UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS HUMANAS
DEPARTAMENTO DE LINGUÍSTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM SEMIÓTICA E LINGUÍSTICA GERAL

**Desambiguação automática de substantivos
em *corpus* do português brasileiro**

VIVIANE SANTOS DA SILVA

DISSERTAÇÃO APRESENTADA AO PROGRAMA DE
PÓS-GRADUAÇÃO EM SEMIÓTICA E LINGUÍSTICA
GERAL DO DEPARTAMENTO DE LINGUÍSTICA DA
FACULDADE DE FILOSOFIA, LETRAS E CIÊNCIAS
HUMANAS DA UNIVERSIDADE DE SÃO PAULO PARA
OBTENÇÃO DO TÍTULO DE MESTRE EM LETRAS.

ORIENTADOR: PROF. DR. MARCOS LOPES

São Paulo
2016

Folha de Aprovação

SILVA, Viviane Santos. **Desambiguação automática de substantivos em *corpus* do português brasileiro**. Tese apresentada à Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo para obtenção do título de Mestre em Letras.

Aprovada em:

Banca Examinadora

Prof. Dr. Instituição:

Julgamento: Assinatura:

Prof. Dr. Instituição:

Julgamento: Assinatura:

Prof. Dra. Instituição:

Julgamento: Assinatura:

Prof. Dr. Instituição:

Julgamento: Assinatura:

Prof. Dr. Instituição:

Julgamento: Assinatura:

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Catálogo na Publicação
Serviço de Biblioteca e Documentação
Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo

S586d Silva, Viviane Santos da Silva
Desambiguação automática de substantivos em corpus
do português brasileiro / Viviane Santos da Silva
Silva ; orientador Marcos Lopes Lopes. - São Paulo,
2016.
125 f.

Dissertação (Mestrado)- Faculdade de Filosofia,
Letras e Ciências Humanas da Universidade de São
Paulo. Departamento de Linguística. Área de
concentração: Semiótica e Linguística Geral.

1. Linguística Computacional. 2. Desambiguação
Lexical Automática. 3. medidas de polissemia. 4.
clusterização de contextos de palavras. I. Lopes,
Marcos Lopes, orient. II. Título.

*Muitas palavras têm muitos significados;
o amor de vocês significa o mundo para mim.*

— Para Silvania, Ju e Luis.

Agradecimentos

Antes de tudo, agradeço à minha família, mãe e Ju, amores da minha vida, pela cumplicidade, pela força, pelo exemplo, pelo amor. Por sempre me apoiarem e me incentivarem a estudar e a fazer e seguir minhas escolhas. Agradeço também à minha segunda família, Lu, Sandra e Paulo, por sempre me acolherem com muito carinho e alegria em sua casa. Agradeço especialmente ao Lu, que me faz me sentir em casa onde quer que a gente esteja e que sempre consegue me fazer ver o lado bom da vida. Obrigada pelas discussões científicas, filosóficas, linguísticas, de games e nerddices, em todo café da manhã (e no restante do dia também).

Agradeço imensamente à Mara e ao Edson pelo incentivo que foi essencial para a minha formação acadêmica, pela enorme generosidade e carinho com que me trataram e pela oportunidade de ter uma formação sólida nos anos anteriores ao meu ingresso na USP. Não poderia também deixar de agradecer aos professores que me inspiraram e me incentivaram ao longo da minha trajetória acadêmica até aqui.

Estou certa de que não seria a mesma pessoa sem ter conhecido os amigos que fiz no curso de Ciências Moleculares, os melhores presentes de toda a convivência intensa e efervescente do curso. Aninha, Chico, Daniel, Furlan, Giulia, João, Júlia, Leo, Otto, Victor: muito obrigada pelas conversas, pelo apoio, pelas noites de pizza, pelas boas risadas, pela paixão pelo conhecimento sem fronteiras, pelas críticas e incentivos e por me ensinarem a enxergar o mundo acadêmico e a vida de formas tão mais ricas.

Ao Marcos Lopes, agradeço por tantos aprendizados, pela confiança e pelo incentivo à minha formação acadêmica ao longo desses anos desde a minha iniciação científica. Agradeço pelas disciplinas ministradas com maestria, estimulando profundas reflexões linguísticas não apenas nas discussões em sala de aula, mas também nas redações das resenhas críticas, sempre minuciosamente corrigidas e que me proporcionaram oportunidades valiosíssimas de desenvolvimento acadêmico. Acrescento que sou muito grata não apenas pelos aprendizados acadêmicos, mas pelos de vida também. Admiro a forma sempre muito acolhedora com que o Marcos trata as

pessoas ao seu redor e agradeço muito por sua disponibilidade, padrões de exigência e comprometimento como orientador.

Gostaria de agradecer ao professor Alexandre Adalardo, do curso de R ministrado no departamento de Ecologia do IB, pela oportunidade da monitoria e aprendizados do curso. Agradeço à Evani Viotti pelo exemplo acadêmico, pelas excelentes disciplinas ministradas, por seus incentivos à pesquisa e ao ensino de qualidade e pela simpatia e disponibilidade com que sempre me tratou. Agradeço aos professores Thomas Finbow e Jairo Nunes pelas oportunidades de ser monitora nas disciplinas de Elementos de Linguística, enriquecendo consideravelmente minha formação. Sou muito grata aos professores Marcelo Ferreira e Marcelo Finger pelas sugestões oferecidas por ocasião do exame de qualificação, as quais foram de grande valia para o desenvolvimento desta pesquisa.

Deixo meus agradecimentos também aos colegas do Grupo de Estudos em Linguística Computacional, por formarem um grupo comprometido e acolhedor — que está crescendo cada vez mais —, pelo interesse e pelo ânimo com que sempre trabalham e pelas discussões férteis que sempre surgem em nossas reuniões.

Agradeço aos colegas linguistas que tive o prazer de conhecer nesses anos, que sempre mostraram grande paixão pela Linguística e por suas pesquisas e que são exemplos de excelência acadêmica. Agradeço, especialmente, aos amigos que tornaram possível o curso de extensão do Departamento de Linguística, que ministramos neste semestre, voltado para professores da rede pública de ensino.

Ao Ben Hur Euzébio, Denise Cristiane dos Santos, Érica Flávia de Lima e Robson Dantas Vieira, obrigada pelo trabalho e paciência no auxílio com os trâmites da universidade. Agradeço, por fim, ao CNPq pelo financiamento de dois anos desta pesquisa.

Resumo

O fenômeno da ambiguidade lexical foi o tópico central desta pesquisa, especialmente no que diz respeito às relações entre acepções de formas gráficas ambíguas e aos padrões de distribuição de acepções de palavras polissêmicas na língua, isto é, de palavras cujas acepções são semanticamente relacionadas. Este trabalho situa-se como uma proposta de interface entre explorações computacionais da ambiguidade lexical, especificamente de processamento de linguagem natural, e investigações de cunho teórico sobre o fenômeno do significado lexical.

Partimos das noções de polissemia e de homonímia como correspondentes, respectivamente, ao caso de uma palavra com múltiplas acepções relacionadas e ao de duas (ou mais) palavras cujas formas gráficas coincidem, mas que apresentam acepções não relacionadas sincronicamente. Como objetivo último deste estudo, pretendia-se confirmar se as palavras mais polissêmicas teriam acepções menos uniformemente distribuídas no *corpus*, apresentando acepções predominantes, que ocorreriam com maior frequência.

Para analisar esses aspectos, implementamos um algoritmo de desambiguação lexical, uma versão adaptada do algoritmo de Lesk (Lesk, 1986; Jurafsky & Martin, 2000), escolhido com base nos recursos linguísticos disponíveis para o português. Tendo como hipótese a noção de que palavras mais frequentes na língua tenderiam a ser mais polissêmicas, selecionamos do *corpus* (Mac-Morpho) aquelas com maiores ocorrências. Considerando-se o interesse em palavras de conteúdo e em casos de ambiguidade mais estritamente em nível semântico, optamos por realizar os testes apresentados neste trabalho apenas para substantivos.

Os resultados obtidos com o algoritmo de desambiguação que implementamos superaram o método *baseline* baseado na heurística da acepção mais frequente: obtivemos 63% de acertos contra 50% do *baseline* para o total dos dados desambiguados. Esses resultados foram obtidos através do procedimento de desambiguação de pseudo-palavras (formadas ao acaso), utilizado em casos em que não se tem à disposição *corpora* semanticamente anotados. No entanto, em razão da dependência

de inventários fixos de acepções oriundos de dicionários, pesquisamos maneiras alternativas de categorizar as acepções de uma palavra.

Tomando como base o trabalho de Sproat & VanSanten (2001), implementamos um método que permite atribuir valores numéricos que atestam o quanto uma palavra se afastou da monossemia dentro de um determinado *corpus*. Essa medida, cunhada pelos autores do trabalho original como *índice de polissemia*, baseia-se no agrupamento de palavras co-ocorrentes à palavra-alvo da desambiguação de acordo com suas similaridades contextuais.

Propusemos, neste trabalho, o uso de uma segunda medida, mencionada pelos autores apenas como um exemplo das aplicações potenciais do método a serem exploradas: a clusterização de co-ocorrentes com base em similaridades de contextos de uso. Essa segunda medida é obtida de forma que se possa verificar a proximidade entre acepções e a quantidade de acepções que uma palavra exhibe no *corpus*.

Alguns aspectos apontados nos resultados indicam o potencial do método de clusterização: os agrupamentos de co-ocorrentes obtidos são ponderados, ressaltando os grupos mais proeminentes de vizinhos da palavra-alvo; o fato de que os agrupamentos aproximam-se uns dos outros por medidas de similaridade contextual, o que pode servir para distinguir tendências homonímicas ou polissêmicas. Como exemplo, temos os *clusters* obtidos para a palavra *produção*: um relativo à ideia de *produção literária* e outro relativo à de *produção agrícola*. Esses dois *clusters* apresentaram distanciamento considerável, situando-se na faixa do que seria considerado um caso de polissemia, e apresentaram ambos pesos significativos, isto é, foram compostos por palavras mais relevantes.

Identificamos três fatores principais que limitaram as análises a partir dos dados obtidos: o viés político-jornalístico do *corpus* que utilizamos (Mac-Morpho) e a necessidade de serem feitos mais testes variando os parâmetros de seleção de co-ocorrentes, uma vez que os parâmetros que utilizamos devem variar para outros *corpora* e, especialmente, pelo fato de termos realizados poucos testes para definir quais valores utilizaríamos para esses parâmetro, que são decisivos para a quantidade de palavras co-ocorrentes relevantes para os contextos de uso da palavra-alvo.

Considerando-se tanto as vantagens quanto as limitações que observamos a partir dos resultados da clusterização, planejamos delinear um método sincrônico (que prescinde da documentação histórica das palavras) e computacional que permita distinguir casos de polissemia e de homonímia de forma mais sistemática e abrangendo uma maior quantidade de dados. Entendemos que um método dessa natureza pode ser de grande valia para os estudos do significado no nível lexical, permitindo o

estabelecimento de um método objetivo e baseado em dados de uso da língua que vão além de exemplos pontuais.

Palavras-chave: Linguística Computacional, Desambiguação Lexical Automática, medidas de polissemia, clusterização de contextos de palavras.

Abstract

The phenomenon of lexical ambiguity was the central topic of this research, especially with regard to relations between meanings of ambiguous graphic forms, and to patterns of distribution of the meanings of polysemous words in the language, that is, of words whose meanings are semantically related. This work is set on the interface between computational explorations of lexical ambiguity, specifically natural language processing, and theoretical investigations on the nature of research on the lexical meaning phenomenon.

We assume the notions of polysemy and homonymy as corresponding, respectively, to the case of a word with multiple related meanings, and two (or more) words whose graphic forms coincide, but have unrelated meanings. The ultimate goal of this study was to confirm that the most polysemous words have meanings less evenly distributed in the corpus, with predominant meanings which occur more frequently.

To examine these aspects, we implemented a word sense disambiguation algorithm, an adapted version of Lesk algorithm (Lesk, 1986; Jurafsky & Martin, 2000), chosen on the basis of the availability of language resources in Portuguese. From the hypothesis that the most frequent words in the language tend to be more polysemic, we selected from the corpus (Mac-Morpho) those words with the highest number occurrences. Considering our interest in content words and in cases of ambiguity more strictly to the semantic level, we decided to conduct the tests presented in this research only for nouns.

The results obtained with the disambiguation algorithm implemented surpassed those of the baseline method based on the heuristics of the most frequent sense: we obtained 63% accuracy against 50% of baseline for all the disambiguated data. These results were obtained with the disambiguation procedure of pseudowords (formed at random), which used in cases where semantically annotated corpora are not available. However, due to the dependence of this disambiguation method on fixed inventories of meanings from dictionaries, we searched for alternative ways of categorizing the meanings of a word.

Based on the work of Sproat & VanSanten (2001), we implemented a method for assigning numerical values that indicate how much one word is away from monosemy within a certain *corpus*. This measure, named by the authors of the original work as *polysemy index*, groups co-occurring words of the target noun according to their contextual similarities.

We proposed in this paper the use of a second measure, mentioned by the authors as an example of the potential applications of the method to be explored: the clustering of the co-occurrent words based on their similarities of contexts of use. This second measurement is obtained so as to show the closeness of meanings and the amount of meanings that a word displays in the corpus.

Some aspects pointed out in the results indicate the potential of the clustering method: the obtained co-occurring clusters are weighted, highlighting the most prominent groups of neighbors of the target word; the fact that the clusters approximate from each other to each other on the basis of contextual similarity measures, which can be used to distinguish homonymic from polysemic trends. As an example, we have the clusters obtained for the word *production*, one referring to the idea of literary production, and the other referring to the notion of agricultural production. These two clusters exhibited considerable distance, standing in the range of what would be considered a case of polysemy, and both showed significant weights, that is, were composed of significant and distinctive words.

We identified three main factors that have limited the analysis of the data: the political-journalistic bias of the corpus we use (Mac-Morpho) and the need for further testing by varying the selection parameters of relevant cooccurrent words, since the parameters used shall vary for other corpora, and especially because of the fact that we conducted only a few tests to determine the values for these parameters, which are decisive for the amount of relevant co-occurring words for the target word.

Considering both the advantages and the limitations we observe from the results of the clusterization method, we plan to design a synchronous (which dispenses with the historical documentation of the words) and, computational method to distinguish cases of polysemy and homonymy more systematically and covering a larger amount of data. We understand that a method of this nature can be invaluable for studies of the meaning on the lexical level, allowing the establishment of an objective method based on language usage data and, that goes beyond specific examples.

Keywords: Computational Linguistics, Word Sense Disambiguation, polysemy index, word senses clusterization.

Sumário

Resumo	vi
Abstract	ix
Lista de Abreviações	xiii
Lista de Figuras	xiv
Lista de Tabelas	xvi
1 Introdução	1
2 Pressupostos teóricos	6
2.1 A ambiguidade no nível lexical	6
2.1.1 Polissemia e homonímia: questões de definição	7
2.1.2 Questões de terminologia	10
2.2 Recursos linguísticos	11
2.2.1 Corpora	12
2.2.2 Subsídios para o aprendizado de máquina	17
2.3 Desambiguação automática de sentidos	18
2.3.1 Método <i>baseline</i> para a desambiguação automática	19
2.3.2 Algoritmo de Lesk	20
2.3.3 Avaliação da desambiguação	22
2.4 Rediscutindo as bases da desambiguação	24
2.4.1 Extração de índices de polissemia	24
3 Metodologia	40
3.1 Ferramentas	40
3.2 Implementação da desambiguação automática	41
3.2.1 Seleção dos lemas	41
3.2.2 Versão adaptada do algoritmo de Lesk	41
3.2.3 Método para a avaliação da desambiguação	43
3.3 Algoritmo de extração dos índices de polissemia	45

3.3.1	Versão adaptada do algoritmo	45
4	Resultados e análises	49
4.1	Frequências de documento	49
4.2	Dados obtidos com as desambiguações	52
4.2.1	Desambiguação dos lemas mais frequentes	52
4.2.2	Desambiguação das pseudo-palavras	57
4.2.3	Discussão dos resultados da desambiguação	62
4.3	Medidas de polissemia obtidas	64
4.3.1	Análises dos pesos dos <i>clusters</i> : alguns exemplos	70
4.3.2	Discussão dos resultados das medidas de polissemia	83
4.4	Comparações entre os métodos implementados	84
4.4.1	Resultados da desambiguação de <i>trabalho</i>	85
4.4.2	Resultados das medidas de polissemia de <i>trabalho</i>	88
4.4.3	Análise comparativa dos métodos apresentados	91
5	Considerações finais	93
5.1	Vantagens e limitações da proposta	93
5.2	Trabalhos futuros	95
	Referências Bibliográficas	96
	Apêndices	101
A	Tabela de lemas desambiguados	102
B	Tabela de pseudo-palavras desambiguadas	106
C	Tabela das medidas de polissemia	108

Lista de Abreviações

PLN Processamento de Linguagem Natural
NLP *Natural Language Processing*

WSD *Word Sense Disambiguation*

ASL Análise Semântica Latente
LSA *Latent Semantic Analysis*

DVS Decomposição de Valor Singular
SVD *Singular Value Decomposition*

IA Inteligência Artificial

NLTK *Natural Language ToolKit*

OWN-PT *Open Wordnet do Português*

Lista de Figuras

2.1	Esquema da polissemia	8
2.2	Esquema da homonímia	9
2.3	Fluxograma de extração do índice de polissemia	30
2.6	Distribuição das palavras co-ocorrentes de <i>vela</i>	36
2.7	Distribuição ponderada das palavras co-ocorrentes de <i>vela</i>	37
2.8	Distribuição ponderada das palavras co-ocorrentes em <i>polar plot</i> . . .	38
2.9	Regressão isotônica e antitônica para os dados de <i>vela</i>	39
4.3	Confiança na desambiguação dos lemas com o Michaelis	53
4.4	Confiança na desambiguação dos lemas com os sinônimos	54
4.5	Comparação dos números de acepções e números de campos	55
4.6	Distribuição dos lemas por acepções e campos semânticos	56
4.7	Janelas contextuais por pesos das acepções	56
4.8	Janelas contextuais por pesos dos campos semânticos	57
4.9	Taxas de acertos com o Michaelis	58
4.10	Taxas de acertos com o Dicionário de Sinônimos	58
4.11	Valores de precisão com o Michaelis	59
4.12	Valores de precisão com o Dicionário de Sinônimos	59
4.13	Densidade de janelas por pesos das acepções	60
4.14	Densidade de janelas por pesos dos campos semânticos	60
4.15	Densidade de janelas com gabaritos dos pesos das acepções	61
4.16	Densidade de janelas com gabaritos dos pesos dos campos	61
4.17	Confiança na desambiguação das pseudo-palavras (Michaelis)	62
4.18	Confiança na desambiguação das pseudo-palavras (sinônimos)	63
4.19	Cobertura das desambiguações com o Michaelis	63
4.20	Cobertura das desambiguações com os sinônimos	63
4.21	Comparação das medidas de polissemia do <i>corpus</i>	65
4.22	Número de acepções no dicionário vs. frequência	66
4.23	Comparação entre números de <i>clusters</i> e números de acepções	67

4.24	Diferenças das medidas de polissemia do <i>corpus</i> (b)	68
4.25	Diferenças das medidas de polissemia do <i>corpus</i> (a)	69
4.26	Co-ocorrências de <i>mercado</i> em coord. cartesianas	71
4.27	Distribuição dos co-ocorrentes de <i>mercado</i>	72
4.28	Co-ocorrências de <i>mercado</i> em coord. polares	73
4.29	Distribuição dos co-ocorrentes de <i>mercado</i> em <i>polar plot</i>	74
4.30	Regressão isotônica e antitônica sobre os dados de <i>mercado</i>	75
4.31	Co-ocorrências de <i>produção</i> em coord. polares	76
4.32	Distribuição dos co-ocorrentes de <i>produção</i>	77
4.33	Distribuição dos co-ocorrentes de <i>produção</i> em <i>polar plot</i>	78
4.34	Co-ocorrências de <i>página</i> em coord. polares	80
4.35	Co-ocorrências de <i>página</i> em coord. polares com <i>tags</i>	80
4.36	Distribuição dos co-ocorrentes de <i>página</i> em <i>polar plot</i>	81
4.37	Regressão isotônica e antitônica sobre os dados de <i>página</i>	82
4.38	Co-ocorrências de <i>casa</i> em coord. polares	83
4.39	Distribuição dos co-ocorrentes de <i>casa</i> em <i>polar plot</i>	84
4.40	Distribuição das acepções nas desambiguações de <i>trabalho</i>	86
4.41	Pesos médios das acepções nas desambiguações de <i>trabalho</i>	87
4.42	Grau de confiança nas acepções selecionadas de <i>trabalho</i>	88
4.43	Co-ocorrências de <i>trabalho</i> em coord. polares com <i>tags</i>	89
4.44	Distribuição dos co-ocorrentes de <i>trabalho</i> em <i>polar plot</i>	90
4.45	Histograma dos co-ocorrentes de <i>trabalho</i>	90

Lista de Tabelas

2.1	Matriz de co-ocorrências relevantes de <i>vela</i>	34
2.2	Matriz P_w de similaridade das palavras relevantes	35
2.3	Tabela dos agrupamentos de co-ocorrentes de <i>vela</i>	37
4.1	Taxas de acertos das desambiguações das pseudo-palavras	57
4.2	Posições e pesos dos <i>clusters</i> de <i>mercado</i>	71
4.3	Posições e pesos dos <i>clusters</i> de <i>produção</i>	77
4.4	Posições e pesos dos <i>clusters</i> de <i>trabalho</i>	89
A.1	Tabela de lemas desambiguados	102
B.1	Tabela de pseudo-palavras	106
C.1	Tabela com as medidas de polissemia	108

1

Introdução

a task-independent set of word senses for a language is not a coherent concept. Word senses are simply undefined unless there is some underlying rationale for clustering, some context which classifies some distinctions as worth making and others as not.

— Adam Kilgarriff
1997

A tarefa de desambiguação automática de palavras (WSD — Word Sense Disambiguation) está entre os principais tópicos de pesquisa da área de Processamento de Linguagem Natural (NLP) (Jurafsky & Martin, 2000; Agirre & Edmonds, 2007; Ravin & Leacock, 2000). Apesar do acelerado desenvolvimento das pesquisas nessa área, a desambiguação ainda se apresenta como um desafio aos pesquisadores. Esse quadro fica evidente quando levamos em consideração a miríade de importantes tarefas de NLP que se apoiam em dados de desambiguação e que costumam atrair pesquisadores tanto em centros de pesquisa e universidades como no mundo corporativo.

Além do interesse prático pela tarefa de WSD, ressalta-se também o interesse gerado pelas possibilidades de análises linguísticas a partir de grandes montantes de dados. O estudo de propriedades linguísticas através de *corpora* mais representativos da língua mostra-se, cada vez mais, uma necessidade, sendo um meio importante de corroborar e testar teorias, assim como de suscitar novas questões linguísticas.

Apesar de a questão da significação no nível lexical estar no centro do problema de desambiguação, ainda há um distanciamento entre os estudos em WSD e áreas como

a lexicografia e a semântica lexical, tanto pelas circunstâncias em que os estudos de WSD se originaram, quanto pelo fato de os estudos teóricos do significado envolverem questões complexas, voltadas para a compreensão do fenômeno da significação e não para a resolução prática de tarefas de processamento da linguagem.

Este trabalho abrange os eixos computacional e semântico, com especial interesse no uso dos dados obtidos com algoritmos de resolução de ambiguidade lexical. Mais especificamente, com o objetivo de estudar as distribuições de aceções de termos ambíguos e suas relações semânticas, entendendo que as reflexões sobre aspectos semânticos teóricos possam também contribuir para os estudos computacionais na busca por soluções que conciliem refinamentos dos métodos de processamento e suas limitações práticas.

