A importância dos recursos lexicais para o processamento automático do português[1]

(The importance of lexical resources for automatic processing of Portuguese)

Magali Sanches Duran[1]

[1]Núcleo Interinstitucional de Linguística Computacional – Universidade de São Paulo (NILC-ICMC-USP São Carlos)

magali.duran@uol.com.br

Abstract: The aim of a lexical description constrains its form and even its content. For this, in spite of Portuguese lexical description presenting an advanced stage in what concerns lexical resources for native speakers, there are challenging gaps in lexical resources for foreign learners and for computational use. Under this assumption, this paper discusses the need of lexical description for the specific purpose of automatic processing of Portuguese. The aim is to inspire lexical research that meets the growing demand of such area. The ideas are presented in a simple way in order to make the reading accessible to linguistic researchers unfamiliar with natural language processing.

Keywords: computational linguistics; computational lexicon; lexical resources.

Resumo: A finalidade da descrição do léxico condiciona a forma e até o conteúdo da descrição. Por isso, embora a descrição do português encontre-se em um estágio avançado no que se refere a recursos léxicos para informar falantes nativos, há ainda lacunas desafiadoras no que diz respeito a recursos léxicos para informar aprendizes estrangeiros e sistemas computadorizados. Partindo desse pressuposto, discute-se neste artigo a necessidade de descrever o léxico para fins de processamento automático do português, com o objetivo de inspirar novos trabalhos que visem a atender a crescente demanda dessa área. Procura-se abordar o tema de forma simples a fim de tornar a leitura acessível a pesquisadores do léxico não familiarizados com o processamento automático de línguas naturais.

Palavras-chave: linguística computacional; léxico computacional; recursos léxicos.

Introdução

As novas tecnologias desenvolvem muitos produtos que embutem uma língua natural. Isso exige que as máquinas estejam preparadas para processar essa língua, interpretando novos insumos – texto escrito ou fala e realizando operações para dar respostas aos usuários. Além disso, a capacidade de armazenamento de dados em meios digitais acarreta um aumento vertiginoso do volume desses dados, o que torna impossível explorá-los e analisá-los apenas com recursos humanos. O processamento automático de línguas naturais (PLN) desenvolve-se como resposta a essas necessidades.

A história do PLN está ligada à história dos computadores e as pesquisas na área foram fomentadas principalmente após a Segunda Guerra Mundial. Se hoje é possível fazer processamentos sofisticados, devemos isso a décadas de trabalho buscando superar as dificuldades de processar línguas naturais.

Muitas tarefas de PLN são pré-requisito para a execução de outras tarefas. Para processar uma língua escrita (a língua falada exige muitos outros pré-requisitos), é neces-

1 A autora deste trabalho agradece o apoio da Fapesp (processo n. 2011-22337-1).

sário, primeiramente, que a máquina reconheça seu alfabeto, os limites de suas unidades lexicais (tarefa conhecida em PLN como tokenização), seja capaz de reduzir as diversas formas flexionadas desses itens a suas formas canônicas (tarefa conhecida como lematização), seja capaz de atribuir rótulos com a classificação morfossintática dessas unidades lexicais (tarefa conhecida como *POS tagging*[2]), seja capaz de reconhecer sintagmas nominais, sintagmas preposicionados e sintagmas verbais, além de reconhecer os limites das sentenças. Só depois de superadas as dificuldades impostas por essas tarefas mais básicas é que o PLN pode se enveredar por tarefas mais complexas, como atribuir rótulos de análise sintática (tarefa conhecida como *parsing*), atribuir rótulos de papéis semânticos (*semantic role labeling* ou SRL) e resolver correferências (*co-reference resolution*), entre outras.

A qualidade da execução dessas tarefas exerce impacto nas grandes tarefas de PLN, como tradução automática, sumarização mono e multidocumento, simplificação textual, sistemas de perguntas e respostas, análise de opiniões e sentimentos.

Até duas décadas atrás, a principal área a fornecer insumos para PLN era a linguística. No início, a abordagem usada pelo PLN era empregar regras definidas por linguistas para que a máquina imitasse o mesmo raciocínio de um humano ao desempenhar uma tarefa (analisar, classificar, traduzir, ler, etc.). Contudo, esse caminho mostrou-se muito lento para atender às necessidades tecnológicas. Por isso, desenvolveu-se outra abordagem: uma vez fornecido um exemplo da tarefa (um *corpus* paralelo, por exemplo, no caso da tradução), os profissionais da computação passaram a utilizar recursos estatísticos para inferir regras que, uma vez automatizadas, produzissem um resultado semelhante ao produzido pelo trabalho humano. A máquina passou a "aprender" a língua não mais com base em regras linguísticas, mas com base em exemplos.

