

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito: 16/02/2004

Assinatura:

Avaliação de Métodos de Extração Automática de Terminologia para textos em Português

Maria Fernanda Teline

Orientadora: *Profa. Dra. Sandra Maria Aluísio*

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências: Área – Ciências da Computação e Matemática Computacional.

USP – São Carlos
Fevereiro/2004

A Deus, a meus pais José Donizeti Teline e Maria Izabel Temporim Teline, a meus irmãos Emerson Luiz Teline e Pamela Cristina Teline e àqueles que contribuíram para que eu pudesse alcançar este objetivo.

Agradecimentos

A Deus pela minha vida e por seu amor infinito que tem me consolado nas situações de aflição, e renovado minhas forças, dando possibilidades para o alcance dos meus ideais.

A meus tão queridos pais e irmãos por acreditarem em minha capacidade e esforço, me incentivando com o amor oferecido a cada dia, e possibilitando esta conquista tão grandiosa.

A meus avós, tios e primos por serem tão especiais e apontarem o melhor caminho, estando sempre acompanhando meus passos.

A meu namorado Ricardo da Costa Funfas, que com seu carinho e amor soube me transmitir paz e tranquilidade a todo instante, sendo paciente e me instruindo nos momentos de dificuldade.

A meus verdadeiros amigos que sempre souberam entender meus defeitos e retribuir o grande sentimento que tenho por eles, em especial a minha amiga Aline Maria da Paz Silva que esteve sempre ao meu lado, sendo compreensiva e atenciosa, compartilhando de cada etapa da realização do meu objetivo, e dando apoio para continuar a caminhada.

A minha orientadora Sandra Maria Aluísio, que esteve me direcionando e me auxiliando nas tomadas de decisões durante os dois anos de mestrado, sendo compreensiva e incentivadora, não permitindo que eu desistisse do meu objetivo.

Aos colegas do NILC, em especial, Aline Maria da Paz Silva, Jorge Pelizzoni e Aline Manfrin, que muito auxiliaram no desenvolvimento desse projeto de mestrado.

Ao CNPq, pelo auxílio financeiro.

Sumário

Lista de Figuras.....	v
Lista de Tabelas	vi
Lista de Quadros	vii
Resumo	ix
Abstract.....	x
Capítulo 1.....	1
Introdução	1
1.1 Motivação.....	5
1.2 Objetivos.....	6
1.3 Organização do texto	6
Capítulo 2.....	9
Terminologia e áreas afins	9
2.1 Um pequeno histórico da Terminologia.....	9
2.2 Pesquisas em Terminologia no Brasil.....	12
2.3 Produtos terminológicos	14
Capítulo 3.....	25
Métodos para extração de termos técnicos	25
3.1 Abordagem estatística	28
3.1.1 Frequência de ocorrência	29
3.1.2 Informação mútua (para associações binárias)	30
3.1.3 Coeficiente <i>log-likelihood</i>	30
3.1.4 Coeficiente <i>dice</i>	31
3.1.5 <i>C-value</i>	31
3.2 Abordagem lingüística	33
3.2.1 Extrator de termos de Heid et al (1996).....	33
3.2.1.1 Descrição do algoritmo	33
3.2.1.2 Aplicação	34
3.2.1.3 Extensão	35
3.2.1.4 Recursos e ferramentas utilizados.....	36
3.2.2 Sistema DEFINDER de Klavans and Muresan (2000; 2001a; 2001b).....	37
3.2.2.1 Descrição do algoritmo	37
3.2.2.2 Aplicação	38
3.2.2.3 Extensão	43
3.2.2.4 Recursos e ferramentas utilizados.....	43
3.3 Abordagem híbrida	43
3.3.1 Extrator de termos de Frantzy and Ananiadou (1997).....	43
3.3.1.1 Descrição do Algoritmo	46
3.3.1.2 Aplicação	48
3.3.1.3 Extensão	49

3.3.1.4 Recursos e ferramentas utilizados.....	49
3.3.2. Extrator de termos de Dias et al (2000).....	49
3.3.2.1 Descrição do algoritmo.....	50
3.3.2.2 Aplicação.....	58
3.3.2.3 Extensão.....	61
3.3.2.4 Recursos e ferramentas utilizados.....	62
3.4 Experimentos: estatístico e híbrido no trabalho de Dias et al (2000).....	62
3.4.1 Abordagem puramente estatística.....	63
3.4.2 Abordagem híbrida.....	66
3.5 Considerações finais.....	66
Capítulo 4.....	71
Recursos e ferramentas utilizados.....	71
4.1 Seleção e preparação do cópuz alvo.....	71
4.2 Lista de termos de referência (da área de Revestimentos Cerâmicos).....	73
4.3 O tokenizador desenvolvido no NILC.....	74
4.4 O etiquetador MXPOST treinado com um cópuz de textos em português.....	74
4.5 O concordanceador AntConc.....	75
4.6 O pacote estatístico NSP.....	75
4.7 O cópuz CópuzEco da área de Ecologia e sua lista de termos.....	81
Capítulo 5.....	83
Implementação e avaliação de quatro métodos estatísticos.....	83
5.1 Geração das listas de unigramas, bigramas e trigramas.....	83
5.2 Os métodos estatísticos implementados.....	84
5.2.1 Método estatístico para unigramas.....	88
5.2.2 Métodos estatísticos para bigramas.....	88
5.2.3 Métodos estatísticos para trigramas.....	90
5.2.4 Discussão dos resultados dos métodos estatísticos.....	91
5.3 Levantamento de candidatos a termos.....	92
Capítulo 6.....	97
Implementação e avaliação de um método lingüístico.....	97
6.1 O método lingüístico implementado e sua avaliação quantitativa.....	97
6.2 Variações do método lingüístico e suas avaliações qualitativas.....	104
6.2.1 Experimento 1: avaliando o etiquetador utilizado.....	107
6.2.2 Experimento 2: extraindo listas de termos.....	111
6.2.3 Discussão dos resultados dos experimentos 1 e 2.....	115
Capítulo 7.....	117
Implementação e avaliação de um método híbrido.....	117
Capítulo 8.....	123
Conclusões.....	123
8.1 Contribuições.....	124
8.2 Limitações.....	124
8.3 Trabalhos Futuros.....	125
Referências Bibliográficas.....	127

A. Telas com exemplos de expressões lingüísticas e seus contextos.....	133
B. Lista de referência separada em unigramas, bigramas e trigramas	135
C. <i>StopList</i>	139

Lista de Figuras

Figura 2.1 - Glossário de biodiversidade	15
Figura 2.2 - Glossário de turismo	16
Figura 2.3 - Glossário de termos do mercado financeiro	17
Figura 2.4 - Dicionário de astronomia	18
Figura 2.5 - Dicionário médico de siglas	19
Figura 2.6 - Glossário de infertilidade	20
Figura 2.7 - Glossário ambiental	21
Figura 2.8 - Glossário de imunologia	21
Figura 2.9 - Glossário de estatística	22
Figura 2.10 - Dicionário de biologia celular e molecular	23
Figura 2.11 - Glossário de direito ambiental internacional	23
Figura 3.1 - Contexto do <i>Span</i> para a palavra pivô <i>Lei</i>	50
Figura 5.1 - Classes de Palavras para Unigramas – Frequência	85
Figura 5.2 - Classes de Palavras para Bigramas – Frequência	85
Figura 5.3 - Classes de Palavras para Bigramas – Informação Mútua	86
Figura 5.4 - Classes de Palavras para Bigramas – <i>Log-likelihood</i>	86
Figura 5.5 - Classes de Palavras para Bigramas – Coeficiente <i>Dice</i>	86
Figura 5.6 - Classes de Palavras para Trigramas – Frequência	87
Figura 5.7 - Classes de Palavras para Trigramas – Informação Mútua	87
Figura 5.8 - Classes de Palavras para Trigramas – <i>Log-likelihood</i>	87
Figura 5.9 - Método Estatístico para Unigramas – Frequência	88
Figura 5.10 - Método Estatístico para Bigramas – Frequência	89
Figura 5.11 - Método Estatístico para Bigramas – Informação Mútua	89
Figura 5.12 - Método Estatístico para Bigramas – <i>Log-likelihood</i>	89
Figura 5.13 - Método Estatístico para Bigramas – Coeficiente <i>Dice</i>	90
Figura 5.14 - Método Estatístico para Trigramas – Frequência	90
Figura 5.15 - Método Estatístico para Trigramas – Informação Mútua	91
Figura 5.16 - Método Estatístico para Trigramas – <i>Log-likelihood</i>	91
Figura 5.17 - Classes de Candidatos para Unigramas – Frequência	93
Figura 5.18 - Classes de Candidatos para Bigramas – Frequência	94
Figura 5.19 - Classes de Candidatos para Bigramas – Informação Mútua	95
Figura 5.20 - Classes de Candidatos para Bigramas – Coeficiente <i>Dice</i>	96
Figura 6.1 - Expressões lingüísticas no singular, masculino e marcadores estruturais utilizados	99
Figura 6.2 - Padrões da lista de referência	100
Figura 6.3 - O método lingüístico implementado	102
Figura 6.4 - Saída do script para “obtid”	107
Figura 6.5 - Tela com as concordâncias da expressão “composiç(ão)(ões) d(o)(a)(s)”	112
Figura 7.1 - O método híbrido implementado	119
Figura A.1 - Tela com concordâncias da expressão “são”	133
Figura A.2 - Tela com concordâncias da expressão “denominad”	134