Dias-da Silva (2006a) avalia que ainda há pouca cooperação entre linguistas e desenvolvedores de projetos de NLP; segundo o pesquisador, os linguistas, muitas vezes, “não percebem a importância da sua contribuição para a construção do corpo de conhecimento interdisciplinar necessário para esse tipo particular de investigação”. Por outro lado, como também nota o pesquisador:

A demanda urgente de aplicativos e a limitação de recursos computacionais são também agravantes que têm impedido que os sistemas de NLP passem a incorporar sofisticações que exijam estudos mais aprofundados e que sejam consistentes com a Teoria Linguística. Essas são justificativas do ponto de vista comercial, mas não se sustentam do ponto de vista científico e tecnológico. (p. 109)

Concentrando-se no problema da desambiguação de sentidos lexicais, este projeto almeja trazer contribuições para o preenchimento da lacuna ainda persistente entre estudos teóricos e computacionais da linguagem. Outra motivação para a realização desta tarefa é a escassez de trabalhos de desambiguação para a língua portuguesa. A pesquisa de Nóbrega & Pardo (2012), que é relativamente recente, apresenta-se como uma das primeiras sobre desambiguação automática de uso geral que propõe implementação para o português brasileiro. Os autores ressaltam que há diversas possibilidades a serem exploradas em WSD para essa língua, tanto do ponto de vista dos recursos a serem explorados, quanto do viés da tarefa, isto é, da aplicação final para a qual a desambiguação é concebida e executada.

De acordo com Silva (2006), as investigações sobre os leques de sentidos que as palavras poderiam exibir remontam, pelo menos, aos estoicos e às *Categorias* de Aristóteles, passando por autores latinos, em seus estudos sobre a Dialética, a Retórica

e a Gramática, assim como pelos comentadores medievais de Aristóteles e pensadores renascentistas, sendo depois objeto de estudos da tradição histórico-filológica que buscava reconstruir uma proto-língua que explicasse as variações e semelhanças formais entre as diversas línguas, chegando à definição do termo *polissemia*, no sec. XIX, pelo semanticista Michel Bréal (1897) em seu *Essai de sémantique: Science des significations*. Silva (2006) comenta que:

Para a Linguística Computacional, em particular, a polissemia constitui efetivamente um desafio muito concreto, já que a identificação e a desambiguação automática dos sentidos das palavras no processamento natural da linguagem é ainda meta não alcançada. (p. 34)

Ainda se discute o quanto os algoritmos de desambiguação propostos na literatura apresentam desempenhos eficazes e o quanto já se avançou efetivamente no estado da arte em desambiguação (Agirre & Edmonds, 2007). Fato é que, apesar de resultados promissores, a tarefa ainda apresenta desafios teóricos e práticos importantes, especialmente, para línguas que dispõem de recursos linguísticos escassos para as tarefas de processamento da linguagem.

Tradicionalmente, a tarefa de desambiguação lexical automática nada mais é do que a determinação dos sentidos das palavras em contexto. Em outras palavras, a desambiguação automática é essencialmente uma tarefa de classificação em que os sentidos são as classes, o contexto fornece os indícios e, a cada ocorrência de uma palavra, assinala-se uma de suas classes possíveis com base nas evidências contextuais (Agirre & Edmonds, 2007). Essa abordagem tradicional compreende a tarefa de WSD como um processo explícito de desambiguação com respeito a um inventário fixo de acepções de palavras.

Na literatura, frequentemente se define polissemia como a associação de dois ou mais sentidos a uma mesma forma linguística, às vezes sem distingui-la da homonímia (seja por homografia, homofonia ou ambas), que corresponderia ao caso de formas coincidentes sem relação semântica (Lyons, 1977). No entanto, as definições de polissemia e de homonímia geralmente aceitas encerram complicações práticas e teóricas. O fenômeno da pluralidade de sentidos associados a uma mesma forma gráfica ou fonológica, apesar de ser frequentemente referido como *polissemia*, envolve matizes que são constantemente ignoradas. Particularmente, adota-se a noção dicionarizada de sentidos e ignoram-se as relações semânticas entre os diferentes usos de uma palavra. Nesse contexto, evidencia-se a necessidade de distinguirmos os fenômenos do significado no nível lexical, delimitando-os com o propósito de

compreender mais claramente as relações entre os significados que uma mesma forma gráfica pode assumir. Assumimos, nesta pesquisa, que *polissemia* e *homonímia* são tipos de ambiguidade lexical que diferem, essencialmente, quanto às relações entre as acepções: um termo polissêmico possui acepções relacionadas entre si e dois termos homônimos apenas compartilham a coincidência de suas formas gráficas (ou sonoras ou ambas), sem quaisquer relações sincrônicas entre suas acepções.

O objetivo geral deste trabalho reside em explorar técnicas de desambiguação automática de palavras (WSD) — para substantivos, mais especificamente — do português brasileiro com o intuito de utilizar os dados resultantes para investigar propriedades de substantivos polissêmicos da língua. A principal propriedade a ser analisada refere-se à distribuição das acepções desses substantivos em *corpus*, assim como às similaridades semânticas entre as acepções. A principal motivação para esta pesquisa residiu em investigar aspectos dos fenômenos da polissemia e da homonímia, a saber:

- a definição de medidas para quantificar acepções de uma palavra e suas relações a partir de um *corpus*
- a confrontação das medidas de polissemia com as frequências das respectivas palavras em *corpus*
- a verificação das distribuições das acepções das palavras polissêmicas em *corpus*, confrontando esses dados de distribuição com as frequências dos itens lexicais que essas acepções definem.

Propusemos, nesta pesquisa, que a desambiguação de sentidos de palavras e a extração de medidas de polissemia por métodos computacionais fossem testadas para o português e que pudessem ser estudadas e avaliadas sob a luz de teorias linguísticas. Em última instância, pretendeu-se utilizar os métodos computacionais de processamento de linguagem em benefício das análises do fenômeno linguístico da ambiguidade lexical, especialmente, nos casos de homonímia e de polissemia.

A exploração mencionada das técnicas de WSD consistiu em implementar algoritmos de desambiguação propostos na literatura, adaptando-os de acordo com a disponibilidade de dados para o português e de acordo com as propriedades linguísticas dos termos a serem desambiguados. Em outras palavras, através de avaliações dos algoritmos de WSD sob uma perspectiva linguística, buscamos chegar a um método que conseguisse gerar resultados satisfatórios de desambiguação a partir dos dados disponíveis, de forma que pudéssemos avaliar as distribuições das acepções

dos substantivos analisados. O interesse em utilizar a desambiguação para investigar as acepções e suas relações é o de contribuir com as análises linguísticas voltadas ao estudo da ambiguidade lexical, particularmente da polissemia e da homonímia, propondo critérios objetivos para diferenciá-las a partir de *corpora*.

Na primeira etapa deste trabalho, nos concentramos em fazer um levantamento bibliográfico dos principais algoritmos propostos na literatura e em selecionar para testes aqueles (idealmente, testaríamos mais de um algoritmo) que melhor se adaptassem aos recursos linguísticos disponíveis em português. Tendo em vista as limitações verificadas no algoritmo que implementamos (ver resultados na seção 2.3.3), fizemos uma nova revisão bibliográfica, mais extensa, buscando alternativas de recursos e de métodos de desambiguação que possibilitassem análises mais produtivas. A partir dessa pesquisa mais ampla, decidimos concentrar nossa investigação em métodos que permitissem computar medidas de polissemia a partir de dados de *corpora* e que não dependessem de medidas extraídas de dicionários.

O método de extração de medidas de polissemia que propomos neste trabalho baseiam-se em Sproat & VanSanthen (2001). As medidas extraídas são o *índice de polissemia* e a *clusterização de co-ocorrentes*. A segunda medida é de especial interesse, pois permite discretizar os contextos em que uma palavra-alvo ocorre a partir de medidas de similaridade entre palavras que co-ocorrem com a palavra-alvo. Essa discretização é feita de tal forma que possibilita estabelecer relações de similaridade pela proximidade entre os contextos clusterizados, além de permitir que se avaliem quais *clusters* são mais representativos no *corpus*.

A partir desse método de clusterização, esboçamos a proposta de um método que permita diferenciar casos de *polissemia* de casos de *homonímia*, em um recorte sincrônico, a partir de critérios objetivos baseados em dados de *corpora*. Os resultados obtidos nesta pesquisa indicam o potencial promissor dessa proposta, não apenas como instrumento para corroborar teorias, mas como ponto de partida para a implementação de métodos de desambiguação que não dependam de dicionários.

2

Pressupostos teóricos

“

Many words have many meanings. The dictionary tell us so. But what does that mean? How many meanings does a word have, and what grounds do we have for saying it is not more or less? [...] How might natural language processing computer systems deal with multiple meanings, or novel meanings?

— Adam Kilgarriff
Polysemy, 1992

Neste capítulo, discutiremos as bases teóricas que sustentam as propostas deste trabalho. Inicialmente, apresentaremos discussões relativas à noção de acepções de uma palavra, explorando o problema da ambiguidade lexical, sob a ótica dos estudos da Semântica Lexical. Em seguida, apresentaremos essas questões no âmbito dos trabalhos sobre desambiguação em Linguística Computacional.

2.1 A ambiguidade no nível lexical

Apesar da quantidade cada vez maior de trabalhos aplicados voltados para a questão da ambiguidade no nível lexical, com soluções que têm trazido resultados promissores (Agirre & Edmonds, 2007), frequentemente se negligenciam nesses estudos discussões fundamentais sobre a semântica dessas ambiguidades, como Kilgarriff (1997) discute amplamente. Não raras vezes, a ambiguidade lexical é tratada de maneira restritiva, abrangendo somente uma parte do fenômeno, genericamente

referida como *polissemia*, e entendida como *a atribuição de mais de um significado para uma mesma forma linguística*.

Entre outros motivos, isso se deu em razão da própria origem da palavra *polissemia*, cunhada por Bréal (1897) com base nos termos gregos πολύζ e σημεῖον¹, *polis* e *semeion*, respectivamente. Todavia, apesar de a noção de base etimológica da palavra *polissemia* poder explicitar parte do fenômeno de ambiguidade lexical, é importante ressaltar que, desde Bréal, desenvolveu-se toda uma tradição de estudos sobre a polissemia e a ambiguidade no âmbito da Semântica Lexical (Geeraerts, 2010). O termo *polissemia* revestiu novos significados ao longo do tempo, adquirindo, inclusive, definições conflitantes entre diferentes teorias (Silva, 2010). Portanto, dentro do panorama de estudos sobre desambiguação, torna-se indispensável que as pressuposições adotadas sobre a ambiguidade sejam apresentadas de forma clara e bem delimitada.

A tarefa computacional de desambiguação de sentidos de palavras (*Word Sense Disambiguation* — WSD) consiste em resolver o problema da ambiguidade gerada no nível lexical. Torna-se necessário, então, circunscrever toda sorte de fenômenos observados dentro desse contexto lexical para que se entenda e se defina adequadamente a tarefa de desambiguação. Por essa razão, discutiremos a terminologia associada à ambiguidade lexical procurando entender seus tipos de realizações e as formas como a literatura as têm descrito. Pretende-se, dessa maneira, refletir sobre o desempenho e as limitações dos métodos de desambiguação baseados em dicionários e discutir as vantagens e as limitações de se considerar refinamentos na definição de ambiguidade lexical para a tarefa de desambiguação.

2.1.1 Polissemia e homonímia: questões de definição

A definição de polissemia figura como um desafio importante tanto do ponto de vista dos estudos teóricos, especialmente para a Semântica Lexical, quanto do ponto de vista dos estudos aplicados, como se pode observar com as pesquisas sobre desambiguação de sentidos dentro da Linguística Computacional. No primeiro caso, a própria noção de significado está longe de ser consensual, o que torna a polissemia um problema de difícil resolução teórica, pois sua compreensão é indissociável do problema de definição do que seriam os significados de uma palavra. No segundo caso, a definição adotada direcionará os potenciais de aplicação do método.

Dias-da Silva (1996) propõe que a ambiguidade lexical seja classificada em duas formas, referidas como categorial e semântica. Consideremos os seguintes exemplos:

¹Em seu texto, Bréal traduz o termo do grego como *numerosas significações*.

- (2.1) a. A *casa* é bonita.
 b. Ela *casa* amanhã.
- (2.2) a. Fui ao *banco* sacar dinheiro.
 b. Meu *banco* foi à falência.
 c. Sentaram-se no *banco* do jardim.

As sentenças em 2.1 exemplificam o caso de ambiguidade lexical categorial, em que as formas gráficas coincidem, mas suas instanciações correspondem a classes gramaticais distintas. Os exemplos em 2.2 são, por sua vez, casos de ambiguidade estritamente semântica. Estes podem ser subclassificados de acordo com as relações entre os sentidos da palavra *banco*: os exemplos 2.2a e 2.2b estariam em relação polissêmica entre si, já que suas acepções estão relacionados; por outro lado, 2.2a e 2.2b estão em relação homonímica com respeito à acepção de *banco* na sentença 2.2c, pois não estabelecem relação semântica com esta acepção.

Entende-se, portanto, que a homonímia abrange casos em que as palavras tenham formas gráficas coincidentes, pertençam à mesma classe gramatical, mas apresentem sentidos não relacionados. A polissemia difere da homonímia na medida em que há relação entre os sentidos: trata-se, então, de uma única forma gráfica à qual estão conectadas múltiplas acepções relacionadas. Procuramos sintetizar essas noções nos esquemas 2.1 e 2.2.

Figura 2.1: Esquema da polissemia

O esquema 2.1, da polissemia, destaca as relações entre os sentidos atribuídos à face do significado de um determinado significante. A noção de significado indicada pelo esquema remonta às teorias saussureanas de significante e significado como as duas faces abstratas do signo linguístico. No esquema 2.2, da homonímia, enfatiza-se a relação entre os significantes, pela coincidência da forma, e não entre os sentidos.

A importância de se fazer essas distinções entre polissemia e homonímia reside no fato de que o método de os métodos de desambiguação. Termos homônimos apresentam contextos de usos muito mais díspares do que se observa dentre acepções de

Figura 2.2: Esquema da homonímia.

um termo polissêmico. Consequentemente, a desambiguação de termos homônimos pode ser mais facilmente realizada em relação a de termos polissêmicos.

Os critérios para diferenciar casos de homonímia e de polissemia são diversos e apresentam, cada qual, méritos e limitações. Segundo Zavaglia (2003), para definir a homonímia, os autores oscilam entre critérios diacrônicos e sincrônicos, critérios sintáticos e morfológicos, variações sociolinguísticas, entre outros.

A identificação da homonímia pelo viés diacrônico baseia-se na etimologia. Como exemplo, pode-se mencionar o caso da palavra *manga*: de acordo com o dicionário Michaelis, a acepção relacionada ao fruto da mangueira é de origem malaiala; a acepção correspondente à peça do vestuário, por sua vez, remonta ao latim. Temos, portanto, que a convergência das formas gráficas se deu ao acaso no português brasileiro.

No entanto, segundo Lyons (1977), há duas críticas principais a esse critério: a primeira diz respeito aos níveis de reconstrução da etimologia e a segunda diz respeito ao recorte sincrônico que os falantes fazem da língua. O autor menciona o exemplo da forma gráfica *port* do inglês, tendo o lexema *port*₁ o sentido de *porto* (derivado do latim *portus*) e o lexema *port*₂, o sentido de *vinho (do Porto)*, sendo esta uma derivação mais recente na língua e originada do nome da cidade portuguesa (Porto). No entanto, como esclarece o texto, o nome da cidade originou-se do mesmo lexema latino que derivou *port*₁². Nesse caso, seriam esses lexemas exemplos de homonímia ou de polissemia? Sobre a questão da reconstrução, comenta Lyons (1977):

Dizer que *port*₁ e *port*₂ são etimologicamente relacionados depende, portanto, de quão longe estamos dispostos a ir, quando temos a evidência,

²Um exemplo do português que transgride a divisão comum entre polissemia e homonímia é o da palavra *cabo*. Os lexemas de *cabo*, de acidente geográfico e de posto militar, têm a mesma origem, mas não são mais identificados atualmente pelos falantes como relacionados.

para traçar a história das palavras. O critério de relação etimológica não é, dessa forma, tão direto quanto pode parecer à primeira vista. (p. 551)

Para a linguística sincrônica, o critério etimológico seria irrelevante: mesmo que haja relação histórica entre dois sentidos de uma palavra, sincronicamente, o falante pode não estabelecer relação entre eles. Lyons (1977) comenta sobre a distinção entre polissemia e homonímia feita de acordo com outro critério:

O segundo critério principal que é tradicionalmente levantado por linguistas e lexicógrafos para a distinção entre homonímia e polissemia é o de sentidos relacionados vs. não-relacionados [...] pode-se considerar que essa seja a única consideração sincronicamente relevante. (p. 551)

De acordo com o que Silva (2010) discute, a definição de polissemia está relacionada a duas questões fundamentais: por um lado, a da discretização de sentidos (acepções) de uma palavra e, por outro, a das relações entre os diferentes sentidos. Entretanto, além do problema inicial de como discretizar acepções, o critério de relação semântica entre as acepções é difícil de ser aplicado e traz consigo o problema inerente da diferenciação sincrônica entre os significados. Mesmo que tenhamos à disposição meios de resolver a distinção entre polissemia e homonímia em termos da diacronia, restaria a questão de como os falantes de fato reconhecem as relações entre os sentidos. De que forma podemos afirmar quando se tratam de sentidos relacionados ou não, sincronicamente?

Propomos, para tratar do aspecto das relações sincrônicas entre sentidos, o uso de métodos que explorem informações contextuais dos itens ambíguos e que possam auxiliar na distinção entre polissemia e homonímia a partir de dados de *corpus* e com métodos objetivos. Exploramos, para tanto, outro aspecto importante na distinção entre polissemia e homonímia: os contextos de uso das palavras, de forma que as acepções seriam distinguidas através de pistas sintagmáticas.

2.1.2 Questões de terminologia

Neste trabalho, entenderemos a noção de *palavra* com base na separação feita na escrita, tendo em consideração que essa escolha é praticável no português, mas não pode ser estendida a todas as outras línguas, não apenas por diferenças nos sistemas de escrita, mas também pelas características morfológicas e semânticas da língua portuguesa. Tomaremos os termos *lexema* e *item lexical* como sinônimos de *palavra*.

Ressaltamos que o termo *lema*, em particular, é utilizado para ressaltar que nos referimos à forma base de uma palavra, sem quaisquer marcas flexionais ou

derivacionais, como a encontramos ao fazer uma consulta em um dicionários. O termo *token*, por oposição, refere-se a palavra instanciada, em uso.

Campos semântico

A questão do campo semântico é entendida nesta pesquisa como uma abstração conceitual que abarcaria acepções, seja de uma palavra ou de palavras sinônimas, de acordo com uma noção comum a essas acepções. Para Baldinger (1966), campo semântico (ou ‘campo semasiológico’ ou, ainda, ‘campo das significações’) é o conjunto das múltiplas significações associadas a uma palavra, considerando-se que elas “têm relações entre si” e, mais que isso, manifestam um *pivô*³ em torno do qual todas se reúnem.

Esses conceitos serão especialmente importantes ao tratarmos das entradas fornecidas pelo Dicionário de Sinônimos e de sua utilização para a tarefa de desambiguação que implementamos. Entendemos que cada entrada desse dicionário agrupa sinônimos de acordo com campos semânticos comuns, como se observa no exemplo a seguir para a palavra *mercado*:

1. **Atividade de compra e venda:** negócio, mercancia, mercadejo, tráfico, comércio.
2. **Cidade importante em comércio:** empório, entreposto, metrópole.
3. **Lugar público para comércio:** quitanda, feira, mercearia, armazém, bazar, loja, praça.

Cada entrada introduz um campo semântico possível para a palavra *mercado*, com as especificações que antecederam a lista de sinônimos desempenhando a função de pivôs dos seus respectivos campos.

2.2 Recursos linguísticos

Neste capítulo, apresentamos mais detalhadamente os recursos linguísticos pesquisados e os selecionados, os algoritmos que implementamos, assim como as justificativas metodológicas para a seleção dos materiais utilizados. Iniciaremos apresentando os *corpora* e as ferramentas computacionais que utilizamos para a implementação do algoritmo de Lesk simplificado e do algoritmo de extração de índices de polissemia e de clusterização.

³Referido como *noeud de significations*, no original

2.2.1 Corpora

A questão da disponibilidade de recursos é um tópico central em quaisquer estudos que envolvam análises linguísticas a partir de *corpus*, não apenas porque o gênero de texto do *corpus* pode influenciar os resultados, mas também porque variações em extensão dos textos, diversidade de autores e de temas, quantidade de locutores, modalidades (escrita ou falada), níveis de formalismo e outras variações sociolinguísticas, formato dos arquivos, anotações, entre outros fatores, provavelmente vão gerar diferenças substanciais nos resultados. Por essas razões, é imperativo adaptar os objetivos de um trabalho aos recursos disponíveis.

No caso da tarefa de WSD, um dos passos fundamentais é a seleção do *corpus* adequado aos objetivos da tarefa: não havendo recursos para a aplicação inicialmente planejada, uma alternativa é lançar mão dos recursos disponíveis, explorando-os com algoritmos adaptados ou combinando recursos linguísticos diversos. Para métodos de desambiguação que contam com algum nível de supervisão, é importante selecionar não apenas o *corpus* bruto em que a tarefa será aplicada, mas também o recurso em que a desambiguação de apoiará, seja ele um dicionário ou um *corpus* com etiquetas semânticas que sirvam de conjunto de dados para o treinamento. E a tarefa torna-se mais desafiadora nos casos de línguas que dispõem de recursos esparsos e ainda limitados como é o caso do português.

Além das demandas mais particulares da tarefa de desambiguação, há também outros critérios que envolvem questões mais gerais do trabalho com *corpora*: para que os resultados sejam coerentes com os objetivos da aplicação, o desejável é que se tenha um *corpus* representativo (da língua em geral ou de domínios específicos, a depender dos interesses do estudo), o que envolve aspectos como a extensão do *corpus* e a sua abrangência de domínios, que são importantes, mas difíceis de quantificar. É desejável também que o *corpus* seja acessível, tanto do ponto de vista da obtenção dos dados, como da organização dos dados disponibilizados.

Investigando a disponibilidade dos *corpora* do português, incluindo os que registram dados do português europeu mesclados aos dados do português brasileiro, verifica-se, segundo comenta Tang (2012), que muitos são de acesso restrito (Sardinha, 2000) ou pagos (Davies & Ferreira, 2006), predominantemente baseados em registros escritos (como se pode verificar na página da Linguateca⁴) ou pouco representativos do português brasileiro em geral (Raso & Mello, 2012), seja por suas dimensões, seja pela abrangência (de tópico e dialetal) dos registros. Motivado por essas questões, Tang iniciou a construção do *corpus* Subtlex-PT-BR, baseado no projeto

⁴http://www.linguateca.pt/corpora_info.html

Subtlex de *corpus* baseados em legendas, concebido por New, Brysbaert, Veronis e Pallier (New *et al.*, 2007) para o francês e feito, depois, para o holandês (Keuleers *et al.*, 2010) e para o inglês (van Heuven *et al.*, 2014), entre outras línguas.

