classe só era criada após uma discussão – e consenso – entre todos os envolvidos na anotação e, a cada alteração, havia revisões retroativas, a fim de garantir uniformidade.

Adicionalmente, tínhamos duas preocupações: (iii) evitar uma classificação muito granular do sentido, o que além de levar a um imenso número de classes, poderia contribuir para uma maior discordância quanto ao conteúdo de cada classe – classes mais gerais têm mais possibilidades de acomodar nuances de sentido; e (iv) não inchar a classe "outros" com usos sistemáticos.

A palavra boca ilustra bem nossa preocupação. Sempre que é usada com o sentido de entrada, atribuímos o sentido de lugar [sema="corpo:lugar"].

1. recebeu livre na **boca** da área
2. terá de se identificar na **boca** do caixa

Observando as demais ocorrências de boca, reparamos que, em alguns casos, o sentido de entrada, espacial, vai se direcionando para o sentido de tempo, vinculado a início: por onde se chega ou por onde se começa, em um deslizamento entre espaço e tempo:

3. Na **boca** da safra, as commmodities estão perdendo o fôlego.
4. A conversão no setor acontecerá na **boca** da entressafra, quando a oferta...
5. Que o faça, no entanto, todos os dias do ano, não apenas quando o país está à **boca** da urna, e nos limites da lei.

Encontramos, até o momento, apenas esses três usos / sentidos nos *corpora*, além do uso / sentido sistemático em boca da noite [6] (88 ocorrências em todo o material do AC/DC). Por isso, não criamos [sema="corpo:tempo"], e as ocorrências [3-5] estão anotadas como [sema="corpo:outros"]. Se encontrarmos ainda mais casos de boca ou de outros lemas associados a tempo, é possível que este novo sema seja criado.

6. Primeiro trabalham, depois vão à escola e depois brincam, no fim do dia, na **boca** da noite .

A documentação do *Esqueleto* (FREITAS, 2013) especifica e exemplifica cada uma das classes. Detalhamos aqui os casos que consideramos menos óbvios ou mais interessantes.

- O [sema="corpo:faculdade"] refere-se sobretudo aos 5 sentidos - visão, olfato, paladar, tato, audição – além da faculdade / capacidade do pensamento, frequentemente associada às palavras do corpo "cabeça", "cérebro" ou "miolo". Aos poucos, o sema faculdade foi se ampliando, e atualmente refere-se também a processos realizados pelo corpo, nomeados pelas partes que os realizam – "pulmão", por exemplo, pode ser usado como sinônimo do processo / capacidade de respiração; e "boca" ou "garganta" podem ser usadas para indicar a faculdade / capacidade de falar [7-10].

7. Mas o Stanley tinha **cabeça** para dinheiro, o que eu nunca tive
8. Cafu e Mazinho constituem um meio-de-campo bom de **pulmão**, dinâmico e criativo
9. Posso ser mau de **boca** mas sou bom de **olho**
10. Existem **ouvidos** atentos para cada acorde, cada nota, cada timbre.

- A presença dos sentidos [sema="corpo:sentimento"] e [sema="corpo:opiniao"] indica que, no *Esqueleto*, consideramos sentimento e opinião duas classes distintas, ainda que frequentemente apareçam juntas. A separação foi uma decisão tomada desde o início do projeto. No entanto, a partir de um dado momento, uma das integrantes do projeto pôs em questão a separação. Em favor da unificação, a constatação de que pesquisadores de várias perspectivas chamaram a atenção para que as emoções e as opiniões (ou seja, a atividade do intelecto e do(s) sistemas associado(s) às emoções) estão indissociavelmente ligados. Por isso, na prática, os trabalhos de "análise de sentimentos" e de "garimpo de opiniões" constituem a mesma (sub)disciplina (veja-se MAIA; SANTOS, 2015). A expressão de atitudes, sentimentos e opiniões está intimamente relacionada com o corpo humano, por (pelo menos) três razões diferentes:

a. porque as emoções (pelo menos algumas, as consideradas por alguns psicólogos como básicas) provocam alterações fisiológicas

b. porque a descrição das emoções ou atitudes passa convencionalmente pela descrição da postura do seu dono (facial ou corporal): ficar de boca aberta, encolher os ombros...

c. porque é uma característica (aparentemente universal) das línguas, visto que estas se originaram em tempos sem conhecimento médico preciso, usar órgãos metonimicamente para sentimentos ou características de personalidade (maus fígados, bom coração...)

De toda forma, no *Esqueleto*, mantemos a distinção, e usamos [sema="corpo:opiniao"] para os casos em que o próprio termo ou expressão do corpo se refere a algo já com uma indicação clara de posicionamento [11]. O sentido [sema="corpo:sentimento"], por sua vez, é atribuído quando não há posicionamento ou julgamento explícitos [12]. Em ambos os casos, localizamos a presença do sentimento e / ou opinião apenas na palavra ou expressão, e não no enunciado completo.

11. Ele é um **pé de valsa**

12. Meu **coração** partiu quando ele se foi

- Distinções e sobreposições entre os sentidos [sema="corpo:lugar"], [sema="corpo:centralidade"] e [sema="corpo:parte"]. O sentido [sema="corpo:lugar"] é atribuído a palavras do corpo que se referem a um lugar, como [13-14]. No entanto, diversas outras ocorrências de “coração”, “seio” e “cérebro” em contextos como [15-16] em que está em jogo a ideia de centralidade, puseram em xeque a utilização do sema corpo:lugar, que comporta apenas uma das dimensões de “centro”, deixando de fora as dimensões de espacialidade e importância.

13. Bem no **coração** da floresta amazônica, a cidade é realmente uma bolha.
14. Ele nasceu em São Pedro Alfa, ao **pé** de Coimbra.
15. Deve, também, chegar ao **coração** da sociedade civil, desmascarando atitudes...
16. plantar a desordem no **seio** da família,

Segundo o dicionário Aulete Digital,[9] a palavra "centro" comporta os seguintes sentidos (entre vários outros):

> (a) Ponto que se situa no meio de uma superfície, de uma área ou de um espaço, tendo exata ou aproximadamente a mesma distância das extremidades ou limites
>
> (b) Localidade, região etc. de grande importância em relação às áreas vizinhas, onde se concentram atividades econômicas, administrativas e/ou políticas etc. (ger. especificadas pelo adj.): *A região sudeste é há anos o centro do país*.
>
> (c) Parte ou aspecto principal, mais importante, mais difícil etc.

A acepção (a) dá conta do sentido de "lugar"; a acepção (b) lida com as dimensões de importância e lugar; e a acepção (c) evidencia apenas a dimensão de importância. Para a ideia de "lugar", já dispomos do sema corpo:lugar. Para lidar com o sentido de "importância", criamos [sema="corpo:centralidade"]:

17. Seu departamento é o **cérebro** da operação
18. Sem revelar o **coração** do plano, Itamar rebatizou o conjunto de medidas...

[9] http://www.aulete.com.br

- O sentido [sema="corpo:parte"], além dos casos clássicos, como "pé da cama" e "costas da cadeira", também engloba partes de um todo que não precisa, necessariamente, ser entendido como um objeto. Inicialmente, chamava-se [sema="corpo:partedeobjeto"], mas a ocorrência de diversos casos de partedeobjeto em contextos em que o todo não se caracterizava como objeto [19-21] nos obrigou a repensar a descrição da classe. Com essa opção, unificamos os casos distintos de parte indiferenciada de algo e parte diferenciada, também em coerência com o princípio de evitar classificações demasiado granulares.

19. o **braço** europeu do Cinema Novo
20. para ser o **braço** de financiamento
21. Os **membros** da expedição reuniram-se

Os demais semas listados no Quadro 1 estão exemplificados na documentação.

3.2.1 Classificações múltiplas e vagueza na anotação

É possível que uma mesma palavra, em um mesmo contexto, admita duas ou mais classificações simultaneamente. Nesses casos, aceitamos todas as possibilidades. A palavra "dorso", por exemplo, corresponde à "parte de cima" em [22]. "Parte de cima", indica, simultaneamente, parte de algo (o que justifica o sentido [sema="corpo:parte"]) e localização espacial (o que justifica [sema="corpo:lugar"]). Esses casos, portanto, recebem ambos os semas. Os exemplos [23-26] também ilustram casos que receberam mais de uma classe semântica.

22. as tochas dos penitentes, e a procissão, estendida na linha de cumeadas, traçou uma estrada luminosa no **dorso** [sema="corpo:lugar_parte"] da montanha
23. uma foto de uma fantástica **boca** [sema="corpo:lugar_parte"] de caverna, que é a Gruta dos Brejões

24. e Yusef foram portanto, na tese de Milroye, os **cérebros** [sema="corpo:faculdade_centralidade"] do atentado
25. A primeira notícia refere uma ilha no **coração** [sema="corpo:lugar_centralidade"] da cidade do Porto.
26. Teve gente que **torceu o nariz** [sema="corpo:sentimento_opiniao"], mas eu gostei

3.2.2 Expressões com várias palavras e a relação entre identificação e classificação

É conhecida a participação de palavras do corpo em expressões, bem como a multiplicidade de nomes para o fenômeno de combinação com muitas palavras e a dificuldade quanto a uma abordagem consensual do que sejam tais expressões – nada muito diferente do construto linguístico "palavra", que pode ser estudado por diferentes ângulos (som, forma e sentido), que nem sempre serão convergentes. Em geral, o termo *locução* corresponde a uma perspectiva que privilegia a unidade em torno do sentido (MATTOSO CÂMARA JR., 1984), o termo "colocação" privilegia a convencionalidade / frequência de coocorrência, motivado sobretudo pela ideia de fluência verbal (SINCLAIR, 1991; MANNING; SCHÜTZE, 1999), e os termos "expressão idiomática*"* e "expressão cristalizada*"* privilegiam a opacidade semântica e a fixidez (XATARA *et al*., 2002, por exemplo). No *Esqueleto*, nomeamos tais combinações de EVP (expressões com várias palavras).[10]

Se consideramos a dimensão da convencionalidade, temos uma EVP como "conhecer=como=a=palma=da=mão", pois a vasta maioria dos usos de "como a palma da mão" acontece com "conhecer". Pelo critério do sentido único, podemos argumentar que "conhecer como a palma da mão" engloba duas ideias, com a especificação do sentido de "conhecer". Assim, pelo critério do sentido único, a EVP é apenas "como=a=palma=da=mão" e essa é a opção do *Esqueleto*. Considerando a enorme quantidade de expressões que incluem alguma palavra do

[10] O nome EVP também foi escolhido para que pudéssemos diferenciar nossas combinações das combinações já marcadas no *corpus* pelo PALAVRAS, que são chamadas de MWE.

léxico do corpo humano, limitamos o escopo das combinações apenas às expressões que correspondem a uma unidade de sentido.[11] Do mesmo modo, em "lembrar de cabeça", "contar de cabeça" e "saber de cabeça", consideramos apenas uma EVP, "de=cabeça". Nos exemplos [27-30] indicamos com = exemplos do que consideramos EVPs.

27. daqui ninguém **arreda=pé**
28. E tem gente que ainda **fica=na=mão**?
29. Uma delas seria o **braço=direito** do traficante Djavan
30. Ela é quem vai **ganhar=o=coração** do personagem Pedro

No entanto, o critério "sentido único" também é escorregadio em certos casos [31-34].

31. Tenha coração, não use peles.
32. Ele é bom, tem um coração enorme e me trata como uma princesa.
33. O que importa é que fulano tem bom coração.
34. Hoje, ter princípios é como ter bom coração.

É possível entender "ter coração" e variações como "ter (bons) sentimentos", e neste caso não haveria EVP, apenas a atribuição de [sema="corpo:sentimento"] a "coração", ou podemos entender como *ser* "generoso", e neste caso temos uma EVP do tipo [sema="corpo:opiniao"].

O mesmo ocorre em "doente da cabeça" [35]. Uma leitura possível atribui a "doente da cabeça" o sentido de louco – um sentido único, portanto. Nesse caso, "doente da cabeça" seria considerado EVP do tipo "outros". Outra leitura atribui apenas a "cabeça" um sentido não convencional – ficar doente das ideias/faculdades mentais ou algo parecido. Mantém-se o sentido de doente, com o sintagma seguinte

[11] O critério da unidade de sentido é do ponto de vista do falante da língua em análise. Assim, por exemplo, não consideramos "ponta dos dedos" uma EVP, ainda que a combinação corresponda a uma única palavra / sentido em inglês (*fingertips*).

especificando a doença. Nessa leitura, não temos a marcação de uma EVP, apenas a anotação de *cabeça* como [sema="corpo:faculdade"].

35. Inclusive ela ficou até doente da cabeça por causa dele.

Idealmente, em ambos os casos ("bom coração" e "doente da cabeça") deveríamos poder considerar todas as leituras possíveis, sem privilegiar alguma delas, mas isso tornaria tecnicamente o processo de anotação de expressões mais complexo do que já é. Nossa preferência, em todos esses casos [36-39], foi por considerar a leitura EVP, privilegiando o sintagma maior.

As EVPs, sobretudo verbais, com frequência admitem variações no formato. Nesses casos, atribuímos um lema único para os casos, isto é ver=com=bons=olhos em [36-37] e abrir=os=olhos=para em [38-39].

36. Ver com bons olhos

37. Ver com muito bons olhos

38. Ter olhos abertos para

39. Abrir os olhos para

Ainda outro ponto quanto à anotação das expressões no *Esqueleto* é a marcação em dois níveis: na EVP e nas palavras do corpo que compõem a EVP. Em [40], "dor de cotovelo*"* é anotada como uma EVP com o sentido de *sentimento*. Nessa EVP, a palavra do corpo, "cotovelo", também é anotada, e recebe [sema="corpo:outros"]. Já em [41], atribuímos o sentido de "sentimento" à EVP (um sentimento de calma; foi para casa acalmar-se), mas "cabeça" recebe [sema="corpo:faculdade"].

40. (...)s quatro últimos emprestam suas vozes em faixas que fazem o ouvinte entrar numa **dor de cotovelo** gostosa

41. (...) foi pra casa **esfriar a cabeça**

Por fim, lembramos que a anotação é sempre feita em contexto. Os trechos [42-44] ilustram três casos com a palavra "olho". As palavras ligadas por = correspondem a uma EVP.

42. O **olho** [sema="corpo:lugar_parte"] do furacão Gonzalo tocou a terra nas ilhas Bermudas.
43. No **olho** [sema="corpo:lugar_parte"]**=do=furacão** [sema="corpo:outros"], FHC faz que não é com ele.
44. Josias é que foi para o **olho** [sema="corpo:lugar"]**=da=rua**[sema="corpo:outros"]

A motivação da anotação em dupla camada é tornar as pesquisas com o léxico do corpo humano mais ricas. Assim, temos a possibilidade de facilitar a procura, em contexto, de expressões do corpo que envolvem um determinado sentido de uma palavra (por exemplo, todas as EVPs que contêm a palavra "boca"*)* ou de buscar as palavras do corpo que participam de expressões (pode-se querer investigar a existência de relação entre palavras do corpo e a classe semântica das EVPs em que participam, por exemplo), como será visto na seção 4.

Como sabemos que nossas interpretações não são definitivas, mas somos forçados, pela anotação, a tomar decisões, lembramos que todas as combinações consideradas EVPs estão listadas na página do projeto ou podem ser recuperadas em uma busca pelo AC/DC. Assim, para além da anotação em *corpus*, as EVPs identificadas estão listadas em um arquivo específico, em que para cada EVP, além da informação da classe semântica, é atribuída uma classe gramatical.

3.2.3 Estudo e validação das classes propostas

Ao longo do processo de anotação e estabilização dos semas, algumas classes nos pareceram especialmente delicadas, como ilustramos no início desta seção.

A fim de verificar o quão consensuais eram as análises propostas, por um lado, e o quão clara era a documentação, por outro, utilizamos a ferramenta Rêve (SANTOS *et al.*, 2015), desenvolvida justamente para permitir a revisão e comparação de diferentes análises. Assim, utilizamos o Rêve, não para aferir uma concordância entre anotadores, mas para, sobretudo, revisar e discutir os casos considerados complicados e, como consequência, fornecer uma medida de nossas escolhas e decisões.

Criamos dois exercícios de anotação distintos, com cerca de 100 frases cada, distribuídas de maneira equilibrada pelas diferentes classes. As frases foram escolhidas de acordo com o grau de dificuldade, isto é: escolhemos preferencialmente frases que, ao longo do processo de anotação, suscitaram dúvidas. Nem todas as frases eram frases difíceis, mas escolhemos uma amostra representativa das dúvidas e discordâncias que tivemos ao longo do processo de anotação. Além disso, evitamos escolher frases em que o sema seria atribuído a uma EVP, para evitar que a discordância na segmentação embaralhasse classificação de semas e identificação da EVP. Em alguns casos isso foi quase impossível, pois há semas que comparecem quase exclusivamente em expressões. Nesses casos, a opção foi por EVPs consensuais, que dificilmente dariam margem a dúvidas quanto à segmentação, como "torcer o nariz" ou "jogar na cara".

Em cada exercício, testamos grupos distintos de classes, mas que, segundo nossa experiência na anotação, continham sobreposições. Além disso, cada grupo incluiu também duas classes em comum, com frases repetidas.

No *Esqueleto* 1, testamos 100 frases distribuídas pelas classes corpo, sentimento, opinião, faculdade, centralidade e outros (além das possibilidades de classificação múltipla).

No *Esqueleto* 2, testamos 92 frases, distribuídas pelas classes corpo, lugar, parte de objeto, centralidade, posição, sentimento, e outros (além das possibilidades de classificação múltipla).

Não consideramos as classes "animal", "vegetal", "medida" e "grupo", por não terem gerado qualquer discussão ao longo do processo de anotação, o que seria indicativo de seu pouco potencial para discordância.

Os participantes do exercício tinham perfis distintos: participaram quatro anotadores: duas com experiência no *Esqueleto* e autoras deste artigo, e dois que tomaram contato com o *Esqueleto* pela primeira vez com o exercício.

Após a primeira rodada de anotação, e observando as discordâncias, refinamos algumas explicações e reanotamos. A alteração do nome "parte de objeto" para apenas "parte", por exemplo, foi consequência dessa discussão posterior.

Apenas o *Esqueleto* 2 passou por uma segunda rodada de anotação. A concordância foi de cerca de 87%, com 15 frases discordantes. Dessas 15, três eram claramente casos em que todas as alternativas / interpretações propostas eram possíveis. A anotação foi então revista e corrigida.

Dos 12 demais casos de discordância, todos envolviam a atribuição da categoria "outros" por alguma das anotadoras, o que, por um lado, revela que nossas intuições quanto à dificuldade dos exemplos selecionados estava correta e, por outro, sugere que os casos em questão eram pouco típicos para a sua classe, e bem poderiam compreender a classe "outros" (o que foi feito, e indicado na documentação). Também nesses casos de discordâncias (com "outros"), boa parte das frases envolvia a atribuição do sentido de *centralidade*.

Em termos gerais, e considerando a variedade de classes, ficamos satisfeitas com os resultados do exercício de anotação / revisão, que indicaram que houve consenso entre a classificação proposta: atingimos cerca de 87% de concordância, sendo que nos casos discordantes não houve propostas radicalmente diferentes, mas antes a atribuição da classe "curinga".

Quanto às concordâncias (79 frases), notamos que em 21 delas havia vagueza envolvida, isto é: se mais de uma classe havia sido proposta por uma participante, e as demais escolheram uma única classe, mas uma classe que já havia sido prevista na classificação múltipla, consideramos que não houve discordância. Nas restantes 58 frases, a concordância foi absoluta, mesmo quando havia mais de uma classe proposta.

3.2.4 Os grupos do corpo

Além das classes semânticas, às palavras do corpo também são atribuídos grupos, de acordo com a zona do corpo a que pertencem. Só recebem informação de grupo as palavras ou expressões do corpo do tipo [sema="corpo"], ou seja, não atribuímos grupo aos usos "especiais".

Atualmente, temos dez grupos, por ordem alfabética: Braços, Cabeça, Cabelos, Interno, Ossos, Pele, Pernas, Sangue, Sexual e Tronco.

É possível também que uma palavra pertença a mais de um grupo. A palavra ombro, por exemplo, pertence aos grupos Tronco e Braço.

A motivação para atribuir uma "zona" corporal proveio de já existir essa possibilidade para os campos cor e vestuário (não sendo, portanto, necessária nenhuma programação adicional) e por ser relativamente fácil essa atribuição (quando as palavras se referem ao corpo humano). Daí poderíamos oferecer aos usuários ainda outro tipo de procuras. Assim, pode-se investigar se existem áreas mais mencionadas que outras, ou áreas que têm mais ou menos usos vinculados ao corpo, por meio da informação dos grupos.

4 Resultados

Embora seja difícil considerar finalizado um projeto com essas dimensões, acreditamos que já temos material suficiente para divulgar resultados sobre a estruturação do léxico do corpo humano.

Revisamos dois *corpora* com características bastante distintas, entrevistas pessoais (Museu da Pessoa, 1.421.677 palavras e 93.479 frases) e obras literárias (OBras, 1.204.436 palavras e 38.011 frases), e os demais foram parcialmente revistos. Nas explorações a seguir, além do material integralmente revisto, relatamos também resultados que levam em conta o *corpus* Floresta (FREITAS *et al*. 2008), um *corpus* majoritariamente jornalístico, e o *corpus* Todos, a união de todos os *corpora* disponibilizados pelo AC/DC.

As Tabelas 1 e 2 apresentam a distribuição dos usos das palavras do corpo por *corpus* e por sentido. A diferença é que, na Tabela 1, não levamos em conta a classe gramatical das palavras do corpo, e na Tabela 2 consideramos apenas as palavras do corpo que são substantivos.

Os resultados são bastante parecidos em todos os campos, exceto em dois aspectos: quando consideramos apenas os substantivos (TABELA 2), a proporção de palavras do corpo mais que quadruplica, o que está de acordo com nossa estratégia de anotação, que privilegiou substantivos. Nesse contexto, a proporção de palavras do corpo no OBras é muito maior que nos demais *corpora*: cerca de 5% de todos os substantivos do *corpus* referem-se a palavras do corpo.

De maneira geral, os dados nas Tabelas 1 e 2 nos permitem inferir que a distribuição das palavras do corpo, nos *corpora*, segue a seguinte ordem: o material mais corporal é o Obras, seguido pelo Floresta, Museu da Pessoa e Todos. No entanto, é importante lembrar que tanto o Floresta como o Todos foram apenas parcialmente revistos.

Tabela 1 – Distribuição de palavras do corpo pelo total de palavras, sem considerar a classe gramatical

Corpus	Total de palavras do corpo	Distribuição dos semas do corpo	
		Tipo de sema	quantidade
OBras	**1,04%**	corpo	**85%** (10649 ocorrências)
		corpo:xxx	**15%** (1931 ocorrências)
MP	**0,17%**	corpo	**50,40%** (1218 ocorrências)
		corpo:xxx	**49,50%** (1198 ocorrências)
Floresta	**0,2%**	corpo	**80%** (12948 ocorrências)
		corpo:xxx	**20%** (3159 ocorrências)
Todos	**0,2%**	corpo	**87%** (2575969 ocorrências)
		corpo:xxx	**13%** (385794 ocorrências)

Tabela 2 – Distribuição de palavras do corpo pelo total de palavras, considerando apenas a classe gramatical dos substantivos

Corpus	Total de palavras do corpo	Distribuição dos semas do corpo	
		Tipo de sema	quantidade
OBras	**5%**	corpo	**85%** (10565 ocorrências)
		corpo:xxx	**15%** (1921 ocorrências)
MP	**1%**	corpo	**50,40%** (1170 ocorrências)

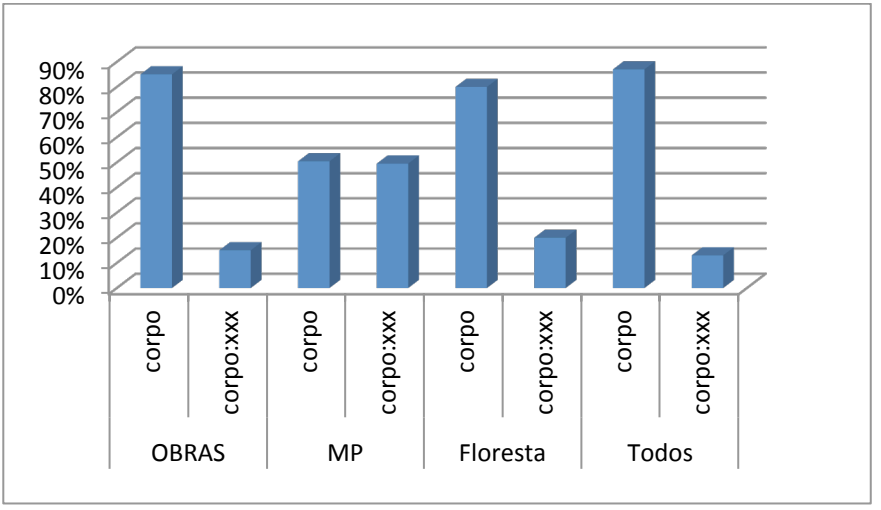
		corpo:xxx	**49,50%** (1146 ocorrências)
Floresta	**1,27%**	corpo	**80%** (12215 ocorrências)
		corpo:xxx	**20%** (3113 ocorrências)
Todos	**0,8%**	corpo	**87%** (2059946 ocorrências)
		corpo:xxx	**13%** (384880 ocorrências)

Outro dado interessante visível nas tabelas é a constância da proporção corpo / corpo:xxx, com cerca de 85% de palavras do corpo para usos corporais. A exceção é o MP, em que a distribuição corpo / corpo:xxx é equilibrada, com 50% das ocorrências para cada um dos usos. O Gráfico 1 apresenta apenas essa distribuição.

Considerando agora apenas o material completamente revisto, temos dois cenários bastante distintos. No OBras, apenas 15% das palavras do corpo não se vincula ao corpo, o que corrobora a ideia de forte presença de descrição nos textos literários analisados.[12] Já no Museu da Pessoa a situação é bem diferente: apenas metade das palavras do corpo se refere ao corpo. No Floresta e no Todos, a proporção de corpo / corpo:xxx é parecida: cerca de 80% das palavras do corpo fazem referência ao corpo, o que nos surpreende um pouco - esperávamos um uso maior de corpo:xxx. No entanto, esses *corpora* não passaram por uma revisão completa, e é possível que esse quadro se altere.

Gráfico 1 – Distribuição dos tipos de sentido das palavras do corpo por *corpus*

[12] Vale mencionar ainda que as obras com maior ocorrência de palavras do corpo são romances brasileiros do chamado período realista/naturalista, especificamente os romances *O Mulato* e *O Cortiço*, de Aluísio Azevedo, *Quincas Borba* e *Dom Casmurro*, de Machado de Assis, e o *Ateneu*, de Raul Pompéia.



O Gráfico 2 apresenta a distribuição dos semas corpo:xxx (pelo total de semas corpo:xxx), considerando apenas o material totalmente revisto.[13] Considerando apenas o OBras, vemos que o sentido mais frequente é o de sentimento – impulsionado pelos usos de "coração"–, seguido de outros e de posição, este último também típico de descrições.

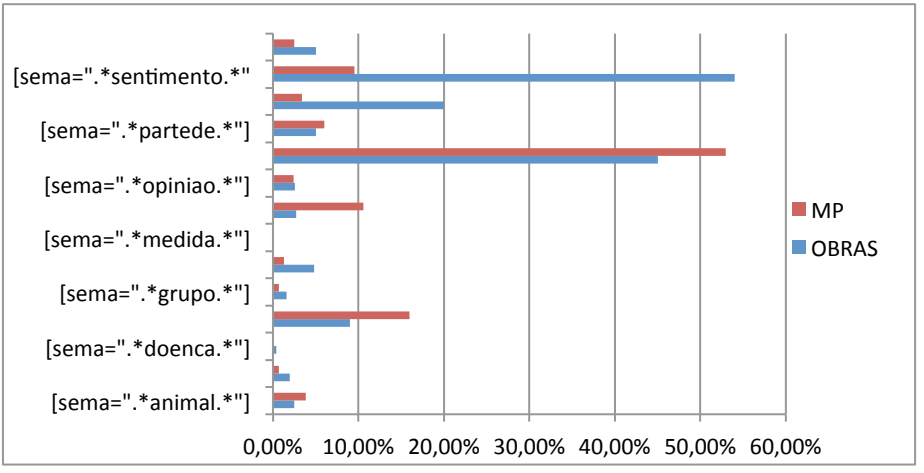
Chamou-nos a atenção o relativamente frequente uso de corpo:vegetal no OBras, e percebemos que a imensa maioria refere-se à palavra *tronco*. No entanto, dessas, boa parte se refere ao tronco em que os escravos eram castigados, o que aparece em obras como *Escrava Isaura*, *O Mulato* e *O Cortiço*. No Museu da Pessoa, o uso mais frequente é [sema="corpo:outros"], seguido de [sema="corpo:faculdade"]; [sema="corpo:movimento"] e [sema="corpo:sentimento"].

No Quadro 2, em uma abordagem qualitativa, apresentamos os lemas que tomam parte em alguns dos semas corpo:xxx. Para o quadro, consideramos o material do OBras, MP e também da Floresta. Como é

[13] Para simplificar, as palavras consideradas vagas entre várias classificações contam em cada uma delas, ou seja, se uma palavra estiver marcada [sema="corpo:opiniao_sentimento"], conta nesta tabela uma vez por corpo:opiniao e outra por corpo:sentimento. As expressões de procura utilizadas foram: [sema=".*movimento.*"], [sema=".*sentimento.*"] etc.

possível observar, há palavras do corpo especialmente maleáveis quanto ao sentido, que participam de todos os semas (ou quase todos), como "pé", "boca" e "mão". No Quadro 2, os lemas estão listados por ordem alfabética, e não por frequência. É importante notar também que, no Quadro 2, estamos considerando apenas os lemas, dissociados das EVPs de que fazem parte. Assim, por exemplo, "dente" integra a EVP "com unhas e dentes". A ideia do quadro é tão somente apresentar a variedade de palavras do corpo utilizada nos diferentes sentidos. É interessante perceber que a ideia de importância / centralidade, que normalmente associaríamos apenas a cabeça / cérebro, também pode estar associada ao coração – que normalmente associaríamos apenas ao sentimento – e ao umbigo.

Gráfico 2 – Distribuição dos semas corpo:xxx no OBras e no Museu da Pessoa



Especificamente quanto às expressões, e ilustrando as possibilidades da anotação em dupla camada, a Tabela 3 apresenta a distribuição dos sentidos das EVPs,[14] ou seja, sem considerar as palavras do corpo que as compõem.

[14] As expressões de procura foram [sema=".*corpoEVP"], [sema=".*corpo:sentimentoEVP"], etc.

Nos três *corpora*, expressões com sentidos que não conseguimos sistematizar – o que corresponde ao "outros" – são as mais frequentes, consistindo em cerca de metade dos sentidos das expressões. Comparando OBras e MP, expressões com o sentido de corpo são muito mais recorrentes no OBras, o que, novamente, é compatível com a ideia de que, neste material, as descrições físicas são bastante usuais. Outro dado interessante é que a distribuição dos sentidos das EVPs segue, de maneira geral, a distribuição dos sentidos total.

Nas Tabelas 4 e 5 apresentamos a distribuição das palavras do corpo nas EVP. Na Tabela 4, listamos, por frequência, as palavras do corpo que mais comparecem em expressões, mas mantendo o sentido de palavra do corpo, independentemente do sentido da expressão (e excluindo as expressões com sentido de corpo humano)[15]; na Tabela 5, fazemos o inverso: listamos as palavras do corpo que mais comparecem em expressões, mas perdem o sentido de corpo.[16]

Quadro 2 – Lista de lemas por semas considerando apenas OBras, MP e Floresta

SEMA	LEMAS
[sema=".*centralidade.*"]	cabeça; coração; cérebro; regaço; seio; umbigo
[sema=".*faculdade.*"]	boca; cabeça; coração; cérebro; língua; mão; nervo; olho; pulmão; orelha; ouvido
[sema=".*lugar.*"]	boca; coração; costas; estômago; face; fronte; olho; pé; seio
[sema=".*movimento.*"]	pé
[sema=".*opiniao.*"]	boca; barriga; cabeça; cara; coração; cotovelo, desmiolado; estômago; língua; mão; nervo; olho; osso; pé; saco
[sema=".*parte.*"]	boca; braço; cabeça; corpo; costas; dente; dorso; espádua; goela; membro; olho; peito; perna; punho; pé; seio

[15] Com a expressão <mwe> []* @[sema=".*corpo_.*EVP"] []* </mwe> conseguimos todas as expressões em que há uma palavra do corpo com o sentido de corpo. Como excluímos as expressões cujo sentido global é corpo, a expressão de busca usada foi <mwe> []* @[sema=".*corpo_.*EVP" & sema!=".*corpo_.*corpoEVP"] []* </mwe>.

[16] Para essa busca, a expressão usada foi <mwe> []* @[sema=".*EVP" & sema!=".*corpo_.*EVP" & sema!=".*corpo_.*corpoEVP"] []* </mwe>.

[sema=".*posicao.*"]	braço; cabeça; cara; costas; face; ilharga; joelho; punho; pé; punho
[sema=".*sentimento.*"]	barriga; boca; cabelo; cabeça; cara; coração; corpo; costas; cotovelo; dedo; dente; estômago; garganta; mão; nariz; nervo; olho; ombro; orelha; ouvido; peito; pele; pé; queixo; sangue; sobrolho; tripa; tropinha; unha; venta
[sema=".*outros.*"][17]	artéria; barriga; boca; braço; busto; cabeça; cara; carne; celular; coração; corpo; costas; célula; dedo; dente; embrionário; embrião; esqueleto; face; franja; língua; manual; mão; olho; osso; ouvido; palma; peito; perna; pulso; pé; rabo; sangue; seio; tronco; umbigo; unha; veia

Tabela 3 – Distribuição das expressões com várias palavras (EVP) nos corpos, sobre o total de EVP

Classificação	**OBras**	**MP**	**todos**
corpo	**8,8%** (76)	2% (11)	**14%** (23354)
corpo:animal	0	0,2% (1)	0,1% (155)
corpo:doenca	0,2% (2)	0 (0)	0,2% (458)
corpo:faculdade	0	0,4% (2)	0,04% (76)
corpo:lugar	0,3% (3)	0,2% (1)	0,07% (123)
corpo:movimento	3% (26)	**12%** (64)	1% (1619)
corpo:opiniao	2,9% (25)	4% (24)	1,2% (2111)
corpo:outros	**45%** (385)	**59%** (301)	**60%** (99275)
corpo:posicao	**28%** (241)	**8%** (40)	**15%** (25378)
corpo:sentimento	**10%** (85)	**11%** (59)	**5%** (9416)

É possível constatar a variação nos lemas. "Cabeça", por exemplo, quando comparece em expressões, costuma a perder o sentido do corpo. "Pé", por outro lado, parece sofrer o efeito inverso: tende a manter o seu sentido de corpo, mesmo quando em expressões. Já as

[17] Como ilustração, e considerando a variedade de lemas em cada corpus consideramos apenas as 25 primeiras ocorrências de cada corpus, sugerindo ao leitor interessado repetir a busca.

palavras "mão" e "olho" participam de expressões de ambas as maneiras.[18]

Tabela 4 – Distribuição das palavras do corpo que mantêm o sentido corporal em expressões

MP		OBras		Floresta	
pé	38,8% (97)	pé	35% (159)	mão	32% (281)
mão	26,4% (66)	mão	22% (100)	pé	21% (189)
cara	7,6% (19)	olho	12% (54)	olho	7% (61)
olho	4,4% (11)	braço	5% (26)	cara	6% (57)
boca	3,2% (8)	boca	4% (21)	palma	3% (33)
braço	2,8% (7)	cara	4% (20)	saco	3% (28)
corpo	2,4% (6)	cabeça	3% (13)	boca	2% (21)
saco	2% (5)	palma	2% (9)	costas	2% (20)
joelho	2% (5)	costas	1% (8)	língua	2% (18)
pele	1,6% (4)	língua	1% (7)	cabeça	2% (17)

Tabela 5 – Distribuição das palavras do corpo que não mantêm o sentido corporal em expressões

MP		OBras		Floresta	
cabeça	27% (63)	joelho	20% (56)	cabeça	17% (109)
mão	19% (45)	olho	9% (25)	olho	16% (101)
olho	11% (26)	mão	8% (22)	mão	14% (90)
cara	5% (12)	cabeça	7% (21)	cara	10% (62)
boca	5% (12)	ouvido	7% (20)	ouvido	8% (53)
coração	5% (11)	coração	6% (17)	corpo	5% (30)

[18] Notamos que a anotação de expressões em dupla camada permite essas e ainda outras possibilidades de busca na interface. Para encontrar expressões que contêm uma palavra específica, por exemplo, mão, a expressão deve ser <mwe> []* @[lema="mão"] []* </mwe>, que devolverá EVPs como *deitar a mão*, *mão na massa*, *de mão beijada*, de *segunda mão*, *abrir mão* etc. Para especificar não apenas a palavra mas também o sentido da EVP, a expressão deve ser <mwe> []* @[lema="mão" & sema=".*sentimento.*"] []* </mwe>.

[19] Expressão de pesquisa: [sema=".*corpo.*"] e pedido de Distribuição de grupo

costas	2% (6)	dente	5% (16)	pé	4% (29)
pé	2% (5)	punho	4% (12)	boca	4% (27)
corpo	1% (4)	boca	4% (12)	punho	3% (20)
sangue	1% (4)	ilharga	3% (10)	dedo	2% (12)
peito	1% (4)	costas	3% (9)	coração	1% (10)

Levando em conta agora apenas as palavras do corpo em seus usos corporais, a Tabela 6 ilustra, para MP, OBras e Floresta, a distribuição de lemas do corpo sobre o total de palavras do corpo (não consideramos corpo:xxx). Consideramos apenas os 10 lemas mais frequentes de cada *corpus*.

Há diferenças significativas entre as ordenações, sobretudo entre um *corpus* literário e um de linguagem oral (que, além disso, são separados por um século: os escritores do OBras são fundamentalmente do século 19; enquanto os entrevistados são todos dos séculos 20-21). Debruçando-nos agora sobre algumas das diferenças mais gritantes: a palavra "olho" (ou "olhos") é muito mais usada no OBras devido, novamente, à descrição de personagens em textos literários – os "olhos" (e também as "mãos") atuam como elemento de descrição física (olhos castanhos), mas também psicológicas (olhos astutos; mãos trêmulas). No Floresta, um *corpus* majoritariamente jornalístico, a ampla frequência da palavra "corpo" refere-se aos óbitos, que optamos por deixar anotado como [sema="corpo"]. Nos três, chama a atenção o grande comparecimento de "mão".

Tabela 6 – Distribuição dos lemas de corpo por *corpus*

MP		OBras		Floresta	
mão	14%	olho	12%	mão	10%
pé	10%	mão	10%	corpo	8%
cabeça	6%	cabeça	6%	olho	7%
perna	5%	coração	5%	cabeça	7%
cabelo	5%	braço	5%	pé	5%
braço	4%	corpo	4%	cara	4%
costas	3%	pé	4%	coração	4%
olho	3%	rosto	3%	cabelo	3%
cara	3%	boca	3%	boca	3%
mama	3%	cabelo	2,6%	braço	2,6%

Finalmente, observamos a distribuição por grupos das palavras de corpo que identificamos, na Tabela 7. Convém indicar que palavras referentes a algo que não diz respeito a uma parte do corpo, como a própria palavra "corpo", estão marcadas com zona (grupo) Não especificada.

Inesperadamente, o grupo Interno é extremamente frequente, na mesma ordem de grandeza que a Cabeça, e o Tronco tem quase o dobro de ocorrências que os membros (quer Braço quer Perna), enquanto que Pele e Cabelo são dos menos usados. Será preciso estudar com mais cuidado todos os gêneros que constituem o corpo total para podermos compreender melhor a causa desses resultados quantitativos. Para um estudo inicial dos gêneros em questão, veja-se também Santos (no prelo).

Tabela 7 – Distribuição das palavras referentes a corpo humano por grupos (zonas)[19]

Grupo	**OBras**	**MP**	**todos**
Braço	2330	345	355828
Cabeça	5130	392	739374
Cabelo	655	84	61770
Interno	1414	173	744269
Osso	94	38	97741
Não especificada	592	49	341575
Pele	90	15	73504
Perna	1331	350	204311
Sangue	262	29	103380
Sexual	81	45	64653
Tronco	1795	238	549171

Por fim, o último levantamento tem como objetivo explorar a ideia do corpo (e do corpo, na língua) como mais uma maneira de

[19] Expressão de pesquisa: [sema=".*corpo.*"] e pedido de Distribuição de grupo

elaborar identificações. Na Figura 1 ilustramos as caracterizações (adjetivos) atribuídas às palavras do corpo, apenas quando estas se referem ao corpo humano.[21] Destacam-se – as palavras maiores – dimensões físicas: espaço (direito, esquerdo), tamanho (baixo, enorme, comprido, curto, pequeno, grande) e cor (branco, vermelho, negro, moreno, negro, castanho, preto). O passo seguinte, que não fizemos aqui, é encontrar, para cada atributo / classificação, as partes do corpo caracterizadas. Nesta breve exploração, consideramos apenas o *corpus* OBras.

Figura 1 – Adjetivos que caracterizam as palavras com sentido de corpo no OBras

[21] A expressão de busca utilizada foi [sema="corpo|corpo_corpoEVP"] @[pos="ADJ.*" & func="<PRED|N<.*"], com pedido de distribuição por lemas.

5 Considerações finais

Apresentamos aqui um estudo voltado ao léxico do corpo humano em língua portuguesa, que tomou como metodologia a anotação de *corpus* e que oferece, como resultado, não apenas análises e interpretações mas também um vasto material, publicamente disponível, para outras explorações.[22]

No decorrer do processo de anotação, criamos classes e categorizamos fenômenos, sempre de um ponto de vista que privilegia o sentido. Ao longo deste texto, explicitamos nossas decisões e motivações. Não temos a ilusão de interpretações definitivas, mas acreditamos ter sido suficientemente explícitos em nossas explicações, permitindo a continuação do debate, o que é facilitado pelo compartilhamento de um mesmo material – especificamente, do mesmo material anotado. Afinal, o corpo humano, e o modo pelo qual nos referimos a ele, interessa a estudos que se alinham a diferentes perspectivas teóricas.

À pergunta inicial do projeto: como se distribui, quanto ao sentido, o léxico do corpo humano em português?, oferecemos a resposta provisória de 13 diferentes sentidos sistemáticos, excetuando-se o próprio sentido corpo e os sentidos que não conseguimos sistematizar, indicados com "outros".

Quanto ao vínculo entre sentimento, opinião e corpo, mostramos que, sem dúvida, é forte e presente na língua portuguesa (como nas demais línguas), mas consideramos as possibilidades dos demais sentidos um rico caminho para explorações futuras, assim como a relação entre cada um dos sentidos e as palavras a eles associadas.

Podemos, então, nos debruçar sobre a segunda pergunta: como falamos do corpo humano em português?, o que só pode ser feito (ou é imensamente facilitado) depois de termos separado as palavras / sentidos do "corpo" dos demais sentidos que também se utilizam das palavras do corpo. Tomando por base os resultados do MP, por exemplo, vemos que

[22] O estudo apresentado aqui foi feito com a versão 2.1 do Obras e versão 5.7 do MP. Pode ser que esse material se amplie no futuro e, com isso, os números precisem ser atualizados.

essa primeira separação é fundamental, visto metade das palavras e expressões do corpo referirem a outros tipos de sentido.

Dissemos no início deste artigo que anotar é uma forma de estudo. A anotação motivada pelo sentido, atividade que precisa compatibilizar sentido e forma, nos força a delimitações que não nos são "naturais", mas antes necessidades decorrentes da própria atividade de anotação. No caso das expressões, altamente frequentes no estudo do léxico do corpo humano, o estudo da anotação opera como uma lente de aumento para a multiplicidade de nuanças que devem ser levadas em conta quando o que está em jogo é o processamento automático da língua e a sua descrição.

Quais as palavras do corpo mais frequentes em expressões? Leitão de Almeida *et al*. (2009) referem, em sua pesquisa, "cabeça", "mão" e "pé". Nossos resultados vão na mesma direção, ainda que apresentem uma variedade de outras palavras também comuns em expressões.

Todo o material criado, bem como os léxicos (listas de palavras e regras) utilizados na anotação, estão disponíveis na página do projeto.[23] Ao longo do artigo, indicamos nas notas de rodapé as expressões que utilizamos na interface de busca, com o intuito de motivar a/os leitoras/es a realizar novas pesquisas e / ou análises. Assim, os interessados nos variados sentido do corpo na língua têm à disposição um material rico e preciso, bastando para isso fazer uso dos diferentes tipos de sema. Para os especialmente interessados nos usos metafóricos, não sistemáticos, a grande quantidade de "outros" indica que há muito ainda por explorar. Para os interessados em como, na língua, identificamos pessoas, quer do ponto de vista físico, quer psicológico, o espaço de exploração é muito vasto.

23 http://www.linguateca.pt/acesso/Esqueleto/

Referências

ARROJO, Rosemary; RAJAGOPALAN, Kanavillil. Noção de literalidade: metáfora primordial. In: ARROJO, Rosemary (Org.). *O signo desconstruído*. São Paulo: Pontes, 1992. p. 47-56.

BAPTISTA, Jorge. Body-part nouns and local grammars. In: DISTER, Anne. (Ed.). *Révue d'Informatique et Statistiques en Sciences Humaines*, v. 36, p. 53-66, 2000.

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

BUTLER, Judith. Corpos que pesam. In: LOURO, Guacira Lopes (Org). *O corpo educado*: Pedagogias da sexualidade. Belo Horizonte: Autêntica, 2000. p.110-125.

CARVALHO, Paula; SARMENTO, Luis; TEIXEIRA, Jorge; SILVA, Mário J. Liars and Saviors in a Sentiment Annotated Corpus of Comments to Political Debates. In*:* ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 49, 2011, Stroudsburg, PA, USA. *Proceedings…,* v. 2. Stroudsburg: Association for Computational Linguistics, 2011. p. 564-568.

COSTA, Luís; SANTOS, Diana; ROCHA, Paulo. Estudando o português tal como é usado: o serviço AC/DC. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY – STIL, 7, 2009, São Carlos. *Proceedings…* São Carlos: Universidade de São Paulo, 2009.

DODDINGTON, George; MITCHELL, Alexis; PRZYBOCKI, Mark; RAMSHAW, Lance; STRASSEL, Stephanie; WEISCHEDEL, Ralph The Automatic Content Extraction (ACE) Program: Tasks, Data, and Evaluation. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 4, 2004, Lisboa. *Proceedings of LREC'2004*. Ed. M. T. Lino; M. F. Xavier; F. Ferreira; R. Costa; R. Silva, Lisboa, Portugal: Universidade Nova de Lisboa, 2004. p. 837-40.

DURAN, Magali S.; ALUÍSIO, Sandra M. Propbank-Br: a Brazilian Treebank annotated with semantic role labels. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 8, 2012, Istanbul. *Proceedings of LREC'2012*. Ed.: N. Calzolari; K. Choukri; T. Declerck; M. U. Dogan; B. Maegaard; J. Mariani; J. Odijk; A. Moreno; S. Piperidis. Istanbul: Lüfti Kirdar Convention & Exhibition Centre, 2012. p. 1862-1867.

ECO, Umberto. *Semiótica e filosofia da linguagem*. São Paulo: Ática, 1991.

FREITAS, Cláudia. *Esqueleto: anotação das palavras do corpo humano*. Primeira edição: 15 nov. 2013. Disponível em: <http://www.linguateca.pt/acesso/Esqueleto/Esqueleto.html>

FREITAS, Cláudia; MOTTA, Eduardo; MILIDIÚ, Ruy L.; CÉSAR, Juliana. Sparkling Vampire... lol! Annotating Opinions in a Book Review Corpus. In: ALUÍSIO, Sandra; TAGNIN, Stella E. (Ed.). *New Language Technologies and Linguistic Research*: A Two-Way Road. Cambridge Scholars Publishing, 2014, p. 128-146.

FREITAS, Cláudia; SANTOS, Diana; SILVA, Rosario. (2012). Corpos e cores: colorindo a descrição da língua portuguesa. In: ENCONTRO DE LINGUÍSTICA DE CORPUS: ASPETOS METODOLÓGICOS DOS ESTUDOS DE CORPORA, 10, 2012, Belo Horizonte. *Anais*... Ed.: D. P. Dutra; H. R. Mello. Belo Horizonte: Faculdade de Letras da UFMG, 2012. p. 76-99.

FREITAS, Cláudia, ROCHA, Paulo; BICK, Eckhard. Um mundo novo na Floresta Sintá(c)tica – o treebank para Português. *Calidoscópio*, v. 6.3, p. 142-148, 2008. DOI: <http://dx.doi.org/10.4013/cld.20083.03>

GREINER, Christine. *O Corpo* – Pistas para estudos indisciplinares. São Paulo: Ed. Annablume, 2005.

KINGSBURY, Paul; PALMER, Martha, MARCUS, Mitch. Adding Semantic Annotation to the Penn TreeBank. In: THE HUMAN LANGUAGE TECHNOLOGY CONFERENCE, 2002, San Diego. *Proceedings*... San Diego, CA, USA, 2002.

KURIYAMA, Shigehisa. *The Expressiveness of the Body, and the Divergence of Greek and Chinese Medicine*. New York: Zone Books, 1999.

LAKOFF, George; JOHNSON, Mark. *Metaphors we Live By*, Chicago: The University of Chicago Press. 1980.

LEITÃO DE ALMEIDA, M. L. *et al*. (Org.) A hipótese de corporificação da categorização e do léxico. In: LEITÃO DE ALMEIDA, Maria Lúcia *et al*. (org.). *Linguística Cognitiva em foco: morfologia e semântica do português*. Rio de Janeiro: Publit, 2009, p. 187-204.

MAALEJ, Zouheir A.; YU, Ning (Ed.). *Embodiment via Body Parts*: Studies from Various Languages and Cultures. *Human Cognitive Processing*, v. 31. Amsterdam and Philadelphia: John Benjamins, 2011. DOI: <http://dx.doi.org/10.1075/hcp.31>

MAIA, Belinda; SANTOS, Diana. Emotions in Language, PhD-course, University of Oslo, 1-5 Jun. 2015. (Mimeo)

MANNING, Christopher; SCHÜTZE, Hinrich. *Foundations of Statistical natural language processing*. Cambridge, MA: The MIT Press, 1999.

MARTINS, Helena. Wittgenstein, the body, its metaphors. *D.E.L.T.A.*, São Paulo, v. 26, p. 479 – 501, 2010. (Edição Especial)

MATTOSO CÂMARA JR, J. *Dicionário de Linguística e Gramática*. Editora Vozes, 1984.

MOTA, Cristina; SANTOS, Diana (Ed.). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*: o segundo HAREM. Linguateca, 2008. Disponível em: <http://www.linguateca.pt/LivroSegundoHAREM/>

MOTA, Cristina; SANTOS, Diana. Emotions in natural language: a broad-coverage perspective. *Linguateca*, Jan. 2015. Disponível em: <http://www.linguateca.pt/acesso/EmotionsBC.pdf>.

ORSI, Vivian; ZAVAGLIA, Claudia. Expressões idiomáticas interditas: uma proposta lexicográfica bilíngue. *Linguasagem Revista Eletrônica de*

Popularização Científica em Ciências da Linguagem, v. 11, p. 1-17, 2010.

PALMER, Martha; GILDEA, Dan; KINGSBURY, Paul. The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics Journal*, v. 31, n. 1, p. 71-106, 2005.

PIAO, S.; ARCHER, D.; MUDRAYA, O.; RAYSON, P.; GARSIDE, R.; McENERY, A.; WILSON, A. A large semantic lexicon for corpus annotation. In: CORPUS LINGUISTICS CONFERENCE, 2005, Birmingham. *Proceedings*... Series on-line e-journal, v. 1, n. 1, Birmingham, UK, July 14-17, 2006. ISSN 1747-9398.

PINTO, Joana P. Conexões teóricas entre performatividade, corpo e identidades. *D.E.L.T.A.*, São Paulo, v. 23, n. 1, p. 1-26, 2007.

RASO, Tomaso; MELLO, Heliana. (Org.). C-ORAL-BRASIL I. *Corpus de referência do português brasileiro falado informal*. Belo Horizonte: Editora UFMG, 2012.

SANTOS, Diana. Podemos contar com as contas? In: ALUÍSIO, Sandra Maria; TAGNIN, Stella E. O. (Ed.). *New Language Technologies and Linguistic Research*: A Two-Way Road. Cambridge Scholars Publishing, 2014, p. 194-213.

SANTOS, Diana. Comparando corpos orais (transcritos) e escritos na Gramateca. In: PARLER LES LANGUES ROMANES / PARLARE LE LINGUE ROMANZE / HABLAR LAS LENGUAS ROMANCES / FALANDO LÍNGUAS ROMÂNICAS. Napoli. *Atti del convegno internazionale GSCP 2014*. Ed.: Camilla Bardel; Anna De Meo. Napoli: Università di Napoli L'Orientale, Il Torcoliere, no prelo.

SANTOS, Diana. Gramateca: corpus-based grammar of Portuguese. In: In: INTERNATIONAL CONFERENCE – PROPOR – COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 11, Oct. 6-8, 2014. *Proceedings*... Ed. J. Baptista; N. Mamede; S. Candeias; I. Paraboni; T. Pardo; Maria das Graças V. Nunes. São Carlos: Springer, Heidelberg, 2014. p. 214-219.

SANTOS, Diana; MOTA, Cristina. A admiração à luz dos corpos. In: SIMÕES, A.; BARREIRO, A.; SANTOS, Diana; SOUSA-SILVA, R.; TAGNIN, Stella E. O. (Ed.) *Linguística, Informática e Tradução*: Mundos que se Cruzam. Homenagem a Belinda Maia, *OSLa*, v. 7, n. 1, p. 57-77, 2015.

SANTOS, Diana; SILVA, Rosário; FREITAS, Cláudia. Pluralidades na cor: contrastando a língua do Brasil e de Portugal. In: SOARES DA SILVA, Augusto; TORRES, Amadeu; GONÇALVES, Miguel. (Ed.). *Línguas Pluricêntricas*: Variação Linguística e Dimensões Sociocognitivas. [Pluricentric Languages: Linguistic Variation and Sociocognitive Dimensions.] Braga: Aletheia, Publicações da Faculdade de Filosofia da Universidade Católica Portuguesa, 2011. p. 555-572.

SANTOS, Diana; MARQUES, R; FREITAS, Cláudia; SIMÕES, A.; MOTA, Cristina. Comparando anotações linguísticas na Gramateca: filosofia, ferramentas e exemplos. *Domínios de Lingu@gem*, v. 9, n. 3, 2015.

SANTOS, Diana; MOTA, Cristina. Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC), 7, 2010, Valletta, Malta. *Proceedings*... Ed. N. Calzolari; K. Choukri; B. Maegaard; J. Mariani; J. Odjik; S. Piperidis; M. Rosner; D. Tapias (Ed.). Valletta, Malta: Mediterranean Conference Centre, 2010. p. 1437-1444.

SILVA, Rosário; SANTOS, Diana. *Arco-íris*: notas sobre a anotação do campo semântico da cor em português. 16 ago. 2012. Disponível em: <http://www.linguateca.pt/acesso/ArcoIris.pdf>.

SINCLAIR, J. *Corpus, concordance, collocation*: Describing English language. Oxford: Oxford University Press, 1991.

SOARES DA SILVA, Augusto. Metáfora, Metonímia e Léxico. *Diacrítica*, v. 7, p. 313-330, 1992.

VALE, Oto. As opiniões nas expressões e a expressão da opinião. In: LAPORTE, Éric; SMARSARO, Aucione; VALE, Oto. (Org.). *Dialogar*

é preciso: Linguística para processamento de línguas. Vitória: PPGEL/UFES, 2013. p. 259-267.

VILLENA ROMÁN, J.; GARCÍA MORERA, J.; MARÍNEZ CÁMARA, E.; JIMÉNEZ ZAFRA, S. M. TASS 2014 – The Challenge of Aspect-based Sentiment Analysis. *Procesamiento del Lenguage Natural*, v. 54, p. 61-68, marzo 2015.

WIEBE, Janyce; WILSON, Theresa; CARDIE, Claire. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, v. 39, n. 2-3, p. 165-210, 2005.

XATARA, Claudia M.; RIVA, Huelinton. C.; RIOS, Tatiane H. C. As dificuldades na tradução de idiomatismos. *Cadernos de Tradução*, Florianópolis, NUT, v. 8, p. 183-194, 2002.

DOI: 10.17851/2237-2083.23.3.681-693

A tradução de suculentos jogos de palavras, sem perder o sabor

Translating juicy wordplays without losing their flavor

Stella E. O. Tagnin
Universidade de São Paulo (USP), São Paulo, SP.
seotagni@usp.br

Resumo: Este artigo discute algumas questões levantadas durante a tradução para o português de títulos de receitas de um livro britânico de sucos para crianças, com especial atenção para as diferenças culturais envolvidas nesse tipo de tradução. A maioria dos títulos baseia-se em jogos de palavras ou algum tipo de metáfora - imagética ou conceitual – que se espera sejam mantidos na tradução, de modo a preservar o apelo que esses títulos têm para o público infantil, pois, em geral fazem referência ao universo da criança. Linguisticamente, os títulos envolvem colocações, rimas e aliterações. Culturalmente, há referências a jogos infantis, canções e heróis que, na maioria das vezes, não são as mesmas na nossa cultura. Alguns títulos são baseados na cor do suco (*Red devil*), outros no ingrediente principal (*Crazy kiwi*), outros ainda no efeito desejado (*Battery charge*). A tradução de jogos de palavras exige certo grau de elaboração semântica, especialmente quando várias camadas de significado estão envolvidas. Serão apresentados vários exemplos que envolvem diferenças culturais entre a Inglaterra e o Brasil, assim como as soluções a que se chegou, sempre na tentativa de manter o jogo de palavras dentro do universo infantil.

Palavras-chave: jogos de palavras; tradução; cultura; títulos; convencionalidade.

Abstract: This paper discusses various issues raised in the Portuguese translation of titles of juice recipes for children from a British recipe book, paying special attention to cultural differences involved in this type of

translation. Most titles involve either wordplay or some kind of metaphor – image or conceptual – which was expected to be retained in the translation so as to preserve the appeal these titles have for children as they usually make reference to the child's universe. Linguistically, titles involve collocations as well as rhymes and alliteration. Culturally, references are made to children's games, songs and heroes which, more often than not, are different across cultures. Some titles are based on the color of the juice (*Red devil*), some on the main ingredient (*Crazy kiwi*), some on the desired effect (*Battery charge*). When these convey prototypical meanings, translation can be quite straightforward. Wordplay, on the other hand, demands a certain degree of semantic elaboration, especially when various layers of meaning are involved. We will present various examples that involve cultural differences between Britain and Brazil, as well as the solutions we arrived at always attempting to somehow reproduce the wordplay in the child's universe.

Keywords: wordplay; translation; culture; titles; conventionality.

Recebido em: 30 de julho de 2015.
Aprovado em: 2 de setembro de 2015.

1 Introdução

Traduzir, em si, como todos sabem, não é uma tarefa trivial. Quando a tradução envolve jogos de palavras, então, a dificuldade é muito maior, principalmente porque nem sempre – ou quase nunca– há uma equivalência linguística entre as duas línguas que permita reproduzir o efeito desejado na língua de chegada. A situação pode se tornar ainda mais complicada quando entram em jogo aspectos culturais típicos de uma língua e inexistentes na outra. Outro agravante ainda poderia ser representado pelo público-alvo, isto é, quando se trata de um público específico.

Foram esses os problemas enfrentados na tradução de um livro britânico de receitas de sucos saudáveis para crianças intitulado *Juices & Smoothies for Kids* (CROSS, 2007a), que recebeu o título *Sucos e Vitaminas para Crianças* em português (CROSS, 2007b).

2 O livro

O livro *Juices & Smoothies for Kids* contém mais de sessenta receitas, cujos títulos pretendem ser apelativos ao público infantil, empregando muitas vezes jogos de palavras, certamente para incentivar as crianças a consumir o alimento. Cada receita é introduzida por um pequeno texto, descrevendo suas características e funções. Por exemplo, já na tradução para o português:

> Perfeito para crianças que gostam de leite, mas têm intolerância à lactose – uma bebida doce sem leite, que vai nutrir e sustentar seu filho, sem levar às alturas seu nível de açúcar no sangue (p. 78).

Do total das receitas, selecionamos os títulos que apresentaram maior dificuldade de tradução. Para facilitar a discussão, esses foram agrupados conforme o problema envolvido, mas devemos lembrar que nenhuma categorização é absolutamente precisa.

3 Revisão da literatura

Quando se pensa em tradução de jogos de palavras, imediatamente pensa-se na tradução de humor e, mais especificamente, na de piadas. De fato, boa parte da literatura foca nesse aspecto (NIEDZIELSKI, 1991; SCHMITZ, 1996; ZABALBEASCOA, 1996, BREZOLIN, 1997; ROSAS, 2002; ATTARDO, 2002, entre muitos outros). Há, porém, vasta literatura sobre a tradução de jogos de palavras em geral, a exemplo de Tagnin (2005) e Delabastita (1997) que reuniu vários ensaios sobre o tema. Outros artigos, alguns abordando filmes ou séries televisivas, são Delabastita (1996), Stewart (2000) e Asimakoulas (2004). Vários trabalhos de conclusão de curso, inclusive de cursos de

graduação e especialização, também têm se dedicado ao tema, dentre eles, Cunha (2004), Silva (2006), Brunet (2014) e Silva (2015).

A tônica dessa produção é que a tradução, nesses casos, deve privilegiar o 'efeito' em detrimento da forma, estratégia que vai ao encontro dos preceitos da teoria funcionalista da tradução (Reiss & Vermeer, 1996) (SCHÄFFNER, 2009). Um dos preceitos dessa teoria é que a tradução deve ser moldada pelo objetivo (*escopo* na terminologia de Vermeer) a que se destina, levando em conta principalmente as diferenças entre a cultura do texto de partida e a do texto de chegada. Certamente, de um modo geral, foi o que orientou a tradução dos títulos em questão, já que se pretendia que esses, em português, despertassem nas crianças brasileiras o mesmo efeito lúdico que, supõe-se, a autora esperava que tivesse nas crianças inglesas.

Outro preceito que regeu as traduções foi o da 'naturalidade' (TAGNIN, 2013) – os títulos teriam de soar 'naturais' para o público-alvo. Por naturalidade entendemos a forma usual de se expressar, a mais provável de ocorrer em determinado contexto. Na tentativa de recriar um jogo de palavras, não nos bastava tentar uma tradução que se aproximasse do jogo original se não representasse um 'modo de dizer' consagrado.

Além disso, conforme alerta Philip (2009), há uma nítida diferença entre o significado de uma palavra em seu contexto convencional de uso e seu sentido num uso não convencional da língua, ou seja, quando é usada de forma criativa. Isso deve ser reconhecido e levado em conta na tradução. Em outras palavras, é preciso primeiro reconhecer que a palavra ou expressão está sendo usada de forma criativa e então tentar recriar esse efeito na tradução. Vejamos, então, como foi esse processo na tradução dos títulos desse livro de sucos para crianças.

4 Os títulos

A autora empregou vários recursos para nomear os sucos sugeridos. Em termos lexicais, empregou colocações em novos contextos; como recursos prosódicos, fez uso de rimas e aliterações; no âmbito cultural, lançou mão de nomes de jogos, canções e personagens

infantis. No entanto, os títulos eram, em sua maioria, motivados; por exemplo, *Go green* e *Yellow submarine* refletiam a cor do suco; *Crazy kiwi* e *Apple aid* incluíam o ingrediente principal; *Sweet and sour* revelava o sabor. Já outros tinham certo viés metafórico, aludindo ao efeito desejado: *Energy bubble* e *Battery charge*. Claramente metafórico era *Kick start*, termo que se refere a dar partida numa moto.

O objetivo das traduções era ater-se, o tanto quanto possível, à 'forma' original, desde que mantivesse, como já dissemos, o efeito 'apelativo'. Em alguns poucos casos, uma tradução literal permitia isso:

Apricot cloud	Nuvem de damasco
Battery charge	Carga na bateria
Crazy kiwi	Kiwi maluco
Doctor, doctor	Doutor, doutor
Hug in a glass	Abraço no copo

No entanto, vários outros demandaram um conhecimento maior da convencionalidade da língua inglesa – o 'reconhecimento' a que se refere Philip (2009) – para evitar uma tradução literal que não soasse natural em nossa língua.

