Núcleo Interinstitucional de Linguística Computacional - NILC

Anotação de Sentidos de Verbos no Córpus CSTNews

Relatório Técnico do NILC
NILC - TR - 14 - 05

Marco A. S. Cabezudo, Erick G. Maziero, Jackson W. C. Souza, Márcio S. Dias, Paula C. F. Cardoso, Pedro P. Balage Filho, Verônica Agostini, Fernando A. A. Nóbrega, Cláudia Dias de Barros, Ariani Di-Felippo, Thiago A. S. Pardo

Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

Série de Relatórios do
Núcleo Interinstitucional de Linguística Computacional

São Carlos, SP, Brasil

RESUMO

Um dos desafios do Processamento de Linguagem Natural (PLN) em nível semântico é a ambiguidade lexical, já que as palavras podem expressar significados distintos em função do contexto em que ocorrem. No PLN, a tarefa responsável por determinar o significado adequado de uma palavra em contexto é a Desambiguação Lexical de Sentido (DLS). Para tanto, a DLS parte de um repositório de sentidos pré-estabelecidos. Nessa tarefa, o uso de córpus anotado é muito útil, pois esse recurso linguístico-computacional permite o estudo mais aprofundado do fenômeno da ambiguidade e o desenvolvimento e a avaliação de métodos de DLS. Neste relatório, relatam-se o processo e os resultados da anotação de sentidos dos verbos presentes no córpus CSTNews, que é um córpus multidocumento de notícias jornalísticas escritas em português brasileiro.

ÍNDICE

1. Introdução

Com a quantidade crescente de informação, em grande parte disponível na Web, e a necessidade de formas mais inteligentes de se aprender e processar tanta informação, o processamento do significado das línguas naturais em seus vários níveis tem sido um dos focos de interesse de pesquisa da comunidade de Processamento de Linguagem Natural (PLN). 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.

Dentre os problemas relativos ao tratamento computacional da semântica das língua naturais, destaca-se a ambiguidade lexical, que resulta da impossibilidade de se identificar o sentido expresso por uma palavra em um contexto *x* dentre os vários que pode expressar, já que se trata, nesse caso, de uma palavra polissêmica. Vale ressaltar 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 seu conhecimento linguístico, de mundo e situacional.

Nas seguintes sentenças, apresentam-se exemplos de ambiguidade lexical com diferentes níveis de complexidade computacional. Nesses exemplos, as palavras em destaque são polissêmicas, ou seja, são palavras que expressam mais de um significado.

1. O professor *contou* a quantidade de alunos.
2. O atacante chutou e o goleiro *tomou um* frango.
3. O banco *quebrou* na semana passada.

Comumente, a desambiguação lexical é feita com base nas demais palavras de conteúdo que coocorrem com aquela cujo sentido precisa ser identificado. Assim, do ponto de vista computacional, a ambiguidade em (1) e (2) é mais 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 das palavras "atacante", "chutar", "goleiro" e "frango" no eixo sintagmático de “tomou” em (2) funciona como pista para a identificação de que o sentido expresso por “tomar” é "sofrer/levar” (um gol). O mesmo não ocorre com "quebrar" em (3). No caso, o contexto sentencial não fornece as pistas linguísticas necessárias para que a máquina identifique o sentido adequado da palavra em questão. Para determinar o sentido adequado, é preciso procurar pistas em outras sentenças e não só naquela em que a palavra ocorre.

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 do 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 e Martin, 2009).

A DLS é comumente realizada por um módulo específico incorporado à análise sintática ou semântica dos processos de interpretação e/ou geração da língua. Tal módulo de DLS é relevante a inúmeras aplicações de PLN. A análise de sentimentos é uma dessas aplicações. Nela, a identificação do conceito 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 (TA) e outras aplicações multilíngues talvez sejam as aplicações de PLN em que a necessidade de um módulo de DLS se faz mais evidente, pois a identificação do sentido de uma palavra vai determinar a escolha de seu equivalente de tradução. Por exemplo, se o verbo "conhecer" expressar o sentido de "encontrar-se (com)", como em *"Eu o conheci na festa"* este deve ser traduzido para "met" em inglês; caso expresse o sentido de "ter conhecimento sobre", como em *"Eu conheço essa teoria"*, a tradução correta é "*know*".

Para o desenvolvimento de métodos de DLS, um córpus em que o significado adequado de cada uma de suas palavras de conteúdo tenha sido explicitado (ou anotado) é um recurso muito importante, pois permite que estratégias de desambiguação automática sejam aprendidas, funcionando como um *benchmark* para a área de DLS.

Para o português, há o córpus multidocumento CSTNews[1] (Aleixo e Pardo, 2008; Cardoso et al., 2011), composto por textos jornalísticos coletados de agências de notícias on-line. Os substantivos comuns desse córpus foram manualmente anotados com os sentidos da WordNet de Princeton (versão 3.0) (WordNet.Pr) (Fellbaum, 1998), o que propiciou o desenvolvimento de métodos gerais de DLS para esse tipo de substantivo do português (Nóbrega, 2013).

Baseando-se em Nóbrega (2013), descreve-se, neste relatório, o processo de anotação semântica dos verbos do CSTNews, a qual será utilizada para a investigação de métodos de DLS. O relatório está estruturado em 5 seções. Na Seção 2, apresentam-se alguns trabalhos relacionados à anotação de sentidos em córpus em português. Na Seção 3, apresenta-se o processo de anotação do córpus CSTNews. Na Seção 4, apresentam-se a avaliação e os resultados da anotação do córpus. E, finalmente, na Seção 5, são feitas algumas considerações finais sobre o trabalho.

2. Trabalhos Relacionados para o Português

Specia (2007) propôs um método de DLS baseado em Programação Lógica Indutiva, caracterizado por utilizar aprendizado de máquina e regras especificadas da lógica proposicional. Focado na TA 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*.

[1] http://nilc.icmc.usp.br/CSTNews/login/?next=/CSTNews/

Para o desenvolvimento do método, construiu-se um córpus paralelo composto por textos em inglês e suas respectivas traduções para o português. Nesse córpus, cada texto original em inglês foi alinhado em nível lexical à sua tradução em português. As possíveis traduções em português usadas para cada verbo do inglês foram extraídas dos dicionários bilíngues DIC Prático Michaelis® (versão 5.1), Houaiss® e Collins Gem® (4a edição). No total, os textos em inglês somam 7 606 150 palavras e as traduções em português somam 7 642 048 palavras. Tais textos foram compilados de nove fontes de diversos gêneros e domínios.

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 córpus formado por 160 notícias jornalísticas extraídas da internet. Cada notícia jornalística passou por um pré-processamento, que consistiu na indexação das palavras de conteúdo aos conceitos da ontologia. A partir do córpus indexado à ontologia, um conjunto de heurísticas identifica o conceito subjacente a cada uma das palavras do córpus. A avaliação desse método de DLS foi feita de forma manual.

No trabalho de Nóbrega (2013), apresenta-se um método de desambiguação de substantivos comuns usando grafos de coocorrência e o algoritmo de Lesk (1986). Para tanto, o CSTNews (Aleixo e Pardo, 2008; Cardoso et al., 2011), córpus multidocumento composto por 50 coleções de notícias jornalísticas em português, foi anotado manualmente. Em especial, essa anotação consistiu na explicitação dos conceitos subjacentes aos substantivos comuns mais frequentes do córpus. A anotação dessa classe gramatical foi motivada pelos estudos sobre o impacto positivo que tem a desambiguação de substantivos comuns em aplicações de PLN (veja, por exemplo, o trabalho de Plaza e Diaz, 2011). Isso ocorre porque, ao serem bastante frequentes nos textos e carregarem boa parte do conteúdo expresso nos mesmos, a desambiguação dos substantivos se mostra relevante para a interpretação textual.

Inicialmente, o objetivo era anotar todos os substantivos comuns das coleções do córpus. No entanto, após a etapa de treinamento dos anotadores, optou-se por diminuir o escopo da tarefa devido à sua complexidade. Assim, a anotação limitou-se aos substantivos comuns mais frequentes, especificamente, aos 10% mais frequentes (totalizando 4366 substantivos). Esse valor foi escolhido ao se observar que palavras do córpus cuja frequência estava abaixo desse limiar não eram representativas para a tarefa de anotação de DLS. Para anotar os substantivos, ou seja, para explicitar os conceitos a eles subjacentes, utilizou-se o repositório de sentidos da WordNet.Pr[2] (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, a indexação dos substantivos em português aos conceitos foi feita com base em um dicionário bilíngue português/inglês. No caso, utilizou-se o WordReference®. Mais destalhes sobre o CSTNews são apresentados na próxima seção.

[2] As *wordnets* são tradicionalmente compostas por *synsets* (do inglês, *synonymy sets*), conjuntos de unidades lexicais sinônimas que expressam um único conceito. Os *synsets* são relacionados por relações semântico-conceituais, como hiperonínia/hiponímia, meronímia, etc.

Outro trabalho de anotação de córpus para o português é o de Travanca (2013). Travanca implementou métodos de DLS para verbos usando regras e aprendizado de máquina. Para tanto, anotou-se manualmente parte do PAROLE, córpus composto por livros, jornais, periódicos e outros textos (Ribeiro, 2003). O subcórpus anotado contém aproximadamente 250000 palavras, sendo 38827 verbos. Dentre eles, 21368 são verbos principais e os demais são verbos auxiliares. A quantidade de verbos ambíguos (anotados com mais de dois sentidos) foi de 12191, o que representa 57.05% do total de verbos principais. 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. O ViPer possui 5037 lemas e 6224 sentidos. Ressalta-se que os lemas do ViPer referem-se apenas aos verbos com frequência 10 ou superior no córpus CETEMPúblico (Rocha e Santos, 2000). O autor não apresenta os valores de concordância obtidos da anotação manual do córpus.

3. Anotação de Córpus

3.1 Considerações iniciais

A anotação teve por objetivo desambiguar as palavras da classe gramatical dos verbos. A escolha pela classe verbal pautou-se no fato de que os verbos, ao expressam um *estado de coisas,* são centrais à constituição dos enunciados (Fillmore, 1968).

Para a tarefa de anotação de sentidos, utilizou-se o CSTNews (Aleixo e Pardo, 2008; Cardoso et al., 2011), córpus multidocumento composto por 50 coleções ou grupos de textos, sendo que cada coleção versa sobre um mesmo tópico. A escolha do CSTNews pautou-se nos seguintes fatores: (i) utilização prévia desse córpus no desenvolvimento de métodos de DLS para os substantivos comuns (Nóbrega, 2013) e (ii) ampla abrangência de domínios ou categorias ("política", "esporte", "mundo", etc.), fornecendo uma gama variada de sentidos para o desenvolvimento de métodos de DLS robustos.