Dentre os *corpora* investigados, cogitamos a compilação de um *corpus* de legendas de filmes e seriados baseada no Subtlex-PT-BR⁵. No entanto, optou-se por concentrar os testes em um *corpus* de extensão relativamente pequena, o *corpus* Mac-Morpho de textos jornalísticos. A escolha desse *corpus* foi feita por três razões principais: o Mac-Morpho tem etiquetas morfológicas que facilitam a identificação de substantivos; está facilmente acessível através da plataforma NLTK⁶ (Natural Language ToolKit) e foi compilado exclusivamente com textos em português brasileiro. A seleção dos *corpora* para este trabalho foi, de acordo com as possibilidades, calcada nas questões mencionadas; porém, procuramos priorizar a acessibilidade dos dados — tentando garantir, entre outros aspectos, maior factibilidade do trabalho e a replicabilidade do mesmo.

A seguir, encontram-se breves descrições e comentários sobre os principais *corpora* pesquisados para este trabalho. Além dos *corpora* textuais, encontram-se também as descrições dos dicionários empregados nas desambiguações.

Mac-Morpho

Componente do projeto Lacio-Web de compilação de *corpora* em português brasileiro e implementação de ferramentas para análises linguísticas, o *corpus* Mac-Morpho é a porção do padrão-ouro do *corpus* de referência Lacio-Ref, com etiquetas morfossintáticas manualmente validadas (Aluísio *et al.*, 2004). Abaixo, encontra-se a descrição geral do *corpus*:

Tipo: Escrito, jornalístico.

Época da produção dos dados: 1994.

Região: São Paulo.

Tamanho: 1.167.183⁷ palavras (Aluísio *et al.*, 2004), distribuídas em dez cadernos do jornal *Folha de S. Paulo*: Mais!, Agrofolha, Esporte, Informática, Dinheiro, Brasil, Cotidiano, Ilustrada, Folha Ciência, Mundo.

Acessibilidade: O *corpus* pode ser integralmente acessado pelo NLTK.

⁵o *corpus* compilado por Tang foi disponibilizado apenas no formato de listas a partir do texto pré-processado. Apesar disso, a metodologia de extração do *corpus* foi documentada e publicada de forma detalhada, além dos dados.

⁶http://www.nltk.org/howto/portuguese_en.html

⁷O número chega a 1.170.095 de acordo com o NLTK.

Etiquetagem: Foi anotado através do *parser* PALAVRAS (Bick, 2000) e, então, revisado manualmente. Segundo Fonseca & Rosa (2013), o Mac-Morpho seria o maior *corpus* manualmente anotado em português brasileiro.

Sites: <http://www.nilc.icmc.usp.br/nilc/projects/lacio-web.htm>,
www.nltk.org/howto/portuguese_en.html.

Subtlex-PT-BR

De acordo com o que foi comentado, a compilação de *corpora* impõe importantes desafios. No entanto, quando se trata de *corpora* orais, as dificuldades tendem a ser ainda maiores, demandando, em muitos casos, a formação de uma equipe de pesquisadores dedicados à tarefa (elaborando o *corpus* e fazendo as transcrições dos dados de áudio), além de voluntários para a produção dos dados, assim como equipamentos de gravação específicos (e, por vezes, caros) e a elaboração de procedimentos adequados, que possibilitem aos voluntários produções mais espontâneas. E a tarefa se torna tanto mais complexa e custosa quanto mais se quiser que o *corpus* seja representativo e bem estruturado.

Diante das limitações apresentadas, uma das alternativas que surgiram foram os *corpora* baseados em legendas de filmes e de seriados, cujas principais vantagens estão, especialmente, nos seguintes fatores: a acessibilidade dos dados, sua proximidade com a fala espontânea, a abrangência de domínios cobertos e as dimensões que esses *corpora* podem atingir. Os itens a seguir apresentam alguns dados gerais sobre o Subtlex-PT-BR compilado por Tang (2012):

Tipo: Escrito, de língua geral, traduzido para o português brasileiro a partir de falas de filmes e de seriados diversos.

Época da produção dos dados: Estima-se que as produções das legendas datem, pelo menos, de 2006⁸, sendo os *uploads* de novos arquivos constantes e diários. No entanto, os filmes e seriados que foram legendados, tratam ou são das mais variadas épocas.

Região: Não especificada.

Tamanho: Cerca de 61 milhões de palavras.

Acessibilidade: Devido a restrições de direitos autorais, os textos integrais não podem ser disponibilizados. O *corpus* está acessível na forma de dados já tratados (ver *links* abaixo): unigramas, bigramas, lemas e etiquetas de partes do discurso.

⁸<https://torrentfreak.com/interview-with-the-founder-of-opensubtitles/>

Etiquetagem: Disponível para lemas (etiquetas de partes do discurso).

sites: <http://www.ucl.ac.uk/llsd/tools>,
<http://zipf.ugent.be/open-lexicons/interfaces/br-pt-lemmas/>.

C-ORAL-BRASIL informal

O C-ORAL-BRASIL é um *corpus* balanceado, que procura representar as três principais tipologias da fala — diálogos, monólogos e conversações —, em contextos públicos e privados (Raso & Mello, 2012). O *corpus* é disponibilizado em arquivos de áudio e transcrito em formato de texto simples, contando com arquivos de alinhamento de texto e som, além de versão do *corpus* lematizada e etiquetada morfossintaticamente. Abaixo, encontra-se a descrição geral do *corpus*:

Tipo: Oral informal (espontâneo).

Época da produção dos dados: 2006 a 2011.

Região: Minas Gerais (majoritariamente, Belo Horizonte).

Tamanho: 208.130 palavras, distribuídas em 139 textos que contém, em média, 1500 palavras.

Acessibilidade: O *corpus* pode ser integralmente obtido com a compra do DVD.

Etiquetagem: Foi anotado através de uma versão adaptada do *parser* PALAVRAS (recebendo contribuição do autor do *parser* original) e contando com revisão manual de excerto do *corpus* escolhido ao acaso.

Site: <http://www.c-oral-brasil.org/>.

O C-ORAL BRASIL foi listado como opção por seu trabalho de compilação, por ser baseado em língua oral e ter sido anotado. Entretanto, o seu tamanho e a restrição dos informantes a falantes de Minas Gerais tornam o *corpus* pouco representativo para os propósitos deste projeto.

Outros corpora

Outros *corpora* foram pesquisados para verificar a viabilidade de utilizá-los. Devido a limitações tais como acesso, dimensões e abrangência dos dados, esses *corpora*⁹ foram descartados. Além dos *corpora* mencionados neste texto, diversos outros podem ser acessados através da Linguateca¹⁰, formados tanto de dados do português europeu, quanto do português brasileiro e, inclusive, de dados mistos.

⁹Tratam-se da compilação de textos da Wikipedia em português, do *corpus* Floresta Sintá(c)tica Afonso *et al.* (2002) e do *Corpus Brasileiro* Berber Sardinha *et al.* (2008).

¹⁰http://www.linguateca.pt/corpora_info.html

Corpus Brasileiro

Tipo: *Corpus* de português brasileiro contemporâneo, língua geral, escrito (variados textos, em geral, formais).

Época da produção dos dados: Não especificada.

Região: Não especificada.

Tamanho: Aproximadamente um bilhão de palavras.

Acessibilidade: O usuário tem acesso a informações sobre ocorrências das palavras de sua busca, além de linhas de concordância onde as palavras ocorrem. Não é dado o acesso aos textos integrais, por questões de direitos autorais.

Etiquetagem: Como todos os *corpus* do AC/DC (Linguatca), o *Corpus Brasileiro* está também anotado sintaticamente pelo PALAVRAS, e contém algumas anotações semânticas.

sites: <http://www.linguatca.pt/acesso/corpus.php?corpus=CBRAS>,
http://www.linguatca.pt/acesso/desc_corpus.php?corpus=CBRAS,
<http://corpusbrasileiro.pucsp.br/cb/Inicial.html>.

Banco de Português

Tipo: *Corpus* do português contemporâneo, composto por textos da imprensa em sua grande parte.

Época da produção dos dados: Contemporâneo.

Região: Não especificada.

Tamanho: Em 2004, continha 223 milhões de palavras (tokens).

Acessibilidade: Apenas uma amostra está disponível para consulta e uso com as próprias ferramentas do site.

Etiquetagem: Não encontramos.

Site: <http://www2.lael.pucsp.br/corpora/bp/index.htm>

2.2.2 Subsídios para o aprendizado semi-supervisionado da desambiguação

Ao iniciar esta etapa, selecionamos dois dicionários como recursos para o algoritmo de desambiguação escolhido: o dicionário eletrônico Michaelis¹¹ e o Dicionário de Sinônimos¹², ambos em suas versões disponíveis pela internet. O uso de dicionários em formato tradicional, como é o caso do Michaelis, pode impor limitações à automatização da desambiguação: as definições das palavras não são regularmente estruturadas para estabelecer relações semânticas com outras palavras, como no caso dos tesauros, e não costuma haver consenso entre os dicionários quanto às acepções atribuídas às palavras. Entretanto, o uso do Michaelis se apresentou como uma das poucas alternativas viáveis para a realização da tarefa.

Outras opções foram consideradas, na busca de recursos que fornecessem maiores informações sobre os lemas a serem desambiguizados, tais como relações semânticas hierárquicas e associativas. Durante a primeira etapa desta pesquisa, o projeto Wordnet-BR¹³ do NILC-São Carlos (Núcleo Interinstitucional de Linguística Computacional) e outros dicionários analógicos foram avaliados. Essas opções, porém, mostraram-se inviáveis para a tarefa: a Wordnet-BR (Dias-da Silva, 2006b; Dias-da Silva *et al.*, 2006) encontra-se incompleta, estando disponível apenas para verbos, e não havíamos encontrado versões eletrônicas de outros tesauros do português brasileiro.

Além do dicionário Michaelis e o de sinônimos, outros dicionários do português foram listados inicialmente como opções: Wikcionário¹⁴, Dicionário Online Houaiss¹⁵, Dicionário Online Caldas Aulete¹⁶, Dicionário Priberam¹⁷, Dicionário inFormal¹⁸, Dicionário Online de Português¹⁹, Dicionário Aurélio²⁰, Dicionário de Sinônimos Lexikon²¹, Dicionário Analógico Lexikon da Língua Portuguesa²², Dicionário de locuções

¹¹<http://michaelis.uol.com.br/>

¹²<http://sinonimos.com.br/>

¹³<http://www.nilc.icmc.usp.br/wordnetbr/>

¹⁴pt.wiktionary.org/

¹⁵houaiss.uol.com.br/

¹⁶www.aulete.com.br/

¹⁷www.priberam.pt/DLPO/

¹⁸www.dicionarioinformal.com.br/

¹⁹www.dicio.com.br/

²⁰www.aureliopositivo.com.br/

²¹lexikon.com.br/dicionario-de-sinonimos

²²lexikon.com.br/dicionario-analogico-da-lingua-portuguesa

e expressões da língua portuguesa²³ e Dicionário de usos do português do Brasil (Borba, 2002).

Em geral, as opções de dicionários que continham informações de relações semânticas dos lemas a serem desambiguados, potencialmente os melhores recursos dentre os listados, apresentaram o fator limitante de estarem disponíveis apenas em versões impressas. A exceção foi o dicionário de sinônimos, que foi selecionado por estar disponível *online*. Os demais dicionários (com exceção do dicionário Aurélio), apesar de estarem disponíveis em suas versões virtuais para consulta, foram descartados por falta de adequação de sua estrutura à tarefa (caso do dicionário inFormal, que apresenta pouco rigor na formatação das entradas) ou por não serem gratuitos (caso do dicionário Houaiss, que exige a assinatura de um provedor de internet para ser acessado ou a compra do CD-ROM do dicionário).

2.3 Desambiguação automática de sentidos

A abordagem padrão da polissemia no âmbito do processamento de linguagem natural tem equacionado o problema do ponto de vista de duas tarefas inevitáveis (Ravin & Leacock, 2000). Uma tarefa é determinar, para cada palavra, os diferentes sentidos a ela associados; a outra é a seleção de sentidos ou desambiguação. Dado que determinada palavra tem n sentidos diferentes, torna-se necessário um procedimento de seleção do sentido que é apropriado ao contexto no qual a palavra é usada. Jurafsky & Martin (2000) definem desambiguação de sentidos de palavras como:

a tarefa de examinar tokens contextualmente e determinar qual sentido de cada palavra está sendo utilizado. WSD é uma tarefa com uma longa história em linguística computacional e [...] é um trabalho não-trivial dada a natureza de muitos sentidos de palavras. Entretanto, há algoritmos robustos que podem atingir altos níveis de acurácia dadas algumas pressuposições razoáveis. Muitos desses algoritmos se apoiam na similaridade contextual para escolher os significados adequados²⁴.

O papel do contexto é um dos aspectos cruciais da desambiguação automática. Como afirmam Jurafsky & Martin (2000), palavras com sentidos similares frequentemente ocorrem em contextos similares, seja em termos de corpora (vizinhança),

²³lexikon.com.br/dicionario-de-locucoes-e-expressoes-da-lingua-portuguesa

²⁴Link para a aula introdutória do curso de Processamento de Linguagem Natural oferecido pelos professores Dan Jurafsky e Christopher Manning através do Coursera: <https://class.coursera.org/nlp/lecture/124>. O vídeo comenta o desenvolvimento atual das principais áreas de NLP, incluindo *word sense disambiguation*.

seja em termos de dicionários tradicionais e analógicos (tendo definições similares ou estando em categorias próximas em hierarquias semânticas).

A tarefa de desambiguação automática de sentidos, tal como é tradicionalmente entendida, pode ser definida como sendo, essencialmente, uma tarefa de classificação: os sentidos das palavras são as classes e os contextos em que as palavras-alvo ocorrem fornecem as evidências para a classificação das respectivas ocorrências de acordo com as classes possíveis para o alvo em questão (Agirre & Edmonds, 2007). De acordo com essa perspectiva, os sentidos são tomados como um inventário fixo de classes discretas dadas por um dicionário, uma base de conhecimento lexical ou uma ontologia (na qual os sentidos correspondem a conceitos lexicalizados por uma determinada palavra ou expressão) (Agirre & Edmonds, 2007). Em sua forma mais básica, algoritmos de WSD tomam como entrada uma palavra em contexto juntamente com um inventário fixo de sentidos potenciais e retornam a acepção adequada para o contexto de uso fornecido. A natureza da entrada e do inventário depende, então, da aplicação final.

A tarefa de WSD pode ser considerada enquanto tarefa intermediária ou como produto final. No primeiro caso, o modo de implementação e avaliação da tarefa dependerão da aplicação que se deseja realizar ao final (p. ex.: tradução automática, *question-answering*, recuperação de informação, classificação textual). No caso de considerarmos a desambiguação isoladamente, geralmente são utilizados conjuntos de definições e de relações semânticas extraídas de dicionários tradicionais e analógicos.

À medida em que este projeto foi avançando, reconsideramos alguns aspectos da tarefa de desambiguação e buscamos novas propostas: inicialmente, seria implementado um algoritmo de desambiguação baseado em grafos de relações semânticas construído a partir do projeto OpenWordNet-PT, com a expectativa de que as relações semânticas fornecessem dados mais informativos para a desambiguação. Optamos, no entanto, por nos concentrar em uma tarefa correlata: a extração de índices de polissemia a partir de *corpora*.

2.3.1 Método *baseline* para a desambiguação automática

Uma primeira medida de referência para os algoritmos de desambiguação são os métodos ditos *baseline*. O principal *baseline* em WSD costuma ser a aplicação do método heurístico que se baseia na escolha da acepção mais frequente (caso essa medida esteja disponível). Considerando-se o custo de se refinar a desambiguação em relação ao *baseline*, espera-se que um algoritmo qualquer de desambiguação mais

elaborado supere, no mínimo, o limiar de desempenho obtido com o *baseline* para um mesmo conjunto de dados.

Algoritmo 2.1: Método *baseline* de aceção mais frequente

Entrada: (*palavra*, *janelas-contextuais*, *aceção-mais-frequente* de *palavra*)

Saída: *limiar-baseline* de *palavra*

```

1 para cada janela em janelas-contextuais faça
2   | gabarito ← GABARITO(palavra, janela, aceção-mais-frequente)
3   | se aceção-mais-frequente == gabarito então
4   |   | acertos ← acertos + 1
5   | fim
6 fim
7 limiar-baseline ← (acertos ÷ número-de-janelas-contextuais)
8 retorna limiar-baseline
  
```

O algoritmo 2.1 representa o método *baseline* da aceção mais frequente. São fornecidas como entradas: a palavra-alvo, o conjunto das janelas contextuais de todas as ocorrências da palavra-alvo no *corpus* e aceção mais frequente da palavra-alvo. A partir de um *corpus* anotado semanticamente, pode-se gerar uma função que compara os resultados do *baseline* com os dados anotados, como faz a função GABARITO apresentada no algoritmo. O limiar do *baseline* é obtido, então, pela razão entre o seu número de acertos e o número total de ocorrências da palavra-alvo.

É importante ressaltar que, se considerarmos qual seria o melhor resultado possível para um desambiguador, a referência passa a ser o nível de concordância entre desambiguadores humanos. Segundo Jurafsky & Martin (2000), há diversos trabalhos que propõem formas de se estabelecer essa medida, com níveis relatados de concordância entre revisores variando entre 70% e 90%. Os melhores resultados de desambiguação reportados atualmente (para o inglês) chegam a taxas de acerto de 97% (Ide & Véronis, 1998). Para o português brasileiro, o trabalho de Nóbrega & Pardo (2012) empregando uma adaptação do método de Lesk (discutido a seguir) obteve resultados, em geral, inferiores ao do método heurístico do sentido mais frequente. Não encontramos outros trabalhos semelhantes para o português brasileiro.

2.3.2 Algoritmo de Lesk

O algoritmo de Lesk (Lesk, 1986) propõe que a desambiguação seja baseada no cruzamento das palavras das aceções do palavra-alvo com os das aceções das palavras co-ocorrentes. Em sua versão simplificada, a base são pistas sintagmáticas:

as palavras das acepções da palavra-alvo são cruzadas diretamente com as palavras que ocorrem nas janelas contextuais da palavra-alvo. Nesta seção, explicitaremos as etapas do algoritmo implementado e os formatos de seus dados de entrada e saída. Dedicamos também uma sub-seção pormenorizando o algoritmo empregado para a avaliação dos resultados das desambiguações.

Algoritmo 2.2: Algoritmo de Lesk Simplificado

Entrada: (*palavra*, *acepções*, *sentença*)

Saída: a melhor acepção de *palavra*

```

1 máx-intersecção  $\leftarrow$  0
2 contexto  $\leftarrow$  conjunto de palavras em sentença
3 para cada acepção em acepções de palavra faça
4   | palavras-acepção  $\leftarrow$  conjunto de palavras da acepção
5   | intersecção  $\leftarrow$  CALCULAINTERSEC(palavras-acepção, contexto)
6   | se intersecção > máx-intersecção então
7   |   | máx-intersecção  $\leftarrow$  intersecção
8   |   | melhor-acepção  $\leftarrow$  acepção
9   | fim
10 fim
11 retorna melhor-acepção

```

Explicando a versão simplificada do algoritmo de Lesk em outros termos, tem-se que: $D = \{A_1, A_2, \dots, A_n\}$ é o conjunto das n acepções de dicionário de uma palavra-alvo p a ser desambiguada; J é a janela contextual de p dada pela sequência $j_1, j_2, \dots, j_k$ de palavras da janela. Sendo $A_i = \{t_1, t_2, \dots, t_m\}$ a i -ésima acepção de p , então, na i -ésima iteração, a intersecção será calculada entre os conjuntos A_i e J . Ao final de n iterações, a acepção escolhida do conjunto D será aquela que apresentar a maior intersecção com a janela J .

No algoritmo de Lesk original, pretendia-se aumentar a quantidade de informações a serem cruzadas com o objetivo de ajudar a refinar a desambiguação. Nele, temos: $D_p = \{A_1, A_2, \dots, A_n\}$, o conjunto das n acepções de dicionário da palavra-alvo p a ser desambiguada e $J = \{j_1, j_2, \dots, j_k\}$, uma janela contextual de p . Para cada palavra $j_k \in J$, temos $D_{j_k} = \{d_1, d_2, \dots, d_j\}$, o conjunto de todas as palavras das definições de j_k no dicionário. A intersecção parcial é calculada entre os conjuntos A_i e D_{j_k} , de forma que a intersecção total de uma determinada acepção A_i será a soma das intersecções parciais calculadas para cada elemento j_k da janela contextual. Ao final, é escolhida a acepção $A \in D_p$ que tiver a maior intersecção total.

Algoritmo 2.3: Algoritmo de Lesk Original**Entrada:** (*palavra*, *acepções*, *sentença*)**Saída:** a melhor acepção de *palavra*

```

1 máx-intersecção  $\leftarrow$  0
2 contexto  $\leftarrow$  conjunto de palavras em sentença
3 para cada acepção em acepções de palavra faça
4   palavras-acepção  $\leftarrow$  conjunto de palavras da acepção
5   intersecção  $\leftarrow$  0
6   para cada palavra em contexto faça
7     entrada-palavra  $\leftarrow$  DEFINICOESDICCIONARIO(palavra)
8     intersecção-palavra  $\leftarrow$  CALCULAINTERSEC(entrada-palavra,
9       palavras-acepção)
9     intersecção  $\leftarrow$  intersecção + intersecção-palavra
10  fim
11  se intersecção > máx-intersecção então
12    máx-intersecção  $\leftarrow$  intersecção
13    melhor-acepção  $\leftarrow$  acepção
14  fim
15 fim
16 retorna melhor-acepção

```

2.3.3 Avaliação da desambiguação

Uma vez que o algoritmo de desambiguação já havia sido definido, tornou-se necessário estabelecer formas de avaliá-lo por meios alternativos, que não envolvessem a verificação direta dos resultados da desambiguação dos lemas, para os quais não tínhamos anotações semânticas que pudessem servir como referência.

O método de desambiguação aplicado foi o de desambiguação de pseudo-palavras, referido em Jurafsky & Martin (2000) e inspirado nos trabalhos de Gale *et al.* (1992) e Schütze (1992). As pseudo-palavras foram formadas por combinações dos lemas a serem desambiguados, de forma que a pseudo-palavra tivesse como *acepções* os conjuntos das definições dos lemas de que foi formada.

Foram criados três conjuntos diferentes de pseudo-palavras para verificarmos se o desempenho do algoritmo dependeria das frequências dos lemas combinados, isto é, da preponderância relativa entre as ocorrências das ‘acepções’ das pseudo-palavras. Para tanto, os cem lemas foram organizados de acordo com suas frequências no *corpus* Mac-Morpho e combinados sob três critérios, formando os grupos A, B e C descritos a seguir. Os conjuntos de pseudo-palavras podem ser consultados no apêndice B.

Grupo A: combinações feitas entre pares consecutivos de lemas por frequência; assim, as pseudo-palavras seriam compostas por lemas de frequências próximas.

Grupo B: os lemas foram pareados com seus opostos na ordem de frequência, isto é, os lemas mais frequentes foram combinados com os lemas menos frequentes.

Grupo C: as combinações foram feitas aleatoriamente.

Para ilustrar o método de desambiguação com pseudo-palavras, utilizaremos os exemplos apresentados a seguir. Nesses exemplos, as ocorrências de *banana* e de *porta* foram substituídas pela pseudo-palavra *banana-porta*, que deve ser desambiguada entre as suas duas ‘acepções’, *banana* e *porta*, para que as desambiguações sejam comparadas com as ocorrências originais das sentenças.