Lista de Tabelas

Tabela 3.1: Cobertura de Dicionários <i>Online</i>	42
Tabela 3.2: Amostra de 3-gramas calculados a partir da palavra pivô <i>Lei</i>	51
Tabela 3.3: 3-gramas de etiquetas correspondentes aos 3-gramas de palavras da Tabela 3.2	51
Tabela 3.4: O Espaço de Probabilidade $(\Omega, A, P[.])$	53
Tabela 3.5: Uma tabela de contingência para bigramas	53
Tabela 3.6: (n-1)-gramas e unidades textuais que estão faltando	55
Tabela 3.7: Resultados comparativos entre ambos os experimentos	60
Tabela 3.8: Resultados comparativos entre ambos os experimentos	61
Tabela 3.9: Termos Base	63
Tabela 3.10: Termos obtidos por composição	64
Tabela 3.11: Termos obtidos por modificação	65
Tabela 3.12: Concordâncias para técnicos responsáveis	66
Tabela 3.13: Ferramentas utilizadas em cada método	69
Tabela 6.1: Resultados do primeiro experimento	108
Tabela 6.2: Resultados do segundo experimento	113

Lista de Quadros

Quadro 4.1: Saída do programa count.pl	77
Quadro 6.1: Trecho da saída do método lingüístico para a expressão “constituído”	101
Quadro 6.2: Precisão, Revocação e Medida F do método lingüístico	103
Quadro 7.1: Precisão, Revocação e Medida F para o método híbrido implementado	120

Resumo

Nas últimas décadas, o grande avanço da ciência e tecnologia com suas invenções, novos materiais, equipamentos e métodos gerou a necessidade da criação de novos nomes, chamados aqui de termos, e alterações nos seus significados, para nomear adequadamente esses avanços, principalmente em áreas dinâmicas como a Ciência da Computação, a Genética e a Medicina. Dado que o desenvolvimento de repertórios terminológicos é um trabalho difícil quando realizado manualmente, lingüistas computacionais, lingüistas aplicados, tradutores, intérpretes, jornalistas científicos têm se interessado pela extração automática de terminologias (EAT) de textos. O crescimento explosivo de dados do tipo texto disponíveis na *Web* foi um fator contribuinte para a facilidade na construção de corpúscos eletrônicos de textos técnicos e científicos, propiciando a implementação de métodos de EAT. A EAT tem sido de grande interesse para todos os tipos de aplicações do Processamento de Línguas Naturais (PLN) que trabalham com domínios especializados e que, conseqüentemente, necessitam de um vocabulário especial. O objetivo desse projeto de mestrado foi avaliar métodos de EAT para o português do Brasil, ainda carente do tratamento automatizado para a criação de terminologias. Especificamente, foram implementados e avaliados métodos de EAT das abordagens estatística, lingüística e híbrida para unigramas, bigramas e trigramas a partir de um corpúscos de textos do domínio de Revestimentos Cerâmicos. Esses métodos empregam recursos simples como (a) uma *stoplist* para eliminar palavras como advérbios, (b) padrões sintáticos para os termos do domínio, por exemplo <**substantivo adjetivo**>, <**substantivo preposição adjetivo**>, levantados após a aplicação de um etiquetador *Part-Of-Speech*, (c) uma *lista de expressões e palavras* características de definições, descrições, classificações como “definido(a)(s) como”, “caracterizado(a)”, “conhecido(a)(s) como”, “significa(m)”, entre outras que são concentradoras de termos. As medidas estatísticas utilizadas nos métodos estatísticos e híbridos para indicar a relevância de termos no domínio são a informação mútua, o *log-likelihood*, o coeficiente *dice* e a frequência. Os métodos propostos foram avaliados pelas medidas de precisão, revocação e medida F, utilizando uma lista de referência da área de Revestimentos Cerâmicos. Os melhores resultados da precisão são do método híbrido para unigramas (7%), bigramas (17%) e trigramas (26%), enquanto que a revocação é melhor nos métodos puramente lingüísticos tanto para unigramas (95%) como para bigramas (90%) e trigramas (100%). Os melhores valores da medida F foram dos métodos híbridos (11%, 17% e 33% para uni, bi e trigramas, respectivamente). Esses valores, embora tenham se apresentado os mais relevantes, foram bastante inferiores àqueles normalmente encontrados na literatura que trata da EAT, cujo desempenho obtido para essa tarefa fica em torno de 60%. Esses valores motivam a busca e implementação de métodos mais avançados para tratar o português, bem como a obtenção de recursos mais elaborados, a fim de encontrar resultados mais significantes para essa tarefa, facilitando, conseqüentemente o trabalho do especialista da área, que vai analisar os candidatos a termos extraídos pelos métodos automáticos, visto que é possível fornecer a ele informações mais precisas (poucas palavras da língua geral) e completas (uma maior quantidade de termos) sobre o corpúscos considerado.

Abstract

During the last decades, the great advance in science and technology and their inventions, new materials, equipment and methods had as one result the necessity of creation of new names, called here terms, and alterations on their meanings, to name adequately these advances, mainly in areas as Computer Science, Genetics and Medicine. Considering that the development of terminological lists is an arduous work if manually executed, computational linguists, applied linguists, translators, interpreters and scientific journalists have been interested on automatic extraction of terminologies (AET) from texts. The sudden growing of data available on the Web was a contributing factor to facilitate the construction of electronic corpus of technical and scientific texts, providing implementation of AET methods. AET is very important for every sort of Natural Language Processing (NLP) applications that works on specialized domains and, consequently, needs special vocabulary. The purpose of this MS project was to evaluate AET methods for Brazilian Portuguese particularly, which is a language still in need of development of automatic treatment for terminology. Specifically, AET methods with statistic, linguistic and hybrid approaches were implemented and evaluated for unigrams, bigrams and trigrams for a corpus of texts in the domain of Ceramic Tiles. These methods use simple resources as (a) *stoplist* to eliminate words as adverbs, (b) syntactic patterns for terms from the domain, as, for instance, <**substantive adjective**>, <**substantive preposition adjective**>, considered after the application of a tagger *Part-Of-Speech*, (c) *list of expressions and words* typical of definitions, descriptions and classifications, like, for instance, “defined as”, “characterized as”, “known as” “that means”, among others that concentrate terms. The statistic measures used by statistic and hybrid methods to indicate the terms relevance in the domain are mutual information, log-likelihood, dice coefficient, and frequency. The methods proposed were evaluated by precision, recall and F-measure, using a reference list in the area of Ceramic Tiles. The best results for precision are from the hybrid method for unigrams (7%), bigrams (17%) and trigrams (26%), while for recall the best results are from purely linguistic methods for unigrams (95%) as well as for bigrams (90%) and trigrams (100%). The best values for F-measure are from hybrid methods (11%, 17% and 33% for uni, bi and trigrams, respectively). These values, although presented as the most relevant ones, were quite inferior when compared to those commonly found in the literature concerned with AET, whose performance obtained for this task is around 60%. These values motivate the search and implementation of more advanced methods for Portuguese treatment, as well as the obtainment of more elaborated resources, in order to find more significant results for this task. In this way, the work of analysis of possible terms extracted by automatic methods done by the specialist of the area becomes much easier, since it is possible to provide him/her more precise (few word from general language) and complete (greater number of terms) information about the corpus under consideration.

Capítulo 1

Introdução

Extração de informação (EI) é o processo de identificar automaticamente tipos específicos de entidades, conceitos, relações ou eventos em textos livres e armazenar esta informação de uma forma estruturada (Yangarber and Grishman, 2000). Sistemas de EI são construídos para diferentes **tarefas** como, por exemplo, identificação e classificação de nomes próprios (Appelt and Israel, 1999), extração de eventos e relações típicas de um domínio de conhecimento (Yangarber and Grishman, 2000), extração de multipalavras (Smadja, 1991; Piao et al, 2003), recuperação de idade de pessoas de um documento, localização das menções sobre um assassino em um jornal (Yangarber and Grishman, 2000) e extração de terminologia (Oh et al, 2000; Bourigault, 1992; Daille, 1996).

Uma possível aplicação da EI é no entendimento das informações expressas em um texto, a partir da extração da informação nele contida na forma de um resumo ou outra forma estruturada, para que o leitor possa, a partir daí, tirar suas próprias conclusões do texto. Extrai-se também informação de um texto para melhorar a precisão da tecnologia de recuperação da informação e para atualizar recursos léxicos. Outra utilização é na manutenção de consistência, com ajuda de dicionários terminológicos ou glossários cujos termos foram extraídos a partir de textos de um domínio, para evitar confusão terminológica e para um melhor entendimento da literatura estudada.

A avaliação do processo de extração de informação utiliza métricas clássicas da área de processamento de sinais, como a Precisão e a Revocação (Recall). **Precisão** é a razão das respostas corretas recuperadas pelo sistema e todas as respostas recuperadas e **Revocação** é a razão de respostas corretas e todas as respostas corretas possíveis (Appelt and Israel, 1999; Hobbs et al, 1997).

Um obstáculo para o uso de sistemas de extração de informação é o custo de desenvolvimento de sistemas de extração para novas tarefas. Um outro problema encontrado são os níveis de desempenho obtidos pelo processo de extração. O desempenho nestas tarefas raramente excede $F = 0.60$, em que F é uma medida que permite colocar ênfase na precisão e na revocação, sendo B um parâmetro que representa a importância relativa da precisão e revocação:

$$F = \frac{(B^2 + 1) * \textit{precisão} * \textit{revocação}}{B^2 * (\textit{precisão} + \textit{revocação})}$$

Se $B = 1$, ambos representam a mesma importância. Se $B > 1$, a precisão é mais relevante, caso $B < 1$, a revocação é mais relevante (Hobbs et al, 1997).

