Núcleo Interinstitucional de Linguística Computacional - NILC
Universidade de São Paulo – USP
Universidade Federal de São Carlos – UFSCar
Instituto Federal de São Paulo - IFSP

Anotação de Sentidos de Verbos no Córpus CSTNews

Relatório Técnico do NILC
NILC - TR - 14 - 05

Marco A. Sobrevilla Cabezudo[1], Erick G. Maziero[1], Jackson W. C. Souza[2], Márcio S. Dias[1], Paula C. F. Cardoso[1], Pedro P. Balage Filho[1], Verônica Agostini[1], Fernando A. A. Nóbrega[1], Cláudia D. Barros[3], Ariani Di Felippo[2], Thiago A. S. Pardo[1]

[1]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
[2]Departamento de Letras, Universidade Federal de São Carlos
[3]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo

Série de Relatórios do Núcleo Interinstitucional de Linguística Computacional NILC-ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil

RESUMO

Um dos desafios do Processamento das Línguas Naturais (PLN) em nível semântico é a ambiguidade lexical, já que as palavras podem expressar significados distintos em função do contexto em que ocorrem. Em PLN, a tarefa responsável por determinar o significado adequado de uma palavra em contexto é a Desambiguação Lexical de Sentido (DLS). 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 línguas 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 e avaliadas, funcionando como um *benchmark* para a área de DLS.

Para o português, há o córpus multidocumento CSTNews (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[1]) (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*.

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

[1] http://wordnetweb.princeton.edu/

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), investigam-se métodos de desambiguação de substantivos comuns, como um método 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). Para anotar os substantivos, ou seja, para explicitar os conceitos a eles subjacentes, utilizou-se o repositório de sentidos da 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 base em um dicionário bilíngue português/inglês. No caso, utilizou-se o WordReference®[2]. Mais detalhes sobre o CSTNews são apresentados na próxima seção.

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.

[2] http://www.wordreference.com/

3. Anotação de Córpus

3.1 Considerações iniciais

A anotação reportada neste relatório 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). Além disso, dá-se continuidade ao trabalho realizado por Nóbrega (2013).

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 (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, 2002).

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 jornalísticas das quais os textos foram compilados 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*. 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 córpus é composto por coleções das seguintes categorias: "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 etiquetação morfossintática automática, realizada pelo etiquetador ou *tagger* MXPOST (Ratnaparkhi, 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.

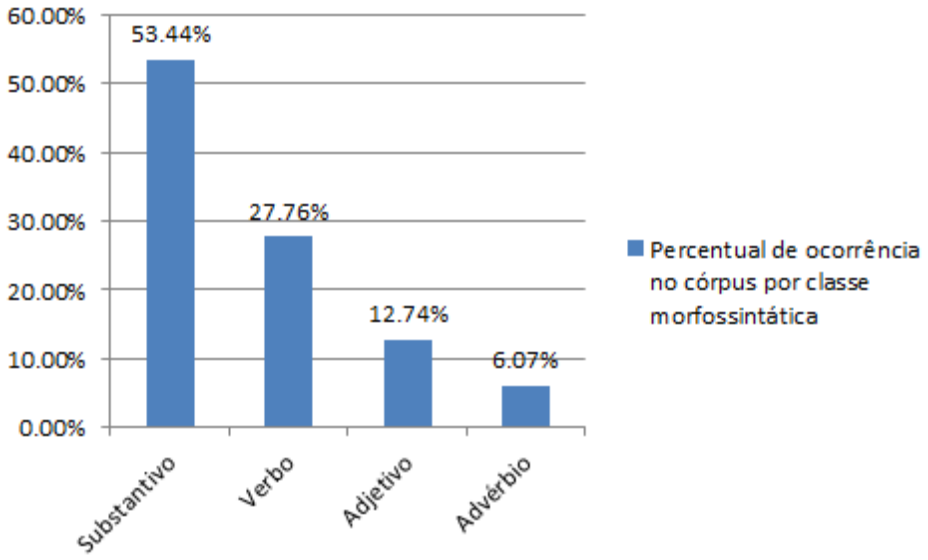

Figura 3.1. Percentual de ocorrência no córpus por classe morfossintática (dados obtidos de Nóbrega, 2013).

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. Mesmo tendo 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 princípios da ciência cognitiva e em um formato computacionalmente tratável, e (ii) abrangência, já que a versão 3.0 possui mais de 155287 unidades lexicais do inglês e 117659 sentidos. 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.

Especificamente, a WordNet.Pr é uma rede em que as palavras e expressões, pertencentes às classes dos nomes, verbos, adjetivos e advérbios, organizam-se sob a forma de *synsets* (do inglês, *synonym sets*). Em outras palavras, pode-se dizer que o *synset* é um conjunto de formas (do inglês, *word forms*) de uma mesma categoria gramatical que podem ser intercambiáveis em determinado contexto, p.ex., {*bicycle, bike, wheel, cycle*}. O *synset*, por definição, é construído de modo a codificar um único conceito lexicalizado por suas formas constituintes.

Entre os *synsets*, codificam-se 5 principais relações lógico-conceituais:

a) Hiperonímia/ Hiponímia: relação entre um conceito mais genérico (o hiperônimo) e um conceito mais específico (o hipônimo), p.ex.: o *synset* {*limusine*} é hipônimo do *synset* {*car, auto, automobile, machine, motorcar*} e, por conseguinte, {*car, auto, automobile, machine, motorcar*} é hiperônimo de {*limusine*}.
b) Antonímia: relação que engloba diferentes tipos de oposição semântica. *Antonímia complementar*: relaciona pares de itens lexicais contraditórios em que a afirmação do primeiro acarreta a negação do segundo e vice-versa, por exemplo: {*alive*} e {*dead*}. *Antonímia gradual*, que relaciona itens lexicais que denotam valores opostos em uma escala como {*small*} e {*big*}. *Antonímia recíproca*, que relaciona pares de itens lexicais

que se pressupõem mutuamente, sendo que a ocorrência do primeiro pressupõe a ocorrência do segundo, como {*buy*} e {*sell*}.

c) Meronímia/ Holonímia: relação entre um *synset* que expressa um "todo" (holônimo), como {*car, auto, automobile, machine, motorcar*}, e outros *synsets* que expressam partes do todo (merônimos), como {*window*}, {*roof*}, {*car door*}, etc.
d) Acarretamento: relação que se estabelece entre uma ação A1 e uma ação A2; a ação A1 denotada pelo verbo x acarreta a ação A2 denotada pelo verbo y se A1 não puder ser feita sem que A2 também o seja. Assim, estabalece-se a relação de acarretamento entre {*snore*} e {*sleep*}. Vale salientar que o acarretamento é uma relação unilateral, isto é, {*snore*} acarreta {*sleep*}, mas {*sleep*} não necessariamente acarreta {*snore*}.
e) Causa: relação que se estabelece entre uma ação A1 e uma ação A2 quando a ação A1 denotada pelo verbo x causa a ação A2 denotada pelo verbo y. Esse é o caso, por exemplo, da relação entre {*kill*} e {*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*}.

3.2 Metodologia de Anotação

Para a anotação em questão, seguiu-se a mesma metodologia de Nóbrega (2013), que engloba diretrizes gerais e etapas específicas.

Quanto às diretrizes, ressalta-se que, diante de uma coleção do CSTNews, os anotadores humanos deveriam seguir 4 diretivas básicas, a saber: (i) escolher um dos textos da coleção para ser anotado; (ii) anotar todas as palavras da classe dos verbos que ocorreram no texto escolhido em (i); (iii) anotar o próximo texto da coleção após a anotação de todos os verbos do texto escolhido em (i) e assim sucessivamente até que todos os textos da coleção tenham sido anotados; (iv) revisar a anotação de todos os textos da coleção e salvá-la no formato e endereço especificados.

Como a WordNet.Pr codifica os conceitos em conjunto de sinônimos do inglês norte-americano, delimitou-se um conjunto de 4 etapas para a anotação de cada verbo distinto, a saber:

a) selecionar um verbo *x* a ser anotado;
b) traduzir o verbo *x* para o inglês;
c) selecionar um *synset* da WordNet.Pr que represente o conceito subjacente a *x*;
d) anotar o verbo *x* com o *synset* escolhido em (c).

Na Figura 3.2, a sequência metodológica composta pelas 4 etapas de anotação semântica e os recursos linguísticos utilizados na tarefa estão representados em um fluxograma. As 4 etapas metodológicas da Figura 3.2 subsidiaram a construção do editor NASP++, isto, é, uma ferramenta de auxílio à anotação semântica, a qual é descrita na próxima subseção. Em outras palavras, o editor NASP++ é a materialização computacional da metodologia definida para a anotação em questão, sendo por meio dela que os verbos do CSTNews foram efetivamente anotados.

Antes, porém, da descrição das funcionalidades da ferramenta NASP++, descreve-se cada uma das etapas metodológicas, sobretudo as etapas A, B e C, enfatizando-se os critérios para a realização de cada uma delas.