A construção de grandes corpora foi essencial para a adoção dessa abordagem. E para aumentar a possibilidade de aprender características da língua em um *corpus*, o PLN passou a requisitar a anotação de *corpus*, ou seja, a atribuição de rótulos ou etiquetas que refletissem a análise do texto. Assim, por exemplo, ao invés de se definir regras para a análise sintática, anotadores com competência em análise sintática são contratados para identificar os segmentos de texto que correspondem a cada papel sintático (cada papel sintático corresponde a uma etiqueta atribuível no processo de anotação sintática de um *corpus*). É claro que, para ser processável computacionalmente, a anotação tem que ser feita com o uso de ferramentas automáticas de anotação (v. DURAN et al., 2010). Uma vez anotado, o *corpus* passa a ser utilizado como *corpus* de treinamento para a máquina inferir regras para atribuição automática das etiquetas.

Por serem apoiados em corpora, os métodos estatísticos são muito bons no que diz respeito a apreender características da dimensão sintagmática da língua. As características da dimensão paradigmática, contudo, onde reside a maior parte do conhecimento léxico, praticamente não é acessada por esses métodos.

A necessidade de recursos léxicos para o PLN é reconhecida mundialmente, tanto por pesquisadores da linguística quanto da computação. As diversas línguas, contudo, se encontram em diferentes estágios no que diz respeito à descrição do léxico para esse fim. No caso do português, embora já existam pesquisas nessa área, ainda há muitas lacunas a serem preenchidas. Para que a tecnologia "fale" nossa língua, é fundamental que a comu-

2 Etiquetador de partes do discurso (POS = *part of speech*)

nidade de linguistas dedicada a essa tarefa se amplie. Espero, com este artigo, contribuir nesse sentido.

Além desta breve introdução, organizo o artigo em quatro seções. Na primeira comento dificuldades de ordem lexical enfrentadas por tarefas típicas de PLN. Na segunda discorro sobre a construção de grandes repositórios lexicais e apresento questões em aberto sobre o léxico do português, que podem inspirar trabalhos de diferentes graus de complexidade. Na terceira são discutidas as contribuições que o processamento automático pode trazer para a pesquisa linguística e, na última, teço algumas considerações finais.

O papel do léxico nas tarefas de PLN

A necessidade de descrição do léxico aparece nas mais diversas tarefas de PLN. Algumas tarefas são mais dependentes desse conhecimento, como a classificação morfossintática, outras menos, como a análise sintática. Nem sempre é óbvio, contudo, o tipo de conhecimento léxico necessário para cada tarefa. Nesta seção procuro ilustrar diferentes tarefas de PLN e suas respectivas demandas em relação à descrição do léxico do português.

Primeiramente é preciso entender o que diferencia a descrição do léxico para fins computacionais. A descrição de uma língua deve levar em conta o pré-conhecimento de quem vai utilizar essa descrição. Por exemplo, no caso de falantes nativos, que adquiriram a língua antes de ter contato com sua descrição, é possível descrever a noção de gênero das palavras a partir de exemplos: "a árvore" é feminino e "o galho" é masculino. A partir daí, usando de inferência, o falante nativo poderá listar todas as palavras de seu léxico pessoal que pertencem ao gênero feminino e todas que pertencem ao gênero masculino.

Já no caso de aprendizes estrangeiros, a situação é outra, pois eles podem ou não ter adquirido a noção de gênero em suas respectivas línguas maternas. Mesmo que tenham essa noção, a tarefa não é tão simples, pois nem sempre o gênero das palavras coincide em diferentes línguas.

No caso da máquina, não há pré-conhecimento que possa ser convocado para a tarefa e, por isso, o léxico precisa ser exaustivamente descrito. Portanto, enquanto um dicionário para fins humanos traz apenas as formas lematizadas dos itens lexicais (masculino singular, para os nomes e infinitivo, para os verbos), os dicionários de máquina devem incluir todas as formas flexionadas dos itens lexicais.