No total, o CSTNews contém 72148 palavras, distribuídas em 140 textos. Os textos são do gênero discursivo "notícias jornalísticas", pertencentes à ordem do relatar[3] (Dolz e Schnewly, 2004). As principais características desse gênero são: (i) documentar as experiências humanas vividas e (ii) representar pelo discurso as experiências vividas, situadas no tempo (capacidade da linguagem) (Lage, 2004). Especificamente, cada coleção do CSTNews contém: (i) 2 ou 3 textos sobre um mesmo assunto ou tema compilados de diferentes fontes jornalísticas; (ii) sumários humanos (*abstracts*) mono e multidocumento; (iii) sumários automáticos multidocumento; (iv) extratos humanos multidocumento; (v) anotações semântico-discursivas, entre outras. As fontes

[3] Dolz e Schnewly (2004) classificam os gêneros textuais em 5 categorias de acordo com algumas regularidades linguísticas, a saber: (i) textos da ordem do relatar, (ii) do narrar, (iii) do expor, (iv) do descrever ações e (i) do argumentar. Na categoria "ordem do relatar", estão agrupados os gêneros pertencentes ao domínio social da memorização e documentação das experiências humanas, como diários de viagem, notícias, reportagens, crônicas jornalísticas, relatos históricos, biografias, autobiografias, testemunhos, etc.

jornalísticas das quais os textos foram compilados correspondem aos principais jornais *online* do Brasil, a saber: *Folha de São Paulo, Estadão, Jornal do Brasil, O Globo* e *Gazeta do Povo*. A coleta manual foi feita durante aproximadamente 60 dias, de agosto a setembro de 2007. As coleções possuem em média 42 sentenças (de 10 a 89) e os sumários humanos multidocumento possuem em média 7 sentenças (de 3 a 14). Ademais, as coleções estão categorizadas pelos rótulos das "seções" dos jornais dos quais os textos foram compilados. Assim, o *corpus* é composto por coleções das seguintes categorias ou domínios: "esporte" (10 coleções), "mundo" (14 coleções), "dinheiro" (1 coleção), "política" (10 coleções), "ciência" (1 coleção) e "cotidiano" (14 coleções).

Como mencionado, os verbos ocupam lugar de centralidade nos enunciados. Isso pode ser constatado, aliás, pela frequência de ocorrência dos mesmos no CSTNews. Na Figura 3.1, apresenta-se a distribuição da frequência de ocorrência das classes de palavras de conteúdo no CSTNews. Para o cálculo dessa distribuição, os textos do CSTNews passaram por um processo de etiqueção morfossintática automática, realizada pelo etiquetador ou *tagger* MXPOST (Rapnaparkhi, 1986). Dessa etiquetação, constatou-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 córpus.

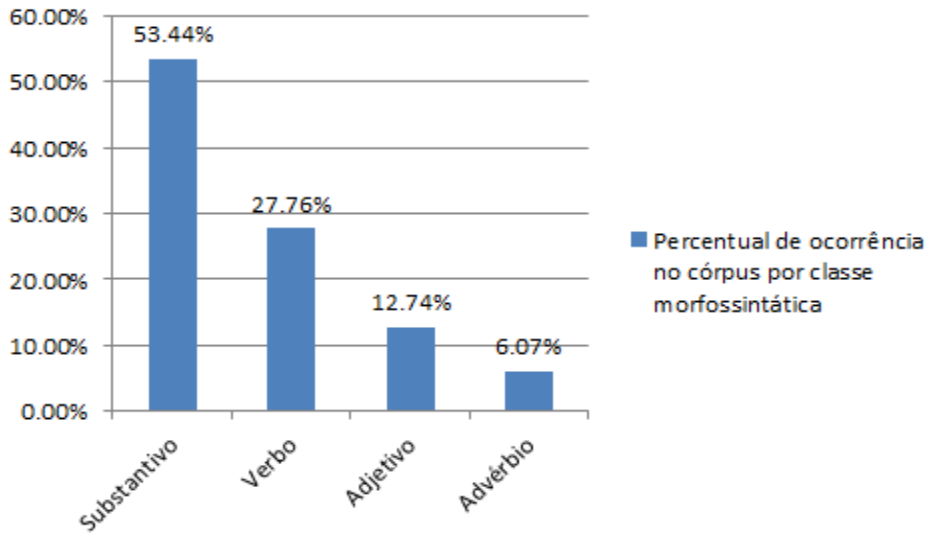

Figura 3.1. Percentual de ocorrência no córpus por classe morfossintática (dados obtidos de Nóbrega, 2013)

Para a tarefa de anotação, alguns recursos lexicais desenvolvidos para o português foram investigados, a saber: (i) TeP (2.0) (Maziero et al., 2008), (ii) onto.PT (Gonçalo Oliveira e Gomes, 2012) e WordNet.Br (Dias da Silva, 2005). Apesar da existência desses recursos, optou-se por utilizar a WordNet.Pr, desenvolvida para o inglês, como repositório de sentidos. Apesar de ter sido desenvolvida para o inglês norte-americano, a WordNet.Pr foi escolhida porque, além de ser o recurso lexical mais utilizado nas pesquisas do PLN, apresenta reconhecida (i) adequação linguística e tecnológica, já que foi construída segundo princípio da ciência cognitiva e em um formato computacionalmente tratável, e (ii) abrangência, já a versão 3.0 possui mais de 155.287

unidades lexicais do inglês e 117659 sentidos (*synsets*). Além disso, ressalta-se que a WordNet.Pr também foi o recurso utilizado por Nóbrega (2013) para o desenvolvimento de métodos de DLS para os substantivos do português.

3.2 Metodologia de Anotação

Para a anotação em questão, seguiu-se a mesma metodologia de Nóbrega (2013), que engloba etapas gerais e específicas. No caso, ambas foram realizadas com o auxílio da ferramenta NASP++, descrita na próxima subseção.

As etapas gerais fazem referência às etapas que todos os anotadores devem seguir para anotar uma coleção de textos. Um ponto importante nesta metodologia é o uso de palavras "pré-anotadas", isto é, se em um texto uma palavra "p" foi anotada com o *synset* "s", todas as outras ocorrências da mesma palavra serão pré-anotadas com o *synset "s".* Com a introdução de palavras "pré-anotadas", visou-se aproveitar o uso do cenário multidocumento, tendo como heurística que uma palavra tende a assumir o mesmo significado em um mesmo contexto (Mihalcea, 2006).

Para anotar uma coleção de textos pertencentes ao córpus CSTNews, seguiram-se os seguintes passos:

1. Escolher um texto de coleção para ser anotado;
2. Anotar todas as palavras indicadas como "verbo" nesse texto, podendo-se corrigir alguns erros do anotador morfossintático e/ou corrigir as palavras pré-anotadas, e, depois disso, anotar o texto seguinte da coleção;
3. Após anotar todos os textos, revisar e salvar os mesmos no formato e endereço especificados.

Para cada um dos textos anotados, foi seguida uma metodologia individual. Na Figura 3.2, apresenta-se um fluxograma com a metodologia de anotação para cada palavra do texto pertencente a uma coleção de textos.

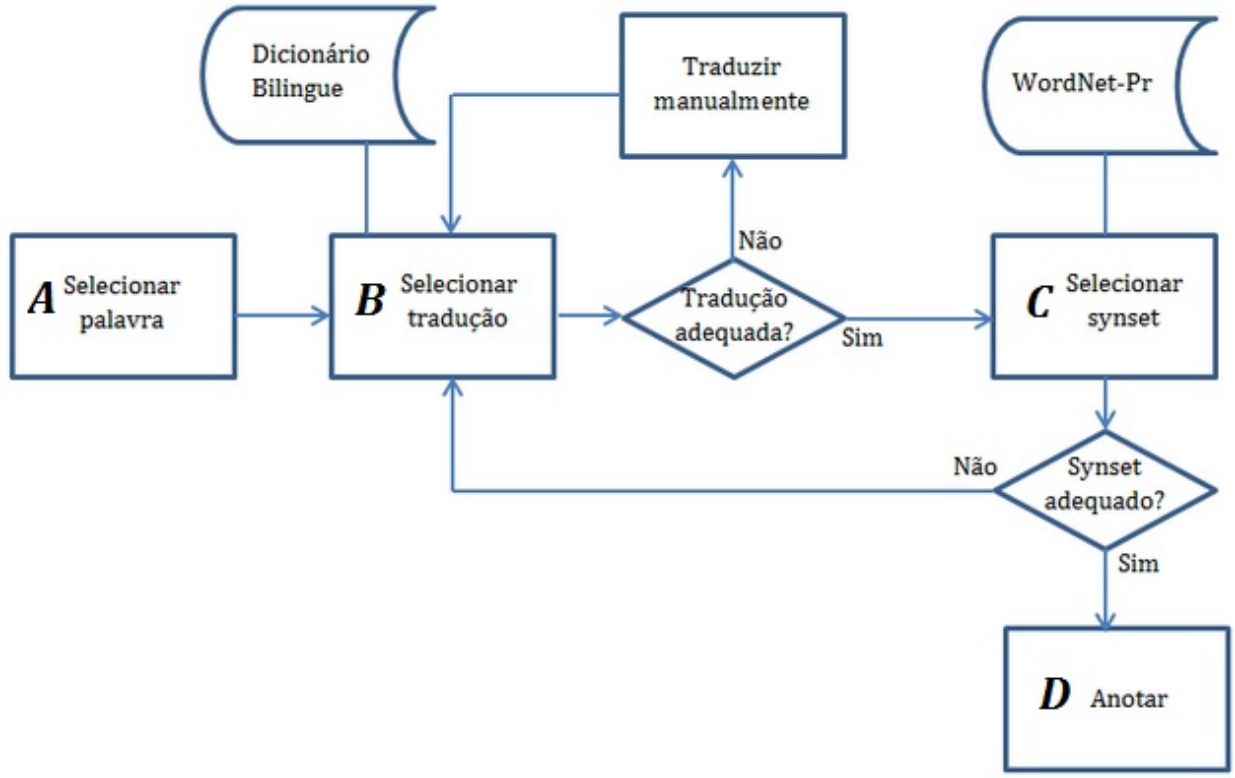

Figura 3.2. Metodologia de Anotação

De modo geral, primeiramente, (A) seleciona-se uma palavra a ser anotada. Com isto, é apresentada uma lista de traduções possíveis do inglês, advindas do dicionário bilíngue WordReference®, para essa palavra, e o usuário deve (B) escolher a tradução mais adequada. Se nenhuma das traduções for adequada para a palavra, então o usuário deve traduzir manualmente a palavra e escolhe-la. Após isto, a é apresentada uma lista de possíveis *synsets*, advindos da WordNet-Pr, para essa tradução, de maneira que o usuário (C) escolha o *synset* mais adequado. Assim, o usuário pode escolher algum dos *synsets* apresentados ou escolher novamente uma tradução melhor. Finalmente, (D) anota-se a palavra com o *synset* escolhido.