O crescimento explosivo de dados do tipo texto disponíveis na *Web* e as vastas quantidades de novos materiais eletrônicos propiciam a criação de novos termos e alterações nos seus significados, principalmente, em áreas dinâmicas tais como Ciência da Computação. Dado que o desenvolvimento de terminologias é um trabalho difícil quando realizado manualmente, lingüistas computacionais, lingüistas aplicados, tradutores, intérpretes, jornalistas científicos têm se interessado pela extração automática de terminologias de textos. A extração automática de terminologias (EAT) tem sido de grande interesse para todos os tipos de aplicações do Processamento de Línguas Naturais (PLN) que trabalham com domínios especializados e que, conseqüentemente, necessitam de um vocabulário especial.

O gargalo da EAT é a sua avaliação, pois exige a opinião de especialistas, sendo esse processo caro e demorado. Por outro lado, contar com recursos como glossários ou dicionários, isto é, com listas de referências, também traz seus riscos, uma vez que tais recursos são incompletos, dada a constante produção de novos termos. Uma saída pode ser o uso de outras medidas, além das tradicionais medidas de precisão e revocação, como a medida de perplexidade que mede quão bem um modelo prediz algum dado, sendo que em PLN usa-se perplexidade para comparar a predição de modelos diferentes de língua sobre um córpus (Pantel and Lin, 2001) ou avaliações em dois estágios primeiramente envolvendo várias medidas e finalmente os especialistas (Ha, 2004).

Dentro do contexto de extração de terminologia de textos, tema deste mestrado, *termos* são unidades lingüísticas, isto é, palavras ou combinações de palavras, designando conceitos ou entidades de um campo altamente especializado da atividade humana. Uma coleção de termos, relacionada com uma área de pesquisa (ou domínio) em particular, usualmente forma um sistema conceitual coerente conhecido como *terminologia* (Bolshakova, 2001). Termos compostos, que correspondem a duas ou mais unidades lexicais, são menos propensos a ambigüidade do que termos simples e aparecem em maior quantidade nos textos especializados, e são mais simples de se extrair. Termos compostos são os preferidos dos métodos de extração automática (Estopà Bagot, 1999).

No início dos anos oitenta, nota-se as primeiras tentativas de se extrair automaticamente unidades terminológicas dos textos especializados, buscando automatizar ou pelo menos semi-automatizar algumas tarefas de certas aplicações terminológicas. A criação de grandes córpus textuais informatizados, no final dos anos oitenta e no decorrer dos anos noventa, fez com que os primeiros programas de extração automática de terminologia começassem a apresentar resultados positivos.

Desde o surgimento do TERMINO¹, considerado o primeiro sistema de extração automática de candidatos a termo, diversos projetos têm sido elaborados com a finalidade de projetar extratores (semi-)automáticos de terminologia de naturezas diferentes. No entanto, mesmo com a grande quantidade de estudos realizados nesta direção, o reconhecimento e a delimitação automáticos das unidades terminológicas a partir de textos ainda não têm apresentado resultados satisfatórios.

A grande maioria dos documentos técnicos e artigos científicos contém termos que são explicitamente ou implicitamente definidos pelos autores e então usados em seus textos. Em oposição aos termos de terminologia aceita que estão fixos no dicionário, é importante que termos recém introduzidos sejam levados em conta para um processamento automático de textos científicos e tecnológicos adequado, pois tais textos apresentam grande quantidade de termos que está em uso e que ainda não foi inserida nos dicionários por ter sido introduzida recentemente ou ter escopo local de aplicabilidade. Tais termos são denominados **termos de autor**.

Em um aspecto diacrônico, não existe uma fronteira bem definida entre termos de dicionário e de autor. Usualmente, a vida de termos começa como termos de autor. Conforme os termos de autor vão sendo utilizados em vários textos de um dado campo, suas frequências aumentam, sendo que um dos critérios de conversão de termo de autor para termo de dicionário pode ser a alta frequência desses termos. As formas usadas para introduzir um termo de autor em um texto variam, resultando em três tipos diferentes de termos de autor, que são, de acordo com Bolshakova (2001): a) termo é explicitamente definido; b) termo é indefinido (sua definição está ausente), mas ele é visualmente exposto; c) termo não é nem definido nem exposto, sendo então escondido.

Estas três formas devem ser consideradas pelos métodos de extração automática. A última delas causa grande dificuldade para certos extratores, em razão de que os extratores geralmente utilizam padrões morfológicos e morfossintáticos para reconhecer e delimitar as unidades terminológicas e, o fato de tais padrões estruturais serem um filtro bastante permissivo para identificar as unidades terminológicas de um determinado domínio impede que tais extratores delimitem todos os termos dos textos especializados. Dessa forma, se forem utilizados padrões referentes somente à forma da unidade, a maioria dos candidatos a termo proposta apresentará delimitações errôneas. Por esta razão, os extratores também devem possuir conhecimento semântico a fim de detectar e delimitar automaticamente as unidades especializadas de forma mais exaustiva e precisa.

Todas as unidades léxicas têm uma frequência associada correspondendo ao número de vezes que elas aparecem em um corpus. A partir desta informação, é possível saber se uma palavra pode ou não ser um termo. Ou seja, substantivos que aparecem mais de um certo número de vezes

¹ TERMINO foi um dos primeiros sistemas de extração automática de terminologia de conhecimento lingüístico. A versão 1.0 deste sistema foi criada em 1989 para o grupo de *Recherche et développement en linguistique computationnelle* (RDLC) do Centro ATO (Analyse de textes par ordinateur) da Universidade de Quebec Montreal.

podem ser considerados termos candidatos; palavras de outras categorias devem ser mantidas a fim de completar o processamento de termos compostos. Existem, porém, estatísticas mais elaboradas para a seleção de candidatos a termos, por exemplo, Informação Mútua, Coeficiente *Log-Likelihood* (Daille, 1996) e Coeficiente *Dice*², que serão descritas neste trabalho. Uma das abordagens para a realização da tarefa de extração usa estatísticas — são os **Sistemas baseados em estatística**. Outra abordagem encontrada na literatura é a **lingüística** em que os sistemas detectam padrões recorrentes de unidades terminológicas complexas, tais como “substantivo–adjetivo” e “substantivo–preposição–substantivo”, por exemplo; e a **híbrida** em que os sistemas começam a detectar algumas estruturas lingüísticas básicas, tal como expressões nominais, e depois de os termos candidatos terem sido identificados, uma estatística relevante é usada para decidir se eles correspondem a um termo. O inverso também é possível, começando-se com uma lista de candidatos levantados estatisticamente, sendo que a informação lingüística, neste caso, é usada para filtrar termos válidos desta lista.

Neste contexto de avaliação de métodos de extração automática de termos está inserido o projeto ExPorTer. Esse projeto foi desenvolvido no *Núcleo Interinstitucional de Lingüística Computacional* (NILC), criado em 1993, sendo ele um grupo interdisciplinar dedicado à pesquisa e ao desenvolvimento de sistemas de PLN. Esse grupo de pesquisadores de lingüística e computação tem desenvolvido recursos e aplicativos para o processamento do português brasileiro, visto que alguns deles foram essenciais para o desenvolvimento desse projeto, sendo que tais recursos, aplicativos, até mesmo outros projetos poderão ser beneficiados com o projeto ExPorTer. No NILC, destacam-se o projeto ReGra: revisor gramatical automático do português brasileiro, apoiado pela FAPESP, CNPq, Finep e Itaotec-Philco S.A., que originou um produto comercializado pela Itaotec e também distribuído com o MS-Office (português) desde 2000; o analisador sintático CURUPIRA³ para o português brasileiro; o projeto de um *thesaurus* do português brasileiro e de uma base de dados lexical; o projeto Universal Networking Language (UNL), patrocinado pelo Instituto de Estudos Avançados da Universidade das Nações Unidas, para o qual o NILC constrói ferramentas de codificação e decodificação de português, apoiado pela FAPESP e, em especial, o projeto Lacio-Web, do edital de Conteúdos Digitais do CNPq, que se propõe a construir e disponibilizar *Cópus do Português* e Ferramentas *Web* de Navegação e Auxílio para Análise Lingüística.

O NILC também possui um etiquetador e um tokenizador (um módulo do CURUPIRA), que foram essenciais para execução de algumas tarefas desse projeto de mestrado.

² <http://www.d.umn.edu/~tpederse/Group01/bsp.txt>

³ <http://www.nilc.icmc.usp.br/nilc/tools/curupira.html>

1.1 Motivação

A tarefa de extração de termos é crucial para várias aplicações, tais como Recuperação de Informação (RI), Sumarização Automática, Indexação e Classificação de Textos, Alinhamento de Textos Bilíngües, Tradução, Recursos Básicos de PLN, Edição Suportada por Computador, Corretores Gramaticais e Geração de Língua Natural (LN).

Na indexação de texto, os termos extraídos constroem um índice remissivo, enquanto que, na extração de informação, as consultas de usuários são respondidas depois que é realizada a comparação dos termos da consulta com os termos dos textos na coleção. Textos que apresentam conjuntos de termos iguais ou similares são classificados dentro do mesmo domínio no campo de classificação de texto. Já na área de pesquisa em alinhamento de textos bilíngües, termos de uma língua são usualmente traduzidos uniformemente em outra língua, dentro do mesmo domínio (Georgantopoulos and Piperidis, 1998).

Embora estas últimas aplicações apresentem diferentes naturezas, é possível observar a importância que a extração de termos exerce sobre elas.