Há, contudo, necessidades menos óbvias de conhecimento léxico no PLN. Um leigo poderia pensar que é muito simples para a máquina, por exemplo, separar um texto em sentenças, pois bastaria observar os pontos finais. Contudo, se o léxico das abreviaturas não estiver descrito, toda vez que aparecer um "Sr." ou um "Dr." na sentença a máquina interpretará como um ponto final (neste exato momento o corretor automático de meu editor de textos cometeu esse erro ao alterar para maiúscula as iniciais das palavras "ou" e "na" que aparecem após os pontos das abreviaturas, alteração que rejeitei, obviamente). Além disso, iniciais de nomes próprios, que não são previsíveis no léxico, impõem dificuldades adicionais a essa tarefa.

A possibilidade de os itens lexicais terem mais de uma função e mais de um sentido é outro grande desafio para as tarefas de PLN, pois gera ambiguidade. Ainda hoje itens lexicais que não costumam ser ambíguos para o ser humano ainda o são no PLN.

Por exemplo, os lematizadores de português (programas que transformam as formas flexionadas das palavras em suas formas canônicas) ainda têm problemas em distinguir a forma *foi* do verbo *ir* da forma *foi* do verbo *ser*. Sob a ótica da desambiguização para fins computacionais, aliás, o estudo dos homônimos ganha maior abrangência. Tanto que um tópico bastante estudado em PLN são os métodos de desambiguação lexical (WSD ou *word sense disambiguation*).

Os analisadores morfossintáticos automáticos, chamados de *POS-taggers* lidam com o problema da ambiguidade "observando" pistas de contexto para decidir a classe de uma palavra. Vejamos, hipoteticamente, a dificuldade de se determinar a classe morfológica de uma palavra como *canto* utilizando pistas do nível sintático (por isso se fala em classificação morfossintática).

Canto pode ser substantivo ou verbo (primeira pessoa do presente do indicativo do verbo *cantar*) e, a princípio, teríamos as seguintes regras:

Primeira regra: se *canto* for precedido de um item lexical das classes dos determinantes (artigos, numerais, pronomes adjetivos), *canto* deve ser classificado como substantivo como mostrado em (1) e (2):

(1) O **canto** $_{\text{substantivo}}$ da cotovia é lindo.

(2) Aquele **canto** $_{\text{substantivo}}$ da sala está sujo.

Segunda regra: se *canto* for precedido do pronome *eu, canto* deve ser classificado como verbo, como mostrado em (3):

(3) Eu **canto** $_{\text{verbo}}$ quando estou feliz.

Contudo, essas regras podem não ser suficientes. Vejamos o exemplo (4), no qual a regra acarreta um erro de classificação:

(4) Esse hino, só o **canto** $_{\text{substantivo}}$ aos domingos. INCORRETO

Para contemplar casos assim, uma nova regra precisaria ser definida:

Terceira regra: na ausência de outro verbo na oração, *canto* deve ser classificado como verbo, o que corrige o erro em (5):

(5) Esse hino, só o **canto** $_{\text{verbo}}$ aos domingos. CORRETO

Mas existem casos em que a ambiguidade é inerente à língua, como no exemplo a seguir:

(6) **Canto** é para espantar a tristeza.

Há duas possibilidades de interpretação:

(7) (O) **canto** $_{\text{substantivo}}$ é para espantar a tristeza.

(8) (Eu) **canto** $_{\text{verbo}}$ é para espantar a tristeza.

Mesmo que novas regras ajudassem a decidir por uma das possíveis interpretações, a máquina estaria eliminando uma ambiguidade que, para os humanos, é real e insuperável sem pistas de um contexto maior.

Outra tarefa que também não é tão simples quanto pode parecer a princípio é o reconhecimento dos limites das unidades lexicais. Embora o espaço entre um grupo de letras seja o separador óbvio dessas unidades, não é um critério útil quando se trata de multipalavras (*multi-word units* ou MWU), ou seja, unidades lexicais compostas por mais de um item lexical e que correspondem a um único significado. A quantidade desse tipo de unidade lexical nas línguas é muito maior do que prevê o senso comum e, por isso, o tema ganhou grande relevância em PLN (v. SAG et al., 2002). A identificação de multipalavras em grandes corpora conta hoje com ferramentas automáticas, como o MWToolkit (RAMISCH et al., 2010), que tomam várias medidas como parâmetro, principalmente a frequência com que dois ou mais itens lexicais ocorrem juntos. É claro que é preciso um olhar humano para filtrar os resultados obtidos com o uso dessas ferramentas, mas a máquina faz o "grosso" do trabalho, processando uma quantidade de textos que seria impossível para um humano processar. Dois tipos de multipalavras que afetam bastante o PLN são comentados a seguir.