3.3 Ferramenta de Anotação: NASP++

A ferramenta NASP++ é uma ferramenta de auxílio à anotação de sentidos. Esta ferramenta usa o MXPOST (Ratnaparkhi, 1986) como *tagger* para o português, usando o modelo para o português proposto por Aires (2000); e a WordNet-Pr, na versão 3.0, como repositório de sentidos. Como a WordNet-Pr foi desenvolvida para o inglês, a NASP++ também faz uso do dicionário bilingüe WordReference®[4] para encontrar as possíveis traduções de uma palavra.

[4] www.wordreference.com

A NASP++ é uma versão atualizada da NASP, proposta por Nóbrega (2013), a qual estava focada na anotação de sentidos de substantivos. As funcionalidades adicionadas nessa versão são as seguintes:

- Anotação de sentidos para verbos: a NASP permitia a anotação somente de substantivos; nessa nova versão, é possível anotar substantivos e verbos.
- Adicionar comentários às anotações feitas: especificamente para anotação de verbos, mas também pode se estender aos substantivos. Atualmente, as opções de comentário são:

    - Sem comentários: quando não existem comentários a serem adicionados para uma palavra anotada (opção colocada por *default*);
    - Não é verbo, erro de anotação: quando a palavra foi erroneamente identificada como verbo;
    - É predicado complexo: quando o verbo pertence a um predicado complexo. Por exemplo, na sentença "A mulher bateu as botas", ao se selecionar "bateu", deve-se escolher o comentário "é predicado complexo".
    - É verbo auxiliar: quando o verbo identificado pelo *tagger* é um verbo auxiliar. Por exemplo, na sentença "Ele **estava** jogando bola", é adicionado o comentário de verbo auxiliar para o verbo "estava".
    - Outros (esta opção pode ser usada também na anotação de substantivos): para demais casos que eventualmente surjam, incluindo dificuldades de anotação.

- Colocar limite de palavras a serem anotadas: na NASP, a quantidade de substantivos que podiam ser anotados estava limitada a 10% do total. Nessa versão, deixou-se aberta a possibilidade de anotar uma porcentagem qualquer, segundo a frequência de ocorrência, tanto para substantivos, quanto para verbos.
- Geração de ontologia: a ferramenta cria uma ontologia que abrange os sentidos anotados e a hierarquia completa deles na WordNet-Pr (esta funcionalidade será detalhada na Seção 3.5).

Na Figura 3.3, apresenta-se a interface principal da ferramenta, composta das seguintes seções:

A. Visualizador de textos que serão anotados
B. Painel para exibição e seleção de traduções
C. Painel para exibição e seleção de *synsets*
D. Painel de comentários

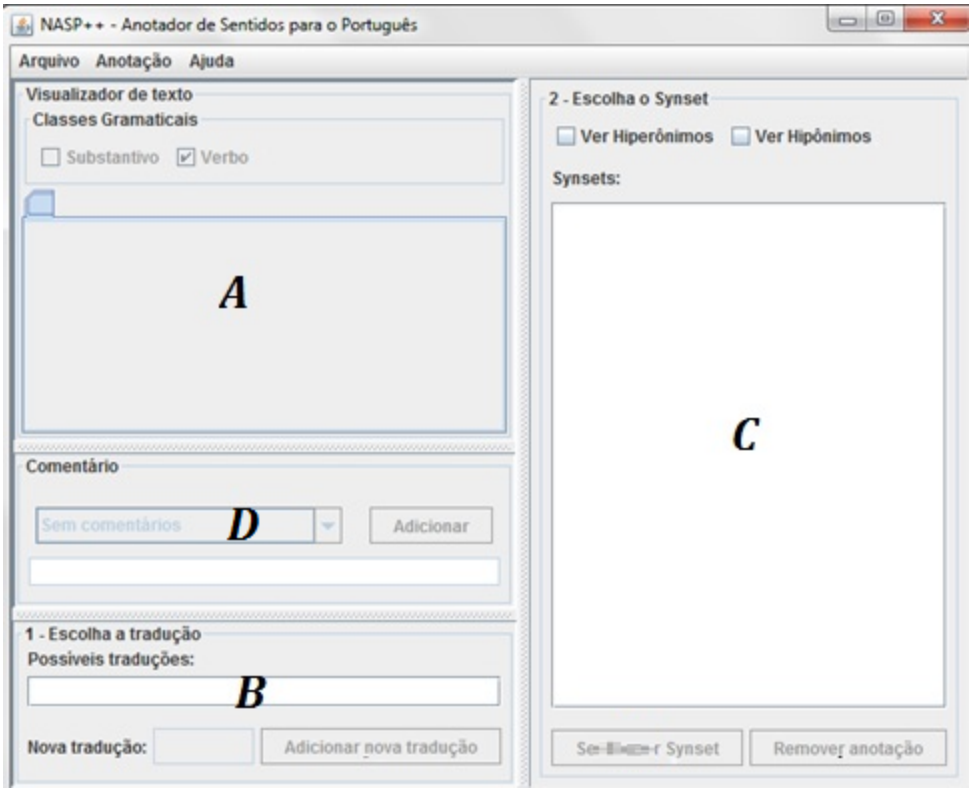

Figura 3.3. Tela principal da NASP++

No que se segue, apresenta-se a metodologia individual seguida usando a ferramenta adaptada/estendida, através de um exemplo de anotação com o verbo “morreram”.

Na Figura 3.4, apresenta-se o visualizador de textos com os os textos já carregados, no qual aparecem em quadro “vermelho” as palavras que devem ser anotadas (os verbos, neste caso). Clica-se, então, na palavra “morreram” para anotá-la, seguindo desta forma o passo “A” da metodologia proposta na Figura 3.2.

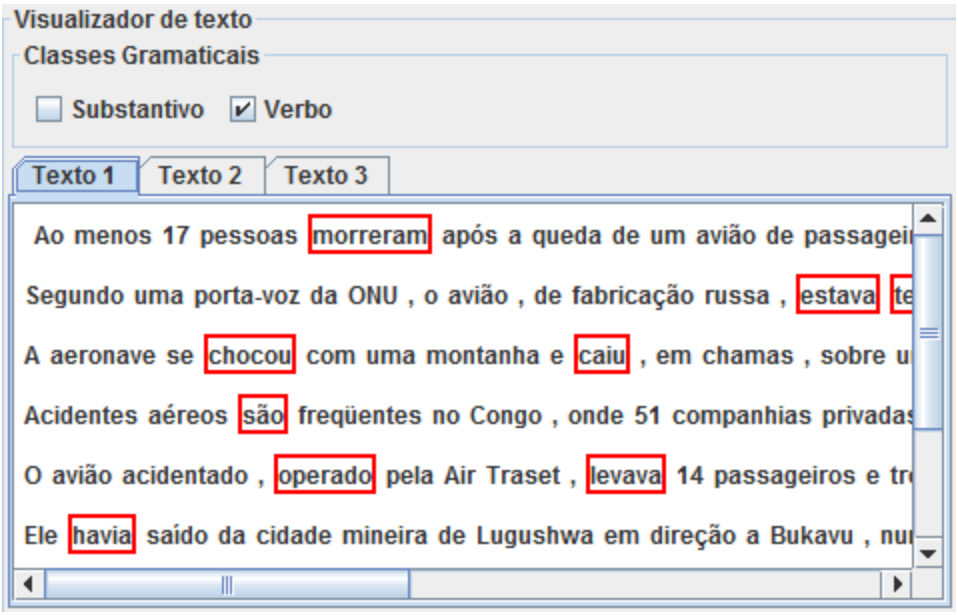

Figura 3.4. Visualizador de textos

Ao se clicar na palavra "morreram", o sistema obtém automaticamente o lema da palavra selecionada e ativam-se automaticamente o painel de comentários e da lista de traduções (como se apresenta na Figura 3.5) para o usuário selecionar a melhor tradução para a palavra (passo "B" da metodologia proposta na Figura 3.2). Para o exemplo, ao selecionar "morreram", o sistema obterá o lema "morrer" e aparecerá o tradução "*die*" (do lema "morrer") como única possível tradução oferecido pelo WordReference®.

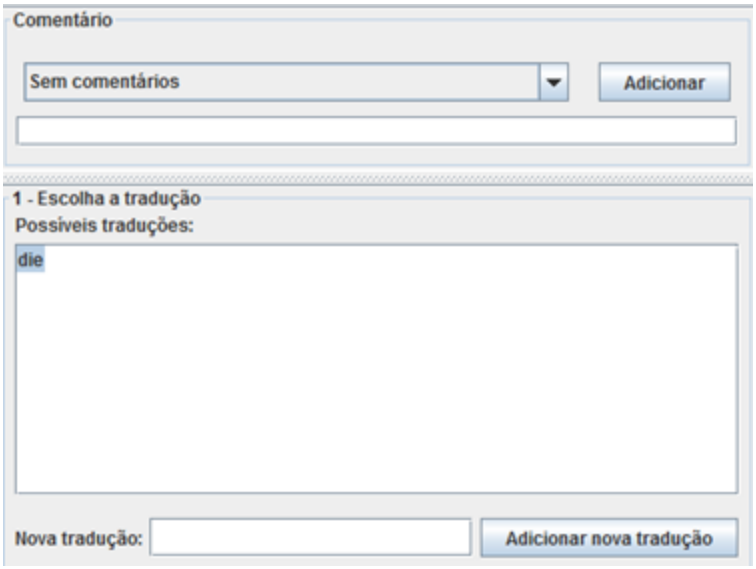

Figura 3.5. Tela de Comentários e Lista de Traduções

Ao escolher a tradução "*die*", apresentam-se na tela de *synsets* (na Figura 3.6) os *synsets* que podem ser escolhidos para indicação do sentido correto da palavra "morreram" em seu contexto.

Para cada um dos *synsets*, mostram-se o conjunto de sinônimos que formam esse *synset,* a glosa[5] e os exemplos (quando existirem).