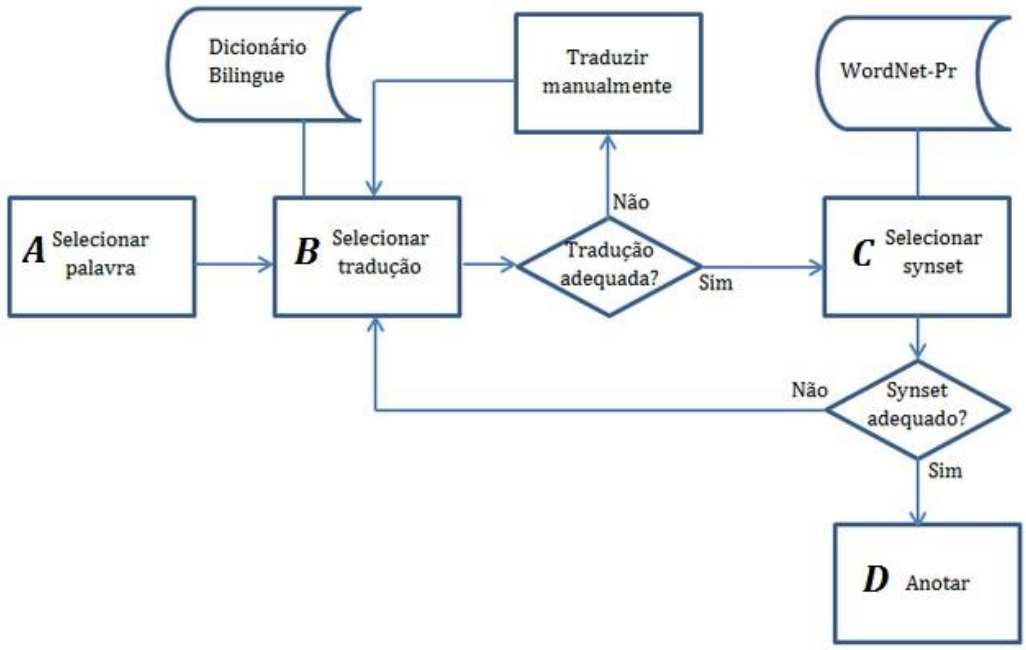

Figura 3.2. Metodologia de anotação.

3.2.1 Seleção dos verbos para anotação (Etapa A)

Como mencionado, a anotação em questão tem início com a seleção de um verbo que ocorre em um dos textos-fonte de dada coleção. Para a adequada identificação dessas unidades lexicais, estabeleceram-se 5 diretrizes ou regras específicas.

A primeira delas estabelece a revisão da identificação dos verbos no texto-fonte. Isso se deve ao fato de que, para agilizar o processo de anotação semântica, optou-se por partir de textos-fonte anotados em nível morfossintático pelo *tagger* MXPOST (Rapnaparkhi, 1986), que, em experimentos realizados por Aires (2000), obteve uma acurácia de 97%. Apesar de bastante preciso, o *tagger* comete erros e, por isso, a etapa de seleção do verbo a ser anotado englobou a tarefa de revisão da anotação morfossintática (ou *tagging*).

Assim, a cada palavra anotada como verbo, verificava-se se de fato a palavra em questão era um verbo. Caso a anotação automática estivesse correta, passava-se para a próxima etapa da anotação semântica. Caso a palavra não fosse de fato um verbo, configurando um caso de ruído, tal anotação era apenas ignorada. 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. 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 havia algum verbo não identificado como tal localizado entre o último verbo (semanticamente) anotado e o próximo a anotar. Se sim, o anotador humano realizava a anotação morfossintática (por meio do editor NASP++) e seguia para a próxima etapa da anotação semântica. Caso contrário, não se seleciona a palavra para ser anotada.

A segunda diretriz dessa etapa da anotação estabelece que os verbos auxiliares deveriam ser anotados como tal, especificamente pelo comentário "verbo auxiliar". Dessa forma, não se atribui sentidos/*synsets* a eles. Ressalta-se que os auxiliares ocorrem nos tempos compostos, os quais são formados por um [verbo auxiliar + forma nominal do verbo principal] (gerúndio, particípio ou infinitivo). Por exemplo, em "Ele havia saído de casa", "havia" é verbo auxiliar e "saído" (particípio) é o verbo principal. Os verbos principais

carregam a carga semântica da forma verbal composta, já os verbos auxiliares são responsáveis por marcar o tempo, o modo, o número e a pessoa daquela forma verbal.

A terceira diretriz estabelece que, nas ocorrências formadas por um tempo composto seguido de infinitivo, o verbo principal (do composto) e o infinitivo deveriam ser anotados, posto que estes 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 conteúdo bem definidos e independentes.

A quarta diretriz estabelece que, nos casos de predicados complexos (isto é, expressões perifrásticas que comumente possuem um equivalente semântico lexicalizado, p.ex.: "fazer uma queixa" → "queixar-se", "dar uma contribuição" → "contribuir" e "tomar conta" → "cuidar"), deve-se: (i) associar ao verbo da expressão o comentário "predicado complexo" e (ii) anotar o verbo com um sentido/*synset* da WordNet.Pr que codifique o significado global do predicado complexo. Assim, em "*Ele dava crédito a ela*", deve-se associar o comentário "predicado complexo" ao verbo "dava" e anotá-lo com um *synset* que represente o sentido do predicado complexo, que é "valorizar" / "confiar". Ressalta-se que a identificação dos predicados complexos foi automática, por meio do 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 fato de um predicado complexo ficava a cargo dos anotadores humanos. Outros 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".

A quinta diretriz de anotação diz respeito à identificação dos verbos no particípio. 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 (2010, pág. 242/243), essa regra ou diretriz estabeleceu 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)."

3.2.2 Tradução dos verbos para o inglês (Etapa B)

Como mencionado, a WordNet.Pr foi utilizada como repositório de sentidos ou conceitos para a anotação semântica ora relatada. Como tais sentidos estão organizados em *synsets*, as 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 *x*, o editor NASP++ recupera todos os *synsets* da WordNet.Pr dos quais *x* 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 ou conceito 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 do equivalente de tradução.

A primeira delas estabeleceu que todas as traduções sugeridas pelo editor fossem analisadas antes da seleção definitiva do equivalente de tradução. Essa regra foi estabelecida com o objetivo de se selecionar a tradução mais adequada em inglês. Por exemplo, se, para um verbo em português *x*, o editor sugerisse 4 traduções possíveis em inglês, *y, w, x,* e *k,* todas elas deveriam ser analisadas. Essa análise pode englobar a consulta a recursos diversos, como o *Google Tradutor*[3], o *Linguee*[4] 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 estabeleceu que, caso o editor não sugerisse uma tradução adequada, o anotador deveria inserir uma manualmente, a partir da qual os *synsets* seriam recuperados automaticamente na sequência. Para sugerir um equivalente de tradução manualmente, sugeriu-se que os anotadores consultassem os mais variados recursos linguísticos. Dentre eles, citam-se os dicionários bilíngues português-inglês, como o *Michaelis Moderno Dicionário Inglês & Português*[5] e os diferentes dicionários disponíveis no site *Cambridge Dictionaries Online*[6] e os serviços *online* como o *Google Tradutor* e o *Linguee*.

A consulta a esses recursos buscou garantir a inserção no editor do equivalente de tradução mais adequado, ou seja, que codificasse de fato o conceito subjacente ao verbo em português e que estivesse armazenado na WordNet.Pr.

3.2.3 Seleção dos synsets (Etapa C)

Uma vez que um equivalente de tradução em inglês tenha sido selecionado, o próximo passo na metodologia consiste em selecionar o *synset* que representa o conceito subjacente ao verbo original em português.

Quanto a essa etapa da metodologia, ressalta-se que, assim que um equivalente de tradução é selecionado, seja indicado pelo editor ou inserido manualmente pelo anotador, deve-se analisar os *synsets* compostos pelo equivalente de tradução para verificar se entre eles há um que seja pertinente. Para essa análise, aliás, deve-se levar em consideração o fato de que a WordNet.Pr, por vezes, apresenta conceitos/*synsets* muito próximos, cuja distinção nem

[3] https://translate.google.com/
[4] http://www.linguee.com.br/
[5] http://michaelis.uol.com.br/
[6] http://dictionary.cambridge.org/

sempre é simples.

Diante de uma lacuna léxico-conceitual, ou seja, a inexistência de um *synset* que representasse o conceito específico subjacente a uma palavra, a segunda diretriz dessa etapa da anotação estabelece 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. Por tanto, ter-se-ia que buscar uma generalização que poderia ser "driblar".

3.3 Ferramenta de Anotação: NASP++

Como mencionado, a metodologia e os recursos ilustrados na Figura 3.2 subsidiaram o desenvolvimento da ferramenta NASP++, que pode ser definida como um editor de auxílio à anotação de sentidos. A NASP++ é, na verdade, uma extensão ou versão atualizada da ferramenta NASP (Nóbrega, 2013), que originalmente fora desenvolvida para a anotação semântica dos nomes ou substantivos.

3.3.1 As funcionalidades do NASP++

A ferramenta disponibiliza para os anotadores humanos as seguintes funcionalidades:

a) Anotação semântica dos conceitos ou sentidos subjacentes às palavras das classes dos nomes e verbos que ocorrem em textos em português;
b) Adição, às anotações, de um dos seguintes comentários:
   - Sem comentários: observação por *default*; 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;
   - É predicado complexo: aplica-se quando o verbo a ser anotado pertence a um predicado complexo. Por exemplo, na sentença "*A mulher bateu as botas*", o verbo "*bateu*" deve ser anotado com o comentário "É predicado complexo".
   - É verbo auxiliar: aplica-se quando o verbo identificado pelo *tagger* é um verbo auxiliar. Por exemplo, na sentença "*Ele estava jogando bola*", 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 ou substantivo), incluindo dificuldades de anotação.
c) Delimitação da quantidade de palavras para anotação: ao contrário da NASP, que restringia a anotação semântica aos substantivos que compunham o conjunto dos 10% mais frequentes da coleção de textos-fonte, a ferramenta NASP++ não possui essa limitação, sendo que qualquer porcentagem dos verbos (e também substantivos) que ocorrem nos textos-fonte pode ser submetida ao processo de anotação semântica;
d) Geração de ontologia: por meio dessa funcionalidade, o editor NASP++ recuperada, da WordNet.Pr, a hierarquia léxico-conceitual à qual cada conceito/*synset* utilizado na anotação pertence e unifica as hierarquias individuais de cada conceito em uma única estrutura hierárquica (cf. Seção 3.4).