1.2 Objetivos

Um dos objetivos deste trabalho foi a obtenção de um conjunto de informações significativas a respeito dos pontos fortes e fracos de vários métodos de extração automática de termos. Para isso, foi realizado um levantamento bibliográfico dos sistemas existentes de extração de informação em geral e de extração de termos em particular. O objetivo principal deste trabalho foi analisar e implementar métodos de extração de terminologia das três abordagens (lingüística, estatística e híbrida) e, através de uma avaliação deles, usando as medidas de precisão, revocação e medida F sob um mesmo córpus e lista de referência, eleger o melhor para textos em português para o domínio de Revestimentos Cerâmicos.

1.3 Organização do texto

O Capítulo 2 apresenta um histórico da Terminologia, além de pesquisas em Terminologia realizadas no Brasil e exemplificação de produtos terminológicos agrupados sob as características: possuir equivalência em mais de uma língua, ser aberto (disponibilização *online* com interação com o usuário) *versus* fechado e forma de apresentação em ordem alfabética *versus* com recursos extralingüísticos. No Capítulo 3, são descritos dois trabalhos sobre a abordagem lingüística e outros dois sobre a híbrida, além de cinco medidas estatísticas para a extração de termos. O Capítulo 4 apresenta os recursos e ferramentas que auxiliaram no desenvolvimento desse trabalho. Os Capítulos 5, 6 e 7 foram dedicados à implementação e avaliação dos métodos das abordagens

estatística, lingüística e híbrida, respectivamente. E, para finalizar, é apresentada uma breve conclusão, trazendo os resultados obtidos com esse projeto de mestrado, bem como suas contribuições e limitações, propondo trabalhos futuros a fim de desenvolver métodos mais elaborados para obtenção de resultados de precisão e revocação mais significativos.

Capítulo 2

Terminologia e áreas afins

A Teoria da Terminologia está passando por transformações nos dias atuais, criando posicionamentos controversos quanto à constituição e ao tratamento dos léxicos terminológicos (Krieger, 2001). Estas transformações se devem ao avanço da ciência e tecnologia as quais já passaram a fazer parte do dia-a-dia das pessoas e, por esta razão, é necessário que a Terminologia acompanhe este avanço, fornecendo à ciência e à tecnologia denominações para seus novos conceitos e invenções.

No uso informal da linguagem a precisão técnica é dispensável, enquanto que no uso especializado tal precisão é fundamental para que os termos estejam de acordo com a estruturação conceitual de um determinado domínio. Isto decorre da necessidade de cientistas efetuarem uma comunicação técnica a respeito de assuntos inseridos dentro de seu domínio específico.

2.1 Um pequeno histórico da Terminologia

A Terminologia não é uma disciplina recente. Identifica-se a existência da prática terminológica desde o século XVIII com os trabalhos de Lavoisier e Berthold no domínio da química ou Linné na botânica e zoologia (Cabré, 1993, p.21) apud (Almeida, 2000). Especialistas de várias áreas interessaram-se em dar um espaço à Terminologia devido ao fato de sentirem a necessidade de, em suas respectivas áreas científicas, denominar e relacionar os conceitos científicos.

Já no século XX, o interesse dos especialistas das diversas áreas não se resume apenas em relacionar denominações a conceitos, mas também em denominar conceitos novos e harmonizar as novas denominações. Isto decorre do grande avanço tecnológico e científico e, conseqüentemente, do surgimento de diversas conceituações as quais necessitavam de uma intensa e acelerada denominação e organização técnica dentro de suas áreas específicas de estudo (Cabré, 1993, p.22) apud (Almeida, 2000).

Wüster, engenheiro e pai da Teoria Geral da Terminologia (TGT), concebe a linguagem científica e técnica como a língua de uso em oposição à “língua literária”, tendo ele o objetivo de transformar essa linguagem técnica em um instrumento eficaz. Propôs, então, o método da normalização para que houvesse uma organização consciente da língua (Felber, 1996, p.19) apud (Almeida, 2000). Ele considerou o termo como um rótulo lingüístico de um conceito. Segundo esta visão clássica, em um domínio cada conceito é, idealmente, associado a um termo o qual é seu rótulo. Isto é o que defende a Teoria Geral da Terminologia. Wüster também define termo como: “Uma unidade terminológica que consiste em uma palavra à qual se atribui um conceito como seu significado, ao passo que para a maioria dos lingüistas atuais, a palavra é uma unidade inseparável

composta de forma e conteúdo”. Com esta definição, Wüster destaca o papel conceitual o qual o termo dá a uma palavra da língua, fornecendo a ela um significado específico. Wüster analisa o termo como tendo um papel denominativo. Wüster (1998) apud Krieger (2001) entende que, primeiramente, tem-se a nomenclatura para, a partir daí, poder-se desenvolver a conceituação acerca do objeto estudado. No entanto, caso a nomenclatura realmente preexistisse ao conceito, não haveria motivação para que o léxico ordinário pudesse ser denominado como um termo técnico ou um termo científico. Só se dá nome ao que já existe, seja na sua substância ou na sua função.

Apesar de sua sistematicidade e coerência para a comunicação padronizada, que é apenas uma das possibilidades da comunicação real, a TGT apresenta princípios pouco satisfatórios no âmbito da comunicação real (Cabré, 1999, p.129) apud (Almeida, 2000).

Nos anos noventa começam a surgir críticas a TGT. De acordo com Cabré et al. (1998, p.36-7) apud (Almeida, 2000), há alguns pontos que identificam a insuficiência da TGT no contexto terminológico científico e técnico atual. São eles: o Logicismo, característica que leva o método de análise a verificar apenas a lógica da realidade, sendo que o que foge a esta realidade não é suficientemente descrito; o Universalismo, o qual faz com que um mesmo termo possa ser utilizado em diferentes áreas e contextos com um significado único, o que pode causar ambigüidade, pouca especificidade e até contradição; o Estatismo, fazendo com que a unidade lexical especializada não tenha o seu significado atualizado conforme evolui a ciência; o Reduccionismo, que é um marco da TGT pelo fato desta não encontrar em si mesma a capacidade de ampliar o significado de um termo científico ou técnico; o Idealismo, que é a crença por parte dos cientistas seguidores da TGT de que um termo surge de um único conceito o qual não se modifica, não foge à idéia inicial.

Em razão dessa insuficiência apresentada pela TGT, alguns lingüistas passaram a substituí-la por uma teoria mais ampla e flexível, a Teoria Comunicativa da Terminologia (TCT), cujo método explica melhor os fenômenos que envolvem a comunicação especializada e também melhor descreve suas unidades representativas, os termos, de forma a abranger toda a sua complexidade.

Esta nova linha de pesquisa está situada entre duas concepções antagônicas sobre a constituição e a natureza das terminologias: a idéia de que termos são unidades de conhecimento e a compreensão de que as terminologias são unidades lexicais, e portanto, componentes naturais dos sistemas lingüísticos. Dessa forma, termos são, ao mesmo tempo unidades lexicais (vocábulos) e ferramentas para estudos científicos (Krieger, 2001).

Segundo Krieger (2001), a unidade lexical básica da terminologia é referenciada de várias formas: *termo técnico-científico*, *termo*, *terminologias*, *unidade lexical terminológica*, *unidade lexical especializada*, *unidade lexical temática*, *vocabulário*.

Segundo Cabré (1993) apud Krieger (2001), “... os termos não formam parte de um sistema independente das palavras, mas que conformam com elas o léxico do falante, mas ao mesmo tempo,

pelo fato de serem multidimensionais, podem ser analisados de outras perspectivas e compartilham com outros signos de sistemas não lingüísticos o espaço de comunicação especializada”. De acordo com Cabré (1993) apud Krieger (2001), os termos não são neologismos mas sim palavras já existentes. Nota-se esta afirmação no dizer da autora “... não formam parte de um sistema independente de palavras...”.

Sendo assim, termos são componentes lexicais ordinários, os quais foram adaptados a um sentido técnico ou científico através de um processo de terminologização a partir da necessidade de um desenvolvimento mais profundo em um determinado campo de estudos. No entanto, há de se observar que esta necessidade surge somente após uma idéia já existente. Não se denomina o que não existe. A noção, pois, pressupõe a denominação.

Existem outras áreas relacionadas à Terminologia. Segundo Teline et al (2003), “a informática e a Terminologia estão ligadas de forma a facilitar o armazenamento e a difusão de dados terminológicos na elaboração de grandes bases de dados especializados, denominadas bancos de terminologia”. A integração entre as duas áreas denomina-se Terminótica, enquanto que a Terminografia é “a atividade de recenseamento, constituição, gestão e difusão dos termos nos campos científicos conforme a necessidade de seu uso”. Já a Terminologia, de acordo com Almeida (2000, p. 36), é “o conjunto de práticas e métodos utilizados na compilação, descrição e apresentação dos termos de uma determinada linguagem de especialidade” e também “o conjunto de premissas e argumentos necessários para a explicação das relações entre os conceitos e os termos, relações estas que são fundamentais para a atividade terminológica coerente”.