Tanto a análise sintática quanto a análise semântica automática ainda sofrem a falta de conhecimento sobre o léxico das locuções adverbiais iniciadas por preposição, como *a torto e a direito*, *de repente*, *para cima*, *em silêncio.* Esse léxico tem sido objeto de alguns estudos (GARRÃO et al., 2008; PALMA, 2009), mas ainda não existe nenhum recurso disponível em que essas locuções estejam listadas e classificadas segundo sua função semântica. Essas multipalavras são importantes porque a preposição, como observou Fillmore (1968) é um importante marcador de caso nas línguas não desinenciais, o que gera ambiguidade entre argumentos previstos e modificadores introduzidos por preposição (v. VILLAVICÊNCIO, 2002). Assim, ao analisar a ocorrência "reclamar de [X]", por exemplo, os analisadores automáticos classificam tudo que aparece no lugar de [X] como objeto indireto (análise sintática) e tema (análise de papéis semânticos) do verbo *reclamar*, como é o caso de *dor-de-cabeça* em "reclamar de dor-de-cabeça". Contudo, para qualquer falante nativo é possível perceber que "reclamar de vez em quando" não pode ser analisado segundo essa mesma regra, ou seja, "vez em quando" não é nem objeto indireto nem tema do verbo *reclamar*. Para que a máquina possa tratar adequadamente casos como esse, é preciso que a sequência "de vez em quando" esteja descrita como uma locução adverbial de tempo.

Também impactante para os analisadores automáticos é o léxico dos predicados complexos (ALSINA et al., 1997), que incluem as construções suporte, cujo significado é composicional e verbos multipalavras (não composicionais ou idiomáticos). Isso porque, por terem significado único, em muitas tarefas esses predicados devem ser tratados como um único verbo. Exemplo do primeiro é *dar queixa de* (queixar-se de) e exemplo do segundo tipo é *dar conta de* (conseguir, ser capaz). Embora haja muitos estudos sobre construções de verbo suporte no português (NEVES, 1996; ATHAYDE, 2001; CONEJO, 2008; SILVA, 2009; HENDRICKX et al., 2010; DUARTE et al., 2010; ABREU, 2011), esses predicados ainda não foram exaustivamente descritos e estabelecer critérios para classificá-los nem sempre é uma tarefa simples, como discutido em Butt, (2003) e Duran et al. (2011).

Os benefícios do PLN para os linguistas

O trabalho de levantamento do léxico para melhorar o processamento automático de uma língua não beneficia apenas o PLN. Vejo duas grandes vantagens para o linguista em trabalhar nessa linha. A primeira delas é que a descrição do português para fins computacionais foca aspectos diferentes daqueles focados na descrição para fins humanos. Para fins humanos, muitos conhecimentos são ignorados (ficam num nível epilinguístico), pois são compartilhados por todos os falantes da língua ou por todos os falantes de línguas naturais. Para fins computacionais, todo conhecimento é necessário, mesmo aquele sobre aspectos que ainda não foram objeto de estudos linguísticos. Nesse sentido, o trabalho em PLN pode inspirar novos estudos linguísticos. Quando um analisador automático é construído, a análise de seus erros evidencia que tipo de conhecimento linguístico está faltando para melhorar seu desempenho. E muitas vezes esse conhecimento faltante nunca foi levantado.

A segunda vantagem é que, uma vez construídos os analisadores automáticos de texto, as pesquisas em *corpus* de língua passam a contar com um leque muito maior de argumentos de busca automática. Por exemplo, se um *corpus* está analisado morfossintática e sintaticamente, é possível definir buscas muito mais precisas do que aquelas que usam apenas palavras-chave, pois todas as etiquetas que enriquecem o *corpus* podem servir de argumentos busca. Essa é a realidade dos corpora acessíveis pelo buscador AC/DC (SANTOS; SARMENTO, 2002) disponível na Linguateca.[3] É possível, por exemplo, pesquisar todas as ocorrências de sentenças que contenham orações subordinadas reduzidas de gerúndio.