3.3.2. A interface gráfica

Na Figura 3.3, apresenta-se a interface gráfica principal da ferramenta, composta pelos seguintes campos: (i) visualizador dos textos-fonte para anotação (A); (ii) painel para exibição e seleção das traduções (B); (iii) painel para exibição e seleção dos *synsets* (C), e (iv) painel para anotação dos comentários (D).

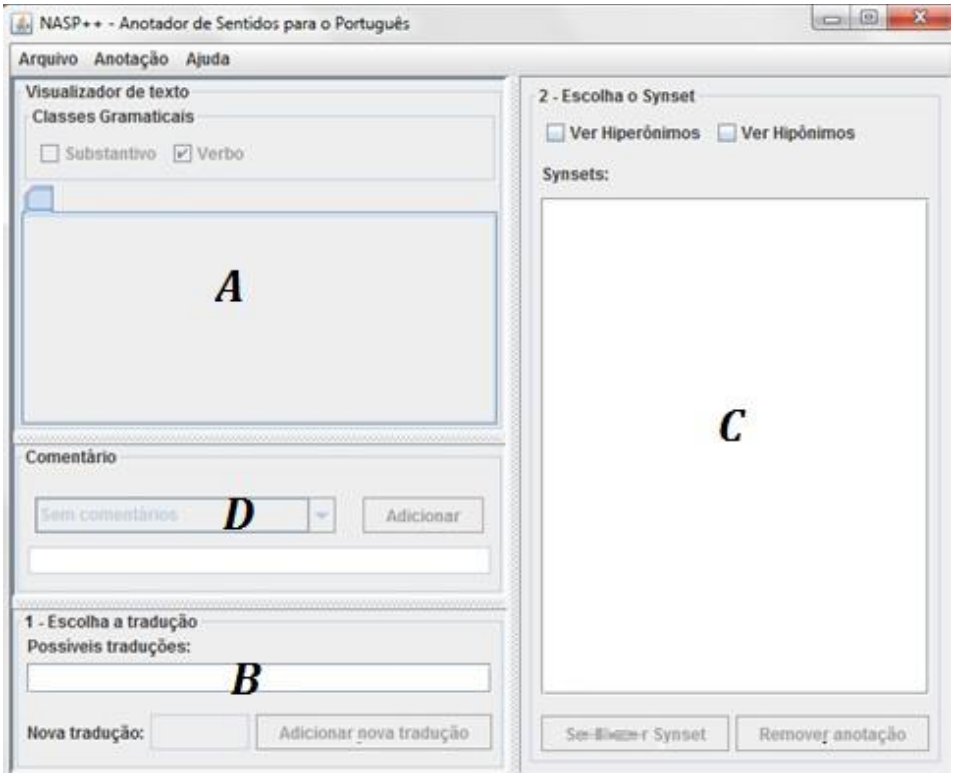

Figura 3.3. Tela principal da NASP++.

Na sequência, ilustra-se cada uma das funcionalidades realizadas por cada um dos campos que compõem a interface principal.

Na Figura 3.4, os três textos a serem anotados foram carregados e são exibidos aos anotadores humanos por meio do campo "visualizador de textos-fonte" (A) (cf. Figura 3.3). Nos textos-fonte exibidos, as palavras destacadas em "vermelho" foram automaticamente identificadas como "verbo" pelo etiquetador morfossintático MXPost. A partir das palavras em destaque, tem-se início o processo de anotação. Por exemplo, no caso do Texto 1 da Figura abaixo, a anotação tem início com o primeiro verbo em destaque, no caso, "morreram".

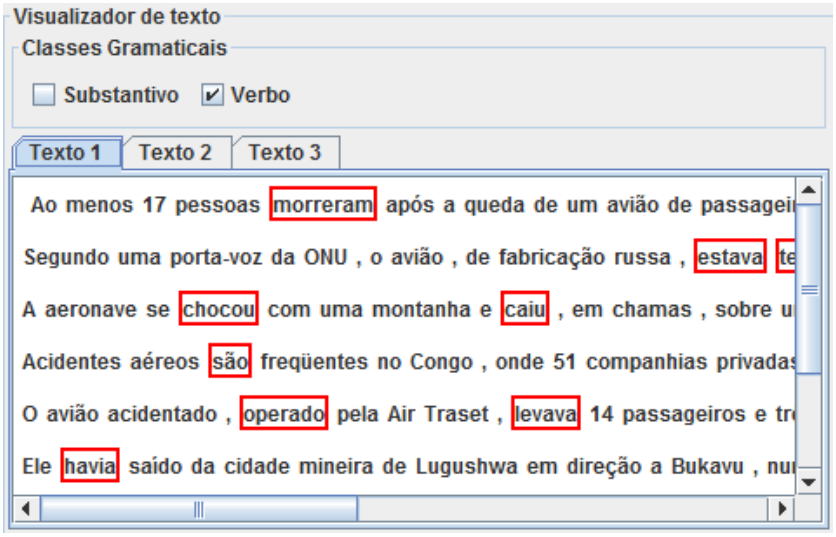

Figura 3.4. Visualizador de textos.

Ainda no campo "visualizador de textos-fonte", clica-se na palavra "morreram" e automaticamente duas tarefas são realizadas pelo editor: (i) ativação do "painel de comentários" (D) (cf. Figura 3.3) e (ii) recuperação, a partir do acesso ao dicionário WordReference®, de todas as possíveis traduções em inglês para a palavra em questão. As possíveis traduções são exibidas no "painel para exibição e seleção das traduções" (B) (cf. Figura 3.3). No caso de "morreram", o editor recuperou somente um equivalente de tradução, "*die*", o qual é exibido ao anotador humano como ilustrado na Figura 3.5. A Figura 3.5 ilustra ainda o painel de comentários ativado.

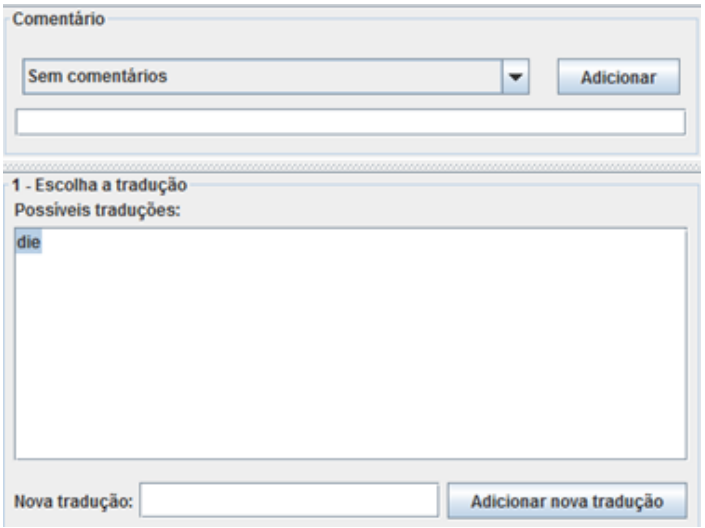

Figura 3.5. Tela de Comentários e Lista de Traduções.

Ao escolher a tradução "*die*", o editor NASP++ recupera automaticamente todos os *synsets* da WordNet.Pr que possuem esse verbo como um de seus elementos constitutivos. Por meio da Figura 3.6, observa-se que o editor recuperou ao menos 4 *synsets* compostos pelo verbo "*die*", sendo que cada um deles representa ou codifica um conceito distinto.

Ressalta-se que o editor NASP++ também recupera a glosa (isto é, definição informal do conceito) e os exemplos em uso de algumas palavras que constituem *synset*. O primeiro *synset* recuperado e exibido ao anotador, por exemplo, foi {*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}, o qual possui a glosa "*pass from physical life and lose all bodily attributes and functions necessary to sustain life*"[7].

Para "*die*" com o sentido em questão, tem-se a frase-exemplo "*She died from câncer*"; para "*perish*", tem-se "*The children perished in the fire*"*;* para "*go*", "*The patient went peacefully*" e, para "*kick the bucket*", a frase-exemplo é "*The old guy kicked the bucket at the age of 102*")[8]. Tanto a glosa quanto as frases-exemplo auxiliam a tarefa de identificar o *synset* que mais corretamente codifica o conceito subjacente ao verbo original em português. Dentre os *synsets* recuperados, cabe ao anotador escolher ou selecionar o que mais adequadamente representa o conceito subjacente ao verbo "morreram" no texto-fonte.