- (2.3) a. *A bananeira exibia cachos ainda muito verdes, mas João não quis esperar e, esticando a mão, arrancou uma **banana-porta** que estava um pouco mais amarela e madura para comer.*
- b. *Ao ouvir a campainha tocar, João rapidamente correu, girou a maçaneta e abriu a pesada **banana-porta** de madeira da entrada da casa.*

Considerando-se que a palavra *banana* tem as acepções $A_b = \{b_1, b_2, b_3\}$ e que a palavra *porta* tem o conjunto de acepções $A_p = \{p_1, p_2, p_3, p_4\}$, a pseudo-palavra *banana-porta* terá duas acepções possíveis: A'_b de *banana* ou A'_p de *porta*, cada qual sendo a concatenação das acepções das respectivas palavras originais. Segundo o dicionário Michaelis, as palavras *banana* e *porta* são dadas pelas seguintes definições (entre outras mais):

ba.na.na (sf)

1. *Bot* Fruto da bananeira. *col:* cacho, penca.
2. Espécie de pescada.
3. *ch* Gesto obsceno que se faz com a mão fechada e braço dobrado.

por.ta (sf)

1. Abertura na parede de um edifício, desde o nível do piso até uma altura conveniente, para permitir a entrada e a saída.
2. Peça de madeira ou outro material que, presa a ombreiras por meio de dobradiças, serve para fechar essa abertura.

3. Peça com que se fecham certos móveis, veículos etc.
4. Casa, habitação, morada: *Andar de porta em porta*.

Como se pode verificar, a entrada *banana* inclui em sua definição *bananeira*, *cachos*, *mão* e algumas palavras funcionais em comum com a primeira janela contextual de *banana-porta* (exemplo 2.3a). A definição de *porta*, por sua vez, apresenta poucas palavras em comum com a primeira sentença e nenhuma dessas palavras chega a ser distintiva do sentido de *porta*. No caso da segunda sentença, exemplo 2.3b, temos os itens *madeira*, *entrada*, *casa*, entre outros, compartilhados entre a sentença e as definições do dicionário para a palavra *porta*. A entrada de *banana*, por sua vez, apresenta pouca intersecção com a segunda sentença.

2.4 Rediscutindo as bases da desambiguação

Até o momento, discutiram-se métodos de desambiguação automática cujas medidas de polissemia — esses métodos, a rigor, não distinguem polissemia e homonímia —, são obtidas através de dicionários. A partir desta seção, passaremos a abordar métodos de extração de índices de polissemia e de clusterização de palavras co-ocorrentes com o propósito de obter e analisar dados que permitirão, em trabalhos futuros, a exploração de métodos de desambiguação baseados em *corpora* e sem a dependência de inventários fixos de acepções.

2.4.1 Extração automática de índices de polissemia

Uma maneira de contornar problemas da seleção de acepções e da desambiguação tem sido o recurso aos dados estatísticos de usos das palavras. Trata-se de caracterizar os sentidos em termos de padrões de contextualização distintos, tendo em conta a probabilidade de uma palavra em determinado sentido co-ocorrer com certas palavras e não com outras (Schütze, 1992). Em relação ao problema específico de desambiguação automática de sentidos, em vez de atribuir um sentido fornecido de antemão a uma palavra, pode-se agrupar palavras co-ocorrentes à palavra-alvo, tomando-as como relacionadas na medida em que partilham contextos semelhantes e definindo esses grupos como correspondentes aos sentidos da palavra. Esse tratamento implica que as palavras são semelhantes em sentido na medida em que os seus padrões de contextualização sejam semelhantes, e que uma palavra é polissêmica na medida em que os usos dessa palavra correspondam a diferentes padrões de contextualização.

Em trabalho intitulado *Method and apparatus for measuring the degree of polysemy in polysemous words*²⁵, Sproat & VanSanten propuseram o cálculo de um índice de polissemia partindo da extração de janelas contextuais das ocorrências de uma dada palavra-alvo em um *corpus*. Apesar da expressão *graus de polissemia* empregada no título, o algoritmo proposto pelos pesquisadores consiste, essencialmente, em retornar valores que indicam o quanto uma palavra se distancia da monosssemia. O índice, no entanto, não foi associado a uma medida absoluta: não se estabelece o limite máximo para os valores dos índices — ou um limiar de referência para distinguir monosssemia de polissemia — e, dada a natureza dos cálculos (que serão detalhados nas próximas seções), os índices são dependentes do *corpus* no qual o algoritmo é aplicado.

Além disso, não se fez distinção entre o que seria considerado polissemia ou homonímia na proposta dos autores: note-se que o exemplo de polissemia utilizada pelos autores foi a palavra *bass*, exemplo recorrente na literatura (Jurafsky & Martin (2000), Yarowsky (1995), Yarowsky (1997), Navigli (2009)) para introduzir a questão da diversidade de sentidos que uma mesma forma gráfica pode assumir em uma língua. A palavra *bass* pode se referir tanto ao timbre grave e ao instrumento musical a ele associado (*contrabaixo*) — noções que estão, assim, em relação polissêmica —, como a um peixe de água doce, a perca. Este último caso configura-se como exemplo de homonímia por relação ao lexema associado ao timbre.

Feitas essas considerações, continuaremos utilizando a expressão “índice de polissemia” para nos referir à saída do algoritmo a fim de manter a terminologia original, com a ressalva de que o valor de saída diz respeito ao quanto podemos considerar uma palavra como não monossêmica relativamente aos índices de outras palavras testados com o mesmo algoritmo. Apesar de termos encontrado poucas referências na literatura que explorassem essa possibilidade de medida para a polissemia (Erk & Padó (2008), Dorow (2006), Deerwester *et al.* (1990)), os autores mencionam uma série de aplicações, como desdobramentos do cálculo dos índices de polissemia, esboçando, inclusive, um método de desambiguação partindo da proposta inicial de extração desses índices.

Acrescentamos, todavia, que o algoritmo fornece uma segunda medida de polissemia que pode ser extraída do *corpus*: a quantidade de agrupamentos de palavras obtidos na etapa de clusterização, que corresponderiam à quantidade de acepções da palavra-alvo. Os autores propõem que as palavras sejam reagrupadas em volta das barras mais distintivas, unindo a cada uma delas os grupos de palavras vizinhas de

²⁵Registrado sob a patente US 6.256.629 B1 em 2001.

cada lado, formando *clusters*. Essa segunda medida é pouco explorada pelos autores e propomos refinamentos para o seu cálculo, além de fazer, ao final, uma comparação entre as duas medidas de polissemia extraídas do *corpus*. Em seus resultados para a palavra *bass*, Sproat & VanSanten obtêm apenas *clusters* que distinguem os sentidos homônimos; os sentidos polissêmicos relacionados ao campo semântico do timbre ficaram agrupados em um único *cluster* (Ver gráfico 2.4 e as listas de palavras correspondentes a cada *cluster* que o acompanham). No entanto, implementamos uma versão adaptada do algoritmo procurando explorar a possibilidade de diferenciação entre casos de polissemia e de homonímia a partir da clusterização, como discutiremos na seção 3.2 e no capítulo 4.

O fluxograma da figura 2.3 apresenta o algoritmo original em linhas gerais, tal como os autores o apresentam em sua patente. O pseudo-código pode ser verificado no algoritmo 2.4.

Versão original do algoritmo de extração de índices de polissemia

As etapas do algoritmo original se dão da seguinte forma: dada uma palavra-alvo w , a partir de um *corpus* C , extraem-se todas as janelas contextuais de tamanho k em torno de cada instância encontrada de w , de forma que tenhamos extraído $k/2$ palavras antes e $k/2$ depois de w para cada janela da palavra-alvo. Tendo o conjunto de todas as palavras q (co-ocorrentes) — excetuando-se a própria palavra-alvo, w — de todas as janelas contextuais de w em C , selecionam-se, dentre esses co-ocorrentes, as palavras relevantes de acordo com critérios de frequência no *corpus* e percentual mínimo da probabilidade de q ocorrer dado que w ocorreu: $P(q|w)$. REL_w será o conjunto das n_w palavras relevantes para w , extraídos com base nas seguintes restrições:

$$\frac{FREQ_w(q)}{FREQ_C(q)} > 0.005 \quad (2.1a)$$

$$FREQ_C(q) > 2 \quad (2.1b)$$

Em outros termos, na equação 2.1, temos que as palavras q relevantes devem ter uma proporção de co-ocorrências com w , em relação ao total de vezes em que ocorrem no *corpus*, que supere um determinado limiar. Os autores comentam que o limiar seria tipicamente dado por 0.005. Entretanto, não são apresentadas referências ou os critérios para a escolha desse valor. O critério do limiar de ocorrências no *corpus*, 2.1b, foi estabelecido pelos autores como 2. Na versão que adaptamos do método,

optamos por um limiar proporcional ao *corpus*, pois o valor absoluto pode tornar a filtragem dos co-ocorrentes relevantes muito permissiva para *corpora* grandes, por exemplo.

Uma vez que todos as n_w palavras relevantes foram selecionados, para cada q , q' em REL_w , a função $CO_k(q|q')$ denotará a frequência de q' nas janelas de tamanho k de todas as instâncias de q . Define-se, então, uma matriz de distâncias D_w , simétrica, $n_w \times n_w$, formada pelos elementos:

$$DIST_{ij}(q, q') = 1 - \frac{P(q|q') + P(q'|q)}{2} \quad (2.2a)$$

$$P(q|q') = \frac{CO_k(q|q')}{Freq_C(q)} \quad (2.2b)$$

$$P(q'|q) = \frac{CO_k(q'|q)}{Freq_C(q')} \quad (2.2c)$$

A equação 2.2a consiste em calcular a média entre as probabilidades condicionais de q e q' (de ocorrerem uma no contexto da outra) e subtraí-la de 1, de forma que, quanto maior a média dessas probabilidades, isto é, quanto mais relacionados estiverem essas duas palavras co-ocorrentes, menor seja a distância calculada entre eles e vice-versa. Portanto, interpretamos essa matriz de distâncias D_w como uma matriz de distâncias entre as palavras co-ocorrentes em um espaço n_w dimensional.

Considerando-se as dimensões da matriz D_w , torna-se necessário submetê-la a um método de redução de dimensionalidade para facilitar a manipulação e representação de seus dados. Todavia, a patente em que nos baseamos apresenta essa etapa de forma bastante vaga, apenas indicando como referência o livro de Torgerson (1958) sobre adaptação de escalas, sem mencionar explicitamente como o método foi implementado ou o que fundamentou a escolha. Optamos, então, por adotar que, na etapa de redução de dimensionalidade da matriz de distâncias dos co-ocorrentes, a redução é realizada através de um procedimento de decomposição conhecido como *Singular Value Decomposition* (SVD), ou *Decomposição de Valor Singular*, o qual decompõe a matriz em três novas matrizes, manipulando-as de forma que se obtenha uma matriz que aproxime a original, contendo as d dimensões com as quais se quer trabalhar. No algoritmo do índice de polissemia, a redução gera uma matriz P de duas colunas, isto é, de duas dimensões. Apresentaremos mais adiante, na seção 19, um exemplo dessa decomposição, explicando-a em mais detalhes, dentro de uma demonstração do passo-a-passo do algoritmo que implementamos.

O método SVD, proposto por Deerwester *et al.* (1990), parece ter se tornado um método usual de redução de dimensionalidade em trabalhos de *Latent Semantic Analysis* e é descrito por Turney *et al.* (2010) como:

uma forma elegante de melhorar as medidas de similaridade com uma operação matemática sobre a matriz de palavras-documentos, baseada em álgebra linear. A operação é a Decomposição de Valor Singular Truncada (SVD). [...] Deerwester *et al.* (1990) mencionam brevemente que a SVD truncada pode ser aplicada tanto para similaridade de documentos quanto para similaridade de palavras, mas que seu foco estava na similaridade de documentos. Landauer & Dumais (1997) aplicaram SVD truncada para calcular similaridade de palavras, atingindo *scores* comparáveis aos dos humanos em questões de múltipla escolha sobre sinonímia do TOEFL. SVD truncada aplicada à similaridade de documentos é chamada de Indexação Semântica Latente (LSI), mas, quando aplicada à similaridade de palavras, ela é chamada de Análise Semântica Latente (LSA).

A justificativa para a necessidade da redução apoia-se, principalmente, no fato de o tratamento dos dados e a interpretação dos resultados serem facilitados, possibilitando também a representação gráfica bidimensional. O esquema da decomposição e sua explicação se encontram na equação 2.3:

$$M = U_0 S_0 V_0 \quad (2.3a)$$

$$\hat{M} = US, \text{ onde} \quad (2.3b)$$

$$U = U_0[0 : n_w, 0 : 2] \quad (2.3c)$$

$$S = S_0[0 : 2, 0 : 2] \quad (2.3d)$$

As matrizes U_0 e a transposta de V_0 têm colunas orto-normais e são os autovetores à esquerda e à direita de M , respectivamente. S_0 é, por convenção, uma matriz diagonal de autovalores positivos dispostos em ordem decrescente. A decomposição pode ser considerada única, exceto por certas permutações de linhas, colunas e sinais.

Implementamos a decomposição SVD tendo como base o artigo de Deerwester *et al.* (1990) e selecionamos as duas colunas (autovetores) de V_0 relacionadas aos dois maiores autovalores de S_0 como eixos de referência para a redução das dimensões do espaço vetorial $n_w \times n_w$ no qual as n_w palavras co-ocorrentes são inicialmente representadas. Como resultado, reduz-se a matriz P_w para uma matriz $P_{n_w \times 2}$, de tal

forma que as palavras mais similares sejam preservadas pela seleção dos autovalores e autovetores associados aos eixos dessas palavras.

A partir da matriz P , procedemos então à mudança de coordenadas, de forma que os valores representados em coordenadas cartesianas sejam expressos em coordenadas polares (expressas por ângulos e raios), ressaltando os ângulos entre as palavras em sua distribuição espacial. Feita a mudança de coordenadas, realiza-se uma subdivisão das palavras, unindo-as em grupos, inicialmente, a cada 4 graus, com os graus variando entre $[0, 360^\circ[$, atribuindo a cada grupo o número total de palavras abarcadas por aquele intervalo de ângulos. Uma vez que esses agrupamentos estão feitos, resta aplicar o método de regressão isotônica sobre esses agrupamentos, que consiste em realizar uma regressão isotônica (estritamente crescente) a partir dos grupos iniciais e mudar para uma regressão antitônica (estritamente decrescente) a partir de um certo ponto que será o de máximo (y_{max}) da regressão, minimizando os erros quadráticos. A soma dos erros quadráticos gerados por essa regressão isotônica será o índice de polissemia da palavra-alvo que estiver sendo testada, de acordo com o *corpus* C utilizado.

Técnicas de regressão isotônica e antitônica são aplicadas para que se obtenha um ajuste de pico único. A soma dos quadrados dos erros gerados é usada como o índice de polissemia. Inicia-se com a regressão isotônica (estritamente crescente) para uma sequência de números $x_1, \dots, x_n$, sendo definida como uma sequência $\hat{x}_1, \dots, \hat{x}_n$, que minimiza a soma dos quadrados das diferenças entre x_i e $\hat{x}_i$, sob a condição de que $\hat{x}_1 \leq \hat{x}_2 \leq \dots \leq \hat{x}_n$. No caso da regressão antitônica, a restrição passa a ser $\hat{x}_1 \geq \hat{x}_2 \geq \dots \geq \hat{x}_n$. Essas duas regressões são, em outras palavras, as curvas (crescentes ou decrescentes) que melhor aproximam os dados originais.

Para cada i , de 1 a n , a regressão isotônica é computada a partir de todos os valores anteriores $(1, \dots, i)$ e a regressão antitônica é calculada para todos os seguintes $(i, \dots, n)$. Para medir os ajustes de cada pico nos dados, basta que se faça a adição das somas dos quadrados das diferenças geradas pelas duas regressões. Define-se, então, o pico principal i como aquele cuja adição das somas é minimizada.

Uma das justificativas para a aplicação desse método é a de que uma palavra que tenha, por exemplo, dois grandes picos (agrupamentos com maior quantidade de palavras) não próximos, intercalados por picos intermediários, forçará um erro maior para o método e, conseqüentemente, aumentará o índice de polissemia em relação a uma palavra monossêmica: o primeiro dos dois maiores picos forçará um ponto de máximo a partir do qual a regressão decrescente vai se distanciar do segundo pico maior.

Figura 2.3: Fluxograma de extração do índice de polissemia

Para o cálculo do número de *clusters* correspondentes ao que pode ser interpretado como número de acepções da palavra-alvo, o algoritmo original sugere que sejam identificados os maiores picos (não é dado um limiar de referência para as dimensões mínimas do que se considerariam picos) e que sejam unidos aos picos os quatro agrupamentos iniciais de quatro graus mais próximos, dois de cada lado do pico: esses agrupamentos em torno dos picos, os *clusters*, corresponderiam às distintas acepções da palavra-alvo. O algoritmo completo está representado no pseudo-código do algoritmo 2.4.

Os exemplos dos autores podem ser vistos nas figuras 2.4 e 2.5. A conclusão de que os picos distinguidos em cada gráfico correspondem a sentidos distintos foi

Algoritmo 2.4: Algoritmo de extração do índice de polissemia

Entrada: (*palavra-alvo*, *texto*)
Saída: *índice de polissemia da palavra-alvo*
número de acepções da palavra-alvo

- 1 *janelas contextuais* $\leftarrow$ EXTRAJANELAS(*palavra-alvo*, *texto*)
- 2 *conjunto de co-ocorrentes* $\leftarrow$ SET(*janelas contextuais*)
- 3 *palavras relevantes* $\leftarrow$ FILTRARELEVANTES(*conjunto de co-ocorrentes*, *janelas contextuais*)
- 4 **para cada palavra em palavras relevantes faça**
- 5 *janelas da palavra* $\leftarrow$ EXTRAJANELAS(*palavra*)
- 6 **para cada co-ocorrente em palavras relevantes faça**
- 7 *i* $\leftarrow$ INDEX(*palavra*, *palavras relevantes*)
- 8 *j* $\leftarrow$ INDEX(*co-ocorrente*, *palavras relevantes*)
- 9 *matriz de co-ocorrentes*[*i*, *j*] $\leftarrow$ CALCULADISTÂNCIAS(*co-ocorrente*, *janelas da palavra*)
- 10 **fim**
- 11 **fim**
- 12 *U*, *S*, *V* $\leftarrow$ DECOMPOSICAOSVD(*matriz de co-ocorrentes*)
- 13 *matriz_reduzida* $\leftarrow$ U[0:*n_w*, 0:2] $\times$ S[0:2, 0:2]
- 14 *matriz_polar* $\leftarrow$ COORDENADASPOLARES(*matriz_reduzida*)
- 15 *barras_4graus* $\leftarrow$ REAGRUPACOOCCORRENTES(*matriz_polar*)
- 16 *índice de polissemia* $\leftarrow$ REGRESSAOISOTONICA(*barras_4graus*)
- 17 *clusters* $\leftarrow$ CLUSTERIZAPORPICOS(*barras_4graus*)
- 18 *número de acepções* $\leftarrow$ ContaClusters(*clusters*)
- 19 **retorna** *índice de polissemia*, *número de acepções*

ratificada manualmente, através da verificação das palavras que compunham os agrupamentos de cada pico. São exibidos dois conjuntos de palavras que ocorreram acumulando dois *clusters*, à direita e à esquerda, em torno de cada pico obtido para a palavra *bass* (figura 2.4). Cada conjunto a seguir corresponde a um dos picos:

1. 1.8, geese, Resources, deer, grows, grouse, lb, chitin, Pacific, gravel, forage, females, eggs, ecological, white-tailed, coast, weigh, forked, 2.1, Suwannee, turkey, valuable, ducks, striped, nest, CAMPBELL, skunk, snake, spotted, sea, raccoon, quail, bituminous, carnivorous, birds, sand, species, Animal, hemlock, Mineral, streams, temperate, bays, lakes, laterally, kg, squirrel, lake, basin, bear, cooler, cm, yellow, reach, reaches, relatives, reserves, transitional, stray, shoulder, reservoirs, pollution, giant, eastern, drainage, desirable, gray, limestone, insufficient, hunting, guards, 0.9, 45, Strait, Bass, North, Florida, Erie, male, currents, female, naturally, 550, Trout, heads, sparse, sport, uneven, compressed, Northern, hair, ground, U., reproduction, repeats, worldwide, Utah, exceeds, 150, prefer, reproduce, belongs, Along, robust, barred, ARTHUR, resources, sportsmen, 250, VOICE, families, differently, Texas, America, extend, Portugal, PATTON, Alabama, Michigan, white, Indigenous, Common, lengths, majority, midway,

Figura 2.4: *clusters* da palavra *bass* em coordenadas polares com suavização

weights, native, westward, Dominions, Kentucky, restricted, leg, Eventually, TURNER, rare, rarely, descended, Usually, reinforced, 1500-1800, stream, generalized, Illinois, Massachusetts, slightly, Made, Probably, horns, Several, None, common, shorter

2. 1750, 17th, drum, theme, tuba, makers, scores, Gilmore, finger, score, song, lyric, treble, intervals, Fyodor, sound, piece, beat, played, accompany, neglected, reed, recorders, voice, shorthand, written, singer, valved, singers, soprano, baritone, brass, heard, loudspeakers, improvisation, organ, marching, contralto, octaves, notes, octave, playing, pitch, guitars, notation, saxophone, note, trombone, Vienna, drums, tone, singing, WIENANDT, Boris, jazz, textures, Wolfgang, tones, cymbals, Ludwig, tenor, OBOE, Donna, performance, BASS, Franz, vocal, ELWYN, baroque, woodwind, flutes, flute, tuned, chorus, chromatic, rhythm, viol, guitar, strings, ensemble, harmonic, fretted, double-reed, string, VIOLIN, symphony, trumpets, violas, violins, chamber, voices, harmony, bop, bugle, alto, accompaniment, d'amore, Wagner, trombones, concert, choir, clarinet, composing, compositions, E-flat, music, performances, performer, performers, instruments, solo, right-hand, melody, melodic, percussion, resonator, CLARINET, orchestras, lute, keyboard, musical, improvised, sounded, operatic, B-flat, harps, MUSIC, instrumental, instrument, Godunov, musicians, reggae, piccolo

As funções apresentadas no pseudo-código 2.4 serão retomadas mais detalhadamente na seção a seguir, após a apresentação da versão adaptada do algoritmo, e exemplificadas através da aplicação do algoritmo para uma palavra-alvo escolhida.

Figura 2.5: Cluster da palavra *oxidation* em coordenadas polares com suavização

Exemplo de implementação do algoritmo

Procurando destrinchar melhor os métodos adotados, elaboramos um exemplo para ser submetido ao algoritmo, passo-a-passo. As entradas abaixo²⁶ para a palavra *vela* serão utilizadas para formar o texto que corresponderá ao *corpus* de exemplo C.

vela 1: Vela ou candeia é uma fonte de luz que consiste de um pavio inserido num combustível sólido, tipicamente parafina. A função do pavio na vela é dar suporte para o combustível ser queimado, produzindo fogo e luz. Além da parafina, a vela pode ser feita de outros combustíveis.

vela 2: Como um barco à vela consegue navegar contra o vento? Totalmente contra o vento não rola, mas dá pra navegar com a vela quase de encontro ao vento.

As janelas contextuais de *vela* deverão conter até 30 co-ocorrentes, ou seja, até 15 palavras de cada lado, podendo haver intersecções entre as janelas. Ressaltamos que não será feito pré-processamento desse texto para eliminar palavras funcionais (na implementação proposta neste trabalho, foram mantidas nas janelas apenas as palavras com as etiquetas de substantivo, adjetivo, verbo e advérbio). Dessa forma, extrairemos as seguintes janelas:

1. vela ou candeia é uma fonte de luz que consiste de um pavio inserido num combustível

²⁶Adaptadas de: <https://pt.wikipedia.org/wiki/Vela> e mundoestranho.abril.com.br, acessados em 13 de junho.