2.2 Pesquisas em Terminologia no Brasil

Existem vários grupos de pesquisa em terminologia em várias universidades brasileiras. Por exemplo, na UFRGS, o projeto TERMISUL⁴, realizado em conjunto com pesquisadores argentinos visa ao desenvolvimento da pesquisa terminológica pura e aplicada. Suas atividades envolvem a elaboração de dicionários terminológicos, glossários técnicos, ferramentas informatizadas para o tradutor técnico e assessoria à elaboração de produtos terminológicos. Uma realização de tal projeto é o “Dicionário de Direito Ambiental: Terminologia das leis do meio ambiente”, que tem por objetivo auxiliar profissionais cuja atuação está relacionada à temática jurídico-ambiental. Ele reúne um conjunto de informações, obedecendo a princípios teóricos e metodológicos da teoria da Terminologia e compreende o exame de aspectos morfológicos e semânticos dos termos, até seu funcionamento pragmático e discursivo analisado a partir do estatuto do Direito Ambiental. A seleção dos termos foi determinada pela natureza multidisciplinar da área, sendo que estes termos se

⁴ www.ufrgs.br/termisul/

constituem no pilar organizacional desta obra pioneira na bibliografia jurídica e singular no universo das línguas latinas. A dimensão lingüística orientou a identificação e a seleção dos termos repertoriados, constituindo-se em um processo complexo, em razão da ausência de fronteiras rígidas entre o léxico especializado e o da língua comum, o que vem a ser um problema crucial para o reconhecimento das terminologias no campo das ciências humanas e sociais. O Dicionário de Direito Ambiental apresenta dois mil e cinco verbetes, com termos equivalentes em espanhol e inglês, cobre 64 anos de produção legislativa brasileira (1934-1998) e a Lei de Base do Ambiente de Portugal e sua legislação complementar.

A versão eletrônica deste dicionário é o TermDic⁵ (Morales, 2001), contendo um banco de dados com 2005 fichas terminológicas, correspondentes à totalidade dos verbetes do livro. O objetivo do TermDic é proporcionar uma forma informatizada do Dicionário, contando com recursos adicionais, entre eles: distribuição dos termos por assuntos; pesquisa e filtragem das fichas terminológicas por diversos critérios (termos completos ou segmentos de termos, texto da definição, assunto, ocorrência, etc.); e impressão de listagem de termos. As entradas dos termos estão organizadas em uma lista e, a partir desta, pode-se escolher um termo e visualizar somente um campo específico ou sua ficha terminológica. É possível imprimir o verbete na ficha terminológica ou copiar um ou mais campos para colar em um outro programa (por exemplo, um editor de texto). Também é possível acessar as remissivas do termo com um único clique.

Além dessas características, é importante salientar que o TermDic apresenta facilidade de uso, que é proporcionada por uma interface simples e amigável em língua portuguesa e um sistema de ajuda para usuários inexperientes.

Um outro grupo de pesquisa criou o Centro Interdepartamental de Tradução e Terminologia da USP – CITRAT⁶, que traz no seu site vários glossários de diversos domínios, que são o resultado de trabalhos de alunos do Curso de Especialização em Tradução (CETRAD) da USP. A disponibilização desses glossários no site do CITRAT tem por finalidade oferecer uma ferramenta de auxílio aos tradutores, e não apresentá-los como fruto de um trabalho terminológico no sentido científico do termo. Cada glossário consta de 100 termos aproximadamente em cada língua, apresentando exemplos autênticos, sendo que estes exemplos não são necessariamente definitórios e nem traduções uns dos outros. Assim, cada exemplo tem somente a intenção de contextualizar o uso do termo em questão e atestar seu uso em textos autênticos, de diversos tipos, e a maioria deles é extraída da Internet, podendo haver incorreções ou lacunas. Alguns dos glossários encontrados no site versam sobre: automação industrial, biodiversidade, biotecnologia, culinária, ecoturismo, finanças, informática, medicina veterinária, moda, odontologia, propaganda e marketing.

⁵ <http://www.ufrgs.br/termisul/termdic.html>

⁶ www.fflch.usp.br/citrat/index.htm

Na Universidade Federal de São Carlos, Departamento de Letras existe um grupo de pesquisa liderado pela professora Gladis M. B. Almeida atuando na aplicação da TCT para a criação de produtos terminológicos; na Unesp de Rio Preto, na UnB e na UFU existem também outros grupos.

2.3 Produtos terminológicos

Existem tantos tipos de recursos quanto formas de aplicações: *thesauri* para indexação automática e recuperação de informação, indexação estruturada para *hyper* documentos, listas especializadas para escrita controlada auxiliada por computador, listas de termos bilíngües para tradução automática, ontologias para Inteligência Artificial, palavras-chave estruturadas para bibliotecas digitais, dicionários e glossários terminológicos. Estes dois últimos recursos serão enfocados nesta seção.

A nomenclatura do dicionário de língua geral tende a abarcar a totalidade das palavras que compõem o léxico comum de uma língua, concentrando-se, de maneira especial, nas formas correntes na época de sua elaboração. Já um dicionário terminológico é entendido como um dicionário de termos de uma área específica do conhecimento ou da experiência humana (Maciel, 1996). Por outro lado, o termo glossário refere-se, geralmente, a termos de um texto, constituindo-se em um repertório das principais palavras de um texto, elencadas ao final de uma obra. Os termos e expressões encontrados em dissertações e teses, por exemplo, devem vir reunidos sob o título de glossário.

Um dicionário terminológico, apesar de ser próprio de uma técnica ou ciência, não é capaz de fornecer cobertura de termos suficiente às mesmas, pois existem subáreas inseridas nestas técnicas ou ciências, sendo que o objetivo do dicionário especializado consiste no tratamento dos termos “superficialmente”, ou seja, não é possível que ele apresente todos os termos das subáreas (que muitas vezes também se subdividem), por haver uma quantidade imensa de termos presentes em cada uma. Para resolver este problema, glossários são construídos a fim de proporcionar cobertura aos termos existentes nas subáreas, de forma bem mais específica do que se encontra nos dicionários terminológicos.

Existem diferentes variedades de glossários e dicionários terminológicos seguindo as características: a) possuir equivalência em mais de uma língua, b) aberto (disponibilização *online* com interação com o usuário) *versus* fechado e c) forma de apresentação em ordem alfabética *versus* com recursos extralingüísticos. Para a primeira característica, dicionários ou glossários que contemplam mais de uma língua, como é o caso da Figura 2.1.

Termo Inglês	Exemplo Inglês	Termo Português	Exemplo Português
Family	Genera are aggregated into families, families into orders, orders into classes, and so up the hierachy (BISBY, 1995:28)	Família	As espécies fazem parte de unidades maiores - os gêneros - que por sua vez se agrupam para formar as famílias. (PEREIRA, 1980:50)
Genus	As the common names sometimes imply, some species are clearly members of recognizable larger aggregations (or the descendants of a common ancestral form) known as genera (singular genus): e.g. date palm, canary date palm, dwarf date palm - species in the date palm genus Phoenix (BISBY, 1995:28)	Gênero	As espécies fazem parte de unidades maiores - os gêneros - que por sua vez, se agrupam para formar famílias. (PEREIRA, 1980:50)
Class	Genera are aggregated into families, families into orders, orders into classes, and so up the hierachy (BISBY, 1995:28)	Classe	As ordens se reúnem em classes e estas em filou ou divisões. (PEREIRA, 1980:50)
Order	Genera are aggregated into families, families into orders, and so up the hierachy (BISBY, 1995:28)	Ordem	Um conjunto de famílias constitui uma ordem. (PEREIRA, 1980:50)
Species	A species is a group of organisms that recognize each other for the purpose of mating and fertilization.(BISBY, 1995:42)	Espécie	O Centro de Pesquisas para a Conservação de Aves Silvestres (CEMAVE) atua desde 1977 com o objetivo de promover a conservação e o manejo de aves silvestres com ênfase nas espécies migratórias... PRNCDB-Br
Intraspecific competition	Competition within the same kind may be viewed as struggling for the opportunity of filling one of the niches vacated by death in the next generation, or intraspecific competition.(COLINVAUX, 1993: 140)	Competição intra-específica	A competição pode ser intra-específica ou interespecífica. (DAJOZ, 1986:186)
Stepping stones	Stepping stones are ecologically suitable patches where an organism can temporarily stop while moving between habitat patches (modified from Forman, 1995). (METZGER & PEREIRA, 1997: 412)	Pontos de ligação	...os fluxos entre a1 e a3 serão maiores para as espécies que são capazes de utilizar corredores finos e "pontos de ligação" (stepping stones). (METZGER, 1999:446-463)

Figura 2.1 - Glossário de biodiversidade⁷.

Neste caso, o glossário apresenta um determinado termo e seu exemplo em duas línguas, primeiramente em inglês (1ª e 2ª colunas) e, em seguida, o termo e seu exemplo em português (3ª e 4ª colunas). Também existem os glossários (e dicionários) que apresentam mais de uma língua de entrada, como os exemplos encontrados nas Figuras 2.2 e 2.3, em que o termo algumas vezes se apresenta em inglês, português, e até mesmo em inglês e português, sendo que a definição do termo é mostrada em português.

BLACK OUT
Período de embargo, ou seja, bloqueado para utilização. Utilizado para bloquear períodos impedidos para voar como milhagem.

BOARDING PASS OU CARTÃO DE EMBARQUE
instrumento que prova que um passageiro realmente voou. É por meio dele que credita-se a milhagem que não foi inserida antes da viagem. Nele estão as informações de embarque.



Status da reserva
a situação da reserva pode ser:
OK = confirmada
RQ = requisitada
WL = lista de espera
RR = reconfirmada
OPEN = em aberto.

Figura 2.2 - Glossário de turismo⁸.