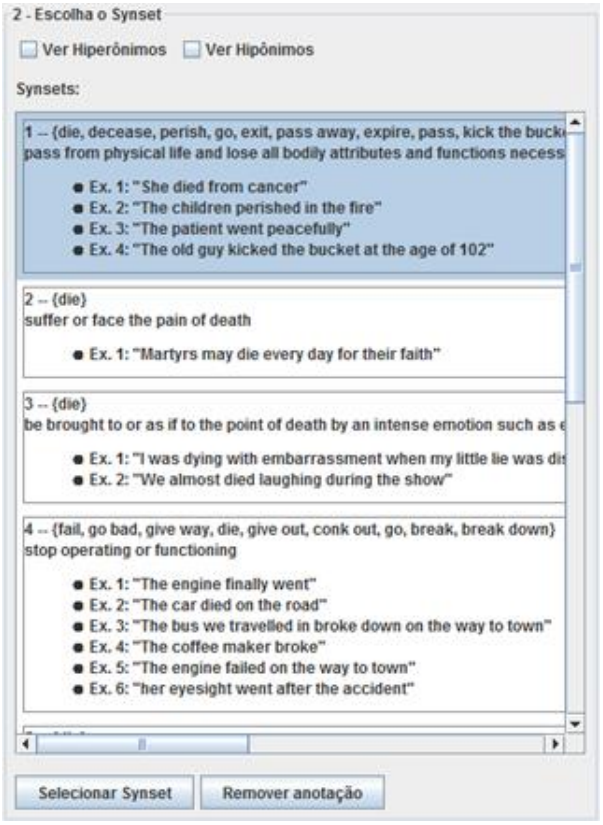

Figura 3.6. Tela de seleção do *synset.*

Caso os *synsets* constituídos pelo equivalente de tradução ("*die*"), as glosas e as frases-exemplo são sejam suficientes para se definir a representação mais adequada do sentido do verbo em português, o editor oferece a possibilidade de visualização dos hiperônimos (e/ou tropônimos) e hipônimos dos *synsets* inicialmente recuperados (que contêm "*die*" como elemento constitutivo), como ilustrado na Figura 3.7.

Nessa Figura, aliás, vê-se que, por exemplo, para o primeiro *synset* recuperado da WordNet.Pr {*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*}, o editor recuperou o hiperônimo direto {*change state, turn*} e 8 tropônimos, ou seja, 8 *synsets* que codificam diferentes formas ou maneiras de "morrer".

[7] Em português, "*passar da vida física e perder todos os atributos corporais e funções necessários para sustentar a vida*" (tradução nossa).
[8] Em português, "*Ela morreu de câncer*", "*As crianças morreram no fogo*", "*O paciente passou/morreu em paz*" e "*O velho morreu com 102 anos de idade*" (tradução nossa).

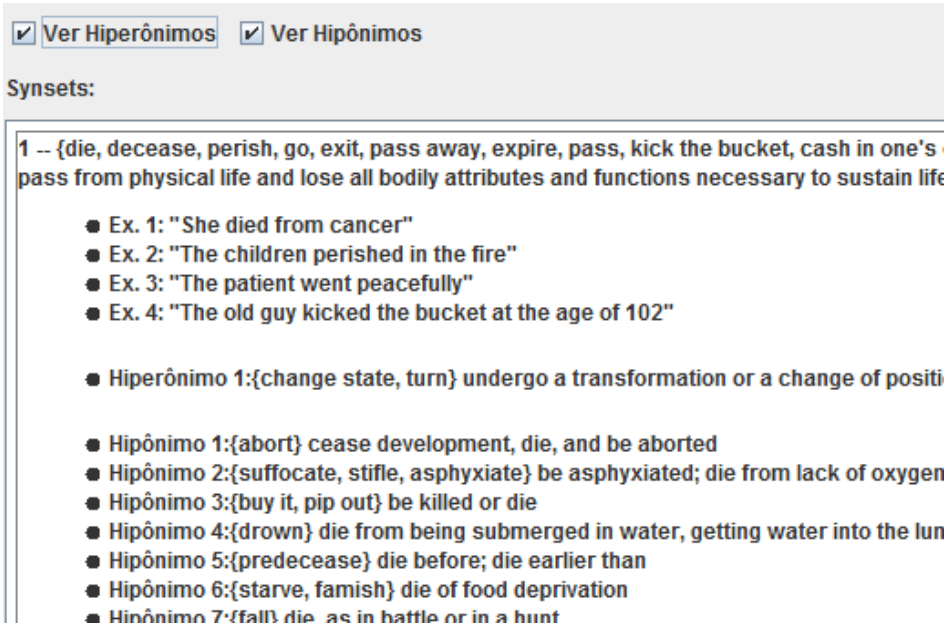

Figura 3.7. Tela de apresentação de hiperônimos e hipônimos.

Para selecionar um dos *synsets* recuperados, o qual será utilizado como rótulo semântico para a anotação da palavra em português, o usuário deve clicar no *synset* em questão, por exemplo, {*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, hoque, croak, snuff it*} e, na sequência, no botão "Selecionar *synset*", que consiste no passo "C" da metodologia de anotação semântica. Ao se clicar no botão "Selecionar *synset*" ou dar dois cliques no *synset* escolhido, o editor exibe uma janela de confirmação, como a ilustrada na Figura 3.8 (passo "D" da metodologia proposta na Figura 3.2). Diante de certeza sobre a escolha do *synset,* o usuário deve clicar no botão "Sim". Diante de dúvidas remanescentes, o usuário pode clicar em "Não" e retornar para a análise dos *synsets.*

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

Uma vez selecionado um *synset*, a palavra sob anotação semântica ("morreram") é destacada no campo "visualizador de textos" em "verde", como ilustrado na Figura 3.9. Esse destaque indica que à palavra em questão foi associado um rótulo semântico, no caso, um *synset.*

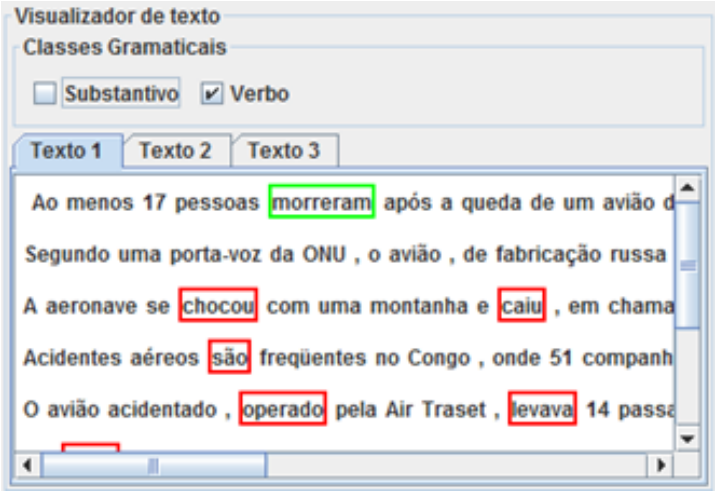

Figura 3.9. Visualizador de textos com o verbo "morrer" anotado.

Partindo-se do pressuposto de que as diversas ocorrências de uma palavra em um texto (ou em textos que abordam mesmo assunto) tendem a ter um mesmo sentido, ressalta-se que, uma vez que uma palavra *x* tenha sido anotada com um sentido *y*, todas as demais ocorrências de *x* também são pré-anotados pelo editor com *y*. No NASP++, as demais ocorrências de *x* pré-anotadas com *y* são destacadas de "amarelo". Na Figura 3.10, por exemplo, vê-se que outra ocorrência de "morreram" foi pré-anotada com o *synset* selecionado para a anotação da primeira ocorrência de "morreram". Ressalta-se aqui que a pré-anotação semântica é realizada para todas as ocorrências do verbo "morrer", independentemente de sua forma flexionada. Assim, caso ocorra "morreu", este também será pré-anotado. Ao anotador humano, cabe a tarefa de verificar se, de fato, o sentido/*synset* pré-anotado é pertinente para as diferentes ocorrências.

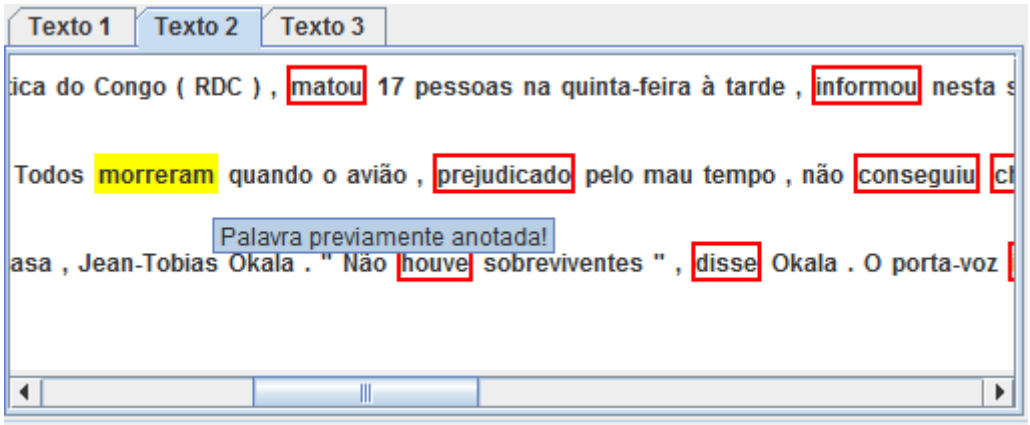

Figura 3.10. Visualizador de textos com o verbo "morrer" previamente anotado.

Após a anotação de todos os verbos pertinentes, a ferramenta permite salvar os textos-fonte anotados no formato de linguagem de marcação XML (do inglês, *Extensible Markup Language*), um dos mais utilizados para a tarefa de anotação de córpus. Na Figura 3.11, ilustra-se a anotação semântica codificada em um arquivo XML.