4.1 Títulos com colocações

As colocações representam uma das categorias convencionais bastante usada em títulos e, muitas vezes, manipulada para efeitos retóricos ou humorísticos. Para reconhecer essa manipulação é necessário estar familiarizado com a colocação. Tomando emprestada uma noção de Fillmore (1979), poderíamos chamar de "tradutor ingênuo" aquele que só entende a língua literalmente e assim a traduz. Vejamos alguns casos de colocações: *Sky high* apresenta 18.500.000[1] resultados no Google, significa 'de alto nível', e é bastante usada como nome próprio: *Sky High Car Audio* (uma loja), *Sky High Wilderness Ranch* (local de aventuras), *Sky High Sports* (academia). Foi também o título de um filme de 2005, cujo título em português era *Sky High* –

[1] Acesso em: 20 jul. 2015.

Escola de Heróis, ou seja, manteve-se o original, acrescido de um texto que alude ao significado de 'alto nível' da colocação, na medida que inclui 'heróis', personagens de 'alto valor', de grandes conquistas.

Um tradutor ingênuo a traduziria como 'céu alto' ou 'alto céu', combinação inexistente em português. Cabe, então, buscar uma colocação consagrada na nossa língua, que possa surtir efeito similar no público-alvo. No nosso caso, optamos por manter a imagem do céu, lançando mão de uma colocação em português, 'céu aberto', que apresenta 688.000 resultados[2] no Google e também é usada como nome próprio: *Pousada Céu Aberto, Céu Aberto Arte e Consciência* (escola de arte), *Céu Aberto* (espaço de saúde).

Outra colocação, *Witches' Brew*, não representou um problema de tradução por termos uma colocação com o mesmo sentido, *Poção mágica*[3]. *Poção mágica* tem 31.700 resultados,[4] mas não devemos esquecer que o volume de material em inglês, disponível na *web*, é infinitamente maior do que o material em português, de modo que essa frequência é válida para nossos objetivos.

Outro título para o qual encontramos um equivalente em português foi *Time out*, termo que, entre outros sentidos, refere-se ao tempo de descanso num jogo, razão pela qual optamos por 'Meio tempo', mantendo o domínio do esporte.

Um caso bastante interessante foi o suco denominado *Traffic light*, que significa 'farol', 'semáforo' ou ainda 'sinaleira', dependendo da região do nosso país. O suco, de fato, emula um farol, pois tem três camadas: uma verde, uma amarela e uma vermelha. Nesse caso, para evitar algum regionalismo, optamos por ressignificar uma expressão relacionada ao trânsito: 'pare, olhe, siga'. Trazendo-a para o universo do livro, criamos *Pare, olhe, beba*, acreditando que o pequeno leitor fará a ligação desse título com um farol de trânsito.

Outro título que envolve um jogo de palavras é *Raspberry crush*, pois *crush* pode ser entendido como uma bebida (existia antigamente um

---

2 Acesso em: 20 jul. 2015.

3 Em inglês, existe *magic brew*, equivalente de 'poção mágica', mas com baixa frequência: 68.600 resultados, contra 483.000 para *witches brew*. Acesso em: 20 jul. 2015.

4 Acesso em: 20 jul. 2015.

refrigerante de laranja chamado *Crush*) ou como o verbo 'to crush', que significa 'amassar, comprimir, quebrar'. Em linguagem informal, entretanto, 'crush' ocorre na expressão *to have a crush on somebody*, ou seja, 'ter uma queda por alguém'. Nesse sentido, podemos correlacionar *crush* com 'afeto'. Em português, o verbo 'amassar' dá origem ao substantivo 'amasso' que, como *crush* em inglês, denota alguma forma de 'afeto': 'dar um amasso em alguém'. Dessa forma, o título em português foi *Amasso de framboesa*.

O suco denominado *Nutty Professor* é o título de um filme com Eddie Murphy, conhecido em português como *Professor Aloprado*. Temos aqui novamente um jogo de palavras, já que *nutty* tem o sentido de 'relacionado a nozes' (*nuts* em inglês) e, em linguagem informal, significa 'maluco, louco', daí a tradução em português. Entretanto, se mantivéssemos como título do suco em português o nome do filme, perderíamos a referência aos ingredientes do suco, vários tipos de nozes. Daí, buscamos algum personagem que pudesse manter essa dupla leitura e lembramos de *Minduim*, nome de uma tira de jornal do cartunista americano Charles Schultz, cujo personagem principal é Charlie Brown. Optando por *Minduim*, mantivemos a ligação com o universo infantil e com os ingredientes do suco.

4.2 Títulos que fazem referência ao universo infantil

Esses envolvem, em geral, personagens, livros ou jogos do mundo infantil. *Purple Tiger* reflete a cor arroxeada do suco, que é feito com amoras e mirtilos, mas também faz referência a um personagem de histórias infantis bastante conhecido em nosso país, o Tigrão, em inglês *Tigger*. Face a isso, denominamos o suco simplesmente *Tigrão*.

O título *Princess Peachy* provém de uma personagem fictícia da série de *videogames Super Mario Bros.,* produzida pela *Nintendo*.[5] Salvo engano, essa personagem não é muito conhecida de nosso público infantil. Entretanto, há uma boneca com o nome de *Pesseguinho* que conjuga, como o título em inglês, o universo infantil e o ingrediente do suco, pêssego.

[5] Cf. <https://pt.wikipedia.org/wiki/Princesa_Peach>.

Dois títulos eram baseados em livros infantis, *Black Beauty* e *Wakey-wakey*. O primeiro é de autoria de Anne Sewell e foi traduzido para o português como *Beleza Negra.* Entretanto, uma busca no Google revelou que essa combinação era mais comumente associada à beleza da mulher negra. Dos 405.000 resultados, no Google, para "beleza negra", 291.000 coocorriam com *mulher*, enquanto apenas 93.900 tinham relação com o livro.[6] A fim de evitar a leitura relacionada à beleza da mulher, optamos por *Corcel Negro,* que também é uma história sobre um cavalo, escrita por Walter Farley (em inglês *Black Stallion*)*,* e que ensejou três filmes dirigidos por Carroll Ballard: *The Black Stallion* (1979), *The Black Stallion Returns* (2003) e *The Young Black Stallion* (2003*),* além da série de TV *The Adventures of the Black Stallion*, produzida entre 1990 e 1993.[7]

O segundo livro, *Wakey-wakey*, de autoria de Dawn Apperlley, traz no título uma expressão empregada para acordar alguém. Como, em português, 'acorda, acorda' não é uma expressão consagrada, optamos por manter a repetição e usar uma expressão relacionada a regozijo e alegria, *Oba-oba*, com a intenção de incentivar a criança a tomar o suco.

Wacky Wizard é um jogo de tabuleiro em que os jogadores, para ganhar, têm de desfazer a magia do bruxo, mas também remete a shows de mágica. Ambas as referências seriam desconhecidas de nosso público. Porém, um personagem conhecido é o *Mágico de Oz*, que mantém a relação com o mundo da magia e foi a opção final para o título desse suco.

Já *Captain Zinger* não apresentou problema, porque esse personagem é conhecido no Brasil como *Capitão Z.*

Em contrapartida, o suco *Jungle Fever* causou muita estranheza, pois essa é uma das denominações de 'malária' ou 'febre amarela' em inglês – referência que nos pareceu incompatível com o nome de um suco para crianças. Uma pesquisa na *web* revelou tratar-se também de um jogo de tabuleiro. Como o suco era amarelo – daí o título – buscamos algum jogo infantil que pudesse preservar essa referência e chegamos a

[6] Acesso em: 20 jul. 2015.
[7] Cf. <https://pt.wikipedia.org/wiki/The_Black_Stallion>.

*Amarelinha***,** que recupera a cor da bebida, além de remeter ao universo dos jogos infantis.

4.3 Título que reflete o efeito desejado

Um dos sucos é denominado *Easy Rider***,** nome de um filme de 1969, traduzido para o português como *Sem Destino*. O efeito pretendido pelo suco é o de aliviar enjoos em viagens, razão pela qual o título do filme em português seria claramente inadequado. Para manter o uso de uma categoria convencional, optamos por *Boa viagem***,** uma fórmula comumente empregada quando alguém vai viajar.

4.4 Título que envolve efeito prosódico

Um dos sucos faz uso de aliteração: *Frisky frog*. É feito com kiwis e uvas verdes, o que lhe dá a coloração verde, motivando o 'sapo' (*frog*, em inglês) do título. Além disso, é um suco energético para que a criança logo esteja revigorada e "brincando por aí". Dessa forma, para manter o efeito prosódico, mantendo o animal, fizemos uso de rima, denominando o suco de *Sapo no papo*.

4.5 Títulos com combinação de estratégias

Run rabbit run é um título que envolve aliteração, é o nome de uma canção de 1939, é o nome de um livro infantil, além de ser o título de um romance de John Updike (obviamente para o público adulto). A referência pretende aludir à característica do suco, pois trata-se de "um excelente laxante natural" (CROSS, 2007b, p. 59). A tradução sugerida foi *Corre-corre*, por ser uma combinação consagrada em nossa língua e remete ao efeito desejado.[8]

4.6 Traduções que exigiram adaptações culturais

[8] A editora alterou o título para *Corre, coelhinho, corre.*

A tradução de alguns títulos demandou adaptações culturais mais específicas. Por exemplo, *Kick Start* remete à forma de dar partida numa motocicleta. Transpusemos essa ideia para o domínio do futebol, assunto com o qual o público infantil brasileiro está mais familiarizado, e traduzimos como *Pontapé inicial*.

Caso similar foi o suco *Super Stripy* que, literalmente, quer dizer 'super listrado'. Visualmente, apresenta três camadas: a inferior é um purê de mirtilos de coloração bastante escura, a do meio é de coalhada seca com mel, portanto branca, e a última é um purê de framboesa, ou seja, avermelhada. As cores lembram o sorvete tricolor chamado *Napolitano* (chocolate, creme e morango) e foi esse o título que lhe demos em português.

Finalmente, *Little Pinky* traz o nome do menor dedo da mão. Entretanto, 'mindinho' não faz qualquer alusão à cor do suco, cuja base é a polpa de melancia, de modo que tem uma cor avermelhada. Na tentativa de manter a referência à cor e incluir um pouco da cultura de nosso país, a tradução sugerida foi *Boto rosa,* personagem de uma lenda amazônica.[9]

5 Considerações finais

Nosso objetivo foi discutir a tradução do inglês para o português de alguns títulos de sucos do livro *Juices & Smoothies for Kids*, de autoria de Amanda Cross. A maioria dos nomes envolve algum jogo de palavras ou manipulação criativa para tornar o suco atraente para o público infantil. Entre as estratégias empregadas, apontamos o uso de colocações em contexto não usual, recursos prosódicos como rimas e aliterações, referências ligadas ao universo infantil, como personagens, nomes de livros e jogos. Alguns nomes são motivados pela cor do suco ou pelo efeito desejado.

Nesse tipo de tradução, conforme já apontaram vários autores, assim como os adeptos da teoria funcionalista da tradução, o importante

[9] A editora substitui-o por *Mix refrescante*.

é priorizar o 'efeito' pretendido no público-alvo, em detrimento da 'forma'. Para tanto, é primordial reconhecer a estratégia utilizada para nomear o suco e tentar reproduzi-la na língua de chegada. Quando isso não é possível, cabe ao tradutor criar um novo título que mantenha o efeito apelativo pretendido. Na maioria dos casos relatados, isso foi obtido; em outros, fizeram-se necessárias adaptações culturais. No geral, acreditamos ter alcançado o objetivo de produzir títulos apelativos ao público infantil.

6 Referências

ASIMAKOULAS, D. Towards a Model of Describing Humour Translation. A Case Study of the Greek Subtitled Versions of Airplane! and naked Gun. *Meta ,* 49, p. 822-842, 2004.

ATTARDO, S. Translation and Humor: an Approach Based on the General Theory of Verbal Humor (GTVH). *The translator*. Editor: J. Vandaele, v. 8, p. 173-194, 2002.

BREZOLIN, A. Humor: sim, é possível traduzi-lo e ensinar a traduzi-lo. *TradTerm - Revista do Centro Interdepartamental de Tradução e Terminologia*, 1, p. 15-30, 1997.

BRUNET, C. D. A tradução do humor sob a perspectiva da Teoria da Relevância. Monografia (Especialização) – Universidade Tuiuti do Paraná, Curitiba, 2014

CROSS, A. *Juices and Smoothies for Kids.* London: Hamlyn, 2007a.

CROSS, A. *Sucos e Vitaminas para Crianças.* Trad. S. E. Tagnin. Barueri, SP: Manole, 2007b.

CUNHA, A. M. *A tradução da comicidade linguística em Alice's Adventures in Wonderland*. Monografia (Graduação em Letras) – Universidade Federal de Juiz de Fora, 2004.

DELABASTITA, D. *Wordplay and Translation.* Manchester: St. Jerome Publishing, 1996.

DELABASTITA, D. *Traductio* - essays on punning and translation. 2. ed. V. 2. Langues et littératures. Manchester: St. Jerome Publishing and Presses Universitaires de Namur, 1997. 296 p.

FILLMORE, C. J. Innocence. A second idealization for linguistics. In: ANNUAL MEETING OF THE BERKELEY LINGUISTICS SOCIETY, 5, 1979, Berkeley. *Proceedings*... Ed.: C. Chiarello. Berkeley, CA: Linguistics Society, 1979. p. 63-76.

NIEDZIELSKI, H. Cultural Transfers in the Translation of Humor. In: LARSON, M. (Ed.). *Translation*: Theory and Practice – Tension and Interdependence. Binghampton, NY: State University of New York, 1991. p. 139-156.

PHILIP, G. Arriving at equivalence. Making a case for comparable corpora in Translation Studies. In: BEEBY, A.; RODRIGUEZ INÉS, P.; SÁNCHEZ-GIJÓN, P. *Corpus Use and Translating*: Corpus Use for Learning to Translate and Learning Corpus Use to Translate. Amsterdam: John Benjamins, 2009.

REISS, K.; VERMEER, H. J. *Fundamentos para um teoría funcional de la traducción.* Trad.: S. Reina; C. de León. Madrid: Akal, 1996.

ROSAS, M. *Tradução de Humor* – Transcriando Piadas. Rio de Janeiro: Lucerna, 2002.

SCHÄFFNER, C. Functionalist approaches. In: BAKER, M.; SALDANHA, G. *Routledge Encyclopedia of Translation Studies*. New York: Routledge, 2009. p. 115-121.

SCHMITZ, J. R. Humor: é possível traduzi-lo e ensinar a traduzi-lo? *TradTerm: Revista do Centro Interdepartamental de Tradução e Terminologia*, 3, p. 87-97, 1996.

SILVA, N. R. *A tradução de jogos de palavras no romance O xangô de Baker Street*: uma revisão do quadro de estratégias de Delabastita com o auxílio da Linguística de Corpus. Tese (Doutorado em Linguística) – FFLCH – Universidade de São Paulo, São Paulo: 2015.

SILVA, N. R. *Um Estudo sobre a Recepção do Humor Traduzido*. Dissertação (Mestrado em Letras) – Universidade Estadual do Ceará, Fortaleza, CE, 2006.

STEWART, D. Conventionality, Creativity and Translated Text: The Implications of Electronic Corpora in Translation. In: OLOHAN, M. *Intercultural Faultlines*. Manchester, Northampton: St. Jerome Publishing, 2000. p. 73-91.

TAGNIN, S. E. O humor como quebra da convencionalidade. *Revista Brasileira de Linguística Aplicada*, 5, n. 1, p. 247-257, 2005.

TAGNIN, S. E. *O jeito que a gente diz*: combinações consagradas em inglês e português. São Paulo: Disal, 2013.

ZABALBEASCOA, P. Translating Jokes for Dubbed Television - Situation Comedies. *The translator* , 2, p. 235-257, 1996.

DOI: 10.17851/2237-2083.23.3.695-726

Topic Modeling for Keyword Extraction: using Natural Language Processing methods for keyword extraction in Portal Min@s

A modelagem de tópicos para extração de palavras-chave: o uso de métodos de processamento natural da linguagem para extração de palavras-chave no Portal Min@s

Arnaldo Cândido Junior
Universidade Tecnológica Federal do Paraná, Curitiba, Paraná, Brasil
arnaldoc@utfpr.edu.br

Célia Magalhães
Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brasil
celiamag@gmail.com

Helena Caseli
Universidade Federal de São Carlos, São Paulo, Brasil
helenacaseli@gmail.com

Régis Zangirolami
Littera TI – Processamento de Dados, São Carlos, São Paulo, Brasil
regisrmz@gmail.com

Abstract: This article aims to evaluate the application of two efficient automatic methods for keyword extraction used by Corpus Linguistics and Natural Language Processing communities for generating keywords from literary texts: WordSmith Tools and Latent Dirichlet Allocation (LDA). These tools have their own specificities and are based on different extraction techniques; thus an analysis focused on their performance was required. This article aims to understand how each method works and to evaluate them when applied to extract keywords from literary works. To this end, we used human

analysis, with knowledge of the field of the texts used. The LDA method was used for extracting keywords through its integration with *Portal Min@s: Corpora de Fala e Escrita,* a general corpora-processing system, designed for different research in corpus linguistics. The experiment outcomes confirm the effectiveness of WordSmith Tools and LDA in extracting keywords from literary corpus. They also show that human analysis of the lists is required at a stage prior to experiments to complement the automatically generated list, crossing WordSmith Tools and LDA results, and that the linguistic intuition of a human analyst about the lists generated separately by the two methods in this study was more favorable to the use of the WordSmith Tools keyword list.

Keywords: keyword extraction, natural language processing, corpus analysis, WordSmith Tools, Latent Dirichlet Allocation, *Portal Min@s*

Resumo: Este artigo tem o objetivo da avaliar a aplicação de dois métodos automáticos eficientes na extração de palavras-chave, usados pelas comunidades da Linguística de *Corpus* e do Processamento da Língua Natural para gerar palavras-chave de textos literários: o *WordSmith Tools* e o *Latent Dirichlet Allocation* (LDA). As duas ferramentas escolhidas para este trabalho têm suas especificidades e técnicas diferentes de extração, o que nos levou a uma análise orientada para a sua performance. Objetivamos entender, então, como cada método funciona e avaliar sua aplicação em textos literários. Para esse fim, usamos análise humana, com conhecimento do campo dos textos usados. O método LDA foi usado para extrair palavras-chave por meio de sua integração com o *Portal Min@s: Corpora de Fala e Escrita*, um sistema geral de processamento de *corpora*, concebido para diferentes pesquisas de Linguística de *Corpus*. Os resultados do experimento confirmam a eficácia do WordSmith Tools e do LDA na extração de palavras-chave de um *corpus* literário, além de apontar que é necessária a análise humana das listas em um estágio anterior aos experimentos para complementar a lista gerada automaticamente, cruzando os resultados do WordSmith Tools e do LDA. Também indicam que a intuição linguística do analista humano sobre as listas geradas separadamente pelos dois métodos usados neste estudo foi mais favorável ao uso da lista de palavras-chave do WordSmith Tools.

Palavras-chave: extração de palavras-chave; processamento natural da linguagem; análise de *corpus*; WordSmith Tools; Latent Dirichlet Allocation; *Portal Min@s*.

Recebido em: 31 de junho 2015.
Aprovado em: 23 de novembro 2015.

1 Introduction

Keywords are a quick and efficient way of indicating the main topics of a text. According to Scott (1996), they are words whose frequency in a text is exceptionally high compared to some standard. Scott (1997) also defines keyword as a word that occurs with unusual frequency, either high or low, compared to a reference corpus. Thus, the analysis of keywords tend to indicate the topic addressed in a certain text or corpus. Keywords are used in scientific papers and literary works to facilitate the cataloging and organization of texts, and thus the search for them. The concept of keyword has become a lot stronger and today it is present in everyone's life, thanks to the Internet and the need to create effective search solutions in an environment where the amount of news, works, articles, blogs, among other types of text, has grown.

Corpus Linguistics (CL) and Natural Language Processing (NLP) communities apply different systems and methods to extract keywords. There are many systems available, such as Kea,[1] Maui indexer,[2] Carrot2,[3] among others. Two systems commonly used by the CL community are WordSmith Tools[4] and Wmatrix.[5] Examples of extraction methods used by the NLP community are the TF-IDF techniques (MANNING; SCHÜTZE, 2000), Latent Semantic Analysis (LSA) (LANDAUER *et al*., 1998) and Latent Dirichlet Allocation (LDA)[6] (BLEI, 2012; DREDZE *et al*., 2008). The latter two methods are also important for Information Retrieval in the context of document extraction, since they allow this extraction based on synonyms and

1 <http://www.nzdl.org/Kea/>. Access on: Oct. 4, 2015.
2 <https://code.google.com/p/maui-indexer/>. Access on: Oct. 4, 2015.
3 <http://project.carrot2.org/>. Access on: Oct. 4, 2015.
4 <http://www.lexically.net/wordsmith/version6/>. Access on: Oct. 4, 2015.
5 <http://ucrel.lancs.ac.uk/wmatrix/>. Access on: Oct. 4, 2015.
6 <http://www.cs.princeton.edu/~blei/lda-c/>. Access on: Oct. 4, 2015.

topics, respectively, which are approximations to search words, helping search engines.

Although there are separate analyses of the techniques used in each method, there is no study comparing methods for keyword extraction in literary works, which is the focus of this paper. There is no evaluation of topic extraction methods for this scenario either. The application of these methods to corpora of fictional texts, for example, must take into account some issues, given the nature of prose fiction. Such corpora, if comprised of novels, in addition to being larger than corpora of academic texts, for example, do not allow prior definition of a fixed vocabulary used in the texts, which happens only with prior knowledge of each text.

This article main goal is to compare two efficient automatic methods for keyword extraction used by CL and NLP communities for generating keywords from literary texts: WordSmith Tools keyword extraction and LDA keyword extraction. A secondary objective was to define whether these methods are effective and accurate when applied to the literary genre.

As both tools have their own specificities and are based on different extraction techniques, an analysis focused on their performance was required. To this end, we have applied human analysis, with knowledge of the field of the texts used. The LDA method was used for keyword extraction from its integration with *Portal Min@s: Corpora de Fala e Escrita*, a general corpora-processing system, designed for different research in corpus linguistics. The Portal has a web interface, which facilitated the efficient analysis of the data presented in this paper.

This text is organized as follows. Section 2 presents the theoretical foundation for this paper, including the tools, methods and corpora used. Section 3 details the experiments. Section 4 presents a discussion of the outcomes. Finally, Section 5 brings the final considerations.

2 Theoretical Foundation

This section provides details on the aforementioned methods, including a presentation of *Portal Min@s*, which shows the

implementation of one of the extraction methods, LDA. In Section 2.1, we describe the methods for keyword extraction commonly used in the two areas of research: CL and NLP. In Section 2.2, the corpus used for the experiments. In Section 2.3, *Portal Min@s*, illustrating the use of the LDA method.

2.1 Methods for Keyword Extraction

Keywords are defined as a set of terms in a document that provides a summary of its content to readers (LIU *et al*., 2010). Several areas, such as Information Retrieval, CL and NLP, make use of keyword extraction for various tasks such as document categorization, clustering and summarization. Several methods for keyword extraction have been proposed and different software implemented. Popular methods and software include WordSmith Tools, TF-IDF, Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) and WMatrix.

WordSmith Tools is a software for Windows that is easy to install and user-friendly. It has three main features: concordancing, keyword extraction, and word listing. To enable extraction, you must first generate the word lists of the texts. WordSmith Tools generates keywords when comparing the list of frequent words of two corpora, being the first corpus called "study corpus" and the second, "reference corpus". The reference corpus is used only to allow comparison, setting an average frequency at which the terms normally occur. Any terms present in the study corpus that show a significant change from the normal frequency set based on the reference corpus, for more and for less, will be highlighted as keywords (SARDINHA, 2000, p. 8). The only requirement of the tool is that the reference corpus is greater than the study corpus. According to Sardinha (2000, p. 12), it is ideal that the reference corpus is at least five times the size of the study corpus because it was observed that smaller reference corpus generates a smaller number of keywords, while those five times greater than the study corpus tend to maintain a reasonable number of keywords. The advantage of using WordSmith Tools is that this method is vastly used by corpus linguists, thus simplifying the process of comparing results on different papers. It

is also worth noting that WordSmith Tools uses mature statistical methodology for keyword extraction.

Other popular approach for keyword extraction is TF-IDF (MANNING; SCHÜTZE, 2000) is a statistical method used to show how important a word is for a text within a corpus. It is the result of multiplying two frequencies of words: the term frequency (TF) and the inverse document frequency (IDF). In practical terms, it compares the frequency at which a word occurs in a text with the frequency at which the same word occurs in the corpus as a whole, enabling the creation of a ranking of frequency of words in a text based on their frequency in the larger whole. It is a very practical method of finding the weight and importance of the words of a text in a specific field of study. It also has the advantage of not relying on a stop-words list, since the most frequent words in the corpus texts have less weight and are last in the ranking of importance. The main advantage of TF-IDF is its simplicity and easiness to implement in software, which is ideal for tool developers and experimenters with background in programming languages.

LSA (LANDAUER, *et al*., 1998) is a method that measures the level of similarity of words and excerpts of texts after analyzing a corpus. It assumes that words with the same meaning, or a similar one, must appear in similar contexts. The method uses a mathematical technique called Singular Value Decomposition (SVD) to build a matrix of word occurrence in the paragraphs and corpus texts, enabling to determine the weight of each word in a text. After creating the occurrence matrix, the LSA must find a minimum version of this matrix to reduce the analysis complexity, minimize noise, and unify the occurrence of words of similar meaning, including the frequency of these words in texts corpus. This method is also commonly known as LSI (Latent Semantic Indexing). It is indicated as suitable for uses in which part of the keywords are already known before the extraction takes place.

LDA, inspired in LSA, is a probabilistic model for generating topics. On this model, Blei (2012) states: “To this end, machine learning researchers have developed probabilistic topic modeling, a suite of algorithms that aim to discover and annotate large archives of documents with thematic information.” (p. 77). That is, this kind of model aims to analyze a corpus and define, through probabilistic algorithms, the topic

addressed in the text. According to Blei (2012, p. 78), the idea behind LDA is that a text deals with multiple subjects, and these subjects can be represented by topics. A topic is formally defined as a distribution of words of a fixed vocabulary, and each document displays its topics in different proportions.

The method has been defined by Blei, Ng and Jordan (2003) as a hierarchical Bayesian model, which seeks to highlight the content of a document by representing its topics, generated through efficient approximation techniques. The document is seen as a mixture of topics following some probability distribution. The method seeks to associate each sentence with at least one topic. The same sentence can be associated with more than one topic. In this case, an estimate is made on the relation of the sentence with each topic in question. All calculations are made based on the bag of words of the document. The LDA method is applied for retrieving information in a first phase that clusters words by topic. In a second phase, the documents are clustered according to their topics (TDK TECHNOLOGIES, 2015). LDA main advantage is its ability to cluster keywords in topics.

Finally, WMatrix is a web platform that works in any system or environment, and which uses the so-called Matrix method. This method makes statistical comparisons in corpora, generating accurate keywords (RAYSON, 2002). It was developed based on a comparison of corpus analysis tools to provide a solution that has all the required and appropriate features for editing and extracting information from a corpus. The idea is not to eliminate the need for further human analysis, but to refine the most the initial analysis and allow a human being to examine a much larger volume of texts with less effort. According to Rayson (2002, p. 153), its main applications are: in the study of vocabulary according to social contexts; contrast native and non-native English speakers; and in the semantic analysis of documents that require software engineering. Wmatrix uses 21 semantic categories[7] to classify a text, including the Portuguese language and others. The categories are comparable to the topics that LSA and LDA raise automatically, but are fixed. Works like McIntyre and Walker (2010) prove the effectiveness of WMatrix in

[7] <http://ucrel.lancs.ac.uk/usas/>. Access on: Oct. 4, 2015.

corpora analysis of poetry and dramatic texts, but as far as we know, there are no empirical studies on its application to long literary texts such as novels. Thus, one of WMatrix advantages relies on the fact it has been applied to keywords extraction in less common genres such as poetry and dramatic texts.

Considering the presented methods and tools, the scope of research work was adjusted to compare WordSmith Tools and LDA, so the analysis of other strategies has been left for future works. WordSmith Tools was selected for being widely used in the linguistic community, considered as a classic tool for extracting keywords. LDA was selected for bringing an innovative approach based on topics, which allows a comparative analysis of its strengths and limitations in relation to a classic strategy.

2.2 Corpora

The corpora described in this section were the basis for the experiments described in Section 3. Two categories of corpora were used: study (literary) and reference (to support keyword extraction).

The corpus of literary texts (study corpus) used in the experiments described here was compiled with translations of *Heart of Darkness*, a Joseph Conrad novel published in 1899, with each version in Portuguese produced by a different translator (Mark Santarrita, in 1996; Regina Regis Junqueira and Hamilton Trevisan, in 1984). The corpus was digitalized from their printed versions in books. The numbers of tokens and sentences in these three documents are shown in Table 1.

Table 1 – Tokens and sentences by translation

File	Tokens	Sentences
HOD_Junqueira	41,808	2,201
HOD_Santarrita	36,681	2,335
HOD_Trevisan	37,896	2,419

The reference corpus used was PLN-BR GOLD, which comprises 1,024 texts and 338,441 tokens and was compiled in the PLN-BR project (MUNIZ *et al*., 2007), having only news and coverages for which *Folha de S.Paulo* (a Brazilian newspaper) has republishing rights. The size of this corpus represents 1% of the largest corpus in the PLN-BR project, the PLN-BR FULL, in order to proportionally preserve the distribution of this largest corpus. It is a stratified random sample proportional to the distribution of the PLN-BR FULL corpus, which covers the years 1994-2005, related to texts published by newspaper *Folha de S.Paulo*.

2.3 *Portal Min@s*

Portal Min@s[8] intends to provide a unified and systematized computational base for processing corpus compiled and made available for linguistics research. The system was motivated by the recent expansion of research in corpus linguistics, which brought large amounts of corpus demanding robust processing. The Portal, as well as other efforts, seek to meet this demand. The differential in relation to other tools is its generalist approach, seeking to act in different contexts, based on the type of corpus (for example, annotated or not) and task (for example, translation and lexicography) or area of study in corpus linguistics.

The tool is free and publicly available to all institutions interested in its use, requiring only a web-hosting server. The advantages of offering a web system are the possibility of simultaneous remote access by multiple users, without the need of installation in each user's computer.

Several features have been implemented, given the generalist focus of the Portal, treating different types of corpus, different languages and different research in corpus linguistics. Examples of features include generation of concordances, alignments for parallel texts and extraction of keywords, focus of this paper. Unlike corpus processing portals that work with a fixed list of corpus (DAVIES, 2005, 2009), the Portal allows the creation of corpus on demand and offers a module to import

[8] <http://fw.nilc.icmc.usp.br:12480/portal/index.jsp>. Access on: Oct. 4th, 2015.

previously existing corpus, provided that they meet the format requirements of the Portal.

The Portal is based on eight core modules for general features to access corpora, and six support modules, created for managing and importing corpora. The core modules are: concordances, alignments, statistics and frequencies, keywords (focus of this paper), annotations, and multimodal corpus. In addition to the core modules, the Portal also has supporting modules, responsible for managing users, as well as importing and managing corpora, namely: importing module, text manager, corpus manager, subcorpus manager, tag manager and user manager.

In particular, the module used in this work allows automatic keyword generation. In this task, the Portal acts as an interface for the LDA-C tool (BLEI, 2003), an implementation of the LDA method with only command line interface available. During the extraction process, the user must inform the number of topics to be extracted. With a single topic, the LDA is similar to other methods used in keyword extraction software. From two or more topics on, however, the tool clusters keywords into coherent topics using statistical techniques.

The keyword extraction module has a practical interface and does not require any technical knowledge from the user, who can perform automatic extraction of texts with only three parameters: the number of topics, the number of keywords for each topic and the analysis text, as illustrated in Figure 1. An optional fourth parameter is a customized list of stop words that can be sent by the user; although *Portal Min@s* has a standard list for Portuguese, English and Spanish, the user has the option of using their own list to perform extraction in a specific text. The list of extracted keywords is displayed on the screen, as shown in Figure 2, but it is also available for download in plain text format.

Figure 1 - Interface of the keyword extraction tool of *Portal Min@s*

Source: <http://fw.nilc.icmc.usp.br:12480/portal/index.jsp>.

Figure 2 - Example of keyword extraction with display of results

Source: <http://fw.nilc.icmc.usp.br:12480/portal/index.jsp>.

Section 3 presents the experiments carried out to compare LDA and WordSmith keyword extraction.

3 Experiments

The idea of the study was to generate keywords through WordSmith Tools and LDA (Section 3.1), and then make comparisons between the keywords generated by analyzing which of them are common and which are specific to each method, and thus understand how each method works and their effectiveness for literary texts (Section 3.2).

In the experiments described in this article, we used only 800 of the 1,024 texts of the reference corpus, PLN-BR GOLD, enough to obtain a subcorpus approximately 5 times greater than the study corpus, following Sardinha's recommendations (2000). The subcorpus had 226,722 tokens and 17,002 sentences.

To apply the methods and compare their keyword extraction, we carried out studies with the literary corpus described in Section 2.2.

3.1 Keyword generation via WordSmith Tools and LDA

WordSmith generates a list of keywords from each corpus' list of word frequencies. The generated list has both positive and negative keywords. Positive keywords are the words of the study corpus that occur more frequently than the average set by the reference corpus, and negative keywords are those below the average rate. For this study, we have considered only positive keywords. Thus, three documents of keywords were generated: one for each translation. The first translation had 61 keywords, the second had 66 and the third had 60.

For the second method, the Latent Dirichlet Allocation (LDA), we have used implementation in language C, provided by one of the authors, called LDA-C (BLEI, 2003). The version also brings a Python script, which helps generating the list. The existing documentation is not complete; therefore, some details of the tool operation are not clear. As a

result, this experiment sought to use as parameters the default values recommended by the documentation.

WordSmith and *Portal Min@s* generate slightly different frequency lists due to variations in the tokenization process. For this reason, we made some changes in the experiments to use the WordSmith frequency list as input to LDA. Two scripts have been developed to adapt the output to the format:

[M] [term_1]:[count_1] [term_2]:[count_2] ... [term_N]:[count_N],

where [M] is the number of single terms in the document, and the number associated with each term means how many times the term occurred in the document (BLEI, 2003). These generated files enable to run LDA-C and generate output files with the keywords. In an exploratory approach, two topics were defined for extraction through the LDA method. An additional script was made to post-process the output and generate the files to be submitted for human analysis.

After applying the tools in the three case studies, 9 files were obtained:

- 3 files containing lists of keywords generated by WordSmith Tools, one for each translation; and
- 6 files containing the topics and their keywords generated by LDA, two files for each translation.

From these 9 files, it was possible to cross data and generate new results to refine the final analysis. Two new processes were performed: the intersection between the lists of keywords generated by each tool separately, highlighting in which case study each keyword occurs (TABLE 2), as well the intersection between the final values of each tool (TABLE 3).

Table 2 - Number of keywords generated in each translation (case) for each tool

	Case 1	Case 2	Case 3	Intersection
WordSmith Tools	61	66	60	93
LDA (2 topics)	150	146	153	227

Table 3 - Keywords generated by each tool, considered as a whole

	WordSmith Tools	LDA	Both
Occurs in 1 case	37	87	11
Occurs in 2 cases	20	56	7
Occurs in 3 cases	36	84	22

Finally, Table 4 presents the 10 best candidates for keywords, i.e. those words generated by both tools and that occur in the three translations (3 cases).

Table 4 - The 10 best candidates for keywords

Kurtz	Homem (Man)	Meu (My)	Tão (So)	Terra (Earth)
Cabeça (Head)	Olhos (Eyes)	Homens (Men)	Marfim (Ivory)	Peregrinos (Pilgrims)

The next step was to analyze these results and come up with a hit rate for each tool. To this end, the candidates generated were analyzed by an expert in the respective text. This analysis sought to determine whether the tools studied are accurate enough to extract keywords from a literary corpus.

3.2 Analysis

The analysis of keyword candidates started with the lists generated by the Keywords module of WordSmith Tools 6.0. The method comprises the analysis of the combined keywords in all the three

translations. The keywords were automatically combined into a single list using a Python script implemented specifically to this task.

In sum, the translations provided 93 unique words, 37 of which were common to all three translations. The keywords were mostly lexical, except for *parecia(m)* (the verb 'to seem' in Portuguese), forms of address and proper names. This list of words is shown in Table 5 (see APPENDIX A).

The list from the LDA keywords for the two topics of each text contained 249 unique words, 78 of which were common to both topics of all three translations. The keywords were mostly lexical, except for *parecia(m)* (the verb 'to seem' in Portuguese), forms of address and proper names. This list of words is shown in Table 6 (see APPENDIX A).

For the analysis of candidates for keywords in the lists generated by LDA, we decided to use an extra approach due to the difficulty in naming a topic for each of the two lists generated for each text, given the similarity of the lists and adversity of words listed. The methodology used in the analysis is described below:

1) the words were clustered into microtopics;
2) some of these topics were discarded because they are more related to textual characteristics (for example, the topic description of actions, relations and utterances that integrated verbs; the topic that comprised first-person pronouns, and topics that we could not identify because they are conjunctions, interjections etc.); and
3) microtopics were clustered into macrotopics, resulting in two macrotopics, although we were not sure whether some of the words they integrated were relevant.

The results of the analysis of lists generated by LDA are presented below.

3.2.1 HOD_Trevisan

For the HOD_Trevisan_tópico1 list, with 99 words,[9] the following microtopics were obtained:

1) Narrative: *me* (me), *minha* (my), *meu* (my), *mim* (me), *meus* (my) (5);
2) Human: *homem* (man), *olhos* (eyes), *cabeça* (head), *voz* (voice), *mãos* (hands), *homens* (men), *Mr* (Mr.), *peregrinos* (pilgrims), *braço* (arm), *jovem* (young), *passos* (steps), *homens* (men), *diabo* (devil), *braços* (arms) (14);
3) Nature: *rio* (river), *floresta* (forest), *mundo* (world), *margem* (shore), *treva* (darkness), *luz* (light), *terra* (earth), *água* (water), *árvores* (trees), *mar* (sea), *sol* (sun), *coisa* (thing), *sombra* (shadow), *coisas* (things), *forma* (shape), *escuridão* (dark), *coração* (heart) (18);
4) Time: *tempo* (time), *vez* (time), *vezes* (times), *momento* (moment), *sempre* (always), *meses* (months), *instante* (instant) (7);
5) Description of action, relations and utterances: *disse* (said), *dizer* (say), *olhar* (look), *poderia* (could), *parecia* (seemed), *encontrava* (found), *fosse* (were), *pareciam* (seemed), *saber* (know), *ficou* (became), *podia* (could), *sei* (know), *ouvi* (heard), *ver* (see) (14);
6) Another description: *direção* (direction), *frente* (front), *acima* (above), *caminho* (way), *movimento* (movement), *volta* (around), *nada* (nothing), *qualquer* (any), *tão* (so), *menos* (less), *sequer* (even), *nenhum* (none), *antes* (before), *ninguém* (nobody), *grande* (large), *claro* (clear), *certo* (right), *longo* (long), *possível* (possible), *pequeno* (small), *selvagem* (wild), *negros* (black), *dois* (two), *duas* (two) (24);
7) Human artifacts: *barco* (boat), *cabine* (cabin), *porta* (door), *companhia* (company) (4);
8) Abstractions: *verdade* (truth), *impressão* (impression), *respeito* (respect), *silêncio* (silence), *existência* (existence), *poder* (power), *razão* (reason), *vida* (life), *morte* (death) (9); and

[9] In this list, '*negros*' (the black) and '*selvagem*' (wild) could also be part of the 'human' topic; '*volta*' (around) could be under 'actions description'.

9) No topic: *oh* (oh), (*no*) *entanto* (but),[10] *porém* (however), *embora* (but) (4).

Topics 1, 5 and 9 were disregarded. Microtopics 2, 4, 7 and 8 were clustered into a
"human nature" macrotopic (34 words in total) and microtopics 3 and 6 were clustered into a 'physical nature' macrotopic (42 words in total). Based on the number of words, we can say that the 'physical nature' macrotopic is predominant in this list.

For the HOD_Trevisan_tópico2 list, with 98 words,[11] the following microtopics were obtained:

1) Narrative: *me* (me), *meu* (my), *minha* (my), *mim* (me), *minhas* (my), *meus* (my) (6);
2) Human: Kurtz, *homem* (man), *administrador* (administrator), *homens* (men), *voz* (voice), *pés* (feet), *espécie* (species), *peregrinos* (pilgrims), *corpo* (body), *olhos* (eyes), *Deus* (God), *rosto* (face), *alma* (soul), *Mr* (Mr.), *tom* (tone) (15);
3) Nature: *forma* (shape), *rio* (river), *terra* (earth), *mar* (sea), *ar* (air), *floresta* (forest), *sol* (sun), *selva* (jungle), *margem* (shore), *colina* (hill), *marfim* (ivory), *coisa* (thing), *coração* (heart), *luz* (light) (14);
4) Time: *tarde* (afternoon), *noite* (night), *dia* (day), *dias* (days), *momento* (moment), *tempo* (time), *instante* (instant), *vezes* (times) (8);
5) Description of action, relations and utterances: *parecia* (seemed), *disse* (said), *tivesse* (had), *dizer* (say), *pode* (can), *poderia* (could), *sabia* (knew), *vi* (saw), *preciso* (need), *exclamou* (exclaimed), *perguntei* (asked), *olhar* (look), *ver* (see), *podia* (could), *imaginar* (imagine), *sei* (know), *falar* (speak), *murmurou* (muttered), *esperar* (wait), *saber* (know), *posso* (could) (21);

[10] The correct translation of "but" is "no entanto". As the align is based on a one-to-one correspondence of lexical items, "but" was aligned only with "entanto".

[11] In this list, '*preciso*' (accurate) and '*volta*' (around) could also belong to the 'actions decription' microtopic, etc. and '*negro*' (black) and '*selvagem*' (wild) could also belong to the 'human' microtopic.

6) Another description: *nada* (nothing), *tão* (so), *mal* (bad), *melhor* (best), *simples* (simple), *maior* (greater), *impossível* (impossible), *longo* (long), *selvagem* (wild), *negro* (black), *capaz* (able), *direção* (direction), *antes* (before), *dois* (two), *duas* (two), *volta* (around), *quilômetros* (kilometers), *lugar* (place), *meio* (middle), *lado* (side), *ponto* (point) (21);
7) Human artifacts: *entreposto* (customs), *barco* (boat), *posto* (post), *trabalho* (work), *cerca* (fence) (5);
8) Abstractions: *silêncio* (silence), *verdade* (truth), *ideia* (idea), *realidade* (reality), *respeito* (respect), *fim* (end) (6); and
9) No topic: (*no*) *entanto* (but), *porém* (however) (2).

Microtopics 1 and 5 were disregarded. The microtopics 2, 4, 7 and 8 (34 words) were reclustered into a macrotopic called 'human nature' and microtopics 3 and 6 (35 words) under the 'physical nature' macrotopic. The two macrotopics have a balanced number of words in this list without predominance of any of them. This list has words not listed in the previous one and the candidates for keywords, according to their frequency, could vary. An example is 'Kurtz', the most representative character of human degradation at the end of the colonization in the 19th century.

For the HOD_Trevisan list, only the words common to the two previous lists were considered, clustered into the two macrotopics. In a total of 31 words, we obtained the following results:

1) Human nature: *homem* (man), *noite* (night), *olhos* (eyes), *peregrinos* (pilgrims), *barco* (boat), *momento* (moment), *dia* (day), *verdade* (truth), *respeito* (respect), *voz* (voice), *tempo* (time), *homens* (men), *instante* (instant) (13); and
2) Physical nature: *luz* (light), *margem* (shore), *direção* (direction), *coração* (heart), *nada* (nothing), *tão* (so), *rio* (river), *frente* (front), *volta* (around), *dois* (two), *forma* (shape), *antes* (before), *floresta* (forest), *longo* (long), *selvage* (wild), *sol* (sun), *terra* (earth), *mar* (sea), *coisa* (thing) (19).

The results show that the words clustered into the "physical nature" macrotopic, comparing the two lists, Topic 1 and Topic 2, have the most occurrences, confirming the dominance of this topic in the text.

Words of higher frequency in this list could also be the candidates for keywords.

3.2.2 HOD_Junqueira

For HOD_Junqueira_tópico1, with 99 words,[12] we obtained the following microtopics:

10) Narrative: *me* (me), *minha* (my), *mim* (me), *meu* (my), *minhas* (my) (5);
11) Human: Kurtz, *senhor* (Mr.), *homem* (man), *cabeça* (head), *voz* (voice), *homens* (men), *Sr.* (Mr.), *gente* (people), *olhos* (eyes), *rosto* (face), *corpo* (body), *mente* (mind), *alma* (soul) (13);
12) Nature: *rio* (river), *marfim* (ivory), *mundo* (world), *coisa* (thing), *mata* (forest), *trevas* (darkness), *forma* (shape), *coração* (heart), *árvores* (trees), *sombra* (shadow), *vista* (view), *margem* (shore), *coisas* (things), *mar* (sea) (14);
13) Time: *tempo* (time), *sempre* (always), (*de*) *repente* (suddenly),[13] *vezes* (times), *noite* (night), *momento* (moment), *vez* (time), *dia* (day) (8);
14) Description of action, relations and utterances: *parecia* (seemed), *falei* (said), *falou* (said), *ver* (see), *olhar* (look), *disse* (said), *falar* (speak), *tivesse* (had), *pareceu* (seemed), *iria* (would go), *ouvir* (hear), *creio* (believe), *podia* (could), *ouvi* (heard), *fiquei* (was), *via* (saw), *ficava* (stay), *voltar* (return), *saber* (know), *ficar* (stay), *vi* (saw), *acho* (think), *ia* (went) (23);
15) Another description: *grande* (large), *certa* (right), *possível* (possible), *velho* (old), *maior* (greater), *doente* (sick), *novo* (new), *tão* (so), *qualquer* (any), *nada* (nothing), *ninguém* (nobody), *mal* (bad), *demais* (too), *maneira* (way), *nenhum* (none), *meio* (middle), *lado* (side), *redor* (around), *direção*

[12] In this list, '*sombra*' (shadow) could also be part of the 'human' topic and '*vista*' (view) could be part of the 'actions description' topic, etc.

[13] The correct translation of "suddenly" is "de repente". As the align is based on a one-to-one correspondence of lexical items, "suddenly" was aligned only with "repente".

(direction), *caminho* (way), *junto* (together), *lugar* (place) (22);
16) Human artifacts: *barco* (boat), *posto* (post), *porta* (door), *casa* (house), *ponto* (point), *sala* (room) (6);
17) Abstractions: *respeito* (respect), *dor* (pain), *vida* (life), *fim* (end), *final* (final), *verdade* (truth) (6); and
18) No topic: *porém* (but), *embora* (although) (2).

Topics 10, 14 and 18 were disregarded. Microtopics 11, 13, 16 and 17 (33 words in total) were clustered into a "human nature" macrotopic and microtopics 12 and 15 (36 words in total) were clustered into a 'physical nature' macrotopic. It can be argued that the 'physical nature' macrotopic, with the largest number of words in this list, would be the predominant topic of the list.

For HOD_Junqueira_tópico2, with 99 words,[14] we obtained the following microtopics:

10) Narrative: *me* (me), *meu* (my), *mim* (me), *meus* (my), *minha* (my), *minhas* (my) (6);
11) Human: Kurtz, *homem* (man), *gerente* (manager), *Sr* (Mr.), *peregrinos* (pilgrims), *mãos* (hands), *cabeça* (head), *gente* (people), *pés* (feet), *senhor* (Mr.), *homens* (men), *mão* (hand), *Deus* (God), *sujeito* (subject), *olhos* (eyes) (15);
12) Nature: *coisa* (thing), *rio* (river), *terra* (earth), *selva* (forest), *ar* (air), *luz* (light), *coisas* (things), *água* (water), *sol* (sun), *mundo* (world), *mata* (forest), *forma* (shape), *marfim* (ivory) (13);
13) Time: *tempo* (time), *momento* (moment), *meses* (months), *noite* (night), *dia* (day) (5);
14) Description of action, relations and utterances: *dizer* (say), *ver* (see), *fosse* (were), *olhar* (look), *disse* (said), *saber* (know), *fiquei* (became), *parecia* (seemed), *pode* (can), *sei* (know), *vi* (saw), *dar* (give), *poderia* (could), *comecei* (started), *podia* (could), *sabia* (knew), *tivesse* (had) (17);
15) Another description: *negro* (black), *tão* (so), *nada* (nothing), *qualquer* (any), *menos* (less), *nenhuma* (none), *mal* (bad), *dois* (two), *duas* (two), *primeira* (first), *meio* (middle), *frente*

[14] In this list, '*negro*' (black) could belong to the 'human' microtopic.

(front), *longe* (far), *lado* (side), *perto* (near), *fundo* (deep), *acima* (above), *próprio* (own), *alto* (high), *novo* (new), *claro* (clear), *melhor* (best), *impossível* (impossible), *profunda* (deep), *perdido* (lost), *lugar* (place), *antes* (before) (27);

16) Human artifacts: *posto* (post), *trabalho* (work), *barco* (boat), *companhia* (company) (4);
17) Abstractions: *silêncio* (silence), *vida* (life), *fato* (fact), *verdade* (truth), *respeito* (respect), *desejo* (desire), *força* (strength), *morte* (death), *ideia* (idea), *fim* (end) (10); and
18) Not identified: *embora* (although), *afinal* (finally) (2).

Microtopics 10, 14 and 18 were disregarded. Microtopics 11, 13, 16 and 17 (34 words) were reclustered into a macrotopic called 'human nature' and microtopics 12 and 15 (40 words) into the 'physical nature' macrotopic. The ‘physical nature’ macrotopic prevails over the ‘human nature’ macrotopic in this list too. It is also observed that this list has words not listed in the previous one and the candidates for keywords, according to their frequency, could vary.

For the list, only the words common to the two previous lists were considered, clustered into the two aforementioned macrotopics. In a total of 33 words, we obtained the following results:

3) Human nature: *senhor* (Mr.), *dia* (day), *noite* (night), *olhos* (eyes), *barco* (boat), *fim* (end), *vida* (life), *posto* (post), *verdade* (truth), *gente* (people), *respeito* (respect), *tempo* (time), *homem* (man), *homens* (men), *cabeça* (head), Kurtz, *momento* (moment) (17); and
4) Physical nature: *coisa* (thing), *novo* (new), *mundo* (world), *antes* (before), *coisas* (things), *mata* (forest), *meio* (middle), *lado* (side), *rio* (river), *mal* (bad), *tão* (so), *maneira* (way), *marfim* (ivory), *forma* (shape), *lugar* (place), *qualquer* (any) (16).

The results show that the “human nature” macrotopic, comparing the two lists, Topic 1 and Topic 2, has a word more than the ‘physical nature’ macrotopic. In this list, we can say that the two topics are balanced, so the predominance of one or the other is not confirmed. Words of higher frequency in this list could also be the candidates for

keywords.

3.2.3 HOD_Santarrita96

For HOD_Santarrita96_tópico1, with 98 words,[15] the following microtopics were obtained:

19) Narrative: *me* (me), *me* (me), *minha* (my) (3);
20) Human: Kurtz, *homem* (man), *olhos* (eyes), *Sr* (Mr.), *senhor* (Mr.), *cabeça* (head), *gerente* (manager), *homens* (men), *voz* (voice), *gente* (people), *peregrinos* (pilgrims), *mão* (hand), *pés* (feet), *espécie* (species), *pé* (foot) (15);
21) Nature: *coisa* (thing), *terra* (earth), *rio* (river), *ar* (air), *mar* (sea), *mato* (forest), *marfim* (ivory), *luz* (light), *escuridão* (dark), *trevas* (darkness), *água* (water), *floresta* (forest), *margem* (shore) (13);
22) Time: *tempo* (time), *dia* (day), *repente* (suddenly), *vezes* (times), *meses* (months), *dias* (days), *noite* (night) (7);
23) Description of action, relations and utterances: *disse* (said), *dizer* (say), *parecia* (seemed), *pode* (can), *olhar* (look), *ver* (see), *sei* (know), *vi* (saw), *deve* (should), *ia* (went), *sabe* (know), *falar* (speak), *queria* (wanted), *murmurou* (muttered), *fosse* (were), *saber* (know), *sentia* (felt), *pareciam* (seemed), *ouvir* (hear), *sabem* (know), *ficava* (was), exclamou (exclaimed) (22);
24) Another description: *barulho* (noise), *nada* (nothing), *lado* (side), *tão* (so), *qualquer* (any), *acima* (above), *dois* (two), *grande* (large), *negros* (black), *meio* (middle), *simples* (simple), *lugar* (place), *duas* (two), *maior* (larger), *mal* (bad), *nenhum* (none), *perto* (near), *claro* (clear), *menos* (less), *modo* (way), *volta* (around), *própria* (own), *grandes* (large), *negro* (black), *demais* (too), *doente* (sick) (26);
25) Human artifacts: *posto* (post), *vapor* (steam), *porta* (door) (3);
26) Abstractions: *vida* (life), *silêncio* (silence), *verdade* (truth), *poder* (power), *fim* (end), *sensação* (sensation), *morte* (death), *fato* (fact), *certeza* (certainty) (9); and

[15] In this list, '*negro*(s)' (the black) could be under the 'human' topic and '*volta*' (around) could be part of the 'actions description' topic, etc.

27) No topic: *afinal* (finally), *oh* (oh) (2).

Topics 19, 23 and 27 were disregarded. Microtopics 20, 22, 25 and 26 (34 words in total) were clustered into a "human nature" macrotopic and microtopics 21 and 24 (39 words in total) were clustered into a 'physical nature' macrotopic. It can be argued that the 'physical nature' macrotopic, with the largest number of words in this list, would be the predominant topic of the list.

For HOD_Santarrita96_tópico2, with 100 words,[16] the following microtopics were obtained:

19) Narrative: *meu* (my), *minha* (my), *me* (me), *meus* (my), *minhas* (my), *mim* (me) (6);
20) Human: Kurtz, *homem* (man), *senhor* (Mr.), *gente* (people), *gerente* (manager), *Sr* (Mr.), *homens* (men), *voz* (voice), *Deus* (God), *cabeça* (head), *mãos* (hands), *alma* (soul), *sujeito* (subject), *peregrinos* (pilgrims), *multidão* (crowd), *face* (face), *nome* (name) (17);
21) Nature: *coisa* (thing), *rio* (river), *coisas* (things), *coração* (heart), *margem* (shore), *selva* (jungle), *sol* (sun), *terra* (earth), *árvores* (trees), *monte* (mound), *marfim* (ivory), *sombra* (shadow), *mundo* (world), *céu* (sky), *água* (water), *luz* (light), *trevas* (darkness) (17);
22) Time: *vez* (time), *noite* (night), *vezes* (times), *momento* (moment), *repente* (suddenly), *dia* (day) (6);
23) Description of action, relations and utterances: *podia* (could), *via* (saw), *parecia* (seemed), *disse* (said), *vi* (saw), *ver* (see), *sabia* (knew), *ouvi* (heard), *creio* (believe), *parece* (seem), *sabem* (know), *pareciam* (seemed), *vi* (saw), *deu* (gave), *perguntei* (asked), *fiquei* (became), *esperava* (waited), *dizia* (said), *pareceu* (seemed) (19);
24) Another description: *grande* (large), *tão* (so), *frente* (front), *qualquer* (any), *antes* (before), *meio* (middle), *alto* (high), *ninguém* (nobody), *fundo* (deep), *modo* (way), *nada* (nothing), *primeira* (first), *certa* (right), *branco* (white), *negro* (black),

[16] In this list, '*negro*' (black) could also belong to the 'human' microtopic.

trás (behind), *menos* (less), *claro* (clear), *dois* (two), *velho* (old), *perto* (near), *próprio* (own), *região* (region) (23)

25) Human artifacts: *vapor* (steam), *casa* (house), *trabalho* (work), *companhia* (*company*), *rebites* (revites), *Europa* (Europe) (6);
26) Abstractions: *verdade* (truth), *ideia* (idea) (2); and
27) Not identified: *jamais* (never), *embora* (although), *oh* (oh), *afinal* (finally) (4).

Microtopics 19, 23 and 27 were disregarded. Microtopics 20, 22, 25 and 26 (31 words) were reclustered into the 'human nature' macrotopic and microtopics 21 and 24 (40 words) into the 'physical nature' macrotopic. The ‘physical nature’ macrotopic prevails over the 'human nature' macrotopic in this list too. It is also observed that this list has words not listed in the previous one and the candidates for keywords, according to their frequency, could vary.

For the list, only the words common to the two previous lists were considered, clustered into the two aforementioned macrotopics. In a total of 35 words, we obtained the following results:

5) Human nature: *homens* (men), *homem* (man), Kurtz, *peregrinos* (pilgrims), *Sr* (Mr.), *cabeça* (head), *verdade* (truth), *repente* (suddenly), *dia* (day), *gente* (people), *gerente* (manager), *noite* (night), *voz* (voice), *senhor* (Mr.), *vezes* (times), *vapor* (steam) (16); and
6) Physical nature: *luz* (light), *rio* (river), *terra* (earth), *coisa* (thing), *trevas* (darkness), *menos* (less), *dois* (two), *claro* (clear), *nada* (nothing), *tão* (so), *grande* (large), *negro* (black), *qualquer* (any), *água* (water), *modo* (way), *marfim* (ivory), *margem* (shore), *perto* (near), *meio* (middle) (19).

The results show that the “human nature” macrotopic, comparing the two lists, Topic 1 and Topic 2, has a word more than the ‘physical nature’ macrotopic. In this list, it can be said that the 'physical nature' topic is predominant. Words of higher frequency in this list could also be the candidates for keywords.

Section 4 discusses the results obtained by applying both methods on the literary corpus, and comparing their performance based on a

quality-oriented analysis.

4 Discussion of Results

The experiment of human analysis of the lists generated by the keyword module (WordSmith) selected 37 keywords from the first list of 93 words, and 78 from the second list of 249 words. Most candidates for keywords in these lists include lexical words like nouns and adjectives, among others, with some occurrences of grammatical words such as the verb *parecer* (to seem). The occurrence of inflected words from this lemma as keywords was expected, since the version of this lemma in English, *seem*, was also noted by Stubbs (2005) as highly frequent in the word list of the novel in English. Returning to the lists of the experiment of human analysis, among the candidates are the 10 keywords presented in Table 4, illustrating the results obtained when considering the two tools used in the first experiment; for example, Kurtz, protagonist whose search and story is told by the sailor-narrator, and ivory, object of travels during the colonization and exploration of continents such as Africa by some of the empires in the 18th and 19th centuries. It also presents words like *cabeça* (head) and *olhos* (eyes) that indicate a fragmented representation of the native population found in the description of the narrator (MAGALHÃES; ASSIS, 2009). However, this framework does not provide list words such as *parecia* (seemed / singular), *pareciam* (seemed / plural), *escuridão* (darkness), among others, indicators of key topics in the texts, such as uncertainty (STUBBS, 2005) and difficulty to understand what you see in a very different violent clash of cultures with a very distinct power position. The experiment of human analysis of word lists generated by the LDA method allowed the clustering of these words into very general topics, such as physical and human nature. It also pointed to the fact that the words found in lists in which the physical nature topic predominates would most likely be keyword candidates. On the one hand, the two general topics, physical nature and human nature, permeate the work, which can be interpreted as a representation of the Western man's confrontation with the devastation of human and physical natures in a context of colonization. For example, some words from

these lists coincide with some of those presented in Table 4, such as Kurtz, pilgrims, earth and ivory. On the other hand, these general topics leave out words generated by the LDA lists (*qualquer* [any]*, nenhuma* [none]*, parecia* [seemed], etc.), also constituents of other more specific topics, but fundamental in the novel, such as the uncertainty and difficulty of understanding the landscape in the novel. Other words such as *escuridão* (dark), *trevas* (darkness), *sombra* (shadow) are key indicators of the specific topics are not necessarily inferred from general topics either.

We can also say that if there are corpus-based research results which show that conjunctions have a low frequency in novels, an investigation of the occurrence of words such as '(no) *entanto*' ('however'), '*porém*' ('but'), '*embora*' ('although') (the latter if it is in fact a conjunction according to the context), the LDA topics lists could be justified in order to verify the adversative conjunction function in this novel.

On Section 5, final remarks on the study are made.

5 Final remarks

The first aim of this study was to compare the extraction of keywords from a literary corpus using two popular tools, with each of them having a specific operation and efficiency proven in other types of texts and documents. In short, we applied the WordSmith Tools and LDA in three different translations of the same literary work. The WordSmith Tools method compares the text word frequency with a reference average frequency and determines keywords according to the words whose frequency is notably larger. LDA, on the other hand, makes a probabilistic analysis of the text word frequency and generates keyword divided into topics. The lists generated by each tool were concatenated, facilitating the analysis of the results.

The second aim was validate the results of the aforementioned experiment, resulting from an experiment of human analysis of keyword lists obtained with the two tools that would indicate the keyword candidates. In this experiment, we can conclude that for long texts, such

as novels, the human analysis of lists is necessary at a stage prior to experiments to complement the list automatically generated, crossing the results of the tools. A suggestion for a future experiment would be to anticipate the human analysis procedure, i.e., before the integration of the results obtained automatically with the two tools, or even before setting the parameters that allow running LDA, generating more specific and robust results for larger literary texts or corpora.

The results indicate that both methods can be applied to the literary text. However, we observed that human analysis and refining of the methods are crucial. We should also mention that the linguistic intuition of human analysts on examining lists generated by each of the two methods in this experiment was more favorable to the use of WordSmith Tools keyword lists.

In this paper, two topics were automatically generated by LDA, which tended to cluster keywords into two macrotopics from a linguistic point of view: human and physical nature. Future research work include: (a) use LDA extraction with three computational topics and analyze the macrotopics obtained; (b) use a larger number of computational topics to see if the resulting lists tend to be clustered according to the linguistic microtopics obtained in the human analysis; and (c) include new methods and tools in the analysis, such as WMatrix.

6 Acknowledgements

The authors thank CAPES for financing the *Portal Min@s: corpora de fala e escrita* (AUX 151/2013) project, which enabled this research, and all partners from the Interinstitutional Center for Computational Linguistics (NILC) and the Experimental Translation Laboratory (LETRA) who supported this project, especially Professor (Ph.D.) Sandra Maria Aluísio, who generously participated in this article, providing relevant comments and suggestions, and researcher Thais Blauth, who collaborated in the analysis of lists generated by WordSmith Tools.

7 References

BERBER-SARDINHA, T. Comparing *corpora* with WordSmith Tools: How large must the reference corpus be? São Paulo, 2000.

BLEI, D. M. Probabilistic Topic Models: Surveying a suite of algorithms that offer a solution to managing large documents archives. Communications of the ACM, s.l., n. 55, p. 77-84, 2012.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent Dirichlet Allocation. Journal of Machine Learning Research, s.l., n. 3, p. 993-1022, 2003.

CONRAD, J. *O coração da treva*. Translated by Hamilton Trevisan. São Paulo: Global Editora e Distribuidora Ltda., 1984.

CONRAD, J. *O coração das trevas*. Translated by Regina Régis Junqueira. Belo Horizonte: Editora Itatiaia Ltda., 1984.

CONRAD, J. *O coração das trevas*. Translated by Marcos Santarrita. Rio de Janeiro: Ediouro S. A., 1996.

DAVIES, M. The advantage of using relational databases for large corpora: speed, advanced queries, and unlimited annotation. International Journal of Corpus Linguistics 10: 301-28, 2005. DOI: <http://dx.doi.org/10.1075/ijcl.10.3.02dav>.

DAVIES, M. Relational databases as a robust architecture for the analysis of word frequency. In What's in a Wordlist?: Investigating Word Frequency and Keyword Extraction, ed. Dawn Archer. London: Ashgate, p. 53-68, 2009.

DREDZE, M.; WALLACH, H. M.; PULLER, D; PEREIRA, F. Generating summary keywords for emails using topics. In *Proceedings of the 13th international conference on Intelligent user interfaces* (IUI '08). ACM, New York, NY, USA, p. 199-206, 2008. DOI: <http://dx.doi.org/10.1145/1378773.1378800>.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. Discourse Processes, 25:259–284, 1998. DOI: <http://dx.doi.org/10.1080/01638539809545028>.