Essa também é a vantagem do PLN para os buscadores automáticos de conteúdo na web. Se um buscador eletrônico "varre" um *corpus* para procurar determinada palavra-chave, ele traz um determinado resultado. Se, além do argumento "palavra-chave" for possível definir outros argumentos de busca, o resultado já virá filtrado. Por exemplo, se vou ao Google e digito a palavra "laranja", obtenho 36.600.000 resultados. Se digito duas palavras-chave, "laranja" + "cor" os resultados caem para 10.500.000, mas nem todas as ocorrências de "laranja" como cor podem ser filtradas dessa maneira. A precisão aumentará muito quando cada ocorrência da palavra "laranja" já estiver automaticamente rotulada como "fruta" ou "cor" (essa é a promessa da web semântica). Isso significa que cada atributo que "ensinarmos" a máquina a identificar poderá ser usado no futuro como argumento de busca em *corpus*.

A construção de recursos lexicais para o PLN

Há pouco mais de uma década foram iniciados grandes projetos de construção de léxicos semânticos para o PLN do inglês: a WordNet (FELLBAUM, 1998), a Framenet (BAKER et al., 1998), a VerbNet (KIPPER et al., 2006) e o Propbank (PALMER et al., 2005). Cada um desses léxicos responde diferentes perguntas. A Wordnet mostra quais são as relações semânticas (sinonímia, antonímia, hiperonímia, etc.) entre os nomes (substantivos, adjetivos e advérbios) e entre os verbos. A Framenet, baseada na semântica de *frames* de Fillmore (1968) mostra como as unidades lexicais podem ser agrupadas em cenários comuns (*frames*) e descreve os papéis semânticos previstos em cada um desses

3 http://www.linguateca.pt/ACDC/

cenários. A Verbnet, baseada nas classes verbais de Levin (1993) agrupa os verbos em classes, de acordo com seu comportamento sintático e semântico. O Propbank, por sua vez, se diz livre de teoria e descreve os papéis semânticos previstos para cada sentido dos verbos. Por meio da combinação desses léxicos, é possível fazer inferências automáticas, ampliando a cobertura individual de cada um deles. Foi esse objetivo que motivou o projeto SemLink,[4] que fez o mapeamento entre eles.

O inglês é a língua que iniciou a grande corrida pelo processamento automático. Por isso, sua experiência com erros e acertos é aproveitada por toda a comunidade científica dedicada a processar automaticamente outras línguas. Existem duas vantagens e uma desvantagem em tomar o PLN do inglês como modelo a ser imitado. A primeira vantagem é contar com uma abordagem já testada em uma língua natural, o que poupa retrabalho e aproveita conhecimento. A outra vantagem é que um recurso projetado da mesma forma em duas línguas possibilita mapeamentos e explorações multilingues. A desvantagem é que a língua que segue o inglês acaba sendo tratada sob a ótica do inglês, inibindo abordagens originais que possam ressaltar o que há de mais genuíno em sua natureza. Acredito, contudo, que à medida que o PLN de uma língua se desenvolva, surjam novas ideias de como tratar questões típicas dessa língua.

No português do Brasil ainda são tímidos os investimentos na construção desses grandes repositórios léxico-semânticos. O projeto mais antigo é a WordNet-Br (DIAS-DA-SILVA et al., 2007). A Framenet, por sua vez, inspira os projetos Framecorp (CHISHMAN et al., 2008) e Framenet Brasil (SALOMÃO, 2009). O Propbank inspirou a construção do *corpus* Propbank-Br (DURAN; ALUÍSIO, 2012) e do Cintil-Propbank (BRANCO et al., 2012) e a VerbNet inspirou a Verbnet-Br (SCARTON, 2011).

Além dos léxicos "espelhos", ou seja, repositórios construídos nos moldes dos repositórios desenvolvidos para a língua inglesa e replicados para outras línguas, há muitas outras lacunas no que diz respeito ao léxico para processamento automático do português. A seguir relaciono algumas das perguntas para as quais não há respostas prontas e completas a fim de serem facilmente empregáveis ao PLN do português:

I. Quais são os nomes eventivos do português, quantos argumentos cada um deles prevê e quais preposições podem introduzir cada um desses argumentos? (Ex: *requisição de* alguma coisa *a* alguém)

II. Quais marcas lexicais estão associadas à expressão da modalidade e do aspecto? (Ex: uso de tempos e modos verbais, uso de auxiliares de modo e aspecto).

III. Qual o léxico utilizado para expressar quantidades? (Ex: *vários*, *diversos*, *muitos*, *inúmeros* etc., além do léxico dos números inteiros e frações).

