Anotação de sentidos de verbos em textos jornalísticos do corpus CSTNews

Revista de Estudos da Linguagem - RELIN,Belo Horizonte : UFMG,v. 23, n. 3, p. 797-832, 2015
http://www.producao.usp.br/handle/BDPI/51102

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".

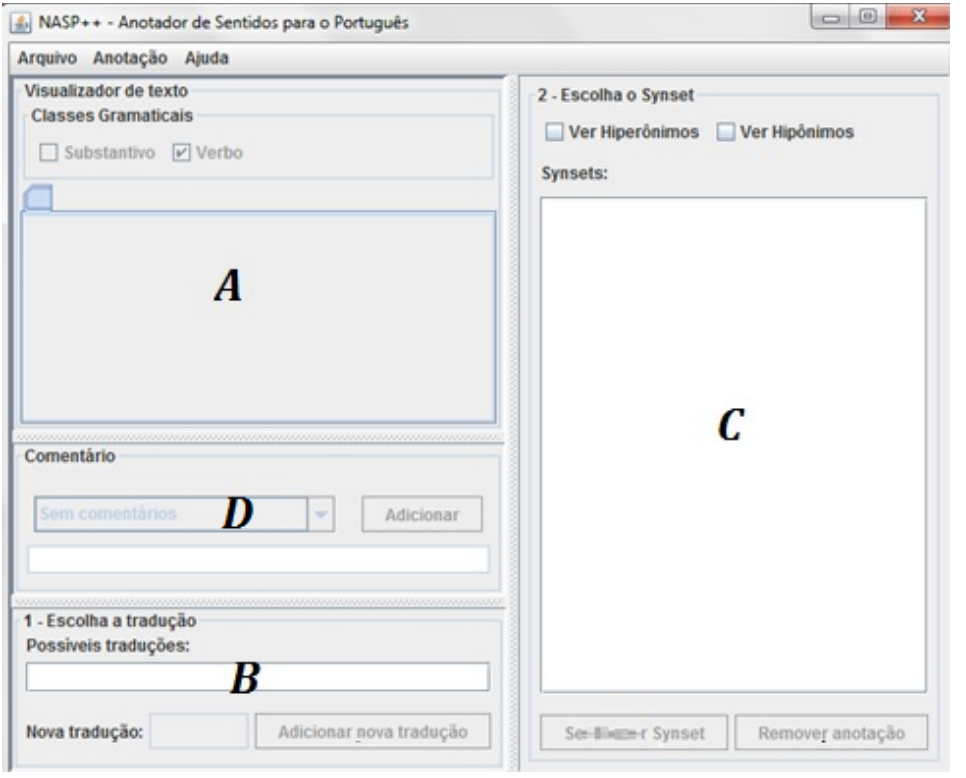
Como parte da metodologia de anotação, em caso de dúvidas, os anotadores puderam usar os recursos mencionados em cada etapa. Caso os anotadores não conseguissem resolver as dúvidas, poder-se-ia fazer uma consulta a toda a equipe de anotação.

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.