2. sólido tipicamente parafina a função do pavio na vela é dar suporte para o combustível ser queimado, produzindo fogo e luz além da parafina
3. a vela pode ser feita de outros combustíveis
4. como um barco à vela consegue navegar contra o vento totalmente contra o vento não rola mas dá pra
5. navegar com a vela quase de encontro ao vento

Agrupando-se as palavras de todas as janelas, obtivemos um total de 76 palavras. Devemos, agora, selecionar aqueles que sejam relevantes. Para isso, considerando a pequena extensão do exemplo, filtraremos as palavras com o mínimo de duas ocorrências. Idealmente, adicionaríamos a restrição do percentual de ocorrências do co-ocorrente nas janelas em relação ao seu total de ocorrências no *corpus*; essa restrição não será feita em razão de esse ser um exemplo mínimo. A princípio, temos a seguinte lista de relevantes: *a, é, parafina, navegar, luz, de, vento, vela, um, contra, o, combustível, pavio, ser*. Nesse caso, as palavras funcionais (palavras com altas frequências gerais e que se tornam pouco distintivas semanticamente, tais como preposições, artigos e verbos de ligação) serão excluídas com base em uma lista de palavras funcionais do NLTK.

Feita a seleção dos relevantes, excluindo-se as palavras funcionais, a lista final conterá as seguintes palavras: *parafina, navegar, luz, vento, vela, contra, combustível, pavio*. Tendo a lista de relevantes, o passo seguinte é calcular a matriz de co-ocorrências 2.1. A partir da matriz de co-ocorrências, calculamos a matriz de similaridade 2.2 com base na equação 3.6a.

	parafina	navegar	luz	vento	contra	combustível	pavio
parafina	0	1	2	0	0	3	2
navegar	1	2	0	5	4	0	0
luz	2	0	0	0	0	2	3
vento	0	5	0	4	4	0	0
contra	0	4	0	4	2	0	0
combustível	3	0	2	0	0	2	3
pavio	2	0	3	0	0	3	2

Tabela 2.1: Matriz de co-ocorrências relevantes de *vela*

A matriz de similaridade é então submetida à decomposição SVD para que seja efetuada a redução de dimensionalidade de $n_w \times n_w$ para o espaço bidimensional. A

	parafina	navegar	luz	vento	contra	combustível	pavio
parafina	0	0.5	1.0	0	0	1.5	1.0
navegar	0.5	1.0	0	2.08	2.0	0	0
luz	1.0	0	0	0	0	1.0	1.5
vento	0	2.08	0	1.33	1.66	0	0
contra	0	2.0	0	1.66	1.0	0	0
combustível	1.5	0	1.0	0	0	1.0	1.5
pavio	1.0	0	1.5	0	0	1.5	1.0

Tabela 2.2: Matriz de similaridade das palavras relevantes

decomposição é feita de tal forma que $P_w = U_0 S_0 V_0$ (ver equação 2.3). Para o nosso exemplo, a decomposição de P_w gerou as seguintes matrizes:

$$U_0 = \begin{bmatrix} -0.12 & 0.41 & 0.61 & 0.03 & -0.52 & 0.05 & 0.41 \\ -0.58 & -0.09 & -0.51 & 0.45 & -0.40 & 0.11 & 0.13 \\ -0.07 & 0.42 & -0.34 & -0.66 & -0.38 & 0.05 & -0.36 \\ -0.58 & -0.13 & 0.22 & -0.20 & 0.13 & -0.73 & -0.11 \\ -0.53 & -0.12 & 0.26 & -0.27 & 0.34 & 0.67 & -0.06 \\ -0.10 & 0.56 & -0.32 & -0.07 & 0.51 & -0.09 & 0.55 \\ -0.10 & 0.56 & 0.15 & 0.51 & 0.19 & 0.02 & -0.60 \end{bmatrix}, S_0 = \begin{bmatrix} 4.99 \\ 4.34 \\ 1.52 \\ 1.10 \\ 0.72 \\ 0.50 \\ 0.17 \end{bmatrix}$$

E temos que $V_0 = U_0^t$, onde U^t é a matriz U transposta. Essa igualdade vem do fato de a matriz P_w decomposta ser simétrica. Em aplicações de LSI, as informações contidas em U_0 e V_0 são de naturezas diferentes, pois a matriz de correlações cruza informações de palavras e de documentos (entendam-se documentos como quaisquer tipos de textos que extrapolem o nível da palavra: sentenças, parágrafos, textos e assim por diante), sendo assimétrica.

Resta agora obter a matriz $P_{2 \times 2}$ resultante da redução de P_w para 2 dimensões:

$$P_{7 \times 2} = U_{7 \times 2} \times S_{2 \times 2}, \text{ sendo: } S_{2 \times 2} = \begin{bmatrix} 4.99 & 0 \\ 0 & 4.34 \end{bmatrix}$$

A matriz $U_{7 \times 2}$ corresponde às duas primeiras colunas da matriz U_0 . Temos, então, a matriz reduzida $P_{7 \times 2}$ calculada conforme a equação indicada anteriormente:

Figura 2.6: Distribuição dos co-ocorrentes relevantes de *vela* de acordo com agrupamentos de similaridade.

$$P_{7 \times 2} = \begin{bmatrix} -0.62 & 1.76 \\ -2.92 & -0.40 \\ -0.37 & 1.80 \\ -2.88 & -0.56 \\ -2.66 & -0.52 \\ -0.51 & 2.42 \\ -0.48 & 2.42 \end{bmatrix}$$

Faz-se, então, a mudança das colunas de $P_{7 \times 2}$ para coordenadas polares, obtendo-se a seguinte matriz de ângulos e raios:

$$P_{polar} = \begin{bmatrix} 1.87 & 70.75 \\ 2.94 & -7.76 \\ 1.84 & 78.37 \\ 2.93 & -11.06 \\ 2.71 & -11.01 \\ 2.47 & 78.17 \\ 2.47 & 78.69 \end{bmatrix}$$

O *plot* dos dados da matriz P_{polar} pode ser visto na figura 2.6. Os raios, na primeira coluna, serão, então, reagrupados a cada 4 graus, de acordo com os ângulos em

que se situarem (segunda coluna da matriz) dentro do intervalo de $[0, 360^\circ[$. Dos agrupamentos em ângulos, ponderado pelos raios das palavras de cada grupo, temos as seguintes barras e respectivas posições (em graus):

Tamanho (normaliz.)	Posição (graus)	Palavras agrupadas
0.83	-12	vento, contra
0.43	-8	navegar
0.27	68	parafina
1.0	76	luz, combustível, pavio

Tabela 2.3: Tabela dos agrupamentos de co-ocorrentes de *vela*

Figura 2.7: Distribuição dos co-ocorrentes de *vela*: as barras expressam a soma dos pesos (raios) correspondentes às palavras que a compõem.

Aplicando-se as regressões isotônica e antitônica aos dados da tabela 2.3, obtemos o erro minimizado pela regressão, interpretado como índice de polissemia. Para a palavra *vela*, no exemplo apresentado, o índice de polissemia foi de 0.84. Salientamos, porém, que o índice, por si, não tem significado. Isso significa que a sua interpretação é feita em termos comparativos e depende do cálculo do índice para outras palavras do *corpus* ou da confrontação com outra medida paralela.

Figura 2.8: Distribuição dos co-ocorrentes em *polar plot*, enfatizando as diferenças angulares entre os agrupamentos (barras).

Por fim, realizando a clusterização das barras (tabela 2.3), obtemos o número de *clusters* e as listas de palavras que compõem cada um dos *clusters* identificados. A clusterização que implementamos foi efetuada da seguinte maneira:

Para o exemplo testado, foram encontrados dois *clusters* de *vela*:

1. parafina, luz, combustível, pavier
2. vento, contra, navegar

No capítulo de resultados, apresentaremos exemplos semelhantes extraídos do *corpus* Mac-Morpho em que testamos o algoritmo, assim como análises mais gerais dos índices de polissemia que calculamos para substantivos com as maiores frequências no *corpus*.

Figura 2.9: Regressão isotônica e antitônica para a extração do índice de polissemia de *vela*: dois picos altos e distantes forçam o aumento do índice.

3

Metodologia

Neste capítulo, trataremos dos detalhes metodológicos das implementações realizadas. Iniciaremos apresentando as ferramentas computacionais que utilizamos para a implementação do algoritmo de Lesk simplificado, pormenorizando os passos do algoritmo e também as formas de avaliação da desambiguação. Passaremos, em seguida, à apresentação do algoritmo de extração de índices de polissemia com base em *corpus*, em sua versão adaptada. Por fim, mostraremos de que forma as medidas de polissemia foram comparadas, explicitando os métodos pelos quais essas medidas foram extraídas e calculadas.

3.1 Ferramentas

Para facilitar o acesso ao *corpus* Mac-Morpho, foi utilizado o NLTK¹ (Natural Language Toolkit), plataforma amplamente empregada na elaboração de programas para o processamento de linguagem natural em Python. O NLTK fornece interface para diversos *corpora* e recursos lexicais, tais como a WordNet de Princeton², além de bibliotecas de processamento textual para classificação de palavras, lematização, extração de raízes das palavras, etiquetagem, análise sintática e raciocínio semântico (*semantic reasoning*).

Os programas de desambiguação e de extração de dados dos dicionários eletrônicos em suas versões *online* foram feitos em Python (versão 2.7.6). As análises estatísticas dos dados obtidos com as desambiguações e os respectivos gráficos dos resultados, apresentados no capítulo 4, foram feitos em R (versão 3.2.1). O programa

¹<http://www.nltk.org/>

²<https://wordnet.princeton.edu/>

de extração das medidas de polissemia e os gráficos e análises dos resultados das extrações foram feitos todos em Python (versão 2.7.6).

3.2 Implementação da desambiguação automática

Apresentaremos, nesta seção, a metodologia referente à desambiguação implementada para substantivos do *corpus* Mac-Morpho a partir da versão adaptada do algoritmo de Lesk simplificado, indicando as palavras escolhidas para serem desambiguadas e o processo de avaliação do desempenho do algoritmo.

3.2.1 Seleção dos lemas

A escolha dos substantivos a serem desambiguados tem como alvo as palavras mais polissêmicas. Restringimos o conjunto de aos cem lemas listados na tabela do apêndice A, com base em suas frequências no *corpus* Subtlex-PT-BR de Tang (2012) e tendo como pressuposto a associação das maiores frequências no léxico a palavras mais polissêmicas (Lee (1990), Crossley *et al.* (2010)).

Idealmente, os cem lemas deveriam ter sido selecionados de acordo com suas frequências de *token* no *corpus* Mac-Morpho — *corpus* em que as desambiguações foram efetuadas. No entanto, o Mac-Morpho não foi lematizado e a contagem das ocorrências acabaria ignorando esse fator. Cogitamos realizar a lematização do Mac-Morpho, tendo em vista tanto as contagens de frequências quanto as desambiguações, que poderiam abarcar flexões dos lemas-alvo. Não obstante, incluir esse procedimento entre as etapas já em andamento poderia atrasar o desenvolvimento dos trabalhos. Outro dado importante que suaviza as consequências dessas escolhas metodológicas é o de que, dos cem lemas selecionados, noventa e cinco estavam entre os duzentos substantivos (incluindo flexões de número e gênero) mais frequentes do Mac-Morpho — que tem um total de 236.462 substantivos, de acordo com o NLTK.

3.2.2 Versão adaptada do algoritmo de Lesk

O método de desambiguação adotado neste trabalho é uma versão simplificada do algoritmo de Lesk. Essa implementação está baseada no que apresentam Jurafsky & Martin (2000). O algoritmo de Lesk simplificado e sua versão, tal como proposta originalmente, estão descritos nos algoritmos 2.2 e 2.3. A versão simplificada recebeu modificações em sua implementação final neste trabalho, visando o refinamento da etapa de seleção de acepções. Essa versão final está apresentada no algoritmo 3.1.

Algoritmo 3.1: Algoritmo de Lesk Simplificado - Versão adaptada**Entrada:** (*palavra*, *sentença*)**Saída:** a melhor acepção de *palavra*

```

1 máx-intersecção  $\leftarrow$  0
2 contexto  $\leftarrow$  conjunto de palavras em sentença
3 para cada acepção em acepções de palavra faça
4   palavras-acepção  $\leftarrow$  conjunto de palavras da acepção
5   idfs-palavras  $\leftarrow$  CALCULAIDFS(palavras-acepção)
6   intersecção  $\leftarrow$  CALCULAIINTERSEC(palavras-acepção, idfs-palavras, contexto)
7   se intersecção > máx-intersecção então
8     máx-intersecção  $\leftarrow$  intersecção
9     melhor-acepção  $\leftarrow$  acepção
10  fim
11 fim
12 retorna melhor-acepção

```

A principal diferença entre o algoritmo de Lesk simplificado e a sua versão adaptada reside no laço em que são calculadas as intersecções entre janelas e acepções: no caso do algoritmo simplificado, os cálculos são feitos entre o conjunto de palavras da janela contextual da palavra-alvo e as palavras contidas em uma determinada acepção dessa palavra; já no caso do algoritmo original, são tomadas as definições de cada palavra contida na janela da palavra-alvo, de forma que as intersecções são calculadas entre essas definições e cada acepção da palavra-alvo.

A versão final do algoritmo de Lesk simplificado (algoritmo 3.1) incluiu o cálculo de pesos para as palavras nas intersecções. Esses pesos foram calculados através das frequências inversas de documentos (*Inverse Document Frequency* — IDF). A frequência de documento de uma palavra, no caso de acepções de dicionário, consiste na contagem das ocorrências dessa palavra em todas as acepções de todas as palavras a serem desambiguadas. Utiliza-se, então, o inverso das frequências de documento para que se reduzam os pesos de palavras pouco distintivas, como as palavras funcionais, e para que palavras que forneçam boas pistas sobre as acepções da palavra-alvo tenham pesos maiores. A distribuição das IDFs pode ser suavizada com a conversão dos seus valores para a escala logarítmica, como sugerem Jurafsky & Martin (2000). No entanto, para os dados em que implementamos o algoritmo de Lesk simplificado, os valores brutos de IDF geraram melhores resultados.

3.2.3 Método para a avaliação da desambiguação

O algoritmo aplicado para as pseudo-palavras segue as etapas descritas na seção anterior: a acepção selecionada será aquela cujas palavras em comum entre dicionário e sentença somem os maiores IDFs, o que garante que a quantidade de pseudo-palavras não obscureça a distintividade das palavras mais significativas das definições.

O levantamento dos pesos das acepções das pseudo-palavras não levou em consideração as distinções internas nas entradas dos lemas. Dessa forma, os resultados das desambiguações tiveram como gabaritos as ocorrências originais dos lemas nas posições em que as pseudo-palavras foram inseridas.

Como explicado na seção anterior, atribuíram-se pesos às palavras das definições fornecidas pelo dicionário através do IDF. No caso das pseudo-palavras, utilizaram-se os mesmo valores de IDF calculados para as desambiguações dos lemas originais, pois não havia indícios de que esses valores sofreriam alterações substanciais.

Ressaltamos que todos os procedimentos de desambiguação descritos foram também realizados com as entradas do Dicionário de Sinônimos. No capítulo 4, estão indicados os resultados das desambiguações e os detalhes das medidas computadas para a avaliação do algoritmo implementado.

Concluída a etapa dos testes de desambiguação, procedemos à etapa de implementação do algoritmo de extração de índices de polissemia e do cálculo do número de acepções com base em *corpus*.

Medidas de desempenho das desambiguações

Tendo em mãos os dados da desambiguação das palavras-alvo e das pseudo-palavras, definiram-se as medidas que serviriam de referência para a avaliação dos resultados. As medidas foram calculadas de acordo com critérios que envolvessem não apenas quantidades de acertos em relação ao total de ocorrências da palavra-alvo, mas que refletissem também outros aspectos, mais específicos das acepções. Por essas razões, selecionaram-se as seguintes medidas para a avaliação dos resultados: grau de confiança, taxa de acertos, cobertura e precisão.

As taxas de acertos foram calculadas com respeito às palavras e consistiram nas razões entre número de desambiguações corretas de uma dada pseudo-palavra pela quantidade de vezes em que ela ocorreu no texto, como indicado na equação 3.1. Em outras palavras, na equação 3.1, r_w (total de acertos) pode ser entendido como a diferença entre n_w (total de ocorrências) e o total de erros.

$$\text{Taxa de acertos} = \frac{r_w}{n_w}, \text{ sendo:} \quad (3.1a)$$

$$w: \text{ palavra-alvo} \quad (3.1b)$$

$$r_w: \text{ total de desambiguações corretas de } w \quad (3.1c)$$

$$n_w: \text{ total de ocorrências de } w \quad (3.1d)$$

Os valores de precisão e de cobertura, por sua vez, são atribuídos especificamente às acepções das palavras desambiguadas. A cobertura, para uma dada acepção, é calculada da seguinte forma: calculou-se a razão do número de vezes em que a acepção foi escolhida corretamente pelo número de vezes em que ela ocorreu no *corpus*. A cobertura é uma medida análoga à taxa de acertos, mas direcionada para cada acepção de uma palavra submetida à desambiguação.

$$\text{Cobertura} = \frac{r_{acep}}{n_{acep}}, \text{ sendo:} \quad (3.2a)$$

$$acep: \text{ acepção de uma palavra-alvo} \quad (3.2b)$$

$$r_{acep}: \text{ total de desambiguações corretas de } acep \quad (3.2c)$$

$$n_{acep}: \text{ total de ocorrências de } acep \quad (3.2d)$$

Os valores de precisão foram calculados como a razão entre o número de acertos de uma acepção e o número total de vezes em que essa acepção foi escolhida. Diferentemente da cobertura, a precisão mede a propensão do método de ter desambiguado corretamente, dada a acepção escolhida. Em conjunto, cobertura e precisão possibilitam avaliar o quanto o método consegue abranger corretamente das ocorrências de uma dada acepção e o quanto a seleção de uma acepção tende a estar correta.

$$\text{Precisão} = \frac{r_{acep}}{e_{acep}}, \text{ sendo:} \quad (3.3a)$$

$$acep: \text{ acepção de uma palavra-alvo} \quad (3.3b)$$

$$r_{acep}: \text{ total de desambiguações corretas de } acep \quad (3.3c)$$

$$e_{acep}: \text{ total de vezes em que } acep \text{ foi escolhida} \quad (3.3d)$$

Para atestar a confiança nas desambiguações realizadas, os pesos das acepções selecionadas foram confrontados com as diferenças entre esses pesos e os pesos das acepções que ficaram em segundo lugar na seleção (equação 3.4).

$$\text{Grau de confiança} = f_{acep}(acep_1) \quad (3.4a)$$

$$f_{acep}(acep_1) = W_{acep}(acep_1) - W_{acep}(acep_2), \text{ sendo:} \quad (3.4b)$$

$$acep_1: \text{acepção escolhida} \quad (3.4c)$$

$$acep_2: \text{acepção que ficou em segundo lugar} \quad (3.4d)$$

$$W_{acep}: \text{função que atribui peso a } acep \quad (3.4e)$$

Estabelecemos que essa seria a medida de confiança adotada em nossas análises. A ideia dessa comparação é a de que não bastaria que a acepção escolhida tivesse um peso alto para que a desambiguação fosse inequívoca, seria necessário também que houvesse uma distintividade considerável entre as acepções concorrentes com os maiores pesos.

3.3 Algoritmo de extração dos índices de polissemia

O algoritmo proposto por Sproat & VanSanten consiste essencialmente em agrupar as palavras co-ocorrentes à palavra-alvo que sejam mais relevantes e, a partir desses agrupamentos, realizar regressões sobre os pontos que expressam os tamanhos desses agrupamentos, fornecendo como saída (isto é, como o índice de polissemia) a soma dos quadrados dos erros minimizados pelas regressões.

3.3.1 Versão adaptada do algoritmo de extração de índices de polissemia

Para implementar a versão que adaptamos do algoritmo, foram feitas algumas modificações especialmente no que diz respeito aos critérios de seleção das palavras relevantes — palavras com número mínimo de ocorrências no *corpus* e que superaram um limiar de co-ocorrências com a palavra-alvo —, ao cálculo da matriz de co-ocorrentes D_w (chamada de matriz de distâncias no trabalho original) e ao cálculo dos *clusters* que corresponderão às acepções da palavra-alvo.

Fizemos diversos testes em nossa versão, procurando o valor que mais equilibrasse a quantidade de palavras relevantes e os resultados da regressão e da clusterização,

de forma que a filtragem de palavras relevantes não fosse muito restritiva (impedindo que selecionássemos um número mínimo de palavras relevantes), ou muito permissiva (levando à inclusão de palavras pouco informativas). Modificamos, então, os dois fatores de filtragem das palavras relevantes: o limiar de ocorrências (equação 3.5b), que passa a exigir que as palavras relevantes tenham ocorrido uma porcentagem mínima de vezes no *corpus* que superasse o número mínimo de duas ocorrências absolutas, como proposto originalmente (equação 2.1b); e um limiar ótimo de co-ocorrências, que estabelecemos experimentalmente, com um número ainda restrito de testes, como sendo de 0.2 (isto é 20 %) — contrastando com os 0.005 do trabalho original (equação 2.1b). Esses aspectos voltarão a ser discutidos no capítulo 4, de resultados.

$$FREQ_w(q)/FREQ_C(q) > \text{limiar} \quad (3.5a)$$

$$FREQ_C(q) > \text{limiar de ocorrências percentual} \quad (3.5b)$$

Quanto ao cálculo da matriz de palavras co-ocorrentes, entendemos que, se a matriz for interpretada como uma matriz de distâncias, isto é, subtraindo-se de 1 (distância máxima) a média das probabilidades condicionais de co-ocorrência, os valores da matriz se aproximariam de zero para palavras muito próximas e aumentaria para as palavras mais distantes. No entanto, o método de decomposição que escolhemos (SVD) é considerado mais apropriado para valores que expressem similaridade³ ou proximidade entre palavras (Deerwester *et al.*, 1990). Pensando de outra forma: se sabemos que duas palavras distam de uma terceira, essas duas palavras podem estar tanto próximas quanto distantes uma da outra; porém, se sabemos que duas palavras estão próximas de uma terceira, saberemos que ambas estão minimamente próximas entre si também. Por essa razão — de os valores de similaridade serem mais informativos — adotamos o cálculo de similaridade em lugar do cálculo das distâncias

³É provável que o método de redução de dimensionalidade adotado no trabalho original tenha privilegiado os autovalores mais baixos da decomposição da matriz, enfatizando a discretização das distâncias entre os grupos de palavras.

para computar a matriz P_w de relações entre os co-ocorrentes, como indicado na equação 3.6a.

$$SIM_{ij}(q, q') = \frac{P(q|q') + P(q'|q)}{2}, \text{ onde } q, q' \in REL_w \quad (3.6a)$$

$$\text{Sendo: } P(q|q') = \frac{CO_k(q|q')}{Freq_C(q)} \quad (3.6b)$$

Uma modificação metodológica importante que implementamos diz respeito ao agrupamento das palavras co-ocorrentes em conjuntos de quatro graus. Uma vez que a mudança de coordenadas cartesianas para coordenadas polares é feita, as palavras são agrupadas em faixas de quatro em quatro graus, no intervalo $[0, 360^\circ[$, totalizando noventa subdivisões. No algoritmo original, faz-se um histograma, isto é, as alturas das barras são dadas diretamente pela quantidade de palavras contidas em cada barra. Em nossa versão, o tamanho das barras do histograma é calculado pela contagem ponderada pelos raios de cada palavra de uma dada faixa. Dessa forma, procuramos atribuir pesos para os agrupamentos que destacassem mais claramente as palavras mais distintivas.