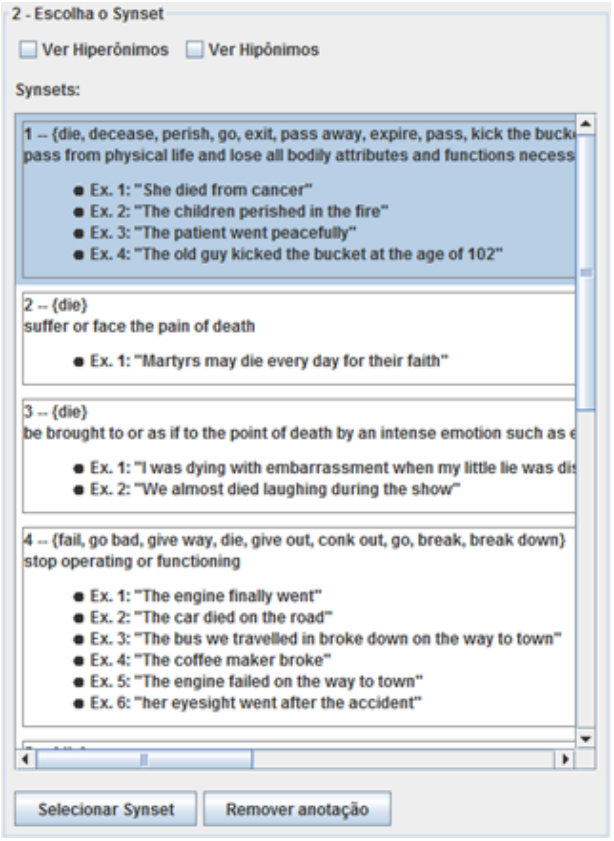

Figura 3.6. Tela de seleção do *synset*

Além disso, em caso de dúvidas de definição do *synset*, é possível ver os hiperônimos e hipônimos do mesmo, como se mostra na Figura 3.7.

[5] Uma glosa, neste caso, é uma definição informal do conceito representado no *synset*.

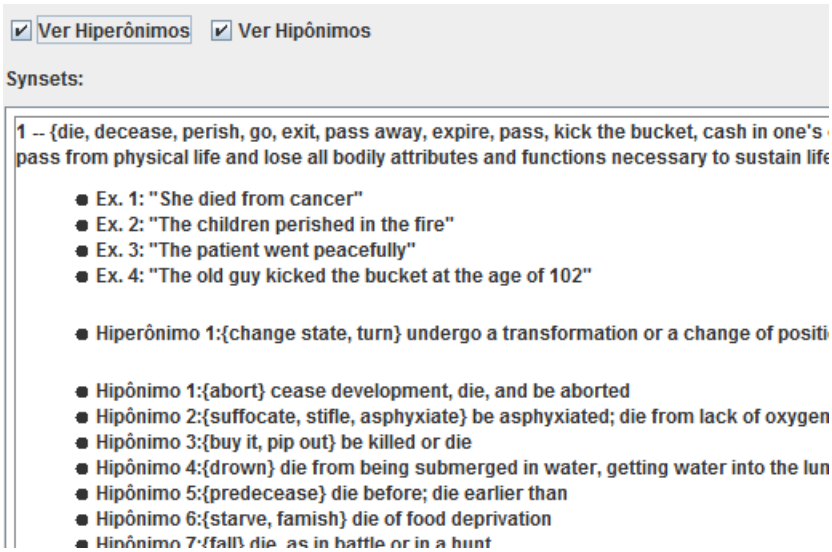

Figura 3.7. Tela de apresentação de hiperônimos e hipônimos

Dentre as opções, o usuário seleciona o *synset* (passo "C" da metodologia proposta na Figura 3.2) clicando no botão "Selecionar *Synset*". Ao se clicar no botão "Selecionar *Synset*" ou dar dois cliques no *synset* escolhido, aparecerá uma janela de confirmação, como a apresentada na Figura 3.8, na qual se tem que clicar em "Sim", no caso de se ter certeza do *synset* escolhido (passo "D" da metodologia proposta na Figura 3.2).

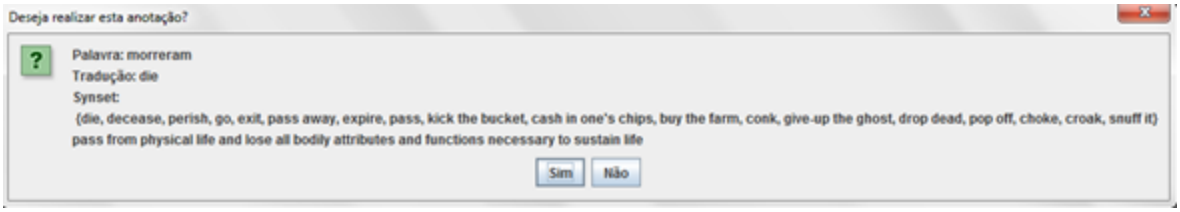

Figura 3.8. Janela de confirmação de escolha do *synset*

Finalmente, no visualizador de textos, aparecerá a palavra "morreram" marcada com quadro de cor "verde" (como apresentada na Figura 3.9), que significa que a palavra já foi anotada.

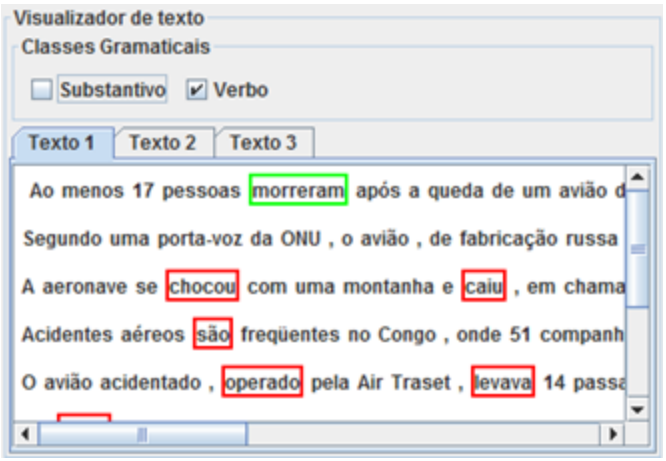

Figura 3.9. Visualizador de textos com o verbo “morrer” anotado

Também, salienta-se que outras ocorrências da mesma palavra (para o exemplo, o verbo “morrer”) são marcadas em cor “amarela”, representando que já foram previamente anotadas, como se apresenta na Figura 3.10.

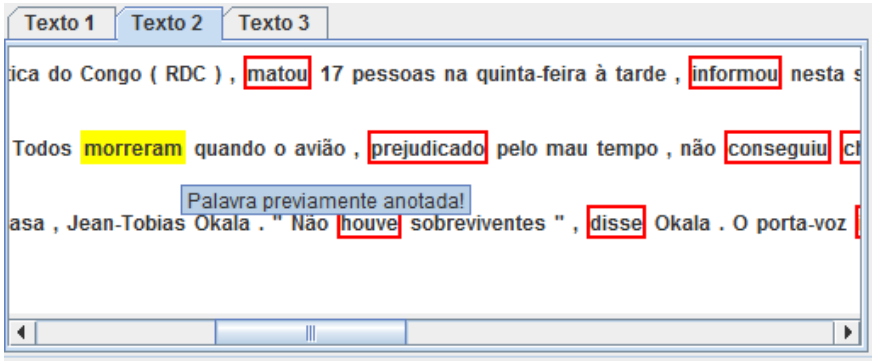

Figura 3.10. Visualizador de textos com o verbo “morrer” previamente anotado

Finalmente, a ferramenta permite salvar os arquivos de anotação em formato de linguagem de marcação XML. Nesse formato, o arquivo é organizado hierarquicamente (apresentado na Figura 3.11), contendo no primeiro nível, elementos gerais da anotação: os anotadores (representados pelo elemento “Annotators”), os limites de palavras (representados pelo elemento “LimitAnnotations”) que podem ser anotadas (para verbo e substantivo) e os arquivos carregados na anotação (representados pelo elemento “Files”).

No elemento marcado como “Files”, estão incluídos todos os textos carregados (elemento “Text”). Cada texto contém o nome do arquivo carregado e a linguagem na qual está escrito o texto. Além disso, cada texto contém parágrafos (representados pelo elemento “p”) e cada parágrafo inclui uma sentença dividida em uma lista de palavras (representadas por “Token”). Cada “token” contém uma lista de atributos, representados pelos elementos a seguir:

- *Word*: contém a palavra em si.

- *Tag*: anotação morfossintática advinda da ferramenta NASP++.
- *MorphoTagPOS*: anotação morfossintática feita pelo POS-*Tagger*.
- *MorphoTag*: por *default*, tem o mesmo valor que MorphoTagPOS, mas pode ser modificado; isto é feito pelo usuário quando existe um erro de anotação.
- *Lemma*: lema de palavra.
- *Comment*: comentários sobre a palavra anotada. Oferecem-se as alternativas de (1) "Sem comentários", (2) "Não é um verbo, erro de anotação", (3) "Verbo auxiliar", (4) "É predicado complexo" e (5) "Outros", sendo que as opções 2, 3 e 4 são especificamente para verbos. Além desse conteúdo, o usuário pode adicionar observações, para ser mais específico sobre o comentário.
- *Type*: o estado de anotação da palavra. Pode ser:

    - ANNOTATED: palavra anotada.
    - VERB_NO_ANNOTATED: verbo não anotado.
    - PREV_ANNOTATED: palavra previamente anotada.
    - NOUN_NO_ANNOTATED: substantivo não anotado.
    - NO_ANNOTATED: palavra não anotada (outras classes gramaticais).