LIU, Z.; HUANG, W.; ZHENG, Y.; SUN, M. Automatic Keyphrase Extraction via Topic Decomposition. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, p. 366–376, 2010.

MAGALHÃES, C; ASSIS, R. C. Representação de atores sociais em corpus paralelo: Heart of Darkness e suas traduções para o português. In COHEN, Maria Antonieta; LARA, Gláucia Muniz Proença. (Org.). *Linguística, tradução, discurso.* Belo Horizonte: Editora UFMG, 2009, p. 201-220.

MANNING, C. D.; H. SCHÜTZE. Foundations of statistical natural language processing. MIT Press, 2000.

MCINTYRE, D.; WALKER, B. How can corpora be used to explore the language of poetry and drama. In: O'KEEFE, A.; McCARTHY, M. (eds.). *The Routledge handbook of corpus linguistics*. London; New York: Routledge, 2010, p. 516-530.
DOI: <http://dx.doi.org/10.4324/9780203856949.ch37>.

MUNIZ, M.; PAULOVICH, F.; Minghim, R.; INFANTE, K.; Muniz, F.; VIEIRA, R.; ALUÍSIO, S. M. Taming the tiger topic: an XCES compliant corpus Portal to generate subcorpus based on automatic text topic identification. In: Corpus Linguistics 2007 Conference, 2007, Birmingham. Proceedings of the Corpus Linguistics 2007 Conference, 2007.

RAYSON, P. E. Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. Lancaster University, 2002.

SCOTT, M. WordSmith Tools Manual. Oxford: Oxford University Press, 1996.

SCOTT, M. PC analysis of keywords - and key keywords. *System*, vol 25, no. 2, 1997, p. 233-245. DOI: <http://dx.doi.org/10.1016/S0346-251X(97)00011-0>.

STUBBS, M. Conrad in the computer: examples of quantitative stylistic methods. *Language and Literature*, vol 14, no. 1, p. 5-24, 2005. DOI: <http://dx.doi.org/10.1177/0963947005048873>.

TDK TECHNOLOGIES. Topic Modeling Explained: LDA to Bayesian Inference. Retrieved on: July 12, 2015, from: <https://www.tdktech.com/tech-talks/topic-modeling-explained-lda-to-bayesian-inference>.

Apêndice A – Generated Keywords

Table 5 – Keywords obtained from WordSmith Tolls

PODIA (COULD)	SELVAGEM (WILD)
LUZ (LIGHT)	SEQUER (EVEN)KURTZ
ADMINISTRADOR (MANAGER)	VERDADE (TRUTH)
MR (MR – Mister)	SR (MR)
TERRA (EARTH)	HOMENS (MEN)
SOMBRA (SHADOW)	PARECIA (SEEMED)
AR (AIR)	POSTO (POST)
MAR (SEA)	NEGRO (BLACK)
VAPOR (STEAM)	TALVEZ (MAYBE)
ÁGUA (WATER)	GERENTE (MANAGER)
ENTREPOSTO (CUSTOMS)	FLORESTA (FOREST)
BARCO (STEAMBOAT)	SENHOR (MR)
MARGEM (SHORE)	MARFIM (IVORY)
SELVA (JUNGLE)	ESCURIDÃO (DARKNESS)
MATA (FOREST)	CORAÇÃO (HEART)
SILÊNCIO (SILENCE)	PARECIAM (SEEMED)
NADA (NOTHING)	PEREGRINOS (PILGRIMS)
VOZ (VOICE)	

Table 6 – Keywords obtained from LDA

PARECIAM (SEEMED)	COLINA (HILL)
VERDADE (TRUTH)	DIABO (DEVIL)
NADA (NOTHING)	TIMONEIRO (HELMSMAN)
MATAGAL (SCRUB)	ÁRVORES (TREES)
RAZÃO (REASON)	RIBANCEIRA (BANK)
MÃO (HAND)	NAVIO (SHIP)
VULTOS (FIGURES)	ROSTOS (FACES)
MATO (JUNGLE)	ATMOSFERA (ATMOSPHERE)
PÉS (FEET)	VOCÊ (YOU)
SOLIDÃO (LONELINESS)	TREVAS (DARKNESS)
LEME (HELM)	PARECEU (SEEMED)
QUILÔMETROS (KILOMETERS)	TEMPO (TIME)
CAPIM (GRASS)	NAVIOS (SHIPS)
SELVAGEM (WILD)	CLARÃO (GLARE)
TÊNUE (FINE)	AR (AIR)
REMANSO (QUIET)	SOL (SUN)
REBITES (RIVET)	FLORESTA (FOREST)
TREVA (DARKNESS)	TOM (TONE)
MARINHEIRO (SAILOR)	QUIETUDE (QUIET)
ROSTO (FACE)	TOLO (SILLY)
NEGROS (THE BLACK)	BRAÇOS (ARMS)
MARGEM (SHORE)	TOCO (STUB)
SOMBRA (SHADOW)	NATIVOS (NATIVES)
CONVÉS (DECK)	TERRÍVEL (TERRIBLE)
VAPOR (STEAM)	CIMA (TOP)
MISTÉRIO (MYSTERY)	SOMBRAS (SHADOWS)
SELVAGENS (WILD)	ALMA (SOUL)
CÁ (HERE)	ESCURIDÃO (DARKNESS)
VOZES (VOICES)	FULGOR (GLOW)
POSTO (POST)	MARLOW
IMPRESSÃO (IMPRESSION)	GRITO (SCREAM)
PADIOLA (LITTER)	NARIZ (NOSE)
MÃOS (HANDS)	CÉU (SKY)
PROFUNDEZAS (DEPTHS)	CABINA (CABIN)
DEMÔNIO (DEMON)	CABINE (CABIN)
LENHA (FIREWOOD)	MONTE (MOUND)
NEVOEIRO (FOG)	RIO (RIVER)
MATA (FOREST)	NUS (NAKED)
HORROR (HORROR)	NEGRO (BLACK)

DOI: 10.17851/2237-2083.23.3.727-768

Comparação linguística e perfilação gramatical sistêmica em um *corpus* combinado

Linguistic comparison and grammatical systemic profiling in a bidirectional parallel corpus

Francieli Silvéria Oliveira
Universidade Federal de Ouro Preto – UFOP, Minas Gerais.
franciellysilveria@hotmail.com

Resumo: Com base nos pressupostos da Linguística de *Corpus* (BERBER SARDINHA, 2000; VIANA, 2011), este trabalho investiga a organização gramatical e semântica de um *corpus* combinado de manual de instrução no par linguístico inglês / português brasileiro, objetivando apresentar a variação linguística característica desse registro, bem como comparar as línguas e descrever a produção textual de significado da tradução técnica. Pesquisas anteriores apresentam que é possível estudar os registros de uma língua por meio da análise de *corpus* (BIBER, 2010). Figueredo (2014) propõe o uso da perfilação gramatical sistêmica para encontrar padrões gramaticais. Mediante esses conceitos, os resultados apresentaram que o manual de instrução tem como padrão linguístico as funções semânticas '*explicar*', '*comandar*', '*classificar*' e '*introduzir*' com suas respectivas funções gramaticais. Com relação à sua tradução técnica, conclui-se que ela é constituída de textos híbridos e multilíngues que, para significar, utilizam-se do pareamento do texto fonte, da língua alvo e de novos significados.

Palavras-chave: comparação linguística; perfilação gramatical sistêmica; metodologia baseada em *corpus*; tradução técnica.

Abstract: This paper investigates the grammatical and semantic organization of a bidirectional parallel corpus of instruction manuals based on Corpus Linguistics. It aims at presenting the linguistic variation of this register as well as describing text production of technical translations. Previous researches show it is possible to study register through corpus analysis (BIBER, 2010).

Figueredo (2014) suggests using grammar systemic profiling to find grammar patterns. Results show that instruction manuals present linguistic patterns of semantic functions '*expounding*', '*commanding*', '*classifying*' and '*introducing*', along with their respective grammar functions. Regarding technical translations, translated manuals are shown to be hybrids and multilingual texts, which mean they use coupling of source text, target language, and new meanings.

Keywords: linguistic comparison; grammatical systemic profiling; corpus-based methodology; technical translation.

Recebido em: 31 de julho de 2015.
Aprovado em: 29 de setembro de 2015.

1 Introdução

Com base nos pressupostos da Linguística de *Corpus* (BERBER SARDINHA, 2000; BIBER, 2010; VIANA, 2011), o presente artigo investiga o funcionamento gramatical e semântico de um *corpus* combinado (STEINER, 2001; NUNES, 2013) do Manual de Instrução, procurando explicar como se dá a construção de significados desse registro e sua generalização para a Tradução Técnica (AZENHA, 1995).

Tendo como objeto de estudo a produção textual de significado (HALLIDAY, 1994), mais especificamente, a 'dinâmica textual' (LEMKE, 1993), objetiva-se apresentar os padrões gramaticais e semânticos que compõem os manuais de instrução no par linguístico inglês / português brasileiro, por meio da descrição de seus perfis sistêmicos (HALLIDAY, 1978; FIGUEREDO, 2014) como forma de compreender a variação linguística que constrói o texto técnico e a tradução técnica. Complementarmente, visa-se construir um modelo multilíngue que contenha os padrões encontrados no *corpus* e que seja capaz de explicar a sua dinâmica textual.

A Linguística de *Corpus* (LC) e Linguística Sistêmico-Funcional (LSF) constituem base teórica e metodológica desta pesquisa. A LC exerce esse papel por investigar a língua empiricamente, utilizando, para

isso, textos produzidos por falantes de forma espontânea (VIANA, 2011) e que, por isso, compõem o *corpus* de investigação. Por essa razão, Biber (2010, p. 241) propõe que o estudo dos registros de uma língua deve ser pautado na análise de *corpus*, pois ele aponta o funcionamento textual dos registros, como também apresenta seus padrões linguísticos; padrões esses que convencionalmente se associam à sua função cultural. Nesse sentido, os padrões gramaticais e semânticos do manual determinam o que é esse registro: uma macro-onda de significado que capacita o falante a desenvolver uma atividade, instruindo-o de forma a não regular seu comportamento.

Os manuais de instrução aqui utilizados são denominados textos técnicos por apresentarem uma abordagem diferente dos textos expositivos e por se tratarem de uma área específica (AZENHA, 1995). Na Tradução Técnica, Azenha (1995, p. 141) afirma que essa área de estudo apresenta problemas, tais como: lexical-terminológicos, grafo-fonológicos, sintáticos, terminológicos, morfossintáticos, semânticos e receptivos. Dessa maneira, para compreender as relações envolvidas nesse tipo de tradução, faz-se necessária uma reflexão mais sistemática sobre o texto. Para essa reflexão, Catford (1965, p. 35) afirma que a comparação linguística em tradução deve se pautar por uma teoria que consiga descrever tanto a língua fonte quanto a língua alvo. Devido a seu potencial de impacto nesse aspecto (MATTHIESSEN, 2001), a Linguística Sistêmico-Funcional é aqui adotada. A LSF (HALLIDAY, 2013), por ser uma teoria abrangente, consegue descrever e analisar a língua em todos os seus estratos, tais como: a gramática, a semântica e o contexto.

Como forma de cumprir o objetivo desta pesquisa, foi compilado um *corpus* combinado (NUNES, 2013) com quatro manuais de instrução. Do *notebook Macbookpro Retina 13 inch* – cf. *website* "Apple"[1] – coletaram-se o texto original em inglês e sua tradução para o português brasileiro. Da Panela de Pressão Solar Tramontina – cf.

[1] Cf. <http://manuals.info.apple.com/MANUALS/1000/MA1663/en_US/macbook_ pro_retina-13-inch-late-2013_qs.pdf>.

website "Tramontina"[2] – coletaram-se o texto original em português brasileiro e a tradução para o inglês.

Para a análise do *corpus*, os sistemas gramaticais TRANSITIVIDADE, MODO e TEMA (HALLIDAY; MATTHIESSEN, 2014) foram utilizados para compreender a produção textual de significado e determinar os padrões gramaticais e semânticos do manual. Mais especificamente para comparar os textos nos pares original / tradução e original / original, a fim de responder como se caracteriza a tradução técnica desse registro: um texto multilíngue que, para significar, utiliza-se do pareamento do texto fonte, da língua alvo e de novos significados.

Com base na metodologia de perfilação gramatical (FIGUEREDO, 2014), o artigo apresenta como resultado o perfil gramatical sistêmico do *corpus* combinado de manuais. Esse perfil mostra as escolhas gramaticais e semânticas mais recorrentes dos manuais de instrução, apresentando, assim, seu funcionamento, além de converter "funções gramaticais em "*tokens* gramaticais" (FIGUEREDO, 2014, p. 18), ampliando as buscas de padrões por meio da metodologia de *corpus*.

Ao final, este artigo demonstra como o arcabouço da LC com a LSF possibilitam apresentar o funcionamento da produção textual de significado do registro: manual de instrução, assim como permite compreender de que modo a tradução técnica do manual de instrução é construída mediante a comparação da variação linguística do *corpus* combinado.

2 A Linguística de *Corpus* na comparação linguística e perfilação sistêmica

Tendo como principal objetivo estudar a língua em seu funcionamento por meio de dados linguísticos reais, produzidos por

[2] Cf. <http://suipnovo.tecnologia.ws/public/upload/product/62516223/62516222FLM001>.

falantes de forma espontânea, a LC ocupa-se da compilação e análise do *corpus* com base em padrões emergentes do uso da língua. Sinclair (1997, p. 31) afirma que "a língua não pode ser inventada; mas apenas capturada". Levando isso em consideração e tomando como base os pressupostos metodológicos da LC, o *corpus* de pesquisa deve ser coletado de forma criteriosa para servir de base para o estudo da língua e da variedade linguística em foco.

Desse modo, Berber Sardinha (2000, p. 338) coloca como condição fundamental para a pesquisa a necessidade do *corpus* possuir dados autênticos. O autor observa que o *corpus* deve ser compilado segundo o propósito da pesquisa, que seus dados devem ser criteriosamente coletados, que devem representar a língua, além de serem legíveis por meio do computador.

Como a LC apresenta métodos para tratar dados linguísticos (VIANA, 2011), ela pode ser usada em vários campos dos estudos da linguagem, como o estudo do léxico, da gramática, no ensino de línguas, no estudo dos registros, nos estudos da tradução, entre outros. No caso deste artigo, a LC foi utilizada no estudo do registro 'manual de instrução' e na sua tradução, por meio da comparação linguística (CATFORD, 1965; KRZESZOWSKI, 1990) e da perfilação sistêmica (FIGUEREDO, 2014).

2.1 Comparação linguística e estudos da tradução

Nos estudos da tradução, a LC cumpre papel fundamental de base metodológica pelo fato de trabalhar com a língua em uso (BERBER SARDINHA, 2002), uma vez que o *corpus* compilado e construído com o propósito de comparar as línguas contribui para a realização de uma pesquisa mais eficaz e replicável (VIANA, 2011).

Nos estudos da tradução, a comparação entre línguas é uma forma de fazer a análise textual das traduções e descobrir os processos linguísticos que as envolvem. Halliday *et al*. (1964) afirmam que todas as línguas são sistematicamente comparáveis, uma vez que elas possuam categorias teóricas que consigam descrevê-las. Diante disso, Catford (1965) apresenta dois tipos de comparação linguística: a equivalência textual e a correspondência formal. A equivalência textual ocorre quando

uma porção de texto na língua fonte, por permutação, é equivalente na língua alvo. A correspondência formal ocorre quando se pode apontar que uma categoria teórica do texto alvo possui uma categoria teórica correspondente na língua fonte. Como forma de ampliar o estudo de Catford, Krzeszowski (1990) afirma que as línguas são estruturalmente comparáveis, na medida que são estruturalmente semelhantes. Neste artigo, as categorias utilizadas na análise do *corpus* de manuais de instrução descrevem o inglês e o português brasileiro (FIGUEREDO, 2011; FIGUEREDO, 2015).

O manual de instrução aqui analisado é um texto técnico, portanto, a sua tradução, que também trata de uma área específica, é uma tradução técnica. A respeito desse campo da tradução, Azenha (1995) propõe que todos os níveis de hierarquia linguística são importantes para o estudo e a tradução do texto técnico, sendo assim, necessária uma reflexão mais sistemática sobre esse tipo de tradução. Para essa reflexão, a LSF foi utilizada na análise do *corpus*.

2.2 Perfilação gramatical sistêmica

A perfilação gramatical é uma metodologia baseada nos pressupostos da LC que permite a identificação de padrões e a análise de *corpora* para a investigação de como as funções gramaticais constroem significado na dinâmica textual (FIGUEREDO, 2014, p. 17). Essa metodologia contribui para os estudos da LC por possibilitar a conversão de categorias gramaticais em "*tokens* gramaticais", permitindo, assim, o processamento da gramática automaticamente.

Uma vez que língua é compreendida pela LC como probabilidade, o estudo dos padrões de uso recorrentes, como também a ocorrência da cosseleção de itens (VIANA, 2011, p. 39) se torna a forma mais eficaz de investigação. A perfilação gramatical segue esses preceitos, analisando as ocorrências gramaticais oracionais e detectando padrões nas cosseleções gramaticais.

Por conseguinte, a perfilação gramatical também contribui na busca de padrões em *corpora*, por meio da qual é possível ainda identificar padrões gramaticais relacionados ao estudo dos registros. Biber (2010, p. 246) aponta que o estudo dos registros de uma língua

deve ser embasado em dados de *corpus*, dados esses que apontam a variação linguística característica dos registros. É mediante a frequência no *corpus* que os padrões linguísticos podem ser encontrados (BIBER *et al*., 2004, p. 376).

A perfilação gramatical se vale da LSF por ela constituir-se de uma teoria que abrange todos os estratos linguísticos e por possuir, em cada estrato, categorias dispostas em rede de sistemas para analisar e descrever a língua (cf. HALLIDAY; MATTHIESSEN, 2014; entre outros). Pela LSF, é possível compreender como as escolhas linguísticas estão dispostas no *corpus* e como elas, pelas suas cosseleções, constroem o significado de cada texto.

3 Metodologia

Neste artigo, é apresentada a construção da tradução técnica do *corpus* de manual de instrução no par linguístico inglês e português brasileiro tendo como base a LC (BERBER SARDINHA, 2000; BIBER, 2010; VIANA, 2011) e a LSF (HALLIDAY, 2002; MARTIN; ROSE, 2007; FIGUEREDO, 2014), como também o perfil sistêmico do *corpus* combinado. Para tanto, a seguir serão apresentados os procedimentos aqui utilizados.

3.1 *Corpus*

Para analisar o funcionamento gramatical dos manuais, foi compilado um *corpus* combinado (STEINER, 2001; KENNING, 2010), definido por Nunes (2013, p. 7) como

> um córpus bidirecional no qual um conjunto de textos originais em uma língua A são traduzidos para uma língua B e originais na língua B são traduzidos para a língua A, de forma que seja possível combinar todas as perspectivas paralelas e comparáveis para análise de dados.

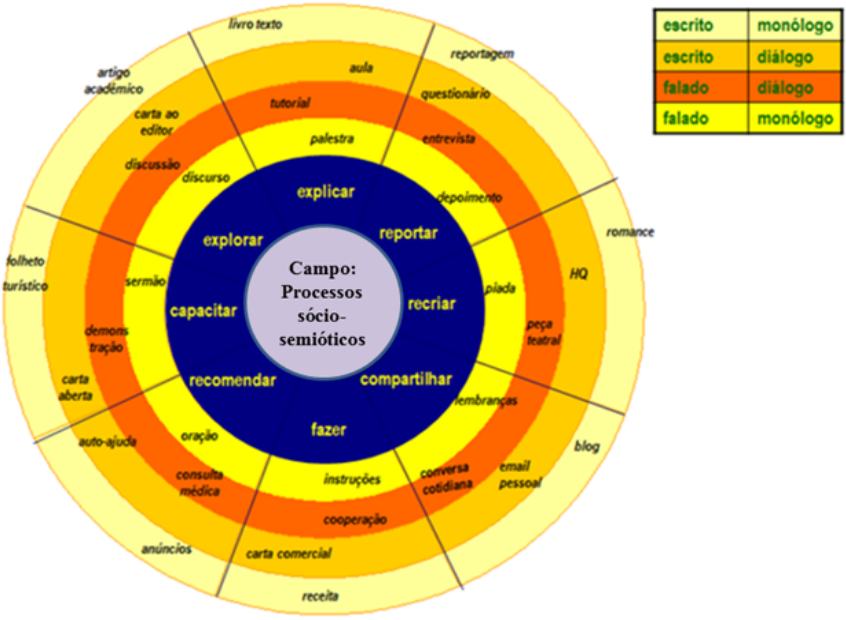
No corpus combinado deste artigo, a língua A mencionada por Nunes se refere ao inglês e a língua B ao português brasileiro. Além disso, a respeito do número de tokens coletados para cada texto, este trabalho segue a proposta metodológica de Biber (1990), que aponta o número mínimo de 1.000 tokens para representar uma amostra autêntica da língua utilizada e para analisar os padrões de frequência de um corpus.

Dessa forma, foram coletados quatro manuais de instrução para o corpus combinado de 1.000 tokens cada um. Em inglês original e tradução para o português brasileiro, foram coletados no website "Apple" os manuais do Macbookpro retina 13 inch, e em português brasileiro original e tradução para o inglês foram coletados no website "Tramontina" os manuais da panela de pressão Solar Tramontina.

Os textos foram compilados de acordo com a tipologia da língua no contexto de cultura (URE, 1989; MATTHIESSEN et al., 2008) e tiveram como base os oito processos sociossemióticos: explicar, relatar, recriar, compartilhar, fazer, recomendar, capacitar e explorar, e os quatro modos de produção: escrito / monólogo, escrito / diálogo, falado / monólogo e falado / diálogo.

Halliday e Matthiessen (2014, p. 35) dividem esses oito processos sociossemióticos em dois grupos, o primeiro em fazer e o segundo em significar: (1) processo sociossemiótico 'fazer': semiotiza significados relacionados com o comportamento e a ação humanos; nesse sentido, a língua é usada para facilitar a execução de uma atividade; (2) 'significar' abarca os demais processos sociossemióticos: caracteriza-se por constituir uma situação na qual os significados são criados. A Figura 1 apresenta os processos sociossemióticos:

Figura 1 – Processos sociossemióticos e os modos de produção com exemplos



Fonte: Adaptado de Matthiessen *et al.*, 2008.

Os textos do *corpus* deste artigo estão localizados no processo sociossemiótico ‘capacitar’ e no modo de produção escrito monólogo. Textos pertencentes a esse processo sociossemiótico podem conter procedimentos simples, como também procedimentos altamente técnicos (HALLIDAY; MATTHIESSEN, 2014).

Os manuais de instrução compilados para esta pesquisa têm como objetivo auxiliar o falante a usar adequadamente o aparelho em questão; no caso, o *notebook* da marca *Apple* e a panela de pressão Solar da Tramontina.

Após a coleta do *corpus*, os manuais de instrução foram etiquetados e armazenados. Para o etiquetamento e armazenamento, este trabalho teve como base as regras do Catálogo da Língua Brasileira (CALIBRA).

3.2 Metodologia de Análise

Cada texto do *corpus* combinado – após coleta nos *websites* e subsequente conversão para formato .txt – foi transferido para a ferramenta computacional *UAM CorpusTool*© (O'DONNELL, 2008). A análise teve como base os sistemas gramaticais TRANSITIVIDADE, MODO e TEMA (HALLIDAY; MATTHIESSEN, 2014). Como o *corpus* desta pesquisa é composto por duas línguas, foram utilizadas categorias do ambiente multilíngues (FIGUEREDO, 2015), dessa maneira, as categorias descritas a seguir representam aquelas utilizadas na análise do *corpus*.

O sistema de TRANSITIVIDADE é responsável por apresentar a língua em ação, por representar gramaticalmente o que acontece no mundo. Para a análise, utilizaram-se suas funções de Processos Material, Mental, Relacional, Verbal e Comportamental (HALLIDAY; MATTHIESSEN, 2014).

A metafunção interpessoal é responsável por apresentar o nível de relação existente entre os interlocutores de um texto. Essa categoria é representada pelo sistema de MODO que possui as funções de Imperativo: Jussivo, Imperativo: Sugestivo, Indicativo: Declarativo, Indicativo: Interrogativo: Polar e Indicativo: Interrogativo: Elemental (FIGUEREDO, 2011).

O sistema de TEMA (MARTIN; ROSE, 2007) gera as funções: *Default*, Tema Elemental, Ângulo: Fonte, Ângulo: Ponto de Vista, Perspectiva e Intensivo.

Posterior à análise da TRANSITIVIDADE, MODO e TEMA, para compreender as fases do discurso que compõem o manual, também foi feita a análise do *corpus* pelo viés das ondas de informação. Para tanto, foi incluído na análise do perfil metafuncional o sistema de MENSAGEM (cf. MARTIN; ROSE, 2007). As ondas de informação representam as fases do discurso em que os significados são reunidos para apresentar um significado maior.

3.3 Metodologia de perfilação sistêmica

Para a criação do perfil sistêmico do *corpus*, foram necessários procedimentos para visualizar a disposição dos significados de cada manual e, assim, realizar a perfilação. Para isso, as categorias da LSF foram colocadas no espaço gramatical segundo o mapa topológico de cada metafunção, estabelecendo-se, assim, suas distâncias relativas (FIGUEREDO, 2014).

As distâncias topológicas, então representadas por distâncias topológicas gramaticais, formam a base da descrição das orações de cada texto. Para exemplificar, o Quadro 1[3] apresenta um exemplo da análise das distâncias topológicas, aqui descritas como um código numérico.

Quadro 1 – Exemplo de descrição das distâncias topológicas no espaço gramatical expressa em código numérico

ORAÇÃO	Transitividade (x)	Modo (y)	Tema (z)	Ondas de Informação
CAP_01_01	14	20	10	1ª
CAP_01_02	14	20	10	
CAP_01_03	39	20	10	
CAP_01_04	14	20	45	
CAP_01_05	39	11	10	2ª
CAP_01_06	16	20	10	
CAP_01_07	14	11	10	

O Quadro 1 aponta o número de cada oração composto pelo código da etiqueta do texto (por exemplo: CAP_01) com o número referente à ordem de cada oração (por exemplo: CAP_01_01). A segunda coluna apresenta as funções da TRANSITIVIDADE do texto, por meio da sua distância topológica relativamente às outras funções (por exemplo: 14 = Processo: Material Transitivo e Transformativo / 39 = Processo: Relacional Identificativo e Intensivo). A terceira coluna apresenta o MODO (por exemplo: 20 = Modo: Declarativo). A quarta apresenta o TEMA (por exemplo: 10 = Tema: *Default* / 45 = Tema: Intensivo). A quinta coluna apresenta cada onda de informação. Esse tipo de quadro revela todas as orações do texto com a disposição da TRANSITIVIDADE, do MODO, do TEMA e das ondas de informação, permitindo, assim, a visualização da mudança entre orações.

[3] As tabelas e quadros desta pesquisa são de nossa autoria.

Consequentemente, as mudanças de significado no manual podem ser visualizadas.

Além da visualização da mudança de significado do manual, para perfilação sistêmica, é necessário o cálculo das frequências relativas de cada categoria supracitada em relação ao *corpus* e às ondas de informação. O resultado desse cálculo aponta as probabilidades das escolhas dos sistemas gramaticais para o *corpus* combinado de manuais em inglês e em português brasileiro.

4 Resultados da análise e discussão da construção de significado do *corpus* combinado

Os quatro textos do *corpus* de manuais foram analisados pelo viés textual por meio das frequências obtidas do TEMA, interpessoal pelas frequências obtidas do MODO, ideacional pelas frequências obtidas da TRANSITIVIDADE e, por fim, pelas frequências das ondas de informação.

As frequências obtidas no TEMA explicam como a relação entre os significados foi construída. As frequências obtidas no MODO mostram a relação entre autor / leitor no discurso dos manuais. As frequências obtidas na TRANSITIVIDADE expõem como a representação das experiências do mundo se apresenta nesse tipo de texto. E a análise das ondas de informação aponta como a relação entre essas três categorias constroem o significado semântico de cada texto.

4.1 Frequências relativas ao TEMA, MODO e TRANSITIVIDADE do manual *Macbookpro Retina 13 inch* em IO

As Tabelas 1, 2 e 3 a seguir apresentam, respectivamente, as frequências relativas ao TEMA, MODO e TRANSITIVIDADE obtidas na análise do Manual *Macbookpro Retina 13 inch* em IO. As análises das categorias foram realizadas de forma semiautomática com o *software UAM CorpusTool* © (O'DONNELL, 2008).

Tabela 1 – Resultados da análise textual do Manual *Macbookpro Retina 13 inch* em IO

CAP_01		
TEMA	Frequência	Nº de mensagens
default	84.3%	75
tema elemental	0.0%	0
ângulo fonte	0.0%	0
ângulo ponto de vista	0.0%	0
tema perspectiva	15.7%	14
tema intensivo	0.0%	0
TOTAL	100%	87

Tabela 2 - Resultados da análise Interpessoal do Manual *Macbookpro Retina 13 inch* em IO

CAP_01		
MODO	**Frequência**	**Nº de mensagens**
jussivo	44.9%	40
sugestivo	2.2%	2
declarativo	52.8%	47
polar	0.0%	0
elemental	0.0%	0
TOTAL	**100%**	**89**

Tabela 3 – Resultados da análise Ideacional do Manual *Macbookpro Retina 13 inch* em IO

(Continua)

CAP_01		
TIPO-DE-PROCESSO	**Frequência**	**Nº de mensagens**
material	74.2%	**66**
TIPO-DE-FAZER		N=66
transformativo	87.9%	58
criativo	12.1%	8
TIPO-DE-IMPACTO		N=66
intransitivo	7.6%	5
transitivo	92.4%	61
mental	9.0%	**8**
TIPO-DE-MENTAL		N=8
superior	87.5%	7
inferior	12.5%	1

Tabela 3 – Resultados da análise Ideacional do Manual *Macbookpro Retina 13 inch* em IO

(Conclusão)

CAP_01		
TIPO-DE-PROCESSO	**Frequência**	**Nº de mensagens**
TIPO-DE-SUPERIOR	N=7	
cognitivo	100.0%	7
desiderativo	0.0%	0
TIPO-DE-INFERIOR	N=1	
perceptivo	100.0%	1
emotivo	0.0%	0
TIPO-DE-FENOMENALIZAÇÃO	N=8	
fenomenalização	75.0%	6
não-fenomenalização	25.0%	2
verbal	**0.0%**	**0**
TIPO-DE-RECEPÇÃO	N=0	
não-recepção	0.0%	0
recepção	0.0%	0
TIPO-DE-VERBAL	N=0	
atividade	0.0%	0
semiose	0.0%	0
relacional	13.5%	**12**
TIPO-DE-RELAÇÃO	N=12	
intensivo	33.3%	4
posessivo	33.3%	4
circunstancial	33.3%	4
MODO-DE-RELAÇÃO	N=12	
atributivo	58.3%	7
identificativo	41.7%	5
existencial	3.4%	**3**
TIPO-DE-EXISTENCIAL	N=3	
criação	0.0%	0
permanência	100.0%	3
TIPO-DE-CRIAÇÃO	N=0	
surgimento	0.0%	0
extinção	0.0%	0
TIPO-PERMANÊNCIA	N=3	
positiva	100.0%	3
negativa	0.0%	0
EXISTÊNCIA	N=3	
introdução	33.3%	1
representação	66.7%	2
TOTAL	**100%**	**89**

4.2 Análise das ondas de informação do manual *Macbookpro Retina 13 inch* em IO

O manual em IO é composto por quinze ondas de informação que juntas exercem a função de fazer com que o falante compreenda e utilize os aplicativos do *Macbookpro Retina 13 inch*. Elas são responsáveis por apresentar o manual, fazer explicações sobre os aplicativos do aparelho, explicar como conseguir mais informações sobre o *notebook* e dar informações sobre a garantia do produto.

Para construir esses significados, esse manual utiliza as funções: (1) EXPLICAR,[4] (2) COMANDAR, (3) CLASSIFICAR, (4) CONVIDAR e (5) INTRODUZIR.

(1) A função EXPLICAR é realizada pelo Processo: Material (14 = Transitivo e Transformativo / 16 – Transitivo e Criativo / 11 – Intransitivo e Transformativo) com o Modo: Declarativo (20) e com o Tema: *Default* (10) ou Tema Perspectiva (41). Ela também pode ser realizada pelo Processo: Mental (56 – Perceptivo e Fenomenalização / 59 = Cognitivo e Fenomenalização) com o Modo: Declarativo (20) e com o Tema: *Default* (10) ou Tema Perspectiva (41).

A função (2) COMANDAR é realizada pelo Modo: Jussivo (11). Ele pode realizar com o Processo: Material (14/16/11) cosselecionando com o Tema: Default ou Tema: Perspectiva, ou com o Processo: Mental (59 – Cognitivo e Fenomenalização / 54 - Cognitivo e Não Fenomenalização) e com o Tema: *Default* (10).

A função (3) CLASSIFICAR ocorre no manual para caracterizar participantes e para auxiliar nas explicações, pois só por meio delas que o falante terá condições de compreender o que está sendo explicado. Estas informações novas são realizadas por Processo: Relacional (39 = Processo: Relacional Identificativo e Intensivo / 37 – Identificativo e Circunstancial / 33 – Atributivo e Intensivo / 32 – Atributivo e Possessivo) que vem sempre acompanhado do Modo: Declarativo (20) e com o Tema: *Default* (10) ou Tema: Perspectiva (41).

---

[4] As funções semânticas estão em caixa alta somente para melhorar a visualização do leitor. Elas não estão aqui representando sistemas como Martin (2013) propõe.

A função (4) CONVIDAR é responsável por estimular o leitor a aprender sobre o notebook, causando, assim, proximidade. Para isso, o Modo: Sugestivo (15) foi utilizado sempre acompanhado do Processo: Material (14/11) e do Tema: *Default* (10).

A função (5) INTRODUZIR insere participantes novos e auxilia nas explicações; para isso, o Processo: Existencial: Permanência (29) é utilizado. No caso do IO, ele sempre se apresenta com o Modo: Declarativo (20) e com o Tema: *Default* (10) ou Tema: Perspectiva (41).

4.3 Frequências relativas ao TEMA, MODO e TRANSITIVIDADE do manual *Macbookpro Retina 13 inch* em PT

As Tabelas 4, 5 e 6 a seguir expõem, respectivamente, as frequências relativas ao TEMA, MODO e TRANSITIVIDADE obtidas na análise do Manual *Macbookpro Retina 13 inch* em PT:

Tabela 4 – Resultados da análise textual do Manual *Macbookpro Retina 13 inch* em PT

CAP_02		
TEMA	**Frequência**	**Nº de mensagens**
Default	82.9%	68
tema elemental	0.0%	0
ângulo fonte	0.0%	0
ângulo ponto de vista	0.0%	0
tema perspectiva	0.0%	0
tema intensivo	17.1%	14
TOTAL	100%	82

Tabela 5 - Resultados da análise Interpessoal do Manual *Macbookpro Retina 13 inch* em PT

CAP_02		
MODO	**Frequência**	**Nº de mensagens**
jussivo	48.8%	40
sugestivo	2.4%	2
declarativo	48.8%	40
polar	0.0%	0
elemental	0.0%	0
TOTAL	100%	82

Tabela 6 – Resultados da análise Ideacional do Manual *Macbookpro Retina 13 inch* em PT

(Continua)

CAP_02		
TIPO-DE-PROCESSO	**Frequência**	**Nº de mensagens**
material	75.6%	**62**
TIPO-DE-FAZER	N=62	
transformativo	90.3%	56
criativo	9.7%	6
TIPO-DE-IMPACTO	N=62	
intransitivo	1.6%	1
transitivo	98.4%	61
mental	8.5%	**7**
TIPO-DE-MENTAL	N=7	
superior	85.7%	6
inferior	14.3%	1
TIPO-DE-SUPERIOR	N=6	
cognitivo	100.0%	5
desiderativo	0.0%	0
TIPO-DE-INFERIOR	N=1	
perceptivo	100.0%	1
emotivo	0.0%	0
TIPO-DE-FENOMENALIZAÇÃO	N=7	
fenomenalização	71.4%	5
não fenomenalização	28.6%	2
verbal	**0.0%**	**0**
TIPO-DE-RECEPÇÃO	N=0	
não-recepção	0.0%	0
recepção	0.0%	0
TIPO-DE-VERBAL	N=0	
atividade	0.0%	0
semiose	0.0%	0
relacional	13.4%	**11**
TIPO-DE-RELAÇÃO	N=11	
intensivo	27.3%	3
posessivo	36.4%	4
circunstancial	36.4%	4
MODO-DE-RELAÇÃO	N=11	
atributivo	63.6%	7
identificativo	36.4%	4
existencial	2.4%	**2**

Tabela 6 – Resultados da análise Ideacional do Manual *Macbookpro Retina 13 inch* em PT

(Conclusão)

CAP_02		
TIPO-DE-PROCESSO	**Frequência**	**Nº de mensagens**
TIPO-DE-EXISTENCIAL	N=2	
criação	0.0%	0
permanência	100.0%	3
TIPO-DE-CRIAÇÃO	N=0	
surgimento	0.0%	0
extinção	0.0%	0
TIPO-PERMANÊNCIA	N=2	
positiva	100.0%	3
negativa	0.0%	0
EXISTÊNCIA	N=2	
introdução	50.0%	1
representação	50.0%	2
TOTAL	**100%**	**79**

4.4 Análise das ondas de informação do manual *Macbookpro Retina 13 inch* em PT

O significado gerado pela junção de TRANSITIVIDADE, MODO e TEMA das mensagens e o significado macro gerado pela junção das mensagens compõem as ondas de informação (MARTIN; ROSE, 2007). Elas são responsáveis por apresentar o manual, fazer explicações sobre os aplicativos do aparelho, explicar como conseguir mais informações sobre ele e dar informações sobre a garantia do produto.

Para construir esses significados, esse manual utiliza as funções: (1) EXPLICAR, (2) COMANDAR, (3) CLASSIFICAR, (4) CONVIDAR e (5) INTRODUZIR.

A função (1) EXPLICAR é realizada pelo Processo: Material (14 = Transitivo e Transformativo / 16 – Transitivo e Criativo) com o Modo: Declarativo (20) e com o Tema: *Default* (10) ou Tema Perspectiva (41). Ela também pode ser realizada pelo Processo: Mental (59 = Cognitivo e Fenomenalização) com o Modo: Declarativo (20) e com o Tema Perspectiva (41).

A função (2) COMANDAR é realizada pelo Modo: Jussivo (11). Ele pode realizar com o Processo: Material (14/16) cosselecionando com o Tema: *Default* ou Tema: Perspectiva, ou com o Processo: Mental (56 – Perceptivo e Fenomenalização / 59 – Cognitivo e Fenomenalização / 54 - Cognitivo e Não Fenomenalização) e com o Tema: *Default* (10).

A função (3) CLASSIFICAR ocorre no manual para caracterizar participantes e para auxiliar nas explicações, pois só por meio delas o falante terá condições de compreender o que está sendo explicado. Essas informações novas são realizadas por Processo: Relacional (33 – Atributivo e Intensivo / 32 – Atributivo e Possessivo / 37 – Identificativo e Circunstancial) que vem sempre acompanhado do Modo: Declarativo (20) e com o Tema: *Default* (10) ou Tema: Perspectiva (41).

A função (4) CONVIDAR é responsável por estimular o leitor a aprender sobre o *notebook*, causando, assim, uma proximidade. Para isso, o Modo: Sugestivo foi utilizado sempre acompanhado do Processo: Material (14/11) e do Tema: *Default* (10).

A função (5) INTRODUZIR insere participantes novos e auxilia nas explicações, para isso, o Processo: Existencial: Permanência (29) é utilizado. No caso do PT, ele se apresenta com o Modo: Declarativo (20) e com o Tema: *Default* (10) ou Tema: Perspectiva (41).

4.5 Frequências relativas ao TEMA, MODO e TRANSITIVIDADE do manual Panela de Pressão Solar em PO

As Tabelas 7, 8 e 9 a seguir mostram, respectivamente, as frequências relativas ao TEMA, MODO e TRANSITIVIDADE obtidas na análise do Manual Panela de Pressão Solar em PO.

Tabela 7 – Resultados da análise textual do Manual Panela de Pressão Solar em PO

CAP_03		
TEMA	**Frequência**	**Nº de mensagens**
default	61.9%	39
tema elemental	0.0%	0
ângulo fonte	0.0%	0
ângulo ponto de vista	0.0%	0
tema perspectiva	38.1%	24
tema intensivo	0.0%	0
TOTAL	100%	63

Tabela 8 – Resultados da análise Interpessoal do Manual Panela de Pressão Solar em PO

CAP_03		
MODO	**Frequência**	**Nº de mensagens**
jussivo	55.6%	35
sugestivo	0.0%	0
declarativo	44,4%	28
polar	0.0%	0
elemental	0.0%	0
TOTAL	**100%**	**63**

Tabela 9 – Resultados da análise Ideacional do Manual Panela de Pressão Solar em PO

(Continua)

CAP_03		
TIPO-DE-PROCESSO	**Frequência**	**Nº de mensagens**
material	77.8%	**49**
TIPO-DE-FAZER	N=49	
transformativo	95.9%	47
criativo	4.1%	2
TIPO-DE-IMPACTO	N=49	
intransitivo	0.0%	0
transitivo	100.0%	49
mental	**9.5%**	**6**

Tabela 9 – Resultados da análise Ideacional do Manual Panela de Pressão Solar em PO

(Conclusão)

CAP_03		
TIPO-DE-PROCESSO	**Frequência**	**Nº de mensagens**
TIPO-DE-MENTAL	N=6	
superior	100.0%	6
inferior	0.0%	0
TIPO-DE-SUPERIOR	N=6	
cognitivo	100.0%	6
desiderativo	0.0%	0
TIPO-DE-INFERIOR	N=0	
perceptivo	0.0%	0
emotivo	0.0%	0
TIPO-DE-FENOMENALIZAÇÃO	N=6	
fenomenalização	100.0%	6
não fenomenalização	0.0%	0
verbal	**3.2%**	**2**
TIPO-DE-RECEPÇÃO	N=2	
não-recepção	100.0%	2
recepção	0.0%	0
TIPO-DE-VERBAL	N=2	
atividade	0.0%	0
semiose	100.0%	2
relacional	7.9%	**5**
TIPO-DE-RELAÇÃO	N=5	
intensivo	60.0%	3
posessivo	40.0%	4
circunstancial	0.0%	0
MODO-DE-RELAÇÃO	N=5	
atributivo	100.0%	4
identificativo	0.0%	0
existencial	**1.6%**	**1**
TIPO-DE-EXISTENCIAL	N=1	
criação	0.0%	0
permanência	100.0%	1
TIPO-DE-CRIAÇÃO	N=0	
surgimento	0.0%	0
extinção	0.0%	0
TIPO-PERMANÊNCIA	N=1	
positiva	100.0%	1
negativa	0.0%	0
EXISTÊNCIA	N=1	
introdução	100.0%	1
representação	0.0%	0
TOTAL	**100%**	**63**

4.6 Análise das ondas de informação do manual Panela de Pressão Solar em PO

O manual em PO é realizado por 18 ondas de informação que, juntas, exercem a função de habilitar o leitor a compreender a forma correta de manusear a panela de pressão, evitando, assim, acidentes domésticos. As ondas de informação são responsáveis por apresentar o manual, advertir o leitor, dar comandos para que o leitor use corretamente a panela, fazer explicações para informar porque o leitor deve obedecer aos comandos e explicar questões envolvidas no uso e na manutenção da panela.

Para construir esses significados, esse manual utiliza as funções: (1) EXPLICAR, (2) COMANDAR, (3) CLASSIFICAR e (4) INTRODUZIR.

A função (1) EXPLICAR é realizada pelo Processo: Material (14 = Transitivo e Transformativo / 16 – Transitivo e Criativo) com o Modo: Declarativo (20) e com o Tema: *Default* (10) ou Tema Perspectiva (41). Ela também pode ser realizada pelo Processo: Mental (59 = Cognitivo e Fenomenalização) com o Modo: Declarativo (20) e com o Tema Perspectiva (41).

A função (2) COMANDAR é realizada pelo Modo: Jussivo (11). Ele pode realizar com o Processo: Material (14/16) cosselecionando com o Tema: *Default* ou Tema: Perspectiva ou com o Processo: Mental (59 – Cognitivo e Fenomenalização) e com o Tema: *Default* (10) ou Tema Perspectiva (41).

A função (3) CLASSIFICAR ocorre no manual para caracterizar participantes e para auxiliar nas explicações. Essas informações novas são realizadas por Processo: Relacional (33 – Atributivo e Intensivo / 32 – Atributivo e Possessivo) que vem sempre acompanhado do Modo: Declarativo (20) e com o Tema: *Default* (10) ou Tema: Perspectiva (41).

A função (4) INTRODUZIR insere participantes novos e significados novos no texto e auxilia nas explicações, para isso, o Processo: Existencial: Permanência (29) é utilizado e o Processo: Verbal (41 – Semiose e Não Recepção). No caso do Processo: Existencial, ele vem com o Modo: Declarativo (20) e com o Tema: *Default* (10) e o

Processo: Verbal vem acompanhado do Modo: Declarativo (20) e com o Tema: *Default* (10) ou o Tema: Perspectiva.

4.7 Frequências relativas ao TEMA, MODO e TRANSITIVIDADE do manual Panela de Pressão Solar em IT

As Tabelas 10, 11 e 12 a seguir apresentam, respectivamente, as frequências relativas ao TEMA, MODO e TRANSITIVIDADE obtidas na análise do Manual Panela de Pressão Solar em IT.

Tabela 10 – Resultados da análise textual do Manual Panela de Pressão Solar em IT

CAP_04		
TEMA	**Frequência**	**Nº de mensagens**
default	54.0%	34
tema elemental	0.0%	0
ângulo fonte	0.0%	0
ângulo ponto de vista	0.0%	0
tema perspectiva	46.0%	29
tema intensivo	0.0%	0
TOTAL	**100%**	**63**

Tabela 11 – Resultados da análise Interpessoal do Manual Panela de Pressão Solar em IT

CAP_04		
MODO	**Frequência**	**Nº de mensagens**
jussivo	61.9%	39
sugestivo	0.0%	0
declarativo	38.1%	24
polar	0.0%	0
elemental	0.0%	0
TOTAL	**100%**	**63**

Tabela 12 – Resultados da análise Ideacional do Manual Panela de Pressão Solar em IT

(Continua)

CAP_04		
TIPO-DE-PROCESSO	**Frequência**	**Nº de mensagens**
material	77.8%	**49**
TIPO-DE-FAZER	N=49	
transformativo	98.0%	48
criativo	2.0%	1
TIPO-DE-IMPACTO	N=49	
intransitivo	0.0%	0
transitivo	100.0%	49
mental	4.8%	**3**
TIPO-DE-MENTAL	N=3	
superior	100.0%	3
inferior	0.0%	0
TIPO-DE-SUPERIOR	N=3	
cognitivo	100.0%	3
desiderativo	0.0%	0
TIPO-DE-INFERIOR	N=0	
perceptivo	0.0%	0
emotivo	0.0%	0
TIPO-DE-FENOMENALIZAÇÃO	N=3	
fenomenalização	100.0%	3
não fenomenalização	0.0%	0
verbal	3.2%	**2**
TIPO-DE-RECEPÇÃO	N=2	
não-recepção	100.0%	2
recepção	0.0%	0
TIPO-DE-VERBAL	N=2	
atividade	0.0%	0
semiose	100.0%	2
relacional	12.7%	**8**
TIPO-DE-RELAÇÃO	N=8	
intensivo	87.5%	7
posessivo	12.5%	1
circunstancial	0.0%	0
MODO-DE-RELAÇÃO	N=8	
atributivo	87.5%	7
identificativo	12.5%	1
existencial	1.6%	**1**

Tabela 12 – Resultados da análise Ideacional do Manual Panela de Pressão Solar em IT

(Conclusão)

CAP_04			
TIPO-DE-PROCESSO	**Frequência**	**Nº de mensagens**	
TIPO-DE-EXISTENCIAL		N=1	
criação	0.0%		0
permanência	100.0%		1
TIPO-DE-CRIAÇÃO		N=0	
surgimento	0.0%		0
extinção	0.0%		0
TIPO-PERMANÊNCIA		N=1	
positiva	100.0%		1
negativa	0.0%		0
EXISTÊNCIAL		N=1	
introdução	100.0%		1
representação	0.0%		0
TOTAL	**100%**		**63**

4.8 Análise das ondas de informação do manual Panela de Pressão Solar em IT

O manual em IT é composto por 18 ondas de informação que auxiliam o leitor a compreender a forma correta de usar a panela de pressão Solar da Tramontina. As ondas de informação são responsáveis por apresentar o manual, advertir o leitor, dar comandos para que o leitor use corretamente a panela, dar explicações para informar por que o leitor deve obedecer aos comandos e explicar questões envolvidas no uso e na manutenção da panela.

Para construir esses significados, esse manual utiliza as funções: (1) EXPLICAR, (2) COMANDAR, (3) CLASSIFICAR e (4) INTRODUZIR.

A função EXPLICAR (1) é realizada pelo Processo: Material (14 = Transitivo e Transformativo / 16 – Transitivo e Criativo) com o Modo: Declarativo (20) e com o Tema: *Default* (10) ou Tema Perspectiva (41). Ela também pode ser realizada pelo Processo: Mental (59 = Cognitivo e Fenomenalização) com o Modo: Declarativo (20) e com o Tema Perspectiva (41).

A função (2) COMANDAR é realizada pelo Modo: Jussivo (11). Ele pode realizar com o Processo: Material (14) cosselecionando com o Tema: *Default* ou Tema: Perspectiva, ou com o Processo: Mental (59 – Cognitivo e Fenomenalização) e com o Tema: Perspectiva (41), ou com o Processo Relacional (33 – Intensivo e Atributivo) cosselecionando com o Tema: *Default* (10) ou Tema: Perspectiva (41).

A função (3) CLASSIFICAR ocorre no manual para caracterizar participantes e para auxiliar nas explicações. Estas informações novas são realizadas por Processo: Relacional (39 - Processo: Relacional Identificativo e Intensivo 33 – Atributivo e Intensivo / 32 – Atributivo e Possessivo) que vem sempre acompanhado do Modo: Declarativo (20) e com o Tema: *Default* (10) ou Tema: Perspectiva (41).

A função (4) INTRODUZIR insere participantes novos e significados novos no texto e auxilia nas explicações, para isso, o Processo: Existencial: Permanência (29) é utilizado e o Processo: Verbal (41 – Semiose e Não Recepção). No caso do Processo: Existencial ele vem com o Modo: Declarativo (20) e com o Tema: *Default* (10) e o Processo: Verbal vem acompanhado do Modo: Declarativo (20) e com o Tema: *Default* (10).

5 Comparação linguística do perfil sistêmico dos manuais

Para a comparação linguística entre os manuais, Catford (1965, p. 27) foi utilizado quanto à comparação entre línguas e Krzeszowski (1990, p. 4) em relação à linguística contrastiva.

Ambos os autores apontam que as línguas podem ser comparadas à medida que elas possuam categorias em comum. Nesse sentido, para a análise do *corpus* desta pesquisa, foram usadas as categorias comuns da TRANSITIVIDADE, MODO e TEMA entre o inglês e português brasileiro. Sendo assim, os manuais são estruturalmente semelhantes e, portanto, comparáveis.

A seguir, apresentam-se as comparações entre os manuais.

5.1 Comparação OI e PT

Comparando os textos pelas categorias gramaticais, é possível apontar equivalência textual e correspondência formal segundo Catford (1965, p. 27) entre o inglês e o português brasileiro.

Para esta análise, o Quadro 2 a seguir apresenta o perfil dos manuais em relação às categorias gramaticais utilizadas na análise e a função semântica gerada pela confluência de TRANSITIVIDADE, MODO e TEMA:

Quadro 2 – Comparação do perfil sistêmico entre os manuais IO e PT

MANUAL INGLÊS ORIGINAL				MANUAL PORTUGUÊS TRADUZIDO			
89 Mensagens / 15 Ondas de Informação				82 Mensagens / 14 Ondas de Informação			
Cosseleções Gramaticais			Funções Semânticas	Cosseleções Gramaticais			Funções Semânticas
Trans.	Modo	Tema		Trans.	Modo	Tema	
14\16\11	20	10\41		14\16	20	10\41	
56\59	20	10\41	EXPLICAR	59	20	41	EXPLICAR
14\16\11	11	10\41		14\16	11	10\41	
59\54	11	10	COMANDAR	56\59\54	11	10	COMANDAR
39\37\33\32	20	10\41	CLASSIFICAR	37\33\32	20	10\41	CLASSIFICAR
14\11	15	10	CONVIDAR	14\11	15	10	CONVIDAR
29	20	10\41	INTRODUZIR	29	20	10\41	INTRODUZIR

Como mostra o Quadro 2, apesar de os manuais possuírem o mesmo número de *tokens*, o manual em IO traz 89 mensagens, ao passo que o manual em PT traz 82 mensagens. A mensagem 82 do manual em PT equivale à mensagem 80 do manual em IO. Mediante esse dado, pode-se afirmar que o manual em português precisa de mais *tokens* para 'significar' do que o manual em inglês e, consequentemente, mais mensagens.

Comparando pelo viés das ondas de informação, o manual em IO possui 15 ondas de informação e o manual em PT 14 ondas de informação. Pela função de cada uma, pode-se afirmar que existe equivalência textual e correspondência formal entre elas, pois a função semântica de uma onda na língua fonte foi traduzida para a mesma

função semântica na língua alvo e, consequentemente, as categorias semânticas foram mantidas. Apesar de as funções "Como usar as pastas." e "Como usar o *Launchpad*" terem sido trocadas de ordem no texto alvo, a equivalência e a correspondência se mantiveram.

O quadro também aponta que as funções semânticas EXPLICAR, COMANDAR, CLASSIFICAR, CONVIDAR e INTRODUZIR são correspondentes na língua fonte e na língua alvo, todavia, a equivalência ocorre parcialmente, pois existem classes utilizadas para formar as funções semânticas em uma língua que não ocorrem na outra língua. Essas classes estão marcadas em negrito no Quadro 2 e estão especificadas a seguir.

O manual em inglês possui mais opções para a cosseleção do que o manual em PT. Para a função EXPLICAR, o manual em IO tem a opção de cosselecionar o Processo 11[5] com o Modo 20 e o Tema 10/41; essa opção não existe no manual em PT. Para a mesma função, a opção de cosseleção com o Processo 56 com o Modo 20 e o Tema 10/41 não possui no PT, assim como a cosseleção do Processo 59 com o Modo 20 e o Tema 10.

Isso também ocorre na função COMANDAR, o manual em IO tem a opção de cosselecionar o Processo 11 com o Modo 20 e o Tema 10/41, opção não existente na tradução. No PT, a cosseleção com Processo 56 com o Modo 11 e o Tema 10 não ocorre no IO.

Na função CLASSIFICAR, existe a opção no manual em IO de cosselecionar o Processo 39 com o Modo 11 e o Tema 10/41, opção esta que não existe no manual traduzido.

As classes gramaticais da TRANSITIVIDADE, MODO e TEMA da língua fonte que foram mantidas na língua alvo apresentam a correspondência formal entre o inglês original e o português brasileiro traduzido no nível gramatical.

[5] Para saber a descrição gramatical do deslocamento no espaço gramatical, consultar análise das ondas de informação da seção anterior.

5.2 Comparação PO e IT

Comparando com base nas classes gramaticais analisadas e nas cosseleções possíveis entre elas, é possível apontar equivalência textual e correspondência formal entre original e tradução.

O Quadro 3 a seguir apresenta o perfil dos manuais com as possíveis categorias gramaticais e as possíveis confluências entre elas, como também a realização semântica de cada uma:

Quadro 3 – Comparação do perfil sistêmico entre os manuais PO e IT

MANUAL PORTUGUÊS ORIGINAL				MANUAL INGLÊS TRADUZIDO			
63 Mensagens / 18 Ondas de Informação				63 Mensagens / 18 Ondas de Informação			
Cosseleções Gramaticais			Funções Semânticas	Cosseleções Gramaticais			Funções Semânticas
Trans.	Modo	Tema		Trans.	Modo	Tema	
14\16	20	10\41		14\16	20	10\41	
59	20	41	EXPLICAR	59	20	41	EXPLICAR
14\16	11	10\41		14	11	10\41	
59	11	10\41	COMANDAR	59	11	41	
33\32	20	10\41	CLASSIFICAR	33	11	10\41	COMANDAR
29	20	10		39\33\32	20	10\41	CLASSIFICAR
41	20	10\41	INTRODUZIR	29	20	10	
				41	20	10	INTRODUZIR

O Quadro 3 aponta que tanto o manual em PO quanto o manual em IT possuem 63 mensagens, porém a 63ª mensagem do manual em PO equivale a 61ª mensagem do manual em IT. Dessa forma, como na comparação da seção anterior, o manual em português brasileiro precisa de mais *tokens* e, por conseguinte, mais mensagens para significar do que o manual em inglês. Além da mesma quantidade de mensagens, os dois manuais possuem 18 ondas de informação.

Comparando as ondas de informação dos manuais, a análise realizada na seção anterior mostra que existe equivalência textual e correspondência formal entre elas, pois a função semântica de uma onda na língua fonte foi traduzida para a mesma função semântica na língua alvo e, consequentemente, as categorias semânticas foram mantidas.

O Quadro 14 também apresenta as funções semânticas EXPLICAR, COMANDAR, CLASSIFICAR e INTRODUZIR do manual em PO que foram mantidas no IT, ocorrendo assim uma correspondência formal. No entanto, a equivalência ocorre parcialmente entre as funções semânticas, pois algumas classes gramaticais que foram utilizadas para formar a função semântica na língua fonte não foram utilizadas na língua alvo. Essas classes estão marcadas em negrito no Quadro 3 e estão especificadas a seguir.

Na função semântica COMANDAR do manual em PO, a opção de confluência entre o Processo 16 com o Modo 11 e o Tema 10/41 não existe no manual em IT. Também na mesma função semântica, a cosseleção entre o Tema 10, o Modo 11 e o Processo 59 no manual em PO não ocorre em IT.

Na função INTRODUZIR, o manual em PO possui a alternativa de confluência entre o Tema 41 com o Modo 20 e o Processo 41 que não ocorre no IT.

Comparando com base no no manual em IT, pode-se afirmar que ele, como na comparação da seção anterior, possui mais opções de cosseleção gramatical do que o manual em PO. Um exemplo disso ocorre na função COMANDAR do manual em IT, ele possui uma opção a mais de cosseleção do que o manual em PO. Esse manual em inglês usa o Processo 33 cosselecionando com o Modo 11 e o Tema 10/41 para dar ordens ao leitor.

Na função semântica CLASSIFICAR do mesmo manual, o Processo 39 cosseleciona com o Modo 20 e o Tema 10/4. Essa opção não existe no manual em PO.

As demais cosseleções gramaticais da língua fonte que foram mantidas na língua alvo apresentam a correspondência formal entre o português brasileiro original e o inglês traduzido no nível gramatical.

5.3 Comparação IO e PO

Pela análise realizada, os manuais em inglês original e em português original, podem ser comparadas em relação às categorias gramaticais utilizadas, as cosseleções entre os sistemas, as funções

semânticas e as ondas de informação, a fim de apresentar a distância e a proximidade entre as línguas.

O Quadro 4 mostra o perfil dos manuais com as categorias gramaticais utilizadas e as confluências entre elas, como também a realização semântica de cada uma.

Quadro 4 - Comparação do perfil sistêmico entre os manuais IO e PO

MANUAL INGLÊS ORIGINAL				MANUAL PORTUGUÊS ORIGINAL			
89 Mensagens / 15 Ondas de Informação				63 Mensagens / 18 Ondas de Informação			
Cosseleções Gramaticais			Funções Semânticas	Cosseleções Gramaticais			Funções Semânticas
Trans.	Modo	Tema		Trans.	Modo	Tema	
14\16**11**	20	10\41		14\16	20	10\41	
56\59	20	**10**\41	EXPLICAR	59	20	41	EXPLICAR
14\16**11**	11	10\41		14\16	11	10\41	
59**54**	11	10	COMANDAR	59	11	10**41**	COMANDAR
39**37**\33\32	20	10\41	CLASSIFICAR	33\32	20	10\41	CLASSIFICAR
14\11	**15**	**10**	**CONVIDAR**	29	20	10	
29	20	10**41**	INTRODUZIR	**41**	**20**	**10\41**	INTRODUZIR

O Quadro 4 mostra que o manual em IO possui 89 mensagens, ao passo que o manual em PO possui 63 mensagens. Assim sendo, o manual do *MacBook Pro* precisa de mais mensagens para significar do que o manual da Panela de Pressão Solar.

Todavia, apesar de possuir mais mensagens, o manual em IO possui menos ondas de informação do que o manual em PO. Enquanto o primeiro apresenta 15 ondas; o segundo possui 18.

Apesar da quantidade de ondas de informação ser diferente e apesar dos manuais tratarem de objetos diferentes, eles apresentam duas funções para as ondas de informação que são semelhantes: a função "Apresentação", que é responsável por introduzir o manual, e a função "Primeiros passos para usar o *MacBook Pro*" e "Primeiros passos para usar a Panela de Pressão Solar", que explica o que é necessário saber para começar a usar o computador ou a panela de pressão.

O Quadro 4 aponta também as funções semânticas do manual em cada língua e suas respectivas cosseleções entre as classes gramaticais da TRANSITIVIDADE, MODO e TEMA.

Comparando os manuais a partir do manual em IO, a função CONVIDAR constituída a partir da confluência entre o Processo 14/11 com o Modo15 e o Tema 10 do manual em IO não ocorre no manual em PO.

Na função EXPLICAR, a confluência entre o Processo 11 com o Modo 20 e o Tema 10/41 não ocorre em português brasileiro. Também nessa mesma função, a confluência entre o Processo 56 com o Modo 20 e o Tema10/41 não ocorre no manual em PO. Com relação ao Tema 10, a confluência entre ele e o Modo20 e Processo 59 não existe em português brasileiro.

Na função COMANDAR, Processo 11 cosseleciona com o Modo 11 e com o Tema 10/41. Essa opção não existe em PO. Nessa mesma função semântica, o Processo 54 cosseleciona com o Modo 11 e com o Tema 10/41. Essa opção não existe em PO.

Na função CLASSIFICAR, o manual em IO possui a opção de cosselecionar com o Modo 20 e com o Tema 10/41, o Processo 39/37. Essas opções não ocorrem no manual em PO.

Com relação à função INTRODUZIR, a confluência entre Tema 41, o Modo 20 e o Processo 29 não ocorre no manual em PO.

Comparando com base no manual em PO, na função COMANDAR, não existe no IO a cosseleção do Tema 41 com o Modo 11 e o Processo 59.

Na função INTRODUZIR, a opção composta pelo Processo 41 confluindo com o Modo 20 e o Tema 10/41 não existe no manual em IO.

Com relação às demais cosseleções gramaticais que não foram mencionadas anteriormente, o Quadro 5 a seguir apresenta as suas ocorrências com suas respectivas funções semânticas.

Quadro 5 – Significados multilíngues do manual de instrução em inglês e português brasileiro

Cosseleções Gramaticais			Funções Semânticas
Trans.	Modo	Tema	
14\16	20	10\41	
59	20	41	EXPLICAR
14\16	11	10\41	
59	11	10	COMANDAR
33\32	20	10\41	CLASSIFICAR
29	20	10	INTRODUZIR

O Quadro 5 aponta as categorias gramaticais e as funções semânticas comuns entre os manuais em inglês e em português brasileiro original. Nesse sentido, todas as classes apresentadas nesse quadro representam a correspondência formal entre as duas línguas para o manual de instrução, ou os seus significados multilíngues (MATTHIESSEN *et al*., 2008).

6 Perfilação gramatical sistêmica do *corpus* combinado

Com base nos resultados das frequências relativas a cada categoria gramatical analisada no *corpus* combinado deste trabalho e por meio da análise das funções semânticas e as ondas de informação, é possível apresentar um perfil gramatical sistêmico que constitui um modelo da variação linguística multilíngue para os quatro manuais de instrução.

O modelo não é nenhum texto específico, criado por usuários do português brasileiro. Ele é, por outro lado, uma generalização das probabilidades maiores de escolhas dos sistemas gramaticais para o *corpus* de manuais. Assim, o perfil a seguir não se refere especificamente a qualquer texto em particular, mas propõe uma previsão para o comportamento da gramática oracional do português brasileiro para os textos caracterizados como manual de instrução.