Por fim, a clusterização desses dados diferiu da solução sugerida originalmente com relação à especificação dos critérios para a formação dos *clusters*, de forma que obtivéssemos uma aproximação para o que seriam as acepções da palavra-alvo. Consideramos que apenas os agrupamentos (barras) com valor mínimo de 0.2 (em uma escala normalizada de 0 a 1) poderiam ser considerados picos e formar *clusters* em torno de si e que apenas as palavras de agrupamentos não nulos seriam clusterizadas. Além disso, os agrupamentos, para serem considerados distintos, deveriam estar a uma distância mínima, em graus; caso contrário, independente dos seus tamanhos, esses agrupamentos vizinhos comporiam o mesmo *cluster*.

As medidas de polissemia extraídas do *corpus* (índice de polissemia e número de *clusters*) serão analisadas e comparadas para 50 substantivos dentre os mais frequentes do *corpus* Mac-Morpho, em suas formas lematizadas, isto é, as formas flexionadas dos substantivos foram descartadas da lista (ordenada pelas frequências) de todos os *tokens* de substantivos encontrados no *corpus*. A lista desses 50 substantivos está na tabela do apêndice C, com as palavras acompanhadas pelos respectivos índices de polissemia, números de *clusters*, frequências absolutas no *corpus* e quantidades de acepções listadas no dicionário Michaelis.

Algoritmo 3.2: Algoritmo de extração do índice de polissemia

Entrada: (*palavra-alvo*, *texto*)**Saída:** *índice de polissemia da palavra-alvo*
número de acepções da palavra-alvo

```

1 janelas contextuais ← EXTRAIJANELAS(palavra-alvo, texto)
2 conjunto de co-ocorrentes ← SET(janelas contextuais)
3 palavras relevantes ← FILTRARELEVANTESMODIFICADA(conjunto de co-ocorrentes,
janelas contextuais)
4 para cada palavra em palavras relevantes faça
5   janelas da palavra ← EXTRAIJANELAS(palavra)
6   para cada co-ocorrente em palavras relevantes faça
7     i ← INDEX(palavra, palavras relevantes)
8     j ← INDEX(co-ocorrente, palavras relevantes)
9     matriz de co-ocorrentes[i, j] ← CALCULASILARIDADE(co-ocorrente,
janelas da palavra)
10  fim
11 fim
12 U, S, V ← DECOMPOSICAOSVD(matriz de co-ocorrentes)
13 matriz_reduzida ←  $U[0:n_w, 0:2] \times S[0:2, 0:2]$ 
14 matriz_polar ← COORDENADASPOLARES(matriz_reduzida)
15 barras_4graus ← REAGRUPACOOCORRENTESPONDERANDO(matriz_polar)
16 índice de polissemia ← REGRESSAOISOTONICA(barras_4graus)
17 clusters ← CLUSTERIZAPORPICOSMODIFICADA(barras_4graus)
18 número de acepções ← ContaClusters(clusters)
19 retorna índice de polissemia, número de acepções

```

4

Resultados e análises

Neste capítulo, apresentamos os resultados da implementação da versão simplificada do algoritmo de Lesk para a desambiguação automática de substantivos do *corpus* Mac-Morpho e os resultados da extração de medidas de polissemia a partir de substantivos desse mesmo *corpus*. Apresentamos também os resultados da avaliação de desempenho da desambiguação automática através do método de desambiguação de pseudo-palavras. Ao final, discutiremos as correlações entre a implementação da desambiguação e a extração de medidas de polissemia, utilizando dados de um exemplo de palavra-alvo submetido aos dois métodos.

4.1 Frequências de documento

Com base no que sugerem Jurafsky & Martin (2000), recorreremos às contagens das frequências inversas de documentos (IDF) para atribuir pesos às palavras das acepções do dicionário para cada palavra-alvo e, conseqüentemente, para selecionar as acepções mais adequadas nas desambiguações. As IDFs são utilizadas para garantir a redução dos pesos de palavras funcionais e conferir pesos maiores às palavras mais distintivas das definições fornecidas pelo dicionário.

No gráficos 4.1 e 4.2, a seguir, estão representadas as distribuições das palavras das entradas dos dicionários de acordo com suas frequências de documento (DFs). As DFs das palavras do dicionário Michaelis estão apresentadas na figura 4.1 em escala logarítmica, para que se possa visualizar melhor a curva da distribuição; a distribuição das DFs pelo Dicionário de Sinônimos está representada pela figura 4.2. O conjunto de palavras distintas encontradas na união de todas as definições dos cem

Figura 4.1: Logaritmos das DFs dos lexemas contidos nas definições do Michaelis. A distribuição se assemelha à curva prevista por Zipf (1945), com grande número de palavras de baixa frequência e poucas de frequências mais altas.

lemas pelo Michaelis foi de 2683. No Dicionário de Sinônimos, foram encontrados 2396 sinônimos distintos.

Para as desambiguações, foram utilizadas as frequências inversas de documento (IDFs) sem a aplicação da função logarítmica. O uso do logaritmo das IDFs, sugerida por Jurafsky & Martin (2000), é normalmente feito para suavizar as diferenças entre os valores de IDF. No entanto, sua aplicação, favorecendo palavras menos distintivas, comprometeria ainda mais algumas desambiguações: em diversos casos, notamos que a desambiguação se apoiou, essencialmente, na intersecção de palavras funcionais entre definições e janelas (ver gráficos 4.7, 4.8). Nesses casos, sem o logaritmo das IDFs, as acepções selecionadas acabaram sendo penalizadas com medidas mais baixas de confiança, isto é, com pesos finais mais baixos do que os que seriam obtidos com a suavização logarítmica.

Figura 4.2: Distribuição das DFs de todos os lexemas das definições dos cem lemas pelo dicionário de sinônimos. A distribuição mais esparsa se deve à quantidade de sinônimos nas entradas de cada lema.

4.2 Dados obtidos com as desambiguações

Nesta seção, apresentamos os resultados obtidos com a implementação da versão simplificada do algoritmo de Lesk, tanto para os dados reais, quanto para as pseudo-palavras utilizadas no método de avaliação da desambiguação. Os resultados, de acordo com a avaliação pelo método das pseudo-palavras, superaram o método *baseline*. Particularmente, no caso das pseudo-palavras formadas ao acaso, o *baseline* atingiu 50% de taxa de acertos, enquanto a desambiguação baseada no algoritmo de Lesk chegou a aproximadamente 64%. No entanto, notamos que uma parcela substancial dos resultados dos dados reais contou apenas com lexemas pouco distintivos para realizar a desambiguação.

4.2.1 Desambiguação dos lemas mais frequentes

A principal medida adotada para a avaliação das desambiguações dos lemas mais frequentes foi o grau de confiança. O desempenho nas desambiguações dos lemas mais frequentes foi também estimado com base nos resultados obtidos com as desambiguações das pseudo-palavras.

As colunas de maiores densidades de pontos nos gráficos 4.3 e 4.4 devem-se às desambiguações que foram apoiadas por valores de IDF de lexemas mais distintivos — à semelhança do que ocorreu ns gráficos de confiança das desambiguações das pseudo-palavras e que também se traduziu nos gráficos de densidade de janelas (4.13, 4.14).

O gráfico da confiança na desambiguação dos lemas a partir das acepções do Michaelis (figura 4.3) indica maior dispersão dos valores confrontados. O resultado difere dos padrões obtidos para as pseudo-palavras pelo fato de a desambiguação real se basear em números de acepções variáveis para cada palavra. Para as pseudo-palavras, as acepções eram mais longas (por conta da concatenação das acepções dos lemas) e sempre dadas aos pares, correspondendo a cada um dos lemas pareados. Acepções mais longas, de acordo com o que relata Lesk (1986) em seu trabalho, tendem a gerar melhores resultados. Isso acontece porque há um maior número de palavras, inclusive as significativas, de conteúdo, compondo essas acepções.

O gráfico 4.5 confronta os números de acepções dos lemas desambiguados com as quantidades de campos semânticos dos mesmos. Verifica-se a tendência de uma relação crescente entre as medianas dos números de acepções e o aumento no número de campos semânticos. Em outros termos, pretendia-se verificar com essas informações se havia relação proporcionalmente direta entre polissemia e número

Figura 4.3: Confiança na desambiguação: medida pelos pesos das acepções dos Michaelis, contrastados com as distâncias entre o primeiro e o segundo colocado na seleção das acepções. A reta em vermelho indica regressão com ajuste de cerca de 78%.

Figura 4.4: Confiança na desambiguação: medida pelos pesos dos campos semânticos, contrastados com as distâncias entre o primeiro e o segundo colocado na seleção dos campos. A reta em vermelho indica regressão com ajuste de cerca de 94%.

Figura 4.5: Comparação dos números de acepções com as quantidades de campos semânticos para os cem lemas desambiguados.

de campos semânticos. No gráfico mencionado (4.5), em particular, os lemas com mais campos semânticos apresentam variação considerável entre seus números de acepções, como se pode verificar pelo aumento dos quartis e das distâncias entre os valores de mínimo e de máximo.

Nos gráficos da figura 4.6, as distribuições dos lemas por número de acepções e por campos semânticos são mostrados mais detalhadamente. A distribuição para o número de acepções apresenta maior concentração de dados, como esperado, para os valores mais baixos: traduzindo o dado do número de acepções em medida para a polissemia, temos que as palavras mais polissêmicas estão distribuídas em número menor. Iniciamos a exploração desses aspectos na segunda fase desta pesquisa, através da extração e análise de medidas de polissemia, cujos resultados serão apresentados na seção 4.3.

Os gráficos apresentados nas figuras 4.7 e 4.8 indicam as distribuições das janelas contextuais dos lemas pelos pesos das acepções selecionadas do Michaelis e pelos pesos dos sinônimos, respectivamente. As distribuições se assemelham, apontando maior concentração dos dados para os pesos abaixo dos IDFs considerados distintivos de significado. Esses resultados indicam que as desambiguações dos lemas estão se apoiando, majoritariamente, em lexemas pouco distintivos, como palavras funcionais.

Figura 4.6: Distribuição dos lemas por números de acepções pelo Michaelis (a) e por números de campos semânticos dados pelo Dicionário de Sinônimos (b).

Figura 4.7: Distribuições das quantidades de janelas contextuais dos lemas pelos pesos das acepções escolhidas do Michaelis.

Figura 4.8: Distribuições das quantidades de janelas contextuais dos lemas pelos pesos dos campos semânticos selecionados do Dicionário de Sinônimos.

4.2.2 Desambiguação das pseudo-palavras

A partir dos dados das desambiguações das pseudo-palavras, avaliamos o desempenho do algoritmo e verificamos se a proporção das acepções de um dado lexema no *corpus* teria impacto sobre os resultados do algoritmo aplicado.

Observando-se as taxas de acertos dos três grupos de pseudo-palavras, obtêm-se as médias de taxas de acertos apresentadas na tabela 4.1.

De maneira geral, os resultados das desambiguações com o dicionário Michaelis foram mais satisfatórios, apresentando taxas de acertos maiores e desvios menores.

Dicionário	Grupo A	Grupo B	Grupo C
Michaelis	0.544 (± 0.045)	0.671 (± 0.204)	0.638 (± 0.185)
Sinônimos	0.540 (± 0.051)	0.504 (± 0.248)	0.522 (± 0.231)

Tabela 4.1: Médias das taxas de acertos das desambiguações para os três grupos de pseudo-palavras. As taxas de acertos estão acompanhadas dos respectivos desvios-padrão.

Figura 4.9: Distribuições das taxas de acertos (acertos por total de ocorrências da pseudo-palavra) das desambiguações das pseudo-palavras do grupo A (a), do grupo B (b) e do grupo C (c), a partir das definições do Michaelis.

Figura 4.10: Distribuições das taxas de acertos (acertos por total de ocorrências da pseudo-palavra) das desambiguações das pseudo-palavras do grupo A a partir do Dicionário de Sinônimos.

Os histogramas das figuras 4.9 e 4.10 mostram, graficamente, os dados sobre os quais foram calculados os valores da tabela 4.1 acima.

Os gráficos das figuras 4.11 e 4.12 mostram a dispersão dos valores de precisão para as desambiguações dos três grupos de pseudo-palavras com cada dicionário. As pseudo-palavras do grupo A apresentam comportamento mais similar entre as aceções: há uma concentração de dados próximos à media. Para os outros grupos, há maior dispersão dos valores de precisão das aceções. Como indicam os gráficos de precisão, esse padrão foi semelhante para os dois dicionários.

Analisamos também as distribuições das janelas contextuais pelos pesos das aceções escolhidas. Na figura 4.13, as curvas das distribuições se mostram semelhantes e podem ser observados dois picos: um primeiro, que concentra as janelas que foram desambiguadas com base em intersecções que continham palavras de baixo IDF; e

Figura 4.11: Distribuições dos valores de precisão (acertos por total de tentativas) das desambiguações das pseudo-palavras do grupo A (a), do grupo B (b) e do grupo (c), a partir das definições do Michaelis.

Figura 4.12: Distribuições dos valores de precisão (acertos por total de tentativas) das desambiguações das pseudo-palavras do grupo A (a), do grupo B (b) e do grupo C (c) a partir do Dicionário de Sinônimos.

Figura 4.13: Distribuição das quantidades de janelas pelos pesos das acepções selecionadas do Michaelis.

Figura 4.14: Distribuição das quantidades de janelas pelos pesos dos campos semânticos selecionados do Dicionário de Sinônimos.

um segundo, que indica os casos em que a desambiguação se apoiou em um lexema altamente distintivo. O máximo peso (IDF) calculado para os lexemas foi de 1068 para o Michaelis e de 584 para o Dicionário de Sinônimos. Esses valores sinalizam as palavras que ocorreram uma só vez no total das definições.

Os gráficos 4.15 e 4.16 apresentam as distribuições das janelas distinguidas para o dicionário Michaelis e para o de sinônimos, respectivamente. No caso do primeiro dicionário, as curvas de acertos e de erros para os três grupos de pseudo-palavras mostraram-se semelhantes, diferindo, essencialmente, nas alturas dos picos de cada gráfico. Para as pseudo-palavras dos grupos B e C, os erros foram maiores para as desambiguações baseadas em pesos baixos das acepções. Com os lexemas do Dicionário de Sinônimos, as desambiguações dos grupos B (4.16b) e C (4.16c) mostraram resultados melhores, isto é, menos janelas foram desambiguadas equivocadamente e as curvas de erro apresentaram comportamento bastante irregular.

Figura 4.15: Distribuição das quantidades de janelas pelos pesos das aceções selecionadas do Michaelis. As linhas em azul e vermelho indicam as distribuições das janelas para os casos em que as aceções escolhidas estavam, respectivamente, corretas e incorretas.

Figura 4.16: Distribuição das quantidades de janelas pelos pesos dos campos semânticos selecionados. As linhas em azul e vermelho indicam as distribuições das janelas para os casos em que os campos escolhidos estavam, respectivamente, corretos e incorretos.

Figura 4.17: Confiança na desambiguação: medida pelos pesos das aceções do Michaelis, contrastados com as distâncias entre o primeiro e o segundo colocado na seleção das aceções. As retas em vermelho indicam regressões com ajustes de cerca de 89% (a), 90% (b) e 90% (c).

Os gráficos das figuras 4.17 e 4.18 mostram as dispersões dos pontos para a medida de confiança nas desambiguações. As retas nos gráficos indicam regressões lineares que formalizam a tendência dos dados nos gráficos: quanto maior o peso da aceção escolhida, maior tende a ser a diferença de peso entre o primeiro e o segundo colocado.

Diferentemente dos valores de precisão, as taxas de cobertura foram mais dispersas para todos os grupos, inclusive, o grupo A — tanto para as desambiguações feitas com o Michaelis, quanto para aquelas feitas com os sinônimos (figuras 4.19 e 4.20).

4.2.3 Discussão dos resultados da desambiguação

As taxas de acertos médias obtidas nas desambiguações com as pseudo-palavras superaram o *baseline* de aceção mais frequente, mesmo considerando-se os desvios-padrões, atingindo, em alguns casos, até 67% de acertos. No caso das pseudo-palavras formadas aleatoriamente, o *baseline* obteve 50% de acertos, enquanto a versão que implementamos do algoritmo de Lesk chegou a 64%.

Considerando-se as desambiguações dos 100 lemas, observamos que os graus de confiança para esses dados seguiram a tendência dos dados das pseudo-palavras: pesos maiores para as aceções escolhidas estariam mais associados a pesos menores para as aceções que ficaram em segundo lugar. O desempenho do algoritmo para essas palavras só pode ser estimado a partir dos dados obtidos com as pseudo-palavras, com a reserva de que as desambiguações das pseudo-palavras tinham a particularidade de serem sempre entre duas ‘aceções’, como explicado na seção 2.3.3.

Figura 4.18: Confiança na desambiguação: medida pelos pesos dos campos semânticos, contrastados com as distâncias entre o primeiro e o segundo colocado na seleção dos campos. A reta em vermelho indica regressão com ajuste de cerca de 93%, 98% e 97% para os grupos A (a), B (b) e C (c), respectivamente.

Figura 4.19: Cobertura (acertos da aceção por ocorrências da mesma) para a desambiguação baseada nas aceções do Michaelis.

Figura 4.20: Cobertura (acertos dos campos por ocorrências dos mesmos) para a desambiguação baseada no Dicionário de Sinônimos.

A implementação da versão simplificada do algoritmo de Lesk obteve melhores taxas de acertos para as avaliações feitas com pseudo-palavras do grupo B, o que pode estar relacionado à disparidade entre as frequências dos lemas que compõem essas pseudo-palavras. As análises das distribuições das janelas por acertos e erros do algoritmo apontou menos erros para pseudo-palavras dos grupos B e C desambiguadas com base nos sinônimos. O grupo C, formado aleatoriamente, obteve bons resultados em geral, por oposição aos resultados gerais do grupo A, que foram os menos expressivos. Esses resultados podem indicar que os vieses dos grupos de pseudo-palavras A e B afetam o desempenho desambiguação. A expectativa que temos é de que os cem lemas que foram desambiguados apresentem todas as acepções predominantes, o que os aproximaria do grupo B de pseudo-palavras. Considerando as limitações para a avaliação do método que implementamos, não foi possível verificar as distribuições das acepções das palavras-alvo no *corpus*.

4.3 Medidas de polissemia obtidas

Comparando-se as medidas de polissemia extraídas do *corpus* com as frequências dos substantivos, não obtivemos correlações entre esses dados, como pode ser observado no gráfico 4.21. O mesmo foi observado para a comparação entre as quantidades de acepções do Michaelis e as frequências dos substantivos no *corpus* (gráfico 4.22). O primeiro resultado contrariou as expectativas que tínhamos em relação aos dados, uma vez que esperávamos forte correlação positiva entre os índices de polissemia e as frequências no *corpus* — hipótese baseada em Miller & Fellbaum (1991). Acreditamos que esses dois resultados possam ser explicados pela pequena quantidade de substantivos testados e pelo fato de o *corpus* utilizado ser pouco representativo em termos de suas dimensões e da variedade de assuntos que abrange.

Não foi observada também correlação entre as medidas de polissemia extraídas do *corpus* e as quantidades de acepções do Michaelis (gráfico 4.23). Esse resultado era esperado, dado que o *corpus* que utilizamos possui um viés jornalístico mais restrito a assuntos político-econômicos e que o dicionário, por sua vez, faz uma cobertura mais abrangente, procurando contemplar toda a diversidade das acepções sincronicamente possíveis de um dado lexema na língua, incluindo casos em desuso.

Comparando-se as medidas de polissemia extraídas do *corpus* entre si, notamos uma aproximação positiva entre as variações das duas medidas, como indicado nos gráficos 4.24 e 4.25. Nos dois gráficos, os dados foram representados com os substantivos ordenados de acordo com seus índices de polissemia, a partir dos

Figura 4.21: Comparação das medidas de polissemia: índice de polissemia, quantidade de *clusters* discretizados e frequência. As palavras estão em ordem decrescente de frequência no *corpus*.

Figura 4.22: Comparação das medidas de polissemia: quantidade de acepções no dicionário Michaelis e frequência no *corpus*. As palavras estão em ordem decrescente de frequência no *corpus*.

Figura 4.23: Comparação entre números de *clusters* e números de acepções

Métricas de polissemia baseadas no corpus para 50 substantivos

Figura 4.24: Raízes dos quadrados das diferenças entre as medidas de índice de polissemia e número de *clusters*. As palavras estão em ordem decrescente pelo índice de polissemia.

maiores valores. O segundo gráfico apresenta os módulos das diferenças entre as duas medidas de polissemia extraídas do *corpus*, apontando uma tendência de maior diferença entre essas medidas para os lexemas menos polissêmicos. No gráfico 4.24, notamos que se atribui 0.3 *clusters* (na escala normalizada) para a maior parte dos substantivos, o que corresponde a 2 *clusters*. Os únicos lexemas marcados como monossêmicos foram *página* e *editoria*. É natural que tais palavras apareçam como monossêmicas quando usadas em um jornal. Isso vem reforçar, aliás, a hipótese de que o viés temático (de texto jornalístico de esfera político-econômica, no caso) seja responsável pela redução polissêmica dos substantivos analisados. Na seção 4.3.1, analisaremos com maior profundidade os dados de alguns dos substantivos testados.

Métricas de polissemia baseadas no corpus para 50 substantivos

Figura 4.25: Raízes dos quadrados das diferenças entre as medidas de índice de polissemia e número de *clusters*. As palavras estão em ordem decrescente de frequência no *corpus*.

4.3.1 Análises dos pesos dos *clusters*: alguns exemplos

Os gráficos 4.24 e 4.25 apresentam os dados que forneceram os principais critérios para a escolha dos substantivos que seriam analisados mais minuciosamente, para servir de ilustração dos procedimentos e para permitir discussões de detalhes específicos. Os lexemas selecionados foram: *mercado*, *produção*, *página* e *casa*. Procuramos, com esses exemplos, contemplar a variedade de resultados e fenômenos observados em nossa amostra de 50 substantivos, observando as possibilidades de análises que o método implementado proporciona, assim como as limitações observadas através dos testes que realizamos.

Mercado

Os co-ocorrentes relevantes do substantivo *mercado* apresentaram a distribuição exibida no gráfico 4.26. Percebe-se que há uma concentração de palavras próximas à origem e algumas poucas que se distanciam desse aglomerado em duas direções opostas. Para outros substantivos, observou-se fenômeno semelhante: aglomeração de diversas palavras próximas à origem e alguns poucos determinando tendências de distanciamento. A interpretação para os aglomerados em torno do ponto (0, 0) vem do fato de que esses lexemas, apesar de terem superado os critérios de relevância para os co-ocorrentes, apresentaram pouca similaridade entre si. A mesma distribuição, em coordenadas polares e contendo as palavras correspondentes a cada ponto, pode ser verificada no gráfico 4.28. Nessa figura, as distinções entre os co-ocorrentes de *mercado* podem ser vistas mais claramente.

Mercado apresentou o maior índice de polissemia e a maior quantidade de *clusters* (um total de 4) dentre os 50 substantivos testados. Não obstante, é possível avaliar a relação entre esses *clusters* e observar as palavras contidas em cada um para melhor analisar os resultados da clusterização.

Os *clusteres* de palavras associadas à *mercado* no *corpus* estão listados abaixo:

1. vencem
2. excedente, taxas, provocam, elevada, financeiro, vendedores, detém, juros, investidores, analistas, projeção, acionário, ágio
3. externo, futuros, projetando
4. interno, cotações, cotação, fecharam, cambial, escassez, desvalorização

A localização dos *clusters* no histograma da figura 4.27 pode ser feita cruzando-se os dados dos ângulos da tabela 4.2 com os do eixo *x* do gráfico.

Figura 4.26: Representação das co-ocorrências de “mercado” em coordenadas cartesianas.