Nesse formato, a anotação é organizada em uma hierarquia de informações. No primeiro nível da hierarquia, tem-se 3 blocos que delimitam as informações relativas a: (i) os anotadores (indicados pela etiqueta XML *Anotadores*), (ii) a classe da palavra que pode ser anotada (verbo e nome) e a porcentagem de palavras dessa classe (expressa em formato de 0 a 1) (indicada pela etiqueta XML "*LimitesAnotacao*") e o (iii) arquivo do texto-fonte a ser anotado (indicado pela etiqueta XML "*Arquivos*").

```
<save>
  <Anotadores>
    <Anotador id="1">Erick</Anotador>
    <Anotador id="2">Fernando</Anotador>
  </Anotadores>
  <LimitesAnotacao>
    <Pos id="NOUN">1.0</Pos>
    <Pos id="VERB">1.0</Pos>
  </LimitesAnotacao>
  <Arquivos>
    <Texto name="D1_C1_Folha_04-08-2006_07h42.tagged" language="PORTUGUESE">
      <p number="0">
        ...
        <Token>
          <Valor>morreram</Valor>
          <Lema>morrer</Lema>
          <Etiqueta>VERB</Etiqueta>
          <MorphoTag>Verbos</MorphoTag>
          <MorphoTagPOS>Verbos</MorphoTagPOS>
          <Comentario content="" obs=""/>
          <Type>ANNOTATED</Type>
          <Traducoes traducao_manual="0">
            <Traducao selecionado="true">die</Traducao>
          </Traducoes>
          <Synsets>
            <Synset selecionado="true">358431</Synset>
            <Synset selecionado="false">2109818</Synset>
            <Synset selecionado="false">1784953</Synset>
            <Synset selecionado="false">434374</Synset>
            <Synset selecionado="false">1829475</Synset>
            <Synset selecionado="false">1785242</Synset>
            <Synset selecionado="false">1555034</Synset>
            <Synset selecionado="false">1074914</Synset>
            <Synset selecionado="false">538323</Synset>
            <Synset selecionado="false">354845</Synset>
            <Synset selecionado="false">224295</Synset>
          </Synsets>
        </Token>
        ...
      </p>
    </Texto>
  </Arquivos>
</save>
```

Figura 3.11. Arquivo de anotação em formato XML.

O bloco de informações encapsuladas entre as etiquetas <Arquivos></Arquivos> engloba: (i) o arquivo referente ao texto-fonte a ser anotado (p.ex.: <Texto name="D1_C1_Folha_04-08-2006_07h42.tagged"></Text>) e a língua na qual está escrito (language="PORTUGUESE") e (ii) a indicação de parágrafo ( ). Ademais, a codificação ilustrada na Figura 3.11 registra ainda que os parágrafos são compostos por sentenças e que cada sentença é composta por palavras ou *tokens* (*tag* <Token></Token>). Cada *token* que pode receber a anotação semântica possui uma lista de atributos, representados pelas seguintes etiquetas XML:

- Valor: codifica a palavra em si.
- Tag: codifica a anotação morfossintática advinda da ferramenta NASP++.
- MorphoTagPOS: codifica um mapeamento da anotação morfossintática feita pelo POS-*Tagger*, podendo ser: Verbos, Substantivos, Adjetivos, Advérbios e Outros.
- MorphoTag: por *default*, tem o mesmo valor que MorphoTagPOS, mas pode ser modificado; isso é feito pelo usuário quando existe um erro de anotação.
- Lema: codifica a forma canônica ou básica da palavra; no caso dos verbos, trata-se do infinitivo.
- Comentario: codifica os comentários sobre a palavra anotada, os quais podem ser (1) "Sem comentários", (2) "Não é um verbo, erro de anotação", (3) "Verbo auxiliar", (4) "É predicado complexo" ou (5) "Outros") (as opções 2, 3 e 4 são específicas para os verbos), ou o usuário pode ainda adicionar observações distintas das previstas pela ferramenta se assim achar pertinente.
- *Type*: codifica o estado de anotação da palavra.
  ANNOTATED: palavra anotada.
  VERB_NO_ANNOTATED: verbo não anotado.
  PREV_ANNOTATED: palavra previamente anotada.
  NOUN_NO_ANNOTATED: substantivo não anotado.
  NO_ANNOTATE: palavra não anotada (outras classes gramaticais).
- Traducoes: codifica as traduções oferecidas pela ferramenta ou adicionadas pelo usuário; o atributo "selecionado" da tradução escolhida recebe o valor "*true*".
- *Synsets*: codifica os *synsets* oferecidos pelas traduções propostas na NASP++; o atributo "selecionado" do *synset* escolhido recebe o valor "*true*".

Uma vez que todos os verbos de uma coleção de textos-fonte tenham sido anotados em nível semântico segundo a metodologia ora descrita e materializada no editor NASP++, essa ferramenta de anotação tem a funcionalidade de gerar uma estrutura conceitual a partir dos *synsets* utilizados na anotação da coleção em questão. A seguir, descreve-se como essa estrutura conceitual é gerada.

3.4 Geração de estruturas conceituais hierárquicas

Uma hipótese que subjaz a criação desta funcionalidade do NASP++ é a de que os conceitos subjacentes às palavras que ocorrem em um mesmo texto (ou coleção de textos sobre um mesmo assunto) tendem a ocupar posições próximas em uma estrutura conceitual. Para ilustrar essa hipótese, considera-se a anotação das palavras A e B nas coleções de textos C1 e C2. Em C1, tem-se que: (i) a palavra A ocorreu 3 vezes com o sentido A1 e (ii) a palavra B foi anotada 3 vezes com o sentido B1. Em C2, a palavra A foi anotada com o sentido A2 e a palavra B não ocorreu.

Na Figura 3.12, ilustram-se as 2 hierarquias conceituais geradas a partir da anotação de A e B em C1 e em C2. Como ilustrada, a hipótese é a de que os conceitos A1 e B1 são próximos, pois ocorreram em um mesmo texto ou coleção, ao passo que A2 é distante de A1 e B1, por ocorrer em outra coleção. Com base nessa hipótese, pode-se inferir que em um novo texto no qual as palavras A e B tenham ocorrido, sendo a palavra B anotada com o conceito B1, há uma probabilidade mais alta de que o conceito subjacente à palavra A seja A1 e não A2, devido à menor distância na estrutura conceitual entre A1 e B1.

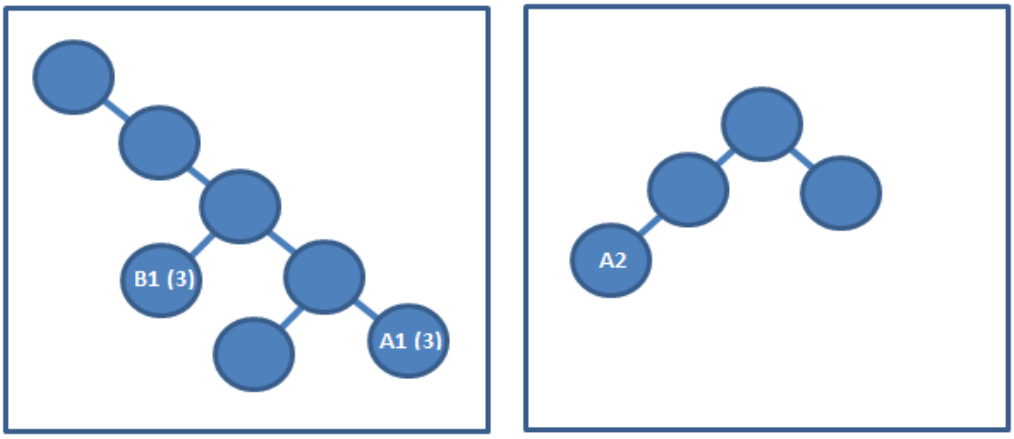

Figura 3.12. Hierarquias conceituais das coleções C1 e C2.

Tendo em vista que esse tipo de inferência pode ser relevante para a tarefa de DLS, desenvolveu-se no NASP++ a funcionalidade denominada "geração de ontologia".

A geração da estrutura conceitual é feita a partir dos conceitos/*synsets* utilizados na anotação das palavras de dada coleção de textos. Especificamente, para cada conceito/*synset x* selecionado para a anotação de uma palavra, a ferramenta automaticamente obtém: (i) os hipônimos imediatos (isto é, conceitos mais específicos) de *x*, (ii) os co-hipônimos de *x*, (iii) o hiperônimo (isto é, conceito mais genérico) imediato de *x*, (iv) os hiperônimos intermediários de *x* e (v) o *unique beginner* de *x*, ou seja, o hiperônimo mais genérico de *x* que inicia a hierarquia da qual faz parte.

Em outras palavras, o NASP++ recupera da WordNet.Pr toda a hierarquia conceitual da qual o conceito/*synset x* é parte integrante, gerando um grafo parcial interno. Esse processo é repetido a cada conceito/*synset* distinto selecionado para anotar uma palavra em português. Ao final, os gráficos parciais, referentes aos diferentes conceitos/*synsets*, são unificados em uma hierarquia final, a qual representa todos os conceitos/*synsets* utilizados na anotação dos textos-fonte de uma coleção. A seguir, são apresentadas as hierarquias geradas para os sentidos anotados de "morrer" e "matar" (Figura 3.13 e Figura 3.14) e a unificação das hierarquias parciais (Figura 3.15).