O Quadro 6 apresenta o perfil sistêmico para o *corpus* combinado no par linguístico inglês e português brasileiro deste artigo.

Quadro 6 – Perfilação sistêmica do *corpus* combinado de manuais de instrução

(Continua)

PERFIL SISTÊMICO DO *CORPUS* COMBINADO									
Número da oração	Trans. (x)	Modo (y)	Tema (z)	Ondas de infor-mação	Número da oração	Trans. (x)	Modo (y)	Tema (z)	Ondas de infor-mação
FRANK_01	14	20	10	1ª	FRANK_01	EXPLICAR			1ª
FRANK_02	14	20	10		FRANK_02	EXPLICAR			
FRANK_03	33	20	41		FRANK_03	CLASSIFICAR			
FRANK_04	14	11	10		FRANK_04	COMANDAR			
FRANK_05	14	20	10		FRANK_05	EXPLICAR			
FRANK_06	14	20	41	2ª	FRANK_06	EXPLICAR			2ª
FRANK_07	32	20	10		FRANK_07	CLASSIFICAR			
FRANK_08	14	20	10		FRANK_08	EXPLICAR			
FRANK_09	59	20	41		FRANK_09	EXPLICAR			
FRANK_10	14	20	10		FRANK_10	EXPLICAR			
FRANK_11	14	11	10		FRANK_11	COMANDAR			
FRANK_12	14	11	10		FRANK_12	COMANDAR			
FRANK_13	59	11	41	3ª	FRANK_13	COMANDAR			3ª
FRANK_14	41	20	10		FRANK_14	INTRODUZIR			
FRANK_15	14	11	10		FRANK_15	COMANDAR			
FRANK_16	14	11	10		FRANK_16	COMANDAR			
FRANK_17	14	20	10		FRANK_17	EXPLICAR			
FRANK_18	33	20	41		FRANK_18	CLASSIFICAR			
FRANK_19	14	20	10		FRANK_19	EXPLICAR			
FRANK_20	32	20	41		FRANK_20	CLASSIFICAR			
FRANK_21	14	20	41		FRANK_21	EXPLICAR			
FRANK_22	16	20	10	4ª	FRANK_22	EXPLICAR			4ª
FRANK_23	29	20	10		FRANK_23	INTRODUZIR			
FRANK_24	14	11	10		FRANK_24	COMANDAR			
FRANK_25	14	11	10		FRANK_25	COMANDAR			
FRANK_26	14	11	10	5ª	FRANK_26	COMANDAR			5ª
FRANK_27	14	11	10		FRANK_27	COMANDAR			
FRANK_28	14	11	41		FRANK_28	COMANDAR			
FRANK_29	14	11	10		FRANK_29	COMANDAR			
FRANK_30	11	11	10		FRANK_30	COMANDAR			
FRANK_31	14	11	10		FRANK_31	COMANDAR			
FRANK_32	14	11	41		FRANK_32	COMANDAR			
FRANK_33	14	11	41		FRANK_33	COMANDAR			
FRANK_34	14	11	10		FRANK_34	COMANDAR			
FRANK_35	14	11	41		FRANK_35	EXPLICAR			
FRANK_36	16	20	41	6ª	FRANK_36	COMANDAR			6ª
FRANK_37	14	20	41		FRANK_37	EXPLICAR			

Quadro 6 – Perfilação sistêmica do *corpus* combinado de manuais de instrução

(Conclusão)

PERFIL SISTÊMICO DO *CORPUS* COMBINADO									
Número da oração	Trans. (x)	Modo (y)	Tema (z)	Ondas de informação	Número da oração	Trans. (x)	Modo (y)	Tema (z)	Ondas de informação
FRANK_38	59	11	10	7ª	FRANK_38	COMANDAR			7ª
FRANK_39	37	20	10		FRANK_39	CLASSIFICAR			
FRANK_40	16	20	10		FRANK_40	EXPLICAR			
FRANK_41	37	20	10		FRANK_41	CLASSIFICAR			
FRANK_42	14	11	10		FRANK_42	COMANDAR			
FRANK_43	14	20	41		FRANK_43	EXPLICAR			
FRANK_44	14	20	10	8ª	FRANK_44	EXPLICAR			8ª
FRANK_45	33	20	10		FRANK_45	CLASSIFICAR			
FRANK_46	59	11	10	9ª	FRANK_46	COMANDAR			9ª
FRANK_47	54	11	10	10ª	FRANK_47	COMANDAR			10ª
FRANK_48	14	11	41		FRANK_48	COMANDAR			
FRANK_49	14	20	10		FRANK_49	EXPLICAR			
FRANK_50	14	20	10		FRANK_50	EXPLICAR			
FRANK_51	14	11	10	11ª	FRANK_51	COMANDAR			11ª
FRANK_52	14	11	41		FRANK_52	COMANDAR			
FRANK_53	14	11	10		FRANK_53	COMANDAR			
FRANK_54	14	11	10		FRANK_54	COMANDAR			
FRANK_55	14	11	41	12ª	FRANK_55	COMANDAR			12ª
FRANK_56	14	11	10		FRANK_56	COMANDAR			
FRANK_57	14	11	10		FRANK_57	COMANDAR			
FRANK_58	14	11	10		FRANK_58	COMANDAR			
FRANK_59	29	20	10	13ª	FRANK_59	INTRODUZIR			13ª
FRANK_60	33	11	41	14ª	FRANK_60	COMANDAR			14ª
FRANK_61	14	11	41		FRANK_61	COMANDAR			
FRANK_62	14	20	10	15ª	FRANK_62	EXPLICAR			15ª
FRANK_63	14	11	41		FRANK_63	COMANDAR			
FRANK_64	14	11	10		FRANK_64	COMANDAR			
FRANK_65	14	11	10		FRANK_65	COMANDAR			
FRANK_66	14	20	41		FRANK_66	EXPLICAR			
FRANK_67	14	20	10	16ª	FRANK_67	EXPLICAR			16ª
FRANK_68	32	20	10		FRANK_68	CLASSIFICAR			
FRANK_69	14	11	10		FRANK_69	COMANDAR			
FRANK_70	16	11	10		FRANK_70	COMANDAR			
FRANK_71	59	20	10		FRANK_71	EXPLICAR			
FRANK_72	14	20	10		FRANK_72	EXPLICAR			
FRANK_73	14	20	10		FRANK_73	EXPLICAR			
FRANK_74	14	20	10		FRANK_74	EXPLICAR			

Esse modelo, também é chamado de Frankenstein por representar o que é mais frequente dos quatro manuais de instrução; é composto por 74 mensagens e 16 ondas de informação. A primeira seção desse quadro aponta as cosseleções gramaticais de cada mensagem; a segunda, mostra

o significado semântico resultante da confluência das classes da TRANSITIVIDADE, MODO e TEMA.

O Frankenstein não é considerado um manual, e sim um modelo criado para compreender como são realizadas pelo falante as escolhas gramaticais e, consequentemente, semânticas (HALLIDAY *et al.*, 1964), a fim de construir o significado do registro: manual de instrução. Para exemplificar o Frankenstein, o Quadro 7 mostra a sua primeira onda de informação, com exemplos do *corpus* combinado. Os exemplos seguem a descrição de cada oração exibida no Quadro 6.

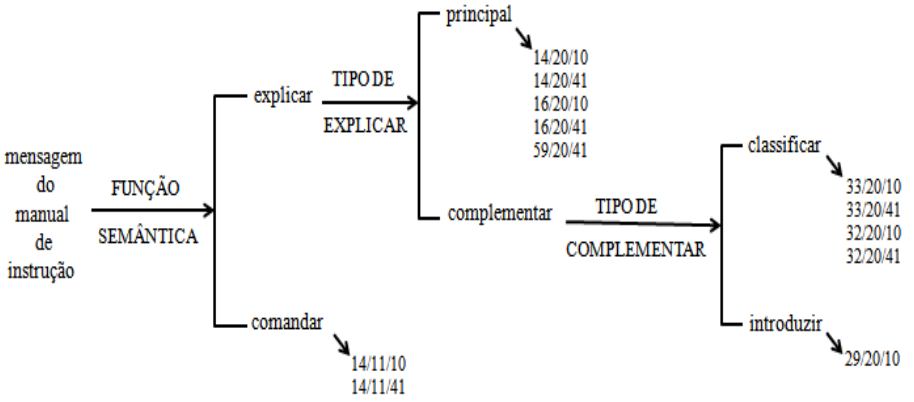
Quadro 7- O Frankenstein com exemplos do *corpus*

ORAÇÃO	EXEMPLOS	TEXTO FONTE
FRANK_01	Você adquiriu a panela de pressão solar Tramontina	PO
FRANK_02	This guide shows you what's on your Mac	IO
FRANK_03	para que eles estejam prontos quando você utilizá-los pela primeira vez.	PT
FRANK_04	Clean the pressure regulator valve (8)	IT
FRANK_05	Este dispositivo impede a abertura da tampa enquanto houver pressão no interior da panela.	PO

O Quadro 7 mostra o Frankenstein criado com partes distintas dos manuais do *corpus*. Apesar de ele não apresentar os mesmos participantes nas orações e de não haver conectividade entre um termo e outro do ponto de vista lexical, esse modelo está gramaticalmente correto, no sentido de que ele representa ideacionalmente, encena interpessoalmente e constrói textualmente a primeira onda de informação do *corpus*, que trata principalmente da identificação do objeto abordado pelo manual. Ademais, faz isso com base nas probabilidades maiores de escolha de funções dos sistemas envolvidos para cada oração.

Por conseguinte, a comparação realizada neste trabalho também apresenta o ambiente multilíngue comum dos quatro textos no par linguístico inglês / português brasileiro com suas respectivas traduções. A Figura 2 mostra o sistema multilíngue da realização semântica pautada em escolhas gramaticais do *corpus* combinado de manuais.

Figura 2 – Sistema Multilíngue da realização semântica pautada em escolhas gramaticais do *corpus* combinado de manuais



Fonte: a autora.

A Figura 2 representa a construção de significados partilhados entre as mensagens do *corpus*. Dessa forma, ela apresenta o que é o manual de instrução no ambiente multilíngue.

Nesse sistema, a condição de entrada é a mensagem do manual de instrução. No primeiro nível de delicadeza do sistema, a mensagem do manual pode apresentar a função semântica 'explicar' ou 'comandar'. Em relação ao TIPO DE EXPLICAR, ele pode ser 'principal' ou 'complementar'. Caso a seleção seja por 'complementar', apresentam-se mais duas classes: 'classificar' ou 'introduzir'. As setas que seguem nas classes apontam as opções gramaticais de cada uma.

7 Conclusões

Tendo como foco principal apresentar como o arcabouço da Linguística de *Corpus* com a Linguística Sistêmico-Funcional permite, pela perfilação gramatical, analisar, comparar e encontrar padrões linguísticos tanto para o registro quanto para a tradução técnica, este artigo demonstrou os padrões gramaticais e semânticos de um *corpus* combinado: manual de instrução.

Mediante a análise do *corpus* e os padrões encontrados (BIBER, 2010), os resultados apontaram que o manual de instrução é uma macro-onda de significado que capacita o falante a desenvolver uma atividade. Mais especificamente, esse registro tem por objetivo instruir o falante e não regular o seu comportamento. Para tanto, o manual utiliza quatro funções semânticas para significar, a saber, a função EXPLICAR, que é composta pelas funções CLASSIFICAR e INTRODUZIR, e a função COMANDAR. Nesse sentido, o manual é uma sequência de explicações intercaladas com comandos em que o objeto do manual e o leitor são negociados, a fim de construir a macro função capacitar.

Os resultados das análises apontaram semelhanças e diferenças entre as línguas, como a função semântica CONVIDAR e sua realização gramatical que ocorreu somente no manual em inglês original, ou a cosseleção na função INTRODUZIR com o Processo: Verbal que ocorreu somente no manual em português original.

Com relação à tradução técnica, cumpre destacar que os manuais traduzidos do *corpus* apresentaram construções gramaticais e semânticas do texto fonte, do texto original da língua alvo e construções que não apareceram nesses dois textos, por exemplo, a cosseleção de 33/11/10 ou 41 na função COMANDAR do manual em inglês traduzido que não ocorre no manual em inglês original. Portanto, a tradução no *corpus* de manuais é um texto híbrido e multilíngue que, para significar, utiliza-se do pareamento do texto fonte, da língua alvo e de novos significados.

Ao final, o perfil gramatical sistêmico do *corpus* (o Frankenstein) apresentou a confluência de significados dos manuais desta pesquisa, mostrando, assim, o que é o manual com todos os significados mais frequentes desse ambiente multilíngue.

Por fim, cabe ressaltar que, para pesquisas futuras, este trabalho contribui como um método de pesquisa que amplia a análise de *corpus* (VIANA, 2011), utilizando a perfilação sistêmica e a comparação linguística. Quanto ao produto final desta pesquisa, o modelo da variação linguística contribui com o estudo dos padrões linguísticos do registro: manual de instrução (BIBER, 2010). E, com relação à tradução, os procedimentos aqui apresentados mostram uma possibilidade de analisar traduções técnicas por meio de um *corpus* combinado (STEINER, 2001; NUNES, 2013), ampliando, assim, os estudos sobre tradução.

Referências

AZENHA, J. Tradução técnica, condicionantes culturais e os limites da responsabilidade do tradutor. In: CONGRESSO BRASILEIRO DE LINGÜÍSTICA APLICADA, 4, 1995, Campinas. *Anais*... UNICAMP, Campinas, SP, 1995. p. 137-149.

APPLE – Macbook Pro retina 13 inch. Disponível em: <http://manuals.info.apple.com/MANUALS/1000/MA1663/en_US/macbook_pro_retina-13-inch-late-2013_qs.pdf>. Acesso em: 22 jan. 2014.

BERBER SARDINHA, T. Linguística de corpus: histórico e problemática. *D.E.L.T.A*., v. 16, n. 2, p. 323-367, 2000.

BERBER SARDINHA, T. *Corpora eletrônicos na pesquisa em tradução.* São Paulo: LAEL, PUC/SP, 2002. p. 15-59.

BIBER, D. M. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing*, Oxford, v. 5, n. 4, p. 257-269, 1990. DOI: <http://dx.doi.org/10.1093/llc/5.4.257>

_____. What can a corpus tell us about registers and genres? In: O'KEEFFE, A.; McCARTHY, M. (Ed.). *The Routledge Handbook of corpus Linguistics*. London, New York: Routledge, 2010. p. 241-254.

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

CATFORD, J. *A linguistic theory of translation*: an essay in applied linguistics. London: Oxford University, 1965.

FIGUEREDO, G. *Introdução ao perfil metafuncional do português brasileiro*: contribuições para os estudos multilíngues. 2011. 385 f. Tese (Doutorado em Linguística Aplicada) – Programa de Pós-Graduação em Estudos Linguísticos, Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte, 2011.

FIGUEREDO, G. Uma metodologia de perfilação gramatical sistêmica baseada em corpus. *Letras & Letras*, Uberlândia, v. 30, n. 2, p. 17-45, jul./dec. 2014. DOI: <http://dx.doi.org/10.14393/LL60-v30n2a2014-2>

FIGUEREDO, G. Um estudo do conjunto multilíngue interpessoal português brasileiro / inglês subsidiado pelos estudos da tradução e pela linguística sistêmico-functional. *Cad. Trad*., Florianópolis, v. 35, n. 1, p. 139-166, 2015.

HALLIDAY, M. A. K. *Language as social semiotic: the social interpretation of language and meaning.* London & Baltimore: Edward Arnold& University Park Press, 1978.

_____. The construction of knowledge and value in the grammar of scientific discourse, with reference to Charles Darwing's the origin of species. In: COULTHARD, M. (Ed.). *Advances in written text analysis.* London and New York: Routledge, 1994.

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

_____. Meaning as choice. In: FONTAINE, L.; BARTLETT, T.; O'GRADY, G. (Ed.). *Systemic Functional Linguistics*: exploring choice. New York: Cambridge University Press, 2013. p. 15-36.
DOI: <http://dx.doi.org/10.1017/CBO9781139583077>

HALLIDAY, M. A. K.; MCINTOSH, A.; STREVENS, P. *The linguist sciences and language teaching.* London: Longman, 1964.

HALLIDAY, M. A. K.; MATTHIESSEN, M. I. M. *Introduction to Functional Grammar*. London and New York: Routledge, 2014.

KENNING, M. 2010. What are parallel and comparable corpora and how can we use them? In: O'KEEFFE, A.; McCARTHY, M. (Ed.). *The Routledge Handbook of corpus Linguistics*. London, New York: Routledge, 2010. p. 487-500.
DOI: <http://dx.doi.org/10.4324/9780203856949.ch35>

KRZESZOWSKI, T. *Contrasting languages.* The scope of contrastive linguistics. Berlin: De Gruyter, 1990.
DOI: <http://dx.doi.org/10.1515/9783110860146>

LEMKE, J. Discourse, Dynamics, and Social Change. *Cultural Dynamics*, v. 6, n. 1, p. 243-275, 1993.

DOI: <http://dx.doi.org/10.1177/092137409300600107>.

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

MARTIN, J. R.; ROSE, David. *Working with discourse*: meaning beyond the clause. 2 ed. London: Continuum, 2007.

MATTHIESSEN, C. The environments of translation. In: STEINER, E. YALLOP, C. (Ed.). *Exploring translation and multilingual text production beyond content*. Berlin, New York: Mouton de Gruyer, 2001. p. 41-124. DOI: <http://dx.doi.org/10.1515/9783110866193.41>

MATTHIESSEN, C.; TERUYA, K.; WU, C. Multilingual studies as a multidimensional space of interconnected language studies. In: WEBSTER, J. (Ed.). *Meaning in Context*: implementing intelligent applications of language studies. London, New York: Continuum, 2008. p. 146-220.

NUNES, L. P. *Relações coesivas e estruturais em corpus combinado*: uma análise de conjunções em textos originais e traduzidos. Belo Horizonte: Universidade Federal de Minas Gerais, 2013. (Trabalho de qualificação para o doutorado). (Mimeo)

O'DONNELL, M. The UAM CorpusTool: software for corpus annotation and exploration. In: BRETONES CALLEJAS, C. M. *et al*. (Ed.). *Applied linguistics now*: understanding language and mind. Almería: Universidad de Almería, 2008. p. 1433-1447.

SINCLAIR, J. Corpus evidence in language description. In: WICHMANN, A. *et al*. (Ed.). *Teaching and language corpora*. London: Longman, 1997. p. 27-39.

STEINER, E. Translations English – German: investigating the relative importance of systemic contrasts and of the text-type "translation". S*PRIKreports*, n. 7, p. 1-49, Oct. 2001. Disponível em: <http://hf.ulo.no/ilos/forkning/prosjekter/sprik/docs/pdf/steiner.pdf>.

TRAMONTINA. Tramontina Solar. Disponível em: <http://suipnovo.tecnologia.ws/public/upload/product/62516223/62516222FLM001.pdf>. Acesso em: 22 jan. 2014.

VIANA, V. Linguística de *corpus*: conceitos, técnicas e análises. In: VIANA, V.; TAGNIN, S. (Org.). *Corpora no ensino de línguas estrangeiras.* São Paulo: HUB Editorial, 2011. p. 25-95.

URE, J. *Text types classified by situational factors*. Atlanta: MS, 1989.

DOI: 10.17851/2237-2083.23.3.769-796

The relevance of the Sketch Engine software to build Field – Football Expressions Dictionary

A relevância da ferramenta Sketch Engine para a construção do Field – dicionário de expressões do futebol

Rove Luiza de Oliveira Chishman
Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, RS, Brasil
rove@unisinos.br

Aline Nardes dos Santos
Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, RS, Brasil
aline.nardes@gmail.com

Diego Spader de Souza
Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, RS, Brasil
dspadersouza@gmail.com

João Gabriel Padilha
Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, RS, Brasil
joaogabrielsl@hotmail.com

Abstract: This paper aims at presenting the role of the Sketch Engine concordancer in the development of Field – Football Expressions Dictionary, a trilingual lexicographic resource based on the notion of frame and on linguistic corpora. Here, some considerations are presented concerning not only the resources available in the software, but also the analyses conducted through it. In the second section, Frame Semantics, a theory developed by Fillmore (1982, 1985) and its lexicographic applications are presented; in the third section, the Corpus Linguistics methodological potential is introduced by approaching some of its main principles, and some features of the study corpus. The fourth section describes the analyses procedures to identify (i) polysemic words and (ii) collocations in our corpus, through Sketch Engine. The results indicate that the software can be used for various lexicography-related purposes, once it enables

users not only to visualize in detail the entire productivity of language, especially through the Word Sketch option, but also to define the lexicography policy supported by corpus evidence.

Keywords: football; Frame Semantics; Corpus Linguistics.

Resumo: O objetivo deste trabalho é apresentar o papel do concordanceador *Sketch Engine* no desenvolvimento do *Field* – Dicionário de Expressões do Futebol, um recurso lexicográfico trilíngue baseado na noção de *Frame* Semântico e em *corpora* linguísticos. Aqui são apresentadas algumas considerações concernentes não apenas aos recursos disponíveis na ferramenta mas também às análises realizadas por meio desse programa. Na segunda seção, a Semântica de *Frames*, teoria proposta por Fillmore (1982, 1985), é apresentada, bem como suas aplicações lexicográficas; na terceira seção, o potencial metodológico da Linguística de *Corpus* é apresentado por meio de alguns de seus princípios basilares, além de se descreverem algumas características dos *corpora* de estudo do projeto. O quarto segmento descreve os procedimentos de análise, visando a identificar (i) as unidades polissêmicas e (ii) as colocações provenientes do *corpus*, por meio do *Sketch Engine*. Os resultados evidenciam que o programa permite aos usuários visualizar detalhadamente toda a produtividade da língua, principalmente por meio do recurso *Word Sketch*.

Palavras-chave: futebol; Semântica de Frames; Linguística de *Corpus*.

Recebido em: 31 de julho de 2015.
Aprovado em: 2 de outubro de 2015.

1 Introduction

Field – Football expressions dictionary – is a trilingual lexicographic resource organized around the notion of semantic frame (FILLMORE, 1982). Its building process involved, amongst other stages, the compilation of three comparable corpora,[1] which were processed

[1] It is important to highlight that this paper focuses only in Portuguese corpora analyses; therefore, the translation stage is not considered in this study.

using the Sketch Engine tool (KILGARRIF, 2004). Considering the precepts of Corpus Linguistics adopted in this project as a methodological resource, the tool had a fundamental role in the process of identifying frames and treating linguistic phenomena such as polysemy and collocations.

Thus, this paper aims at presenting the functionalities of Sketch Engine in what concerns the development of Field, showing its potential to the process of organizing frames through corpus attestations, as well as treating polysemic units and collocations. In order to illustrate the course of our analysis and demonstrate the results achieved, we approach the use of three Sketch Engine resources: Concordance, Word Sketch, and Collocations.

Having this in mind, we organize this paper in four sections: in the second section, we present the Cognitive Linguistics research program, which includes Frame Semantics and its applications. In the third section, we discuss some important aspects concerning Corpus Linguistics, the Sketch Engine tool and the corpora compiled for this investigation. In the third section, we present the analyses in two distinct moments, focusing on two phenomena: (a) polysemy and (b) collocations. The last section of this paper presents final remarks and perspectives for future researches.

2 Theoretical framework

Founded in the second-half of the 1970's, Cognitive Linguistics started as a movement of rupture between a group of linguists and the Chomskyan approach to the study of language. Lakoff (1987), Langacker (1987; 2006), Talmy (1987), Fillmore (1982; 1985) and Fauconnier (1985), among others, are part of this group. The dissatisfaction these researchers with Chomsky's model resided in the minor role semantics and pragmatics were playing along different views towards cognition and its relationship with language and meaning. It is important to point out that the Chomskyan paradigm is, in its fashion, cognitive. The fundamental difference in relation to the cognitivism that emerged in the late seventies is that this notion is broadened by the inclusion of notions

from cognitive psychology, such as the Prototype Theory (ROSCH, 1973; 1975a; 1975b).

Cognitive Linguistics defends that semantics arises from encyclopedic knowledge; meaning is built through our experiences while living and discovering the world around us. Assuming that meaning is encyclopedic, Cognitive Linguistics adopts a usage-based perspective. Such affirmative tells us that within this paradigm, the sharp distinction that separates semantics and pragmatics is no longer accepted: that is to say it is not possible to define strictly what belongs to language, and what belongs to the usage of language. To Fauconnier (2003, p. 1-2),

> Cognitive linguistics recognizes that the study of language is the study of language use and that, when we engage in any language activity, we draw unconsciously on vast cognitive and cultural resources, call up models and frames, set up multiple connections, coordinate large arrays of information, and engage in creative mappings, transfers, and elaborations. Language does not 'represent' meaning; it prompts for the construction of meaning in particular contexts with particular cultural models and cognitive resources.

According to Langacker (1987), the distinction between semantics and pragmatics is quite artificial, and to think of a "viable" approach to semantics means to consider one that rejects false dichotomies – a semantic model that presents encyclopedic nature, for that matter. Frame Semantics is considered to be an innovative theory of this point of view within the realm of Cognitive Linguistics.

2.1 Frame Semantics

Frame Semantics is a theory developed by Charles J. Fillmore (1982; 1985) today considered one of the alternatives Cognitive Linguistics presents to the study of meaning. To Miriam Petruck (2001, p. 1), Frame Semantics can be described as "[…] a research program in empirical semantics which emphasizes the continuities between language

and experience […]". In this sense, this theory seeks to investigate the relations between the senses of words of a given language and the experiences speakers go through in life. In other words, Frame Semantics understands that our knowledge of a language depends on the way we perceive the world. In what concerns Frame Semantics, "[…] the meaning dimension is expressed in terms of the cognitive structures (frames) that shape speakers' understanding of linguistic expressions." (FILLMORE; BAKER, 2010, p. 317).

A frame, therefore, can be described as an *experience schematization* (EVANS; GREEN, 2006) that structures the information we have regarding the elements that compose a given scene or situation. Fillmore (1982, p. 111) considers the frame to be "[…] any system of concepts related in such a way that to understand any one of them you have to understand the whole structure in which it fits […]". When a concept is introduced in a text or in a conversation, all concepts related to it are automatically activated, or, in Fillmore's terms, *evoked*. We can think of the word "waiter", for example: according to Fillmore's concept of frame, to comprehend the concept of "waiter" is to comprehend the whole scene in which this particular concept is inserted, including other related words, such as "menu", "check", "client" etc. One of the purposes of the research undertaken by Frame Semantics refers to investigate the reasons why a community relates a given category to a word. Those reasons are to be included in the description of the meaning of such a word (PETRUCK, 2001). From this perspective, we are able to notice that Fillmore's theory does not share the views of more formal approaches to Semantics when it considers the role of factors such as experience / encyclopedic knowledge in the process of conceptualizing meaning.

Fillmore proposes the concept of frame as a way to cover a set of notions already in existence in the literature, which influenced his work – for instance, schema (BARTLETT, 1932), and script (SCHANK; ABELSON, 1977). It is therefore evident that the idea that underlies the semantic frames suggested by Fillmore was already present in linguistics, though under different denominations. It is worth mentioning that even the word "frame" was already in use by other researchers in different areas of knowledge: Minsky (1974) and Goffman (1975), in Artificial

Intelligence and Sociology respectively. To Minsky, when we are confronted by a new situation, we access a mental structure that provides us information concerning how to act and what to do in the situation – this structure is the frame. Goffman sees the concept through a very similar way, stating that frames are structures used to depict the various ways we behave in different situations. Through these observations, we are able to say that Fillmore's frames are very close to Minsky's and Goffman's.

However, according to Petruck (2001), the concept used in Frame Semantics is influenced by the notion of case frame, used by Fillmore in his Case Grammar, developed during the decade of 1960 – in 1968, more precisely. In the article "The Case for Case", the linguist stablishes the notion of cases, which are similar to semantic roles: in a sentence, the verb selects a determined number of cases, forming a set that Fillmore refers to as case frame.

Another important aspect of Frame Semantics worth mentioning concerns the nature of the frame as responsible for the depiction of a conventional, typical situation. Such characteristic leads us to talk briefly about the concept of prototype. Fillmore (1982) treats this notion using an example with the word *breakfast*, showing us that the understanding of this word is inseparably related to the cultural habit of having a meal during the first part of the day, after a period of sleep, whose menu is made up of a given kind of food and so on (FILLMORE, 1982). It is noticeable, however, that such statement motivates speakers to use the concept of breakfast in unusual situations: one can sleep until noon, eat a toast, drink some juice, and call this a breakfast. At the same time, one can wake up at seven in the morning, eat pizza and call this breakfast as well.

2.3 Frame Semantics and the development of lexicographic resources

Because of the possibility of organizing words around the situations in which they occur, Frame Semantics has been used as a basis for lexicographic resources. In this respect, it is essential to mention FrameNet, a pioneer project initiated by Fillmore and a group of linguists and lexicographers that presents semantic information about the lexicon

of the English language based on frames. According to Ruppenhofer *et al*. (2010, p. 5), The Berkeley FrameNet project is "[…] an on-line lexical resource for English, based on Frame Semantics and supported by corpus evidence". FrameNet works essentially with the documentation of the syntactic and semantic possibilities of lexical combination in English, taking into account all the senses words may present (RUPPENHOFER *et al*., 2010): its database has over 10 thousand lexical units. A lexical unit, according to Ruppenhofer *et al*. (2010), is the pairing of a given word and a meaning.[2] In the FrameNet context, it means that, in order to be a lexical unit, the word must be connected to a frame. In cases of polysemy, when a word is related to more than one meaning (thus related to more than one frame), each sense possibility is a lexical unit.

Besides FrameNet, the Kicktionary project (SCHMIDT, 2009) is also worth mentioning. As well as FrameNet, it is a database whose purpose is to describe the language of football in English, French and German, using corpora evidence. Kicktionary groups the football frames in scenes that organize them according to their similarities and differences. Is it important to emphasize, however, that these resources were designed having in mind a specialized user that comprehends the notions concerning Fillmore's theory. In virtue of this aspect, during the compilation of Field, a dictionary for the general public, we discarded the exhibition of some linguistic information in favor of a friendly interface, even though all data associated with semantic frames were mapped and analyzed. For the same reason, Field does not present semantically annotated examples,[3] preferring to occult information related to the elements of each frame.

The next section approaches the structure of Field Dictionary.

2.3.1 Field: a user-friendly, frame-based dictionary of football expressions

Field is a trilingual dictionary of football language guided by the notion of scenarios, which are an adaptation of Fillmore's concept of

[2] This concept is firstly proposed by Cruse (1986).

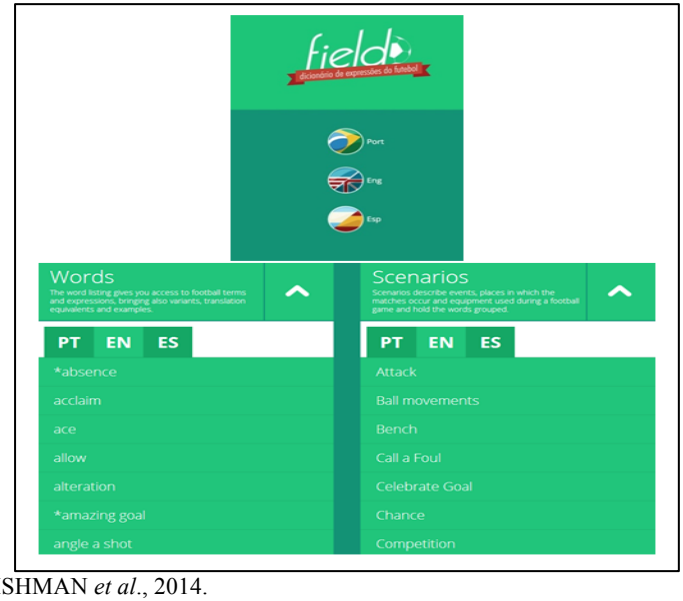
[3] In FrameNet context, semantic annotation means the "assignment of semantic role tags to syntactic constituents." (FILLMORE; PETRUCK, 2003, p. 359).

frame. The main goal of Field creators is not only to offer a word list, but also to organize words regarding their context of use, grouping them according to the situation in which they appear. Field has around 600 word entries, and approximately 40 scenarios; the website has optimized versions for tablets and smartphones.

In the context of Field, frames are called scenarios and are used to structure and organize the entries: football events such as SHOT, INFRACTION and GOAL are scenarios around which words and expressions are organized. This way, if users search for the word *bicicleta* (bicycle kick), they have access to information regarding the corresponding scenario (SHOT), apart from the translation to English and Spanish. By clicking on this scenario, they will find other related words, such as *bomba* (screamer) and *bico* (toe poke).

When accessing the dictionary for the first time, users can choose the desired language, which will also be applied to the lists of entries and scenarios. Field dictionary has two listings: one for words and one for scenarios.

Figure 1 – Field interface: select language screen and lists of words / scenarios

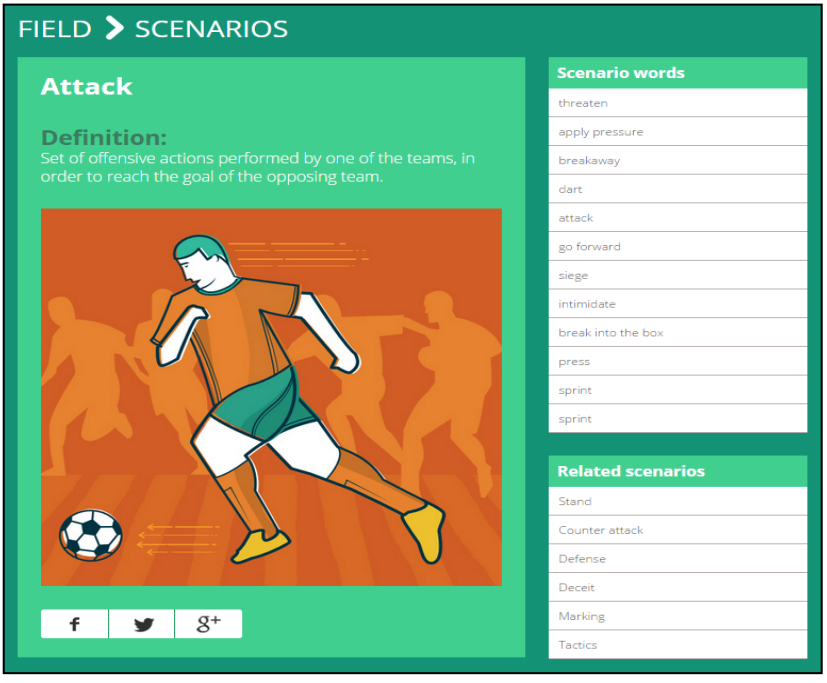


Source: CHISHMAN *et al*., 2014.

When searching for a word entry in Portuguese, users have access to information such as audio, word class and variants (including spelling differences and words with similar meaning), as well as translation to English and Spanish. The work of translation considered three situations: (i) cases of direct equivalence in the target language, (ii) cases of partial equivalence, and (iii) cases with no translation equivalent. Entries that illustrate the second situation (partial or inaccurate equivalence) bring a possible translation or a suggestion, and are marked with an asterisk. Entries to which no equivalence was found are identified with the abbreviation NT and have an explanatory note. This space is also used to explain the use of words that do not have direct equivalents of translation in one of the languages.

Concerning scenario entries, they represent events, match places or equipment used during the game, displaying all the words belonging to them. It is provided a definition and an illustration for each scenario.

Figure 2 – an example of frame / scenario



Source: CHISHMAN *et al*., 2014.

In the next section, we present the methodology adopted in the present investigation.

3 The methodologic potential of Corpus Linguistics: the concordance Sketch Engine and the corpora of the Field Dictionary

Corpus Linguistics is a branch in Linguistics that emerged in the second half of the twentieth century (BERBER SARDINHA, 2000). Having an empiricist character, its main objective is studying language through real-usage samples extracted from texts that exist in the world, that is, that were not invented with the purpose of illustrating a linguistic phenomenon. The word *corpus* relates to a "set of textual linguistic data collected properly with the purpose of serving the research of a given language or linguistic variety" (BERBER SARDINHA, 2000, p. 325). This set of data can be handled through computational tools that allow the researcher to deal with significant quantities of data – an example of such tools is the Sketch Engine concordancer.

Therefore, since the arising of personal computers, in 1980, to investigate phenomena such as polysemy and collocations from corpora means organizing corpora in accordance with the purposes of the investigation, as well as to process the data electronically through specific software. However, it is important to consider that electronic storage of corpora had already been created by linguists since the 1950's – despite the limited technologies they had access to –, due to the fact that American structuralists used to collect and analyze real speech data. In accordance to McCarthy and O'Keefe (2010), between the 1980's and the 1990's, Corpus Linguistics took the shape it has today, thanks to the development of computational resources. Nonetheless, is interesting to note that predecessors such as Sinclair *et al*. (1987), the developers of the COBUILD corpus, had to use punched cards technology when it was already considered outdated. The reason why they had to do it was the lack of user-friendly tools for linguists at the time; available resources could only be explored by computer experts (MCCARTHY; O'KEEFE, 2010). This scenario has completely changed since the development of software for linguists such as *Wordsmith* (SCOTT, 1996), *Sketch Engine* (KILGARRIF, 2004), *Unitex* (PAUMIER, 2002) and *AntConc* (ANTHONY, 2014) – these two are free, it is worth to mention.

A corpus linguistics research can be either corpus-driven, via the sole application of statistical measures (GRANGER; PAQUOT, 2008), or corpus-based – as in this study –, when investigations combine corpus data with other theoretical premises, in order to test their assumptions or even to improve them. In addition, Corpus Linguistics as a methodology "[…] does not go to the extreme of rejecting intuition while attaching importance to empirical data." (MCENERY *et al*. *apud* EVISON, 2010, p. 132). Therefore, within the context of Field, the usage of corpora was combined with Frame Semantics premises, as well as with the decisions that constituted what Atkins and Rundell (2008) call Style Guide.[4] This approach is also called "linguistics-with-corpus methodology"[5] (SANTOS, 2008, p. 52).

Regarding the building of our corpora, considering there was not a football language corpus in Brazilian Portuguese available, it was necessary to compile our own. The texts that form the Field Dictionary corpus were collected from the official websites of Brazilian football teams, from news websites and from Twitter profiles. The texts that interested our purposes were the ones that described how the matches had occurred. Known in the sports domain as *match report*, this genre resumes the main moments in a match: who scored the goals, how the goals were scored, which team won, what were the main moments of the match etc. These features ensured that we could verify how the football frames are organized by the occurrences of typical evocators as *shot* and *goal*. Among the texts available on the sites, there were some that did not fit our research, once they do not evoke football frames, and were, thus, ignored – their content was based on institutional news of the clubs, like the facilities, the promotions, the products related to the teams etc.

By following these features, our corpus can be considered representative of the football events in brazilian Portuguese, concerning our research purposes. Our collection of texts can be considered *big*, once it totalizes one million words to each language – Portugese, English and Spanish. Size of corpora is not a consensus amongst the authors, so it is important to say that we consider ours a big corpus based on the sum proposed by Berber Sardinha (2000). Once collected, the texts were

[4] A Style Guide is composed by all the instructions that guide the building of a dictionary (ATKINS; RUNDELL, 2008).

[5] In Portuguese: "metodologia da linguística com *corpora*".

converted to the .txt format (UTF-8), then processed by the *parser* PALAVRAS (which labelled the texts with morpho-syntactic information) and, finally, processed by the software Maestro, a pre-requisite for uploading the corpus to the Sketch Engine tool.

The Sketch Engine is a tool that allows the creation, the manipulation and the study of corpora. Through the search options, the user is taken to the *concordances*, which consist of lines based on fraction of texts in which the queried word or expression (the so-called *node word*), appear highlighted, as well as its co-texts (portions of texts that surround the node word). Other useful resources present in Sketch Engine are *Word Sketch* – which presents schematically the syntactic realizations of the lexical items in the corpus – and *Collocations*, where the user can visualize the words that tend to occur together in the corpus. These functionalities are shown in more detail in the next chapter.

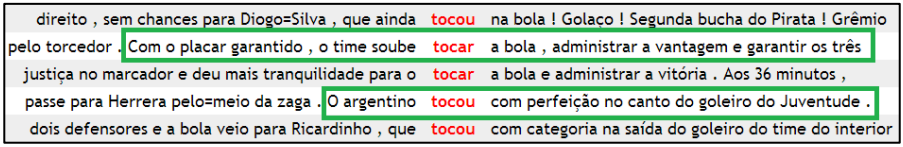
4 Analysis and results

4.1 The identification of polysemic lexical units using Sketch Engine

The Sketch Engine tool made it possible to identify polysemic lexical units in the corpus. In general lines, polysemy is the phenomenon whereby a linguistic form can have more than one related meanings, and it is a fundamental feature of human language (ULLMANN, 1961). It is important to mention that the considerations concerning polysemy here presented are part of a masters dissertation (PADILHA, 2015) written in the context of FIELD – Football Expressions Dictionary. Its main objective was to verify how polysemy presents itself in the football language. One of the main findings in this paper is that the most frequent polysemous items analysed represent *complex categories* radially shaped, confirming Langacker's (1987) and Lakoff's (1987) hypothesis, respectively. That is to say: the polysemous senses of words are stored in cognition in respect to a *prototypical sense*, the first one that comes to the mind of speakers when they hear or think of a word, through which all the other related senses are generated, in this case, by extension, as we intend to show.

Through a simple query in the *concordance* menu, we realized that the verb *tocar*, for example, presents polysemic behaviour, judging by the co-text that surrounds the realizations of this verb:

Figure 3 – KWIC-list of the verb *tocar* (part 1)[6]

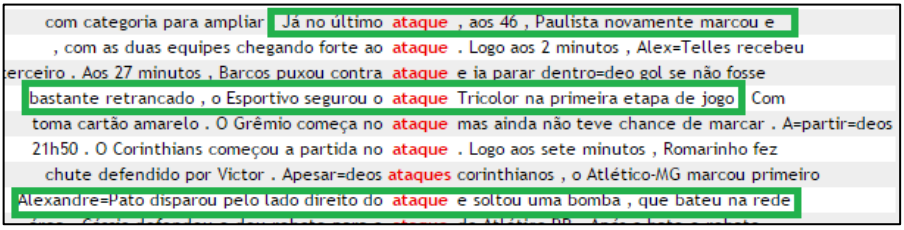


Source: KILGARRIF *et al*., 2004.

Right above, highlighted in green, it is possible to see two senses related to *tocar*: in [...] *o time soube tocar a bola* [...], the frame Pass is evoked, once it concerns the transfer of the ball possession between two players from the same team. In the second case, [...] *o argentino tocou com perfeição no canto do goleiro do Juventude*, Shot is the frame evoked by *tocar*: in this sense, this verb conceptualizes the action in which a player kicks the ball against the adversary team's goal.

Another case of polysemy that we find interesting to mention relates to the noun *ataque*, for which there were accounted three relates senses:

Figure 4 – KWIC-list of the verb *tocar* (part 2)[7]



Source: KILGARRIF *et al*., 2004.

---

[6] Free translation of the highlighted examples: 1) Possessing a fair scoreboard, the team could *pass* the ball around, hold the lead and guarantee the three…; 2) The Argentinian *placed* the ball perfectly into Juventude's goalkeeper corner.

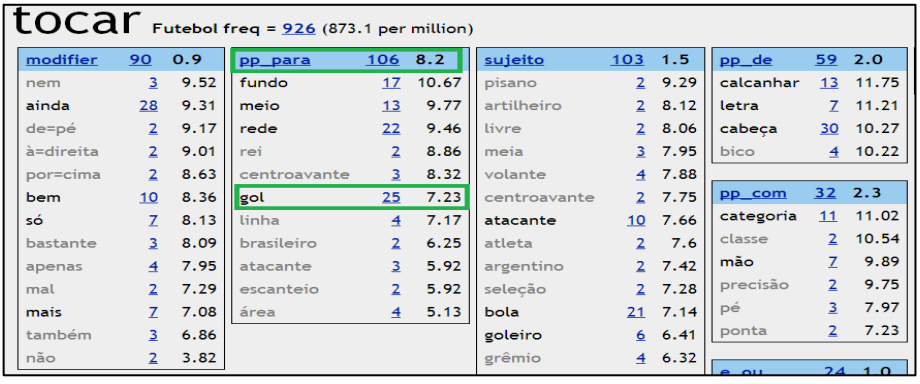
[7] Free translation of the highlighted examples:
1) In the last *attack move*, at the 46th minute, Paulista scored another goal and…;
2) …very defensive, Esportivo held back the Tricolor's attack in the first half of the match;
3) Alexandre Pato bolted to the right wing of the *attack* and fired a screamer, which hit the net.

The first highlighted sentence shows the sense of *ataque* that conceptualizes the *action of attacking the adversary team in order to score a goal*, evoking the Attack frame. The second sense related to *ataque* observed through the *concordance* query conceptualizes not the action, but the *group of players responsible for attacking the other team*. This comes to the fore by the excerpt *ataque tricolor*, which conceptualizes these players from Grêmio, referred to like that because it is one of the brazilian clubs whose uniforms have three colors. In this sense, *ataque* evokes the Actors frame. The third and last sense recorded for *ataque* relates to the *place in which the players responsible for scoring goals act.* This sense evokes the Field frame.

Though the *concordance* query allowed us to identify the senses of polysemous items, as we tried to show above, it would take us an endless amount of time to analyse all the instances for the senses of *tocar* and *ataque*. It would be so due to two basic reasons: the first one is that the tool shows *all* the sentences grouped, in a way that the user cannot separate one sense from the others; the second reason is that the sentences are not shown completely, just part of them (even considering that the tool allows the user to extend the co-texts to the left and right sides, it would still take a considerable amount of time from the user). Considering these issues, we relied on another Sketch Engine resource: the *Word Sketch*.

The *Word Sketch* option provides a more detailed search by allowing the user to access all the syntactic realizations of the queried word, as shown in the box below:

Figure 5 – Word Sketch for the verb *tocar* (part 1)



Source: KILGARRIF *et al*., 2004.

In the upper image, we have the sketch obtained for the verb *tocar*. Within each box it is possible to see a syntactic relation and its occurrences in the corpus. Highlighted in green, we can observe the realization of *tocar* when it is complemented by the PP *para*, totalizing 106 occurrences, of which 25 involve the noun *goal* as an object – *tocar para o gol*, sense that evokes the Shot frame, as we said before. Other senses related to *tocar* emerge through the objects *bola* and *mão*:

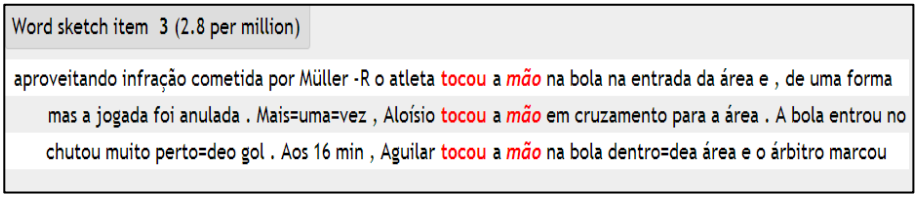
Figure 6 – Word Sketch for the verb *tocar* (part 2)

objeto	100	1.1
bola	75	8.98
mão	3	8.31
10	2	8.09
travessão	3	7.92
trave	2	6.4
camisa	2	6.36

Source: KILGARRIF *et al*., 2004.

Tocar a bola can relate not only to a pass, but also to a shot, as we have stated. On the other hand, the object *mão* involves another sense, and, consequently, another frame. By clicking on the number right beside the word, we have access to the concordances that show this sense:

Figure 7 – KWIC-list of the verb *tocar* (part 3)



Word sketch item 3 (2.8 per million)

aproveitando infração cometida por Müller -R o atleta tocou a *mão* na bola na entrada da área e , de uma forma

mas a jogada foi anulada . Mais=uma=vez , Aloísio tocou a *mão* em cruzamento para a área . A bola entrou no

chutou muito perto=deo gol . Aos 16 min , Aguilar tocou a *mão* na bola dentro=dea área e o árbitro marcou

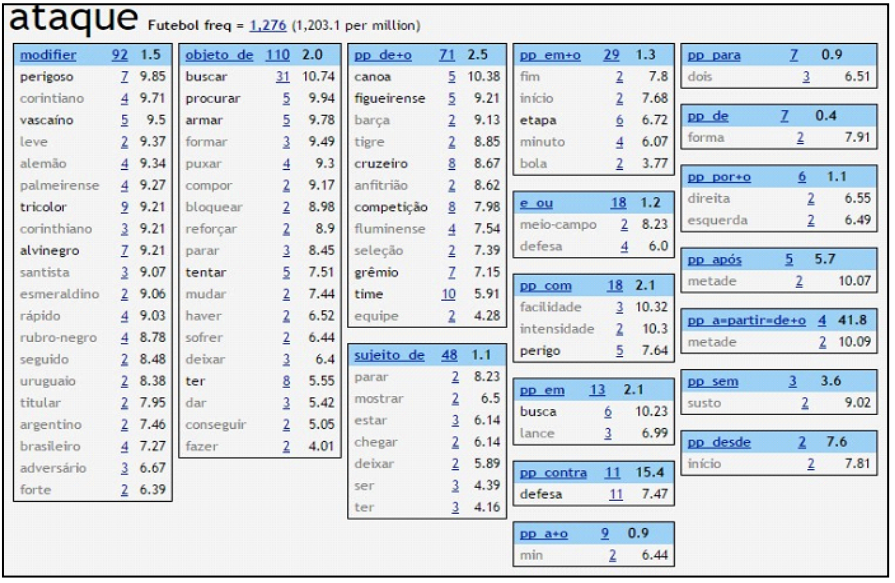
Source: KILGARRIF *et al*., 2004.

By checking the three occurrences in which *mão* occurs as a complement for *tocar*, we could observe that there is a third sense for this verb: this sense of *tocar* conceptualizes the moment a player touches the ball with his hand, which is an infraction, unless this player is the

goalkeeper. The frame evoked by the third sense, thus, is the Infraction one.

Right below, the word sketch for *ataque* is shown:

Figure 8 – Word Sketch for the noun *ataque* (part 1)



Source: KILGARRIF *et al*., 2004.

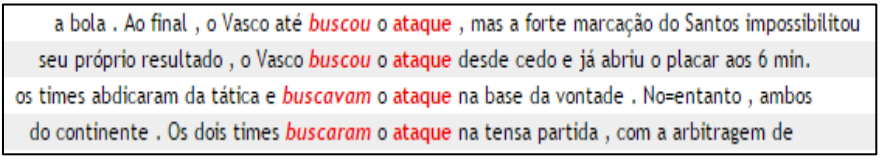
The relation *objeto de* (object_of) shows us all verbs that take *ataque* as an object in our corpus and their occurrence – 110 times. The verb *buscar* (search for) is the more prominent in this list, followed by *procurar* (look for) and *armar* (set) as the image below shows:

Figure 9 – Word Sketch for the noun *ataque* (part 2)

Source: KILGARRIF *et al*., 2004.

Appearing as an object of *buscar* 31 times, the high occurrence of this pattern showed us the first sense of *ataque*, according to our corpus: *action of attacking the adversary team*, as the concordances below show:

Figure 10 – KWIC-list of the verb *tocar* (part 3)



Source: KILGARRIF *et al*., 2004.

In all the concordance lines in the image, *ataque* conceptualizes an *action,* it is possible to observe: in the first two lines, this action is performed by a team – *Vasco* – that tried to score a goal in the match, as expected. In the following lines, the subjects of *buscar* are *os times* (the teams) and *os dois times* (the two teams), respectively, conceptualizing the act of attacking performed by the players responsible for this function in the game. This sense occurred 124 times in our corpus, being the most recurrent one. The second sense related to *ataque*, *group of players that perform the action of attacking*, according to our analysis, appears 52 times, being followed by the third sense – *part of the field* – that was counted 25 times in our data.

It is quite significant to mention that, even though the *Word Sketch* filter shows *all* the combinations for the nouns and verbs queried, the analysis must rely on the researcher proficiency when identifying the senses registered for such a language. To put it simply, the tool presents the combinations and the exemplars of such combinations, but it is the researcher that judges what counts or not as a sense, once the sense is not in the form, as the data reveal.

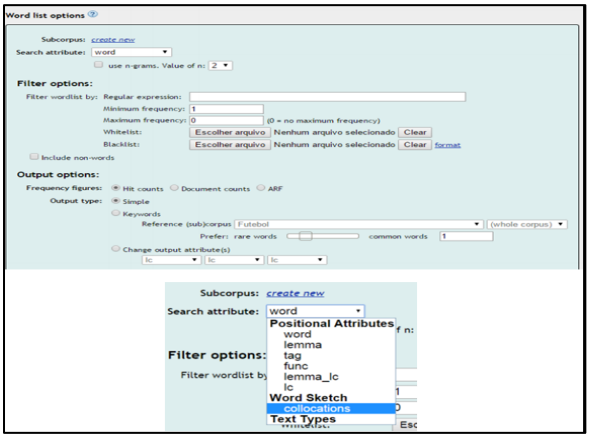
In the next section, we present the study of the collocations in the football language through the Sketch Engine concordancer.

4.2 *The* Collocations *tool and the identification and treatment of collocations*

Sketch Engine was also useful when it comes to the identification of collocations – groupings of words that usually appear together, such as *close friend* and *key question* (TAGNIN, 2013)[8] – in our *corpus*. It is important to say that this section is based on the results of a master's degree dissertation (SOUZA, 2015) developed in the context of *Field: Football Expressions Dictionary*. The objective of the study was to evaluate the role of the FrameNet methodology in the lexicographic treatment of collocations in the aforementioned dictionary. The hypothesis is that the FrameNet concept of *lexical unit* (a pairing between a linguist form and a meaning / frame, as we have previously said) allows collocations to be in the main list of headwords of the dictionary. Therefore, this section provides information regarding the methodological steps taken in the thesis dissertation and reports some of its results.

First, it is fair to start by addressing some characteristics of the *Collocations* tool of Sketch Engine, used in the study. Under the *Word list options* menu (FIGURE 9), *Collocations* allows the search to generate a list of all lexical combinations in a corpus, organizing the results based on a frequency criterion.

Figure 11 – The *Word list options* menu and *Collocations* filter



Source: KILGARRIF *et al*., 2004.

[8] For more information concerning collocations, cf. Tagnin (2013) and Hausmann (1989).

By clicking the options in *Search attribute*, the resource *Collocations* appears available. It is important to emphasize that in *Frequency figures* (below *Output options*), we maintain the option *Hit counts* selected, once we want the software to show how many times each one of the collocations appear in the corpus. For this particular study, we used the Brazilian Portuguese corpus *Futebol*, compiled for *Field*.

The search for collocations resulted in a vast list of combinations, and not all of them were structures we could consider football collocations. For the ongoing research, we selected the 500 most frequent results for a manual analysis, which consisted in eliminating the structures that did not have place in the study (non-collocations such as *que não* [that don't] for example). After this first scrutiny of the data provided by Sketch Engine, 74 collocations of football language were identified. The first highlight we were able to see in the data corresponds to the very high incidence of verbal structures: out of 74 collocations, 44 have a verb as one of its components, therefore meaning 59,4% of this subcorpus. Beforehand, we can consider this aspect a reflection of the language, as football is a dynamic sport, made of actions and events.

From this final list, in this section we present an overall quantitative and qualitative analysis of six verbal collocations related to the event of scoring a goal and three concerning the event of passing the ball.

The most frequent collocation in the corpus fits in this category: *abrir (o) placar* [open (the) score] with 784 occurrences. See the example below:

(1) Com o jogo bastante concentrado no meio-campo, o Grêmio TIME conseguiu **abrir o placar** aos 31 minutos TEMPO.
With the game highly centered in the midfield, Grêmio TEAM *was able to* ***open the score*** *at 31 minutes* TIME.

In the sentence above, we can see elements such as the *team* that scored a goal and *time*, fitting the collocation within the Score Goal frame in *Field*.

The next two collocations are *marcar gol* [score goal] and *fazer gol* [make goal] with 441 and 401 occurrences respectively. One

interesting semantic aspect of these structures is that they can work as substitutes of one another, as we can see in the examples:

(2) Juanfran JOGADOR **marcou gol** e deixou tudo igual no estádio El Madrigal.
Juanfran PLAYER ***scored a goal*** *and left it all the same in the El Madrigal stadium.*

(3) Zagueiro JOGADOR volta e **faz gol**.
The defender PLAYER *comes back and* ***makes a goal****.*

Having identified the element of a player who scores, these collocations can also integrate the Score Goal frame.

It is also interesting to notice the presence of the collocation *marcar pênalti* [sign penalty] in our data. Unlike what happens in the previous structures, in which the verbs refer to an action performed by a player or team, the verb here is related to the referee, who defines the penalty for a given offense made by some of the players in the game to the other team. See the example:

(4) O árbitro ÁRBITRO **marcou pênalti** e o veterano Blanco JOGADOR converteu a cobrança com categoria, ampliando o marcador PLACAR.
The referee REFEREE ***signaled penalty*** *and the veteran Blanco PLAYER converted the charging with category, expanding the marker MARKER.*

Differently from the other collocations, which are all part of the Score Goal frame, this particular structure, even though related to the same big event, belongs to the Referee Decisions frame as it is, in fact, depicting one of the actions done by referees in the match.

The next collocation, *balançar a rede* [swing the net], whose frequency is 245 occurrences, presents a change of focus, a change in the perspective through which the moment of goal is perceived. See the examples below:

(5) Quando Vagner Love JOGADOR **balançou a rede** do Palmeiras, neste domingo (18.11), em Volta Redonda, não foram só os torcedores do Flamengo que vibraram muito.

When Vagner Love PLAYER ***swung*** *Palmeiras'* ***net****, this Sunday (18.11), in Volta Redonda, it was not only Flamengo's crowd that celebrated.*

It seems that while the previous collocations focused more on the ball and the player, the collocation *balançar a rede* focuses on the result of the action in which the ball takes part, performed by the player towards the net. We highlight the fact that collocations need to be analyzed taking into consideration the category it is expressing. In a football match, what else could swing the net if not the ball? Maybe something else, like an object or even a player. However, even though that is possible, this particular collocation would not be used as it is related to a very specific event: scoring a goal.

The last collocation that conceptualizes the moment of goal is *sofrer gol* [concede goal], with 90 cases in the corpus. We notice that this collocation is on the same level of *marcar gol* and *fazer gol*, but referring to the opposing team's point of view, as we are able to see in (6):

(6) Mas Antunes JOGADOR não conseguiu acabar com a sina de **sofrer gols** logo no início.

But Antunes PLAYER *failed to end the fate of* ***conceding goals*** *early on.*

Another case in which verbal collocations designate two contrary points of view of a given action happens with the following structures: *dar (o) passe* [give pass], *fazer (o) passe* [make pass] e *receber (o) passe* [receive pass], with 118, 53 and 198 occurrences respectively. The first two collocations conceptualize the perspective of the player who has the ball and is passing it on to another player from their team. The third, however, depicts the point of view of the player to whom the ball is passed. The interesting feature about these collocations is how the frame elements are organized in the sentences, considering that different perspectives will possibly conceive the arrangement of such elements under different ways. See the examples from the corpus:

(7) Ivanildo JOGADOR QUE PASSA desceu pela meia direita, **fez o passe** para Magique JOGADOR QUE RECEBE, que apareceu na cara do goleiro e tocou para o fundo das redes.

*Ivanildo PASSER went down by the right wing, **made the pass** to Magique RECEIVER, who appeared in the front of the goalkeeper and kicked to the net.*

(8) Leandro JOGADOR QUE PASSA **faz o passe** para o centroavante JOGADOR QUE RECEBE.

*Leandro PASSER **does the pass** to the center-forward RECEIVER.*

(9) O jovem JOGADOR QUE PASSA fez um gol e **deu passe** para outro JOGADOR QUE RECEBE.

*The young PASSER scored a goal and **gave a pass** to another RECEIVER.*

(10) Ganso JOGADOR QUE PASSA **deu passes** precisos que criaram boas chances.

*Ganso PASSER **gave** precise **passes** that created good chances.*

(11) Com 25 minutos, Edenílson JOGADOR QUE RECEBE **recebeu passe**, invadiu a área e quando se preparava para tentar fazer o gol, foi derrubado por Élton.

*With 25 minutes, Edenível RECEIVER **received the pass**, invaded the area and while he was preparing to attempt a goal, was knocked down by Élton.*

(12) Primeiro, ele JOGADOR QUE RECEBE **recebeu passe** de Dadá JOGADOR QUE PASSA e tocou sem chances para o camisa 1 do Coxa.

First, he RECEIVER received the pass from Dadá PASSER and shot with no chances to Coxa's no. 1.

The first feature we can see through these examples is that, even though a pass situation will always involve at least two players, not always both of them will appear in the sentence, as we are able to notice in (10) and (11). Another interesting aspect also concerns the example

(10): in many cases in which we have the plural form *passes*, it is followed by an adjective. The data indicates that the singular form is often used to describe a single moment, a specific pass situation, while the plural is more directed to the *performance* the player had along the play, his style and competence. It is an evaluation not only of the pass or passes, but of the player as well.

5 Final Remarks

The present paper shows how the Sketch Engine tool contributed to the development of Field. It is highly valid to emphasize the treatment of the corpora through this software, considering not only the size of these collections, but also the possibility that it offers to explore the linguistic properties of these texts. The tool makes it possible not only to identify the most frequent words in the corpora, but also the way these words relate to each other. In addition, this study demonstrates the fundamental role of Corpus Linguistics in this corpus-based approach, providing a precise methodology for verifying empirically polysemic units and collocations.

One of the main advantages of using Sketch Engine is that its functionalities work in an integrated form, offering different levels of analysis: users can either perform a basic, fast query through the concordance option, or a more detailed one through Word Sketch. What is worth emphasizing is that, in our understanding, these resources do not superpose each other, but work together, once the user is always taken to the concordances in the end, as we demonstrated above.

Beyond its validity concerning the studies of polysemy and collocations, we emphasize the relevance of the Sketch Engine concordancer to the frame construction, considering that the *Word Sketches* show the syntactic-semantic behaviour of the frame evokers through empiric evidences.

Moreover, we would like to stress that, if on the one hand Sketch Engine provides several kinds and quantity of information, allowing different levels of analysis, on the other hand, it also poses a challenge to users, once they have to deal with large amount of data, not considering the task of compiling a corpus of their own – an activity that demands a considerable knowledge not only of the tool, but also of programing.

Finally, the results indicate that the software can be used for various lexicography-related purposes, once it enables users not only to visualize in detail the entire productivity of language, especially through the Word Sketch option, but also to define the lexicography policy supported by corpus evidence, as we intended to demonstrate.

To the next stage of the research, we foresee the construction of new corpora, having in mind the continuity of Field and also the creation of a dictionary devoted to the Olympic games.

References

ANTHONY, L. *AntConc (Version 3.4.3)* [Computer Software]. Tokyo, Japan: Waseda University, 2014. Retrieved on: September 15th, 2015, from: <http://www.laurenceanthony.net/>.

ATKINS, S.; RUNDELL, M. *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press, 2008.

BARTLETT, F. Remembering: A study in Experimental and Social Psychology. Cambridge: Cambridge University Press, 1932.

BERBER SARDINHA, T. *Linguística de Corpus: histórico e problemática. D.E.L.T.A.*, [S. l.], v. 16, n. 2, p. 323-367, 2000. Retrieved on: March 4th, 2015, from: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502000000200005&lng=en&nrm=iso>.

CHISHMAN, R. L. O. *et al*. *Field – Dicionário de Expressões do Futebol*. São Leopoldo: Unisinos, 2014. Retrieved on: Apr. 25th, 2015, from: < http://dicionariofield.com.br/>.

CRUSE, D. A. *Lexical Semantics*. New York: Cambridge University Press, 1986.

EVANS, V.; GREEN, M. *Cognitive linguistics*: an introduction. Edinburgh: Edinburgh University Press, 2006.

EVISON, J. What are the basics of analyzing corpus? In: MCCARTHY, M.; O'KEEFE, A. (Ed.) *The Routledge handbook of Corpus Linguistics*. London/New York: Routledge, 2010.
DOI: <http://dx.doi.org/10.4324/9780203856949.ch10>.

FAUCONNIER, G. *Mental Spaces*. Cambridge: MIT Press, 1985.

_____. Cognitive Linguistics. In: NADEL, L. *Encyclopedia of Cognitive Science*. London: Macmillan, 2003.

FILLMORE, C. Frame Semantics. *Linguistics in the Morning Calm*. Ed.: The Linguistic Society of Korea, Seoul: Hansinh Publishing Co., p.111-137, 1982.