- *Translations*: traduções oferecidas pela ferramenta e adicionadas pelo usuário. No caso de alguma tradução ser escolhida, receberá o valor "*true*" no atributo "*selected*".
- *Synsets*: *synsets* oferecidos pelas traduções propostas na NASP++. No caso de algum *synset* ser escolhido, receberá o valor "*true*" no atributo "*selected*".

```
<save>
  <Annotators>
    <Annotator id="1">Erick</Annotator>
    <Annotator id="2">Fernando</Annotator>
  </Annotators>
  <LimitAnnotations>
    <Pos id="NOUN">1.0</Pos>
    <Pos id="VERB">1.0</Pos>
  </LimitAnnotations>
  <Files>
    <Text name="D1_C1_Folha_04-08-2006_07h42.tagged" language="
      <p number="0">
        ...
        <Token>
          <Word cp="">morreram</Word>
          <Tag>VERB</Tag>
          <MorphoTag>Verbos</MorphoTag>
          <MorphoTagPOS>Verbos</MorphoTagPOS>
          <Lemma>morrer</Lemma>
          <Comment>
            <Content/>
            <Obs/>
          </Comment>
          <Type>ANNOTATED </Type>
          <Translations manual_translation="true">
            <Translate selected="true">die</Translate>
          </Translations>
          <Synsets>
            <Synset selected="true">358431</Synset>
            <Synset selected="false">2109818</Synset>
            <Synset selected="false">1784953</Synset>
            <Synset selected="false">434374</Synset>
            <Synset selected="false">1829475</Synset>
            <Synset selected="false">1785242</Synset>
            <Synset selected="false">1555034</Synset>
            <Synset selected="false">1074914</Synset>
            <Synset selected="false">538323</Synset>
            <Synset selected="false">354845</Synset>
            <Synset selected="false">224295</Synset>
          </Synsets>
        </Token>
        ...
      </p>
    </Text>
  </Files>
</save>
```

Figura 3.11. Arquivo de anotação em formato XML

3.4 Procedimentos usados na metodologia

Para a execução correta da metodologia mencionada na Seção 3.2, alguns procedimentos foram incluídos em cada etapa da metodologia. A seguir, detalham-se cada um deles:

3.4.1 Durante a etapa de seleção de palavra a ser anotada

- Procedimento 1: Erros no etiquetador morfossintático

Sabendo que o etiquetador morfossintático (MXPOST) pode errar, se deve considerar a possibilidade de verificar se as palavras etiquetadas são realmente verbos ou não, dado que, por exemplo, no caso da sentença "Ele havia **saído** da cidade mineira" o etiquetador poderia mostrar a palavra "saído" como se não fosse um verbo, sendo que é um verbo. Também pode ser o caso que uma palavra seja etiquetada como verbo, mas não seja de fato um. Contudo, a ferramenta tem a capacidade de deixar o usuário anotar essas palavras.

- Procedimento 2: Anotação de verbos auxiliares

Os verbos auxiliares têm pouco conteúdo semântico próprio. Por exemplo, na sentença "Ele **havia** saído de casa." o verbo "haver" não veicula a ideia principal da ação, portanto, é um verbo auxiliar. Para a melhor identificação de verbos auxiliares, forneceu-se uma lista de verbos auxiliares e como eles acontecem em uma sentença. Esta lista é fornecida pelo relatório técnico de verbos auxiliares no português do Brasil (Duran e Aluisio, 2010). No caso de uma palavra a ser anotada ser um verbo auxiliar, optou-se por adicionar um comentário no verbo, do tipo "Verbo auxiliar" (item "D" da Figura 3.3). Dessa forma, a palavra torna-se automaticamente anotada, não sendo necessário lhe atribuir um sentido.

- Procedimento 3: Anotação de verbos "independentes"

Os verbos "independentes" acontecem quando, dentro de uma mesma sentença, podem existir dois ou mais verbos consecutivos e eles apresentam conteúdo semântico próprio. Para este tipo de casos, os verbos "independentes" devem ser anotados normalmente. Por exemplo, na sentença "Ele havia **prometido retornar**", o verbo "havia" é um verbo auxiliar, mas os outros dois verbos ("prometer" e "retornar") têm significados claros (analisados individualmente) e, portanto, devem ser anotados com seus respectivos *synsets*.

- Procedimento 4: Anotação de Predicados Complexos

Um predicado complexo é aquele que é composto por 2 ou mais elementos para expressar um sentido, por exemplo, "dar abrigo a" tem o significado de "abrigar" e "tomar conta" de "cuidar". Para a anotação, a ferramenta NASP++ foi capaz de sugerir quais poderiam ser predicados complexos, seguindo os estudos feitos por Duran et al. (2011). Contudo, apesar da indicação feita pela ferramenta, ficava a cargo dos usuários confirmar se um conjunto de palavras formava um predicado complexo ou não. A maneira de anotar predicados complexos era a seguinte: (1) identificar o verbo do predicado complexo, (2) adicionar um comentário do tipo "É predicado complexo" e, (3) seguindo a metodologia individual, anotar o sentido referente ao predicado complexo e não só ao verbo. Por exemplo, na sentença "Ele dava crédito a ela.", tem-se o predicado complexo "dar crédito": o usuário deve selecionar o verbo "dar", adicionar o comentário de "É predicado complexo" e associar o verbo ao sentido do predicado complexo ("valorizar" / "confiar").

- Procedimento 5: Diferença entre verbo no particípio e adjetivo

Para efeitos da anotação, optamos por distinguir uma palavra como um verbo no particípio se o anotador conseguir trocar da voz passiva para a ativa. Por exemplo, na sentença “A água está poluída pelos agrotóxicos”, a sentença está na voz passiva; então, o que fazemos é tentar trocar para voz ativa, ficando da seguinte forma: “Os agrotóxicos poluem a água”; portanto, “poluído” é anotado como um verbo. Já na sentença “A água está poluída”, a palavra “poluída” é um adjetivo que fornece uma característica da água, e não é possível trocar para a voz ativa. Neste caso, a palavra não é anotada.

3.4.1 Durante a etapa de seleção de tradução

- Procedimento 1: Erros de tradução

Existem palavras que não são bem traduzidas pelo dicionário bilíngue, em consequência não apresentam tradução, nesse caso o usuário realizará a tradução manual.

3.4.2 Durante a etapa de seleção de *synset*

- Procedimento 1: Escolha do *synset*

Para escolher o *synset* mais adequado se recomenda ler e analisar todos os *synsets* mostrados pela ferramenta e seus respectivos hiperônimos e tropônimos, de tal maneira que se tenha um nível de certeza alto ao anotar, sendo que dois *synsets* podem possuir uma alta similaridade produzindo erros na escolha.

- Procedimento 2: Lacunas lexicais

Em casos que exista uma palavra muito específica que não tenha tradução e/ou *synset* explícito, poderá usar-se uma generalização da mesma palavra com o propósito de encontrar a o sentido correto. Por exemplo, tem-se a palavra “jantar” a qual tem como tradução “have a dinner” que não possui *synset* indexado na WordNet.Pr. Por tanto, ter-se-ia que buscar uma generalização que poderia ser “comer”.

- Procedimento 3: Palavras pré-anotadas

Para as palavras pré-anotadas, o anotador primeiro precisa checar se o *synset* pré-anotado é o correto e confirmar a anotação, ou, caso contrário, seguir a metodologia apresentada na Seção 3.2.

3.4.3 Dúvidas durante a anotação

O uso de ferramentas de suporte foi essencial para a tarefa de anotação de sentidos de verbos. Como parte da metodologia de anotação, em caso de dúvidas, os anotadores puderam usar o Google Translate[6] e o Linguee[7] como dicionários bilíngues. Desta forma, os anotadores podiam escolher melhor as traduções para cada palavra e, também, em caso de dúvida, saber o significado das glosas dos *synsets* da WordNet-Pr. Já no caso em que não conseguiam resolver as dúvidas, podiam consultar a toda a equipe de anotação.

3.5 Geração de ontologias on-*line*

Uma das funcionalidades já mencionadas é a geração ou extração de ontologias *on-line.* O benefício que traz essa funcionalidade é a análise do comportamento dos sentidos dos verbos. Uma hipótese que subjaz a criação desta funcionalidade é que os sentidos de verbos mais próximos na ontologia tendem a co-ocorrer em um texto. Na Figura 3.12, apresentam-se duas ontologias geradas ao anotar as palavras hipotéticas "A" e "B" nas coleções de textos hipotéticas C1 e C2. Na coleção C1, a palavra "A" foi anotada com o sentido "A1" 3 vezes. A palavra "B" foi anotada com o sentido "B1" 3 vezes. Pode-se ver no gráfico que as duas palavras estão próximas na ontologia. Já na coleção C2, a palavra "A" foi anotada com o sentido "A2" e não apareceu a palavra B. Além disso, vemos que o sentido "A2" está no lado oposto na ontologia gerada. Com isto, pode-se inferir que, se em um texto novo co-ocorressem as palavras A e B, e o sentido da palavra B fosse "B1", por proximidade na ontologia, o sentido escolhido da palavra A seria "A1" e não "A2". Esse tipo de inferência pode ser valioso para a tarefa de DLS.

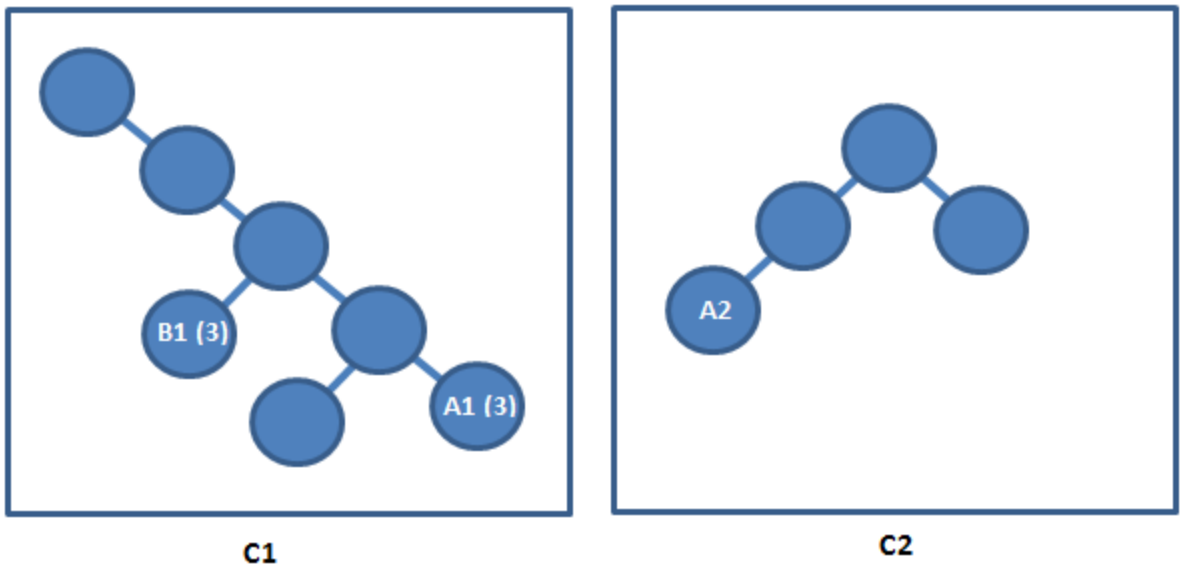

Figura 3.12. Ontologias geradas para as coleções C1 e C2

[6] https://translate.google.com.br/
[7] http://www.linguee.com.br/

A forma pela qual a ontologia é construída é apresentada a seguir: quando o anotador seleciona o *synset* da Wordnet-Pr que corresponde a uma palavra (na Figura 3.13, seguindo o exemplo da anotação, usaremos a palavra "morreram"), a ferramenta obtém automaticamente os filhos, os irmãos e os hiperônimos do mesmo, recursivamente recuperando os hiperônimos intermediários até se chegar ao nível mais alto na hierarquia da WordNet-Pr (aos chamados *unique beginners*, portanto), juntando-os mediante relações de hiponímia e hiperonímia. Com essas relações extraídas, cria-se um grafo internamente (como na Figura 3.13); depois, ao selecionar o um novo *synset* na anotação, o grafo criado para essa escolha é anexado ao grafo anteriormente criado. Finalmente, gera-se um grafo com todas as anotações feitas. O grafo representa a ontologia de sentido de verbos anotados na coleção de textos.