IV. Quais são os verbos obrigatoriamente reflexivos (*aventurar-se*), opcionalmente reflexivos (*esquecer-se*) e reflexivos recíprocos (*encontrar-se*)?

V. Quais são os predicados complexos idiomáticos e seus respectivos sinônimos? (Ex: *ir embora=partir*).

VI. Quais são os predicados complexos formados por verbo suporte e seus respectivos sinônimos? (Ex: *dar palestra=palestrar*).

VII. Quais substantivos são formados pelo particípio passado? (Ex: *o passado*, *o aposentado*).

VIII. Qual o léxico da expressão do tempo?

4 http://verbs.colorado.edu/semlink/

À guisa de ilustração, mostro a dificuldade de responder de forma exaustiva essa última pergunta. Há trabalhos sobre a expressão do tempo (como o de HAGÈGE et al., 2008), mas muitas vezes a pesquisa é financiada por empresas privadas e, por isso, nem sempre o léxico levantado fica disponível para a comunidade.

Primeiramente é preciso modelar o conhecimento do tempo. Poderíamos, por exemplo, dizer que a expressão do tempo pode ser subdivida semanticamente em:

Tempo preciso	=>	em 1922
Período de tempo	=>	de dezembro a janeiro
Origem no tempo	=>	desde o último inverno
Fim no tempo	=>	até o final das aulas
Duração	=>	durante o carnaval
Frequência	=>	muitas vezes

A partir daí levantaríamos o léxico utilizado para cada uma dessas classes. Tomando apenas a classe da frequência de tempo, teríamos:

Palavras simples:

sempre, diariamente, semanalmente, quinzenalmente, mensalmente, anualmente, eventualmente frequentemente, raramente, esporadicamente etc.

Expressões fixas:

às vezes, de vez em quando, vez por outra, vezes seguidas etc.

Expressões variáveis:

N vezes, N vezes por X

onde:

N = léxico de quantidade (poucas; muitas; várias, inúmeras, (léxico dos números inteiros));

X = unidades de medida de tempo (segundo, minuto, hora, dia, semana, mês, ano, século etc.).

Orações adverbiais temporais introduzidas por:
toda vez que, sempre que.

É esse tipo de modelagem do conhecimento léxico que facilita sua formalização e seu subsequente aproveitamento no processamento da língua.

Considerações finais

No que diz respeito à qualidade, os resultados do processamento automático ainda não se comparam aos resultados do trabalho de especialistas em língua. Porém a capacidade de processamento da máquina é infinitamente superior à humana nos quesitos tempo e quantidade, o que, por si só, justifica sua aplicação. Muitos artefatos modernos dependem desse processamento e, nesse sentido, a máquina não substitui o homem, mas, sim, opera onde seria humanamente impossível operar.

A descrição do léxico para fins computacionais é essencial para melhorar o processamento automático do português, mas ainda são poucos os pesquisadores dedicados a essa tarefa, principalmente no Brasil. Acredito, contudo, que a causa disso seja a falta de consciência sobre a demanda, pois embora tenhamos muitos programas que abarcam os estudos do léxico, poucos deles estudam o léxico sob a perspectiva da linguística computacional.

É possível, inclusive, que haja resultados de pesquisa passíveis de serem aproveitados no PLN do português, mas que não tenham sido divulgados ainda. A fim de reunir em um local de fácil acesso recursos léxicos para processamento automático do português, foi construído um portal chamado PortLex[5] aberto a contribuições de diversas origens.

REFERÊNCIAS

ABREU, D. T. B. *A semântica de construções com verbos-suporte e o paradigma Framenet*. 2011. Dissertação (Mestrado em Linguística) – Universidade do Vale dos Sinos, São Leopoldo, RS, Brasil, 2011.

ALSINA, A.; BRESNAN, J.; SELLS, P. *Complex Predicates*. Stanford, CA, EUA: CSLI Publications, 1997.

ATHAYDE, M. F. Construções com verbo-suporte (funktionsverbgef¨uge) do português e do alemão. *Cadernos do CIEG Centro Interuniversitário de Estudos Germanísticos*, Universidade de Coimbra, Coimbra, Portugal, n. 1, 2001.

BAKER, C.F.; FILLMORE, C. J.; LOWE. J. B. The Berkeley FrameNet Project. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS AND 17TH INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 36, Montreal, Canada, 1998. *Proceedings...* v. 1. Stroudsburg, PA, EUA: Association for Computational Linguistics, 1998. p. 86-90.