⁷ <http://www.fflch.usp.br/citrat/>

⁸ <http://www.uol.com.br/folha/turismo/preparese/glossario.shtml>

A	B	C	D	E	F	G	H	I	J	K	L	M
N	O	P	Q	R	S	T	U	V	W	X	Y	Z

U	
UEE	Sigla de União Econômica Européia. Designa o mercado comum da Europa.
Underwriters	Instituições financeiras especializadas em operações de lançamento de ações no mercado primário. No Brasil, tais instituições são, em geral, bancos múltiplos ou bancos de investimento, sociedades distribuidoras e corretoras que mantêm equipes formadas por analistas e técnicos capazes de orientar os empresários, indicando-lhes as condições e a melhor oportunidade para que uma empresa abra seu capital ao público investidor, por meio de operações de lançamento.
Underwriting	A tradução literal é subscrição. Os bancos de investimento montam operações financeiras nas quais intermediam a colocação (lançamento) ou distribuição de ações, debêntures ou outros títulos mobiliários, para investimento ou revenda no mercado de capitais, recebendo uma comissão (fee) pelos serviços prestados, proporcional ao volume do lançamento.
Unidade Monetária Européia	Índice composto por 10 diferentes moedas européias.

Figura 2.3 - Glossário de termos do mercado financeiro⁹.

Uma variedade bastante interessante de dicionários (e glossários) está relacionada com a característica *online*, anteriormente citada, visto que tais dicionários apresentam a funcionalidade de permitir que usuários adicionem conteúdos não presentes em tais dicionários (glossários), ou indicar a falta de termos que serão posteriormente inseridos. Exemplos destes são os dicionários encontrados nas Figuras 2.4 e 2.5.

⁹ <http://www.investshop.com.br/ajd/glossario.asp?letra=U>

averted vision

Dictionary of Astronomy

Keywords: none

When you look squarely at something, you are using a part of the retina of your eye that is not as sensitive to low light levels as the parts that are off to the side. Thus to see faint objects, don't look straight at them. Center them in the field of view of your telescope, but fix your stare part way out to the edge of the field. People sometimes ask which way to avert -- that is, which way away from the center of the field to move their gaze. Try several.

Added: 2000/06/16 at 18:01:20

none

Top ▲

back focal length

Dictionary of Astronomy

Keywords: none

The distance from the last optical element of a system encountered by the light passing through it, to the focal plane. Opinions differ on whether flat mirrors and diagonal prisms count as "optical elements", for determining back focal length. I think they do not. With classical all-lens optical systems, like refractor objectives and many camera lenses, back focal length is approximately what you get if you put a measuring stick against the lens cell and measure the distance to the focal plane. But with things like Schmidt-Cassegrains, the term is more confusing -- for that system, back focal length is measured from the secondary, which is way up inside the tube.

Added: 2000/06/16 at 18:01:20

none

Figura 2.4 - Dicionário de astronomia¹⁰.

¹⁰ <http://www.yourdictionary.com/diction5.html>

Siglas

Dicionário médico de SIGLAS (iniciais), ACRÓNIMOS (extremidades de palavra) e ABREVIATURAS. Caso detecte alguma falta nesta lista, por favor, envie-nos sugestões de correcção para ✉ info@opapeldomedico.com

Tudo/ A/ B/ C/ D/ E/ F/ G/ H/ I/ J/ K/ L/ M/ N/ O/ P/ Q/ R/ S/ T/ U/ V/ W/ X/ Y/ Z

Pesquisar



#, número de ordem

#, fractura óssea

μ, micron

μCi, microcurie

μg, micrograma

μL, microlitro

Figura 2.5 - Dicionário médico de siglas¹¹.

A maioria dos dicionários e glossários mais comuns está organizada em ordem alfabética (ordem de apresentação). Exemplos de dicionários e glossários que se encaixam nessa característica são encontrados nas Figuras 2.2, 2.3, 2.6, 2.7, 2.8 e 2.9.

¹¹ <http://www.opapeldomedico.com/siglas.asp>

Glossário da infertilidade (de A a E)

Aborto Espontâneo: A interrupção da gravidez ate vinte semanas da gestação e ou o feto deve estar pesando menos que 500 gramas.

Aborto habitual: Um termo que se refere a uma condição em que a mulher teve três ou mais abortos.

Aborto Incompleto: Um aborto depois do qual restos ovulares remanesçam dentro do útero. Neste caso uma Curetagem deve ser realizada para remover o material evitando complicações.

Aborto Retido: Um aborto onde o feto morre dentro do útero mas não há nenhum sangramento ou sinais de deslocamento. Uma Curetagem será necessária para remover o feto morto e prevenindo complicações.

Aborto Terapêutico: Um procedimento usado para interromper uma gravidez antes que o feto possa sobreviver por si só.

Aborto: A morte do feto entre a vigésima semana de gestação e o nascimento.

Aborto: Perda espontânea do embrião ou feto .

Figura 2.6 - Glossário de infertilidade¹².

¹² <http://www.abdelmassih.com.br/drvida/glossario1.html>

Glossário

A | B | C | D | E | F | G | H | I | J | K | L | M |

N | O | P | Q | R | S | T | U | V | W | X | Y | Z |

Abiótico: é o componente não vivo do meio ambiente. Inclui as condições físicas e químicas do meio.

Aceiro: prática utilizada por bombeiros e agricultores no combate e prevenção de incêndios florestais. Consiste numa faixa de terra aberta em volta da área que está sendo queimada ou que se quer proteger, mantida livre de vegetação, com capina ou poda, a qual impede a invasão do fogo.

Adubo verde: vegetal incorporado ao solo com a finalidade de adicionar matéria orgânica que vai se transformar, parcialmente, em húmus, bem como em nutrientes para a planta. Os adubos verdes podem consistir de ervas, gramíneas, leguminosas, etc.

Aeróbico: ser ou organismo que vive, cresce ou metaboliza apenas em presença do oxigênio.

Figura 2.7 - Glossário ambiental¹³.

a b c d e f g h i j k l m n o p q r s t u v w
x y z

Fator anti-hemofílico (AHF): Nome genérico dos produtos de fator VIII para terapia de reposição na hemofilia A.

Fator de coagulação: substância de natureza proteica que se encontram no plasma, normalmente de forma inativa, e que são responsáveis pelo processo de coagulação (hemostasia). Ao todo são treze os fatores de coagulação.

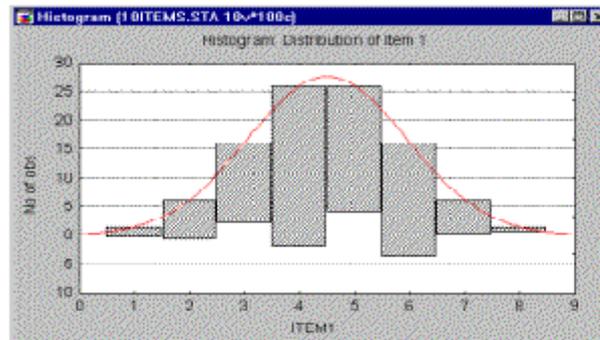
Fracionamento pelo método de Cohn: Método para separar e purificar o plasma em seus componentes, baseado na solubilidade em etanol, sob baixas temperaturas.

Figura 2.8 - Glossário de imunologia¹⁴.

¹³ <http://www.sema.rs.gov.br/sema/html/glossa.htm>

¹⁴ <http://www.aventisbehring.com/brazil/Functions/Glossary/portuguese/index.asp?Alphachar=f>

Hanging Bars Histogram. The hanging bars histogram o identify the areas of the distribution where the discrepancies occur. While the standard way of presenting the normal distr fitting normal curve over a histogram, the hanging bars histog observed frequencies for consecutive ranges of values from t



If the investigated distribution can be well approximated by t straight, horizontal line.

Harmonic Mean. The *Harmonic Mean* is a "summary" st

$$H = n * 1/\sum(1/x_i)$$

Figura 2.9 - Glossário de estatística¹⁵.

Além desses dicionários e glossários, aqueles que incluem siglas também se caracterizam pela forma de apresentação (ordem alfabética), como é o caso do dicionário médico de siglas (Figura 2.5), que tem como objetivo definir as siglas encontradas no domínio da medicina.

No entanto, alguns glossários (e dicionários) se apresentam através de recursos extralingüísticos, possuindo a organização própria do domínio, permitindo assim, um melhor entendimento dos termos que são mostrados. No dicionário de biologia celular e molecular (Figura 2.10), um desenho é apresentado a fim de identificar a função de cada unidade, permitindo uma boa visualização do processo que acontece nas células e moléculas. Já o glossário de direito ambiental internacional, que se encontra na Figura 2.11, apresenta um mapa de atos internacionais, proporcionando, dessa forma, o conteúdo de maneira concisa, auxiliando, assim, a compreensão do mesmo.

¹⁵ <http://www.statsoftinc.com/textbook/glosfra.html>

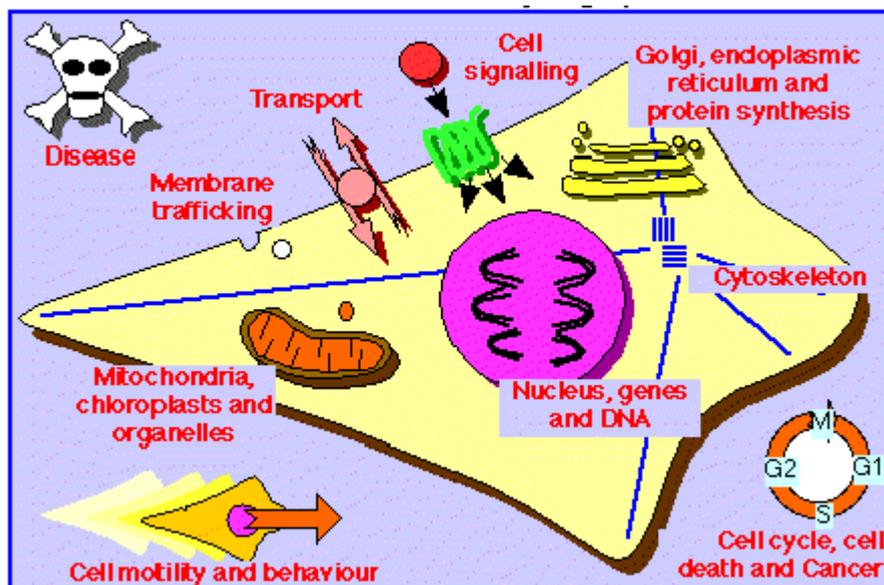


Figura 2.10 - Dicionário de biologia celular e molecular¹⁶.