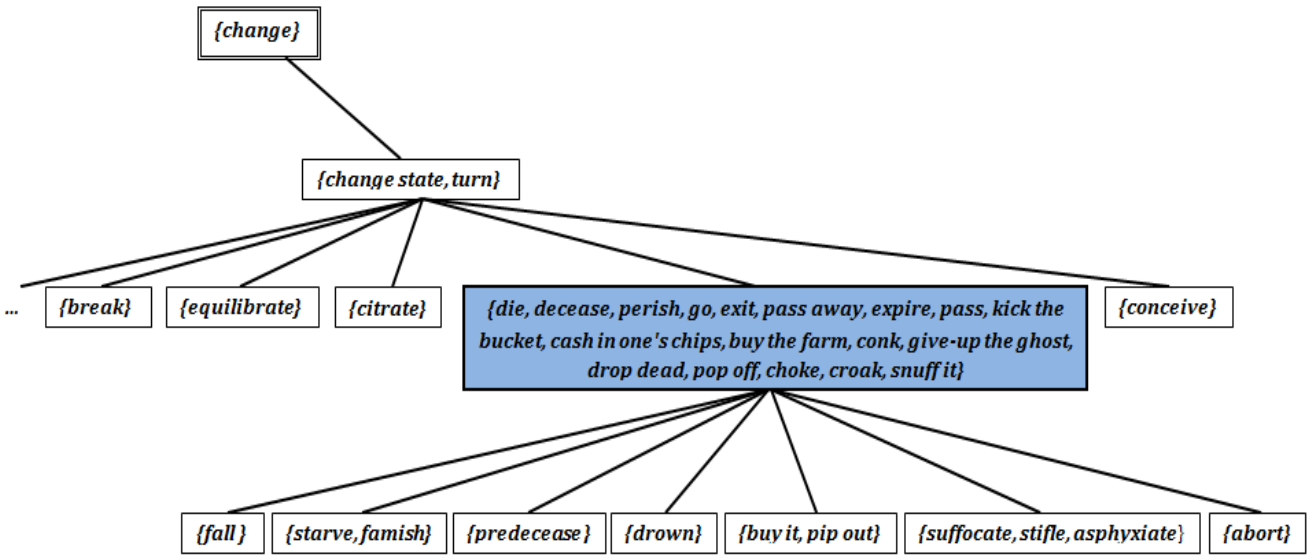

Figura 3.13. Hierarquia gerada para um *synset* do verbo "morrer".

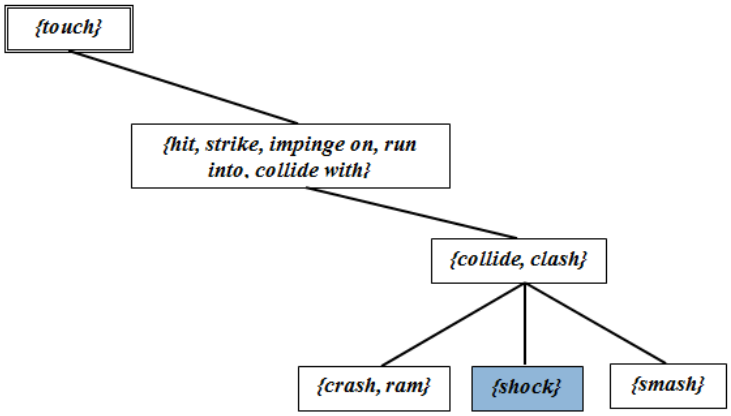

Figura 3.14. Hierarquia gerada para um *synset* do verbo "matar".

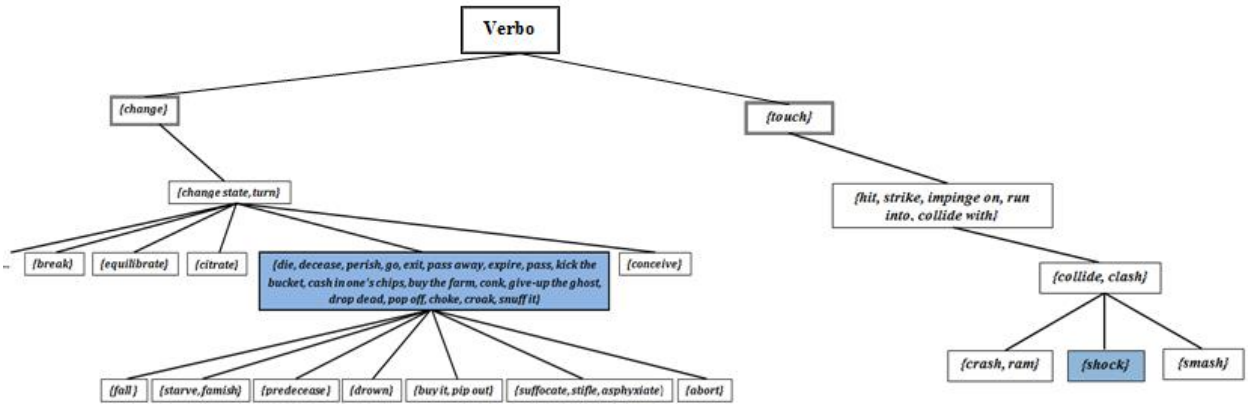

Figura 3.15. Exemplo de unificação de hierarquias parciais.

Para cada coleção do CSTNews, o NASP++ gerou automaticamente uma estrutura conceitual. Ademais, as estruturas conceitos das coleções foram unificadas, o que gerou uma estrutura final que representa os conceitos subjacentes aos verbos de todo o córpus CSTNews.

Na sequência, apresentam-se a avaliação da anotação e os seus resultados.

4. Avaliação e Resultados

4.1 Visão geral da anotação

A anotação foi realizada durante 7 semanas e meia, sendo que a primeira metade da primeira semana foi destinada ao treinamento e teste da ferramenta NASP++ pelos anotadores. Cada sessão de anotação durou aproximadamente 1 hora.

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, sendo cada grupo responsável por uma coleção do córpus. 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, garantindo que a tarefa não fosse tendenciosa. Com isso, buscou-se compartilhar o conhecimento dos anotadores, atingindo um padrão de anotação.

Na Tabela 4.1, apresenta-se a distribuição quantitativa da anotação dos verbos principais, verbos auxiliares, predicados complexos e dos erros de anotação.

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

Tabela 4.1. Estatísticas da anotação de verbos do córpus CSTNews.

Quanto à anotação dos verbos principais, salienta-se que foram anotados 5082 *tokens*[9], conforme evidencia a tabela. Dos 5082 *tokens*, há 844 *types*. Na anotação dos 844 *types*, foram selecionadas 787 traduções e 1047 *synsets* diferentes.

Na Figura 4.1, apresenta-se a quantidade de *synsets* distintos selecionados para cada uma das palavras (*types*) do córpus. No que se refere a isso, destaca-se:

a) a quantidade de *synsets* distintos para a anotação de um verbo variou entre 1 e 18;
b) apenas 8 casos foram anotados com mais de 10 *synsets* distintos;
c) a maioria dos casos foi anotada com apenas um *synset*, ou seja, na anotação de 508 *types*, apenas um *synset* foi selecionado pelos anotadores;
d) a média de *synsets* por *type* é 1.92 e o desvio padrão é 1.87.

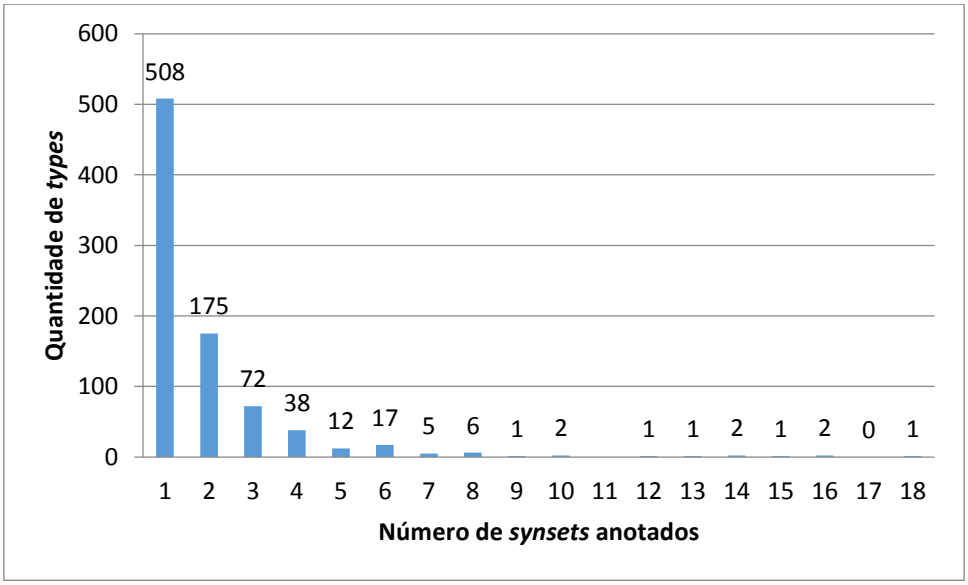

Figura 4.1. Distribuição de *synsets* por *type* no córpus.

Na Figura 4.2, apresenta-se a quantidade de *synsets* distintos selecionados para os *types* nas coleções. Com base na Figura 4.2, observa-se:

[9] *Token* é uma palavra corrida (do inglês, *running words*) e, por isso, cada ocorrência de uma palavra em um texto é um *token*. *Type* equivale a uma palavra distinta. Na sentença "*A Joana conversou com a Maria. Já a Patrícia conversou com a Joana, mas não falou com a Maria*", por exemplo, há 19 *tokens* e 10 *types*. Os 10 *types* são "a", "com", "conversou", "falou", "já", "joana", "maria", "mas", "não" e "patrícia".

a) os anotadores selecionaram entre 1 e 4 *synsets* distintos para a anotação dos *types* de uma mesma coleção;
b) os anotadores selecionaram apenas 1 *synset* para a maioria dos *types* (2671);
c) a média de *synsets* por *type* é 1.07 e o desvio padrão é 0.30.