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

_____.; BAKER, C. A frames approach to semantic analysis. In: HEINE, B.; NARROG, H. (Ed.). *The Oxford Handbook of Linguistic Analysis*. New York: Oxford University Press, 2010. p. 313-339.

FILLMORE, C. J.; PETRUCK, M. R. L. FrameNet Glossary. *International Journal of Lexicography*, Oxford, v.16, n.3, p. 359-361, 2003. Retrieved on: June 20th, 2015, from: <http://ijl.oxfordjournals.org/content/16/3/359.full.pdf>.

GOFFMAN, E. *Frame Analysis*: An Essay on the Organization of Experience. Cambridge, MA (U.S.): Harvard University Press, 1975.

GRANGER, S.; PAQUOT, M. Disentangling the phraseological web. In: GRANGER, S.; MEUNIER, F. (Ed.). *Phraseology*: an interdisciplinary perspective. Amsterdam: John Benjamins, 2008. p. 27-49. DOI: <http://dx.doi.org/10.1075/z.139.07gra>.

HAUSMANN, F. J. Le dictionnaire de collocations. In: HAUSMANN, F.J.; REICHMANN, O.; WIEGAND, H. E.; ZGUSTA, L. (Org.). *Wörterbücher, Dictionaries, Dictionnaires*. Ein Internationales Handbuch zur Lexikographie, v. 1. Berlin: Walter de Grujter, 1989, p. 1010-1019.

KILGARRIFF, A. *et al*. *The Sketch Engine*. Lorient: Euralex, 2004. Retrieved on: 28 jun. 2014, from: <http://www.sketchengine.co.uk/>.

LAKOFF, G. *Women, fire, and dangerous things*: what categories reveal about mind. Chicago: The University of Chicago, 1987.
DOI: <http://dx.doi.org/10.7208/chicago/9780226471013.001.0001>.

LANGACKER, R. *Foundations of Cognitive Grammar*. V. 1. Stanford: Stanford University Press, 1987.

_____. Cognitive Grammar. Introduction to Concept, Image, and Symbol. In: GEERAERTS, D. (Ed.) *Cognitive Linguistics*: basic readings. Berlin/New York: Mouton de Gruyter, 2006.

LYONS, L. *Semantics*. V. 2. Cambridge: Cambridge University Press, 1977.

MCCARTHY; M.; O'KEEFE, A. Historical perspective: what are corpora and how have they evolved? In: MCCARTHY, M.; O'KEEFE, A. (Ed.) *The Routledge handbook of Corpus Linguistics*. London/New York: Routledge, 2010.

MCENERY, A.; XIAO, R; TONO, Y. *Corpus-based Language Studies*. London: Routledge, 2006.

MINSKY, M. A framework for representing knowledge. *Artificial Intelligence Memo*, n. 306, Cambridge: Massachusetts Institute of Technology, 1974.

PADILHA, J. G. M. *A polissemia na linguagem do futebol*: uma proposta de aproximação entre redes lexicais e frames semânticos. 2015. 134 f. Dissertação (Mestrado em Linguística Aplicada) – Programa de Pós-Graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos (UNISINOS), São Leopoldo, 2015. Disponível em: <http://www.repositorio.jesuita.org.br/handle/UNISINOS/3774>. Acesso em: 25 abr. 2015.

PAUMIER, S. *Manuel d'utilisation du logiciel Unitex*. IGM, Université de Marne-la Vallée, 2002. Retrieved on: September 15th, 2015, from: <http://www-igm.univ-mlv.fr/~unitex/manuelunitex.pdf>.

PETRUCK, M. R. L. *Frame semantics*. Berkeley: University of California, 2001.

RUPPENHOFER, J. *et al*. *FrameNet II*: Extended Theory and Practice. Berkeley, California: International Computer Science Institute, 2010.

ROSCH, E. Natural categories. *Cognitive Psychology*, n. 4, p. 328-350, 1973. DOI: <http://dx.doi.org/10.1016/0010-0285(73)90017-0>.

_____. Cognitive Reference Points. *Cognitive Psychology*, n. 7, p. 532-547, 1975a.

_____. Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology*: General, v. 104, n. 3, p. 192-233, 1975b. DOI: <http://dx.doi.org/10.1037/0096-3445.104.3.192>.

SANTOS, D**.** Corporizando algumas questões. In: TAGNIN, S. E. O.; VALE, O. A. (Org.). *Avanços da linguística de corpus do Brasil.* São Paulo: Humanitas, 2008.

SCHANK, R. C.; ABELSON, R. P. *Scripts, plans, goals, and understanding*: An inquiry into human knowledge structures. Hillsdale, NJ: Lawrence Erlbaum, 1977.

SCHMIDT, T. The Kicktionary: a multilingual lexical resource of football language. In: BOAS, H. C. (Ed.) *Multilingual FrameNets in computational lexicography*: Methods and applications. Berlin/New York: Mouton de Gruyter, 2009, p.102-132.

SCOTT, M. *Wordsmith Tools.* Oxford: Oxford University Press, 1996.

SINCLAIR, J. M. *et al*. *Cobuild English Dictionary*. London/Birmingham: Collins Cobuild, 1987.

SOUZA, D. S. *Jogada de letra*: um estudo sobre colocações à luz da Semântica de Frames. 2015. Dissertação (Mestrado em Linguística Aplicada) – Programa de Pós-Graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos (UNISINOS), São Leopoldo, 2015. Disponível em: <http://www.repositorio.jesuita.org.br/handle/UNISINOS/3924>. Acesso em: 25 abr. 2015.

TAGNIN, S. E. O. *O jeito que a gente diz*: combinações consagradas em inglês e português. São Paulo: Disal Editora, 2013.

TALMY, G. Beyond foreground and background. In: TOMLIN, E. R. *Coherence and grounding in discourse* (Typological studies in language, 11), Amsterdam: John Benjamins, 1987.

ULLMAN, S. *Semântica*: a ciência do significado. Lisboa: Fundação Calouste Gulbenkian, 1961.

DOI: 10.17851/2237-2083.23.3.797-832

Anotação de Sentidos de Verbos em Textos Jornalísticos do *Corpus* CSTNews[1]

Verb Sense Annotation in News Texts in the CSTNews Corpus

Marco Antonio Sobrevilla Cabezudo
Universidade de São Paulo (USP), São Paulo, SP, Brasil.
marcosbc@icmc.usp.br

Erick Galani Maziero
Universidade de São Paulo (USP), São Paulo, SP, Brasil.
erickgm@icmc.usp.br

Jackson Wilke da Cruz Souza
Universidade Federal de São Carlos (UFSCar), São Carlos, SP, Brasil.
jackcruzsouza@gmail.com

Márcio de Souza Dias
Universidade de São Paulo (USP), São Paulo, SP, Brasil.
marciosouzadias@gmail.com

Paula Christina Figueira Cardoso
Universidade de São Paulo (USP), São Paulo, SP, Brasil.
paulastm@icmc.usp.br

Pedro Paulo Balage Filho
Universidade de São Paulo (USP), São Paulo, SP, Brasil.
balage@icmc.usp.br

Verônica Agostini
Universidade de São Paulo (USP), São Paulo, SP, Brasil.
agostini@icmc.usp.br

[1] Todos os autores declaram que participaram igualmente do processo de anotação do *corpus* e da escrita do artigo, o qual relata a anotação realizada e seus resultados. Todos os autores assumem a responsabilidade pelo conteúdo do artigo.

Fernando Antônio Asevedo Nóbrega
Universidade de São Paulo (USP), São Paulo, SP, Brasil.
fasevedo@icmc.usp.br

Cláudia Dias de Barros
Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP), SP, Brasil.
claudias84@gmail.com

Ariani Di Felippo
Universidade de São Paulo (USP), São Paulo, SP, Brasil.
arianidf@gmail.com

Thiago Alexandre Salgueiro Pardo
Universidade de São Paulo (USP), São Paulo, SP, Brasil.
taspardo@icmc.usp.br

Resumo: Um dos problemas mais difíceis de serem tratados no Processamento de Linguagem Natural (PLN) é a ambiguidade lexical, pois as palavras podem expressar sentidos distintos de acordo com o contexto no qual elas ocorrem. Em PLN, a tarefa responsável por determinar o sentido adequado de uma palavra em contexto é a Desambiguação Lexical de Sentido (DLS). Nessa tarefa, o uso de *corpus* anotado é muito útil, pois esse recurso linguístico computacional permite o estudo mais aprofundado da ambiguidade, assim como o desenvolvimento e a avaliação de métodos de DLS. O presente trabalho relata o processo de anotação de sentidos dos verbos em textos jornalísticos presentes no *corpus* CSTNews, usando-se a WordNet de Princeton como repositório de sentidos. As contribuições deste trabalho incluem a disponibilização de um recurso linguístico que serve de base para futuras pesquisas em DLS para o português, além de detalhar o processo de anotação e seus resultados.

Palavras-chave: Linguística de *Corpus*; Desambiguação Lexical de Sentido; Português Brasileiro.

Abstract: One of the hardest problems in Natural Language Processing (NLP) is the lexical ambiguity, as words may express different senses depending on the context in which they occur. In NLP, Word Sense Disambiguation (WSD) is the task that aims at determining the proper meaning of a word in its context. In this task, the use of a sense annotated corpus is useful because this

computational linguistic resource enables further study of the ambiguity phenomenon and the development and evaluation of WSD methods. This paper describes the verb sense annotation process in news texts in the CSTNews corpus, using Princeton WordNet as sense repository. Besides detailing the annotation process and its results, the contributions of this work include the availability of a linguistic resource that may be the basis for future research in WSD for Portuguese.

Keywords: Corpus Linguistics; Word Sense Disambiguation; Brazilian Portuguese.

Recebido em: 1º de agosto de 2015.
Aprovado em: 6 de novembro de 2015.

1 Introdução

Atualmente, existe uma quantidade imensa e crescente de informação disponível, principalmente *on-line*. Um dos principais motivos para isso foi a chegada da *Web* 2.0, na qual o usuário está envolvido na criação de conteúdo, que pode ser livremente disponibilizado em variados *blogs*, *microblogs*, fóruns e redes sociais. Consequentemente, pessoas e empresas têm necessidade de formas mais inteligentes para ler e apreender tanta informação e, assim, poder tomar decisões sobre seus negócios e desejos de forma mais informada.

Nesse cenário, para facilitar a aquisição de conteúdo disponível, a área de Processamento de Linguagem Natural (PLN) tem investido na criação de ferramentas e aplicações computacionais de processamento textual, como sistemas de tradução automática, sumarização de textos e recuperação e extração de informação. Para isso, como Jurafsky e Martin (2009) apresentam, a área deve reconhecer, representar apropriadamente e lidar com a informação veiculada nos níveis linguísticos da morfologia, da sintaxe, da semântica, do discurso e da pragmática, assim como na intersecção entre eles.

Enquanto alguns níveis de processamento da língua se encontram bem delimitados e com ferramentas com precisão satisfatória já desenvolvidas (como é o caso dos etiquetadores morfossintáticos e dos

analisadores sintáticos), outros níveis ainda necessitam de mais estudo e investimento de pesquisa, por exemplo, a semântica, ou seja, o significado veiculado pelo texto e suas partes. O processamento da língua nesse nível pode permitir o desenvolvimento de ferramentas e sistemas computacionais que realizam a interpretação de um texto de entrada com desempenho mais próximo ao do humano, produzindo resultados mais satisfatórios, portanto.

Entre os problemas relacionados ao tratamento computacional da semântica das línguas naturais, destaca-se a ambiguidade lexical, caracterizada pela multiplicidade de sentidos possíveis das palavras, que é o fenômeno conhecido por polissemia. Ressalta-se que, do ponto de vista humano, as ambiguidades são raras, pois os humanos conseguem facilmente interpretar o significado adequado de uma palavra polissêmica com base em conhecimento linguístico, de mundo e situacional. Do ponto de vista computacional, o processo de interpretação do significado não é tão imediato, pois se faz necessário instrumentar o computador com métodos de desambiguação e conhecimento apropriados.

Como ilustração da tarefa supracitada, apresentam-se, nas sentenças seguintes, exemplos de ambiguidade lexical com diferentes níveis de complexidade. Nesses exemplos, as palavras em destaque são polissêmicas, ou seja, são palavras que têm mais de um significado.

1) O professor *contou* a quantidade de alunos.
2) *Tomar* uma cápsula uma hora antes das refeições principais.
3) O atacante chutou e o goleiro *tomou* um frango.
4) O banco *quebrou* na semana passada.

Analisando os exemplos, pode-se notar que, do ponto de vista humano, as sentenças (1) e (2) não apresentam ambiguidade. Porém, do ponto de vista computacional, as sentenças (1) e (2) apresentam ambiguidade, mas esta pode ser facilmente resolvida porque as pistas linguísticas estão no contexto sentencial. A ocorrência da palavra "quantidade" no contexto sentencial de “contou” em (1) ajuda a determinar que o sentido adequado é “enumerar”. Da mesma forma, a ocorrência de “cápsula” em (2) ajuda a determinar que o sentido

adequado para “tomar” é “ingerir”. Já na sentença (3), o verbo “tomar” poderia apresentar certa ambiguidade para os humanos com pouco conhecimento sobre esportes. A ocorrência das palavras "atacante", "chutar", "goleiro" e "frango" ao redor do verbo “tomar” em (3) funciona como pista para a identificação de que a expressão usada é “tomar um frango” e o sentido expresso por esta expressão é “pensar que a bola é de fácil domínio e acabar levando gol”. Finalmente, a fácil identificação do sentido não ocorre com "quebrar" em (4), pois não é possível saber se se refere ao fato de uma instituição financeira falir ou a um assento se partir em pedaços / fragmentar. Do ponto de vista humano, é necessário ter conhecimento de mundo (da situação na qual é apresentada essa sentença) para poder identificar o sentido do verbo. Do ponto de vista computacional, o contexto sentencial não fornece as pistas linguísticas necessárias para que uma máquina identifique o sentido adequado da palavra em questão.

A tarefa cujo objetivo é tratar a ambiguidade lexical, escolhendo o sentido mais adequado para uma palavra dentro de um contexto (sentença ou porção de texto maior), é chamada Desambiguação Lexical de Sentido (DLS). Na forma mais básica, os métodos de DLS recebem como entrada uma palavra em um contexto determinado e um conjunto fixo de potenciais sentidos, chamado repositório de sentidos (RS), devendo retornar o sentido correto que corresponde à palavra (JURAFSKY; MARTIN, 2009).

A DLS é usada como uma tarefa intermediária, incorporada à análise sintática ou semântica dos processos de interpretação e / ou geração da língua. Essa tarefa é relevante a inúmeras aplicações de PLN. A análise de sentimentos (AKKAYA *et al.*, 2009) é uma dessas aplicações. Nela, a identificação do sentido subjacente às palavras de um texto sob análise pode auxiliar a determinação da opinião expressa pelo texto, se positiva ou negativa, ou mesmo se o texto expressa ou não uma opinião. A tradução automática (IDE; VÉRONIS, 1998) e outras aplicações multilíngues talvez sejam as aplicações de PLN em que a necessidade da DLS se faz mais evidente, pois a identificação do sentido de uma palavra vai determinar a escolha de sua tradução. Por exemplo, se o verbo “jogar” expressar o sentido de “divertir-se, entreter-se com um jogo qualquer”, como em “Eu joguei baralho a noite inteira”, este deve

ser traduzido para "*play*" em inglês; caso expresse o sentido de "deslocar (algo) pelo ar (até determinado ponto), usando força muscular ou alguma arma", como em "O atleta conseguia *jogar* pesos a uma distância de mais de seis metros", a tradução correta deve ser "*throw*".

Entre os recursos usados no desenvolvimento de métodos de DLS, salienta-se o uso de um repositório de sentidos. Esse repositório de sentidos é usado para fornecer os possíveis sentidos que uma palavra pode assumir. Os repositórios podem ser representados por: dicionários, que organizam as palavras segundo seus significados; *thesaurus*, que são estruturas que organizam as palavras usando relações de sinonímia e antonímia; e *wordnets*, que são repositórios frequentemente utilizados em que se organizam as palavras segundo seus sentidos e as relações existentes entre eles. A *wordnet* mais usada na literatura é a WordNet de Priceton (referenciada simplesmente por WordNet-Pr) (FELLBAUM, 1998), que foi desenvolvida para a língua inglesa. Já para o português, esforços realizados deram origem à WordNet-Br (DIAS DA SILVA, 2005) (DIAS DA SILVA *et al*., 2008), à OpenWordNet-Pt (DE PAIVA, 2012) e à Onto.PT (GONÇALO OLIVEIRA *et al*., 2012), entre outras iniciativas da área.

Outro recurso muito importante para o desenvolvimento de métodos de DLS é a presença de um *corpus* anotado com sentidos, já que, com base nele, pode-se conduzir estudos teóricos sobre o fenômeno linguístico em mãos, tanto para caracterização linguística quanto para formalização de estratégias de tratamento, podem-se desenvolver e treinar sistemas de DLS e se realizar avaliações para aferir os resultados dos métodos investigados. Um *corpus* anotado, portanto, é essencial para o desenvolvimento da área e pode servir como um *benchmark* para a tarefa. Para o inglês, pode-se citar, por exemplo, o Semcor, que é o *corpus* com anotações de sentido mais usado na literatura. Foi criado pela Universidade de Princeton e inclui 352 textos extraídos do *corpus* Brown (KUCERA; FRANCIS, 1967), e possui anotações da classe gramatical, lema, e sentidos provenientes da WordNet-Pr. Vale citar também o OntoNotes (PRADHAN *et al*., 2007), que é um projeto desenvolvido em parceria entre a Raytheon BBN Technologies, a University of Colorado, a University of Pennsylvania e a University of Southern California. O objetivo desse projeto foi criar um grande *corpus*

anotado semanticamente em vários idiomas. Esse *corpus* abrange vários gêneros textuais (notícias, conversas telefônicas, *weblogs* e entrevistas, entre outros) escritos em inglês, chinês e árabe. Para o português, têm-se alguns *corpus* específicos para certos domínios e aplicações, como os apresentados por Specia (2007) e Machado *et al*. (2011), e outros mais gerais, como os apresentados por Nóbrega e Pardo (2014) e Travanca (2013).

Specia (2007) propôs um método de DLS baseado em Programação Lógica Indutiva, caracterizado por utilizar aprendizado de máquina e regras em lógica proposicional. Focado na tradução português-inglês, esse método foi desenvolvido para a desambiguação de 10 verbos bastante polissêmicos do inglês, a saber: "*ask"*, "*come"*, "*get"*, "*give"*, "*go"*, "*live"*, "*look"*, "*make"*, "*take"* e "*tell"*. Para o desenvolvimento do método, construiu-se um *corpus* paralelo composto por textos em inglês e suas respectivas traduções para o português. Nesse *corpus*, cada texto original em inglês foi alinhado em nível lexical a sua tradução em português. Tais textos foram compilados de nove fontes: a bíblia, uma versão bilíngue dos documentos do parlamento europeu, algumas traduções de livros de ficção, um conjunto de artigos da edição *on-line* do jornal *New York Times*, um conjunto de mensagens de entrada e saída utilizadas pelo sistema operacional Linux, um conjunto de resumos de teses e dissertações em Ciências da Computação do ICMC – Universidade de São Paulo, um manual do usuário da linguagem de programação PHP, documentos da ALCA (Área de Livre Comércio das Américas) e um conjunto de sentenças incluídas no romance *The Red Badge of Courage*.

Machado *et al*. (2011) apresentaram um método para desambiguação geográfica (especificamente, desambiguação de nomes de lugares) que utiliza uma ontologia composta por conceitos de regiões, chamada OntoGazetteer, como fonte de conhecimento. Para a avaliação do método, os autores utilizaram um *corpus* formado por notícias jornalísticas extraídas da *web*. Cada notícia jornalística passou por um pré-processamento, que consistiu na indexação das palavras de conteúdo aos conceitos da ontologia. Com base no *corpus* indexado à ontologia, um conjunto de heurísticas identifica o conceito subjacente a cada uma das palavras do *corpus*.

No trabalho de Nóbrega e Pardo (2014), investigaram-se métodos variados de desambiguação de substantivos comuns. Para tanto, o CSTNews (ALEIXO; PARDO, 2008) (CARDOSO *et al*., 2011), *corpus* multidocumento composto por 140 notícias jornalísticas em português, agrupadas em 50 coleções, foi anotado manualmente. Em especial, essa anotação consistiu na explicitação dos sentidos subjacentes aos substantivos comuns mais frequentes do *corpus*. A anotação dessa classe gramatical foi motivada pelos estudos sobre o impacto positivo que a desambiguação de substantivos comuns tem em aplicações de PLN (veja, por exemplo, o trabalho de PLAZA; DIAZ, 2011). Para anotar o sentido de cada substantivo, utilizou-se como repositório de sentidos a WordNet-Pr (versão 3.0) (FELLBAUM, 1998). Dado que os conceitos estão armazenados na WordNet-Pr sob a forma de conjuntos de unidades lexicais sinônimas do inglês (os *synsets*), a indexação dos substantivos em português aos conceitos foi feita com o auxílio de um dicionário bilíngue português-inglês. No caso, utilizou-se o WordReference®.

Outro trabalho de DLS que usou *corpus* anotado para o português (europeu) é o de Travanca (2013). Travanca investigou alguns métodos de DLS para verbos. Para tanto, foi anotado manualmente uma porção do *corpus Parole* (RIBEIRO, 2003), o qual é composto por livros, jornais, periódicos e outros textos. O repositório de sentidos utilizado por Travanca foi o ViPer (BAPTISTA, 2012), que armazena várias informações sintáticas e semânticas sobre os verbos do português europeu, contendo somente verbos com frequência 10 ou superior no *corpus* CETEMPúblico (ROCHA; SANTOS, 2000).

Neste artigo, relata-se o processo de anotação de sentidos para os verbos de textos jornalísticos em português brasileiro no corpus CSTNews, descrito inicialmente por Sobrevilla-Cabezudo *et al*. (2014), e sua respectiva avaliação, visando-se ao aprofundamento nas pesquisas teóricas e práticas em DLS para o português. As contribuições desse trabalho residem (i) na caracterização do fenômeno de ambiguidade lexical de verbos no gênero jornalístico, (ii) na sistematização e descrição do processo de anotação conduzido e (iii) na disponibilização de um *corpus* anotado para subsidiar pesquisas futuras na área. Segue-se na linha de trabalho adotada por Nóbrega e Pardo (2014), de forma que a

anotação produzida nesse trabalho consiste na única anotação de propósito geral de sentidos de verbos para o português brasileiro.

Este artigo, então, está estruturado da seguinte maneira: na Seção 1, Introdução do objeto de estudo, na Seção 2, apresenta-se a metodologia usada para a anotação de *corpus*; na Seção 3, apresentam-se os resultados e a avaliação da anotação. Finalmente, na Seção 4, apresentam-se algumas considerações finais.

2 Anotação de *corpus*

2.1 Considerações iniciais

Para a tarefa de anotação a ser realizada, utilizou-se o CSTNews (ALEIXO; PARDO, 2008) (CARDOSO *et al*., 2011), *corpus* multidocumento composto por 50 coleções ou grupos de textos, sendo que os textos de cada coleção abordam um mesmo tópico. A escolha do CSTNews pautou-se nos seguintes fatores: (i) utilização prévia desse *corpus* no desenvolvimento de métodos de DLS para os substantivos comuns (NÓBREGA; PARDO, 2014) e (ii) ampla variedade de domínios ou categorias (“política”, “esporte”, “mundo”, etc.), proporcionando uma gama variada de sentidos para o desenvolvimento de métodos de DLS robustos. No total, o CSTNews contém 72.148 palavras, distribuídas em 140 textos. Os textos pertencem ao gênero “notícias jornalísticas”. Os textos das coleções têm em média 42 sentenças (e a quantidade mínima de 10 e a máxima de 89).

Especificamente, cada coleção do CSTNews contém: (i) dois ou três textos, que abordam um mesmo assunto, compilados de diferentes fontes jornalísticas; (ii) sumários humanos (*abstracts*) mono e multidocumento; (iii) sumários automáticos multidocumento; (iv) extratos humanos multidocumento; (v) anotações semântico-discursivas; (vi) alinhamento sentencial; entre outras camadas de anotação linguística. As fontes jornalísticas das quais os textos foram selecionados correspondem a alguns dos principais jornais *online* do Brasil, a saber: *Folha de São Paulo*, *Estadão*, *Jornal do Brasil*, *O Globo* e *Gazeta do Povo*.

As coleções estão classificadas pelos rótulos das "seções" dos jornais dos quais os textos foram selecionados. Assim, o *corpus* é composto por coleções das seguintes categorias: "esporte" (10 coleções), "mundo" (14 coleções), "dinheiro" (uma coleção), "política" (10 coleções), "ciência" (uma coleção) e "cotidiano" (14 coleções).

O foco da presente anotação foi desambiguar as palavras identificadas como verbos. Essa escolha foi feita devido ao fato de estudos concordarem que o verbo é uma classe gramatical de grande relevância na estrutura e na construção de uma sentença (veja, por exemplo, o trabalho clássico de FILLMORE, 1968). Isso pode ser verificado pela frequência de ocorrência dessa classe de palavras no CSTNews. No trabalho de Nóbrega (2013), apresenta-se a distribuição da frequência de ocorrência das classes de palavras de conteúdo no CSTNews. Para o cômputo dessa distribuição, os textos do CSTNews passaram por um processo de etiquetação morfossintática automática, realizada pelo etiquetador morfossintático, ou *tagger* MXPOST (RATNAPARKHI, 1996), usando o modelo treinado para o português proposto por Aires (2000). Dessa etiquetação, verificou-se que a classe verbal é a segunda mais frequente (27,76%). Os substantivos compõem a classe mais frequente, com 53,44% das palavras de conteúdo do *corpus*.

Para este trabalho, seguindo-se o trabalho de Nóbrega e Pardo (2014), a WordNet de Princeton (WordNet-Pr) foi usada como repositório de sentidos. Outras alternativas que poderiam ser utilizadas foram os recursos lexicais desenvolvidos para o português, a saber: (i) TeP 2.0 (MAZIERO *et al*., 2008), Onto.PT (GONÇALO OLIVEIRA *et al.*, 2012) e WordNet.Br (DIAS DA SILVA, 2005). Os motivos que levaram ao uso da WordNet-Pr são descritos a seguir:

- é um dos recursos mais difundidos e usados na literatura;
- é considerada como uma ontologia linguística (DI FELIPPO, 2008, p. 18), isto é, abrange o conhecimento geral do mundo, com conceitos representados em língua natural (no caso, o inglês); e
- dá-se continuidade ao trabalho feito por Nóbrega e Pardo (2014) para a língua portuguesa do Brasil.

A WordNet-Pr é uma base de dados lexical desenvolvida inicialmente para o inglês pela Universidade de Princeton. Abrange substantivos, verbos, adjetivos e advérbios, que são organizados em conjuntos de sinônimos que representam os sentidos de uma palavra. Esses conjuntos de sinônimos são chamados *synsets*. Um *synset* é também acompanhado pela glosa, que é a descrição informal do sentido do *synset*, e, em alguns casos, por uma sentença de exemplo. Por exemplo, o *synset* da palavra "*die*" (cuja tradução em português é "morrer", neste caso) é composto pelos sinônimos "*die*", "*decease*", "*perish*", "*go*", "*exit*", "*pass away*", "*expire*", "*pass*", "*kick the bucket*", "*cash in one's chips*", "*buy the farm*", "*conk*", "*give-up the ghost*", "*drop dead*", "*pop off*", "*choke*", "*croak*" e "*snuff it*"; possui a seguinte glosa: "*pass from physical life and lose all bodily attributes and functions necessary to sustain life*", que faz referência a "sair da vida física e perder todos os atributos e funções necessários para a vida"; e o seguinte exemplo: "*She died from cancer*", que em português se traduz como "Ela morreu de câncer".

2. 2 Metodologia de anotação

A metodologia de anotação usada foi a proposta no trabalho de Nóbrega e Pardo (2014). Ela é dividida em duas partes, uma metodologia geral e outra individual. A metodologia geral faz referência às etapas que todos os anotadores devem seguir para anotar uma coleção de textos. Os passos que formam a metodologia geral são os seguintes:

1) escolher um texto da coleção para ser anotado;
2) anotar todas as palavras indicadas como "verbo" nesse texto e, depois disso, anotar o texto seguinte da coleção; e
3) após anotar todos os textos, revisar e salvá-los.

Salienta-se nessa metodologia que, no segundo passo, foi necessário seguir a seguinte sequência: primeiramente, ler o texto completo para ter uma noção do contexto do qual estava se tratando e, em segundo lugar, revisar todos os verbos com eventuais anotações prévias (provenientes de anotações anteriores de outras ocorrências do

mesmo verbo) e confirmar ou não a anotação, para depois anotar os verbos restantes. Nesse processo de anotação, optou-se por usar a heurística de um sentido por discurso, isto é, se um verbo for anotado com um determinado sentido em um texto (discurso), todas as outras ocorrências do mesmo verbo, no mesmo texto, são pré-anotadas com o mesmo sentido (daí a existência de anotações prévias).

A metodologia individual esteve relacionada ao processo de anotação de cada verbo. Essa metodologia foi realizada em quatro passos (ou etapas) que são mencionados a seguir:

1) selecionar o verbo a ser anotado. Ao selecionar o verbo a ser anotado, são apresentadas, como apoio, as possíveis traduções fornecidas pelo dicionário bilíngue WordReference®;
2) selecionar uma tradução. Após apresentada a lista de possíveis traduções, o anotador deve escolher uma tradução entre as possíveis, sendo apresentados os *synsets* da WordNet-Pr pertencentes à tradução selecionada;
3) selecionar um *synset.* Após apresentada a lista de possíveis *synsets,* e o anotador revisá-los, o anotador deve selecionar o *synset* mais apropriado à palavra selecionada; e
4) finalmente, anotar o sentido da palavra.

Embora a anotação das palavras tenha seguido a metodologia mencionada, alguns passos foram tratados de maneira particular. A revisão das palavras identificadas como verbos no texto-fonte se deve ao fato de que se partiu de textos-fonte anotados em nível morfossintático pelo *tagger* MXPOST (RATNAPARKHI, 1996), que, em experimentos realizados por Aires (2000), obteve uma acurácia de 97%. Apesar da alta precisão, o *tagger* incorre em erros e, por isso, a etapa de seleção do verbo a ser anotado engloba a tarefa de revisão da anotação morfossintática.

Assim, para cada palavra anotada como verbo, verifica-se se de fato a palavra em questão é um verbo. Caso a anotação automática seja correta, passa-se para o próximo passo da anotação. Caso a palavra não seja de fato um verbo, configurando um caso de ruído, tal anotação é ignorada e coloca-se um comentário de erro de anotação por parte do

tagger (por meio do editor NASP++ para apoio à anotação, apresentado posteriormente neste artigo). Por exemplo, na sentença "e o governo decretou toque de recolher", a palavra "recolher" faz parte do substantivo "toque de recolher" e, portanto, não é considerada nos seguintes passos da anotação e adiciona-se o comentário de "erro de anotação". Depois disso, analisa-se a próxima palavra anotada automaticamente como verbo, e assim por diante. Dando sequência à anotação do texto-fonte, verifica-se se há algum verbo não identificado como tal, localizado entre o último verbo anotado com sentido e o próximo a anotar. Se sim, o anotador humano realiza a anotação morfossintática (também por meio do editor NASP++) e segue para a próxima etapa da anotação sobre essa palavra. Caso contrário, não se seleciona a palavra para ser anotada.

Também se estabeleceu que os verbos auxiliares devem ser anotados como tal, especificamente pelo comentário "verbo auxiliar" (por meio do editor NASP++). Dessa forma, não se atribui sentidos a eles. A razão para essa regra foi que os verbos auxiliares não possuem uma carga semântica forte e sua principal função é marcar tempo, modo, número, pessoa e o aspecto do verbo ao qual auxiliam. Por isso, não foram considerados nessa anotação. Por exemplo, em "Ele havia saído de casa", "havia" é verbo auxiliar e "saído" (particípio) é o verbo principal.

Nas ocorrências formadas por um tempo composto seguido de infinitivo, o verbo principal (do composto) e o infinitivo deveriam ser anotados, posto que esses veiculam conteúdo próprio. Por exemplo, em "Ele havia prometido retornar", o verbo "havia" é auxiliar e, por isso, recebe uma anotação própria que evidencia sua função como tal, mas o verbo principal do composto ("prometido") e a forma no infinitivo que ocorre na sequência ("retornar") devem receber uma anotação semântica por expressarem conteúdos bem definidos e independentes.

Nos casos de predicados complexos (isto é, expressões perifrásticas que comumente possuem um equivalente semântico lexicalizado, por exemplo: "fazer questão" e "insistir", "dar um passe" e "passar" e "tomar conta" e "cuidar"), devia-se: (1) associar ao verbo da expressão o comentário "predicado complexo" (por meio do editor NASP++) e (2) anotar o verbo com um sentido / *synset* da WordNet-Pr que representasse o significado do predicado complexo. Assim, em "Ele dava crédito a ela", devia-se associar o comentário "predicado

complexo" ao verbo "dava" e anotá-lo com um *synset* que representasse o sentido do predicado complexo, que é "confiar". Ressalta-se que a identificação dos predicados complexos foi automática, por meio do editor NASP++, com base especificamente na lista de predicados estabelecida por Duran *et al*. (2011). A confirmação (ou não) de que a expressão identificada pelo editor se tratava de um predicado complexo ficou a cargo dos anotadores humanos. Contudo, apesar da lista de predicados complexos usada, foram identificados outros predicados complexos durante o processo de anotação no *corpus*. Alguns dos predicados complexos encontrados foram:

- "Levantar o caneco", cuja tradução utilizada foi "*win*" (no contexto esportivo);
- "Soltar uma bomba", cuja tradução foi "*kick*" (no contexto esportivo);
- "Bater falta", cuja tradução foi "*kick*" (no contexto esportivo);
- "Sentir falta", cuja tradução foi "*miss*".

Outra questão de anotação diz respeito à identificação dos verbos no particípio e sua diferenciação de adjetivos. Isso se deve ao fato de que a identificação das formas terminadas em "-ado (os / a / as)" ou "-ido (os / a / as)" como verbos no particípio ou adjetivos nem sempre é fácil. Assim, recuperando Azeredo (2000, p. 242-243), estabeleceu-se que:

> O particípio é sintaticamente uma forma do verbo apenas quando, invariável e com sentido ativo, integra os chamados tempos compostos ao lado do auxiliar ter. Fora daí, o particípio se torna um adjetivo [...], tanto pela forma – já que é variável em gênero e número –, quanto pelas funções, pois, assim como o adjetivo, pode ser adjunto adnominal (cf. o livro novo / livro rabiscado) ou complemento predicativo, quando constitui a chamada voz passiva (cf. Estas aves são raras / Estas aves são encontradas apenas no pantanal) (AZEREDO, 2000, p. 242-243).

Em relação ao segundo passo, como mencionado, a WordNet-Pr foi utilizada como repositório de sentidos para a anotação relatada. Como tais sentidos estão organizados em *synsets*, os quais são compostos por palavras e expressões sinônimas do inglês, os verbos em português a serem anotados precisaram ser traduzidos para o inglês.

A partir de um verbo em inglês, o editor NASP++ recupera todos os *synsets* da WordNet-Pr dos quais o verbo é elemento constitutivo e os disponibiliza aos anotadores como possíveis rótulos semânticos a serem usados para a anotação do verbo em português, cabendo ao humano selecionar, entre os *synsets* automaticamente recuperados, o que mais adequadamente representa o sentido subjacente ao verbo original em português.

Para traduzir os verbos para o inglês, o editor NASP++ acessa o dicionário bilíngue WordReference® e, a partir desse acesso, exibe aos anotadores humanos as traduções possíveis em inglês da palavra original em português. Diante da tradução automática dos verbos, estabeleceram-se duas regras para a seleção da tradução. A primeira delas estabelece que todas as traduções sugeridas pelo editor devem ser analisadas antes da seleção definitiva da tradução. Essa regra foi criada com o objetivo de se selecionar a tradução mais adequada em inglês. Por exemplo, se, para um verbo em português, o editor sugerisse quatro traduções possíveis em inglês, todas elas deveriam ser analisadas. Essa análise pode englobar a consulta a recursos diversos, como o Google Tradutor,[2] o Linguee[3] e outros dicionários bilíngues, com o objetivo de selecionar a palavra em inglês que mais adequadamente expressa o sentido do verbo em português. A segunda regra ou diretriz estabelece que, caso o editor não sugira uma tradução adequada, o anotador deveria inserir uma manualmente, a partir da qual os *synsets* seriam recuperados automaticamente na sequência. Para indicar uma tradução manualmente, sugeriu-se que os anotadores consultassem os mais variados recursos linguísticos. Entre eles, citam-se os dicionários bilíngues português-inglês, como o Michaelis Moderno Dicionário Inglês & Português, os

[2] Disponível em: <https://translate.google.com.br/>.
[3] Disponível em: <http://www.linguee.com.br/>.

diferentes dicionários disponíveis no *site Cambridge Dictionaries Online*[4] e também os serviços *online* como Google Tradutor e Linguee.

A consulta a esses recursos buscou garantir a inserção no editor da tradução mais adequada, ou seja, que codificasse de fato o sentido subjacente ao verbo em português e que estivesse armazenada na WordNet-Pr.

Em relação ao terceiro passo da metodologia individual, ressalta-se que, assim que uma tradução é selecionada, seja indicada pelo editor ou inserida manualmente pelo anotador, deve-se analisar os *synsets* compostos pela tradução para verificar se entre eles há um que seja apropriado. Para essa análise, além disso, deve-se levar em consideração o fato de que a WordNet-Pr, por vezes, apresenta *synsets* muito próximos, cuja distinção nem sempre é simples.

Um problema relacionado a esse passo foi a ocorrência de lacunas léxico-conceituais, isto é, a inexistência de um *synset* que represente o sentido específico subjacente a uma palavra. A segunda regra dessa etapa estabeleceu que um *synset* hiperônimo (ou seja, mais genérico) fosse selecionado. Por exemplo, o verbo "pedalar" na sentença "O Robinho pedalou..." não possui *synset* indexado na WordNet-Pr. Portanto, ter-se-ia que buscar uma generalização desse verbo, que poderia ser "driblar".

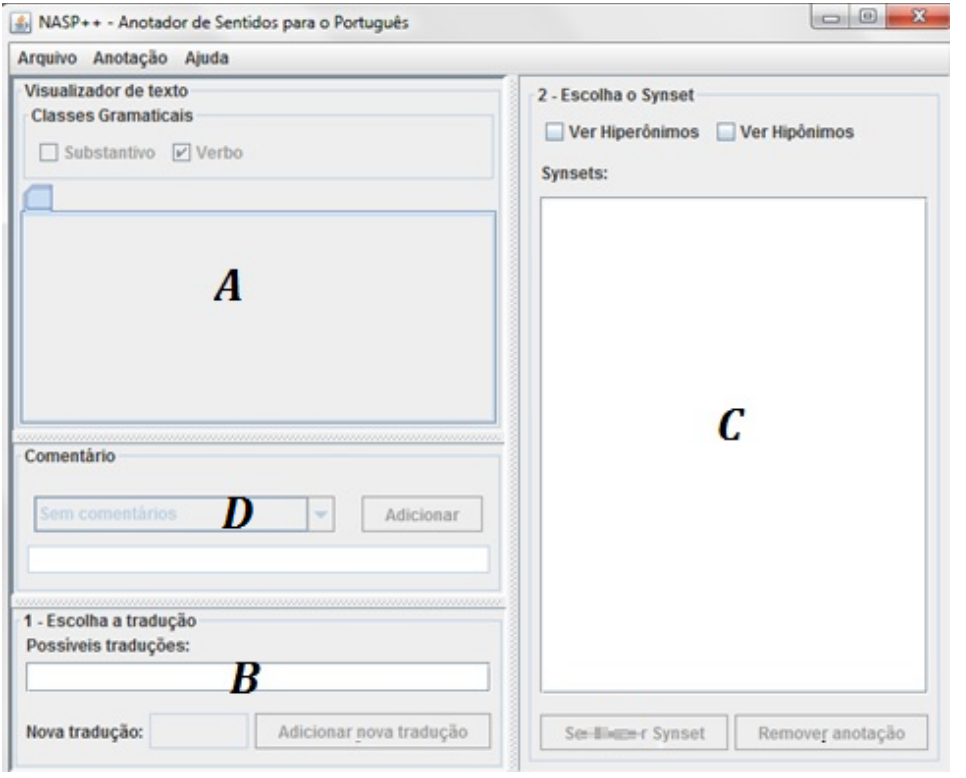
Como parte da metodologia de anotação, em caso de dúvidas, os anotadores puderam usar os recursos mencionados em cada etapa. Caso os anotadores não conseguissem resolver as dúvidas, poder-se-ia fazer uma consulta a toda a equipe de anotação.

Como mencionado, a metodologia e os recursos ora descritos contribuíram para o desenvolvimento da ferramenta NASP++, que pode ser definida como um editor de auxílio à anotação de sentidos. A NASP++ captura todos os passos da anotação de sentidos. Na Figura 1, representam-se todos os passos, começando-se pela (A) Seleção dos verbos a serem anotados, seguida pela (B) Seleção da tradução adequada para o verbo selecionado e, finalmente, (C) Seleção do *synset* adequado para a tradução. Além disso, podem-se ver algumas funcionalidades na seção D da figura, na qual podem ser adicionados comentários aos verbos selecionados.

[4] Disponível em: <http://dictionary.cambridge.org/pt/dicionario/ingles-portugues/>.

A NASP++ é, na verdade, uma versão atualizada da ferramenta NASP (NÓBREGA, 2013; NÓBREGA; PARDO, 2014), que originalmente foi desenvolvida para a anotação de sentidos dos nomes ou substantivos.

Figura 1 – NASP++



Fonte: SOBREVILLA-CABEZUDO *et al*. (2014).

As funcionalidades adicionadas nessa nova versão são descritas a seguir:

a) anotação de sentidos das palavras das classes dos substantivos e verbos que ocorrem em textos em português;
b) adição, às anotações, de um dos seguintes comentários:

- sem comentários: observação padrão; aplica-se quando não há observações a serem feitas sobre a anotação;
- não é verbo, erro de anotação: aplica-se quando a palavra a ser anotada foi erroneamente etiquetada como verbo pelo *tagger*;
- é predicado complexo: aplica-se quando o verbo a ser anotado pertence a um predicado complexo. Por exemplo, na sentença "O jogador levantou o caneco", o verbo "levantar" deve ser anotado com o comentário "É predicado complexo", pois a expressão "levantar o caneco" faz referência a "ganhar";
- é verbo auxiliar: aplica-se quando o verbo identificado pelo *tagger* é um verbo auxiliar. Por exemplo, na sentença "Ele estava brincando na rua", o verbo "estava" deve ser anotado com o comentário em questão;
- outros: aplica-se quando há outros tipos de observação a serem feitos sobre o processo de anotação semântica de uma palavra (seja ela verbo, seja ela substantivo), incluindo dificuldades de anotação;

c) delimitação da quantidade de palavras para anotação: ao contrário da NASP, que restringia a anotação de sentidos aos substantivos que pertenciam a um conjunto dos 10% mais frequentes da coleção de textos-fonte, a ferramenta NASP++ não possui essa limitação, por isso é que qualquer porcentagem dos verbos (e também substantivos) que ocorrem nos textos-fonte pode ser submetida ao processo de anotação; e

d) geração de ontologia: por meio dessa funcionalidade, o editor NASP++ recupera, da WordNet-Pr, a hierarquia léxico-conceitual à qual cada *synset* utilizado na anotação pertence e unifica as hierarquias individuais de cada conceito em uma única estrutura hierárquica. A finalidade dessa funcionalidade é prover um recurso (a ontologia) para outras tarefas, como sumarização automática, ou representar como estão

distribuídos os sentidos dos verbos, de acordo com o domínio ao qual pertencem.

3 Resultados da anotação

3.1 Visão Geral dos Resultados

A anotação foi realizada diariamente, em sessões de uma hora, durante sete semanas e meia, e a primeira metade da primeira semana foi destinada ao treinamento e teste da ferramenta NASP++ pelos anotadores.

Cada coleção do CSTNews foi anotada uma única vez por um único grupo de anotadores, com exceção das coleções utilizadas para obter os valores de concordância, as quais foram anotadas por todos os grupos.

No total, participaram 10 anotadores. No caso, esses anotadores eram linguistas computacionais com graduação em Linguística / Letras ou Ciência da Computação. A cada sessão de anotação, os anotadores foram organizados em grupos de dois a três anotadores, e cada grupo ficou responsável por uma coleção do *corpus*. Os grupos foram sempre compostos por linguistas e cientistas da computação, de tal forma que, em cada dia de anotação, havia configurações diferentes de linguistas e cientistas da computação em cada grupo, de forma a (i) evitar vícios (*bias*) de anotação e (ii) permitir o compartilhamento de experiências de anotação. Com isso, buscou-se compartilhar o conhecimento dos anotadores, atingindo um padrão de anotação.

Na Tabela 1, apresentam-se dados gerais obtidos da anotação. Salienta-se que as 5.082 instâncias de verbos principais que foram anotadas representam 844 verbos principais diferentes (*types*), e que, para esses, foram indicadas 787 traduções e anotados 1.047 *synsets* diferentes. O item “Erros de anotação” refere-se às palavras erroneamente indicadas como verbos pela ferramenta NASP++ e que foram manualmente corrigidas.

Tabela 1 – Estatísticas da anotação do corpus CSTNews

	Total	Verbos principais	Predicados complexos	Verbos auxiliares	Erros de anotação
# instâncias anotadas	6494	5082	146	949	317

Comparando-se com as estatísticas apresentadas por Nóbrega e Pardo (2014) para os substantivos (apresentadas na Tabela 2), ressalta-se que os verbos possuem uma maior variação de sentidos, tanto no *corpus* todo quanto em coleções de textos individuais. Com esses resultados, infere-se que a tarefa de anotação para os verbos é mais difícil do que para os substantivos. Uma razão é porque os verbos são mais polissêmicos, como afirmam Miller *et al*. (1990).

Tabela 2 – Comparação entre as estatísticas da anotação no trabalho de Nóbrega e Pardo (2014) e no presente trabalho

	Substantivos	Verbos
Número máximo de *synsets* anotados por *type* no *corpus*	5	18
Número máximo de *synsets* anotados por *type* em uma coleção (e não no *corpus* todo)	3	4
Média do número de *synsets* possíveis (disponíveis para escolha na anotação) por *type*	6	12
Porcentagem de *types* ambíguos	77%	82.1%

Algumas das dificuldades encontradas na anotação são discutidas a seguir.

Apesar da existência da lista de predicados complexos fornecida pela NASP++, a detecção de predicados complexos foi uma tarefa difícil. Por exemplo, a ferramenta indicava como predicado complexo a expressão “ficaram feridas” e, portanto, segundo as diretrizes de anotação, dever-se-ia anotar o verbo ‘ficar’ com o sentido da expressão.

No entanto, durante a anotação, alguns anotadores anotaram a palavra "ficaram" como verbo auxiliar, gerando discordâncias.

Outra dificuldade da anotação de sentido dos verbos foi a ausência de *synsets* que adequadamente representassem certos conceitos expressos pelos verbos em português. Esses casos representam as chamadas lacunas léxico-conceituais. Por exemplo, o verbo "pedalar", com o sentido de "passar o pé por sobre a bola, em especial, por repetidas vezes, com o objetivo de enganar seu marcador", como em "Robinho pedalou, driblou o zagueiro e chutou", não é lexicalizado em inglês. Para esses casos, é possível muitas vezes identificar um *synset* aproximado, porém, ele será composto por palavras ou expressões de outras classes de palavras. Por exemplo, na WordNet-Pr, tem-se o conceito "impulsionar, mover-se em uma direção particular" ("*propel*"), cujo *synset* correspondente é composto pelas unidades nominais [*dribble, carry*]. Para esses casos, a diretriz de anotação é a de generalização, portanto, nessa sentença, o verbo "pedalar" foi generalizado para "driblar" e foi selecionado o *synset* correspondente na WordNet-Pr ([*dribble, carry*]).

Outro problema de falta de *synsets* foi para o verbo "poder". Não foram encontrados *synsets* adequados para nenhuma das traduções possíveis ("*can*", "*may*", "*could*" e "*might*"), pois o verbo é usado como modal na maioria dos casos (por exemplo: "Não podemos fazer nada.") e não possui sentidos indexados na WordNet-Pr. Portanto, esses verbos não foram anotados. Tampouco foi encontrado *synset* para o verbo "vir" na sentença "O ano que vem...", pois o verbo é parte da expressão fixa "que vem", cujo sentido é "próximo, que se segue imediatamente" e a função é adjetival.

Além disso, os anotadores identificaram alguns verbos (ou predicados complexos) no CSTNews com conceitos subjacentes muito específicos, que, por isso, não estavam contemplados na WordNet-Pr. Portanto, a anotação desses verbos foi feita por meio de um processo de generalização, que consistiu na tradução do verbo para uma tradução mais genérica e, na sequência, na seleção de um *synset* que adequadamente representasse o conceito subjacente ao item lexical generalizado. Por exemplo, a expressão "tomar um frango" (por exemplo: "para a defesa do camisa 1 argentino, que quase bateu roupa e

tomou um frango."), que no domínio do futebol significa que "o goleiro toma um gol por falha grave cometida por ele", foi traduzida para a tradução genérica "*mistake*" ("errar") e, com base nele, selecionou-se o *synset* apropriado. A mesma diretriz foi seguida para a expressão "dar uma meia lua", a qual foi traduzida para "*dribble*" ("driblar").

Ressalta-se que os exemplos citados são de conceitos relativos a domínios específicos ou especializados e, nesses casos, o nível de expertise dos anotadores quanto aos domínios pode influenciar a anotação. Caso os anotadores fossem especialistas em futebol, talvez a escolha da tradução e do *synset* tivesse sido diferente.

3.2 Avaliação

A avaliação realizada considerou quatro medidas. A primeira delas foi a medida Kappa (CARLETTA, 1996). Essa medida computa o grau de concordância entre os anotadores em determinada tarefa, descontando-se a concordância ao acaso. As outras três medidas de avaliação usadas, que computam de forma direta o número de concordâncias entre os anotadores, são descritas a seguir:

- concordância total: número de vezes em que todos os anotadores concordaram, em relação ao total de instâncias anotadas;
- concordância parcial: número de vezes em que a maioria dos anotadores concordou (a metade do número de anotadores ou mais), em relação ao total de instâncias anotadas; e
- concordância nula: número de vezes em que houve discordância total ou em que não se configurou uma maioria na anotação, em relação ao total de instâncias anotadas.

A avaliação de concordância entre anotadores é importante para verificar a qualidade da tarefa realizada e poder ter confiança nos dados, para que conclusões possam ser obtidas. O nível de concordância esperado varia de tarefa para tarefa e, quanto mais subjetiva a tarefa, menor a concordância.

Devido à tarefa de anotação apresentar uma etapa de tradução para poder obter o sentido da WordNet-Pr, fez-se necessária a avaliação da concordância: (1) na etapa de tradução; (2) na etapa de escolha do *synset*; e (3) na combinação da seleção da tradução com seu respectivo *synset*.

A avaliação foi realizada a partir da anotação de três coleções do CSTNews, as mesmas utilizadas por Nóbrega e Pardo (2014) para a avaliação da anotação de sentidos dos substantivos, referenciadas por C15, C29 e C50. Na avaliação, cada coleção foi anotada por quatro grupos diferentes de anotadores, obtendo-se os resultados apresentados nas Tabelas 3, 4 e 5.

Tabela 3 – Valores de concordância para a coleção C15

	Kappa	Total (%)	Parcial (%)	Nula (%)
Seleção da tradução	0.591	42.11	52.63	5.26
Seleção do *synset*	0.483	35.53	56.58	7.89
Seleção de tradução+*synset*	0.421	28.95	63.16	7.89

Tabela 4 – Valores de concordância para a coleção C29

	Kappa	Total (%)	Parcial (%)	Nula (%)
Seleção da tradução	0.659	48.82	48.82	2.36
Seleção do *synset*	0.514	35.43	58.27	6.30
Seleção de tradução+*synset*	0.485	32.28	60.63	7.09

Tabela 5 – Valores de concordância para a coleção C50

	Kappa	Total (%)	Parcial (%)	Nula (%)
Seleção da tradução	0.695	55.50	44.04	0.46
Seleção do *synset*	0.529	34.40	60.55	5.05
Seleção de tradução+*synset*	0.516	33.95	60.09	5.96

Quanto à medida Kappa nas Tabelas 3, 4 e 5, nota-se que os valores obtidos para cada um dos 3 critérios de avaliação aumentaram a cada experimento, que seguiu a sequência C15 → C19 → C50. Por exemplo, quanto ao critério "seleção de tradução+*synset*", obteve-se: (i) 0.421 na anotação da C15 (1ª concordância), (ii) 0.485 na anotação da C29 (2ª concordância) e (iii) 0.516 na anotação da C50 (3ª concordância) (ou seja, 0.421 < 0.485 < 0.516). Uma possível justificativa para o aumento no valor de concordância entre os anotadores pode ser a experiência adquirida por eles durante o processo de anotação de sentidos, isto é, quanto maior a familiaridade com as regras / diretrizes e a ferramenta de anotação, maior foi o nível de concordância. Outra possibilidade diz respeito à familiaridade dos anotadores com os temas tratados nas coleções C15, C29 e C50. Supondo-se que o conhecimento do assunto por parte dos anotadores é um fator importante para um bom desempenho na anotação, pode-se sugerir a hipótese de que o tema abordado em C15, isto é, "explosão em um mercado em Moscou", é eventualmente menos familiar aos anotadores do que o de C29, ou seja, "pagamento de indenização pela igreja católica", o qual, por sua vez, é menos familiar que o de C50 ("proposta do governo sobre cobrança de imposto"). Tal hipótese, no entanto, necessita de verificação posterior.

Quanto às outras medidas de concordância, salientam-se os valores altos obtidos para as concordâncias total e parcial, com baixos valores de concordância nula. Isso mostra que, em muitos casos, os anotadores tiveram dúvidas e discordâncias na anotação e que também puderam convergir em muitos outros casos. Algumas causas para a concordância parcial alta podem ter sido a identificação de verbos no particípio, a identificação de predicados complexos e a identificação dos

verbos auxiliares. Por exemplo, no fragmento "foi cancelada", a palavra "foi" foi anotada como verbo auxiliar por alguns anotadores e como um verbo principal em algumas ocasiões por outros anotadores; e a palavra "cancelada" foi ora anotada como adjetivo (tratando-se, portanto, de um erro de anotação do *tagger*) e ora como verbo principal.

Outro ponto a destacar é que a concordância média do critério "seleção da tradução" é superior à do critério "seleção do *synset*". Esse resultado era esperado, pois a tradução é uma tarefa mais usual e direta que a desambiguação lexical de sentido. Sobre o critério "seleção de tradução + *synset*", vê-se que o valor médio da concordância é o menor. Isso se deve ao fato de que diferentes traduções podem fazer referência ao mesmo *synset* e diferentes *synsets* podem ser referenciados pela mesma tradução, ou seja, há mais possibilidades de combinação entre traduções e *synsets*, o que impacta nos resultados da concordância.

Outra avaliação realizada foi a comparação com os resultados obtidos no trabalho de Nóbrega e Pardo (2014) para os substantivos. Como foi mencionado, as coleções usadas na anotação de verbos foram as mesmas, para, assim, poder fazer uma avaliação o mais justa possível. Salienta-se que, na anotação de substantivos, foram anotados apenas 10% do total de substantivos, no entanto, podem-se usar esses valores como uma amostra para essa comparação. Na Tabela 6, apresentam-se os valores de concordância obtidos nos dois trabalhos.

Tabela 6 – Valores de concordância obtidos por Nóbrega e Pardo (2014) e neste trabalho

	Substantivos				Verbos			
	Kappa	Total (%)	Parcial (%)	Nula (%)	Kappa	Total (%)	Parcial (%)	Nula (%)
Tradução	0.853	82.87	11.08	6.05	0.648	48.81	48.50	2.69
Synset	0.729	62.22	22.42	14.36	0.509	35.12	58.47	6.41
Tradução-*Synset*	0.697	61.21	24.43	14.36	0.474	31.73	61.29	6.98

Analisando os resultados da Tabela 6, nota-se que os valores de concordância para os substantivos são, na maioria, superiores aos verbos. Esse resultado era esperado, devido à maior complexidade que os verbos apresentam (já que, no caso dos verbos, teve-se de identificar se era um verbo principal; caso contrário, descartar os verbos ou anotar com outros tipos como verbo auxiliar ou predicado complexo) e ao maior grau de polissemia presente nos verbos. Contudo, os valores obtidos na anotação de sentidos de verbos são aceitáveis no cenário da DLS.

No processo de anotação ora descrito, também é interessante analisar os casos em que ocorreram concordância total, parcial e nula, para melhor entendimento do fenômeno em mãos. Alguns casos de concordância total para os verbos, para ilustração, estão listados a seguir:

"morreram", "reduzir", "hospitalizada, "investigar", "têm", "acreditar", "informou", "disseram", "abusados", "convencer" e "começa".

A respeito das palavras que apresentaram com concordância total, listam-se sentenças retiradas do *corpus* nas quais algumas das palavras ocorrem, especificamente, "morreram", "hospitalizada" e "investigar".

a. "Nove pessoas **morreram**, três delas crianças, e..."
b. A maioria dos feridos, entre os quais há quatro com menos de 18 anos, foi **hospitalizada**.
c. A procuradoria de Moscou anunciou a criação de um grupo especial para **investigar** o acidente."

Algumas das razões aventadas para a concordância total na anotação de tais palavras são:

- os verbos expressam conceitos claros; por exemplo, na sentença em (a), pode-se facilmente determinar o sentido do verbo "morrer", que é "perder todos os atributos e funções corporais para manter a vida";

- o verbo possui uma tradução direta para o inglês; no caso acima, "*die*".

- os vários sentidos que a tradução pode expressar são bem delimitados e distintos e estão codificados na WordNet-Pr por *synsets* (assim como glosas e frases exemplo) bem formulados; no caso, entre os 11 conceitos que "*die*" pode expressar, identifica-se facilmente o *synset* [*die, decease, perish, go, exit, pass away, expire, pass, kick the bucket, cash in one's chips, buy the farm, conk, give-up the ghost, drop dead, pop off, choke, croak, snuff it*], cuja glosa é ("sair da vida física e perder todos os atributos corporais e funções necessários para sustentar a vida") ("*pass from physical life and lose all bodily attributes and functions necessary to sustain life*").

- tais palavras em média expressam poucos conceitos distintos; por exemplo, em (b), o verbo "hospitalizada" expressa apenas um conceito e, por isso, é elemento constitutivo de apenas um *synset* ([*hospitalize, hospitalize*]) na WordNet-Pr; em (c), o verbo "investigar" pode expressar dois conceitos, codificados na WordNet-Pr pelos *synsets* [*investigate, inquire, enquire*] e [*investigate, look into*].

A seguir, mostra-se uma lista de verbos que obtiveram concordância parcial:

"marcados", "exige", "revelados", "encerrar", "afirmou", "comentou", "peço", "acontecer", "avançar", "mexe", "isolado", "começou", "informou", "descartaram"," "declarou".

A seguir, exibem-se algumas sentenças nas quais alguns desses verbos apareceram, acompanhadas dos *synsets* indicados pelos anotadores:

a. "(...) nesta segunda-feira em uma explosão ocorrida em um mercado de Moscou, **informou** a polícia."

- *[inform] : impart knowledge of some fact, state of affairs, or event to.*
- *[inform] : act as an informer.*

b. para Virgílio, o governo ainda pode **avançar** na proposta.

- *[progress, pass on, move on, march on, go on, advance]: move forward, also in the metaphorical sense.*
- *[throw out, advance]: bring forward for consideration or acceptance.*
- *[shape up, progress get on, get along, come on, come along, advance]: develop in a positive way.*

Uma razão pela qual a concordância obtida pode ter sido parcial é que os *synsets* selecionados, apesar de distintos, possuíam certa proximidade conceitual, o que evidencia novamente a dificuldade de delimitação do conceito subjacente ao verbo em português. Um detalhe a salientar é que, ao analisar os *synsets* escolhidos, percebe-se que em alguns casos qualquer um deles poderia ser aceito como sentido correto da palavra anotada. Um estudo mais aprofundado seria útil porque se dois (ou mais) *synsets* diferentes fossem válidos, aumentaria o nível de concordância apresentado anteriormente.

Finalmente, mostra-se uma lista de palavras que obtiveram concordância nula:

"localizado", "conseguiram", "surgirem", "somariam", "enfrentar", "entenderam", "haverá", "adiantaram", "tramitando", "levar", "daria", "assinalaram", "veio", "caminhe" e "aceitamos".

Para essa lista, apresentam-se sentenças extraídas do *corpus* em que alguns dos verbos ocorrem, em particular, "localizada", "adiantaram" e "assinalaram". Cada sentença é seguida pelos diferentes *synsets* selecionados pelos anotadores, os quais resultaram em "concordância nula".

a. "A bomba detonou no interior de uma cafeteria **localizada** no setor denominado "Evrazia" do mercado Cherkizov."

- *[put, set, place, pose, position, lay]: put into a certain place or abstract location*
- *[locate, place, site]: assign a location to*
- *[set, localize, localise, place]: locate*
- *[situate, locate]: determine or indicate the place, site, or limits of, as if by an instrument or by a survey.*

b. "(...) fontes da polícia moscovita **adiantaram** que ela teria acontecido provavelmente por causa da explosão acidental de um bujão de gás."

- *[inform]: impart knowledge of some fact, state or affairs, or event to*
- *[submit, state, put forward, posit]: put before*
- *[advance, throw out] : bring forward for consideration or acceptance*
- *[announce, declare] : announce publicly or officially*

c. "As autoridades policiais de Moscou **assinalaram** que no recinto do mercado..."

- *[inform]: impart knowledge of some fact, state or affairs, or event to*
- *[state, say, tell]: express in words*
- *[notice, mark, note]: notice or perceive*
- *[announce, declare]: announce publicly or officially*

Alguns das razões pelas quais a concordância obtida pode ter sido nula são as seguintes:

- os verbos expressam conceitos relativamente vagos, de difícil delimitação; em (c), tem-se um exemplo paradigmático de verbo (no caso, "assinalar"), cujo sentido é de difícil delimitação;
- os anotadores utilizaram traduções distintas, o que pode ser explicado pela dificuldade de se delimitar ou definir o sentido; por exemplo, para "localizada" em (c), foram utilizados "*locate*" e "*localize*".
- a seleção das traduções distintas pode ter levado os anotadores a selecionarem *synsets* diferentes; isso foi observado na anotação dos verbos em (a), (b) e (c).
- da mesma forma que na concordância parcial, existe proximidade conceitual entre *synsets* diferentes.

4 Considerações finais

A anotação semântica das palavras de um *corpus* é uma tarefa bastante complexa, dada a dificuldade de delimitar os conceitos subjacentes às palavras. Para o caso dos verbos, essa complexidade fica evidente principalmente pelo alto grau de polissemia das palavras que pertencem a essa classe. Como consequência, os níveis de concordância entre os anotadores são relativamente baixos (em relação a outras classes gramaticais). Contudo, a criação de um *corpus* cujos verbos possuem anotação de sentido fornece um recurso linguístico que possibilita avançar as pesquisas sobre a tarefa de DLS para o português, que tem sido relativamente pouco explorada, devido à falta de recursos linguísticos adequados. Salienta-se que, só recentemente, alguns recursos focados nos verbos, seus sentidos e / ou seu comportamento, têm sido desenvolvidos, por exemplo, a VerbNet.Br (SCARTON, 2013), o Verbo-Brasil (DURAN; ALUISIO, 2015) e o VerbLexPor (ZILIO, 2015).

A disponibilização desse *corpus* anotado, com textos jornalísticos de vários domínios, deve subsidiar pesquisas em DLS de propósito geral, ou seja, que não sejam voltadas para aplicações e domínios específicos, como até recentemente ocorria para a língua portuguesa.

O *corpus* anotado e a ferramenta de edição NASP++ estão disponíveis para uso da comunidade de pesquisa e podem ser encontradas na página *web* do projeto SUCINTO (cf. <www.icmc.usp.br/~taspardo/sucinto/>), projeto maior que englobou esta pesquisa e que trata do desenvolvimento de recursos, ferramentas e aplicações para acesso mais inteligente à informação, em particular, na área de sumarização automática de textos.