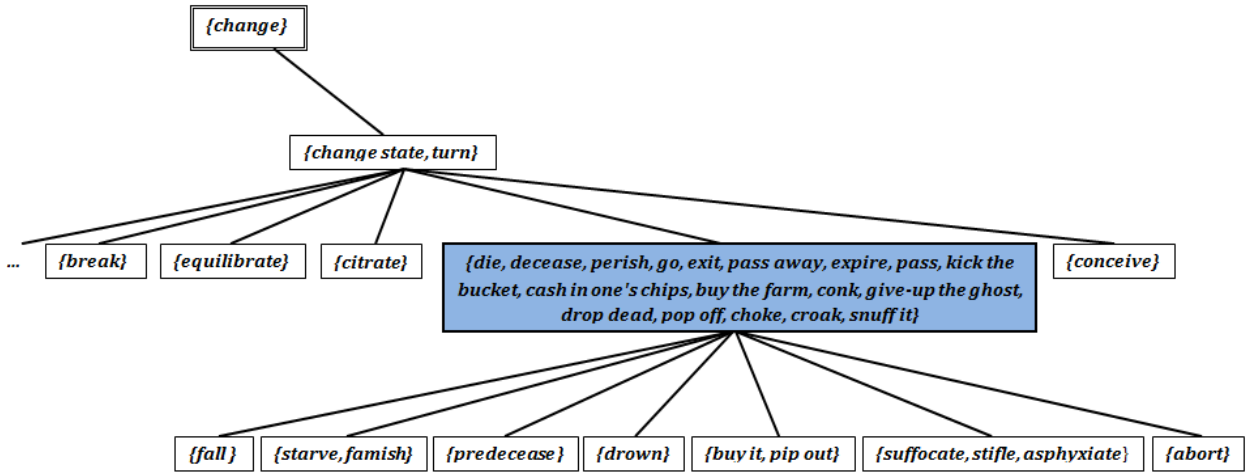

Figura 3.13. Hierarquia adicionada na ontologia ao anotar o sentido de "*morreram*"

Por fim, para a análise futura do comportamento dos sentidos dos verbos, adotou-se a seguinte estratégia: para cada uma das coleções anotadas, gerou-se automaticamente uma ontologia. No final, foram unidas as ontologias de todos as coleções para gerar uma única ontologia dos verbos no córpus CSTNews.

4. Avaliação e Resultados

4.1 Visão geral da anotação

Durante a anotação, foram desambiguadas todas as palavras identificadas como verbos e também as palavras que os anotadores concordaram que eram verbos. O processo de anotação durou 7 semanas e meia, sendo que a primeira metade da semana foi destinada ao treinamento e teste da ferramenta por parte dos anotadores. A anotação aconteceu em reuniões diárias com uma hora de duração.

Cada coleção de textos do CSTNews foi anotada uma única vez por um determinado grupo de anotadores, com exceção das coleções que foram usadas para obter os valores de concordância e, portanto, foram anotadas por todos os grupos.

O agrupamento dos anotadores foi realizado de maneira que membros de diferentes áreas de conhecimento (linguistas e cientistas da computação) trabalhassem juntos. Além disso, procurou-se que os grupos sempre estivessem formados por diferentes pessoas. Dessa forma, pretendia-se nivelar o conhecimento dos anotadores e, assim, a tarefa não começasse a ficar tendenciosa, com "bias", já que, se um grupo permanece anotando várias coleções seguidas, é possível gerar um viés na anotação, afetando-se a qualidade da mesma.

Na Tabela 4.1, apresenta-se a quantidade de instâncias anotadas, incluindo os verbos principais, verbos auxiliares, predicados complexos e erros de anotação.

	Total	**Verbos principais**	**Predicados complexos**	**Verbos auxiliares**	**Erros de anotação**
# instâncias anotadas	**6494**	5082	146	949	317
# porcentagem instâncias	**100 %**	78.26%	2.25%	14.61%	4.88%

Tabela 4.1. Estatísticas da anotação de instâncias do córpus CSTNews

Na Tabela 4.1, salienta-se que as 5082 instâncias de verbos principais que foram anotadas representam 844 verbos principais diferentes. Também, para estes 844 verbos diferentes, foram anotadas 787 traduções e 1047 *synsets* diferentes.

Avaliando a distribuição de *synsets* por palavra no córpus (o número de *synsets* usados para anotar uma palavra em um córpus) apresentada na Figura 4.1, teve-se que foram anotados entre 1 e 18 *synsets* para uma mesma palavra. Contudo, salienta-se que a maioria das palavras (508) foi anotada apenas com um *synset*.

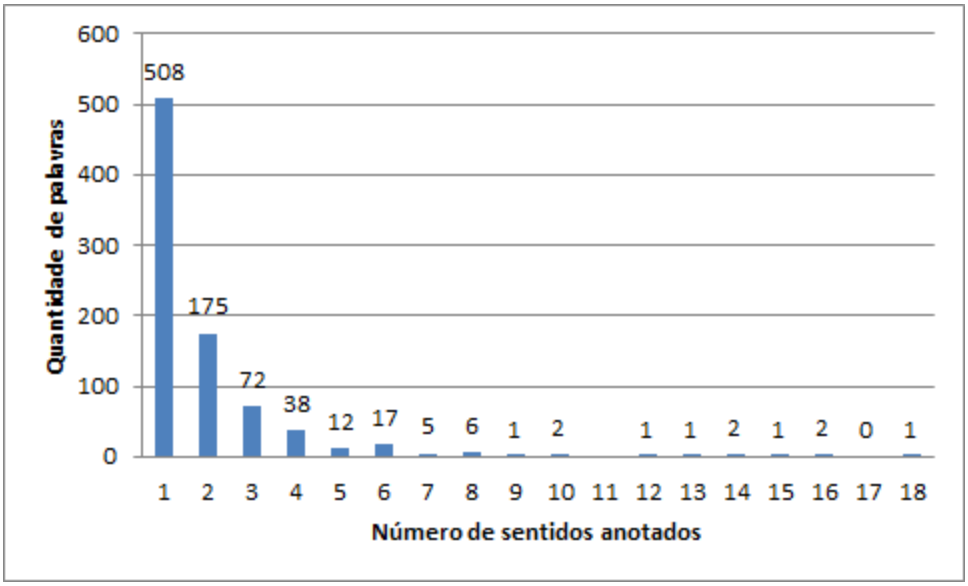

Figura 4.1. Distribuição de *synsets* por palavra no córpus

Na Figura 4.2, apresenta-se a distribuição de *synsets* por palavra por coleção (o número de *synsets* usados para anotar uma palavra dentro de uma coleção). Observa-se que, dentro de uma única coleção de textos, para uma palavra, foram anotados entre 1 e 4 *synsets* diferentes. Como na distribuição anterior, a maior quantidade de palavras (2671) por coleção possuíram somente 1 *synset* anotado.

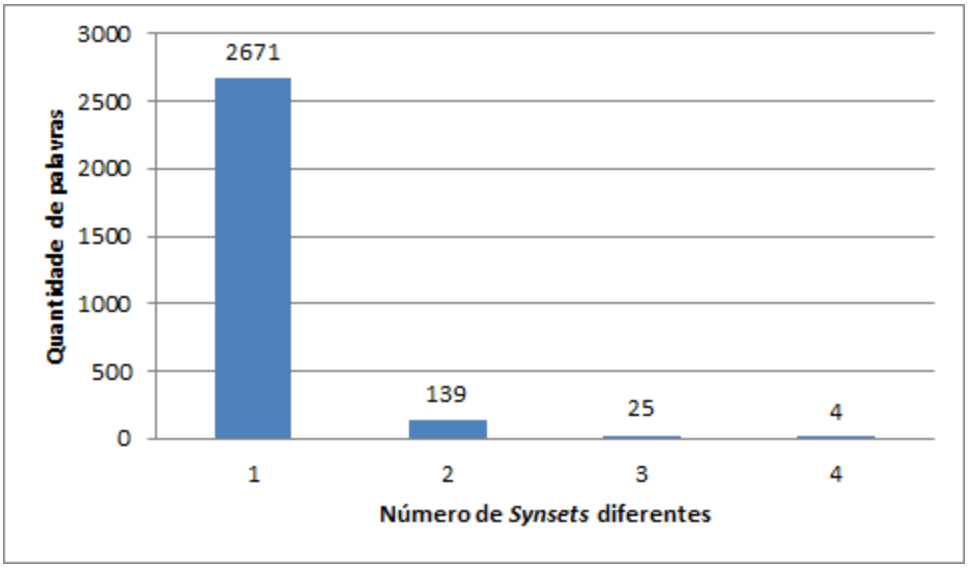

Figura 4.2. Distribuição de *synsets* por palavra por coleção

Comparando-se estas estatísticas com as obtidas por Nóbrega (2013) (como apresentado na Tabela 4.2), ressalta-se que os verbos possuem uma maior variação de sentidos, ou seja, são mais

polissêmicos do que os substantivos, o que está de acordo com os estudos feitos por Miller et al. (1990).

Número máximo de *synsets* anotados por palavra	**Em Nóbrega (2013), para substantivos**	**Neste trabalho, para verbos**
No córpus	5	18
Em uma coleção	3	4

Tabela 4.2. Número máximo de *synsets* por palavra no córpus e por coleção no córpus do trabalho de Nóbrega (2013) e do presente trabalho

Para medir a dificuldade da tarefa de anotação de sentidos de verbos, podemos analisar a quantidade de possíveis *synsets* para cada palavra a ser desambiguada. Com esses valores, podemos saber o nível de ambiguidade dessas palavras. Salienta-se que, ao se trabalhar com palavras escritas em português e usar a WordNet-Pr 3.0 (feita para o inglês) como repositório de sentido, tem-se usado a WordReference® como dicionário bilíngue. Portanto, os possíveis *synsets* retornados provêm das traduções fornecidas pelo WordReference® para cada palavra (quando o usuário não forneceu, manualmente, uma tradução alternativa). Na Figura 4.3, apresenta-se distribuição do número de palavras por número de possíveis *synsets*. Por exemplo, vê-se que 6 palavras têm 39 possibilidades de *synsets*.