BRANCO, A.; CARVALHEIRO, C.; PEREIRA, S.; SILVEIRA, S. SILVA, J.; CASTRO, S.; GRAÇA, J. A PropBank for Portuguese: the CINTIL-PropBank. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC'12), 8, Istambul, Turquia, 23-25 de maio de 2012. *Proceedings...* Paris: European Language Resources Association (ELRA). p. 1526-1521. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2012/pdf/373_Paper.pdf>. Acesso em: 30 jun. 2012.

BUTT, M. The Light Verb Jungle. In: Aygen, G.; Bowern, G.; Quinn, C.; (ed.). *Harvard Working Papers in Linguistics*. v. 9, Papers from the GSAS/Dudley House Workshop on Light Verbs. Cambridge, MA, EUA: Harvard University, 2003. p. 1-49.

5 <http://www2.nilc.icmc.usp.br/portlex/index.php/pt/>

CHISHMAN, R. L. O.; BERTOLDI, A.; LERMEN, L.; PADILHA, J. G. Corpus e Anotação Semântica: um Experimento para a Língua Portuguesa a partir da Semântica de Frames. In: SIMPÓSIO BRASILEIRO DE SISTEMAS MULTIMÍDIA E WEB (WebMedia 2008), XIV, WORKSHOP DE TECNOLOGIA E INFORMAÇÃO, V, Vila Velha, 2008, *Anais*... Porto Alegre: Sociedade Brasileira de Computação (SBC), 2008. v. II. p. 321-325.

CONEJO, C. R. *O verbo-suporte "fazer" na língua portuguesa*: um exercício de análise de base funcionalista. 2008. Dissertação (Mestrado em Letras) – Universidade Estadual de Maringá, PR, Brasil.

DIAS-DA-SILVA, B. C. The WordNet.Br: an Exercise of human language technology research. In: South Jeju Island. In: INTERNATIONAL WORDNET CONFERENCE, 2007. *Proceedings*... Brno: Masaryk University Press, 2007. v. 3. p. 301-303.

DUARTE, I.; GONÇALVES, A.; MIGUEL, M.; MENDES, A.; HENDRICKX, I.; OLIVEIRA, F.; CUNHA, L. F.; SILVA, F.; SILVANO, P. Light verbs features in European Portuguese. In: INTERDISCIPLINARY WORKSHOP ON VERBS: The Identification and Representation of Verb Features (Verb 2010), Pisa, Italy, Nov. 2010.

DURAN, M. S.; ALUÍSIO, S. M. *Proceedings*... Pisa, Itália: Universidade de Pisa, 2010. Disponível em: <http://linguistica.sns.it/Workshop_verb/papers/Duarte_verb2010_submission_66.pdf>. Acesso em: 30 jan. 2011.

_____. Propbank-Br: a Brazilian treebank annotated with semantic role labels. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC'12), 8, Istambul, Turquia, 2012. *Proceedings*... Paris: European Language Resources Association (ELRA), Disponível em: <http://www.lrec-conf.org/proceedings/lrec2012/pdf/272_Paper.pdf>. Acesso em: 13 jul. 2012.

DURAN, M. S.; AMANCIO, M. A.; ALUÍSIO, S. M. Assigning Wh-Questions to Verbal Arguments: Annotation Tools Evaluation and Corpus. In: CONFERENCE ON INTERNATIONAL LANGUAGE RESOURCES AND EVALUATION, 7, Valletta, Malta, 2010. *Proceedings*… Paris: European Language Resources Association (ELRA), 2010, p. 1445-1451. Dsiponível em: <http://www.lrec-conf.org/proceedings/lrec2010/summaries/564.html>. Acesso em: 07 jul. 2010.

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, 7, Portland-OR, USA, 2011, *Proceedings*... Portland: Association for Computational Linguistics, 2011. p. 74-82. Disponível em; <http://www.aclweb.org/anthology-new/W/ W11/W11-0812.pdf>. Acesso em: 20 fev. 2011.

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

FILLMORE, C. The Case for Case. In: BACH, E.; HARMS, R. T. (Ed.) *Universals in Linguistic Theory*. New York, NY, EUA: Holt, Rinehart, and Winston, 1968. p. 1-88.

GARRÃO, M.; QUENTAL, V.; CAMINADA, N.; BICK, E. The Identification and Description of Frozen Prepositional Phrases through a Corpus-Oriented Study. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE: PROPOR, 8, Aveiro, Portugal, 2008. *Proceedings*... Berlin Heidelberg: Springer-Verlag, p. 220–223.