<i>cluster</i>	peso	ângulos (graus)
1	0.60	-54 a -52
2	2.87	-48 a -8
3	0.61	-4 a 4
4	2.53	48 a 68

Tabela 4.2: Tabela para localização dos *clusters* de *mercado* no gráfico 4.27. Os *clusters* estão numerados de acordo com as listas de palavras dos *clusters* apresentadas anteriormente.

Nota-se que há barras que não compõem nenhum *cluster*, como a barra próxima dos 25 graus no gráfico 4.27. Em geral, essa situação pode ocorrer quando a barra está isolada (não tendo barras adjacentes dentro de uma faixa de até 16 graus para a direita e para a esquerda), não podendo ser agrupada a nenhum *cluster*, ou quando não atinge o valor mínimo para a formação de um *cluster* (que estabelecemos como 0.25, na escala normalizada).

A tabela 4.2 indica os *clusters* 2 e 4 como os mais distintos. Como se pode observar no gráfico 4.28, as principais palavras responsáveis pelos pesos desses dois *clusters* são *taxas* e *juros*, para o *cluster* 2, e *desvalorização* e *cambial*, para o *cluster* 4. Interpreta-se que, quanto mais afastadas da origem em relação ao eixo *x* (raios), mais distintivos

Figura 4.27: Distribuição dos co-ocorrentes de *mercado* em agrupamentos de 4 graus. O eixo x , correspondente aos ângulos, foi representado no intervalo entre -180 e 180 para que os co-ocorrentes próximos à origem (que representam a maior parte dos dados) fossem mostrados no centro do gráfico. Os pesos correspondem às somas normalizadas das distâncias radiais de cada co-ocorrente que compõe uma barra (agrupamento de 4 graus).

Figura 4.28: Co-ocorrências de *mercado* em coordenadas polares.

e relevantes serão as palavras. As distâncias entre os *clusters* podem ser melhor observadas no gráfico 4.29.

De acordo com o dicionário Michaelis, a palavra *mercado* é definida da seguinte forma:

1. Lugar público onde se compram mercadorias postas à venda.
2. Ponto onde se faz o principal comércio de certos artigos.
3. Centro de comércio.
4. O comércio.
5. *Econ.* Esfera das relações econômicas de compra e venda, de cujo ajuste resulta o preço.
6. *Econ.* Meio onde certos produtos são aceitos; centro de comércio.

No entanto, ao contabilizar a quantidade de acepções dicionarizadas dos substantivos testados, consideramos apenas as acepções numeradas e ignoramos as colocações. Nesse caso, a quantidade de acepções de *mercado*, de acordo com o Michaelis, foi de 6.

Nota-se que a acepção de *mercado* enquanto *lugar público onde se compram mercadorias* (relacionada às 3 primeiras acepções pelo Michaelis) não foi contemplada pelos dados do *corpus* — muito provavelmente por conta do viés político-econômico do mesmo e, possivelmente, em razão de suas dimensões.

Co-ocorrências de "mercado" em clusteres de 4°

Figura 4.29: Distribuição dos co-ocorrentes de *mercado* em agrupamentos de 4 graus.

As regressões que geraram o índice de polissemia de *mercado* estão representadas no gráfico 4.30. Como se pode verificar, a existência de diversos picos altos e distantes forçou a regressão a priorizar o pico maior, minimizando grosseiramente os erros em relação aos demais picos. Por essa razão, o índice de polissemia de *mercado* acabou sendo alto em relação aos índices dos demais substantivos testados. As distâncias entre os picos e os seus altos pesos evidenciam o fato de o método implementado ter associado 4 *clusters* à palavra *mercado*. Por outro lado, como observado no gráfico 4.29, os *clusters* mantêm alguma proximidade em termos dos ângulos que os distanciam, o que parece estar em concordância com o fato de que os lexemas dos *clusters* distinguidos guardam um nível de associação semântica que os aproxima mais do fenômeno da polissemia do que da homonímia.

Figura 4.30: Regressão isotônica e antitônica sobre os dados do histograma dos co-ocorrentes de *mercado*.

Figura 4.31: Representação das co-ocorrências de *produção* em coordenadas polares.

Produção

O substantivo *produção* obteve índice de polissemia de pouco mais de 0.2 na escala normalizada para os 50 substantivos testados. Esse valor, como pode ser verificado no gráfico 4.24 posicionou a palavra acima, apenas, das 10 palavras com os menores índices, aproximando-o mais, portanto, das palavras mais próximas da monossemia. No entanto, como pode ser observado no gráfico 4.31, há três grupos de palavras que se distanciam consideravelmente do valor zero no eixo dos raios.

Como resultado da clusterização que realizamos, obtivemos 2 *clusters* distintos para *produção*, listados a seguir. A tabela 4.3 indica os ângulos pelos quais os *clusters* podem ser localizados no histograma 4.32.

1. hispânicos, literária, étnicos
2. cana

Com a representação em *polar plot*, circunferência trigonométrica, observamos, mais claramente, a diferença entre os ângulos dos dois *clusters* distinguidos para *produção*: 90 graus. Esse resultado indica uma diferenciação considerável entre os *clusters*, o que segue a expectativa de distinção clara entre os mesmos, sem se

<i>cluster</i>	peso	ângulos (graus)
1	1.0	-106
2	0.68	0

Tabela 4.3: Tabela para localização dos *clusters* de *produção* no gráfico 4.32. O eixo x, correspondente aos ângulos, foi representado no intervalo entre -180 e 180 para que os co-ocorrentes próximos à origem (maior parte dos dados) fossem mostrados no centro do gráfico. Os pesos correspondem às somas, normalizadas, das distâncias radiais de cada co-ocorrente que compõe uma barra (agrupamento de 4 graus).

Figura 4.32: Distribuição dos co-ocorrentes de *produção* em agrupamentos de 4 graus.

Co-ocorrências de "produção" em clusteres de 4°

Figura 4.33: Distribuição dos co-ocorrentes de *produção* em agrupamentos de 4 graus.

configurar como o extremo da diferenciação (180 graus), que corresponderia a casos inequívocos de homonímia¹.

A seguir, estão apresentadas as acepções da palavra *produção* enumeradas pelo dicionário Michaelis. Identificando-se, manualmente, essas acepções listadas com os *clusters* distinguidos a partir dos dados do *corpus*, verificamos a relação entre o *cluster* 1 e a acepção 3 do Michaelis, e entre o *cluster* 2 e, provavelmente, as acepções 1, 2, 4 e 5.

1. Ato ou efeito de produzir; fabricação, manufatura; extração; geração.
2. Coisa produzida naturalmente ou pelo trabalho; produto.
3. Obra literária ou artística.
4. Quantidade produzida por unidade produtora ou unidade de tempo; rendimento: Produção de uma máquina; produção por hectare; produção horária; produção anual.
5. Quantidade local produzida (falando-se de qualquer mercadoria): A produção de arroz do Brasil excede o consumo.

¹De acordo com o método implementado, homonímia ocorreria nos casos em que os lexemas co-ocorrentes à palavra-alvo se agrupariam em *clusters* bastante distintivos (picos altos) e distantes. Notamos que a distância entre esses *clusters* se torna maior quanto menos relacionados estiverem os grupos de palavras que compõem cada *cluster*.

6. *Econ. polít.* Obtenção de bens comerciais, quer matérias-primas (agricultura, mineração, pesca etc.), quer manufaturas destas, com destino ao mercado e ao próprio consumo.
7. *Dir.* Ação de produzir ou de apresentar testemunhas, títulos ou documentos etc., em um processo.
8. *Anat.* Prolongamento.
9. *Med.* Desenvolvimento de um sintoma mórbido.
10. *Med.* O próprio sintoma mórbido.
11. *Med.* Aparecimento de um fenômeno mórbido.
12. *Med.* Hipertrofia de um tecido normal.

De todo modo, as correlações sugeridas podem ser facilmente questionadas, dado que os *clusters*, especialmente o de *cana*, foram formados por pouco lexemas, fornecendo poucas pistas contextuais para a determinação das acepções correspondentes a cada *cluster*. Além disso, algumas definições do dicionário são vagas ou mesmo contêm diversas palavras polissêmicas. A título de exemplo, a acepção 2 do Michaelis *Coisa produzida naturalmente ou pelo trabalho; produto*. contém os lexemas *coisa* e *trabalho* e poderia ser relacionada à terceira acepção se fosse entendida como *obra produzida pelo trabalho artístico*.

Página

O lexema *página* foi um dos dois únicos apontados como monossêmicos na clusterização. Apesar disso, é a palavra mais frequente do *corpus*, juntamente com *editoria*. É possível que *página* seja recorrente no *corpus* por conta do viés do mesmo — situação semelhante se daria para *editoria* —, sendo repetidamente utilizado como referência às páginas do próprio jornal. Todavia, apresenta-se o exemplo de *página* essencialmente para discutir o caso de monossemeia e por falta de outros exemplos para esse caso dentre os substantivos testados.

Como pode ser observado no gráfico 4.36, grande parte dos co-ocorrentes se concentra próxima à origem, com alguns poucos pontos com distâncias destacadas e com pouca variação angular com relação ao ponto (0, 0). As palavras que se destacam podem ser vistas no gráfico 4.35.

No entanto, observamos que as palavras que formaram o *cluster* de *página* apresentam pouca associação semântica. A questão está relacionada às situações de uso, provavelmente, muito variadas e discrepantes em que *página* é empregada no jornal, como pode ser verificado na lista de lexemas que formaram o único *cluster* relativo a *página* listados adiante.

Figura 4.34: Representação das co-ocorrências de *página* em coordenadas polares.

Figura 4.35: Representação das co-ocorrências de *página* em coordenadas polares.

Co-ocorrências de "página" em clusteres de 4°

Figura 4.36: Distribuição dos co-ocorrentes de *página* em agrupamentos de 4 graus.

1. cubanos, virão, especial, pão, geometria, psíquico, pág, bi, enviado, previsões, cinemateca, sobem, terror, plásticas, tel., vôlei, tradução, capa, tango, soltar, conheça, astronomia, indica, grito, crônicas, economistas, utopia, filmar, evita, atriz, págs, erramos, fornece, aconselhamento, fecha, entrelinhas, colunista, trad., vence, sucursal, exhibe, leões, redação, épico, reciclagem, afasta, pág., islâmico, enviados, reforça, saiba, elogio, continuação, obtém, sábados, investiga, lidera, investe, expõe, critica, pega, testa, eleva, cai, discute, frase, inédito, decide, emissoras, namoro, ecológica, desce, definem, favorito, semifinal, segurar, pênaltis, entenda, colaborou, unanimidade, neve, líbero, ilustrado, teme, reduz, papa, editoras, tênis, plásticos, editoria, reage, ataca, goleada, bate, morre

Figura 4.37: Regressão isotônica e antitônica sobre os dados do histograma dos co-ocorrentes de *página*.

Figura 4.38: Representação das co-ocorrências de *casa* em coordenadas polares.

Casa

Para o substantivo *casa*, que somou 415 ocorrências no *corpus*, os dois *clusters* distinguidos basearam-se apenas em uma palavra cada: *acordar* e *clicar*. Identificaram-se duas interpretações possíveis para esse resultado: os limiares de seleção de co-ocorrentes relevantes podem ter sido excessivamente rigorosos para a quantidade de dados disponíveis ou o *corpus* é pouco extenso e, por essa razão, pouco representativo. O mesmo resultado que observamos para *casa* se repetiu para diversas outras palavras dentre as 50 testadas.

4.3.2 Discussão dos resultados das medidas de polissemia

De forma geral, observamos que os dados para as medidas de polissemia (índice de polissemia e número de *clusters*) ainda são escassos e que a formação dos *clusters*, para a maioria das palavras, conta com poucos itens em sua composição. No entanto, a possibilidade de análise das relações de similaridade entre os *clusters* a partir de dados extraídos de forma automática do *corpus* mostrou resultados promissores, como procuramos ilustrar com alguns dos exemplos da seção 4.3.1.

As vantagens do método, por outro lado, estão especialmente no fato de que os *clusters* podem ser avaliados não apenas do ponto de vista de suas distâncias relativas, mas também de acordo com os seus pesos, isto é, do quanto são representativos. Em

Figura 4.39: Distribuição dos co-ocorrentes de *casa* em agrupamentos de 4 graus.

alguns casos, um *cluster* de poucas palavras pode ser mais representativo do que um outro formado por várias palavras pouco representativas daquele grupo.

Para as palavras testadas, não se verificou correlação entre as medidas de polissemia extraídas do *corpus* e os números de acepções distinguidos pelo Michaelis ou as frequências em *corpus*. No entanto, observamos que o escopo do dicionário é mais geral e que o *corpus* apresenta um viés político-jornalístico que restringe as possibilidades das acepções.

4.4 Comparações entre os métodos implementados

O algoritmo de extração de medidas de polissemia (especialmente a clusterização de co-ocorrentes) e a versão simplificada do algoritmo de Lesk que implementamos, apesar de não servirem estritamente às mesmas finalidades, geram produtos comparáveis na medida em que associam informações contextuais das palavras-alvo aos sentidos que ela pode assumir. Como forma de comparar e discutir os potenciais desses métodos, apresentamos como exemplo os resultados obtidos com a palavra *trabalho* para os dois procedimentos.

4.4.1 Resultados da desambiguação de *trabalho*

Como base para a desambiguação, utilizamos a entrada do substantivo *trabalho* como então definida pelo dicionário Michaelis. As 24 acepções enumeradas pelo dicionário estão listadas a seguir.

1. ato ou efeito de trabalhar
2. exercício material ou intelectual para fazer ou conseguir alguma coisa ocupação em alguma obra ou ministério
3. esforço labutação lida luta
4. aplicação da atividade humana a qualquer exercício de caráter físico ou intelectual
5. tipo de ação pelo qual o homem atua de acordo com certas normas sociais sobre uma matéria a fim de transformá-la
6. esmero ou cuidado que se emprega na feitura de uma obra
7. a composição ou feitura de uma obra
8. a própria obra que se compõe ou faz
9. obra literária ou artística
10. maneira como alguém trabalha
11. maneira como funciona um aparelho
12. movimento que se produz nos materiais industriais
13. ação dos agentes naturais
14. discussão ou deliberação para preparar resoluções
15. produto da multiplicação de uma força pela distância percorrida pelo ponto de aplicação na direção da força
16. ação de uma força que põe em movimento um corpo que lhe opõe resistência
17. esforço empregado em vencer uma resistência
18. conjunto dos fenômenos que constituem o parto
19. fenômeno orgânico que se opera no âmago dos tecidos
20. incômodo infortúnio tormento ou sucesso infeliz
21. roubo assalto execução de um delito
22. os exames discussões e deliberações de uma corporação de uma repartição de um escritório de uma oficina
23. aflições cuidados

Figura 4.40: Distribuição das acepções selecionadas de *trabalho* quanto às ocorrências da palavra no *corpus*.

24. empreendimentos

No *corpus* Mac-Morpho, o substantivo *trabalho* ocorreu, ao todo, 553 vezes. Os resultados que obtivemos com a desambiguação pela versão simplificada do algoritmo de Lesk mostraram que as acepções 5 e 22 foram associadas a uma parcela substancial das ocorrências de *trabalho* (cerca de 35% e 22%, respectivamente), como se pode verificar na figura 4.40, enquanto as demais acepções não superaram os 10%.

Por outro lado, os pesos médios associados às acepções não seguiram variação correlata à das atribuições das acepções às ocorrências de *trabalho*, como indicam os resultados apresentados no gráfico 4.41. Apenas 4 das acepções (9, 10, 13 e 20) superaram a faixa de 60% do peso máximo atribuído a uma acepção de *trabalho*, que foi de 1379. Para os dados gerais de desambiguação dos substantivos, o peso máximo superou a faixa de 2500.

Os valores de confiança para a desambiguação de cada uma das ocorrências de *trabalho* estão apresentados na figura 4.42. Seguindo a tendência dos dados gerais do *corpus*, a confiança mostrou-se maior, quanto maior o peso da acepção selecionada, isto é, desambiguações baseadas em pesos maiores apresentaram acepções concorrentes com pesos relativamente baixos. A regressão linear calculada sobre esses dados apresentou coeficiente angular de 0.84, indicando a tendência de as diferenças entre

Figura 4.41: Distribuição dos pesos médios das aceções selecionadas de *trabalho* no *corpus*. A normalização foi feita com base nos valores obtidos para a palavra *trabalho*.

a aceção selecionada e a sua concorrente se aproximarem dos valores máximos, caso em que a reta teria coeficiente angular 1.

No gráfico 4.42, observa-se que os dados se distribuem em três faixas, de acordo com os valores dos pesos das aceções selecionadas. Na primeira faixa, próxima da origem, há uma concentração relativamente alta de aceções selecionadas com menos que 30% do peso máximo. As duas faixas seguintes, situadas nos valores 0.4 e 0.8 do eixo x, correspondem aos casos em que as desambiguações se basearam em 1 e 2 termos distintivos.

Para exemplificar pontualmente os resultados obtidos com as desambiguações, selecionamos uma das janelas contextuais em que a palavra *trabalho* foi desambi-guada. De acordo com o algoritmo que implementamos, essa ocorrência de *trabalho* corresponde à aceção 9 dada pelo dicionário Michaelis, relacionada à noção de *obra literária ou artística*. A janela está apresentada a seguir:

- (4.4) *não era apenas folclórico. Pereira foi o primeiro a gravar Cartola e Paulo Vanzolin, entre outros, sem contar a história musical das quatro maiores escolas de samba do Rio de Janeiro: Salgueiro, Portela, Mangueira e Império Serrano. “O melhor do **trabalho** dele é o que ele não chegou a fazer”, diz o compositor*

Figura 4.42: Grau de confiança nas aceções seleccionadas de *trabalho*

*e músico Marcus Vinicius responsável pela direção **artística** da Discos Marcus Pereira e atualmente dono do estúdio Spalla. O que há por ser feito hoje é muito*

Para esse exemplo, particularmente, o peso da aceção escolhida ficou entre os mais altos da palavra *trabalho*. Somada a isso, a diferença entre a aceção escolhida e a segunda opção situa essa desambiguação entre aquelas com as maiores medidas de confiança. Em valores absolutos, o peso foi de 1068 para a primeira aceção e de 137 para a segunda colocada, o que corresponde a 0,77 e 0,1 na escala normalizada para os dados de *trabalho*. Como destacado no trecho, a palavra que baseou a escolha da aceção foi *artística*.

4.4.2 Resultados das medidas de polissemia de *trabalho*

Dentre as medidas de polissemia atribuídas à palavra *trabalho*, destacaremos a da clusterização, tanto quanto ao número de *clusters* gerados, quanto em relação às composições dos *clusters* e seus pesos. A partir dos parâmetros estabelecidos para a extração do índice de polissemia, foram selecionados poucos co-ocorrentes relevantes de *trabalho*. Consequentemente, no passo de clusterização, os dois *clusters* que obtivemos foram formados por poucas palavras. Os grupos de palavras que formaram os dois *clusters* identificados para a palavra *trabalho* estão listados a seguir

Figura 4.43: Representação das co-ocorrências de *trabalho* em coordenadas polares.

<i>cluster</i>	peso	ângulos (graus)
1	1.0	-108
2	0.52	176

Tabela 4.4: Tabela para localização dos *clusters* de *trabalho* no gráfico 4.45.

e correspondem, essencialmente, aos agrupamentos em faixas de 4 graus feitos com os co-ocorrentes em uma etapa anterior à clusterização.

1. postos, desempregados, fechamento, carteiras, representando
2. ignorância, geometria, acidentes

A distância entre os dois *clusters* foi de cerca de 90 graus e seus pesos podem ser definidos, principalmente, em função da relevância das palavras na composição de cada um, de acordo com os parâmetros de relevância de co-ocorrentes que estabelecemos e que foram apresentados na seção 3.3. A influência dos pesos de cada palavra sobre os pesos dos *clusters* pode ser verificada no gráfico 4.43, especialmente, no caso da palavra *geometria*. Quanto mais os pontos estiverem distantes do eixo y, maiores serão seus pesos. Os *clusters* podem ser observados nos gráficos 4.44 e 4.45.

O método de clusterização discretizou 2 contextos que podem ser atribuídos à palavra *trabalho*: um que se aproxima das noções de *trabalho* pertencentes ao

Co-ocorrências de "trabalho" em clusteres de 4°

Figura 4.44: Distribuição dos co-ocorrentes de *trabalho* em agrupamentos de 4 graus.

Histograma dos co-ocorrentes de "trabalho"

Figura 4.45: Histograma dos co-ocorrentes de *trabalho* em agrupamentos de 4 graus.

campo semântico de *ocupação profissional* e outro que, apesar de ser composto por palavras pouco relacionadas, teve o termo *geometria* como principal componente, aproximando-se do que o dicionário sugere como campo semântico de *obra*. O Dicionário de Sinônimos lista 9 campos semânticos associados à palavra *trabalho*. Os campos e seus respectivos conjuntos de sinônimos estão apresentados a seguir.

1. **Ocupação profissional:** serviço, emprego, ofício, atividade, expediente, exercício, ocupação, mister, tarefa.
2. **Lide:** empreitada, faina, esforço, luta, azáfama, lide, labuta, afã, canseira, farfúncia, laboração, labor, lida, quefazer, trabalhão, afazer, trabalhadeira.
3. **Obra:** realização, obra, composição, empreendimento, elaboração, empresa, labor, feitura, produção, projeto.
4. **Cuidado:** esmero, empenho, dedicação, capricho, zelo, cuidado.
5. **Encargo:** compromisso, função, dever, incumbência, obrigação, responsabilidade, encargo.
6. **Aflicção:** apreensão, pesar, padecimento, aflicção, inquietação, pena, preocupação.
7. **Deliberações:** discussões, deliberações, exames.
8. **Atividade:** movimento, ato, energia, treino, ação, movimentação, atividade, treinamento.
9. **Feitiçaria:** oferenda, feitiçaria, bruxaria, mandinga.

4.4.3 Análise comparativa dos métodos apresentados

O método de clusterização identificou apenas duas acepções da palavra *trabalho*, em contraposição às 21 acepções distinguidas pelo método de desambiguação. Apesar dessa disparidade, observamos que as acepções fornecidas pelo dicionário Michaelis são discretizadas de acordo com critérios diversos e sem que se estabeleçam relações entre as acepções. Contrastando-se a entrada de *trabalho* no dicionário Michaelis com a do Dicionário de Sinônimos, verificamos que este distingue 9 campos semânticos da palavra que poderiam englobar praticamente toda, se não toda, a miríade de acepções que o Michaelis diferencia.

Além disso, notamos que o algoritmo de desambiguação que implementamos atribuiu poucas janelas à maioria das acepções, apesar de algumas dessas atribuições terem exibido altos graus de confiança. Por outro lado, as acepções que foram atribuídas a maiores quantidades de janelas obtiveram, em média, pesos com valores baixos. Esses resultados indicam que estas desambiguações foram baseadas em termos poucos distintivos e, provavelmente, com baixos graus de confiança.

Retomando o que comentamos anteriormente neste capítulo, a deficiência dos algoritmos de desambiguação baseados em dicionário reside no fato de as acepções de dicionário não conterem, necessariamente, palavras que ocorrem nos contextos de uso das palavras-alvo, dado o caráter da função do dicionário de fornecer descrições metalinguísticas. Como apresentado para os resultados gerais das desambiguações, avaliamos que as desambiguações contaram, para uma parcela substancial dos dados, com termos pouco distintivos.

Salientamos, entretanto, que dicionários que fornecem exemplos de usos² podem enriquecer métodos de desambiguação como os de Lesk, trazendo palavras que co-ocorrem com a palavra-alvo de maneira distintiva, mesmo que os exemplos não sejam elaborados de forma sistemática.