5 Agradecimentos

Parte dos resultados apresentados neste artigo foram obtidos por meio do projeto intitulado "Processamento Semântico de Textos em Português Brasileiro", patrocinado pela Samsung Eletrônica da Amazônia Ltda., nos termos da lei federal brasileira No. 8.248/91. Também apresentamos nossos agradecimentos à FAPESP e à CAPES pelo apoio a esta pesquisa.

6 Referências

AIRES, R.V.X. *Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil*. 2000. 166 p. Dissertação. (Mestrado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2000.

AKKAYA, C.; WIEBE, J.; MIHALCEA, R. Subjectivity Word Sense Disambiguation. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, 2009, Singapore. *Proceedings*... Singapore: Association for Computational Linguistics, 2009. p. 190-199. DOI: <http://dx.doi.org/10.3115/1699510.1699535>

ALEIXO, P.; PARDO, T.A.S. CSTNews: um *corpus* de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (Cross-document Structure Theory). *Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação*, Universidade de São Paulo, n. 326, São Carlos, SP, 2008.

AZEREDO, J. C. *Fundamentos de Gramática do Português*. 1. ed. São Paulo: Jorge Zahar, 2000.

BAPTISTA J. ViPEr: A Lexicon-Grammar of European Portuguese Verbs. In: INTERNATIONAL CONFERENCE ON LEXIS AND GRAMMAR, 31, 2012, Nové Hrady. *Proceedings*... Jan Radimsky, Nové Hrady, Czech Republic, 2012. p. 10-16.

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: RST BRAZILIAN MEETING, 3, 2011, Cuiabá. *Proceedings*... Cuiabá, Sociedade Brasileira de Computação, 2011. p. 88-105.

CARLETTA, J. C. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics*, Cambridge, v. 22, n. 2, p. 249-254, 1996.

DE PAIVA, V.; RADEMAKER, A.; DE MELO, G. OpenWordNet-PT: An Open Brazilian Wordnet for Reasoning. In: COLING, 2012, Mumbai. *Proceedings*... Demonstration Papers, The COLING 2012 Organizing Committee, 2012. p. 353-360.

DIAS DA SILVA, B. C. A construção da base da WordNet.br: Conquistas e desafios. In: WORKSHOP IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 3, 2005; CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 25, 2005. São Leopoldo. *Proceedings*... 2005, p. 2238-2247.

DIAS DA SILVA, B. C.; DI FELIPPO, A.; NUNES, M. G. V. The automatic mapping of Princeton WordNet lexical-conceptual relations onto the Brazilian Portuguese WordNet database. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 6, 2008, Marrakech. *Proceedings*... Marrakech: European Language Resources Association, 2008. p. 1535-1541.

DI FELIPPO, A. *Delimitação e Alinhamento de Conceitos Lexicalizados no Inglês Norte-americano e no Português Brasileiro.* 2008. 253 p. Tese,

Faculdade de Ciências e Letras, Universidade Estadual Paulista, São Paulo, 2008.

DURAN, M. S.; RAMISCH, C.; ALUÍSIO, S. M.; VILLAVICENCIO, A. Identifying and Analyzing Brazilian Portuguese Complex Predicates. In: WORKSHOP ON MULTIWORD EXPRESSIONS: FROM PARSING AND GENERATION TO THE REAL WORLD, 2011, Portland. *Proceedings*... Portland, US: Association for Computational Linguistics, 2011. p. 74-82.

DURAN, M. S.; ALUÍSIO, S. M. Automatic Generation of a Lexical Resource to support Semantic Role Labeling in Portuguese. In: THE FOURTH JOINT CONFERENCE ON LEXICAL AND COMPUTATIONAL SEMANTICS, 2015, Denver, Colorado. *Proceedings*... Denver, Colorado, 2015. p. 216-221.

FELLBAUM, C. *WordNet An Eletronic Lexical Database*. 1. ed. Cambridge. MIT Press, 1998.

FILLMORE, C.J. The Case for Case. E. Bach and R. T. Harms, eds., *Universals in linguistic theory*, New York, Holt, Rinehart & Winston, 1968.

GONÇALO OLIVEIRA, H.; ANTÓN PÉREZ, L.; GOMES, P. Integrating lexical-semantic knowledge to build a public lexical ontology for Portuguese. In: INTERNATIONAL CONFERENCE ON APPLICATIONS OF NATURAL LANGUAGE PROCESSING AND INFORMATION SYSTEMS, 17, 2012, Berlin. *Proceedings*... Berlin: Springer-Verlag, 2012. p. 210-215.
DOI: <http://dx.doi.org/10.1007/978-3-642-31178-9_23>

IDE, N.; VÉRONIS, J. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, Cambridge, v. 24, n. 1, p. 2-40, 1998.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing*: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2 ed. Englewood Cliffs, New Jersey: Prentice-Hall, 2009.

KUCERA, H.; FRANCIS, W.N. *Computational analysis of present-day American English*. 2. ed. Providence: Brown University Press, 1967.

MACHADO, I. M.; DE ALENCAR, R. O.; CAMPOS, R.; DAVIS, C. A. An ontological gazetteer and its application for place name disambiguation in text. *Journal of the Brazilian Computer Society,* v. 17, n. 4, p. 267-279, 2011. DOI: <http://dx.doi.org/10.1007/s13173-011-0044-4>

MAZIERO, E. G.; PARDO, T. A. S.; DI FELIPPO, A.; DA SILVA, B. C. D. A base de dados lexical e a interface *web* do tep 2.0 – thesaurus eletrônico para o português do Brasil. In: WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 6, 2008, Vila Velha, ES. *Anais*... (TIL 2008) Vila Velha, 2008. p. 390-392.

MILLER, G. A.; BECKWITH, R.; FELLBAUM, C.; GROSS, D.; MILLER, K. J. Introduction to WordNet: An online lexical database. *International Journal of Lexicography,* v. 3, n. 4, p. 235-244, 1990. DOI: <http://dx.doi.org/10.1093/ijl/3.4.235>

NÓBREGA, F. A. A. *Desambiguação Lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento.* 2013. 106 p. Dissertação (Mestrado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2013.

NÓBREGA, F. A. A.; PARDO, T. A. S. General Purpose Word Sense Disambiguation Methods for Nouns in Portuguese. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE, 11, 2014, São Carlos, SP. *Proceedings of the PROPOR 2014 PhD and MSc/MA Dissertation Contest*, São Carlos, 2014. p. 94-101.

PLAZA, L.; DIAZ, A. Using semantic graphs and word sense disambiguation techniques to improve text summarization. In: CONGRESO DE LA SOCIEDAD ESPAÑOLA PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL, 27, 2011, Huelva. *Proceedings*… Huelva, España, 2011. p. 97-105.

PRADHAN, S. S.; HOVY, E.; MARCUS, M.; PALMER, M.; RAMSHAW, L.; WEISCHEDEL, R. OntoNotes: A Unified Relational Semantic Representation. In: INTERNATIONAL CONFERENCE ON SEMANTIC COMPUTING, 4, 2007. *Proceedings*... Irvine, CA: IEEE, 2007. p. 517-526. DOI: <http://dx.doi.org/10.1109/icsc.2007.83>

RATNAPARKHI, A. A Maximum Entropy Part-Of-Speech Tagger. In: EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING CONFERENCE, 1996, Pennsylvania. *Proceedings*... Pennsylvania, 1996. p. 133-142.

RIBEIRO, R. *Anotação Morfossintáctica Desambiguada do Português*. 2003. 78 p. Dissertação. (Mestrado) – Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, 2003.

ROCHA, P. A.; SANTOS, D. CETEMPúblico: Um *corpus* de grandes dimensões de linguagem jornalística portuguesa. In: ENCONTRO PARA O PROCESSAMENTO COMPUTACIONAL DA LÍNGUA PORTUGUESA ESCRITA E FALADA, 5, 2000. *Conferências*... Atibaia: Maria das Graças Volpe Nunes ed., 2000. p. 131-140.

SCARTON, C. E. *VerbNet.Br*: construção semiautomática de um léxico verbal online e independente de domínio para o português do Brasil. 2013. 242 p. Dissertação (Mestrado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2007.

SPECIA, L. *Uma abordagem híbrida relacional para a desambiguação lexical de sentido na tradução automática*. 2007. 245 p. Tese (Doutorado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2007.

SOBREVILLA-CABEZUDO, M. A.; MAZIERO, E. G.; SOUZA, J. W. C.; DIAS, M. S.; CARDOSO, P. C. F.; BALAGE FILHO, P. P.; AGOSTINI, V.; NÓBREGA, F. A. A.; DE BARROS, C. D.; DI FELIPPO, A.; PARDO, T. A. S. Anotação de Sentidos de Verbos em Notícias Jornalísticas em Português do Brasil. *Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação*, Universidade de São Paulo, n. 402. São Carlos, SP, 2014.

TRAVANCA, T. *Verb Sense Disambiguation.* 2013. 72 p. Dissertação (Mestrado) – Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, 2013.

ZILIO, L. *Verblexpor: um recurso léxico com anotação de papéis semânticos para o português*. 2015. 196 p. Tese (Doutorado em Linguística) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, 2015.

DOI: 10.17851/2237-2083.23.3.833-881

Collocations workbook: um material de apoio pedagógico on-line baseado em *corpus* para o ensino de colocações em inglês

Collocations workbook: a corpus-based online pedagogical support material for the teaching of English collocations

Adriane Orenha-Ottaiano
Universidade Estadual Paulista (UNESP), São José do Rio Preto, SP, Brasil.
adriane@ibilce.unesp.br

RESUMO: Esta pesquisa defende que a seleção de colocações para o ensino de língua estrangeira deve ser inicialmente realizada com o propósito de atender aprendizes de uma L1 específica e, desse modo, sustenta que materiais de ensino devam ser elaborados a começar por uma seleção cuidadosa de colocações que retratem as dificuldades específicas desses aprendizes (MACKIN, 1978). Nesse sentido, este artigo visa tratar da compilação de um *workbook on-line* de colocações na língua inglesa, baseado em corpus, no intuito de permitir que professores de inglês como LE trabalhem com as colocações em sala de aula de modo mais eficiente e auxiliem seus alunos a empregá-las de modo mais preciso e produtivo. Temos também como público-alvo tradutores aprendizes e profissionais, principalmente aqueles cuja L1 é o português. As atividades estão sendo desenvolvidas com base nas dificuldades dos alunos brasileiros no uso de colocações, segundo dados observados em traduções produzidas por alunos universitários, na direção português-inglês, que compõem um *corpus* paralelo, o *Corpus de Aprendizes de Tradução*, bem como de redações, escritas por alunos universitários, que compõem o *Corpus de Aprendizes de Língua Estrangeira*. Para extrair os dados, utilizamos o programa *WordSmith Tools* (SCOTT, 2008), possibilitando-nos levantar os padrões colocacionais mais frequentemente empregados pelos alunos. O *Corpus of Contemporary American English* (DAVIES, 1990-2012) também é utilizado, para verificar a frequência e recorrência dos padrões colocacionais extraídos, assim como levantar outras colocações para serem incluídas nos

exercícios que irão compor o *workbook*. A proposta objetiva contribuir para a conscientização do aspecto colocacional e, principalmente, o desenvolvimento da fluência na língua inglesa.

Palavras-chave: *workbook*; colocações; *corpus*; material pedagógico; ensino de colocações.

ABSTRACT: This study argues that the selection of collocations should be geared to targeting L2 learners of a particular L1 background and thus teaching material should be designed with a careful selection of collocations focusing on their specific difficulties (MACKIN, 1978). Bearing that in mind, this article aims at discussing the compilation of an online corpus-based collocational workbook, in order to allow English as a foreign language teachers to work with the collocations in the classroom more effectively and help learners use them more accurately and productively. Our public audience is also learner and professional translators, mainly the ones whose L1 is Portuguese. The activities are being developed based on the difficulties the Brazilian learners had regarding the use of collocations, raised from the translations carried out by university students in the Portuguese-English direction, which comprise a *Translation Learner Corpus,* as well as essays also written by university students which constitute a *Learner Corpus*. For extracting the data, the computing program *WordSmith Tools* (SCOTT, 2008) is used, enabling us to raise the most frequent collocational patterns employed by the learners. We also check frequency and recurrence of collocational patterns extracted as well as raise more collocations to be included in the workbook on *The Corpus of Contemporary American English* (DAVIES, 1990-2012). The proposal of designing an online workbook has the objective of contributing to Brazilian students' awareness of collocational aspects and mainly to the improvement of fluency in English.

Keywords: workbook; collocations corpus; pedagogical material; teaching collocations.

Recebido em: 14 de agosto de 2015.
Aprovado em: 20 de novembro de 2015.

1 Introdução

Segundo explica Kjellmer (1991, p.125), "o léxico mental de qualquer falante nativo é formado não apenas por unidades de palavras simples, mas também por combinatórias de palavras ou colocações".[1] Para o autor, "o domínio dos dois tipos de estrutura constitui parte essencial do equipamento linguístico do falante (seja na produção oral, seja na produção escrita) e lhe permite mover-se rapidamente, e com o mínimo de esforço, durante a sua exposição de uma estrutura pré-fabricada para outra"[2] (KJELLMER, 1991, p. 125). Nesselhauf (2005) explica que pode ser de interesse por parte de muitos falantes não-nativos e aprendizes de uma LE reduzir o esforço de processamento de unidades fraseológicas mais complexas, de modo a desenvolver a fluência, já que naturalmente necessitam reduzir este esforço, a fim de transmitir de modo mais claro sua mensagem. E, para isso, para desenvolver essa fluência, faz-se necessário ter controle de um grande repertório de unidades pré-fabricadas (WRAY, 2002).

Em nossa experiência de ensino de inglês como língua estrangeira (LE), deparávamos com muitas frustrações por parte dos alunos, os quais costumam dizer que, durante suas viagens ao exterior ou em qualquer situação de comunicação com um falante da língua inglesa durante suas reuniões de negócios, não conseguiam produzir a mesma linguagem desse falante, sentindo-se limitados linguisticamente, devido à dificuldade de combinar as palavras de maneira adequada. Por exemplo, podem conhecer a palavra *check*, porém, quando desejam utilizar uma unidade fraseológica, em inglês, correspondente a "fazer um cheque", geralmente não sabem qual verbo combinar. Dada a interferência da L1, arriscam-se a dizer *to make or to do a check*, em vez de *to write a check*.

---

1 "The mental lexicon of any native speaker contains single-word units as well as phrasal units or collocations". [Todas as traduções deste trabalho são de nossa responsabilidade.]

2 "Mastery of both types is an essential part of the linguist equipment of the speaker or writer and enables him to move swiftly and with little effort through his exposition from one prefabricated structure to the next".

Outra dificuldade dos aprendizes brasileiros ocorre em relação às colocações adjetivas. Neste trabalho, definimos colocações como combinações lexicais recorrentes e arbitrárias, que não são expressões idiomáticas, mas que o significado de uma de suas partes é contextualmente restrito àquela combinação específica, com base em Heylen e Maxwell (1994). Tomemos como exemplo o adjetivo *wet*. Caso desejem causar um efeito e usar, com o adjetivo "molhado" ou "encharcado", um advérbio intensificador como "extremamente", dificilmente produzirão a colocação *soaking wet* ou *dripping wet*. O mesmo ocorre com o adjetivo *naked*, cujo colocado, além de *completely* ou *entirely*, pode ser *stark* ou *buck*, empregado para causar maior efeito, formando a combinatória *stark naked* ou *buck naked*. Tal situação se dá, em virtude de a primeira atitude do aluno ser a de transferir a combinação da L1 para a LE, o que pode acarretar uma não compreensão por parte do falante da língua inglesa ou, muitas vezes, causar-lhe certa estranheza. Além disso, o sentimento de estar linguisticamente "limitado" pode apontar para a necessidade que alguns tipos de aprendizes têm de se sentir um membro de determinados grupos linguísticos, tais como os executivos acima mencionados, ou tradutores e professores em formação, que fazem parte do público-alvo de nossa investigação. Wray (2002, p. 75) corrobora essa afirmação ao mencionar que o conhecimento de uma grande gama de unidades pré-fabricadas, e nessa esfera incluem-se as colocações, servem para indicar que aquele indivíduo pertence a determinado grupo linguístico, ou seja, os aprendizes "satisfazem o desejo de soar [e escrever] como os demais falantes daquela comunidade".

Vale lembrar que estamos tratando, neste artigo, de um tema bastante específico (ensino de colocações) e de uma área igualmente específica (a Fraseologia) e que, embora tenhamos mencionamos o aprendizado de padrões léxico-gramaticais empregados por nativos da língua inglesa, seguramente não temos a intenção de desconsiderar pesquisas que abordam, por exemplo, o ensino de inglês como língua franca, que não tem como foco as normas da língua inglesa no processo de seu aprendizado e cujos propósitos de pesquisa, bem como público-alvo, parecem ser bastante diferentes de grande parte de nosso público-alvo. A própria concepção de Inglês como Língua Franca nos leva a

notar que também se trata de um ambiente de ensino e aprendizagem da língua inglesa relativamente diferente do nosso, a qual descrevemos a seguir, por meio da definição de Seidlhofer (2001, p. 146, *apud* SMIT, 2010, p. 46), sobre o que é inglês como língua franca no sentido estrito da palavra:

> um sistema adicional de aquisição de língua que serve como meio de comunicação entre falantes de diferentes L1 ou uma língua, por meio do qual os membros de diferentes comunidades linguísticas podem se comunicar, mas que não é a língua materna de nenhum dos falantes envolvidos.[3]

Observamos que, além do ambiente de ensino e aprendizagem de inglês, os próprios interesses e objetivos de grande parte de nosso público-alvo – professores em formação, professores de língua, bem como tradutores em formação ou tradutores profissionais – não se enquadram na concepção de inglês como língua franca e que, portanto, apesar de considerar sua relevância no ensino de inglês, não compartilharemos seus conceitos neste trabalho.

Ainda sobre as dificuldades relacionadas ao aprendizado de colocações, podemos mencionar o fato de que, na grande maioria das vezes, o aprendiz de uma LE não tem consciência de que essas combinações são fixas na língua. Em sua "ingênua" concepção de linguagem (FILLMORE, 1979), o aprendiz acredita que pode combinar as palavras de maneira livre e aleatória. Desse modo, um aprendiz de uma LE poderá ter dificuldades em produzir algumas colocações – a não ser que a colocação correspondente na LE seja literal–, caso não a tenha aprendido de modo explícito ou não haja observado seu uso antes, ou se não tiver sido criadas oportunidades para maximizar oportunidades de aprendizado de modo implícito, já que sua combinabilidade não se dá por critérios semânticos, mas por critérios arbitrários, ditados pela

[3] "(…) an additionally acquired language system that serves as a means of communication between speakers of different first languages or a language by means of which the members of different speech communities can communicate with each other but which is not the native language of either."

convencionalidade na língua. Hill (2000) igualmente trata dessa questão, ressaltando a importância de conscientizar os aprendizes da existência das colocações na língua, acrescentando a relevância de, ao propor atividades de conscientização, desenvolver também a autonomia no aprendizado das colocações, uma vez que seu processo de aprendizagem é bastante longo. Nesselhauf (2005) também aborda a não consciência (*unawareness*) dos aprendizes de notar a fixidez na combinabilidade dos elementos que formam uma colocação, assim como sua falta de consciência em relação ao aspecto convencional da língua, ao citar, por parte de aprendizes alemães, o modo exagerado, repetitivo e, na maioria das vezes, inapropriado, do adjetivo *big*, por exemplo, *play a big role* ou *take biggest pleasure out of*.

Pawley e Syder (1983, p. 200) relatam um fato que pode ilustrar uma situação na qual os aprendizes podem não ter aprendido unidades fraseológicas de modo mais explícito ou, ainda, que os alunos, mesmo depois de concluído alguns estágios de aprendizagem da língua inglesa, em sua concepção mais "ingênua", podem não ter percebido a fixidez de tais combinatórias: pode ser observado "nos estágios iniciais de colocar em prática os 'conhecimentos do livro didático' (não importando se o livro é bom ou se o estudo foi bastante profundo), é uma experiência bastante comum descobrirmos que a maior parte da produção dos aprendizes soa, aos ouvidos do nativo ou falante fluente da língua inglesa, como não idiomática"[4] – o termo *idiomático*, nesse caso, tem o significado de *natural*, *típico da linguagem*. Os autores (*Op. cit.)* também observam que, ao comparar a produção de falantes nativos moderadamente fluentes – "inseguros e hesitantes" – com a de aprendizes moderadamente fluentes dessa mesma língua, os falantes nativos fazem consideráveis pausas de hesitação entre sequências de palavras razoavelmente longas, enquanto os aprendizes o fazem depois de cada duas ou três palavras. Esse cenário nos leva a entender que a desigualdade entre os dois grupos consiste em uma diferença na automação da combinabilidade das palavras e, desse modo, na automação das colocações. Dessa forma, Pawley e Syder notam que,

[4] "In the early stages of putting one's 'book knowledge' into practice (no matter how good the book and how diligent the study of it), it is common experience to find that most of one's productions are, to the native ear, unidiomatic".

embora os alunos tenham sido expostos, durante seu processo aprendizado, a uma gama de unidades fraseológicas, tais como as colocações, se não houver instâncias de um ensino explícito dessas unidades, auxiliando-os a ter mais consciência de como as palavras combinam, quais são seus elementos, bem como de seu aspecto fixo e convencional na língua, poderá não haver produção oral nem escrita mais fluente ou natural.

Nesse sentido, podemos notar que, quanto maior o repertório de colocações aprendidas, mais a comunicação pode ser agilizada. Esse domínio facilita a comunicação, já que, uma vez adquiridas e armazenadas, o discurso dá-se com mais rapidez e fluência. O aprendiz, ao utilizar uma colocação ou unidade multipalavra não esperada pelo falante fluente da língua inglesa, provocará um desvio em seu discurso. Esse tipo de desvio pode ocorrer até mesmo em uma conversa entre nativos, porém, com frequência muito mais baixa. A quebra na convencionalidade na própria L1 também causa estranheza ao ouvinte. No entanto, muitas vezes essa quebra é proposital e pode ser um recurso muito utilizado em propagandas, ou como uma forma de humor e, somente poderá ser compreendida, se o aprendiz conhecer a combinatória, a fim de que perceba como se deu a quebra e entender o humor da frase, fato este imprescindível para um tradutor ou um intérprete. Por exemplo, um falante fluente da língua inglesa, ao comentar a respeito de uma tradução que apresentou sérios erros, diz, em tom de humor, que o tradutor *committed a translation* (*cometeu uma tradução*), em vez de usar a colocação correta (*did a translation*), querendo dizer que o tradutor "acabou" com o texto traduzido, fazendo uma alusão ao verbo *to commit* em inglês, que possui uma prosódia semântica negativa. Conforme dissemos, o aprendiz, o tradutor e o professor têm de conhecer os padrões da língua inglesa, para que compreendam a quebra colocacional, cujo intuito era exclusivamente de causar um impacto ou criar uma situação de humor.

De acordo com Sinclair (1991), se escolhemos formas não esperadas, a atenção do receptor será direcionada para a forma, e não para o conteúdo. O mesmo não ocorrerá se optarmos por uma forma pré-fabricada, que não prejudica a concentração do receptor. A respeito desse assunto, cabe mencionar a pesquisa realizada por Altenberg e Eeg-

Olofsson (1990, p. 2) sobre a compreensão e produção de discurso, tendo como perguntas de pesquisa: "Como falantes competentes de uma língua (ou gênero) conseguem falar de maneira mais fluente do que (ao que parece) sua limitada capacidade de processamento possa permitir? Como intérpretes simultâneos conseguem, de modo análogo, tais proezas miraculosas?".[5] Altenberg e Eeg-Olofsson (1990) concluem que falantes envolvidos em uma interação espontânea, precisam constantemente de expressões facilmente recuperadas do léxico mental e de grandes repertórios de "modos preferenciais" de se dizer as coisas. A esse respeito, Pawley e Syder (2000), Wray (2002) e Nesselhauf (2005) acrescentam que evidências psicolinguísticas mostram que o cérebro humano está muito mais apto a memorizar do que processar e que, desse modo, ter à sua disposição um grande número de unidades pré-fabricadas reduz seu esforço de processar e, portanto, favorece a fluência na língua.

Hausmann (1985), por sua vez, salienta que todo falante estrangeiro necessita conhecer as "palavras satélites" que gravitam em torno das básicas, a fim de entender o funcionamento típico da língua. Fontenelle (1994) igualmente acredita que os aprendizes de uma LE devam ser conscientizados a respeito das colocações e que nós, professores de uma LE, devemos ensiná-los a memorizar essas combinações fixas na língua. Dessa forma, de acordo com o autor, devemos conscientizar os aprendizes dessas restrições e ensiná-los a identificá-las e fazer uso dessas "expressões semicongeladas" que fazem parte da competência fraseológica do falante nativo.

Para justificar o que acabamos de mencionar, vale comentar o que Granger (1998b) descobriu após examinar um corpus de redações em inglês, escritas por falantes nativos da língua inglesa e falantes não nativos (falantes franceses, com nível de conhecimento em inglês avançado). Após aplicar um teste de combinações de palavras a 112 informantes, a autora constatou que aprendizes avançados têm uma consciência muito vaga do que é uma colocação convencional. Raramente detectam desvios de padrões colocacionais. A autora

[5] "How do competent speakers of a language (or genre) manage to talk more fluently than (it appears) their limited processing capacity should permit? How do simultaneous interpreters achieve their seemingly miraculous feats?"

menciona que a colocação *fully different* foi rapidamente aceita. Porém, a combinatória *bitterly cold* foi, na maior parte, rejeitada. Observou também um sobreuso de advérbios, tais como *completely* e *totally,* por parte dos falantes não-nativos. Essa observação nos pareceu bastante pertinente, uma vez que chegamos a um resultado bastante semelhante com uma pesquisa feita com aprendizes de inglês brasileiros.[6] Por não saberem, ou por não terem consciência de que haja um advérbio específico para determinado verbo ou adjetivo, acabam por utilizar sempre os mesmos intensificadores. Os aprendizes brasileiros investigados empregam os advérbios *very* e *completely* com uma frequência bastante alta, o que, embora em alguns casos não seja errado, pode tornar o discurso enfadonho e pouco criativo.

A fim de comprovar a importância do ensino de colocações em aulas de inglês como LE, Bahns e Eldaw (1993), realizaram uma pesquisa com 58 alunos universitários que estudaram inglês por num período de 7 a 9 anos antes de entrar para a universidade. Como instrumento de pesquisa, foram aplicados um exercício de tradução e um exercício de preenchimento de lacunas. Chegaram à conclusão de que 1) os alunos não aprendem as colocações de forma implícita e que seu conhecimento colocacional não se desenvolverá da mesma forma que seu conhecimento vocabular; 2) o conhecimento a respeito das colocações é necessário para que o falante possa comunicar-se de modo mais eficiente, já que nem sempre é possível parafrasear as combinatórias; 3) é preciso, sim, ensinar as colocações, tendo como critério a seleção daquelas que não sejam facilmente parafraseadas.

Assim, apesar de comumente empregada na língua e de seu relevante papel no processo de aprendizagem, as colocações ainda são pouco exploradas no ensino de LE, conforme mostram diversas pesquisas na área. Devido à dificuldade na produção, e não na compreensão, o ensino explícito de colocações ou, pelo menos, a maximização de sua exposição ou seu aprendizado em sala de aula,

---

[6] Realizamos pesquisa semelhante mediante a análise de redações escritas por alunos universitários do Curso de Licenciatura em Letras e Bacharelado em Letras com Habilitação para Tradutor, que compõem o *Corpus* de Aprendizes de Língua Escrita (CALE) e as traduções para o inglês que formam o *Corpus* de Aprendizes de Tradução (CAT), traduzidos por alunos do Curso de Tradução.

mostra-se fundamental para um aprendiz que deseja ser fluente em inglês, bem como conhecer diferentes possibilidades de se expressar de modo mais natural, preciso e, em alguns casos, criativo. Em virtude de sua relevância no contexto de ensino-aprendizado, torna-se imprescindível enfatizar seu aprendizado sistemático, por meio de um material didático voltado especificamente para tais combinatórias e, portanto, contribuir para a conscientização do aspecto convencional e colocacional da língua e para a produção de um discurso mais fluente.

Em pesquisa realizada na instituição de origem (Projeto Trienal), durante o triênio 2010-2013, analisamos as escolhas colocacionais, a influência da L1 nessas escolhas, bem como as quebras colocacionais presentes nas traduções e nos textos escritos (redações) dos alunos do Curso de Letras e de Tradução, que respectivamente compõem o *Corpus* de Aprendizes de Língua Escrita (CALE) e o *Corpus* de Aprendizes de Tradução (CAT), com o auxílio das ferramentas dos programas computacionais *WordSmith Tools* (SCOTT, 2008) e *AntConc* (ANTHONY, 2012). Em seguida, comparamos as opções colocacionais sugeridas pelos alunos com as colocações mais frequentemente empregadas por falantes fluentes da língua inglesa, constantes nos *corpora on-line British National Corpus*[7] e *Corpus of Contemporary American English*,[8] que representam a língua em seu contexto real de uso e mais comumente empregada por falantes fluentes da língua inglesa. Com base nas dificuldades colocacionais enfrentadas pelos alunos, selecionamos algumas colocações que deverão fazer parte do material proposto e, dessa maneira, atender às necessidades de brasileiros aprendizes de inglês como LE, como também sanar suas dificuldades, possibilitando um aprendizado sistemático dessas unidades fraseológicas.

Além de começar das palavras-chave e colocações extraídas dos *corpora* de aprendizes (CALE e CAT), visamos selecionar e analisar palavras com base na lista dos lemas mais frequentes (60.000 e 100.000)

---

[7] Disponível em: <http://www.natcorp.ox.ac.uk/>, o *British National Corpus* é composto de 100 milhões de palavras, das quais disponibiliza 50 milhões na *Internet*.

[8] Disponível em: <http://corpus.byu.edu/coca/>, o *Corpus of Contemporary American English* é formado por 450 milhões de palavras.

no *The Corpus of Contemporary American English* (*COCA*), com vistas a, com base nesse levantamento, extrair outros padrões colocacionais nos *corpora* de referência acima citados. O propósito dessa seleção e análise é garantir que um maior número de colocações frequentemente empregadas por falantes da língua inglesa estejam inseridas no *workbook*.

Nesse sentido, o objetivo principal deste artigo é tratar da compilação e da relevância do *Online English Collocations Workbook*, baseado em *corpus*, com foco no ensino das colocações (adjetivas, verbais, nominais e adverbiais) mais frequentemente empregadas na língua inglesa, por meio de atividades e jogos interativos. Dada a escassez de materiais de apoio que contemplem unidades fraseológicas do tipo colocação, a obra proposta busca também preencher essa lacuna. Além disso, vale mencionar que poderá ser utilizada não apenas por alunos da instituição em que atuamos mas também por aprendizes de todo o país, principalmente pelo fato de ser acessada via *Internet*, caracterizando novamente a proposta como inédita tanto em âmbito nacional quanto, de certo modo, internacional.

Por questões de delimitações orçamentárias, esta pesquisa contempla a primeira e a segunda fase de elaboração do *workbook*, disponibilizando, assim, o *Memory Game* (Jogo da Memória) e a atividade *Gap Fill* ("Completar espaços"), que consiste em completar frases ou trechos de textos por meio de colocações pré-definidas. Como fase seguinte, pretendemos, em pesquisa futura, adicionar outros tipos de jogos, a fim de oferecer diferentes possibilidades de memorização das colocações, também buscando atender a diversos estilos e formas de aprendizagem.

2 Fundamentação Teórica

Neste capítulo, abordamos os conceitos teóricos que embasam o desenvolvimento da pesquisa. Tratamos da convencionalidade e de sua importância na compreensão dos fraseologismos. De modo mais específico, apresentamos o conceito de colocação, algumas de suas características, bem como sua taxonomia. Além disso, faz-se necessário

discutir a conceituação de *corpus* e *corpora* de aprendizes, de sua definição e importância para o ensino de línguas e para a tradução.

2.1 Convencionalidade, fraseologismos e colocações

O termo "convencionalidade" na língua está relacionado ao uso que fazemos, ao interagir, de uma série de convenções, expressões e blocos de palavras já preestabelecidos e consagrados em nossa comunidade. É por isso que, quando esbarramos em alguém, por exemplo, dizemos "(me) desculpe". Essa expressão faz parte de um costume social consagrado e previsível; não fazer uso dela, em tal situação, implicaria falta de educação, pois uma pessoa considerada educada pediria desculpas automaticamente, segundo rezam as regras e convenções sociais já estabelecidas.

Outro exemplo de convencionalidade na língua é a expressão idiomática "Matei dois coelhos com uma cajadada só". Logo vem à mente de nosso interlocutor / falante nativo a ideia de que conseguimos realizar duas coisas ao mesmo tempo. Não passa pela cabeça de um falante nativo que, literalmente, tenhamos matado dois coelhos com uma cajadada, ou seja, não há uma relação analógica entre o significante e o real significado dessa expressão, uma vez que seu significado metafórico já está cristalizado na língua. Observamos, desse modo, um exemplo de convencionalidade em uma situação social, no primeiro caso, e em uma situação linguística, no segundo.

A convencionalidade também está ligada ao aspecto da arbitrariedade na língua. Ao empregar a expressão "Matar dois coelhos com uma cajadada só", usamos, em português, um animal – o "coelho". Em inglês, há uma expressão equivalente: *To kill two birds with one stone*, em que o "coelho" é substituído por uma ave – o "pássaro". Observamos que até mesmo a forma de matar é diferente nas duas línguas: em português com uma "cajadada" e, em inglês, com uma "pedra". O mesmo ocorre com a expressão "trabalhar como um burro", no português, e *to work like a dog* ou *to work like a horse*, em inglês. O animal associado ao trabalho árduo em português é o "burro" e, no inglês, o "cavalo" ou o "cachorro". Por esses exemplos, nota-se que tais

escolhas deram-se de maneira arbitrária e, assim, ficaram convencionalizadas nas duas línguas.

O termo convencionalidade foi empregado por Fillmore (1979) para designar "o conjunto dos elementos linguísticos, cuja coocorrência não é explicada sintática ou semanticamente, mas, sim, pelo uso". Fillmore (1979) também chama a atenção para o fato de a convencionalidade estar intimamente relacionada à fluência em uma língua, ou seja, o desconhecimento de unidades convencionais pode fazer com que o aprendiz não se comunique de maneira eficiente. A convencionalidade está relacionada aos usos e costumes sociais já preestabelecidos e consagrados pela comunidade. Está igualmente associada às normas de procedimento em determinadas situações e à linguagem dentro dessa mesma comunidade.

Ao tratar da convencionalidade, cabe discutir os fraseologismos na língua e no campo do aprendizado de uma LE. Fillmore (1979, p. 66) apresenta o aprendiz de uma LE como o "falante ingênuo", ou seja, aquele falante que desconhece o fator convencionalidade na língua. Consoante o autor, "[...] o ouvinte / falante ingênuo não conhece expressões idiomáticas lexicais e expressões idiomáticas frasais, colocações lexicais, fórmulas situacionais, comunicação indireta, ou as estruturas esperadas de textos de um dado tipo".[9] Ou seja, esse aprendiz faz apenas uma leitura composicional e não idiomática das estruturas linguísticas da língua que está aprendendo, podendo comprometer sua compreensão e produção – oral ou escrita.

Tagnin (2002) expande a noção de Fillmore (1979) ao tradutor, defendendo que "tradutor ingênuo" é aquele que desconhece a convencionalidade de uma língua, de modo que não é capaz de detectar sua ocorrência no texto que traduz, deixando, dessa maneira, de recuperá-la no texto traduzido. Ou seja, o "tradutor ingênuo" também pode fazer uma leitura composicional e não idiomática das estruturas linguísticas do texto de partida, o que, para a tradução de documentos jurídicos, pode comprometer a compreensão do leitor ou ainda ter consequências mais sérias. De acordo com a autora, o tradutor, ao se

[9] "[…] the innocent speaker/hearer does not know lexical idioms, phrasal idioms, lexical collocations, situational formulas, indirect communication, or the expected structures of texts of given types".

prender ao texto fonte, pode não notar que, entre várias formas gramaticais, haja uma opção preferencial. Caso essa escolha não seja a mais adequada na língua de chegada, sua tradução pode não soar "natural".

No que tange aos fraseologismos sob o ponto de vista cognitivo, é sabido que o falante nativo conta com um repertório mais ou menos fixo de expressões armazenadas em seu léxico mental (PAQUOT, 2008; NESSELHAUF, 2005; WRAY, 2002; TER-MINASOVA, 1992). Para simplificar a produção, ele as resgata de maneira automática como um bloco só – de acordo com o seu grau de competência linguística – e não lexema por lexema. Ou seja, não precisará produzi-las novamente no momento de seu discurso:

> As unidades fraseológicas têm a mesma função das palavras no discurso. No processo de fala, você não junta meramente as palavras separadas, em uma sucessão linear; você também usa unidades 'prontas', blocos pré-fabricados (i.e., unidades fraseológicas) que já existem na língua como um bloco inteiro e que funcionam, no discurso, como uma única palavra. (TER-MINASOVA, 1992, p. 535).[10]

Dessa maneira, o que parece ser espontâneo é, na verdade, uma forma estereotipada, fixa e repetitiva e, se o falante não tiver um repertório amplo ao seu dispor, seu discurso poderá ficar comprometido, no sentido de que não conseguirá se expressar, seja por não combinar as palavras conforme os padrões linguísticos da LE, por exemplo, durante uma entrevista ou reunião de negócios, seja pelo fato de ter que pausar muitas vezes durante sua fala, exatamente por não ter um grande reportório de padrões fraseológicos esperados para aquele contexto,

[10] "Phraseological units function in speech in the same way as words. In the process of speaking, one does not merely bring separate words together in linear succession; one also uses "ready-made" units, prefabricated blocks (i.e. phraseological units) that already exist in language as a global whole and function in speech as one word".

tendo que construir essas unidades fraseológicas no momento de sua fala.

Altenberg e Eeg-Olofsson (1990, p. 2) revelam que "as 'prefabs' desempenham um papel importante no discurso oral, pois podemos dizer que agem como uma espécie de 'piloto automático' que o falante pode ligar, a fim de ganhar tempo para se dedicar aos aspectos criativos e sociais do processo de fala".[11] Assim, aprendemos nossa L1 dessa maneira, ou seja, em blocos pré-fabricados, como combinatórias prontas que são produzidas de modo automático, sem reflexão, de forma inconsciente. Isso se faz necessário, pois, se a cada vez que discursássemos tivéssemos de formar novamente uma combinação de palavras, além do tempo que perderíamos para elaborar essa combinação, nosso discurso se tornaria enfadonho. Sob o ponto de vista do ouvinte, como ele poderia decodificar um discurso se, a todo momento, o falante criasse novos blocos de palavras? Dessa maneira, o processo de comunicação se tornaria bastante cansativo, tanto por parte do falante quanto do ouvinte.

Observamos, dessa forma, que o ensino de unidades multipalavras e, mais especificamente, de colocações, é de fundamental importância para um falante que deseja ser fluente em uma LE, ou seja, é imprescindível enfatizar seu aprendizado sistemático:

> Uma língua que é colocacionalmente rica, é também mais precisa. Isso ocorre porque a maioria das palavras da língua inglesa – especialmente as mais comuns – adotam uma grande variedade de significados, alguns bem distintos e alguns que variam entre si por graus. O significado preciso em qualquer contexto é determinado por aquele contexto: pelas palavras que circundam e combinam com a palavra núcleo – pela colocação. Um aluno que escolher a melhor colocação se expressará de maneira muito mais clara e será capaz de transmitir não apenas

[11] "Prefabs play a major part in spontaneous spoken interaction, […] where they can be said to act as a kind of 'autopilot' which the speaker can switch on to gain time for the creative and social aspects of the speech process".

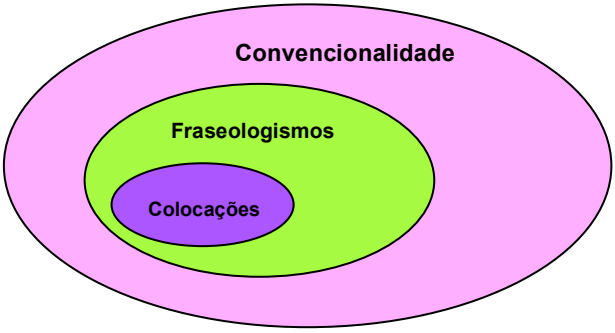
> um significado geral mas também algo bem preciso (*MCINTOSH*; *FRANCIS*; POOLE, 2009, p. v).[12]

Dessa maneira, em uma situação jurídica, por exemplo, o aprendiz deverá conhecer a colocação *mitigating factors* ou *mitigating circumstances* (em "circunstâncias atenuantes"), já que se trata de colocações bastante específicas, cujo colocado (*mitigating*) raramente coocorre com outras palavras. Isso se deve, pois, se tentar explicar somente que foi citado um fato que abranda a culpa do réu, seu discurso não terá o mesmo impacto e especificidade. Da mesma forma que, se quiser dizer que "o local estava um breu", de tão escuro que estava, e empregar somente *the place* was *very dark*, poderá não ser tão específico ou claro caso tivesse empregado a colocação *pitch dark* (*the place was pitch dark*).

Com base nesse cenário, passamos a tratar das colocações sob a ótica da Fraseologia. Trata-sede um dos tipos de fraseologismos que, por sua vez, estão inseridos no âmbito da convencionalidade, consoante a figura abaixo:

[12] "Language that is collocationally rich is also more precise. This is because most single words in the English language – especially the more common words – embrace a whole range of meanings, some quite distinct, and some that shade into each other by degrees. The precise meaning in any context is determined by that context: by the words that surround and combine with the core word – by collocation. A student who chooses the best collocation will express himself much more clearly and be able to convey not just a general meaning, but something quite precise".

Figura 1 - Convencionalidade, Fraseologismos e Colocações



Fonte: ORENHA-OTTAIANO, 2004.

Estando no âmbito da convencionalidade, as colocações, bem como alguns tipos de fraseologismos, representam um problema de produção, e não de compreensão, ou seja, é perfeitamente possível compreendê-las, na maioria dos casos, como em *place an order* e *pay a compliment* – desde que o aluno conheça o significado de *order* e *compliment*. Entretanto, um falante não nativo poderia ter mais dificuldade para produzi-las e muitas vezes, influenciado por sua língua-mãe, segundo também apontam pesquisas de Nesselhauf (2005) e Paquot (2008), acabaria cometendo quebras colocacionais, como faz ao produzir a colocação "fazer um pedido" por *make an order* e "fazer um elogio" por *make a compliment*, compreensíveis por parte de um falante fluente da língua inglesa, embora possam soar um tanto estranhas. Diferentemente de uma expressão idiomática, que esse mesmo falante não nativo poderia ter grande dificuldade não só de produzi-la mas também de compreendê-la, como em: *Peter is a bit thin on the top*. Se o aprendiz ainda não tiver se deparado com essa expressão, tampouco observar o contexto em que estiver inserida, provavelmente não compreenderá seu significado: "Peter é careca".

De acordo com Firth (1957), colocações são palavras que mostram uma coocorrência habitual sob o ponto de vista sintagmático. Halliday (1961, p. 75), discípulo de Firth (1957), apresenta uma definição bastante semelhante ao dizer que "uma colocação é uma

associação sintagmática de itens lexicais [...]". Halliday (1961), assim como Firth (1957), salientou a importância de um nível lexical distinto do gramatical, para explicar aspectos que a gramática não era capaz de explicar. O autor menciona que podemos dizer tanto *strong argument* (= "forte argumento") quanto *powerful argument* (= "poderoso argumento"). No entanto, dizemos preferencialmente *strong tea* (relacionado à cor do chá) e *powerful car* (= "um carro potente"), enquanto *powerful tea* e *strong car* não são frequentes. A combinação *powerful tea* poderia até ser empregada no sentido de se tratar de um "chá poderoso", com poderes talvez curativos. Já o adjetivo *strong* combinado com *car* poderia se referir a uma característica física do "carro".

Mediante o exposto, buscamos tecer algumas considerações a respeito da convencionalidade na língua e apresentar alguns conceitos acerca das colocações. Na subseção seguinte, tratamos da definição, delimitação e taxonomia das colocações, sob a ótica de Hausmann (1984, 1985).

2.2 A Proposta de Hausmann: delimitação e taxonomia das colocações

Outro autor que muito contribuiu para a descrição e definição de colocações foi Hausmann (1985), cujas ideias ainda não foram condensadas em uma única obra. Vários autores alemães citam seus trabalhos, mas nossos comentários contaram com o apoio de artigos em inglês e em português.

Hausmann (1985) tem importância fundamental para esta pesquisa, por ter descrito as colocações sob pontos de vista que também abordamos em nosso trabalho, segundo nos mostra Heid *et al.* (1991, p. 12):

- a construção de entradas de dicionários (HAUSMANN,1988), e de dicionários de colocações especializados;

- a aprendizagem de vocabulário: a relevância das colocações para o aprendizado e os aspectos pedagógicos das colocações em dicionários (HAUSMANN, 1984); e
- o estudo do fenômeno das colocações e suas implicações gerais para a Lexicografia (HAUSMANN, 1985).

O pesquisador faz uma distinção básica entre combinações fixas e não-fixas. As combinações fixas são as expressões idiomáticas e os compostos ("mercado negro", por exemplo, uma vez que são lexemas que correspondem a um conceito). Já as combinações não-fixas referem-se à coocorrência dos signos.

O critério que distingue a coocorrência dos signos é a "afinidade", utilizada por Hausmann no sentido de disponibilidade da combinação, na medida que chama as colocações de "produtos semiacabados de uma língua" (*semi-finished products of a language*). Segundo o autor, o ponto essencial das colocações reside em seu "*status* de disponibilidade mental como um todo, como um bloco só, e não como uma criação produzida *ad hoc* por um falante"[13] (HAUSMANN, 1984, *apud* HEID *et al.*, 1991, p. 15).

Para Hausmann (1984), o aspecto da lexicalização é muito importante para a definição de colocação e sua inserção em obras lexicográficas. Segundo ele, os usuários de uma língua simplesmente "reutilizam" os "produtos semiacabados" da língua quando utilizam as colocações.

No que se refere aos dois elementos da colocação – a base e o colocado –, consoante a terminologia de Hausmann (1984), esses não possuem o mesmo *status* semântico na combinação. Na verdade, há uma hierarquia entre esses elementos, pois um determina e o outro é determinado. Aquele que determina é chamado de base, que é o elemento autônomo, enquanto o outro, o determinado, é chamado de colocado, que somente pode ser interpretado semanticamente quando na colocação. Simplificando, a base é aquilo que já sabemos e o colocado aquilo que

[13] "(…) status of mental disponibility as a whole, not as a creation produced ad hoc by a speaker".

estamos buscando, por exemplo, em *issue shares*, *shares* é a base e *issue* é o colocado.

Resumidamente, a base é:

- um elemento independente;
- semanticamente autônoma;
- traduzível, independentemente de seu uso na colocação; e
- determina padrões lexicais que podem combinar com ela.

Quanto ao colocado:

- funciona como um conceito modificador;
- é semanticamente interpretável somente dentro da colocação;
- sua tradução depende do uso na colocação;
- é escolhido por uma dada base para formar uma colocação (HEID *et al*., 1991).

No que se refere à taxonomia das colocações, Hausmann (1985) sugere uma classificação, a qual apresentamos abaixo, com exemplos tirados dos *corpora* de nossa pesquisa:

VERBAIS – com quatro formas básicas:

- Verbo **colocado** + Substantivo **base**: *acquire shares*; *develop an illness*
- Substantivo **base** + Verbo **colocado**: *investments dropped*
- Verbo **colocado** + Preposição + Substantivo **base**: *dispose of shares*;

- Verbo **colocado** + Partícula Adverbial[14] + Substantivo **base**: *set up a business*
- Verbo **colocado** + Adjetivo **base**: *grow strong*

NOMINAIS – com duas formas básicas:

- Substantivo **base** + Substantivo **colocado**: *share subscription*;
- Substantivo **colocado** + Preposição + Substantivo **base**: *holder of shares; an attack of flu*

ADJETIVAS – com uma forma:

- Adjetivo **colocado** + Substantivo **base**: *bearer shares*; *life-threatening illness*

ADVERBIAIS – com três formas básicas:

- Advérbio **colocado** + Adjetivo **base**: *fully eligible*
- Verbo **base** + Advérbio **colocado**: *drop dramatically*
- Advérbio **colocado** + Verbo **base**: *fully paid*; *duly appointed*

Conforme mencionamos, o *workbook* tratará de todos os tipos de colocação.

[14] No caso dos *phrasal verbs* em inglês.

2.3 *Corpus* e *Corpora* de Aprendizes: definição e importância para o ensino de línguas e para a tradução

Com referência ao termo *corpus*, objeto da Linguística de *Corpus*, Berber Sardinha (2004, p.18) define-o 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 análise (BERBER SARDINHA, 2004, p.18).

Tognini-Bonelli (2001, p. 55) sugere uma definição que se mostra bastante coincidente com a definição anterior. Segundo a autora, "(…) um *corpus* é uma coletânea de textos autênticos e computadorizados, passível de análise ou processamento automático ou semi-automático".[6] Ela acrescenta ainda que os textos são selecionados de acordo com critérios explícitos, a fim de apreender as regularidades de uma língua, de uma variedade de língua, ou de uma sub-língua.

Na área de ensino-aprendizagem de línguas estrangeiras, são inegáveis os benefícios advindos do uso de *corpora* em sala de aula, sem a pretensão de descartar outras metodologias, abordagens e ferramentas já utilizadas. Segundo defende Bernardini (2004, p. 31-32), "(...) *corpora* e ferramentas de análise de *corpus* podem oferecer um dos instrumentais mais potentes disponíveis até hoje para atividades de aprendizagem para

---

[6] "(...) a corpus is taken to be a computerized collection of authentic texts, amenable to automatic or semi-automatic processing or analysis".

descoberta em sala de aula".[15] Dessa forma, o uso de *corpora* em sala de aula pode facilitar o aprendizado de uma LE, uma vez que permite o aluno observar os itens lexicais em contexto e o insere no processo ensino-aprendizagem.

No tocante a *corpora* eletrônicos de aprendizes, de acordo com Granger, Hung e Petch-Tyson (2002, p. 7), referem-se a "coletâneas eletrônicas de dados textuais em uma segunda língua ou em uma LE, compilados de acordo com critérios de desenho explícitos, a fim de atender a um propósito específico, no que se refere à aquisição de segunda língua ou ao ensino de LE".[16] Cabe lembrar, que os referidos dados textuais podem ser formados por textos escritos ou orais (por meio de transcrições).

Para inserir um texto em um *corpus* de aprendizes, é necessário seguir uma série de critérios, previamente estabelecidos, para que todos os textos que compõem este *corpus* apresentem uma homogeneidade quanto ao gênero textual, à temática, à extensão dos textos e ao nível de língua. De acordo com o projeto ICLE (*International Corpus of Learner English*), coordenado por Granger (1993), por exemplo, os textos dos aprendizes deverão ser, obrigatoriamente, argumentativos. Há também alguns temas, entre os quais deverão escolher, tais como: *Crime does not pay*, *Feminism has done more harm to the cause of women than good*, *Pollution: a silent conspiracy* etc. Além disso, as redações deverão conter de 500 a, no máximo, 1.000 palavras, entre outros dados.

Entre os corpora de aprendizes, podemos citar o ICLE, coordenado por Granger (1993),[17] composto por redações de alunos de nível intermediário superior e avançado; o *Longman Learner's Corpus*, sob a supervisão de Summers (1999); o *Cambridge Learners' Corpus*, sob a responsabilidade de Nicholls (1999); e o *Multilingual Learner*

---

[15] "(…) corpora and corpus analysis tools seem to provide one of the most powerful tools made available to date for classroom discovery learning activities".

[16] "Computer learner corpora are electronic collections of authentic FL/SL textual data assembled according to explicit design criteria for a particular SLA/FLT purpose".

[17] O projeto ICLE abarca vários corpora, compilados em vários países (um total de 21 países, conforme informação no *site* <http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm>). No Brasil, a vertente do ICLE é o Br-ICLE, coordenado pelo Prof. Tony Berber Sardinha, no LAEL, Pontifícia Universidade Católica de São Paulo (PUC/SP).

Corpus, coordenado pela Profa. Stella E. O. Tagnin, da Universidade de São Paulo (USP), entre vários outros.

No que tange ao uso de *corpora* de aprendizes para o ensino de LE, pesquisadores da área (GRANGER, HUNG, PETCH-TYSON, 2002; BERBER SARDINHA, 2001; TONO, 1999; GRANGER, 1998a) destacam várias aplicações como, por exemplo, a descrição da interlíngua e a preparação de materiais didáticos. Para a área da Lexicografia, a exploração de corpora de aprendizes pode favorecer a compilação de dicionários, de modo geral. Berber Sardinha (2001) também faz referência às vantagens oferecidas por um corpus de aprendizes, no que diz respeito à identificação das áreas em que os aprendizes tendem a usar "estratégias de evitação" e, por conseguinte, deixam de explorar profundamente o potencial da língua alvo.

É possível, também, investigar assuntos relacionados a "*foreign soundingness*", conceito dado por Pravec (2002), na escrita de aprendizes de inglês como LE (ou qualquer outra língua), que pode ser detectado ao levantar e analisar, por exemplo, quais estruturas gramaticais, linguísticas, lexicais, discursivas ou pragmáticas são empregadas de forma restrita (subuso) ou excessiva (sobreuso) por parte dos referidos aprendizes, em relação a falantes nativos.

No que se refere à influência da L1 na produção escrita ou oral dos aprendizes de uma LE, por exemplo, a questão da transferência, os *corpora* de aprendizes podem contribuir de maneira significativa, conforme aponta Tono (1999). De modo mais específico, pesquisadores como De Cock (1998), Granger (1998b), Granger, Hung e Petch-Tyson (2002), Paquot (2008), Nesselhauf (2005), entre outros, tratam da influência da L1 na produção de unidades multipalavras (*multiword sequences*). Nesta pesquisa, será possível observar a influência da L1 na produção da unidade multipalavra do tipo colocação na língua inglesa.

Além disso, *corpora* de aprendizes possibilitam analisar o desempenho de um falante quase nativo em relação ao falante fluente da língua inglesa e a identificar dificuldades específicas do aprendiz de uma dada língua de partida. Leech (1998) também menciona que, por meio da exploração de *corpora* de aprendizes, é possível detectar quais são as áreas nas quais os aprendizes de um dado país parecem necessitar mais de ajuda para desenvolver sua produção na língua-alvo.

Segundo pode ser notado, os *corpora* de aprendizes, por fornecerem um instrumental eficaz de apresentação e exploração de dados, abriram novas perspectivas para investigações em ensino e aquisição de LE, possibilitando a pesquisadores e professores da área, diferentes maneiras de contribuir para uma melhoria no ensino-aprendizagem de LE.

Em relação a corpora de aprendizes de tradutor, são vários os autores que apontam sua relevância (BOWKER, 2000; BERBER SARDINHA, 2001; ORENHA-OTTAIANO, 2012, entre outros), no sentido de promover melhor compreensão do processo tradutório, bem como facilitar a tarefa do próprio aprendiz de tradutor na busca pela melhor opção tradutória. Segundo Orenha-Ottaiano (2012), o uso de *corpora* para a formação de tradutores tem se mostrado extremamente eficiente, principalmente, no que tange à conscientização do aspecto fraseológico e convencional da língua, assim como às especificidades da terminologia de determinadas áreas de especialidade.

3 Metodologia

Com base no arcabouço teórico relatado, esta seção trata da metodologia de pesquisa que está sendo empregada para a realização desta investigação. Abordaremos os procedimentos para a compilação dos *corpora* de aprendizes (CALE e CAT) e, na sequência, dos passos metodológicos para a extração das colocações, por meio do programa computacional *WordSmith Tools* (SCOTT, 2008), versão 5.0, e de novas combinatórias, com base na lista de frequência do *COCA* e com o auxílio do referido *corpus* de referência.

3.1 A Compilação do *Corpus* de Aprendizes de Tradução (CAT)

O CAT, por ser um *corpus* paralelo, é formado por textos originalmente escritos em inglês e seus respectivos textos traduzidos pelos alunos envolvidos na pesquisa (2º e 3º ano do Curso de Tradutor, em um total de 18 alunos), no Laboratório para o Ensino de Línguas Estrangeiras, em horário extraclasse. Os participantes tinham

aproximadamente duas horas para traduzir um texto de aproximadamente 800 palavras. Ao finalizar, enviavam as traduções para a pesquisadora em arquivos identificados.

Quanto à tipologia dos textos, trata-se de textos jornalísticos, retirados de jornais ou revistas de grande circulação no país, ou seja, todos os textos originais foram escritos em português e vertidos para o inglês. Em relação aos temas, foram selecionados tópicos discutidos internacionalmente, tais como: "Morte de Kadafi"; "Um ano depois do Terremoto no Japão"; "Crise financeira na Grécia e na Europa"; "Desemprego"; "Bullying"; "Legalização da maconha"; "Eleições nos Estados Unidos", "Aborto", "Eutanásia" etc. Outros gêneros de textos estão sendo inseridos ao *corpus* nos últimos meses, que também servirão de extração e posterior análise de colocações. Cabe lembrar, que tanto os textos originais quanto os textos traduzidos são salvos em extensão .txt, a fim de serem processados pelo programa *WordSmith Tools* (SCOTT, 2008).

3.2 A Compilação do Corpus de Aprendizes de Língua Escrita (CALE)

O CALE é formado pelas redações dos alunos dos cursos de Letras e Tradutor, segundo mencionado. Baseado no projeto ICLE, coordenado por Granger (1993), os textos dos aprendizes são, na sua maioria, argumentativos. Da mesma forma que o ICLE, há alguns temas que os alunos devem seguir para escrever seus textos, temas esses previamente determinados pela professora-pesquisadora, na expectativa de que tal seleção suscitasse ideias para a escrita dos textos. Alguns dos temas selecionados foram:

Quadro 1 - Temas para a elaboração de textos escritos

1º ano	2º ano	3º ano	4º ano
Only one child – lonely child?	*Stories about things that went wrong*	*Abortion*	*Genetically-modified food*
Friendship	*Longevity*	*Brain power*	*Global Warming*
Country life or city life?	*Marriage and tradition*	*Financial Crises in Europe*	*Addiction/Drugs*
Finding the ideal job	*One year after Japan's earthquake*	*Culture Shock*	*Euthanasia*
Staying healthy	*Unemployment*	*Bullying*	*Capital Punishment*
My last vacation	*Lies and Truth*	*Positive and negative effects of TV viewing*	*Advertisement*
	Happiness	*Stereotypes*	*Same-sex marriage*
		Religion	*Censorship*

Fonte: nossa autoria.

Além desses, outros temas da atualidade, presentes nos jornais, revistas e materiais didáticos, foram selecionados para esta pesquisa.

Quanto à extensão dos textos escritos, as redações deveriam conter, para este estudo:

- de 200 a 350 palavras para os alunos do primeiro ano de ambos os cursos;
- de 300 a 450 palavras para os alunos do segundo ano;
- de 400 a 650 palavras para os alunos do terceiro ano; e
- de 800 a 1.000 palavras para os alunos do quarto ano.

Os textos também foram salvos em formato .txt, e enviados pelos alunos à docente-pesquisadora, via *e-mail*. Coube à pesquisadora deixar claro para os alunos que suas redações seriam utilizadas para fins de pesquisa e que seus nomes seriam mantidos em sigilo. Dessa forma, os alunos assinaram um documento permitindo a utilização de seus textos para essa finalidade. Essa decisão está apoiada no texto da *British Association for Applied Linguistics* (THOMPSON, 2005), cuja sugestão é de que a autorização de dados seja gravada nos mesmos meios da coleta de dados: por vídeo ou áudio, ou, no caso do CALE e do CAT, na forma escrita.

3.3 Procedimentos para levantamento e análise das colocações para sua inserção no *Workbook*

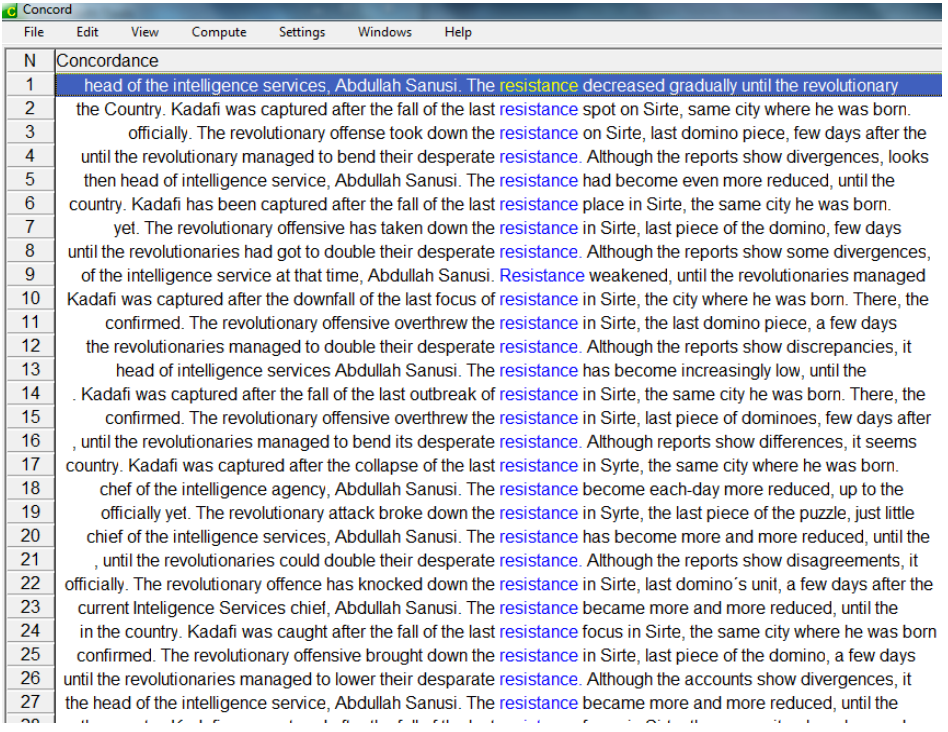
Por questões de delimitação, apresentamos, neste item, o levantamento de algumas colocações extraídas apenas do CAT, uma vez que, por conta desse levantamento, outras combinatórias serão extraídas do *COCA*, tomando como ponto de partida a base dessas colocações (HAUSMANN, 1985). De posse dessas colocações, estão sendo elaborados os exercícios ou jogos colocacionais, os quais empregam as referidas colocações extraídasA partir da observação de algumas palavras-chave do CAT, apresentadas no quadro abaixo, especificamente dos textos traduzidos para o inglês, selecionamos duas (*resistance* e *primary*) e, a partir desta seleção, geramos linhas de concordâncias, por meio da ferramenta *Concord*, para fins de levantamento e análise das colocações.

Quadro 5 - Quatorze primeiras palavras-chave do CAT

1. Primaries	8. Caucus
2. Dictator	9. Revolutionary
3. Resistance	10. Elections
4. Republican	11. Contribution
5. Priors	12. Announce
6. Marijuana	13. Leader
7. Delegates	14. Regime

A título de exemplificação, apresentamos as linhas de concordância do termo *resistance*:

Figura 2 - Linhas de concordância da palavra-chave *resistance*



A palavra-chave *resistance* foi traduzida, por exemplo, de um dos contextos dos textos originais: "A ofensiva revolucionária **derrubou a resistência** (...])". Para essa colocação em português, levantamos as colocações em inglês no CAT, abaixo elencadas na Tabela 1. A fim de verificar se as escolhas colocacionais feitas pelos alunos são recorrentes na língua inglesa, verificamos sua recorrência no *COCA*. No entanto, em virtude de serem colocações verbais muito específicas, de um contexto mais especializado, a frequência levantada foi muito baixa, não sendo possível afirmar se eram realmente frequentes na língua inglesa. Dessa maneira, essas colocações foram também investigadas na *Web*, por meio da ferramenta de busca *Google*. Vale mencionar, que as buscas foram

realizadas com os verbos no tempo verbal em que se encontravam os verbos no contexto analisado, ou seja, no passado (Simple past – Ex.: *took down the resistance*), com exceção da combinação *upend the resistance*, uma vez que preservamos a forma como foi utilizada na tradução, segundo apresentamos a seguir:

Tabela 1 - Lista dos possíveis colocados do nódulo *resistance*

RESISTANCE - CAT	RESISTANCE – Google
[...] took down the resistance	6
[...] overthrew the resistance	0
[...] broke down the resistance	212.000
[...] knocked down the resistance	6
[...] brought down the resistance	278.000
[...] knocked out the resistance	4
[...] dropped the resistance	152.000
[...] toppled the resistance	3
[...] upend the resistance	0
[...] took out the resistance	266.000

Conforme pode ser notado na Tabela 1, em razão de se tratar de um *corpus* paralelo, fica muito interessante verificar as diversas possibilidades tradutórias sugeridas pelos alunos: por meio de uma colocação em português ("derrubar a resistência"), tivemos 10 opções tradutórias dos 18 alunos envolvidos na pesquisa, ou seja, algumas dessas opções foram empregadas por mais de um aluno. Vale mencionar a rapidez com que essas colocações foram geradas pelo programa *WordSmith Tools*.