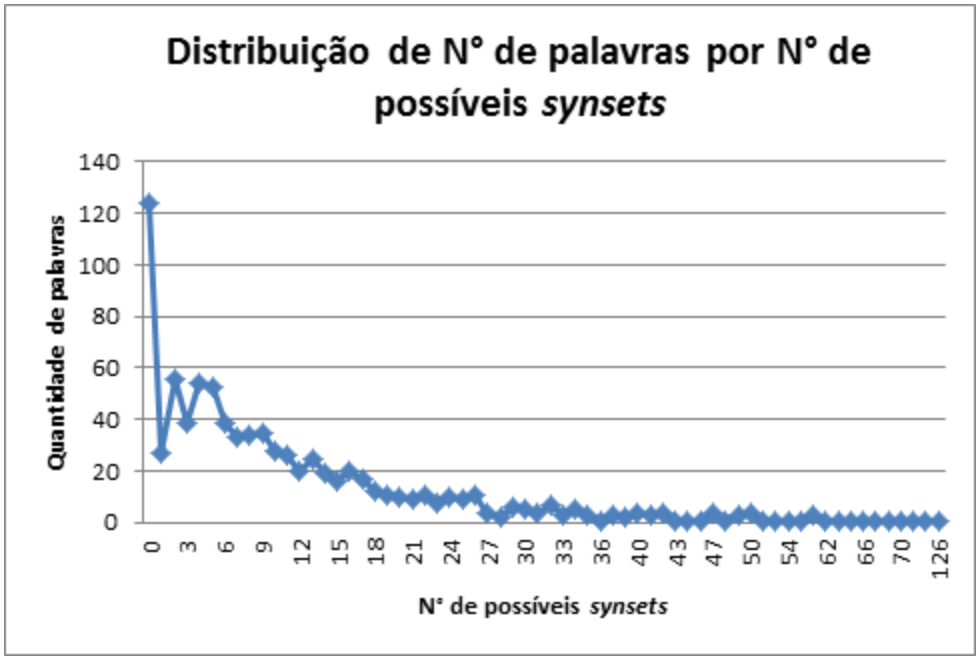

Figura 4.3. Distribuição de número do número de palavras por número de possíveis *synsets*

Analisando esta distribuição, pode-se observar que o nível de ambiguidade das palavras é decrescente, isto é, a quantidade de palavras que tem uma menor quantidade de possíveis *synsets*

é superior à quantidade de palavras com maior quantidade de possíveis *synsets*. A média do número de possíveis *synsets* para cada palavra foi 12 e, das 844 palavras anotadas, 693 (82.11%) possuíram duas ou mais possibilidades de desambiguação (possíveis *synsets*) e, dentre estas, 276 (32.7%) superaram a média de possíveis *synsets*, considerando-se assim como palavras altamente ambíguas.

Comparando os resultados desta distribuição com os obtidos por Nóbrega (2013), apresentados na Tabela 4.3, podemos observar que a tarefa de anotação de sentidos para os verbos é mais difícil do que para os substantivos. Além disso, existe uma maior quantidade de verbos ambíguos do que substantivos ambíguos. Tudo isto concorda com os estudos feitos por Miller et al. (1990), que diz que os verbos são mais polissêmicos dos que os substantivos, vindo disso a dificuldade apresentada. Por outro lado, é interessante notar que há mais substantivos altamente ambíguos do que verbos altamente ambíguos no córpus em questão.

	Em Nóbrega (2013), para substantivos	**Neste trabalho, para verbos**
Média do número de possíveis *synsets* por palavra	6	12
Porcentagem de palavras ambíguas	77%	82.11%
Porcentagem de palavras altamente ambíguas	42%	32.70%

Tabela 4.3. Comparação da distribuição de possíveis *synsets* por palavra entre o trabalho de Nóbrega (2013) e o presente trabalho

Algumas das dificuldades encontradas na anotação são discutidas a seguir.

Apesar da lista de predicados complexos fornecida pela NASP++, a detecção de predicados complexos foi uma tarefa difícil. Por exemplo, a ferramenta mostrava como predicado complexo a expressão “ficaram feridas”, mas, durante a anotação, a palavra “ficaram” foi anotada como verbo auxiliar e também como um predicado complexo.

Outros predicados complexos encontrados foram:

- *“*Levantar o caneco”, cuja tradução é “*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*”.

Também ocorreu falta de *synsets*, por exemplo, para a sentença encontrada "...ficaram desabrigadas...". O anotador anotou "desabrigadas" como verbo, mas não foi encontrado um *synset* para essa palavra (que fosse da categoria dos verbos). 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.

Tampouco foi encontrado *synset* para o verbo "vir" na sentença "O ano que vem...", já que o verbo é parte de uma expressão fixa que quer dizer "o próximo ano" e não tem tradução direta como verbo no inglês.

Algumas palavras que aparentavam ser da classe dos verbos não foram anotadas por pertencerem a outra categoria gramatical. Por exemplo, na sentença "e o governo decretou toque de **recolher**", a palavra "recolher" faz parte do substantivo "toque de recolher" e, portanto, não foi anotada.

Também foram encontrados verbos muito específicos que não tinham tradução ou *synset* indexado na WordNet-Pr. Portanto, optou-se por usar a generalização dos mesmos, quando possível. Alguns dos verbos (ou predicados complexos) foram:

- "tomar frango": não foi encontrada tradução para o inglês. Assim, foi usada a palavra "errar" e usada a tradução "*mistake*" para poder encontrar o *synset* mais apropriado.
- "dar uma meia lua": da mesma maneira, foi necessário usar um termo geral, que é "driblar", e usar a tradução "*dribble*".

Certamente, em todos esses casos, pairam as decisões de anotação dos anotadores humanos envolvidos. Não se descarta a possibilidade de que anotadores com mais conhecimento do domínio de esportes pudessem, por exemplo, lidar melhor com a escolha das traduções e dos *synsets* de termos como os citados anteriormente.

4.2 Avaliação

Na avaliação da anotação, foi usada a medida Kappa (Carletta, 1996). Esta medida calcula o grau de concordância entre os anotadores para uma mesma tarefa, descontando-se a concordância ao acaso. Outras medidas de avaliação usadas foram:

- Concordância Total: número de vezes em que todos os anotadores concordaram, em relação ao total de instâncias.
- Concordância Parcial: número de vezes em que a maioria dos anotadores concordou, em relação ao total de instâncias.
- Concordância Nula: número de vezes em que não houve maioria de concordância, ou em que houve discordância total, em relação ao total de instâncias.

Devido a tarefa de anotação ter uma etapa de tradução para poder obter o sentido da WordNet-Pr, fez-se necessária a avaliação da concordância na (1) etapa de tradução, na (2) etapa de escolha do *synset* e na (3) seleção da tradução com seu respectivo *synset*.

Para avaliar a anotação, foram escolhidas 3 coleções de textos do córpus CSTNews. Para efeitos de comparação com o trabalho de Nóbrega (2013), escolheu-se as mesmas coleções (identificadas por 15, 29 e 50) utilizadas pelo autor no cálculo de concordância. Cada coleção foi anotada por 4 grupos diferentes de anotadores, obtendo-se os resultados apresentados nas Tabelas 4.4, 4.5 e 4.6.

	Kappa	**Total (%)**	**Parcial (%)**	**Nula (%)**
Tradução	0.591	42.11	52.63	5.26
Synset	0.483	35.53	56.58	7.89
Tradução-*Synset*	0.421	28.95	63.16	7.89

Tabela 4.4. Valores de concordância para a coleção15

	Kappa	**Total (%)**	**Parcial (%)**	**Nula (%)**
Tradução	0.659	48.82	48.82	2.36
Synset	0.514	35.43	58.27	6.30
Tradução-*Synset*	0.485	32.28	60.63	7.09

Tabela 4.5. Valores de concordância para a coleção 29

	Kappa	**Total (%)**	**Parcial (%)**	**Nula (%)**
Tradução	0.695	55.50	44.04	0.46
Synset	0.529	34.40	60.55	5.05
Tradução-*Synset*	0.516	33.95	60.09	5.96

Tabela 4.6. Valores de concordância para a coleção 50

Um detalhe a salientar é o incremento dos valores da medida Kappa desde a coleção 15 até a 50. Acreditamos que isto se deva ao fato de que as coleções foram anotadas em datas diferentes, na coleção 15 ainda se apresentava certo erro, produto do aprendizado dos anotadores. Conforme se foi avançando na anotação, esses valores foram melhorando.

Deve-se também aventar a hipótese de que o tema abordado na coleção 15 (“explosão em um mercado em Moscou”) pode ser eventualmente menos familiar aos anotadores do que o da coleção 29 (“pagamento de indenização pela igreja católica”), que, por sua vez, pode ser menos familiar do que o da coleção 50 (“proposta do governo sobre cobrança de imposto”). Supondo-se que o domínio do assunto é um fator importante para um bom desempenho na anotação, tal hipótese necessita de verificação posterior.

Na Tabela 4.7, apresentam-se os valores de concordância médios para a anotação de sentidos de verbos. Pode-se observar que o valor Kappa obtido para concordância entre *synsets* anotados é de 0.509, que é considerado aceitável no cenário da DLS. O valor de concordância para as traduções supera ao de *synsets.* Isto era esperado, já que a tradução é uma tarefa menos subjetiva que a desambiguação lexical de sentido. Finalmente, na concordância para os pares tradução-*synset*, o valor é menor, devido a diferentes traduções poderem fazer referência ao mesmo *synset* e diferentes *synsets* poderem ser referenciados pela mesma tradução.

Avaliando-se as outras medidas de concordância, salienta-se a porcentagem alta de concordância parcial. Isto mostra que, mesmo com a Kappa apresentando resultados aceitáveis, os anotadores tiveram dúvidas na anotação. Algumas das causas que ocasionam esse cenário são a identificação de verbos em particípio e também a identificação dos verbos auxiliares. Por exemplo, no fragmento “foi cancelada”, a palavra “foi” foi anotada como verbo auxiliar e também como um verbo principal em algumas ocasiões; e a palavra “cancelada” foi ora anotada como se fosse um adjetivo (indicando, portanto, que se trata de um erro de anotação do *tagger*) e também como se fosse um verbo principal. Outro ponto importante é que os níveis de concordância nula sempre foram baixos.

	Kappa	**Total (%)**	**Parcial (%)**	**Nula (%)**
Tradução	0.648	48.81	48.50	2.69
Synset	0.509	35.12	58.47	6.41
Tradução-*Synset*	0.474	31.73	61.29	6.98

Tabela 4.7. Valores de concordância gerais

Comparando-se com o trabalho de Nóbrega (2013), cujos resultados são apresentados na Tabela 4.8, podemos ver que os valores de concordância para os substantivos são, na maioria, superiores aos verbos. Este resultado era esperado, devido a maior complexidade que os verbos apresentam e ao maior grau de polissemia presente nos verbos.