HAGÈGE, C.; BAPTISTA, J.; MAMEDE, N. Proposta de Anotação e Normalização de Expressões Temporais da Categoria TEMPO para o HAREM II. In: MOTA, C.; SANTOS, D. (Ed.). *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas*: O Segundo HAREM. Lisboa, Portugal: Linguateca, 2008. p. 261-274. Disponível em: <http://www.linguateca.pt/HAREM/actas/Capitulo_15-MotaSantos2008.pdf>. Acesso em: 15 abr. 2013.

HENDRICKX, I.; MENDES, A.; GONÇALVES, A.; DUARTE, I. Complex predicates annotation in a corpus of Portuguese. In: ACL LINGUISTIC ANNOTATION WORKSHOP, 4, Uppsala, Sweden, 2010. *Proceedings*... Taberg: Taberg Media Group, 2010. p. 100-108.

KIPPER, K.; KORHONEN A., RYAN N.; PALMER, M. Extending VerbNet with Novel Verb Classes. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2006), 5, Genova, Itália, 2006. *Proceedings*...Paris: European Language Resources Association (ELRA), 2006, p. 1027-1032. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2006/>. Acesso em: 23 abr. 2009.

LEVIN, B. *English Verb Classes and Alternation*, a Preliminary Investigation. Chicago: The University of Chicago Press, 1993.

NEVES, M. H. M. Estudo das construções com verbos-suporte em português. In: KOCK, I. G. V. (Org.). *Gramática do português falado VI*: Desenvolvimentos. Campinas: EdUnicamp; São Paulo: Fapesp, 1996. p. 201-231.

PALMA, C. *Expressões Fixas Adverbiais*: descrição léxico-sintática e subsídios para um estudo contrastivo Português-Espanhol. 2009. Dissertação (Mestrado em Linguística) – Universidade do Algarve, Portugal.

PALMER, M.; GILDEA, D.; KINGSBURY, P. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, Cambridge, MA USA, v. 31, n. 1, p. 71-105, March 2005.

RAMISCH, C.; VILLAVICENCIO, A.; BOITET, C. Multiword expressions in the wild? The mwetoolkit comes in handy. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS (COLING 2010) – Demonstrations, 23, Beijing, China, 2010. *Proceedings*… Association for Computational Linguistics, p. 1041-1049. Disponível em: <http://www.aclweb.org/anthology/C/C10/C10-2120.pdf>. Acesso em: 23 jan. 2011.

SAG, I. A.; BALDWIN, T., BOND, F.; COPESTAKE, A.; FLICKINGER, D. Multiword Expressions: A Pain in the Neck for NLP. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS AND INTELLIGENT TEXT PROCESSING (CICLing'02), Cidade do México, 17-23 de fevereiro de 2002, *Proceedings*... London, UK: Springer-Verlag, 2002. p. 1-15.

SALOMÃO, M. M. FrameNet Brasil: Um trabalho em progresso. *Calidoscópio*, São Leopoldo, v. 7, n. 3, p. 171-182, set/dez 2009.

SANTOS, D.; SARMENTO, L. O projecto AC/DC: acesso a corpora/disponibilização de corpora. In: MENDES, A.; FREITAS, T. (Ed.). *Actas do XVIII Encontro Nacional da Associação Portuguesa de Linguística.* Lisboa: APL, 2002. p. 705-717.

SCARTON, C. E. VerbNet.Br: construção semiautomática de um léxico computacional de verbos para o português do Brasil. In: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL 2011), 8, Cuiabá-MT, 24-26 de outubro de 2011. *Proceedings*… São Paulo: Sociedade Brasileira de Computação. Disponível em: <http://www.nilc.icmc.usp.br/til/stil2011_English/stil/artigos/Long/STIL2011_P3.pdf>. Acesso em: 30 jun. 2012.

SILVA, H. M. F. Verbos-suporte ou expressões cristalizadas? *Soletras*, Rio de Janeiro, v. 9, n. 17, p.175-182, 2009.

VILLAVICENCIO, A. Learning to distinguish PP arguments from adjuncts. In: CONFERENCE ON NATURAL LANGUAGE LEARNING (CoNLL-2002), 6, Taipei, Taiwan, 2002. *Proceedings*… New Brunswick, NJ, EUA: Association for Computational Linguistics, 2002, p. 84-90.