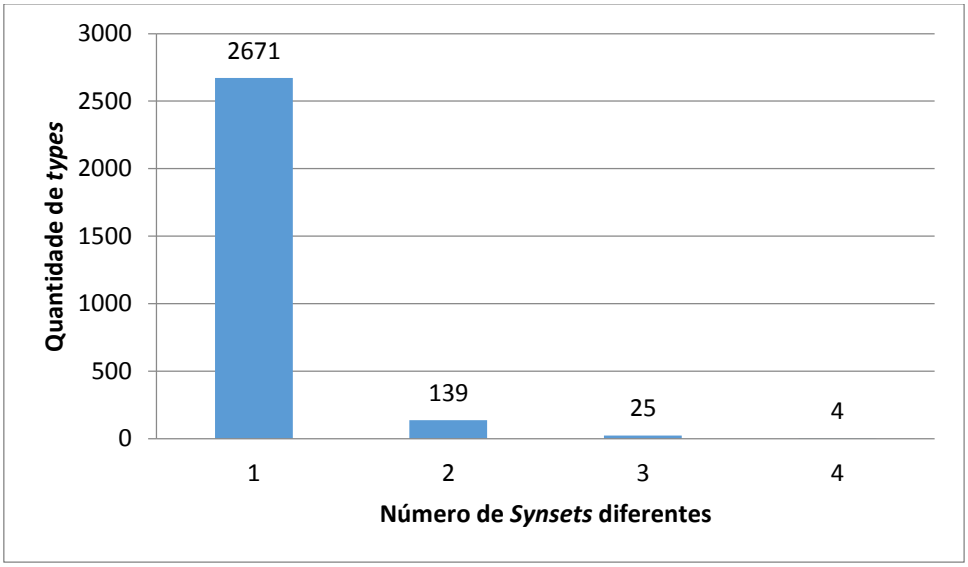

Figura 4.2. Distribuição de *synsets* por *type* nas coleções.

Na Tabela 4.2, tem-se a comparação das estatísticas referentes à anotação dos verbos com as obtidas por Nóbrega (2013) na anotação dos substantivos. Nessa comparação, os verbos possuem maior variação de sentidos, ou seja, são mais polissêmicos que os substantivos. Esse dado empírico, aliás, já havia sido mencionado por Miller et al. (1990).

Número máximo de *synsets* anotados por palavra	**Substantivos (Nóbrega, 2013)**	**Verbos**
No córpus	5	18
Em uma coleção	3	4

Tabela 4.2. Variação de número de *synsets* para substantivos e verbos.

A anotação semântica das palavras é uma tarefa que pressupõe a desambiguação lexical de sentido. Partindo do princípio de que, quanto mais polissêmica for uma palavra, mais difícil é a tarefa de desambiguação, buscou-se calcular a dificuldade da tarefa de anotação pela quantidade de conceitos que as palavras anotadas podiam representar.

Para calcular o grau de dificuldade da anotação de uma palavra/verbo, considerou-se o número de *synsets* nos quais essa palavra ocorre na base da WordNet.Pr (3.0), ou seja, o número de conceitos que essa palavra pode representar segundo a referida base léxico-conceitual. Vale ressaltar que todo *synset* é uma representação de um conceito e, por conseguinte, se uma palavra (*type*) *x* for elemento constitutivo de 10 *synsets* distintos, por exemplo, isso significa que ela pode expressar 10 conceitos distintos.

Na Figura 4.3, tem-se a distribuição dos *types* do córpus em função do número de conceitos/*synsets* possíveis armazenados na WordNet.Pr.

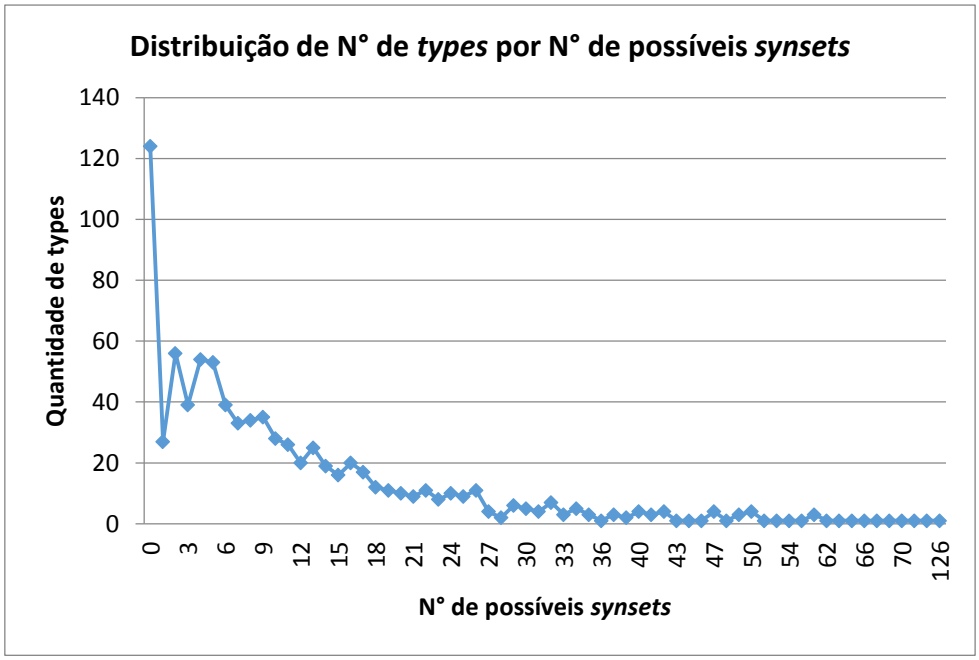

Figura 4.3. Distribuição do número de *types* por número de possíveis *synsets*.

Analisando a distribuição da Figura 4.3, ressalta-se que:

a) os verbos do córpus podem representar, em média, 12 conceitos distintos de acordo com os dados recuperados da WordNet.Pr;
b) 693 (82.11%) dos 844 verbos distintos (*types*) anotados podem representar/codificar 2 ou mais conceitos (*synsets*);
c) 276 verbos (32.7%) dos mesmos 844 anotados superam a média de 12 conceitos/*synsets*, os quais são considerados bastante polissêmicos.

Na Tabela 4.3, tem-se a comparação dos resultados aqui obtidos com os de Nóbrega (2013) na anotação dos substantivos do CSTNews. Dessa comparação, vê-se que a tarefa de anotação de sentidos dos verbos é mais difícil que a tarefa de anotação semântica dos substantivos, uma vez que os verbos são mais polissêmicos, podendo expressar mais conceitos que os substantivos (cf. Miller et al., 1990). No CSTNews, os verbos codificam em média 12 conceitos na WordNet.Pr, enquanto que os substantivos, 6. Além disso, a quantidade de verbos do CSTNews com 2 ou mais *synsets* equivale a aproximadamente 82% do total de verbos (*types*), enquanto que a quantidade de substantivos com 2 ou mais *synsets* equivale a 77% dos substantivos (*types*). Isso quer dizer que há mais verbos ambíguos do que substantivos ambíguos no CSTNews. Por outro lado, há mais substantivos altamente ambíguos (que têm mais *synsets* possíveis do que a média) do que verbos altamente ambíguos no córpus em questão.

	Substantivos (Nóbrega, 2013)	**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 substantivos e verbos.

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 indicava como predicado complexo a expressão "ficaram feridas" e, portanto, segundo as diretrizes de anotação, dever-se-ia anotar o verbo ficar com sentido da expressão. No entanto, durante a anotação, alguns anotadores anotaram a palavra "ficaram" como verbo auxiliar, gerando discordâncias.

Outra dificuldade da anotação semântica dos verbos foi a ausência de rótulos, representados por *synsets*, que adequadamente representassem certos conceitos expressos pelos verbos em português. Esses casos são 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 (p. ex.: "Não podemos mexer em nada."). 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.

Ademais, 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 um equivalente mais genérico e, na sequência, na seleção de um *synset* que adequadamente representeasse o conceito subjacente ao item lexical generalizado. Por exemplo, a expressão "tomar um frango" (p.ex.: "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 o equivalente genérico "*mistake*" ("errar") e, a partir

dele, 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.

4.2 Avaliação

Na avaliação da anotação, utilizou-se a medida Kappa (Carletta, 1996). Essa medida calcula o grau de concordância entre os anotadores em determinada tarefa, descontando-se a concordância ao acaso. Outras medidas de avaliação usadas, mais simples, também foram usadas. Essas, mostradas abaixo, não descontam a concordância ao acaso, mas computam de forma direta o número de concordâncias entre os anotadores.

- Concordância Total: número de vezes em que todos os anotadores concordaram em relação ao total de *tokens*;
- Concordância Parcial: número de vezes em que a metade ou a maioria dos anotadores concordou em relação ao total de instâncias;
- Concordância Nula: número de vezes em que a maioria dos anotadores não concordou.

Especificamente, avaliou-se a concordância entre os anotadores com relação a 3 parâmetros: (i) seleção da tradução, (ii) seleção do *synset* e (iii) seleção da tradução e do *synset*. A avaliação desses 3 parâmetros foi feita a partir da anotação de 3 coleções do CSTNews, as mesmas utilizadas por Nóbrega (2013) para a avaliação da anotação semântica dos substantivos, a saber: C15, C29 e C50. Na avaliação, 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 (%)**
Seleção da tradução	0.591	42.11	52.63	5.26
Seleção do *synset*	0.483	35.53	56.58	7.89
Seleção da tradução+*synset*	0.421	28.95	63.16	7.89

Tabela 4.4. Valores de concordância para a C15.