	Kappa	**Total (%)**	**Parcial (%)**	**Nula (%)**
Tradução	0.853	82.87	11.08	6.05

Synset	0.729	62.22	22.42	14.36
Tradução-*Synset*	0.697	61.21	24.43	14.36

Tabela 4.8. Valores de concordância da anotação feita por Nóbrega (2013)

A seguir, como ilustração, apresenta-se uma lista de algumas palavras que obtiveram concordância total e também algumas sentenças que incluem estas palavras:

morreram	investigar	acabado	considerou	têm
reduzir	acreditar	informou	disseram	anunciou
hospitalizada	cometidos	abusados	convencer	começa

(1) "Nove pessoas **morreram**, três delas crianças, e..."

(2) A maioria dos feridos, entre os quais há quatro com menos de 18 anos, foi **hospitalizada**.

(3) A procuradoria de Moscou anunciou a criação de um grupo especial para **investigar** o acidente."

Algumas das razões aventadas para a concordância total para essas palavras foram:

- Palavras com *synsets* bem diferenciados e os verbos encontrados nas sentenças possuírem um sentido claro. Por exemplo, na sentença 1, é fácil determinar o sentido do verbo "morrer" como "perder todos os atributos e funções corporais para manter a vida" (*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}: pass from physical life and lose all bodily attributes and functions necessary to sustain life*).
- Palavras com menor número de *synsets* possíveis. Por exemplo, na sentença 2, a palavra "hospitalizada" possui um *synset* associado (*{hospitalize, hospitalise}: admit into a hospital*) na WordNet-Pr e, na sentença 3, s palavra "investigar" possue 2 possíveis *synsets* na WordNet-Pr e esses *synsets* representam sentidos bem diferenciados (*{investigate, inquire, enquire}: conduct an inquiry or investigation of* e *{investigate, look into} : investigate scientifically*, sendo escolhido o primeiro).

A seguir, mostra-se uma lista de palavras que obtiveram concordância nula:

localizado	estimado	fossem	levada	aceitamos
registrada	conseguiram	surgirem	somariam	deixou

entraram	enfrentar	entenderam	haverá	
adiantaram	tramitando	levar	daria	
assinalaram	veio	caminhe	aceitamos	

Apresentam-se algumas sentenças, extraídas do córpus, que incluem verbos com concordância nula:

(1) "A bomba detonou no interior de uma cafeteria **localizada** no setor denominado "Evrazia" do mercado Cherkizov."
Synsets usados na anotação:
- *{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*

(2) "...fontes da polícia moscovita **adiantaram** que ela teria acontecido provavelmente por causa da explosão acidental de um bujão de gás."
Synsets usados na anotação:
- *{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*

(3) "As autoridades policiais de Moscou **assinalaram** que no recinto do mercado..."
Synsets usados na anotação:
- *{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:

- Foram usadas diferentes traduções para as palavras e os *synsets* utilizados, associados a essas traduções, eram diferentes, mesmo possuindo alguma relação entre eles. Por exemplo, na sentença 1, foram usadas as traduções "*locate*" e "*localize*" para a palavra "localizada", e, para cada uma, usaram-se *synsets* diferentes. Nas sentenças 2 e 3, a mesma situação ocorreu.
- As palavras não tinham um sentido tão bem definido, o que fez com que os anotadores colocassem traduções diferentes e/ou *synsets* diferentes.

- Os *synsets* utilizados possuíam certa similaridade, fazendo com que se gerasse confusão entre os anotadores.

5. Considerações finais

A construção de um córpus com sentidos anotados é uma tarefa difícil de ser feita. Para o caso dos verbos, dada a complexidade, advinda do alto grau polissêmico, os níveis de concordância entre os anotadores é relativamente baixo. Contudo, a criação de um córpus anotado com sentidos de verbos abre a possibilidade de se pesquisar com nível de profundidade maior a área de DLS para o português, que tem sido pouco relativamente estudada, principalmente quando se consideram os trabalhos para outras línguas.

Um detalhe a salientar é que, apesar do uso de um repositório de sentidos feito para o inglês, os níveis de concordância não foram tão prejudicados, o que era esperado, já que a WordNet-Pr é considerada também uma ontologia por alguns autores, abrangendo o conhecimento geral do mundo. Contudo, uma das limitações encontradas durante a anotação é a falta de traduções. Apesar de a ferramenta de anotação usar um dicionário bilíngue, existem palavras e/ou expressões próprias de um contexto específico que não são encontradas nos dicionários bilíngues tradicionais, por exemplo, o "pedalar" no contexto de esportes, entre outras. A criação e introdução de dicionários bilíngues específicos poderiam trazer benefícios à tarefa de anotação.

Outro detalhe interessante é a ocorrência de lacunas lexicais, que são palavras do português, as quais não possuem sentidos, diratemente relacionados, indexados na WordNet-Pr. Contudo, embora não existam sentidos na WordNet-Pr, podem-se usar generalizações dos sentidos do português, para assim poder encontra o sentido correto no repositório. Por exemplo, "tomar frango" não possue sentido na WordNet-Pr, portanto, teve-se que generalizar o termo para "errar" e depois encontrar o *synset* associado.

Como foi mencionada, uma das dificuldades encontradas é a identificação de predicados complexos, verbos auxiliares e a distinção de verbos no particípio de adjetivos. A tarefa de detecção de predicados complexos é um problema difícil de ser tratado, já que é também uma área de pesquisa propriamente dita. No caso dos verbos auxiliares e a distinção de verbos no particípio, métodos baseados em regras podem ser implementados para melhorar a ferramenta de anotação e, assim, ajudar os anotadores na tarefa.

Finalmente, o córpus anotado e a ferramenta NASP++ estão disponíveis na página do projeto SUCINTO, em www.icmc.usp.br/pessoas/taspardo/sucinto/resources.html

Referências Bibliográficas

Aires, R. V. X. (2000). Implementação, adaptação, combinação e avaliação de etiquetadores para o português do Brasil. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brasil.

Aleixo, P.; T. A. S. Pardo (2008). CSTNews: um córpus 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, no. 326. São Carlos-SP, Maio, 12p.

Baptista, J. (2012). ViPEr: A Lexicon-Grammar of European Portuguese Verbs, in *Proceedings of the 31st International Conference on Lexis and Grammar*, Nove Hrady, Czech Republic, pp. 10-16.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics 22*, Cambridge, MA, USA, pp. 249-254.

Cardoso, P. C. F.; E. G. Maziero; M. Jorge; E. M. Seno; A. Di Felippo; L. H. Rino; M. G. V. Nunes; T. A. S. Pardo (2011). CSTNews - a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian Portuguese, in *Proceedings of the 3rd RST Brazilian Meeting*, Cuiabá, MT, Brasil, pp. 88-105.

Dias-da-Silva, B. C. (2005). A construção da base da wordnet.br: Conquistas e desafios, in *Proceedings of the Third Workshop in Information and Human Language Technology (TIL 2005), in conjunction with XXV Congresso da Sociedade Brasileira de Computação*. São Leopoldo, RS, Brasil, pp. 2238–2247.

Duran, M. S.; S. M. Aluísio (2010). Verbos Auxiliares no português do Brasil. Relatório técnico, Brasil.

Duran, M. S.; C. Ramisch; S. M. Aluísio; A. Villavicencio (2011). Identifying and Analyzing Brazilian Portuguese Complex Predicates, in *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, Portland, OR, USA. pp. 74-82.

Fellbaum, C. (1998.) WordNet: An Electronic Lexical Database. 2. Ed. Cambridge (Mass.): MIT Press.

Fillmore, C. J. (1968). The Case for Case, in *Universals in Linguistic Theory*, New York, USA. pp. 1-88.

Gonçalo Oliveira, H.; L. Antón Perez; P. Gomes (2012). Integrating lexical-semantic knowledge to build a public lexical ontology for Portuguese, in *Proceedings of the 17th International Conference on Applications of Natural Language Processing and Information Systems*, Groningen, The Netherlands, pp. 210-215.

Jurafsky, D.; J. H. Martin(2009). Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd Ed.). Prentice Hall. Pearson. 988p.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell pine cone from an ice cream cone, in *Proceedings of 5th Annual International Conference on Systems Documentation*, New York, NY, USA, pp. 24-26. Association for Computing Machinery.

Machado, I. M.; R. de Alencar; J. Campos; R. De Oliveira; C. A. Davis (2011). An ontological gazetteer and its application for place name disambiguation in text, in *Journal Brazilian Computional Society*, pp. 267-279.

Maziero, E. G.; Pardo, T. A. S.; Felippo, A. D.; Da Silva, B. C. D. (2008). A base de dados lexical e a interface web do tep 2.0 - thesaurus eletrônico para o português do Brasil, in *Anais do VI Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, pp. 390-392.

Miller, G. A.; Beckwith, R.; Fellbaum, C.; Gross, D.; Miller, K. (1990). Introduction to Wordnet: An on-line lexical database, in International Journal of Lexicography, pp. 235-244.

Miller, G. A. (1995). WordNet: A Lexical Database for English. Communications of the ACM 38, New York, NY, USA, pp. 39-41.

Mihalcea, R. (2006). "Knowledge-Based Methods for WSD", in *Word Sense Disambiguation: Algorithms and Applications*, pp. 107-132. Springer.

Nóbrega, F. A. A. (2013). Desambiguação Lexical de sentidos para o português por meio de uma abordagem multilíngue mono e multidocumento. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brasil.

Plaza, L.; A. Diaz (2011) Using semantic graphs and word sense disambiguation techniques to improve text summarization, in *XXVII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Huelva, Espanha, pp. 97-105.

Ratnaparkhi, A. (1996).A Maximum Entropy Par t-Of-Speech Tagger, in *Proceedings of the Empirical Methods in Natural Language Processing Conference*, University of Pennsylvania.

Rocha, P. A.; D. Santos (2000). CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa, in *V Encontro para o processamento computacional da língua portuguesa escrita e falada*, Atibaia, São Paulo, ICMC/USP, pp. 131-140.

Ribeiro, R. (2003). Anotação Morfossintáctica Desambiguada do Português, Dissertação de Mestrado, Instituto Superior Técnico, Lisboa, Portugal.

Specia, L. (2007). Uma Abordagem Híbrida Relacional para a Desambiguação Lexical de Sentido na Tradução Automática. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brasil.

Travanca, T. (2013) Verb Sense Disambiguation, Dissertação de Mestrado, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.