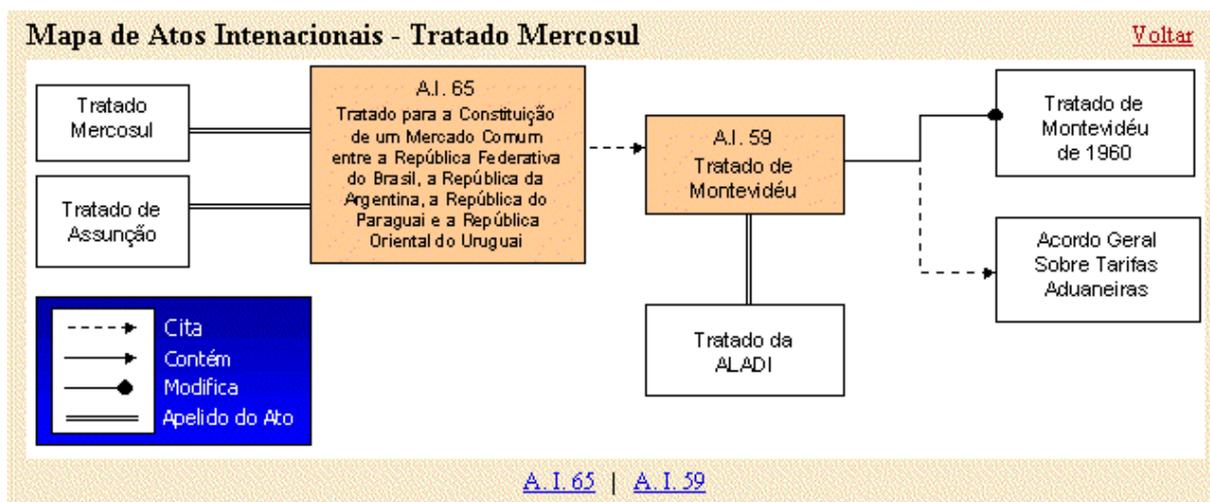


Figura 2.11 - Glossário de direito ambiental internacional¹⁷.

Outros dicionários, que também apresentam recursos extralingüísticos (incorporam figuras), como nos casos das Figuras 2.2 e 2.9, que são incluídos com o intuito de permitir um bom entendimento do termo que está sendo descrito.

¹⁶ <http://on.to/dictionary>

¹⁷ <http://www.ufrgs.br/termisul/mapapai/mapapai.html>

Capítulo 3

Métodos para extração de termos técnicos

Dados o grande volume de informação técnica disponível nesta última década e, o crescente uso da WWW como fonte de pesquisa e depósito de textos técnicos e científicos, esforços manuais para a extração de terminologia de *cópus*¹⁸ se tornaram ineficazes. Atualmente, os sistemas de extração automática de terminologia têm sido largamente utilizados, considerando a importância desses para muitas aplicações tais como tradução humana ou automática, indexação, construção de *thesaurus*, organização do conhecimento, entre outras. Um sistema de extração automática de candidatos a termo¹⁹ (SEACAT) é formado por um conjunto de programas para o reconhecimento de unidades terminológicas de *cópus* (Estopà Bagot, 1999). O principal objetivo dos SEACAT é a automatização da fase de seleção de todas as unidades terminológicas de um texto especializado, proporcionando, assim, rapidez e sistematicidade ao trabalho terminológico. Dada a importância de tais sistemas, este capítulo é dedicado a eles.

Os SEACAT são tradicionalmente classificados conforme a metodologia que utilizam para reconhecer as unidades terminológicas, em sistemas que:

- utilizam apenas métodos baseados em conhecimento estatístico;
- utilizam apenas métodos baseados em conhecimento lingüístico;
- utilizam métodos baseados em conhecimento estatístico e lingüístico.

Os métodos baseados em conhecimento estatístico geralmente detectam as unidades terminológicas de acordo com a frequência em que elas ocorrem em um *cópus*. Existem métodos estatísticos que utilizam desde simples frequências a estatísticas mais complexas, cuja função é identificar os candidatos a termo.

Os métodos estatísticos são dependentes do tamanho do *cópus* que utilizam, diferentemente dos métodos lingüísticos. Dessa forma, se o *cópus* de aplicação é pequeno, gera-se muito silêncio, que consiste no número de termos não encontrados do total de termos existentes em um texto; mesmo quando o *cópus* apresenta milhões de ocorrências, há sempre uma porcentagem de palavras que não podem ser recuperadas em razão de sua baixa frequência de uso no *cópus*.

Os métodos estatísticos também são responsáveis por gerar bastante ruído, que corresponde ao número de candidatos a termo que não apresenta valor terminológico, isto é, aquelas palavras que não apresentam significado especializado em textos especializados, sendo pertencentes à língua

¹⁸ Sejam eles construídos de textos da WWW ou textos impressos.

¹⁹ Os termos candidatos devem ser, posteriormente, validados por humanos.

geral. Muitas dessas palavras aparecem nos textos com uma alta frequência, sendo responsáveis pela geração da grande quantidade de ruído (Estopà Bagot, 2001).

Embora possuam os problemas levantados acima, os métodos estatísticos são independentes da língua, sendo essa mais uma característica que os diferencia dos métodos lingüísticos.

Os sistemas baseados em conhecimento lingüístico utilizam diferentes recursos que contêm diferentes informações lingüísticas para a extração dos termos, que são:

- Lexicográficas: dicionários de termos e lista de palavras auxiliares (“stopwords”)
- Morfológicas: padrões de estrutura interna da palavra
- Morfossintáticas: categorias morfossintáticas e funções sintáticas
- Semânticas: classificações semânticas
- Pragmáticas: representações tipográficas e informações de disposição do termo no texto

Este tipo de conhecimento utilizado faz com que os sistemas baseados em conhecimento lingüístico se apliquem somente a uma língua e às vezes até mesmo a uma única variante, pois a sua utilização em textos em uma língua diferente exige um estudo lingüístico prévio e necessita de um novo projeto para alguns dos módulos do sistema.

De acordo com a Estopà (1999), a grande quantidade de ruído gerada (entre 55% e 75%) é um dos problemas principais dos sistemas que trabalham apenas dados morfológicos, morfossintáticos, sintáticos e/ou léxicos. Nem todas as palavras que são consideradas pelo sistema como unidades terminológicas polilexicais o são, já que a maioria dos mesmos padrões corresponde também a unidades léxicas e fraseológicas que não apresentam uso especializado. Em alguns casos elas correspondem a unidades de uso especializado, mas não terminológicas, como as unidades fraseológicas especializadas ou as combinações que apresentam muitas recorrências (colocações²⁰); e em outras vezes tais padrões são expressões discursivas (como “o objetivo deste trabalho”, “nesta seção”), sem caráter especializado.

Por essa razão, pesquisadores compartilham da idéia de que o emprego de algum tipo de conhecimento semântico é a única forma de reconhecer e delimitar as unidades terminológicas de um texto especializado.

Os sistemas baseados em conhecimento híbrido utilizam o conhecimento estatístico juntamente com o lingüístico. A aplicação do conhecimento híbrido torna o sistema mais eficiente, visto que ele condiciona os resultados.

Existem dois tipos de métodos híbridos: aqueles que aplicam o conhecimento estatístico primeiro e depois o lingüístico, e aqueles que utilizam a estatística apenas como um complemento

²⁰ Colocações são expressões que consistem de duas ou mais palavras que co-ocorrem em um texto; correspondem ao modo mais natural de expressar conceitos.

da lingüística. No primeiro caso, acontecem os mesmos problemas de silêncio encontrados nos sistemas puramente estatísticos. Já no segundo, os resultados finais podem se apresentar melhores em razão da estatística auxiliar no momento do processo de detecção, reafirmando ou recusando a condição de termo de uma unidade lingüística.

Se for considerado não somente o tipo de conhecimento que os extratores utilizam, mas também a diversidade de informação utilizada e a ordem de utilização, obtém-se uma proposta de classificação mais elaborada²¹ dos sistemas de extração automática de termos:

²¹ Esta proposta combina a de Estopà Bagot (1999) com informações utilizadas nos sistemas apresentados nas próximas seções.