De maneira similar ao algoritmo de Lesk, o método de clusterização de co-ocorrentes possibilita que se verifiquem os pesos das palavras que compõem os *clusters*. No caso do algoritmo de Lesk, os pesos (dados pelos IDFs) de cada uma das palavras das acepções do dicionário contribuíram para o peso total das respectivas acepções. No exemplo 4.4.1 de janela contextual, a palavra *artística* foi a mais relevante, isto é, com maior peso, para a desambiguação de *trabalho* e está destacada em negrito no exemplo.

O método de clusterização pode ser uma importante ferramenta para a identificação de contextos em que a palavra-alvo ocorre, partindo exclusivamente dos dados do *corpus*. Outro fator importante é o grau de similaridade, que pode ser verificado entre os *clusters* encontrados, de acordo com as distâncias (ângulos) entre eles. A vantagem de termos os graus de similaridade é que essa informação permite averiguar quais ocorrências de uma palavra-alvo se aproximam mais da polissemia e quais são homônimas. Ocorrências homônimas seriam mais facilmente distinguíveis do que as polissêmicas porque aquelas tendem a compartilhar menos co-ocorrentes distintivos. Dessa maneira, o desempenho das desambiguações pode ser avaliado quanto a esses aspectos.

²A nova versão eletrônica do dicionário Michaelis foi reformulada sob vários aspectos, entre eles, a inclusão de mais exemplos de uso para as acepções listadas.

5

Considerações finais

Tendo em vista os principais objetivos desta pesquisa, avaliamos que os resultados obtidos, especialmente com a extração de índices de polissemia e *clusters*, apontam potenciais promissores, mas que precisam ser investigados com mais dados, isto é, para mais palavras e com *corpora* maiores, de conteúdos temáticos diversificados. Por essas razões, pretendemos, em trabalhos futuros, implementar e estender para o *corpus* Subtlex os testes aqui apresentados.

5.1 Vantagens e limitações da proposta

O método de desambiguação automática implementado, baseado no algoritmo de Lesk, evidenciou uma limitação inerente a quaisquer métodos de desambiguação baseados em dicionário: os dicionários não apresentam, necessariamente, as palavras que co-ocorrem com o termo-alvo, de forma distintiva, em contextos de uso. O fato de o *corpus* utilizado (Mac-Morpho) ser tematicamente enviesado tornou o uso do dicionário, como recurso de supervisão, um fator ainda mais limitante.

Os critérios para as divisões de entradas e distinções entre acepções nos dicionários tradicionais são alvos de discussões e críticas na literatura (Lyons (1977), Kilgariff (1997)). Uma das grandes dificuldades dos lexicógrafos ao elaborarem dicionários está em determinar critérios objetivos que permitam dizer que duas acepções são distintas ou que uma determinada forma gráfica necessita de mais de uma entrada no dicionário em função de questões etimológicas. Essas questões deixam claras as dificuldades de se partir dos inventários de significados desses recursos, especialmente, ao se manipular *corpora* que tratem de conteúdos especializados.

Pretendia-se analisar as distribuições das acepções dos substantivos testados com o intuito de verificar quais teriam distribuições mais uniformes e quais tenderiam a ter acepções predominantes. Esses dados seriam confrontados com as frequências dos substantivos no *corpus* e com suas quantidades de acepções de acordo com o dicionário. No entanto, essa etapa do trabalho ficou prejudicada pelo desempenho obtido com a desambiguação: apesar de o método ter superado o *baseline*, a maioria das desambiguações se apoiou em palavras pouco distintivas, como discutido na seção 4.1 e na subseção 4.2.1, o que foi causado pelas poucas intersecções entre janelas contextuais e acepções do dicionário e, principalmente, pelo fato de as intersecções terem sido de termos de baixos IDFs, isto é, de baixos pesos, como é o caso de palavras funcionais.

O método de clusterização proposto juntamente ao de extração de índices de polissemia foi incluído aos testes deste projeto como uma forma de minimizar a dependência de recursos externos ao *corpus*. Os resultados indicaram o potencial do método para a distinção entre polissemia e homonímia através de informações contextuais, incluindo não apenas a clusterização das palavras co-ocorrentes de acordo com sua similaridade contextual, mas também os pesos dessas relações e dos próprios *clusters*.

A relevância dos índices de polissemia foi mantida em segundo plano neste trabalho em razão de os testes terem sido realizados para um grupo reduzido de substantivos e de a importância desses valores ter sido posta em termos da correlação com os números de *clusters*, aspecto pouco explorado pelos autores do trabalho original.

Notamos, entretanto, a sensibilidade do método aos parâmetros de seleção de termos relevantes e entendemos que esses valores necessitam de investigação mais ampla para serem estabelecidos de forma adequada. Deerwester *et al.* (1990) sugerem, a partir de evidências experimentais de seu trabalho, que a filtragem permita que cerca de cem palavras relevantes sejam selecionadas. Os limiares que estipulamos para selecionar palavras co-ocorrentes relevantes, tanto o de mínimo de ocorrências, quanto o de co-ocorrências com a palavra-alvo, mostraram-se mais restritivos do que o sugerido pelos autores. Consideramos que esse seja um dos aspectos que necessita de mais investigações em trabalhos futuros, pois os testes foram realizados para um conjunto ainda restrito de palavras, utilizando um *corpus* enviesado e relativamente pequeno.

5.2 Trabalhos futuros

Recentemente, tivemos contato com o projeto OpenWordNet-PT¹ (de Paiva *et al.*, 2012), cujos conteúdos podem ser acessados abertamente e vêm sendo constantemente expandidos. Iniciamos o pré-processamento dos dados dessa WordNet, mas os testes não puderam ser concluídos até o fechamento desta dissertação: pretendíamos implementar a proposta de desambiguação desenvolvida por Agirre *et al.* (2014) baseada no algoritmo *PageRank* (Page *et al.*, 1999), cujos cálculos de ranqueamento são utilizados para desambiguação de termos a partir de passeios aleatórios sobre um grafo formado pelas relações semânticas da WordNet.

Diante dos resultados obtidos com o algoritmo de Lesk, procuramos esses recursos como forma de obter resultados mais robustos com a desambiguação. A inclusão da rede de relações semânticas somada ao uso de um *corpus* mais representativo é relatada como fonte de bons resultados (Agirre *et al.*, 2014). Esse método de desambiguação baseado em grafo será explorado em trabalhos futuros, assim como a extensão do método de extração dos índices de polissemia para o uso na tarefa de desambiguação proposta por Sproat & VanSanten (2001). Com métodos de desambiguação mais precisos, esperamos poder realizar a análise de distribuição de acepções em *corpus* para palavras polissêmicas.

¹<http://wnpt.brcloud.com/wn/>; código e dados disponíveis em: <https://github.com/own-pt/openWordnet-PT>

Referências Bibliográficas

- AFONSO, Susana, BICK, Eckhard, HABER, Renato & SANTOS, Diana. “Floresta sintá (c) tica: A treebank for portuguese.” *In: LREC*, 2002.
- AGIRRE, Eneko, DE LACALLE, Oier Lopez & SOROA, Aitor. “Random walks for knowledge-based word sense disambiguation.” *Computational Linguistics*, vol. 40(1), 57–84, 2014.
- AGIRRE, Eneko & EDMONDS, Philip (Eds.). *Word Sense Disambiguation: Algorithms and Applications*, vol. 33. Springer, 2007.
URL <http://www.wsdbook.org/>
- ALUÍSIO, Sandra M, PINHEIRO, Gisele Montilha, MANFRIN, Aline MP, DE OLIVEIRA, Leandro HM, GENOVES JR, Luiz C & TAGNIN, Stella EO. “The lácio-web: Corpora and tools to advance brazilian portuguese language investigations and computational linguistic tools.” *In: LREC*, 2004.
- BALDINGER, Kurt. “Semasiologia e onomasiologia.” *ALFA: Revista de Linguística*, vol. 9, 1966.
- BERBER SARDINHA, Tony, MOREIRA FILHO, JL & ALAMBERT, E. “O corpus brasileiro.” *Comunicação ao VII Encontro de Lingüística de Corpus*, 2008.
- BICK, Eckhard. *The parsing system "Palavras": Automatic grammatical analysis of Portuguese in a constraint grammar framework*. Aarhus Universitetsforlag, 2000.
- BORBA, Francisco da Silva. *Dicionário de usos do português do Brasil*. São Paulo: Editora Ática, 2002.
- BRÉAL, Michel. *Essai de sémantique: (science des significations)*. Librairie Hachette, 1897.
- CROSSLEY, Scott, SALSBUURY, Tom & MCNAMARA, Danielle. “The development of polysemy and frequency use in english second language speakers.” *Language Learning*, vol. 60(3), 573–605, 2010.

- DAVIES, Mark & FERREIRA, Michael. “Corpus do português (45 million words, 1300s-1900s).” *National Endowment for the humanities*, 2006.
- DE PAIVA, Valeria, RADEMAKER, Alexandre & DE MELO, Gerard. “Openwordnet-pt: An open Brazilian Wordnet for reasoning.” In: *Proceedings of COLING 2012: Demonstration Papers*, pp. 353–360. Mumbai, India: The COLING 2012 Organizing Committee, 2012. Published also as Techreport <http://hdl.handle.net/10438/10274>.
URL <http://www.aclweb.org/anthology/C12-3044>
- DEERWESTER, Scott, DUMAIS, Susan T., FURNAS, George W., LANDAUER, Thomas K. & HARSHMAN, Richard. “Indexing by latent semantic analysis.” *Journal of the American Society for Information Science*, vol. 41(6), 391–407, 1990.
URL <http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.108.8490>
- DIAS-DA SILVA, Bento Carlos. *A face tecnológica dos estudos da linguagem: o processamento automático das línguas naturais*. Ph.D. thesis, Universidade Estadual Paulista, Araraquara, 1996.
- . “O estudo linguístico-computacional da linguagem.” *Letras de Hoje*, vol. 41(2), 103–138, 2006a.
- . “Wordnet. br: An exercise of human language technology research.” In: *Proceedings of the Third International WordNet Conference—GWC*, pp. 22–26, 2006b.
- DIAS-DA SILVA, Bento Carlos, DI FELIPPO, Ariani & HASEGAWA, Ricardo. “Methods and tools for encoding the wordnet.br sentences, concept glosses, and conceptual-semantic relations.” In: *Computational Processing of the Portuguese Language*, pp. 120–130. Springer, 2006.
- DOROW, Beate. *A graph model for words and their meanings*. Ph.D. thesis, 2006.
- ERK, Katrin & PADÓ, Sebastian. “A structured vector space model for word meaning in context.” In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 897–906. Association for Computational Linguistics, 2008.
- FONSECA, Erick Rocha & ROSA, Joao Luis G. “Mac-morpho revisited: Towards robust part-of-speech tagging.” In: *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, pp. 98–107, 2013.

- GALE, William A, CHURCH, Kenneth W & YAROWSKY, David. "Work on statistical methods for word sense disambiguation." In: *Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language*, vol. 54, p. 60, 1992.
- GEERAERTS, Dirk. *Theories of Lexical Semantics*. Oxford University Press, 2010.
URL <http://books.google.pt/books?id=AHU{ }2mFy8iEC>
- IDE, Nancy & VÉRONIS, Jean. "Introduction to the special issue on word sense disambiguation: the state of the art." *Computational linguistics*, vol. 24(1), 2–40, 1998.
- JURAFSKY, Daniel & MARTIN, James H. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*, 2000.
- KEULEERS, Emmanuel, BRYBAERT, Marc & NEW, Boris. "Subtlex-nl: A new measure for dutch word frequency based on film subtitles." *Behavior research methods*, vol. 42(3), 643–650, 2010.
- KILGARRIFF, Adam. *Polysemy*. Ph.D. thesis, University of Sussex, 1992.
- . "I don't believe in word senses." *Computers and the Humanities*, vol. 31(2), 91–113, 1997.
- LANDAUER, Thomas K & DUMAIS, Susan T. "A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review*, vol. 104(2), 211, 1997.
- LEE, Christopher J. "Some hypotheses concerning the evolution of polysemous words." *Journal of Psycholinguistic Research*, vol. 19(4), 211–219, 1990.
- LESK, Michael. "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone." In: *Proceedings of the 5th annual international conference on Systems documentation*, pp. 24–26. ACM, 1986.
- LYONS, John. *Semantics*, vol. 2, 1977.
- MILLER, George A & FELLBAUM, Christiane. "Semantic networks of english." *Cognition*, vol. 41(1), 197–229, 1991.
- NAVIGLI, Roberto. "Word sense disambiguation: A survey." *ACM Computing Surveys (CSUR)*, vol. 41(2), 10, 2009.

- NEW, Boris, BRYSSBAERT, Marc, VERONIS, Jean & PALLIER, Christophe. “The use of film subtitles to estimate word frequencies.” *Applied Psycholinguistics*, vol. 28(04), 661–677, 2007.
- NÓBREGA, Fernando Antônio A & PARDO, Thiago A Salgueiro. “Explorando métodos de uso geral para desambiguação lexical de sentidos para a língua portuguesa.”, 2012.
- PAGE, Lawrence, BRIN, Sergey, MOTWANI, Rajeev & WINOGRAD, Terry. “The pagerank citation ranking: bringing order to the web.”, 1999.
- RASO, Tommaso & MELLO, Heliana. *C-oral-Brasil: corpus de referência do português brasileiro falado informal. I*, 2012.
- RAVIN, Yael & LEACOCK, Claudia. *Polysemy: Theoretical and Computational Approaches: Theoretical and Computational Approaches*. Oxford University Press, 2000.
- SARDINHA, Tony Berber. “Linguística de corpus: histórico e problemática.” *Delta*, vol. 16(2), 323–367, 2000.
- SCHÜTZE, Hinrich. “Context space.” In: *AAAI fall symposium on probabilistic approaches to natural language*, pp. 113–120, 1992.
- SILVA, Augusto Soares da. *O mundo dos sentidos em português: polissemia, semântica e cognição*. Almedina, 2006.
- . “Polissemia e contexto: o problema dura da diferenciação de sentidos.” *Estudos Linguísticos/Linguistic Studies*, vol. 5, 353–367, 2010.
- SPROAT, Richard William & VANSANTEN, Jan Pieter. “Method and apparatus for measuring the degree of polysemy in polysemous words.”, 2001. US Patent 6,256,629.
- TANG, Kevin. “A 61 million word corpus of brazilian portuguese film subtitles as a resource for linguistic research.” Tech. rep., UCL Working Papers in Linguistics 24: 208–214, 2012.
- TORGERSON, Warren S. “Theory and methods of scaling.”, 1958.
- TURNEY, Peter D, PANTEL, Patrick *et al.* “From frequency to meaning: Vector space models of semantics.” *Journal of artificial intelligence research*, vol. 37(1), 141–188, 2010.

- VAN HEUVEN, Walter JB, MANDERA, Pawel, KEULEERS, Emmanuel & BRYLSBAERT, Marc. "Subtlex-uk: A new and improved word frequency database for british english." *The Quarterly Journal of Experimental Psychology*, vol. 67(6), 1176–1190, 2014.
- YAROWSKY, David. "Unsupervised word sense disambiguation rivaling supervised methods." In: *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pp. 189–196. Association for Computational Linguistics, 1995.
- . "Homograph disambiguation in text-to-speech synthesis." In: *Progress in speech synthesis*, pp. 157–172. Springer, 1997.
- ZAVAGLIA, Claudia. "Ambigüidade gerada pela homonímia: revisão teórica, linhas limítrofes com a polissemia e proposta de critérios distintivos." *Delta*, vol. 19(2), 237–266, 2003.
- ZIPF, G K. "The meaning-frequency relationship of words." *The Journal of general psychology*, vol. 33(33), 251–256, 1945.

Apêndices

Lemas desambiguados

Lista dos cem lemas de substantivos mais frequentes do *corpus* Subtlex-PT-BR de Tang (2012). Os dados do *corpus* podem ser acessados em <http://zipf.ugent.be/open-lexicons/interfaces/br-pt-lemmas/>. Os lemas da tabela abaixo estão ordenados de acordo com a ordem decrescente de suas frequências em *corpus* e acompanhados de suas diversidades contextuais, isto é, medidas de frequência de documento baseadas na fração de arquivos de legendas distintos em que o lema ocorreu.

Tabela A.1: Tabela dos cem lemas mais frequentes do Subtlex-PT-BR, em ordem decrescente, e suas respectivas diversidades contextuais relativas

Rank	Lemas	Diversidade Contextual
1	coisa	0.89
2	dia	0.80
3	senhor	0.48
4	vez	0.84
5	homem	0.70
6	tempo	0.94
7	casa	0.88
8	favor	0.89
9	pai	0.47
10	ano	0.70
11	noite	0.86
12	deus	0.77
13	filho	0.49
14	vida	0.89

Tabela A.1: Cont. da tabela de lemas desambiguados

Rank	Lemas	Diversidade Contextual
15	cara	0.70
16	pessoa	0.70
17	amigo	0.67
18	mãe	0.49
19	verdade	0.85
20	lugar	0.87
21	garoto	0.45
22	hora	0.71
23	mulher	0.55
24	mundo	0.82
25	problema	0.65
26	dinheiro	0.64
27	nome	0.80
28	trabalho	0.79
29	irmão	0.38
30	carro	0.54
31	amor	0.61
32	merda	0.32
33	mão	0.56
34	certeza	0.75
35	tipo	0.70
36	parte	0.74
37	criança	0.42
38	minuto	0.50
39	fora	0.83
40	gente	0.60
41	cidade	0.57
42	lado	0.72
43	momento	0.69
44	volta	0.77
45	frente	0.72
46	olho	0.60
47	cabeça	0.67
48	semana	0.53

Tabela A.1: Cont. da tabela de lemas desambiguados

Rank	Lemas	Diversidade Contextual
49	arma	0.37
50	medo	0.65
51	família	0.55
52	história	0.54
53	porta	0.58
54	cuidado	0.49
55	forma	0.59
56	caminho	0.64
57	guerra	0.35
58	quarto	0.52
59	papai	0.31
60	terra	0.37
61	morte	0.53
62	sorte	0.63
63	palavra	0.43
64	cima	0.62
65	razão	0.60
66	caso	0.53
67	água	0.52
68	coração	0.49
69	mês	0.47
70	manhã	0.60
71	jeito	0.57
72	querido	0.48
73	mamãe	0.29
74	polícia	0.37
75	jogo	0.41
76	esposo	0.43
77	fim	0.57
78	corpo	0.47
79	ideia	0.49
80	chefe	0.36
81	ajuda	0.61
82	filme	0.26

Tabela A.1: Cont. da tabela de lemas desambiguados

Rank	Lemas	Diversidade Contextual
83	diabo	0.34
84	pergunta	0.46
85	boca	0.50
86	escola	0.38
87	música	0.32
88	maneira	0.48
89	número	0.41
90	marido	0.38
91	luz	0.40
92	prazer	0.42
93	festa	0.35
94	causa	0.55
95	adeus	0.36
96	sangue	0.42
97	paz	0.46
98	poder	0.63
99	licença	0.42
100	país	0.35

B

Pseudo-palavras desambiguadas

Tabela B.1: Tabela das pseudo-palavras organizadas de acordo com os métodos de formação.

Grupo A	Grupo B	Grupo C
frente-razão	diabo-semana	tempo-minuto
cara-quarto	terra-frente	maneira-ano
marido-sangue	mãe-maneira	favor-coisa
coisa-problema	caminho-água	número-adeus
nome-lado	volta-mão	lugar-gente
querido-deus	garoto-casa	mulher-noite
lugar-número	sorte-número	boca-dinheiro
forma-cidade	adeus-tempo	vez-papai
dia-ano	papai-história	irmão-semana
ajuda-medo	chefe-carro	guerra-ideia
porta-manhã	palavra-família	senhor-cabeça
filho-terra	quarto-polícia	fora-arma
amigo-pai	medo-filme	olho-poder
licença-cuidado	fora-cidade	certeza-escola
hora-verdade	irmão-problema	cidade-morte
ideia-adeus	cuidado-vez	esposo-licença
irmão-certeza	cima-ano	deus-homem
boca-jeito	arma-lugar	garoto-merda
palavra-senhor	festa-verdade	volta-festa
amor-chefe	cabeça-luz	frente-pai

Tabela B.1: Cont. da tabela de pseudo-palavras

Grupo A	Grupo B	Grupo C
história-trabalho	porta-noite	diabo-mão
filme-morte	paz-gente	caso-quarto
mãe-caminho	marido-mulher	amor-chefe
fim-mulher	esposo-dia	sangue-música
guerra-família	merda-mundo	prazer-hora
tempo-país	favor-vida	jeito-dia
fora-minuto	filho-razão	caminho-carro
favor-olho	criança-hora	vida-problema
vez-caso	cara-momento	terra-história
cima-esposo	pergunta-tipo	corpo-querido
festa-criança	minuto-forma	sorte-família
mamãe-papai	senhor-guerra	forma-verdade
prazer-pergunta	licença-caso	parte-mamãe
garoto-coração	sangue-fim	manhã-paz
semana-mundo	jeito-nome	fim-mundo
sorte-arma	querido-mês	casa-mãe
luz-mão	prazer-música	tipo-marido
polícia-momento	amor-poder	trabalho-causa
parte-mês	certeza-coisa	amigo-criança
volta-cabeça	ideia-país	país-ajuda
jogo-vida	amigo-pessoa	polícia-momento
homem-noite	coração-dinheiro	filho-porta
carro-poder	deus-parte	cima-filme
causa-paz	mamãe-trabalho	lado-luz
merda-diabo	causa-corpo	coração-mês
gente-corpo	ajuda-morte	razão-cara
dinheiro-casa	boca-lado	medo-água
escola-pessoa	manhã-homem	pessoa-nome
água-maneira	pai-escola	palavra-pergunta
tipo-música	olho-jogo	jogo-cuidado

Medidas de polissemia

Os índices de polissemia estão em escala normalizada, as demais medidas da tabela estão em valores absolutos.

Tabela C.1: Tabela com as medidas de polissemia calculadas para os 50 substantivos mais frequentes em suas formas lematizadas.

Substantivos	Índices (pol.)	Clusters	Acepções	Frequências
mercado	1.817	4	6	693
candidato	1.412	3	3	362
mês	1.334	2	5	602
número	1.212	3	14	370
valor	1.18	4	18	308
produção	1.059	2	12	434
tempo	0.989	2	21	616
parte	0.985	2	21	612
mundo	0.985	2	17	513
sistema	0.983	2	15	393
empresa	0.976	3	6	558
equipe	0.955	2	2	450
forma	0.955	2	21	447
caso	0.955	2	15	445
casa	0.955	2	14	415
campanha	0.955	2	7	396
diretor	0.955	2	5	381
obra	0.955	2	14	365

Tabela C.1: Cont. da tabela das medidas de polissemia

Substantivos	Índices (pol.)	Clusters	Acepções	Frequências
região	0.955	2	11	324
processo	0.955	2	14	322
lugar	0.955	2	21	320
time	0.925	2	2	402
estado	0.891	2	13	478
média	0.853	2	7	312
inflação	0.81	3	6	394
grupo	0.806	2	17	506
lado	0.758	2	19	313
governo	0.721	2	12	1094
presidente	0.691	2	2	899
dinheiro	0.663	2	7	386
projeto	0.618	2	4	391
livro	0.6	2	9	567
país	0.565	2	4	821
preço	0.491	2	6	443
julho	0.486	2	1	330
jogo	0.479	2	27	423
vida	0.478	2	16	430
programa	0.469	2	6	578
história	0.448	2	10	532
vez	0.443	3	5	434
cidade	0.369	2	4	453
período	0.347	2	11	312
real	0.322	2	10	335
plano	0.315	2	16	404
trabalho	0.265	2	24	551
semana	0.263	2	6	486
ano	0.253	2	5	1259
dia	0.116	2	8	1156
página	0.075	1	9	2016
editoria	0.067	1	1	1981