Podemos observar que 40% das possibilidades tradutórias acima elencadas são recorrentes em inglês e comumente empregadas nessa língua: *broke down the resistance* (212.000 ocorrências), *brought down the resistance* (278.000), dropped down the resistance (152.000) e *took out the resistance* (266.000 ocorrências). No entanto, as demais combinações empregadas pelos alunos de tradução parecem não ser recorrentes: *overthrew the resistance* (0 ocorrência no *Google*), *knocked*

down the resistance (6 ocorrências), *knocked out the resistance* (4 ocorrências), e *toppled the resistance* (3 ocorrências) e *upend the resistance* (0 ocorrência)

Esse resultado, somado a outros casos semelhantes levantados durante a análise das colocações baseada no CAT com base em outras palavras-chave, pode indicar que, mesmo aprendizes com nível de conhecimento de língua mais avançado, podem não ter consciência do aspecto colocacional da língua, considerando que os alunos investigados tiveram acesso à *Internet* e a dicionários (impressos e *on-line*) para realizar as traduções.

Os resultados dessa investigação serviram de motivação para a compilação do *workbook* proposto, evidenciando sua relevância para o ensino de inglês como LE. No item seguinte, trataremos dos aspectos relacionados à elaboração do *Online English Collocations Workbook*.

4 Da elaboração do *Online English Collocations* W*orkbook*

O objetivo principal de nossa investigação é a compilação de um *Online English Collocations Workbook,* contendo colocações mais frequentemente empregadas na língua inglesa, com base nas colocações levantadas nos *corpora* de aprendizes e no *corpus* de referência *on-line*, atendendo, desse modo, às necessidades dos brasileiros aprendizes de inglês como LE. O material de apoio pedagógico proposto apresenta-se como uma resposta às dificuldades enfrentadas por aprendizes brasileiros da língua inglesa, no que concerne ao aprendizado de colocações. Os exercícios que compõem o *workbook* são de cunho prático, visando auxiliar aos alunos incorporá-las a seu léxico mental, principalmente, por meio de jogos e de forma interativa e divertida.

Outro motivo que justifica sua compilação é a escassez de materiais de apoio que contemplem unidades fraseológicas do tipo colocação. Mackin (1978) já defendia a criação de materiais como o proposto:

> O que os professores de inglês como LE precisam, para um treinamento eficaz de competência

> colocacional (na condição de que [as turmas] sejam homogêneas em relação à L1), são livros de exercícios que contenham uma seleção de colocações direcionadas às dificuldades específicas dos aprendizes (...). Tais materiais permitiriam-nos realmente ensinar colocações e, dessa maneira, diminuir, pelo menos até certo ponto, o longo e árduo processo de aquisição da competência colocacional, durante anos de estudo, leitura e observação da língua (MACKIN, 1978, p. 151).[18]

Dessa maneira, é possível notar a relevância da obra proposta para alunos de inglês como LE. Esperamos, por meio desta pesquisa, poder contribuir para que os aprendizes atinjam o caminho da fluência, produzindo textos (orais ou escritos) mais idiomáticos e naturais.

Segundo citamos na introdução desta proposta, esta investigação pretende, devido à complexidade na elaboração dos exercícios colocacionais e da plataforma *on-line*, bem como a questões de delimitações orçamentárias, contemplar a primeira e a segunda fase de elaboração do *workbook*, disponibilizando, dessa maneira, o jogo da memória (*Memory Game*) e a atividade "Completar espaços" (*Gap Fill*), que consiste em completar frases ou trechos de textos por meio de colocações pré-definidas. Essa plataforma está sendo elaborada pelo Grupo de Banco de Dados (GBD), sob coordenação do Prof. Dr. Carlos Roberto Valência, do Departamento de Computação, da UNESP.

Vale mencionar que alguns dos exercícios já compilados (*Memory game* e *Gap fill*) já foram aplicados em sala de aula para os professores inscritos no Curso de Extensão (de Aperfeiçoamento) "Linguística de *Corpus* e Fraseologia aplicadas à prática pedagógica de Professores de Língua Inglesa da Rede Pública", com carga horária de 180 horas, de agosto de 2013 a julho de 2014. Este curso faz parte de um

---

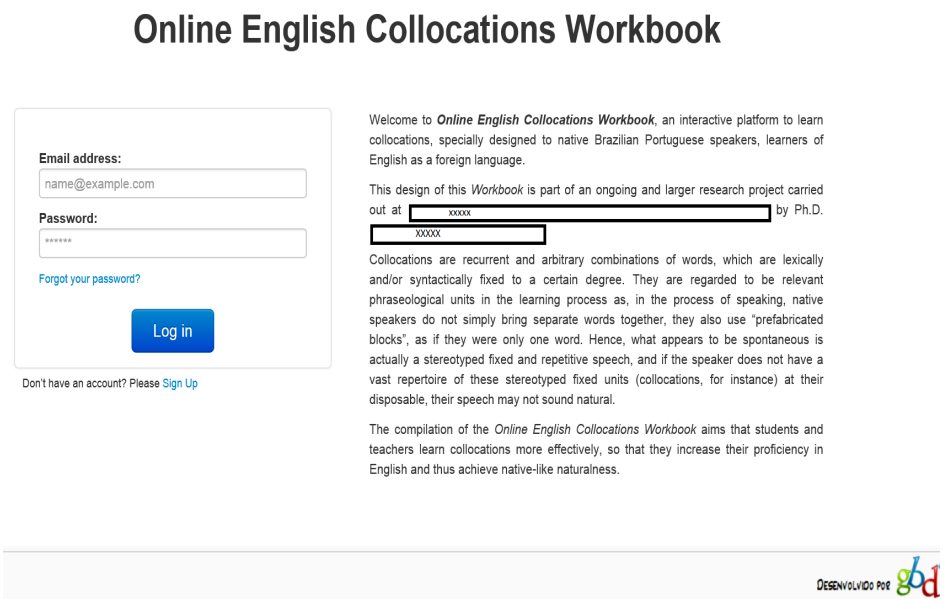
[18] "What EFL teachers need for an effective training of collocational competence in their classes (as long as these are homogenous with regard to the L1) are workbooks presenting a selection of collocations geared to the specific difficulties of learners (…). Such material would allow us to actually teach collocations and thus shorten, at least to a certain extent, the long and laborious process of acquiring collocational competence through years of study, reading, and observation of the language".

Projeto de Extensão de mesmo nome, aprovado pela PROEX em 2013, sob nossa coordenação. Além disso, os professores avaliaramm os exercícios e nos deram *feedback* referente à sua aplicabilidade e eficácia. Alguns desses professores, segundo consta, buscarão aplicar as atividades do *workbook* em suas aulas na rede pública e privada de ensino.

Para melhor ilustrar como está sendo realizada a compilação do *Online English Collocations Workbook*, apresentamos a seguir um dos jogos (*Memory Game*) e uma das atividades já disponíveis na plataforma do *Workbook* (*Gap Fill*) e algumas funcionalidades referentes à criação e ao gerenciamento dos jogos, gerenciamento de usuários, entre outros aspectos.

Ao acessar o material, o consulente abre a uma tela, explicando o que é o *Online English Collocations Workbook*, conforme figura abaixo:

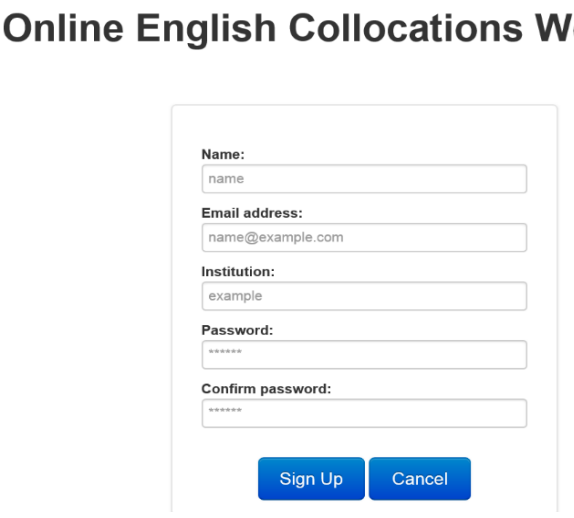
Figura 3 - Plataforma de acesso ao *Workbook on-line*



Para obter acesso ao *workbook*, haverá dois tipos de usuários: 1) a coordenadora do projeto, responsável pelo gerenciamento e cadastro de

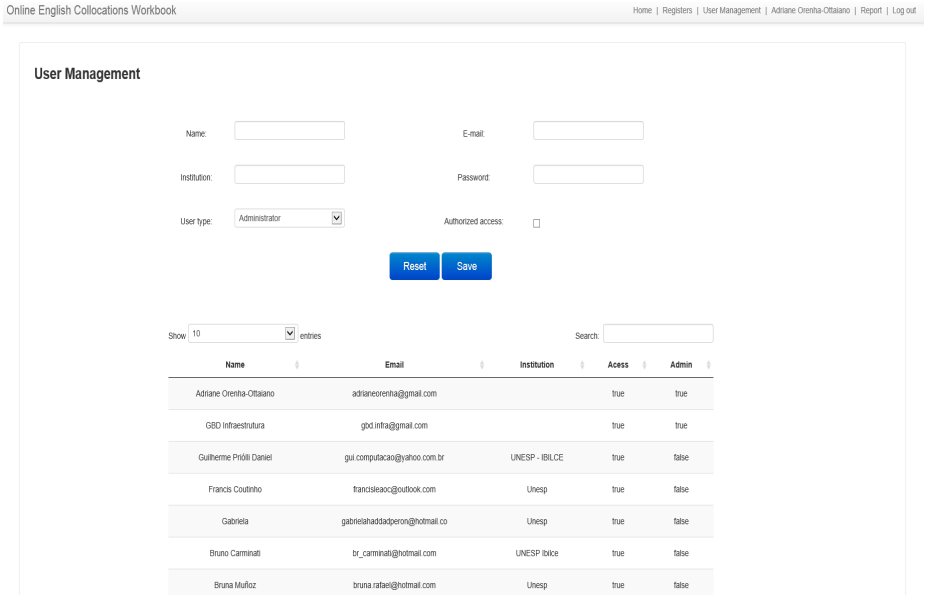
novos exercícios e pela liberação do acesso aos usuários e gerenciamento de seus dados; 2) o público de modo geral, que terá acesso ao sistema para fazer os exercícios e jogos de colocação, mediante cadastro. Desse modo, para acessar o *workbook,* o usuário deverá fazer um cadastro, conforme a figura a seguir:

Figura 4 - Tela de cadastro para acesso ao *Online English* Collocations *Workbook*



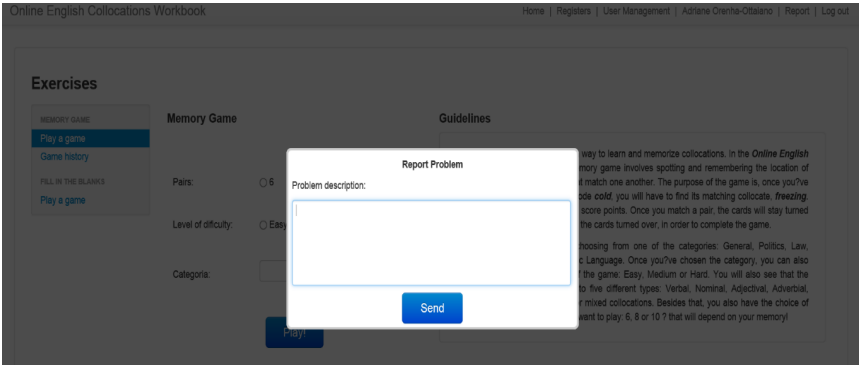
O administrador do sistema (a coordenadora do projeto) receberá a mensagem e poderá autorizar ou não o acesso daquele usuário à plataforma, além de gerenciar os usuários cadastrados, segundo tela abaixo:

Figura 5 - Tela do Administrador para gerenciamento dos usuários do *Workbook*



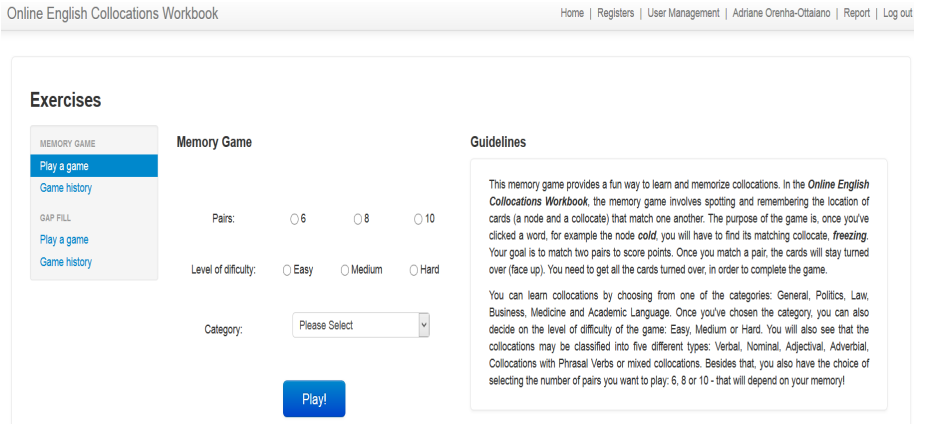
A figura a seguir mostra a aba *Report*, por meio da qual o usuário pode reportar problemas ou erros encontrados ao realizar os exercícios ou jogos e enviá-los para a equipe do GBD, que os discutirão com a coordenadora do projeto:

Figura 6 - Tela de teste do usuário para reportar problemas



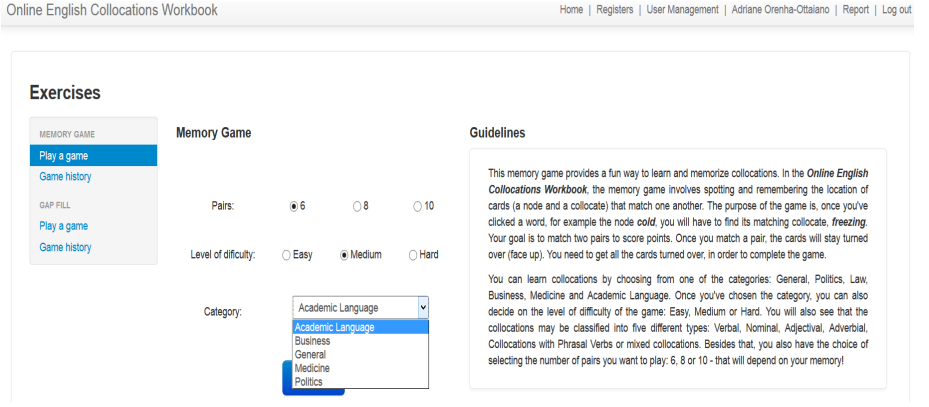
Durante as reuniões com GBD, no processo de definição e provimento de recursos das interfaces do sistema, ficou decidido que tanto o *Memory game* quanto o *Gap Fill* trarão atividades bastante interativas e com diferentes níveis de dificuldade (fácil, médio e difícil), segundo mostra a figura abaixo:

Figura 7 - Tela com os níveis de dificuldade das atividades colocacionais



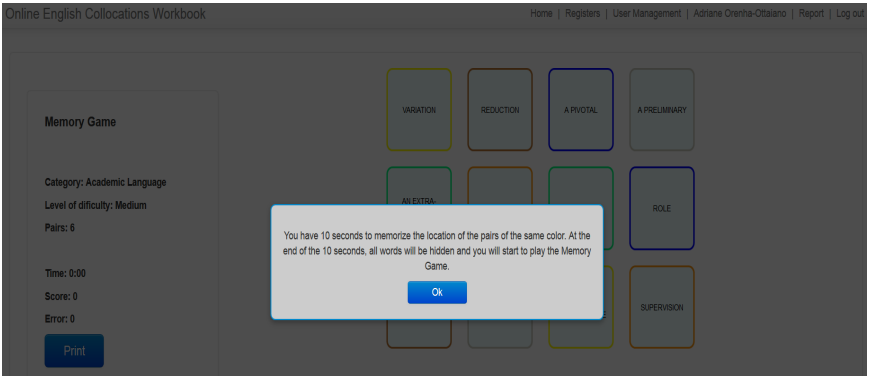
Além dos níveis, o usuário poderá optar pelo número de pares que gostaria de jogar (vai depender de sua memória!), e também por categorias das quais o usuário poderá selecionar *Academic Language*, *Business*, *General*, *Medicine* e *Politics*, a fim de trabalhar com colocações de diferentes áreas ou da língua geral (*General*). Futuramente, novas categorias serão inseridas:

Figura 8 - Tela com as categorias focadas no glossário



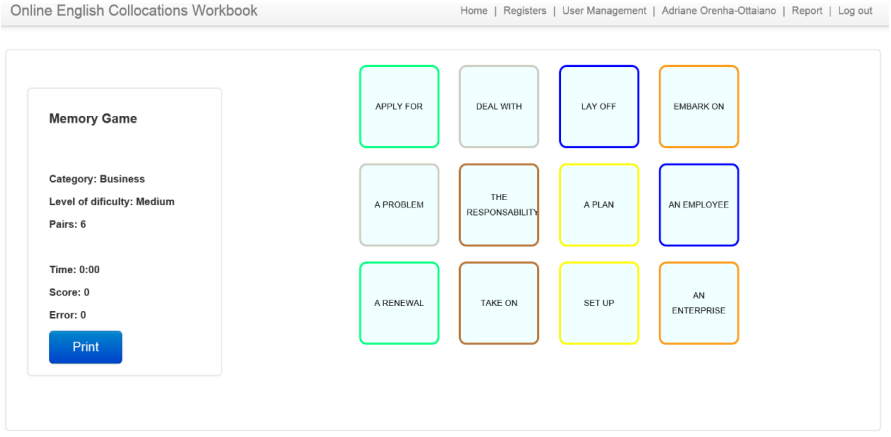
No que se refere ao *Memory Game*, ao clicar em *Play a game*, o usuário receberá uma mensagem de que terá 10 segundos para tentar memorizar os pares de palavras que formarão uma colocação e deverá clicar em OK para visualizá-las: *You have 10 seconds to memorize the location of the pairs of the same color. At the end of the 10 seconds, all words will be hidden and you will start to play the Memory Game.*

Figura 9 - Tela de aviso para memorização das palavras que formam as colocações



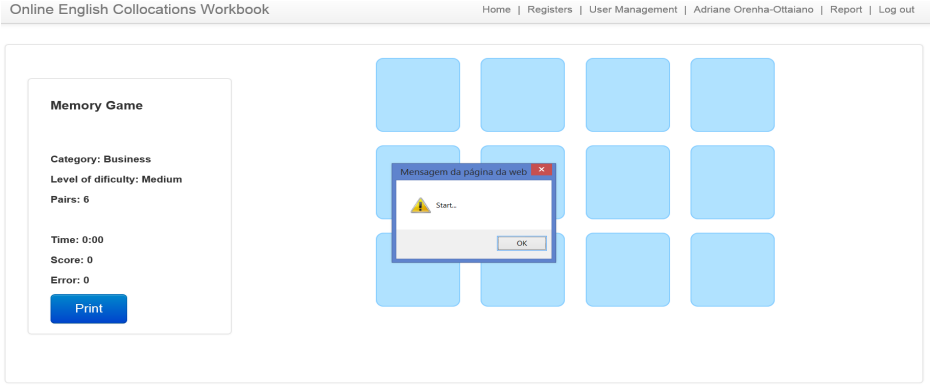
Abaixo, apresentamos a tela com todas as palavras que formarão colocações. Conforme pode ser notado, cada par de palavras está em uma cor, a fim de mostrar ao usuário a colocação formada e facilitar a memorização de sua localização no quadro:

Figura 10 - Tela para memorização das colocações



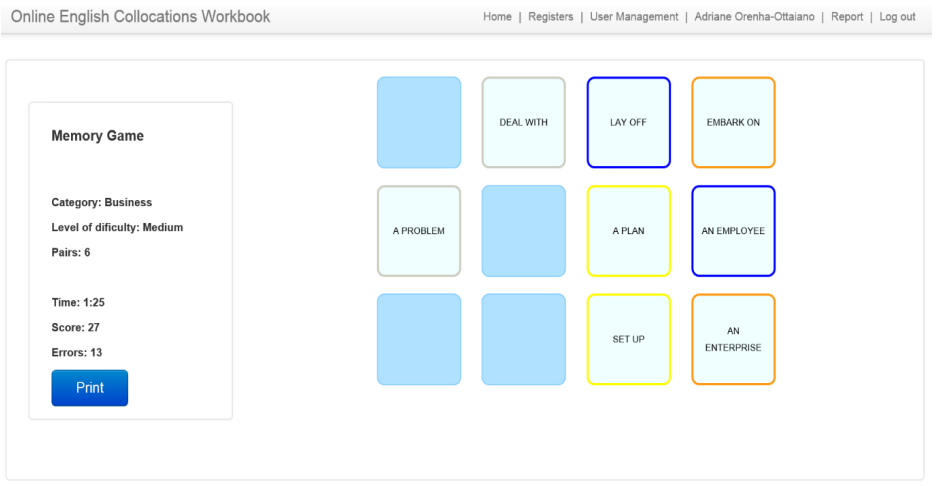
Quando o tempo dessa parte da atividade chega ao fim, aparece uma tela informando que o usuário poderá iniciar o jogo.

Figura 11 - Tela informando que jogo pode ser iniciado



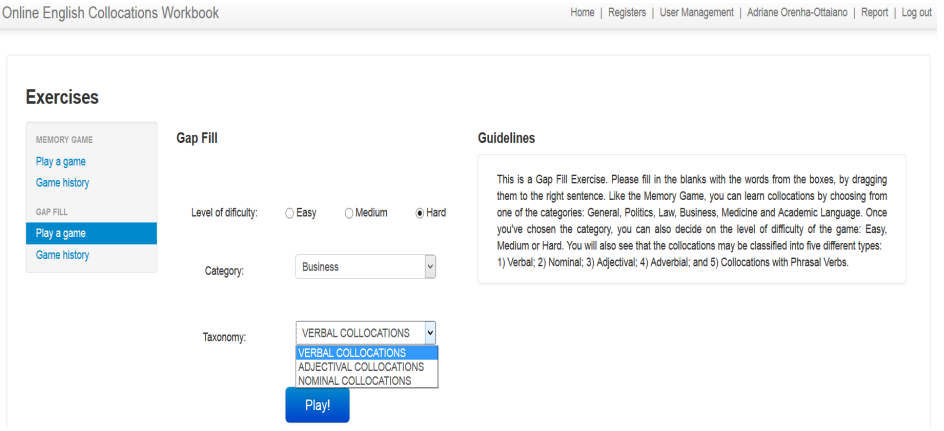
A figura a seguir mostra o usuário jogando o *Memory Game*, em busca dos pares que formarão as colocações. Observe que, no lado esquerdo da tela do jogo, há um resumo de escolhas que o usuário pode fazer: a categoria (*Category*), o nível de dificuldade (*Level of difficulty*), o número de pares selecionado (*Pairs*), bem como o tempo para realizar o jogo (*Time*), sua pontuação (*Score*) e o número de erros (*Errors*):

Figura 12 - Jogando o *Memory Game*



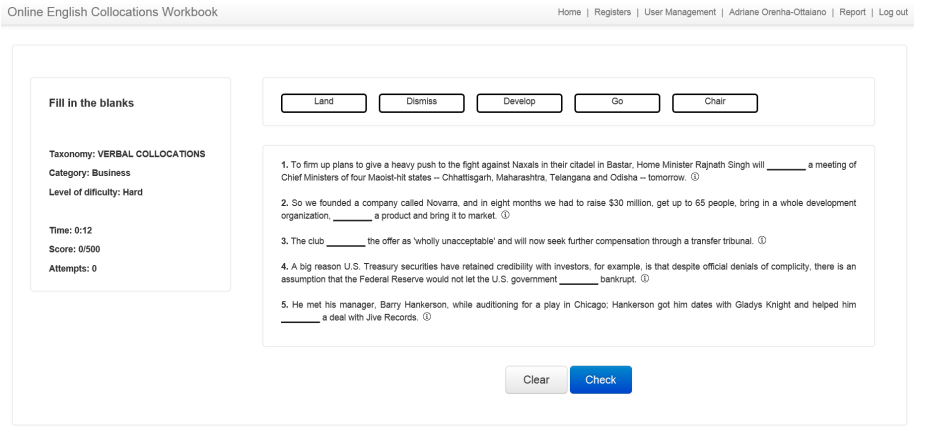
Na figura seguinte, apresentamos a tela do exercício *Gap fill*, seguindo o mesmo padrão do *Memory Game*. A única diferença é que ele já oferece ao usuário a opção de escolher, além dos níveis de dificuldade e da categoria, a tipologia das colocações (*taxonomy*) que gostaria de trabalhar: *Verbal*, *Nominal*, *Adjectival* e *Adverbial* – já providenciamos a correção dos referidos tipos de colocações na plataforma. Vale ressaltar que essa opção também será oferecida para o *Memory Game* em breve.

Figura 13 - Tela de seleção de itens para realizar o exercício *Gap Fill*



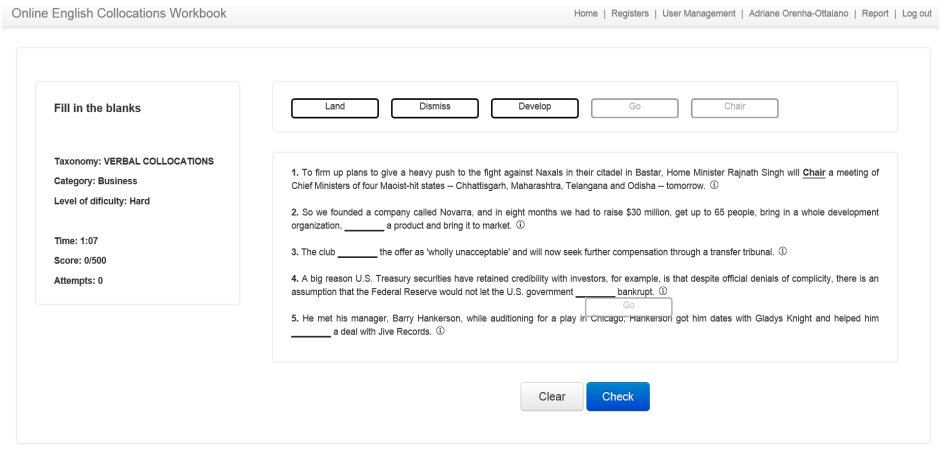
Na Figura 14, é possível ver uma tela em que o usuário irá iniciar o exercício *Gap Fill*. O usuário foi orientado a arrastar, neste caso, os verbos que formarão as colocações contextualizadas nos exercícios, para preencher as lacunas. Muitas das colocações que compõem os exercícios de exercício *Gap Fill* são trabalhadas no *Memory Game* e que, neste momento, o usuário tem a oportunidade de observá-las novamente em contexto.

Figura 14 - Tela com exemplo de um exercício do tipo exercício *Gap Fill*



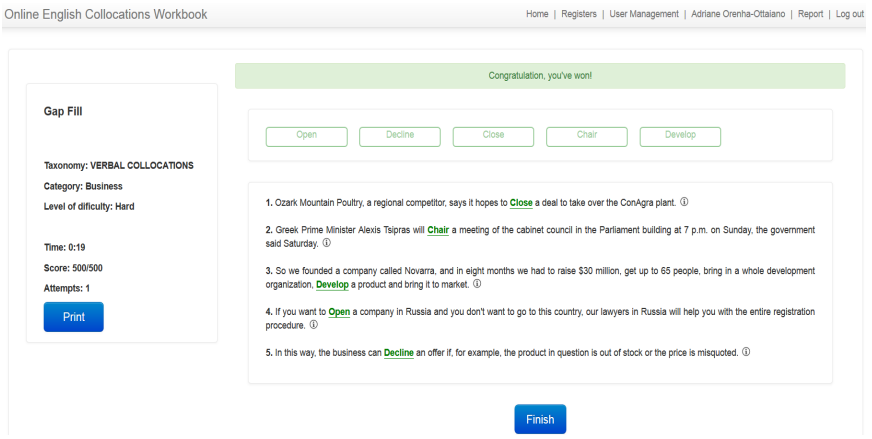
A Figura 15 mostra como os verbos são arrastados pelo usuário até a oração que formaria a colocação:

Figura 15 - Usuário arrastando um verbo que formará uma colocação em contexto



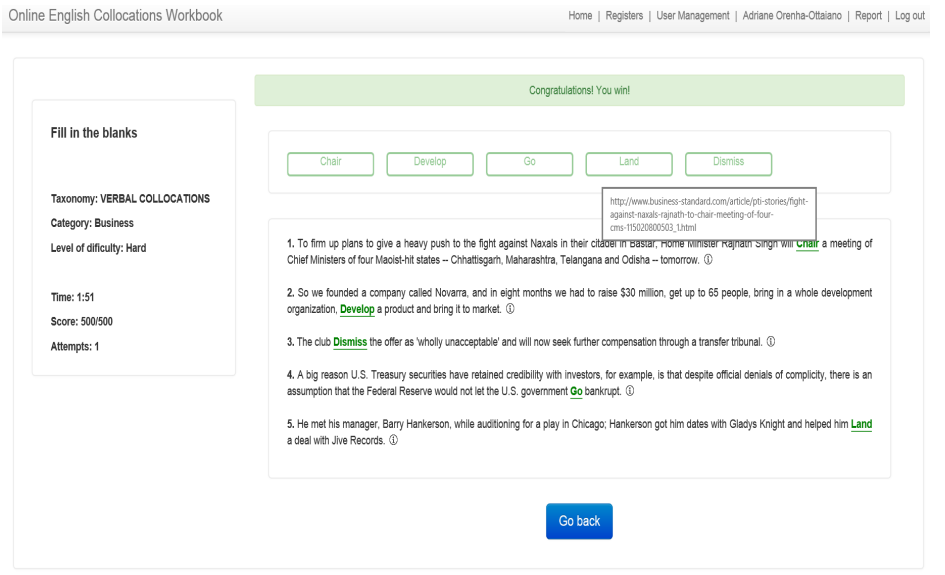
Uma vez completado o exercício, o usuário pode clicar em *Check*, a fim de verificar se suas opções estariam corretas. Ao fazer isso, obterá o seguinte resultado:

Figura 16 - Correção do exercício de *Gap Fill*



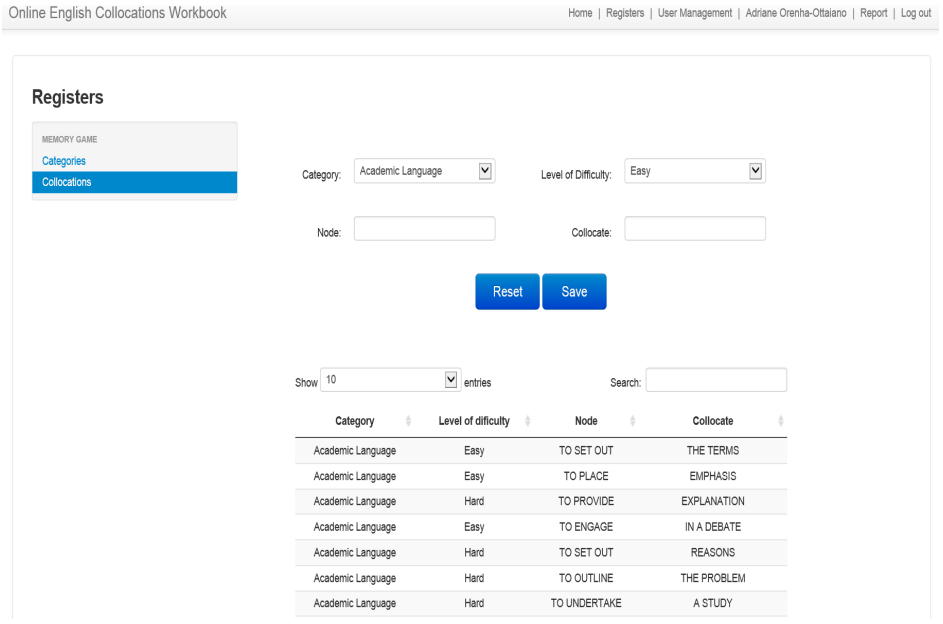
Segundo pode ser observado na Figura 17, ao clicar no ícone ⓘ, o usuário terá acesso à fonte do contexto em que as colocações empregadas foram retiradas:

Figura 17 - Fonte do contexto em que as colocações empregadas foram retiradas



Se o aprendiz clicar em *Registers*, na tela de abertura do *Workbook*, ele terá acesso à lista de todas as colocações com as quais ele já trabalhou:

Figura 18 - Lista de todas as colocações estudadas



Além do *Memory Game* e do *Gap Fill*, outros tipos de exercícios ou jogos serão elaborados e inseridos na plataforma do *Workbook*, por exemplo, um jogo de *Domino* (Dominó).

5 Considerações Finais

Conforme o relato de elaboração do *Online English Collocations Workbook*, consideramos que a pesquisa aqui relatada tenha relevância científica para o público-alvo, dada sua proposta original e seu caráter inédito. Até o presente momento, desconhecemos um material de apoio publicado no Brasil que, primeiramente, enfoque as colocações e que, em segundo, seja direcionado especificamente a brasileiros aprendizes de inglês como língua estrangeira, de modo a viabilizar um aprendizado mais sistemático e, ao mesmo tempo, interativo, das colocações. Isso se dá, uma vez que as atividades estão sendo desenvolvidas com base nas

dificuldades que alunos brasileiros tiveram em relação ao uso de colocações, segundo dados observados em traduções de alunos universitários que compõem o *Corpus de Aprendizes de Tradução*, bem como de redações, também escrita por alunos universitários, que compõem o *Corpus de Aprendizes de Língua Estrangeira*. Dessa maneira, esta investigação propõe preencher uma lacuna referente à escassez de obras fraseológicas no país, além de disponibilizar um material didático como uma resposta às dificuldades enfrentadas por aprendizes brasileiros da língua inglesa quanto ao aprendizado de colocações.

Nesse sentido, o material pedagógico aqui exposto poderá contribuir para a divulgação da importância do ensino de colocações, especialmente pelo fato de estar disponível *on-line*, possibilitando mais visibilidade das pesquisas realizadas aqui no Brasil, possibilitando um acesso a um número maior de aprendizes e professores. Outro fator determinante de sua relevância se fundamenta na demanda já constatada para a disponibilização do referido material de apoio na *Web*, por parte de alunos universitários e, principalmente, de professores da rede pública e privada.

6 Referências

ALTENBERG, B.; EEG-OLOFSSON, M. Phraseology in Spoken English: presentation of a Project. In: AARTS, J.; MEIJS, W. (Ed.). *Theory and practice in Corpus Linguistics*. Amsterdam: Randpi, 1-26, 1990.

ANTHONY, L. *AntConc*. 2012. Disponível em: <http://www.antlab.sci.waseda.ac.jp/ software.html>. Acesso em: 17 fev. 2014.

BAHNS, J.; ELDAW, M. Should we teach EFL students collocations? *System*, v. 21, n. 1, p. 101-114, 1993.

BERBER SARDINHA, A. P. *Linguística de Corpus*. Barueri, SP: Editora Manole, 2004.

____. O corpus de aprendiz Br-ICLE. In: XI InPLA, 2001, São Paulo. *Intercâmbio 10*. São Paulo: EDUC, p. 227-239. Disponível em: <http://www2.lael.pucsp.br/~tony/temp/ publications/2001bricle-interc.pdf >. Acesso em: 18 fev. 2010.

BERNARDINI, S. Corpora in the classroom: an overview and some reflections on future developments. In: SINCLAIR, J. McH. *How to use corpora in language teaching*, 2004. DOI: <http://dx.doi.org/10.1075/scl.12.05ber>

DAVIES, M. *The Corpus of Contemporary American English*: 450 million words, 1990-present. Disponível em: <http://corpus.byu.edu/COCA/ >. Acesso em: 20 de fev. 2014.

DE COCK, S. An automated approach to the phrasicon of EFL learners. In: GRANGER, S. (Org.). *Learner English on Computer*. London and New York: Longman, 1998, p. 67-79.

FILLMORE, C. J. Innocence: a second idealization for Linguistics. *Berkeley Linguistic Society*, v. 5, p. 63-76, 1979.

FIRTH, J. R. Modes of Meaning. In: FIRTH, J. R. (Ed.). *Papers in Linguistics* – 1934-1951. Oxford: Oxford University Press, 1957.

FONTENELLE, T. Towards the construction of a collocational database for translation students. *Meta XXXIX*, v. 39, n. 1, 1994.

GRANGER, S. The International Corpus of Learner English. In AARTS, J.; DE HAAN, P.; OOSTDIJK, N. (Ed.) *English Language Corpora: Design, Analysis and Exploitation*. Amsterdam: Rodopi, 1993, p. 57-69.

_____. *Learner English on Computer*. London and New York: Longman, 1998a.

_____. Prefabricated patterns in advanced EFL writing: collocations and formulae. In: COWIE, A. (Ed.). *Phraseology*: theory, analysis and applications. Oxford: OUP, 1998b. p. 145-160.

GRANGER S.; HUNG, J.; PETCH-TYSON, S. *Computer learner corpora, second language acquisition and foreign language teaching*.

Amsterdam/Philadelphia: John Benjamins, 2002. DOI: <http://dx.doi.org/10.1075/lllt.6>

HALLIDAY, M. A. K. Lexical Relations. *System and Function in Language*: selected papers. Oxford, England: Oxford University Press, p. 73-83, 1961.

HAUSMANN, F. J. Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. *Praxis des neusprachlichen Unterrichts*, 31, 1984, p. 395 – 406.

_____. Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In: BERGENHOLTZ, H; MUGDAN, J (Ed.). *Lexikographie und Grammatik*. Tubingen: Niemeyer, 1985.

_____. Grundprobleme des zweisprachigen Wörterbuchs. In: INTERNATIONAL SYMPOSIM ON LEXICOGRAPHY, 3., May 14-16, 1986, Tubingen. *Proceedings*… Ed. K. Hyldgaard-Jensen; K. A. Zettersten, University of Copenhagen, Tubingen: Niemeyer, 1988. p. 137-154.

HEID, U.; MARTIN, W; POSCH, I. An Overview of approaches towards the description of collocations. *Feasibility of standards for collocational description of lexical items. EUROTRA 7* - Report, Stuttgart/Amsterdam, 1991.

HEYLEN, D.; MAXWELL, K. Lexical functions and the translation of collocations. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 13., 1994, Kyoto. *Proceedings*... Kyoto, Japan, 1994. p. 298-305.

HILL, J. Revising priorities: from grammatical failure to collocational success. In: LEWIS, M. (Ed.). *Teaching Collocation*: further developments in the lexical approach. Croatia: Heinle, 2000. p. 47-69.

KJELLMER, G. A mint of phrases. In: AIJMER, K.; ALTENBERG, B. *English corpus linguistics.* Studies in honor of Jan Svartvik. London: Longman. p. 111-127.

LEECH, G. Learner corpora: what they are and what can be done with them. In: GRANGER, S. *Learner English on Computer*. London and New York: Longman, 1998. p. xiv-xx.

MACKIN, R. On collocations: 'words shall be known by the company they keep'. In: STREVENS, P. (Ed.). *In Honor of A. S. Hornby*. Oxford: Oxford University Press, 1978.

MCINTOSH, C.; FRANCIS, B.; POOLE, R. (Ed.). *Oxford Collocations Dictionary for Students of English*. 2nd ed. Oxford: Oxford University Press, 2009. 963 p.

MEUNIER, F.; GRANGER, S. (Ed.). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam & Philadelphia: John Benjamins, 2008.
DOI: <http://dx.doi.org/10.1075/z.138>

NESSELHAUF, N. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins, 2005.

ORENHA-OTTAIANO, A. English collocations extracted from a corpus of university learners and its contribution to a language teaching pedagogy. Acta Scientiarum. Language and Culture 34 (2), 214-251. 2012. DOI: <http://dx.doi.org/10.4025/actascilangcult>

_____. *A compilação de um glossário bilíngue de colocações, na área de jornalismo de Negócios, baseado em corpus comparável***.** São Paulo, 2004. 246 f. Dissertação (Mestrado em Estudos Linguísticos e Literários), FFLCH/USP.

PAQUOT, 2008. Exemplification in learner writing: A cross-linguistic perspective. In: MEUNIER, F.; GRANGER, S. (Ed.). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam & Philadelphia: John Benjamins, 2008, p. 101-119. DOI: <http://dx.doi.org/10.1075/z.138>

PAWLEY, A.; SYDER, F. H. Two Puzzles for linguistic theory: nativelike selection and nativelike fluency. In RICHARDS, Jack C.;

SCHMIDT, Richard W. (Ed.). *Language and Communication*. London/New York: Longman, pp.191- 225, 1983.

PAWLEY, A.; SYDER, F. H. The One-Clause-at-a-Time Hypothesis. In RIGGENBACH, H. (Ed.), *Perspectives on Fluency*. Ann Arbor: The University of Michigan Press, 2000.

PRAVEC, N. A. Survey of learner corpora. In: *ICAME Journal*, No. 26, p.81-114, 2002.

SCOTT, M. *WordSmith Tools*: version 5.0. Oxford: Oxford University Press, 2008.

SEIDLHOFER, B. Closing a conceptual gap: the case for a description of English as a lingua franca. International Journal of Applied Linguistics 11 (2): 133–158, 2001. DOI: <http://dx.doi.org/10.1111/1473-4192.00011>

SINCLAIR, J. McH. *Corpus, concordance and collocation*. Oxford: Oxford University Press, 1991.

SMIT, U. *English as a lingua franca in higher education: a longitudinal study of classroom discourse*. Germany: DeGruinter, 2010, p. 46. DOI: <http://dx.doi.org/10.1515/9783110215519>

TAGNIN, S. E. O. Corpora and the Innocent Translator: how can they help him. In: THELEN, M (Ed.). *Translation and Meaning*, Part 6**,** *Proceedings of the Lodz Session of the 3rd Maastricht-Lodz Duo Colloquium on Translation on Meaning*, Lodz, Poland, September 22-24, 2000, Maastricht: Universitaire Pers Maastricht, p. 489-496, 2002.

TER-MINASOVA , Svetlana G. The Freedom of Word-Combinations and the Compilation of Learners' Dicitonaries . In: EURALEX '92 – INTERNATIONAL CONGRESS ON LEXICOGRAPHY, 5. 1992, Tampere. *Proceedings*... Ed. H. Tommola, K. Varantola,
T. Salmi-Tolonen, J. Schopp. Tampere, Finland: University of Tampere, 1992. p. 533- 539.

THOMPSON, P. Spoken language corpora. In: WYNNE, M. (Ed.) *Developing Linguistic Corpora: a Guide to Good Practice.* Oxford: Oxbow Books, p. 59-70, 2005. Disponível em

<http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm>. Acesso em: 14 jan. 2010.

TOGNINI-BONELLI, E. *Corpus linguistics at work*. Studies in Corpus Linguistics, v. 6. Amsterdam: John Benjamins Publishing Company, 2001. DOI: <http://dx.doi.org/10.1075/scl.6>

TONO, Y. Using learner corpora in ELT and SLA research. In: AILA '99 – THE ROLES OF LANGUAGE IN THE 21st CENTURY: UNITY AND DIVERSITY, 12., 1-6 August 1999, Tokyo. *Proceedings of the Symposium*: The roles of corpora in language teaching and language engineering. Tokyo, Japan: Waseda University Press, 1999.

WRAY, A. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press, 2002. DOI: <http://dx.doi.org/10.1017/CBO9780511519772>

DOI: 10.17851/2237-2083.23.3.883-922

O papel da pausa na segmentação prosódica de *corpora* de fala

The role of pause in the prosodic segmentation of spoken corpora

Tommaso Raso
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brasil
tommaso.raso@gmail.com

Maryualê Malvessi Mittmann
Universidade do Sul de Santa Catarina (UNISUL), Tubarão, Santa Catarina, Brasil
maryuale@gmail.com

Anna Carolina Oliveira Mendes
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais, Brasil
annacarol.om@gmail.com

Resumo: O trabalho apresenta uma análise da pausa silenciosa enquanto critério de segmentação da fala em unidades comunicativamente autônomas. O objetivo deste trabalho é contextualizado com uma discussão sobre a unidade de referência da fala e a noção de pausa. Argumenta-se a favor de uma segmentação da fala em unidades pragmáticas de natureza acional, delimitadas no fluxo da fala por meio de fronteiras prosódicas. Realizou-se análise estatística em amostra aleatória de fala espontânea extraída do *corpus* C-ORAL-BRASIL. Por intermédio de um modelo de regressão não linear, procurou-se identificar durações de pausas consistentemente preditivas de fronteiras de unidades de referência. Os resultados mostram que não há nenhum valor de duração de pausa que possa ser utilizado como critério confiável para a segmentação da fala em unidades comunicativamente autônomas.

Palavras chave: fala; corpora; segmentação; unidade prosódica; ilocução.

Abstract: This work presents an analysis of silent pauses when applied as criterion for the segmentation of speech into communicatively autonomous units. The aim of this work is contextualized in a discussion about the unit of reference for speech and the notion of pause. We argue that speech should be

parsed into pragmatic units of actional nature, which are signaled within the speech flow by prosodic boundaries. We performed a statistic analysis on a random spontaneous speech sample extracted from C-ORAL-BRASIL corpus. Non-linear regression models were applied so as to identify pause durations that could consistently predict unit boundaries. Results show that there isn't any pause duration value that could be used as a reliable criterion for the segmentation of speech into communicatively autonomous units.

Keywords: speech; corpora; segmentation; prosodic unit; illocution.

Recebido em: 25 de setembro de 2015.
Aprovado em: 23 de outubro de 2015.

1 Introdução

Investigamos a relação entre pausas por um lado e fronteiras de enunciados e unidades tonais por outro, conforme definidos adiante. Essa investigação baseia-se no *corpus* C-ORAL-BRASIL (RASO; MELLO, 2012) e insere-se em uma discussão acerca da unidade de referência da fala, ressaltando a importância de uma segmentação da fala adequada para representar fenômenos próprios dessa diamesia[1] e questionando diferentes propostas desenvolvidas na literatura.

Primeiramente, discutimos brevemente o conceito de unidade de referência da fala e as principais propostas. Em seguida, aprofundamos em uma das propostas, também apresentando a perspectiva da *Language into Act Theory*, que norteia a elaboração dos *corpora* de fala espontânea C-ORAL-ROM e C-ORAL-BRASIL. Depois, apresentamos a arquitetura do C-ORAL-BRASIL; e, por fim, apresentamos a metodologia e os resultados da pesquisa.

[1] Para o conceito de *diamesia*, veja-se Berruto (1993) e Rossi (2011).

2 O problema da segmentação da fala

Algumas características tornam a fala intrinsecamente distinta da escrita e, portanto, a nosso ver, não é possível transpor categorias analíticas da escrita para a fala. A discussão sobre diferenças entre fala e escrita é ampla. Remetemos a Raso (2013) para uma síntese e a proposta que embasa este trabalho. Apenas lembramos que essas diferenças ocorrem porque a fala e a escrita são transmitidas por meios e canais diferentes, e são decodificadas por sentidos diferentes. Isso gera consequências pragmáticas e linguísticas muito distintas na organização comunicativa. Característica própria da fala, inevitavelmente ausente na escrita, é a transmissão da informação pelo canal sonoro. Desse canal faz parte não somente o conteúdo segmental (articulado, percebido e decodificado de modo muito diferente da realização e decodificação dos grafemas), mas também a prosódia, com sua enorme bagagem de informações que, na escrita, desaparece ou é veiculada por procedimentos físicos e cognitivos diferentes. Isso torna impossível a transposição da fala para a escrita sem que haja, de fato, uma perda enorme de informações e a transferência de algumas delas para categorias diferentes. Por isso, consideramos impossível estudar adequadamente a fala recorrendo exclusivamente a transcrições, sem considerar as informações veiculadas por meio do som, principalmente, os efeitos da prosódia.

A necessidade de se recorrer constantemente ao áudio para avaliar os fenômenos linguísticos da fala impõe que os *corpora* disponibilizem o alinhamento do texto ao som. O alinhamento evidencia ainda mais a necessidade de definir o domínio mínimo dos principais fenômenos linguísticos, ou seja, a unidade de referência da fala maior do que a palavra.

Naturalmente, é oportuno identificar tal unidade com base nos princípios que governam a comunicação oral, e não transpondo mecanicamente estruturas que funcionam na análise da escrita. A unidade de referência deve ser entendida como o âmbito de funcionamento das principais relações entre os elementos linguísticos, nas suas várias ordens: sintática, semântica ou pragmática.

Um exemplo pode esclarecer essa questão. Imaginemos a sequência *João vai pro Rio até amanhã*. Ela pode receber diferentes segmentações, entre as quais:

(i) *João vai pro Rio até amanhã //* (uma única unidade de referência);
(ii) *João vai pro Rio // até amanhã //* (duas unidades de referência);
(iii) *João // vai pro Rio até amanhã //* (também duas unidades de referência);
(iv) *João // vai pro Rio // até amanhã //* (três unidades de referência).

Junto à segmentação, atribui-se um valor acional às unidades. Em (i), podemos ter uma asserção, uma pergunta, uma manifestação de surpresa ou outras ações; em (ii), temos duas ações, a primeira poderia ser uma asserção, uma resposta, um pedido de confirmação ou outra, e a segunda, uma despedida, uma outra asserção, um outro pedido de confirmação ou outra; em (iii), também temos duas ações, mas relativas a conteúdos locutivos diferentes, podendo a primeira ser um chamamento, ou um pedido de confirmação, ou uma resposta ou outra, e a segunda, uma ordem, uma pergunta ou outra; em (iv), temos três ações, a primeira podendo ser um chamamento, um pedido de confirmação, uma expressão de surpresa ou outra; a segunda, podendo ser uma ordem, uma expressão de incredulidade, ou de surpresa ou outra; e a terceira, podendo ser uma despedida, uma asserção, uma pergunta ou outra.

Somente após identificarmos as unidades e atribuirmos a elas um valor acional podemos dizer se *João* é sujeito ou o que tradicionalmente é chamado de "vocativo", e se a forma verbal *vai* deve ser analisada como terceira pessoa do presente do indicativo ou segunda pessoa do imperativo. Parece-nos evidente que tanto a segmentação quanto a atribuição de valor acional são guiados pela percepção de marcas de natureza prosódica.

Frequentemente dá-se o nome de *enunciado* a essa unidade de referência, mas sob essa denominação são apresentados objetos diferentes. Na tradição pragmática (AUSTIN, 1962), o termo *enunciado* faz referência ao que é considerada a mínima unidade comunicativa,

capaz, portanto, de veicular uma mensagem mínima e autônoma, com um núcleo comunicativo de natureza acional. Como identificar e descrever as estruturas linguísticas que realizam essa função é ainda objeto de discussão. Quatro são as principais perspectivas.

2.1 Diversas visões sobre a unidade de referência da fala

2.1.1 A sentença falada

Na perspectiva sintaticista, define-se enunciado como uma "sentença falada". Consideramos as duas principais definições de sentença: (a) como predicação verbal relativa a um SN sujeito, hierarquicamente subordinado ao SV (HARRIS, 1962); (b) como a "máxima projeção do núcleo verbal" (CHOMSKY, 1970).

Com base no *corpus* LABLITA de italiano falado (CRESTI, 2000), Cresti e Gramigni (2004) observam que estruturas do tipo (a) constituem menos de 5% dos enunciados, ou seja, uma frequência muito baixa para ser considerada fenômeno relevante para a fala. Mello, Raso, Bossaglia e Santana (em preparação), analisando o *minicorpus* extraído do *corpus* C-ORAL-BRASIL (RASO, 2012a; MELLO, 2014; PANUNZI; MITTMANN, 2014), observam que em Português Brasileiro (PB) as predicações constituem cerca de 15% dos enunciados, mas aquelas preenchidas com um SN pleno e um SV pleno representam pouco mais de 1%. Cresti e Gramigni (2004) já haviam observado que no italiano a maioria das predicações são constituídas por sujeitos pronominais e / ou verbos de ligação. As predicações em PB são bem mais frequentes que em italiano, mas ainda aparecem em quantidade muito reduzida. Essas diferenças não devem ser atribuídas a características dos *corpora*, que são comparáveis, mas às características acentuais das duas línguas. O PB, principalmente na variedade mineira (representada no C-ORAL-BRASIL), é mais acentual do que o italiano, o que permite hospedar nas unidades tonais maior quantidade de material fonológico, possibilitando estruturas sintáticas mais complexas.

Estruturas do tipo (b) são consideravelmente mais comuns nos dois *corpora*, variando entre 62% e 70%, dependendo da língua e de fatores sociolinguísticos. Em consonância com o evidenciado por Biber *et al*. (1999) para o inglês, Cresti e Gramigni (2004) observam, no

italiano, que 38% dos enunciados são não verbais. Para o PB, Raso e Mittmann (2012) notam que 27% dos enunciados não apresentam qualquer forma verbal e apenas 55% apresentam forma funcionalmente verbal[2] como núcleo do enunciado. A definição de enunciado como máxima projeção de V mostra-se, portanto, incapaz de explicar mais de um terço das unidades presentes em *corpora*.

Ambas as definições do enunciado como "sentença falada" deixam de fora um número grande demais de casos para serem consideradas satisfatórias na análise da fala.

2.1.2 O turno

Na perspectiva dialógica, adota-se o turno como unidade de referência. Em *corpora* de fala espontânea como o C-ORAL-ROM (CRESTI; MONEGLIA, 2005) para espanhol, francês, italiano e PE, o C-ORAL-BRASIL para o PB e o Santa Barbara *Corpus* (DU BOIS *et al.*, 2005) para o inglês americano, observa-se que essa delimitação corresponde a objetos muito heterogêneos. Turnos podem abranger uma única palavra (ou mesmo interjeição), ou estruturas extremamente complexas, com duração de muitos minutos. Devido a tamanha heterogeneidade, o turno, apesar de constituir uma unidade natural da fala, não é adequado para a delimitação da unidade mínima significativa maior do que a palavra.

Os exemplos a seguir ilustram sequências de turnos em uma conversação e um diálogo, e um fragmento de texto prevalentemente monológico.[3]

---

[2] Uma parte significativa dos enunciados com verbo pertencem a duas tipologias: a) enunciados em que o verbo aparece como expansão de um núcleo de outra natureza, por exemplo um SN modificado por uma relativa; b) formas verbais que retomam o verbo do enunciado anterior para expressar confirmação ou negação, como no exemplo fictício a seguir:
*AAA: você ligou para João //
*BBB: liguei //

[3] A sigla de três letras precedidas de asterisco e seguida de dois pontos indica o falante do turno; os números entre colchetes indicam o número dos enunciados no texto; os símbolos "<>" indicam sobreposição de fala; as barras duplas indicam final de enunciado; as barras simples indicam final de unidade tonal; o símbolo "+" indica enunciado interrompido; a barra simples entre colchetes e seguida por um

Ex. 1 - bfamcv05 [75-95][4]

*JOS: [75] <e as> mordomia que es têm //
*CAR: [77] na hora que fizer cinco / nós vamo parar cinco minutos / viu // [78] porque / <tem jeito não> //
*JOS: [79] <ou mais / né> // [80] ué / onde é que essa bola foi //
*CAR: [81] <machucou / ô> //
*JOS: [82] <no &go [/2] lá no quintal do vizinho> //
*CEL: [83] espim / cara //
*CAR: [84] cuidado aí que <tem> [/1] tem coisa aí //
*MAR: [85] <possível> / aí //
*CAR: [86] vai dar pra jogar //
*CEL: [87] não / vai / sô //
*JOS: [88] furou nada não //
*MAR: [89] caco de vidro //
*CEL: [90] não / foi <espinho mesmo> //
*CAR: [91] <nũ é não> / ali tem / ora-pro-nóbi //
*MAR: [92] <não / é de mão / meu> filho //
*CEL: [93] <tá a bosta / mesmo aqui o'> //
*JOS: [94] não // [95] tanto faz //

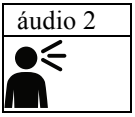
O exemplo 1 mostra uma alternância de turnos em uma conversação fortemente interativa, com turnos extremamente curtos: trata-se de quatro amigos jogando futebol. Em 89 palavras, temos 17

número indica *retracting* (o número indica a quantidade de palavras retratadas, ou seja virtualmente deletadas, pelo falante); o símbolo "&" indica palavra não terminada; o símbolo "hhh" indica riso ou tosse ou som sem valor linguístico; a sequência "&he" indica tomada de tempo, frequentemente chamada também de pausa preenchida.

[4] Cada exemplo é acompanhado da referência do texto e dos enunciados aos quais se refere, sempre relativos ao *corpus* C-ORAL-BRASIL. A sigla de identificação do texto é composta por uma letra que identifica a língua (em nosso caso sempre "b"), por três letras que identificam o contexto, se familiar-privado ("fam") ou público ("pub"), por duas letras que identificam a tipologia interacional, monólogo ("mn"), diálogo ("dl") ou conversação ("cv"), esse último entendido como um diálogo com mais de dois participantes, e por um número que identifica o texto dentro da categoria definida pela sigla. O(s) número(s) entre colchetes indicam o enunciado ou a sequência de enunciados.

turnos. Observe-se que apenas três foram segmentados em dois enunciados (delimitados pela barra dupla), enquanto 14 são formados por um único enunciado. Em exemplos assim, a coincidência entre turno e enunciado é quase total. Observe-se também que seis enunciados e cinco turnos não apresentam elemento verbal, e outros dois apresentam verbos não nucleares.

Ex. 2 - bpubdl03 [118-135]



*GUI: [118] volta aqui //
*GUI: [119] faz força //
*GUI: [120] mais //
*GUI: [121] beleza //
*GUI: [122] contrai o abdômen //
*GUI: [123] joga o tronco só um pouquinho pra frente //
*GUI: [124] aí //
*GUI: [125] beleza //
*GUI: [126] descansou //
*GUI: [127] vou baixar um pouquinho mais //
*GUI: [128] vai //
*GUI: [129] pera aí / deixa eu passar a faixa //
*GUI: [130] aí //
*GUI: [131] vai / força //
*GUI: [132] aqui //
*GUI: [133] pra frente // [134] isso //
*GUI: [135] pesado //

O exemplo 2 mostra uma alternância de 17 turnos do mesmo falante em 41 palavras. Trata-se de um diálogo entre *personal trainer* e cliente, durante uma sessão de ginástica. Aqui, não há alternância de falante; GUI produz turnos alternados a comportamentos não verbais do ouvinte. Evidentemente não podemos considerar esse um trecho monológico. O falante não segue um plano textual independente da interação. Ao contrário, o trecho é fortemente interativo, mas um dos participantes interage de maneira não verbal. Os enunciados do falante não podem, portanto, ser considerados como pertencentes ao mesmo turno, mas como turnos que se alternam a turnos não verbais, mas "agidos", pelo interlocutor. Observe-se que apenas um turno foi segmentado em dois enunciados e que oito turnos e nove enunciados não possuem verbo.

Ex. 3 - bfammn06 [6-17]

*JOR: [6] bom / eu tive / a minha / formação / profissional / dentro da / área de engenharia / depois que eu fiz escola técnica no Rio de Janeiro / e passei pela administração na Fundação Getúlio Vargas // [7] na + [8] já tem algum tempo que eu tô formado / naquela época / o mercado de trabalho era totalmente diferente de hoje // [9] hhh a [/1] as multinacionais estavam entrando dentro do país / e procuravam / alunos dentro das próprias universidades // [10] e assim eu iniciei minha vida profissional / na área técnica de engenharia elétrica // [11] um belo dia durante o almoço / o gerente de recursos humanos de uma multinacional / me informou que havia uma vaga na área comercial da empresa / e / se eu tinha interesse // [12] eu &fo + [13] eu &es + [14] informei a ele que eu tava preste a me formar / e / estava trabalhando dentro duma / área que eu gostava // [15] mas / ele me informou / que o salário seria quase o dobro do que eu ganhava // [16] e aquilo mexeu muito comigo // [17] e aí / eu consegui / a [/1] com a experiência que eu tinha dentro da multinacional / concorrer à vaga e &f [/1] isso me facilitou / e eu passei pra área comercial da empresa pra vender / disjuntores / transformadores / motores de / corrente contínua / corrente alternada / isoladores / e / relés de proteção secundária / e assim foi iniciando a minha vida comercial //

O exemplo 3 mostra apenas um turno de uma situação tendencialmente monológica (na fala espontânea informal não existe monólogo perfeito). Em 104 palavras, temos apenas um turno, segmentado em 17 enunciados. Observe-se que, com exceção de dois interrompidos, todos os enunciados possuem núcleo verbal.[5] A natureza do texto, a sua diafasia, se revela, portanto, decisiva na composição sintática dos enunciados.

Esses exemplos mostram como o turno, mesmo sendo uma unidade natural da fala, não serve para identificar a unidade de referência da fala, entendida como o âmbito no qual identificar uma intenção comunicativa mínima e autônoma do falante e as relações estreitas dos elementos linguísticos. Isso nos induz a procurar algo entre a palavra e o turno que constitua o domínio das principais relações linguísticas em uma unidade comunicativa mínima.

[5] Para uma análise comparativa da estrutura de textos dialógicos e monológicos, veja-se Raso e Mittmann (2012) e Mittmann (2013).

2.1.3 A pausa

Frequente é a definição de enunciado como sequência entre duas pausas do mesmo falante. Primeiramente, precisa-se fazer algumas considerações sobre o que entendemos por pausa. Comumente, distingue-se entre "pausa silenciosa" e "pausa preenchida" (ZELLNER, 1994; SORIANELLO, 2006; MERLO; BARBOSA, 2012). A pausa silenciosa seria constituída por um silêncio no fluxo locutivo; voltaremos a esse tipo de pausa, objetivo principal do trabalho. A pausa preenchida é considerada uma interrupção cognitiva que, entretanto, é realizada com alguma vocalização. Quem agrupa em um mesmo conceito os dois tipos de pausa, de fato, considera a vocalização um elemento superficial, que não merece um estatuto diferenciado por si só. A nosso ver, pausa silenciosa e pausa preenchida não podem ser tratadas como duas formas de uma única categoria. Tomar uma decisão definitória com base em uma suposta identidade cognitiva em presença de correlatos físicos diferentes não nos parece metodologicamente confiável.

A pausa silenciosa é uma interrupção do fluxo da fala, mesmo que provisória, identificável perceptualmente sem ambiguidade. Pode gerar efeitos involuntários (como perda do turno) ou efeitos voluntários de ordem comunicativa (como ressaltar um trecho de fala ao criar expectativa). A pausa preenchida é um fenômeno principalmente de disfluência, ainda mais se nessa categoria se inserem as repetições. Pausa silenciosa e pausa preenchida podem ou não se sobrepor funcionalmente. Mesmo certas expressões linguísticas podem ter funções parecidas com algumas das funções da pausa, silenciosa ou preenchida.

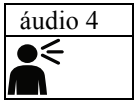
Consideramos importante distinguir entre o nível da identificação de fenômenos próprios da fala em termos de características físicas, e o nível da interpretação funcional desses fenômenos, que pode variar em contextos distintos. Assim, diferenciamos esses casos, deixando o termo "pausa" unicamente para aquela silenciosa. As vocalizações serão consideradas tomadas de tempo, marcadas em nossa transcrição com o símbolo "&he", e na etiquetagem funcional das unidades tonais com =TMT= (*time taking*); as repetições são consideradas casos de *retracting*; os *retractings* são correções que os falantes fazem, com ou sem repetições; em nossa notação, são marcadas com uma barra simples entre colchetes e um número que indica a quantidade de palavras retratadas; os

retractings são etiquetados como =EMP= (*empty*), para marcar que se trata de unidades informacionalmente vazias. Tanto as tomadas de tempo quanto as palavras retratadas devem poder ser isoladas na computação do número total das palavras, já que não possuem o mesmo valor informativo do resto (MONEGLIA; RASO, 2014; PANUNZI; MITTMANN, 2014).