	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 da tradução+*synset*	0.485	32.28	60.63	7.09

Tabela 4.5. Valores de concordância para a C29.

	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 da tradução+*synset*	0.516	33.95	60.09	5.96

Tabela 4.6. Valores de concordância para a C50.

Quanto à medida Kappa nas Tabelas 4.4, 4.5 e 4.6, observa-se que os valores obtidos para cada um dos 3 parâmetros aumentaram a cada experimento de avaliação, que seguiu a sequência C15 → C19 → C50. Por exemplo, quanto ao parâmetro "seleção da tradução, obteve-se: (i) 0.591 na anotação da C15 (1ª concordância), (ii) 0.659 na anotação da C29 (2ª concordância) e (iii) 0.695 na anotação da C50 (3ª concordância) (ou seja, 0.591 < 0.659 < 0.695).

Uma possível justificativa para esse aumento pode ser a experiência adquirida pelos anotadores durante o processo de anotação semântica, 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 a temática das 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 aventar 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.

Na Tabela 4.7, apresentam-se os valores médios de concordância referentes às 3 coleções. Observa-se que o valor Kappa obtido para o parâmetro "seleção do *synset*" foi de 0.509, valor considerado aceitável no cenário da DLS. A concordância média do parâmetro "seleção da tradução" é superior à do parâmetro "seleção do *synset*". Isso, aliás, era esperado, pois a tradução é uma tarefa mais usual e direta que a desambiguação lexical de sentido. Finalmente, sobre o parâmetro "seleção da tradução + seleção do *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.

Quanto às outras medidas de concordância, destacam-se os valores altos obtidos para a "concordância parcial". Isso mostra que, mesmo com uma Kappa aceitável, os anotadores tiveram dúvidas na anotação. Algumas das causas podem ter sido a identificação de verbos no particípio 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. Outros valores a destacar na Tabela 4.7 são os referentes à "concordância nula", que foram baixos.

	Kappa	**Total (%)**	**Parcial (%)**	**Nula (%)**
Seleção da tradução	0.648	48.81	48.50	2.69
Seleção do *synset*	0.509	35.12	58.47	6.41
Seleção da tradução+*synset*	0.474	31.73	61.29	6.98

Tabela 4.7. Valores de concordância gerais.

Comparando com o trabalho de Nóbrega (2013), cujos resultados são apresentados na Tabela 4.8, vê-se que os valores de concordância para os substantivos são, na maioria, superiores aos dos verbos. Esse resultado também era esperado, já que os verbos são mais polissêmicos, o que dificulta a identificação do conceito/*synset* correspondente.

	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).

No processo de anotação ora descrito, os anotadores concordaram totalmente sobre a anotação semântica de uma série de palavras. Algumas delas, para ilustração, estão listadas em (4):

(4) "morreram", "reduzir", "hospitalizada, "investigar", "acabado", "considerou", "têm", "acreditar", "informou", "disseram", "anunciou", "cometidos", "abusados", "convencer" e "começa"

Em (5), listam-se sentenças retiradas do córpus nas quais algumas das palavras em (4) ocorrem, especificamente, "morreram", "hospitalizada" e "investigar":

(5) 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 em português expressam conceitos claros; por exemplo, na sentença em (5a), pode-se facilmente determinar o sentido do verbo "morrer", que é "perder todos os atributos e funções corporais para manter a vida";
- O verbo em português possui um equivalente de tradução direto; no caso, "*die*";

- Os vários conceitos que o equivalente de 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, dentre 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 ("passar 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 (5b), o verbo "hospitalizada" expressa apenas um conceito e, por isso, é elemento constitutivo de apenas um *synset* ({hospitalize, hospitalize}) na WordNet.Pr; em (5c), o verbo "investigar" pode expressar 2 conceitos, codificados na WordNet.Pr pelos *synsets* {investigate, inquire, enquire} e {investigate, look into}.

A seguir, em (6), mostra-se uma lista de palavras que obtiveram concordância nula:

(6) "localizado", "estimado", "fossem", "levada", "aceitamos", "registrada", "conseguiram", "surgirem", "somariam", "deixou", "entraram", "enfrentar", "entenderam", "haverá", "adiantaram", "tramitando", "levar", "daria", "assinalaram", "veio", "caminhe" e "aceitamos"

Apresentam-se, em (7), sentenças extraídas do córpus em que alguns dos verbos em (6) ocorrem, especialmente, "localizada", "adiantaram" e "assinalaram". Cada sentença é seguida pelos diferentes *synsets* selecionados pelos anotadores, os quais resultaram em "concordância nula".

(7)

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 em português expressam conceitos relativamente vagos, de difícil delimitação; em (7c), tem-se um exemplo paradigmático de verbo (no caso, "assinalar") cujo sentido ou conceito é de difícil delimitação;
- Os anotadores utilizaram equivalentes de tradução distintos, o que pode ser explicado pela dificuldade de se delimitar ou definir o conceito; por exemplo, para "localizada" em (7c), foram utilizados "*locate*" e "*localize*";
- A seleção de equivalentes de tradução distintos pode ter levado os anotadores a selecionar *synsets* diferentes; isso foi observa na anotação dos verbos em (7a), (7b) e (7c);
- 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.

5. Considerações finais

A anotação semântica das palavras de um córpus 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 dessa classe. Como consequência, os níveis de concordância entre os anotadores são relativamente baixos. Contudo, um córpus cujos verbos possuem anotação de sentido é 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.

No processo de anotação ora descrito, destaca-se que a utilização de um dicionário bilíngue para acessar os rótulos conceituais em inglês (os *synsets*) não prejudicou a concordância entre os anotadores.

Contudo, uma das limitações da anotação foi a ausência no dicionário bilíngue de equivalente de tradução adequados. Apesar de a ferramenta NASP++ sugerir equivalentes de tradução com na consulta a um dicionário bilíngue, observou-se que em muitos casos a ferramenta não sugeria traduções ou sugeria traduções inadequadas. A não indicação de possíveis traduções pelo NASP++ pode resultar do fato de que os verbos, quando pertencentes a domínios especializados (p.ex.: "pedalar" no domínio dos esportes), expressam conceitos não registrados nos dicionários bilíngues de língua geral, como o utilizado neste trabalho. A criação e inclusão de dicionários bilíngues específicos poderiam trazer benefícios à tarefa de anotação.

Outra limitação diz respeito às lacunas lexicais, que se caracterizam pela inexistência de uma palavra em uma língua *x* que represente um conceito lexicalizado em uma língua *y*. Assim, muitos verbos em português expressam conceitos que, em inglês, são codificados por elementos linguísticos de outro tipo, que não lexicais. Para tais verbos em português, não é possível identificar na WordNet.Pr um *synset* correspondente ao conceito expresso pelo verbo em português, uma vez que os conceitos na WordNet.Pr são codificados por conjuntos de unidades lexicais sinônimas (os *synsets*). Nesses casos, os verbos foram anotados com

conceitos generalizados, ou seja, com *synsets* que representam conceitos hiperônimos. Por exemplo, o conceito subjacente à expressão "tomar um frango" não tem representação na WordNet.Pr e, por isso, a essa expressão foi associado um conceito mais genérico, no caso, "errar".

Como foi mencionado, uma das dificuldades da anotação foi a identificação dos predicados complexos, verbos auxiliares e verbos no particípio. 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 verbos no particípio, métodos baseados em regras podem ser implementados para melhorar a ferramenta de anotação.

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.

Agradecimentos

Os autores agradecem à FAPESP, à CAPES, ao CNPq e à Samsung Eletrônica da Amazônia Ltda pelo apoio a este trabalho.

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. 166p.

Aleixo, P.; Pardo, T. A. S. (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.

Azeredo, J. C. (2000). *Fundamentos de Gramática do Português*. Jorge Zahar Editor. 283p.

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

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

Cardoso, P. C. F.; Maziero, E. G.; Jorge, M.; Seno, E. M.; Di Felippo, A.; Rino, L. H.; Nunes, M. G. V.; Pardo, T. A. S. (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*, pp. 88-105. Cuiabá, MT, Brasil.

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*, pp. 2238–2247. São Leopoldo, RS, Brasil.

Dolz, J.; Schneuwly, B. (2004). *Gêneros orais e escritos na escola*. Mercado de Letras. 278 p.

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

Fellbaum, C. (1998.) *WordNet: An Electronic Lexical Database*. 2. Ed. Cambridge (Mass.). MIT Press. 423p.

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

Gonçalo Oliveira, H.; Antón Perez, L.; Gomes, P. (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*, pp. 210-215. Groningen, The Netherlands.

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

Lage, N. (2002). *Estrutura da Notícia*. (5ª Ed.). São Paulo. 64p.

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*, pp. 24-26. New York, NY, USA. Association for Computing Machinery.

Machado, I. M.; de Alencar, R.; Campos, J.; De Oliveira, R.; Davis, C. A. (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*, pp. 39-41. New York, NY, USA.

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. 126p.

Plaza, L.; Diaz, A. (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*, pp. 97-105. Huelva, Espanha.

Ratnaparkhi, A. (1996). A Maximum Entropy Par t-Of-Speech Tagger, in *Proceedings of the Empirical Methods in Natural Language Processing Conference*, pp. 133-142, Pennsylvania, USA.

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

Ribeiro, R. (2003). Anotação Morfossintáctica Desambiguada do Português, Dissertação de Mestrado, Instituto Superior Técnico, Lisboa, Portugal. 98p.

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. 269p.

Travanca, T. (2013). Verb Sense Disambiguation, Dissertação de Mestrado, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal. 94p.