Em vários arcabouços, as pausas são consideradas marca de fronteira de enunciado. De fato, uma pausa silenciosa igual a ou maior que 200ms foi considerada como fronteira em alguns *corpora* de fala, como no *Dutch Corpus* (BUHMANN *et al.*, 2002) e uma pausa igual a ou maior que 100 ms foi considerada fronteira em *corpora* japoneses (MARUYAMA, 2012, 2015). A vantagem disso é evidente: estabelecendo uma duração de silêncio como marca de fronteira, pode-se realizar a segmentação da fala automaticamente. Contudo, também as desvantagens são evidentes. Não há uma medida *a priori* ancorada a algum aspecto de natureza perceptual ou linguística que defina a duração mínima e máxima de um silêncio que possa, por si só, ser considerado fronteira de enunciado.

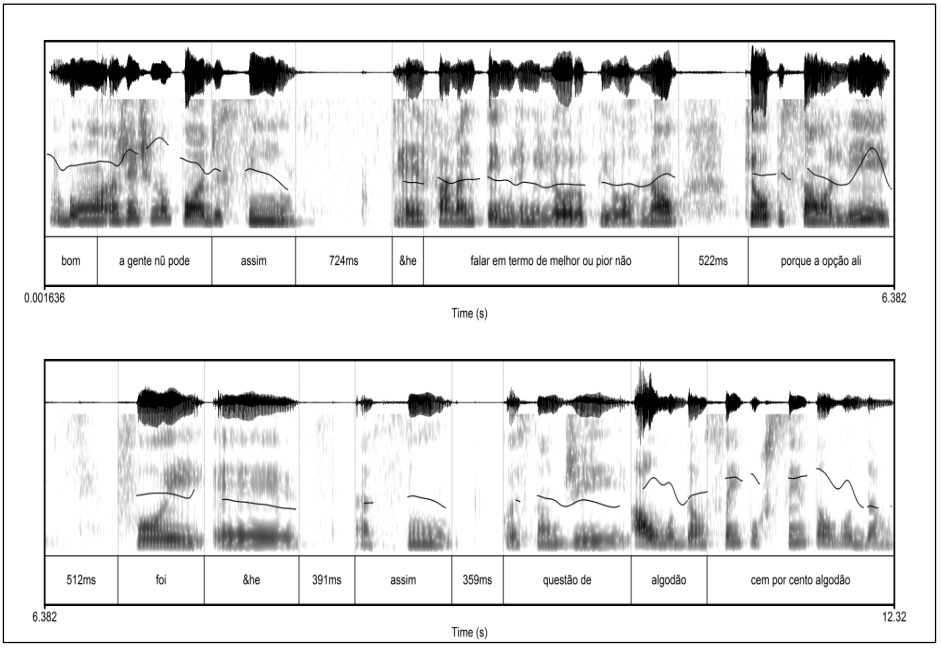
Os exemplos seguintes mostram casos de pausas, tomadas de tempo, *retractings* e coocorrência dos três. Neles as pausas são sempre maiores que 200ms e, à oitiva, resultam claramente não fronteiriças, no sentido de não separar duas unidades percebíveis como comunicativamente autônomas. Nesses casos, as pausas constituem apenas silêncios internos à mesma unidade comunicativa mínima. Outros exemplos mostram como unidades comunicativamente autônomas podem ocorrer na fala sem serem separadas por pausas. O intuito é duplo: mostrar que pausas silenciosas e preenchidas não podem ser consideradas duas formas do mesmo fenômeno e mostrar que a pausa não é nem suficiente nem necessária para marcar fronteira de enunciado.

Exemplo 4 - bfamcv19 [37]



*MAE: bom / a gente nũ pode / assim / &he / falar em termo de melhor ou pior não / porque a opção ali / foi / &he / assim / questão de / algodão / cem por cento algodão //

Figura 1 – Sinal de áudio, espectrograma, F0 e transcrição do exemplo 4

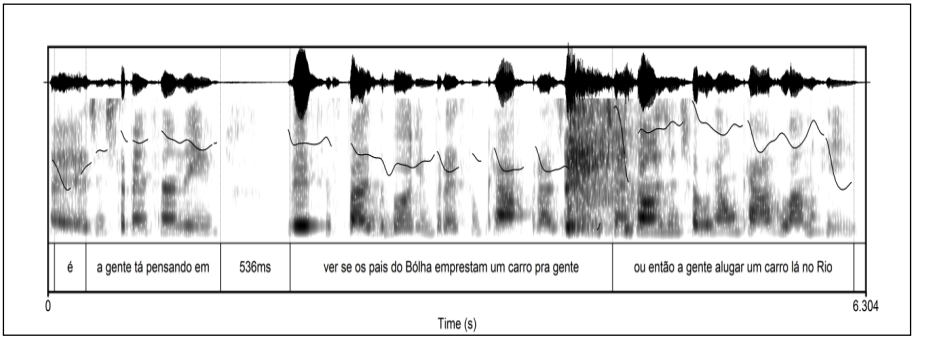


Esse enunciado apresenta pausas longas e tomadas de tempo: primeiro uma pausa de 724ms, seguida de uma rápida tomada de tempo; depois, uma segunda pausa de 522ms; uma terceira de 512ms, por sua vez, seguida de uma palavra (foi), uma tomada de tempo e uma pausa de 391ms, ainda seguida de uma palavra (assim) e uma outra pausa de 359ms. A rigor, ainda se segue uma pausa menor (147ms), após duas palavras (questão de) antes do fim do enunciado.

Exemplo 5 - bfamcv29 [37]

*ELI:é / a gente tá pensando **em / ver** se os pais do Bólha emprestam um carro pra gente / ou então a gente alugar um carro lá no Rio //

Figura 2 – Sinal de áudio, espectrograma, F0 e transcrição do exemplo 5



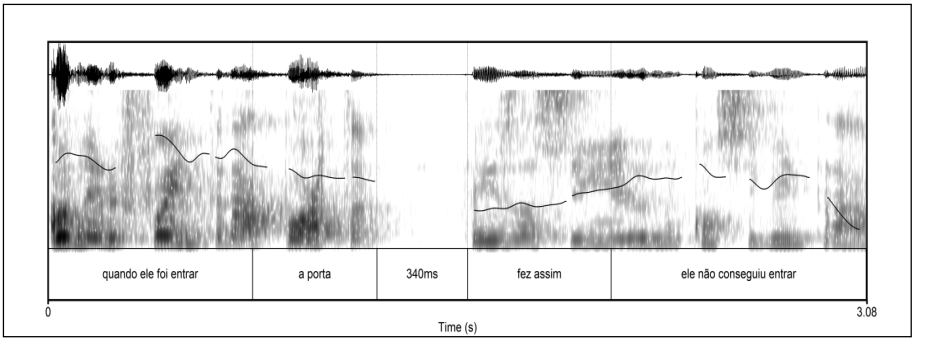
No exemplo 5, temos, entre outras, uma pausa de 536ms entre a preposição e o verbo por ela regido, evidenciando ainda mais que duração de pausa não é correlato confiável para determinar a existência de fronteira de unidade mínima comunicativamente completa, por estar em uma posição impossível de ser considerada fronteiriça do ponto de vista sintático, e não somente pragmático e prosódico.

Exemplo 6 - bfamdl14 [02]

áudio 6	áudio 6a
	

Quando ele foi entrar / **a porta / fez assim** / ele não conseguiu entrar //

Figura 3 – Sinal de áudio, espectrograma, F0 e transcrição do exemplo 6



No exemplo 6, uma pausa de 340ms está entre os elementos que,

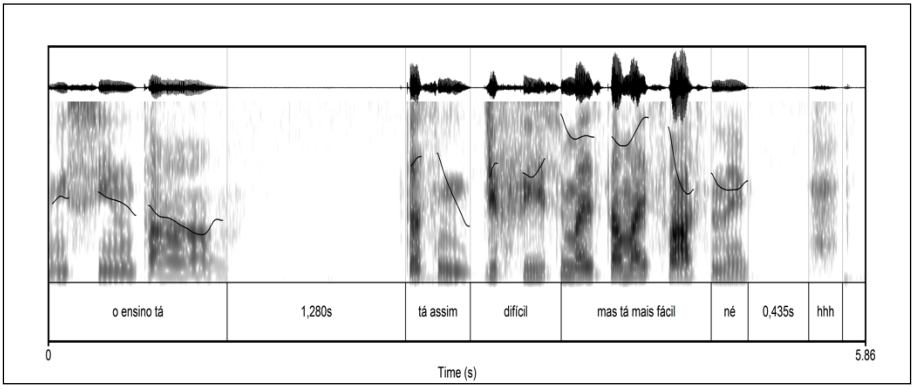
em uma análise sintática, seriam considerados o sujeito e o verbo. Mesmo em uma análise prosódica, essa posição não poderia ser considerada fronteira de enunciado, como mostram as oitivas do enunciado (áudio 6) e do trecho até a pausa (áudio 6a).

Exemplo 7 bpubdl11 [113]

áudio 7	áudio 7a	áudio 7b
		

o ensino tá [/1] tá assim / difícil / mas tá mais fácil / né hhh //

Figura 4 – Sinal de áudio, espectrograma, F0 e transcrição do exemplo 7



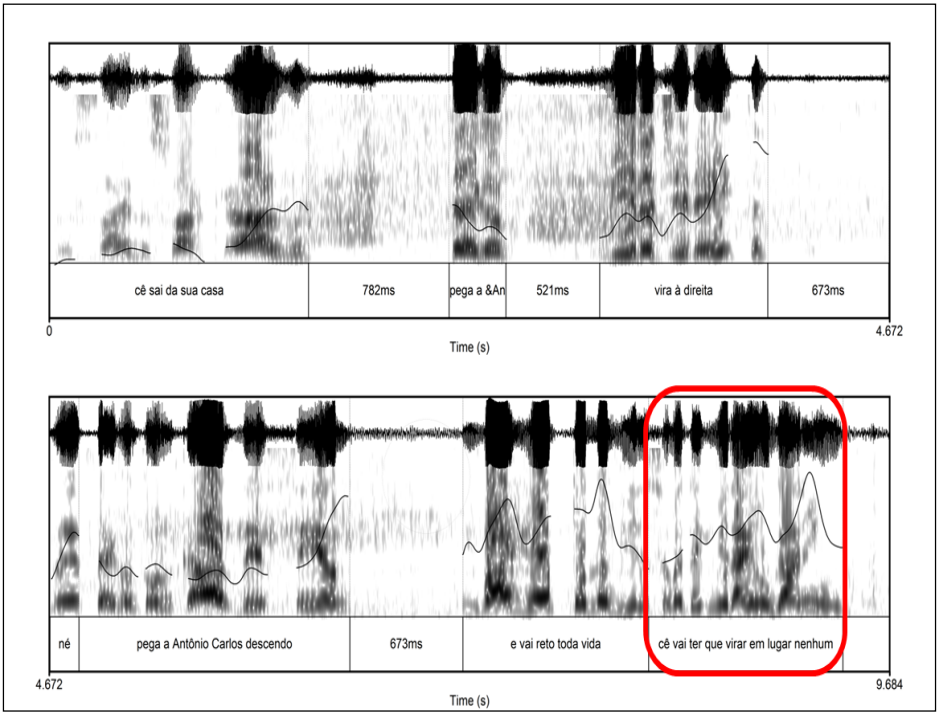
O exemplo 7, além da pausa, apresenta repetição de uma palavra. Aqui temos dois fenômenos não reduzíveis a um: a repetição, que poderia ter acontecido sem pausa (como mostra o áudio 7a editado), e o silêncio de 1,28s. Nesse caso também, a pausa não pode constituir marca de fronteira de enunciado, como mostra o áudio 7b.

Exemplo 8- bfamdl08 [34-35]

áudio 8


*AND: cê sai da sua casa / pega a &An [/3] vira à direita / é / pega a Antônio Carlos descendo / e vai reto toda vida // cê vai ter que virar em lugar nenhum //

Figura 5 – Sinal de áudio, espectrograma, F0 e transcrição do exemplo 8



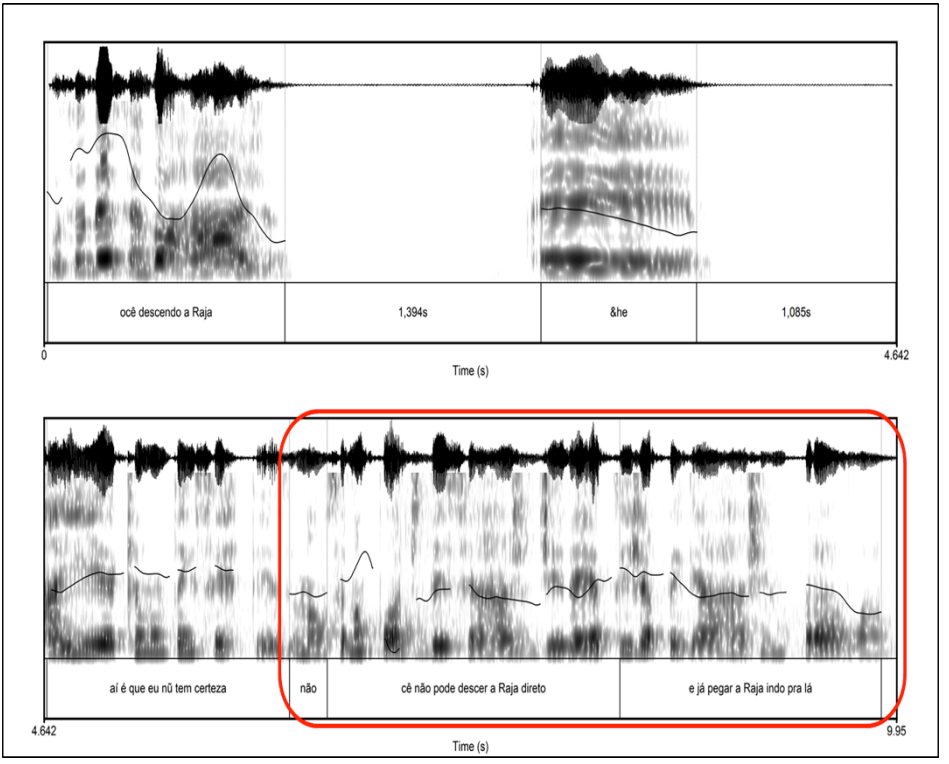
No exemplo 8, mostramos que a pausa não é uma marca acústica necessária para a segmentação da fala. Temos aqui dois enunciados, como mostrado na Figura 5, na qual o segundo enunciado está destacado. Mesmo sem haver pausa entre eles, podemos interpretá-los em isolamento (áudios 8a e 8b). Contudo, temos diversas pausas dentro do primeiro enunciado: três delas ultrapassam os 600 ms, e uma, em coincidência com o *retracting*, é de 521ms. O *retracting* pode ser considerado um provável engatilhador da pausa, mas constitui um fenômeno separado dela.

Exemplo 9 - bfamcv14 [165-166]

áudio 9

*LCS: ocê descendo a Raja / &he / aí é que eu nũ tem certeza // não / cê não pode descer a Raja direto / e já pegar a Raja indo pra lá //

Figura 6 – Sinal de áudio, espectrograma, F0 e transcrição do exemplo 9



Em 9 também temos duas unidades terminadas. Novamente, não há pausas entre elas, mas duas pausas muito longas dentro da primeira: 1.394ms e 1.085ms. Além disso, entre as pausas, há uma tomada de tempo de 847ms. Não podemos falar de uma única pausa de 3326ms. Os dois fenômenos, pausa e tomada de tempo, devem permanecer separados. Cada um pode aparecer sozinho, como frequentemente acontece. A tomada de tempo, nesse caso, tem provavelmente a função de evitar um silêncio prolongado demais, que geraria um risco alto de perda do turno ou até de encerramento da comunicação.

Exemplos assim não são excepcionais; ao contrário, são comuns. Seja pela função, seja pela frequência, esses casos mostram que a pausa não é um critério adequado para marcar fronteira entre enunciados. Isso é confirmado por uma grande quantidade de casos em que podemos

perceber fronteiras entre unidades autônomas, sem que o fluxo da fala seja interrompido por pausas, como vimos nos exemplos 8 e 9.

2.2 A perspectiva pragmática e a *Language into Act Theory*

Vários autores falam em *pausa virtual*, para se referir a fronteiras entre enunciados não coincidentes com uma interrupção física, mas, sim, com uma pausa "cognitiva". Nespor e Vogel (1986), por exemplo, contrastam o que consideram pausa *virtual* ou pausa perceptual e a pausa real. A pausa perceptual refere-se ao que é percebido como pausa, podendo foneticamente corresponder a uma variedade de fenômenos, como mudança de *pitch* e de duração, e que só algumas vezes corresponde à cessão da fonação. Já a pausa real refere-se à existência de porção de silêncio no sinal da fala. Se, por um lado, isso representa o reconhecimento de que pausa, entendida como silêncio dentro do fluxo da fala, não é um parâmetro adequado para identificar fronteiras entre os enunciados no fluxo da fala, por outro lado, o conceito de *pausa virtual* deixa sem resposta algumas questões: (a) do ponto de vista formal, por que utilizar a expressão *pausa virtual*, que sugere que se trata de uma pausa que não é realmente uma pausa, ou seja, por que denominar um fenômeno de natureza cognitiva, a fronteira, com um termo que normalmente faz referência a algo físico, adicionando um adjetivo (virtual) que indica que o substantivo deve ser entendido como faltante de seu correlato físico?; (b) do ponto de vista mais substancial, o que seria essa *pausa virtual*?; que configuração de parâmetros físicos (*pitch,* duração, intensidade dos fones pré e pós fronteiriços) geram um objeto nomeado de forma que sugere ser apenas virtualmente o que o nome designaria (como pausa sem pausa)?; e (c) do ponto de vista pragmático, ou seja, mais diretamente comunicativo, o que torna um enunciado autônomo e, portanto, suscetível de ser reconhecido como unidade mínima de referência com capacidade comunicativa?

A seguir tentaremos responder às perguntas mediante de conceitos oriundos da *Language into Act Theory*. Para isso, resumiremos a teoria rapidamente. A bibliografia indicada permite maior aprofundamento.

2.2.1 *Language into AcT Theory* (L-AcT)

Partindo de uma perspectiva pragmática, L-AcT define enunciado como "a menor unidade pragmática e prosodicamente autônoma" (CRESTI, 2000). É uma teoria *corpus driven* (CRESTI 2000; MONEGLIA 2011; RASO, 2012c; MONEGLIA; RASO, 2014), fruto de décadas de observação de *corpora* de fala espontânea, que retoma o conceito de ato de fala de Austin (1962) e analisa a fala com base em unidades demarcadas por meio da prosódia. Segundo essa teoria, o que confere autonomia pragmática a um enunciado é o fato de ele veicular uma ilocução, e o que delimita os enunciados é um tipo de fronteira de unidade tonal (CRYSTAL, 1975), realizada acusticamente pelo que denominamos *quebra prosódica terminal*, marcada nas transcrições com a barra dupla (//) (MONEGLIA; CRESTI, 1997).

Pesquisas baseadas em *corpora* de fala espontânea revelaram algumas regularidades:

i. os falantes segmentam naturalmente a fala em unidades pragmática e prosodicamente autônomas: os enunciados. Enunciados contêm uma unidade tonal-informacional nuclear (ilocucionária), necessária e suficiente para veicular a autonomia pragmática;
ii. o enunciado pode ser constituído por mais de uma unidade tonal, que, em princípio, constitui uma unidade informacional, mas apenas uma delas deve necessariamente ser de natureza ilocucionária;
iii. a prosódia funciona como interface entre locução e ilocução, determinando o significado do que é dito: (a) marcando fronteira entre enunciados, ou seja, as unidades de referência da fala maiores do que a palavra; (b) marcando, com as *quebras prosódicas não terminais* (/), fronteiras entre eventuais unidades internas ao enunciado; e c) marcando a função de cada unidade informacional, seja ela de natureza ilocucionária (nesse caso, a prosódia marca também o tipo de ilocução), seja ela de natureza não ilocucionária (marcando a função específica); e

iv. as relações sintáticas estreitas têm como domínio a unidade informacional. As relações entre as unidades informacionais são de natureza semântica e marcadas prosodicamente (CRESTI, 2014).

Uma quebra prosódica terminal marca, portanto, o limite entre uma unidade de referência e outra. As relações informacionais são internas a essa unidade, e as relações sintáticas estreitas são internas às unidades informacionais que compõem a unidade de referência. A unidade de referência articula-se em volta de um núcleo ilocucionário, que pode (mas não necessariamente deve) ser precedido e / ou seguido por unidades não ilocucionárias. O que revela a unidade ilocucionária são principalmente dois fatores: a) a unidade ilocucionária é interpretável em isolamento. É ela que confere autonomia ao enunciado; e b) nenhuma das unidades não ilocucionárias é interpretável em isolamento, somente em conjunto com a unidade ilocucionária.

Esse critério está na base da segmentação dos *corpora* C-ORAL-ROM, C-ORAL-BRASIL e, com algumas variações, do Santa Barbara *Corpus* (DU BOIS *et al.*, 2005), do CorpAfroAs (METTOUCHI; CHANARD, 2010) e do CoSIH (IZRE'EL; HARY; RAHAV, 2001).

2.2.2 Quebra prosódica, unidade tonal e unidade de referência da fala

A noção de *quebra prosódica* nos parece mais adequada do que aquela de *pausa virtual*. Do ponto de vista terminológico, *quebra* remete a uma ruptura que gera fronteiras entre unidades distintas, enquanto *pausa virtual* remete a um efeito cognitivo em ausência de seu veículo físico. Contudo, ainda precisamos definir o conceito de *quebra*. Pesquisas em várias línguas mostram que fronteiras são percebidas de modo constante e homogêneo pelos falantes (HARRIS; UMEDA; BOURNE, 1981; MO; COLE; LEE, 2008; SCHUETZE-COBURN; SHAPLEY; WEBER, 1991; CARLSON; HIRSCHBERG; SWERTS, 2005; SWERTS; COLLIER; TERKEN, 1994). Para verificar a confiabilidade da segmentação de *corpora* com base na percepção de quebras prosódicas é necessário recorrer a testes estatísticos de medição do acordo entre segmentadores. O teste normalmente usado é o Kappa de Fleiss (1971), que mede a probabilidade que o acordo seja devido ou não

ao acaso. Para o C-ORAL-ROM e o C-ORAL-BRASIL o teste foi aplicado com metodologias um pouco diferentes. A experiência do C-ORAL-BRASIL (por ser posterior ao C-ORAL-ROM) testou o acordo entre três segmentadores antes do início da segmentação e depois da primeira revisão das transcrições. A metodologia de aplicação dos testes é descrita detalhadamente em Moneglia *et al*. (2010), Raso e Mittmann (2009) e Mello *et al*. (2012). O teste mostrou um acordo de 0,86 quanto às quebras não terminais e de 0,87 quanto às terminais, e o valor geral de 0,86. Trata-se de um valor de acordo considerado excelente (o valor máximo é 1). Esse resultado mostra a saliência perceptual das quebras prosódicas. Considere-se que os casos de desacordo nunca são devidos à percepção, por um ou mais juízes, de uma quebra avaliada como terminal *versus* a percepção de ausência de quebra por outros juízes. Os desacordos concentram-se na percepção de quebra terminal *versus* percepção de quebra não terminal e de quebra não terminal *versus* ausência de quebra. Isso nos oferece algumas indicações:

i. existe evidência perceptual indiscutível das quebras prosódicas;
ii. existe evidência perceptual indiscutível que permite aos falantes a atribuição do valor terminal ou não terminal da fronteira, dependendo do tipo de quebra prosódica; e
iii. existem quebras mais salientes e menos salientes.

Se interpretadas à luz da L-AcT, as quebras teriam duas funções importantíssimas: (a) a segmentação entre unidades mínimas interpretáveis autonomamente, ou seja, entre unidades de referência da fala, que podemos chamar de *unidades terminadas*; e (b) a eventual segmentação interna dessas em unidades tonais-informacionais, uma das quais necessariamente deve ser ilocucionária.

Mais complexo é definir o conceito de quebra prosódica (em nosso entender, o principal fenômeno que delimita as fronteiras na fala) em termos de seus correlatos físicos. A pausa produz uma fronteira, mas há fronteiras mesmo em ausência de pausas. A literatura aponta vários parâmetros, entre os quais o *reset* da curva de F0 é considerado um dos principais (ao lado do alongamento da sílaba pré-fronteiriça); mas é possível perceber fronteiras mesmo sem evidência de *reset* da F0, por efeito da combinação de outros parâmetros, como forte e repentina

variação na intensidade, na duração e na velocidade de fala (ALBANO LEONI; MATURI, 2002). Entender como se combinam os parâmetros que constituem a quebra prosódica é de grande interesse para o estudo da estruturação da fala, visto a forte relevância perceptual da quebra prosódica e a grande importância da unidade tonal como domínio de muitos fenômenos linguísticos.

Apesar de não sermos ainda capazes de detalhar as combinações de parâmetros que produzem as quebras prosódicas, é indubitável que atribuímos forte relevância à segmentação prosódica do fluxo de fala. Essa segmentação parece ter dois níveis hierarquicamente distintos: (a) a unidade tonal, que empacota a informação segmental em uma unidade prosódica; e (b) a *unidade terminada* que representa uma unidade mínima autônoma no fluxo da fala.

2.2.3 Unidade de referência e ilocução

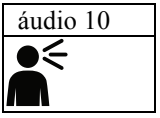
Os exemplos a seguir, juntamente ao exemplo 8, mostram (i) uma pequena sequência de unidades terminadas compostas por uma única unidade, necessariamente de natureza ilocucionária; (ii) um caso de unidade terminada complexa, formada por várias unidades diferentes em volta de uma unidade ilocucionária; e (iii) um caso (exemplo 8) de unidade terminada formada por mais unidades tonais, entre as quais mais de uma de natureza ilocucionária: a primeira unidade terminada apresenta quatro ilocuções de instrução. Tudo isso nos permite observar o seguinte:

i. a unidade ilocucionária é essencial para realizar uma unidade terminada, ou seja, a unidade terminada não é simplesmente uma cadeia concluída por uma quebra terminal, mas, sim, uma unidade de natureza acional: sem seu núcleo acional ela não se sustenta. Isso é de extrema importância, porque permite entender uma qualidade de natureza pragmática constitutiva da unidade terminada, ou seja, da unidade de referência da fala;
ii. existem unidades terminadas com mais de uma ilocução, nas quais é realizado um sinal prosódico de continuidade depois da produção de cada ilocução, com exceção da última; esse sinal de continuidade sinaliza que, apesar da unidade já possuir um

> núcleo acional, ela não pode ser considerada terminada. Chamamos *stanza* esse tipo de unidade.[6]

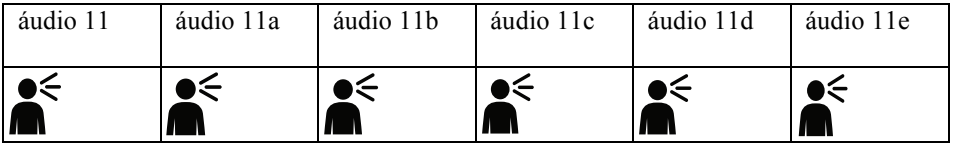
Disso, concluímos que a unidade terminada deve possuir duas propriedades: a presença de pelo menos uma ilocução e a marca prosódica de quebra terminal.

Exemplo 10 bfamdl02 [85-87]



*BAL: que cê acha colocar isso aqui //
*BEL: bom // deu certinho //

Exemplo 11 bfammn01 [19]



Que / na hora que ele lá envinha voltando do [/1] do comércio / que e' foi fazer a compra dele / **a cobra percebeu o cheiro dele** / na hora que ele lá envinha no [/1] no trilho //

O exemplo 10 mostra como enunciados simples são autônomos em virtude da capacidade de veicularem uma força ilocucionária. O exemplo 11 mostra que uma unidade ilocucionária é necessária para a constituição do enunciado (marcada em negrito), mas que é possível termos outras unidades não ilocucionárias que contextualizam a ilocução, sem ser interpretáveis em isolamento. Isso pode ser verificado escutando cada unidade sozinha nos áudios 11a, 11b, 11c, 11d (a única autônoma) e 11e. O exemplo 8 apresenta uma *stanza* com mais unidades ilocucionárias, quase todas, menos a última, acompanhadas por um sinal prosódico de continuidade.

---

[6] Para mais informações sobre a *Stanza*, além da bibliografia sobre a L-AcT já indicada, veja-se Cresti (2010). Veja-se a bibliografia sobre L-AcT para os casos, não mencionados aqui, de ilocuções padronizadas.

3 Relação entre pausas e quebras prosódicas

Até aqui realizamos uma discussão de natureza teórica, ilustrada por diversos exemplos, questionando a utilização da pausa como principal (ou único) parâmetro para identificar fronteiras de enunciados em *corpora* orais. Também discutimos o conceito de unidade de referência da fala e a importância de uma segmentação que represente adequadamente a realidade dos eventos de fala em *corpora* orais.

Nesta seção investigamos sistematicamente os resultados obtidos como consequência da utilização da pausa como critério de segmentação do fluxo da fala em um *corpus* oral. Tal investigação se justifica com o interesse crescente na compilação de *corpora* de fala. Decorre daí a necessidade de aprimorar técnicas de processamento dos dados que permita compilar *corpora* mais rapidamente e com maior fundamentação científica.

A divisão do contínuo da fala em unidades discretas pode ser realizada de forma manual ou automática. A segmentação manual é feita pelo exame do sinal acústico (onda, espectrograma, curva melódica). É um processo demorado, que requer muito treinamento aos segmentadores, além de ser sujeito aos erros naturalmente decorrentes de um processo manual. Na segmentação automática, o áudio é segmentado por uma máquina em unidades definidas operacionalmente. Apresenta como principais vantagens a rapidez e o fato de os erros serem previsíveis e sistemáticos, podendo ser, portanto, reduzidos. Esse sistema de segmentação é amplamente utilizado para a segmentação fonética da fala em unidades menores do que a palavra (SVENDSEN; SOONG, 1987; van HEMERT, 1991; BARBOSA, 2006), mas é ainda muito pouco aplicado para segmentar unidades maiores do que a palavra, ou seja, enunciados e unidades tonais.

A pausa é o parâmetro acústico que mais recebe atenção na identificação de fronteiras no fluxo da fala, sendo comumente associada a fronteiras de enunciados. A simplicidade de identificar interrupções no sinal acústico justifica sua utilização como principal parâmetro na segmentação automática da fala. Entretanto, considerados os exemplos contrários extraídos do *corpus*, nos perguntamos em que proporção as pausas correspondem a fronteiras prosódicas na fala espontânea, e se seria possível relacionar algum valor de duração de pausa a um

determinado tipo de fronteira prosódica (de enunciados ou unidades tonais).

Por meio de uma investigação sistemática do *corpus* C-ORAL-BRASIL, segmentado manualmente em enunciados e unidades tonais, procuramos identificar a coincidência entre pausas e fronteiras de enunciado; e / ou entre pausa e fronteira de unidade tonal interna ao enunciado. Adicionalmente verificamos se há algum valor de duração de pausa que coincida o máximo possível com as fronteiras de enunciados e o mínimo possível com fronteiras de unidades tonais. Desse modo, poderíamos produzir uma segmentação automática baseada em pausas que, pelo menos, reduzisse o trabalho dos segmentadores.

3.1 A fonte para a coleta dos dados

Os dados da pesquisa foram coletados no C-ORAL-BRASIL (RASO; MELLO, 2012). Para uma descrição detalhada da metodologia do *corpus*, veja-se Raso (2012b), Raso e Mello (2014) e Mello (2014). Para uma descrição da metodologia de compilação do C-ORAL-ROM, veja-se Moneglia (2005). Aqui resumimos as principais características do *corpus*, mostrando a natureza dos dados que embasam a pesquisa.

O C-ORAL-BRASIL representa a fala espontânea, definida como a fala planejada ao mesmo tempo em que é executada (NENCIONI, 1983). A arquitetura representa a variedade diatópica da área metropolitana de Belo Horizonte e privilegia a variação diafásica, com atenção também à diastratia. O *corpus* é de fala informal, com cerca de 210.000 palavras. Os textos têm, em média, 1.500 palavras, com poucos textos significativamente menores ou maiores. As situações dividem-se em 75% de contexto privado / familiar, 25% de contexto público. Cada contexto é dividido igualmente em monólogos, diálogos e conversações. A variação diafásica se completa com a maior variedade situacional possível. A razão para privilegiar a variação diafásica é evidente: em uma concepção da fala como comportamento verbalizado, o tipo de ação realizada é decisiva para a sua variação e estruturação. Portanto, a realização de ilocuções variadas é função da variedade dos comportamentos dos falantes e, consequentemente, da variedade das situações comunicativas.

A segmentação é feita em turnos e unidades terminadas.[7] Nessas, há ainda a segmentação em unidades tonais, *retractings* e unidades interrompidas. O texto e o sinal sonoro estão alinhados pelo *software* Winpitch (MARTIN, 2015), e a unidade de alinhamento corresponde ao que é considerada como unidade de referência da fala.

3.2 Metodologia

A amostra utilizada na pesquisa foi composta por trechos abrangendo nove quebras terminais de um mesmo falante, ignorando casos em que a quebra coincidia com fronteira de turno. Os trechos foram extraídos de cada um dos 139 textos do *corpus*, utilizando-se a técnica de amostragem aleatória simples sem reposição. A amostra analisada totalizou 1.251 fronteiras terminais e 2.580 fronteiras não terminais.

Para operacionalizar a análise, definiu-se como "silêncio" qualquer período de interrupção do sinal não provocado pela existência de consoantes desvozeadas, e como "pausa", qualquer duração convencionalizada de silêncio que pudesse ser utilizada para marcação de fronteiras em um modelo de segmentação automática da fala. Estabelecemos que silêncios de duração mínima de 10ms já seriam considerados pausas, definindo, a partir desse valor mínimo, intervalos de 10 em 10 ms, até chegar a 200 ms. O limite de 200 ms foi estabelecido porque, de modo geral na literatura, um silêncio de 200 ms é sempre considerado pausa, e frequentemente intervalos menores também são considerados pausas.

Com isso, convencionalizamos como pausas intervalos de silêncio maiores ou iguais a 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190 e 200 ms. Cada pausa encontrada na amostra foi medida e contabilizada de acordo com esses intervalos (uma pausa de 35 ms, por exemplo, é igual ou maior do que 10, 20 e 30 ms). Assim, podemos saber quantas fronteiras seriam identificadas na segmentação automática assumindo-se diferentes valores de duração das pausas.

---

[7] As unidades terminadas podem ser enunciados (possuem um único núcleo ilocucionário) ou *stanzas* (possuem mais de um núcleo ilocucionário).

O objetivo principal era verificar, para cada intervalo, quantas pausas coincidem com fronteiras julgadas por anotadores humanos como marcadas por quebras terminais e quantas delas coincidem com fronteiras marcadas por quebras não terminais. Adicionalmente, quantas fronteiras não coincidem com qualquer pausa conforme convencionalizada.

Para a análise acústica, foi utilizado o *software* Praat (BOERSMA; WEENINK, 2014); os dados foram anotados em planilhas.

Considerando-se um modelo de segmentação automática da fala com base em pausas, a ideia é verificar se existe um valor de duração mínimo de pausa que seria consistentemente associado à fronteira de tipo terminal. Assim, por exemplo, um silêncio observado no sinal de áudio que tivesse duração de 55ms seria utilizado para assinalar fronteira de enunciado caso a pausa fosse estabelecida tanto em 10ms quanto em 20ms, 30ms, 40ms ou 50ms. Já um silêncio observado de 12ms somente seria associado a fronteiras caso a pausa fosse estabelecida como sendo de 10ms.

Com isso, torna-se possível aferir qual intervalo mínimo apresenta a maior coincidência com as fronteiras de tipo terminal, já que sabemos que não é produtivo simplesmente realizar uma equivalência entre qualquer pausa e fronteiras de tipo terminal, como já demonstrado nos exemplos. Imaginando que a pausa pudesse ser considerada uma marca confiável de fronteira de enunciado, é necessário encontrar um determinado valor de duração de pausa que coincida, de maneira estatisticamente significativa, com as fronteiras de enunciados e, ao mesmo tempo, não coincida com as fronteiras de unidades internas ao enunciado (não terminais). E mesmo se aceitarmos uma taxa de coincidência não significativa estatisticamente, a metodologia de segmentação por pausa seria o mais eficaz possível se as pausas coincidissem o máximo possível com as fronteiras terminais, e o mínimo possível com as fronteiras não terminais.

Os percentuais de coincidência entre pausas e fronteiras terminais ou não terminais em cada ponto foram utilizadas para gerar dois modelos de regressão não linear (GRÁFICO 3), os quais atingiram coeficientes de predição de 0,99.

3.3 Resultados

Na Tabela 1, apresenta-se o cômputo do número de fronteiras terminais e não terminais que coincidiram com pausas de valor igual ou maior a cada duração de pausa delimitada. Por exemplo, na décima linha do quadro, identifica-se que foram detectadas 810 fronteiras terminais e 660 fronteiras não terminais coincidentes a silêncio de tamanho igual ou maior a 100ms. Isso significa que a quantidade de coincidências com fronteiras, a cada linha, é inclusiva (conforme explicado na seção anterior), já que cada valor de pausa é compreendido como a duração mínima de silêncio requerida para que um sistema de segmentação automática o identifique como uma fronteira de tipo terminal.

O que vemos é que, à medida que decresce a duração mínima de silêncio, tomado como referência para identificar uma pausa, aumenta a quantidade de fronteiras terminais coincidentes com pausas, mas, ao mesmo tempo, aumenta também a quantidade de fronteiras não terminais coincidentes com a pausa. Em um sistema de segmentação automático baseado em pausas, isso significaria um aumento do número de fronteiras não terminais erroneamente consideradas como fronteiras terminais. Por exemplo, se consideramos o maior valor de pausa (200ms) como a duração mínima de silêncio necessária para identificação de fronteira terminal, temos 746 coincidências com as fronteiras terminais e outros 587 casos de atribuição de fronteira terminal a fronteiras que, perceptualmente, foram julgadas como não terminais.

Tabela 1 – Número de fronteiras terminais e não terminais coincidentes com pausas

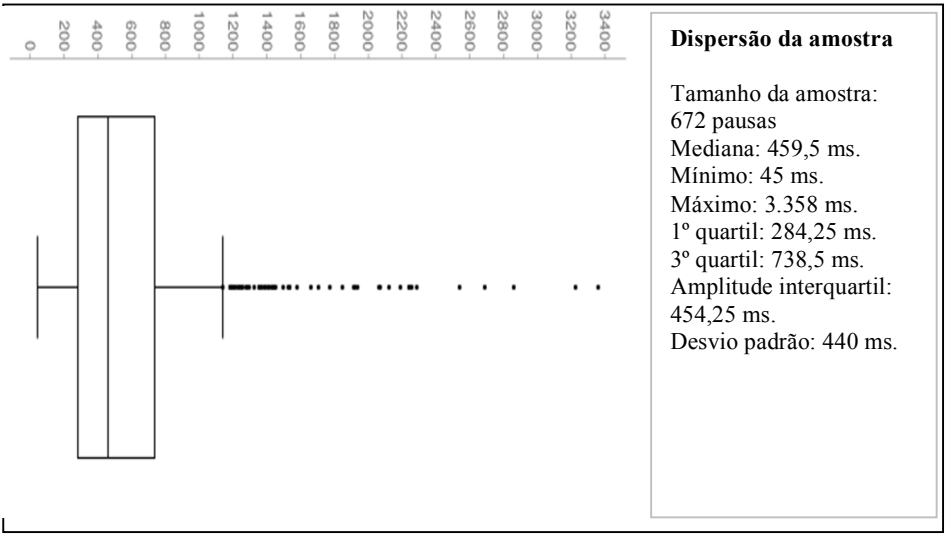
Duração da pausa	Coincidência terminais	Coincidência não- terminais
>= 10 ms	842	675
>= 20 ms	842	675
>= 30 ms	841	675
>= 40 ms	839	675
>= 50 ms	837	673
>= 60 ms	833	673
>= 70 ms	827	668
>= 80 ms	823	667
>= 90 ms	815	663
>= 100 ms	810	660
>= 110 ms	807	655
>= 120 ms	799	646
>= 130 ms	792	639
>= 140 ms	785	630
>= 150 ms	776	629
>= 160 ms	767	623
>= 170 ms	763	615
>= 180 ms	759	607
>= 190 ms	750	595
>= 200 ms	746	587

Se tomamos como parâmetro de segmentação a menor duração de pausa (10ms) como sendo o valor mínimo necessário para identificar uma fronteira terminal, conseguimos identificar no processo automático mais fronteiras terminais (842), mas, ao mesmo tempo, aumentamos também o número de fronteiras não terminais às quais seria erroneamente atribuído o estatuto de terminal (675). Assim, se por um lado temos um benefício de 96 reconhecimentos de fronteiras terminais a mais, por outro lado, temos um aumento de erros com relação às fronteiras não terminais

de 88 fronteiras. Isso sem considerar que a duração de silêncio de 10ms, aquela em que se obtém o melhor resultado na marcação das fronteiras terminais, é normalmente considerado insuficiente para a pausa. Ademais, mesmo aceitando um silêncio de duração tão pequeno como pausa, continuaríamos não capturando 33% das fronteiras que foram marcadas como terminais no C-ORAL-BRASIL, com uma grande consistência entre os anotadores, pois essas não coincidem com pausas convencionalizadas.

Das 1.251 fronteiras terminais analisadas, 409 não coincidem com nenhum valor de silêncio, o que equivale a 33%. Do total de 1.627 enunciados analisados, 913 enunciados apresentam fronteiras não conclusivas em seu interior, dos quais 565 (62%) não apresentam pausas.

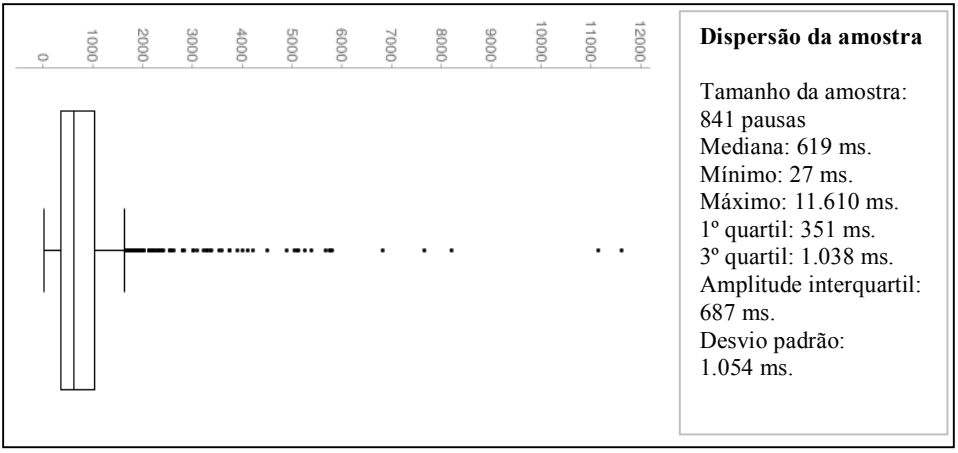
Gráfico 1 – Distribuição das durações de pausas coincidentes com fronteiras não terminais



O Gráfico 1 mostra as durações de pausas da amostra coincidentes com fronteiras identificadas como não terminais. Verifica-se, pelos valores de dispersão da amostra, que a menor duração de silêncio coincidente com fronteira não terminal foi de 45ms, e o valor máximo de silêncio coincidente com fronteira equivalente a 3.229 ms. Podemos

observar que há grande quantidade de pausas não coincidentes com fronteira terminal com duração de silêncio superior a 200ms.

Gráfico 2 – Distribuição das durações de pausas coincidentes com fronteiras terminais



O Gráfico 2 traz as durações de pausas coincidentes com fronteiras identificadas pelos transcritores do *corpus* como terminais. Nota-se que as pausas coincidentes com fronteiras terminais apresentam uma dispersão ainda maior do que aquelas coincidentes com fronteiras não terminais. Os valores de dispersão mostram que o valor mínimo de pausa coincidente com fronteira terminal foi de 27 ms, e o valor máximo equivalente a 11.610 ms. Comparando-se os dados dos Gráficos 1 e 2, podemos observar a grande sobreposição em termos das durações de pausas coincidentes com fronteiras não terminais e terminais. Metade (50%) das pausas tem durações entre 351 ms e 1.038 ms, o que engloba a maioria dos valores de pausa coincidentes com fronteiras não terminais. O desvio padrão é de 1.054 ms (praticamente 1 s de duração). Nota-se que as pausas na fala espontânea são distribuídas de modo bastante disperso, sendo, portanto, muito difícil correlacionar a duração com a marcação de um tipo particular de fronteira.

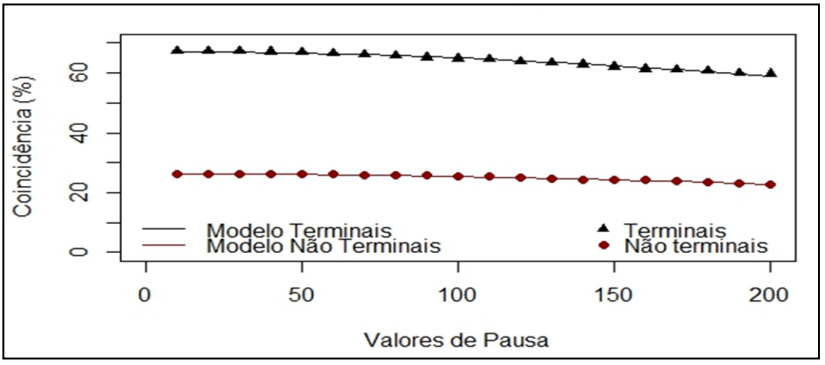
Para averiguar se há algum valor de duração de silêncio que possa ser utilizado como valor mínimo para marcação de fronteiras de

enunciados na fala espontânea, foram criados dois modelos não lineares que indicam a relação entre os diferentes valores mínimos convencionalizados para pausa e a taxa de coincidência com fronteiras de tipo terminal e não terminal.

As curvas expostas no Gráfico 3 descrevem a relação entre o tamanho convencionalizado da pausa e a coincidência entre pausa e fronteira terminal. Os ícones indicam os valores observados e as linhas contínuas, os modelos aproximados. Os modelos estimados (linhas contínuas) para o Gráfico 3 são representados pelas equações seguintes:

- para as fronteiras terminais: $y = 67{,}19 * \exp(-0{,}00000327*x^2)$
- para as fronteiras não terminais: $y = 26{,}33 * \exp(-0{,}00000353*x^2)$

Gráfico 3 – Percentual de coincidência entre pausas e fronteiras terminais e não-terminais, por duração mínimo de pausa



No Gráfico 3, a curva superior indica a relação entre pausas de diferentes durações e fronteiras terminais do *corpus*. Na vertical, temos as porcentagens de coincidência entre pausas e fronteiras. Nota-se que quando o valor da pausa tende a 0ms, há uma coincidência limite máxima de 67,2% entre fronteira terminal e pausa. Para a curva das fronteiras não terminais esse limite é de 26%. A partir daí, as duas curvas decaem.

Analisando o modelo para as fronteiras não terminais, observamos que a coincidência entre pausa e fronteira chega ao valor mínimo de 19,2% quando o valor mínimo de duração de pausa é de 300 ms ou mais.

A coincidência entre pausas e fronteiras não-terminais chega abaixo dos 5% somente com pausas com duração de, pelo menos, 687 ms, resultado consistente com a distribuição das durações de pausas mostradas no Gráfico 1, no qual observamos que 75% das pausas têm duração de até 738,5 ms.

Ainda no Gráfico 3, para o valor de pausa maior ou igual a 100ms, observa-se 65% de coincidência com as fronteiras terminais e 25,4% de coincidência com as fronteiras não-terminais no *corpus*. A segmentação, segundo esse critério, deixaria de identificar, portanto, 35% das fronteiras terminais existentes e marcaria como terminais 25,41% das fronteiras que, na nossa segmentação, são não terminais.

Verifica-se, então, que se variando o tamanho de pausa, as duas curvas apresentam comportamentos semelhantes: ao aumentar o tamanho de pausa, aumentam as coincidências nos dois casos; ao diminuir o tamanho de pausa, ocorre o inverso. Logo, não é possível estabelecer uma duração de pausa que tenha a máxima coincidência com fronteiras terminais e a mínima coincidência com fronteiras não terminais.

4 Conclusão

Os dados expostos indicam que

i. o critério de segmentação por pausa não se mostra adequado para a segmentação de textos de fala espontânea. Não somente os exemplos mostrados nesse trabalho, mas também os modelos estatísticos realizados sobre amostra aleatória extraída do *corpus* mostram que não há uma real correspondência entre pausas e fronteiras de enunciados e unidades menores (já validadas quanto ao acordo entre anotadores); a pausa não parece constituir a única (e talvez nem seja a principal) estratégia de marcação de fronteiras na fala espontânea; e

ii. não existe uma duração de pausa que possa ser eficazmente empregada para marcação de fronteira terminal de unidade de referência. O aumento na captura de fronteiras terminais, dada uma certa duração de pausa, corresponde também a superestimativa de fronteiras, e o aumento no reconhecimento de fronteiras não terminais corresponde a subestimativa de fronteiras.

As conclusões desta pesquisa são, a nosso ver, muito claras: qualquer definição de pausa relacionada a um correlato físico de silêncio não pode ser automaticamente considerada marca de fronteira de unidade de referência da fala. O conceito de quebra prosódica é mais eficiente para essa função demarcativa tão importante. Contudo, a questão da segmentação da fala não se conclui aqui. Precisamos entender melhor como se realiza a quebra prosódica em termos de seus correlatos físicos. Sabemos que as quebras possuem uma fortíssima saliência perceptual, mas não sabemos exatamente de que maneira atuam os correlatos físicos que as produzem.

Nesse sentido, um desdobramento natural, já encaminhado do trabalho apresentado aqui, é o desenvolvimento de um modelo computacional que considere diversos parâmetros acústicos utilizados na identificação de fronteiras. Com uma abordagem *corpus driven* e o conceito de aprendizagem de máquina, visa-se compreender melhor as configurações físicas das quebras, mediante uma amostra de fala espontânea segmentada em enunciados e unidades tonais por múltiplos juízes, da qual seriam extraídos dados relativos a diversos parâmetros acústicos (por exemplo, F0 média de sílaba pré e pós fronteiriça e duração silábica). Com base nos resultados obtidos será possível delinear as características prosódicas das fronteiras e calcular sua probabilidade de ocorrência.

Acreditamos que uma resposta, mesmo que parcial, à pergunta acerca das propriedades físicas constitutivas das fronteiras teria dois efeitos fundamentais para o estudo da fala: o primeiro seria permitir uma segmentação (semi)automática de *corpora* de fala. Para quem sabe quão grande é o trabalho de compilação de *corpora* de fala, é fácil imaginar a vantagem que isso traria: seria muito mais viável compilar *corpora*, não de poucas centenas de milhares de palavras, mas de milhões de palavras,

com os óbvios reflexos disso para a pesquisa em diferentes áreas. O segundo efeito seria de natureza teórica: conhecer mais sobre os correlatos físicos das fronteiras entre as unidades de referência da fala abriria perspectivas importantes para compreender melhor a percepção da fala e, por consequência, seus mecanismos neurolinguísticos.

Referências

ALBANO LEONI, F.; MATURI, P. *Manuale di fonetica*. Roma: Carocci, 2002.

AUSTIN, J. *How to do things with words*. London: Oxford University Press, 1962.

BARBOSA, P. A. *Incursões em torno do ritmo da fala*. São Paulo: Pontes, 2006. (Fapesp).

BERRUTO, G. Varietà diamesiche, diastratiche, diafasiche. In: A SOBRERO, A (Org.). *Introduzione all'italiano contemporaneo***:** La variazione e gli usi. Roma: Laterza, 1993. p. 37-92.

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

BOERSMA, P.; WEENINK, D. *Praat*: doing phonetics by computer. 2014. Programa de computador, versão 5.4.19. Disponível em: <http://www.praat.org/>. Acesso em: 16 set. 2015.

BUHMANN, Jeska *et al*. Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 3, 2002, Las Palmas. *Proceedings...* Las Palmas: Elra, Universidad de Las Palmas de Gran Canaria, 2002. p. 1 - 7. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2002/pdf/96.pdf>. Acesso em: 20 set. 2015.

CARLSON, R.; HIRSCHBERG, J.; SWERTS, M. Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication***,** Nara, v. 46, n. 3-4, p. 326-333, jul.

2005. Elsevier BV. Disponível em: <http://api.elsevier.com/content/article/PII:S0167639305000932?httpAccept=text/xml>. Acesso em: 22 set. 2015.
DOI: <http://dx.doi.org/10.1016/j.specom.2005.02.013>

CAVALCANTE, F.; RAMOS, A. Um *minicorpus* de inglês americano etiquetado informacionalmente, em preparação.

CHOMSKY, N. Remarks on nominalization. In: JACOBS, R.; ROSENBAUM, P. (Org.) *Reading in English Transformational Grammar*. Waltham: Ginn, 1970. p. 184-221.

CRESTI, E. *Corpus di Italiano parlato*. v. 1. Firenze: Accademia della Crusca, 2000.

CRESTI, E., La Stanza: un'unità di costruzione testuale del parlato. In: CONGRESSO DELLA SOCIETÀ INTERNAZIONALE DI LINGUISTICA E FILOLOGIA ITALIANA, 10, 2008, Basilea. *Atti*… Org.: A. Ferrari. Firenze: Cesati, 2010. p. 713 - 732.

CRESTI, E. Syntactic properties of spontaneous speech in the Language into Act Theory. In: RASO, T.; MELLO, H. *Spoken Corpora and Linguistic Studies*. Amsterdam/Philadelphia: John Benjamins, 2014. p. 365-410.
DOI: <http://dx.doi.org/10.1075/scl.61.13cre>

CRESTI, E.; GRAMIGNI, P. Per una linguistica corpus based dell'italiano parlato: Le unità di riferimento. In: CONVEGNO NAZIONALE "IL PARLATO ITALIANO", Napoli. *Atti*… Ed.: F. Albano Leoni, *et al*. Napoli: D'Auria, CD-ROM, 2004. p. 1–26.

CRESTI, E.; MONEGLIA, M. (Orgs.). *C-ORAL-ROM: Integrated reference corpora for spoken Romance Languages*. Amsterdam/Philadelphia: John Benjamins, 2005.
DOI: <http://dx.doi.org/10.1075/scl.15>

CRYSTAL, D. *The English tone of voice*. London: Edward Arnold, 1975.

DU BOIS, J. *et al*. *Santa Barbara Corpus of Spoken American English,* Parts 1-4. Philadelphia, PA: Linguistic Data Consortium, 2000–2005.

FLEISS, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, v. 76, p. 378-382, 1971.
DOI: <http://dx.doi.org/10.1037/h0031619>

HARRIS, Z. S. *String analysis of sentence structure***.** The Hague: Mouton, 1962.

HARRIS, M.; UMEDA, N; BOURNE, J. Boundary perception in fluent speech. *Journal of Phonetics*. n. 9, p.1-18, 1981.

IZRE'EL, S.; HARY, B.; RAHAV, G. Designing CoSIH: The Corpus of Spoken Israeli Hebrew. *International Journal of Corpus Linguistics*, n. 6, p. 171-197, 2001.

MARTIN, Philippe. Winpitch Pro W8. v 7.2.00. Pitch Instruments. Disponível em: <http://www.winpitch.com/>, 2015. Acesso em: 25 set. 2015.

MARUYAMA, T. Speech segmentation by clausal and non-clausal boundaries in Japanese. INTERNATIONAL CONFERENCE ON COGNITIVE SCIENCE, 5, Kaliningrad. *Abstracts… V. 2*: Workshop Spoken discourse corpora as a window on cognitive mechanisms of speech production. Kaliningrad, June 2012. p. 787-783

MARUYAMA, T. Two-Level Utterance Units: Cognitive and Communicative Aspects of Spontaneous Speech. LABLITA, 9; LEEL, 4, INTERNATIONAL WORKSHOP, Belo Horizonte, 2015. *Conferências*... Units of Reference for Spontaneous Speech Analysis and their correlations across languages, Belo Horizonte, August 6, 2015.

MELLO, H. R., Methodological issues for spontaneous speech corpora compilation: The case of C-ORAL-BRASIL. In: RASO, T.; MELLO, H. R. *Spoken Corpora and Linguistic Studies.* Amsterdam/Philadelphia: John Benjamins, 2014. p. 27-68.
DOI: <http://dx.doi.org/10.1075/scl.61.01mel>.

MELLO, H. R. *et al*. Transcrição e segmentação prosódica do *corpus* C-ORAL-BRASIL: critérios de implementação e validação. In: RASO, T.; MELLO, H. R. (Ed.) *C-ORAL – Brasil I*: *Corpus* de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012. p. 125-176.

MELLO, H.; RASO, T.; BOSSAGLIA, G.; SANTANA, T. Enunciados e estruturas de predicação no corpus C-ORAL-BRASIL. em preparação.

MERLO, S.; BARBOSA, P. A. Séries temporais de pausas e de hesitações na fala espontânea. *Caderno de Estudos Linguísticos.* v. 1, n. 54, p. 11-24, 2012. Disponível em: <http://revistas.iel.unicamp.br/index.php/cel/article/view/2569>. Acesso em: 20 set. 2015.

METTOUCHI, A.; CHANARD, C. From Fieldwork to Annotated Corpora: the CorpAfroAs Project. *Faits de Langue-Les Cahiers.* n. 2, p. 255-265, 2010.

MITTMANN, M. M. O C-ORAL-BRASIL e o estudo da fala informal: um novo olhar sobre o Tópico no Português Brasileiro. Tese - (Doutorado em Estudos Linguísticos) Faculdade de Letras da Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, 2012, 248 p., Belo Horizonte, 2012. 248 p. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/ handle/1843/LETR-97YMKT>. Acesso em: 20 set. 2015.

MITTMANN, M. M. Análise da estruturação de diálogos e monólogos na fala informal: quantificando as diferenças, *Domínios de Lingu@gem*, v. 7, p. 338-372, 2013.

MITTMANN, M.; RASO, T. The C-ORAL-BRASIL Informationally Tagged Mini-Corpus. In: MELLO, H. R.; PANUNZI, A.; RASO, T. (Org.). *Illocution, modality, attitude, information patterning and speech annotation*. Firenze: Firenze University Press, 2012. p. 151-183.

MO, Y.; COLE, J.; LEE, E. Naïve listeners' prominence and boundary perception. In: SPEECH PROSODY INTERNATIONAL CONFERENCE, 4, 2008, Campinas. *Proceedings...* Org.: P. A. BARBOSA; S. MADUREIRA; C. REIS (Org.). Campinas: Isca, 2008. p. 735 - 738. Disponível em: <http://www.isca-speech.org/archive/sp2008/sp08_735.html>. Acesso em: 20 set. 2015.

MONEGLIA, M. The C-ORAL-ROM resource. In: CRESTI, E.; MONEGLIA, M. (Org.). *C-ORAL-ROM*: integrated reference corpora for spoken Romance languages. Amsterdam/Philadelphia: John Benjamins, 2005. p. 1-70. DOI: <http://dx.doi.org/10.1075/scl.15.03mon>

MONEGLIA, M. Spoken corpora and pragmatics. *Revista Brasileira de Linguística Aplicada*, Belo Horizonte, v. 11, n. 2, p. 479-519, 2011. Disponível em: <http://www.periodicos.letras.ufmg.br/rbla/arquivos/335.pdf>.

MONEGLIA, M.; CRESTI, E. L´intonazione e i criteri di trascrizione del parlato adulto e infantile. In: BORTOLINI, U.; PIZZUTO, E. (Org.). *Il Progetto CHILDES Italia***.** Pisa: Del Cerro, 1997. p. 57–90.

MONEGLIA, M. *et al.* Evaluation of consensus on the annotation of terminal and non-terminal prosodic breaks in the C-ORAL-ROM *corpus*. In: CRESTI, E.; MONEGLIA, M. (Org.). *C-ORAL-ROM*: integrated reference corpora for spoken Romance languages. Amsterdam: John Benjamins, 2005. p. 257-276.

MONEGLIA, M.; RASO, T. Notes on Language into Act Theory. In: RASO, T; MELLO, H. R. *Spoken Corpora and Linguistic Studies*. Amsterdam / Philadelphia: John Benjamins, 2014. p. 468-495.

MONEGLIA, M. *et al*. Challenging the perceptual relevance of prosodic breaks in multilingual spontaneous speech corpora: C-ORAL-BRASIL/C-ORAL-ROM. In: SPEECH PROSODY, Chicago. *Proceedings*... Chicago, 2010. p. 1-4.

NENCIONI, G. *Di scritto e di parlato*: discorsi linguistici. Bologna: Zanichelli, 1983.

NESPOR, M.; VOGEL, I. *Prosodic phonology.* Dordrecht: Foris, 1986.

PANUNZI, A.; MITTMANN, M. The IPIC resource and a cross-linguistic analysis of information structure in Italian and Brazilian Portuguese. In: RASO, T.; MELLO, H. R. Spoken Corpora and Linguistic Studies. Amsterdam / Philadelphia: John Benjamins, 2014. p. 129-151. DOI: <http://dx.doi.org/10.1075/scl.61.05pan>.

RASO, T. Minicorpus retirado do corpus C-ORAL-BRASIL e etiquetado informacionalmente. Annotation of information patterns according to Language into Act Theory in DB IPIC - first release, 2012a. Disponível em: <http://lablita.dit.unifi.it/ipic/>.

RASO, T. O corpus C-ORAL-BRASIL. In: RASO, T.; MELLO, H. R. (Org.). *C-ORAL-BRASIL I*: *Corpus* de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012b. p. 55-90.

RASO, T.; O C-ORAL-BRASIL e a Teoria da Língua em Ato. In: RASO, T; MELLO, H. R. (Org.). *C-ORAL-BRASIL I*: *Corpus* de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012c. p. 91-124.

RASO, T. Fala e escrita: meio, canal, consequências pragmáticas e linguísticas. *Domínios de Lingu@gem*, v. 7, p. 12-46, 2013.

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

RASO, T.; MITTMANN, M. As principais medidas da fala. In: RASO, T.; MELLO, H. R. (Org.). *C-ORAL – BRASIL I*: *Corpus* de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012. p. 177-222.

RASO, T.; MITTMANN, M. Validação estatística dos critérios de segmentação da fala espontânea no *corpus* C-ORAL-BRASIL. *Revista de Estudos da Linguagem*, v. 17, p. 73-92, 2009. DOI: <http://dx.doi.org/10.17851/2237-2083.17.2.73-91>.

ROSSI, F. La variazione diamesica. In: SIMONE, R. (Org.). *Enciclopedia dell'Italiano*. Milano: Treccani, 2011. Disponível em: <http://www.treccani.it/enciclopedia/variazione-diamesica_(Enciclopedia_dell'Italiano)/>. Acesso em: 20 set. 2015.

SCHUETZE-COBURN, S; SHAPLEY, M; WEBER, E. G. Units of intonation in discourse: a comparison of acoustic and auditory analyses. *Language and speech*. v. 34, n. 3, p. 207-234, 1991. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/1843524.0023830991034>. Acesso em: 20 set. 2015.

SORIANELLO, P. *Prosodia.* Roma: Carocci, 2006.

SVENDSEN, T.; SOONG, F. On the automatic segmentation of speech signals. IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING - ICASSP '87, Dallas, 1987.

Conferences... Dallas, US: Institute of Electrical & Electronics Engineers (IEEE), 1987. p. 77-80. Disponível em: <http://dx.doi.org/10.1109/ICASSP.1987.1169628>. Acesso em: 20 set. 2015.

SWERTS, M.; COLLIER, R.; TERKEN, J. Prosodic predictors of discourse finality in spontaneous monologues. *Speech Communication*, v. 15, n. 1-2, p. 79-90, 1994. Disponível em: <http://www.sciencedirect.com/science/article/pii/0167639394900434>. Acesso em: 20 set. 2015.

VAN HEMERT, J. P. Automatic segmentation of speech. *Ieee Transactions On Signal Processing***,** Piscataway, v. 39, n. 4, Institute of Electrical & Electronics Engineers (IEEE), p. 1008-1012, abr. 1991. DOI: <http://dx.doi.org/10.1109/78.80941>.

ZELLNER, B. Pauses and the temporal structure of speech. In: KELLER, E. (Org.) *Fundamentals of speech synthesis and speech recognition*. Chichester: John Wiley, 1994. p. 41-62.