- | |
|---|
| <ul style="list-style-type: none">a. Métodos que usam conhecimento estatístico<ul style="list-style-type: none">a.1 Cálculo de frequênciasa.2 Informação mútuaa.3 Coeficiente <i>log-likelihood</i>a.4 Coeficiente <i>dice</i>a.5 <i>C-value</i>b. Métodos que usam conhecimento lingüístico<ul style="list-style-type: none">b.1 Informações morfoestruturaisb.2 Informações sintáticasb.3 Informações semânticasb.4 Informações pragmáticasc. Métodos de conhecimento híbrido<ul style="list-style-type: none">c.1 Primeiro conhecimento estatístico e depois lingüísticoc.2 Primeiro conhecimento lingüístico e depois estatístico |
|---|

Nesse capítulo serão apresentadas cinco estatísticas léxicas utilizadas em sistemas estatísticos, que são: Frequência de ocorrência, Informação mútua, Coeficiente *log-likelihood*, Coeficiente *dice* e *C-value*, e dois exemplos de métodos lingüísticos e outros dois de híbridos, baseados, respectivamente, nos trabalhos de Heid et al (1996), Klavans and Muresan (2000; 2001a; 2001b); Frantzy and Ananiadou (1997) e Dias et al (2000b).

3.1 Abordagem estatística

Em razão da escassez de trabalhos atuais descrevendo sistemas ou algoritmos estatísticos, o que se justifica pelo maior uso da abordagem híbrida atualmente, nesta seção serão apresentadas apenas as principais medidas estatísticas. Tais medidas são mostradas a seguir juntamente com referências nas quais foram descritas detalhadamente:

- Frequência de ocorrência (Manning and Schütze, 1999; Daille, 1996)
- Informação mútua (Pantel and Lin, 2001)
- Coeficiente *log-likelihood* (Pantel and Lin, 2001)
- Coeficiente *Dice*
- *C-value* (Frantzy and Ananiadou, 1997)

Uma exceção é o trabalho atual de Dias et al (2000) que se utiliza da medida de associação das unidades de um n-grama ($\forall n, n \geq 2$) chamada Esperança Mútua e o procedimento de seleção de candidatos a termos chamado LocalMaxs. Nesse trabalho, os autores realizaram 2 experimentos: um puramente estatístico (aplicando as medidas em um cópuz sem anotação morfossintática) e outro

híbrido utilizando um cópús marcado morfossintaticamente. Esse trabalho será apresentado na seção 3.3.2 e, os resultados dos 2 experimentos serão discutidos na seção 3.4.

Antes de iniciar a descrição dos métodos estatísticos, será definido o modelo probabilístico de linguagem, para que o modelo de linguagem n-grama possa ser definido e posteriormente utilizado nos métodos.

Toda linguagem consiste de uma seqüência de palavras, assim, o modelo probabilístico de linguagem proporciona a probabilidade da próxima palavra, dadas as palavras precedentes. Uma quantidade considerável de textos é utilizada para treinar os modelos de linguagem e determinar os parâmetros de tal modelo na modelagem estatística. Dentre tais modelos, o modelo de linguagem mais usado é o n-grama. Um modelo de linguagem n-grama utiliza a história das $n-1$ palavras imediatamente precedentes para computar a probabilidade de ocorrência P da palavra em questão. Na prática, o valor de n se limita a 2 (modelo bigrama) ou 3 (modelo trigrama) (Zhao, 1999).

3.1.1 Frequência de ocorrência

Muitos sistemas se utilizam de frequência, pois certamente ela é a medida mais simples e popular de se encontrar termos em um cópús. Se duas palavras ocorrerem muitas vezes juntas, existe, então, uma evidência de que elas apresentam uma função especial.

Nota-se que, apenas selecionar, por exemplo, os bigramas que ocorrem mais freqüentemente em um cópús não parece ser um método muito interessante, pois a maioria deles corresponde a pares de palavras funcionais, como artigos e preposições.

A frequência de ocorrência é independente do domínio e não requer recursos externos. Esse método, porém, apresenta mais uma restrição considerando que termos com baixa frequência podem também ser termos válidos.

3.1.2 Informação mútua (para associações binárias)

Informação Mútua é uma medida da quantidade de informação que uma variável contém sobre uma outra, sendo ela a redução da incerteza de uma variável randômica devido ao conhecimento da outra. A definição de informação mútua é:

$$mi(x, y) = \frac{P(x, y)}{P(x) * P(y)}$$

onde x e y são palavras ou termos, $P(x)$ e $P(y)$ são, respectivamente, probabilidades de x e y , que correspondem às frequências das palavras x e y em um cópús de tamanho N , e $P(x,y)$ é a probabilidade que as palavras x e y ocorram juntas adjacientemente.

Esta medida foi usada inicialmente para extração de colocações. Existe uma sobreposição entre as colocações e os termos técnicos: as colocações têm uma composicionalidade limitada, e os termos técnicos aceitam um número limitado de modificadores.

Quando todas as ocorrências de x e y são adjacentes umas às outras, a informação mútua é a maior, deteriorando-se, portanto, em contagens de baixa frequência. A fim de amenizar esse problema, a medida *log-likelihood* é utilizada, em razão de ela se apresentar mais robusta para eventos com baixas frequências. Tal medida é descrita a seguir.

3.1.3 Coeficiente *log-likelihood*

A medida *log-likelihood*, por se apresentar mais robusta para eventos de baixa frequência, é utilizada a fim de amenizar o problema da informação mútua quando esta apresenta contagens de baixa frequência. Considerando que $C(x, y)$ é a frequência de dois termos (x e y) que são adjacentes em algum corpus (onde (*) representa o caractere “coringa”), é possível definir a razão *log-likelihood* de x e y como:

$$\log L(x, y) = ll\left(\frac{k_1}{n_1}, k_1, n_1\right) + ll\left(\frac{k_2}{n_2}, k_2, n_2\right) - ll\left(\frac{k_1+k_2}{n_1+n_2}, k_1, n_1\right) - ll\left(\frac{k_1+k_2}{n_1+n_2}, k_2, n_2\right)$$

onde $k_1 = C(x, y)$, $n_1 = C(x, *)$, $k_2 = C(-x, y)$, $n_2 = C(-x, *)$, e
 $ll(p, k, n) = k \log(p) + (n - k) \log(1 - p)$

Assim como ocorre com a informação mútua, a razão de *log-likelihood* é a maior quando todas as ocorrências de x e y são adjacentes umas às outras. Porém, a razão também é alta para dois termos frequentes que são raramente adjacentes.

3.1.4 Coeficiente *dice*

A medida de associação coeficiente *dice* apresenta uma interpretação similar à informação mútua, visto que ela é definida como:

$$Dice(x, y) = \frac{2 \text{freq}(x, y)}{\text{freq}(x) + \text{freq}(y)}$$

onde, assim como acontece com a informação mútua, x e y são palavras ou termos, $\text{freq}(x, y)$ representa a frequência em que as palavras x e y ocorrem juntas adjacientemente, e $\text{freq}(x)$ e $\text{freq}(y)$ são, respectivamente, frequências de x e y em um corpus de tamanho N.

Essa medida produzirá escores normalizados entre 0 e 1, sendo que valores próximos de 1 indicam uma forte relação (dependência) entre as duas palavras (Tiedemann, 1997).

O coeficiente *dice* depende apenas da frequência do bigrama e das palavras do bigrama. Diferentemente do que ocorre com a informação mútua, essa medida não depende do tamanho da amostra²².

3.1.5 C-value

C-value é uma medida estatística para a extração de termos compostos, e suas características serão agora descritas.

O procedimento para extração tem o início com as cadeias de tamanho máximo. O único parâmetro considerado na possibilidade de um candidato a termo é sua frequência no córpus. Assim, se a é uma cadeia candidata e $f(a)$ sua frequência, $C-value(a) = f(a)$.

Em seguida, é realizada a extração de cadeias menores, sendo que três parâmetros são considerados para cada uma delas:

- 1) a frequência total de ocorrência da cadeia no córpus;
- 2) sua frequência de ocorrência em termos candidatos maiores (já extraídos);
- 3) o número destes termos candidatos maiores.

O primeiro parâmetro se deve ao fato de que termos técnicos possuem a tendência de aparecer com altas frequências, embora uma alta frequência não seja garantia, enquanto que o segundo e o terceiro são instituídos a fim de evitar subcadeias de termos a serem extraídas de forma errônea como termos, em razão de suas ‘altas’ frequências de ocorrência. Por exemplo, “Loja de Material para Construção”, “Material para Construção”, “Loja de Acessórios Femininos”, “Acessórios Femininos”.

Portanto, o requisito para que uma subcadeia de um termo candidato seja também um termo candidato é satisfeito se ela mostra independência ‘suficiente’ com relação aos termos candidatos maiores, dos quais ela é uma subcadeia. Então, enquanto uma alta frequência de uma cadeia candidata em termos candidatos maiores representa um ‘menos’, se o número destes termos maiores é grande, a subcadeia apresenta independência, representando, assim, um ‘mais’. A combinação destes conceitos se resume na seguinte medida:

$$C - value(a) = f(a) - \frac{t(a)}{c(a)}$$

onde

a é a cadeia examinada

$f(a)$ é a frequência total de ocorrência de a no córpus

²² <http://www.d.umn.edu/~tpederse/Group01/bsp.txt>

$t(a)$ é a frequência de ocorrência de a em termos candidatos maiores (já extraídos)

$c(a)$ é o número dos termos candidatos maiores

É importante lembrar que o que foi mostrado acima descreve apenas uma possibilidade, pois a medida extrai uma lista de candidatos a termos cuja avaliação final deve ser feita manualmente.

3.2 Abordagem lingüística

3.2.1 Extrator de termos de Heid et al (1996)

As ferramentas descritas no trabalho de (Heid et al, 1996) dão suporte à extração de candidatos a termo a partir de dicionários *online* e proporcionam contextos e outras informações relativas a termo. O objetivo, ao tornar tais ferramentas disponíveis, é aumentar a eficiência do processo de construção de glossário, fornecendo um suporte melhor aos tradutores na terminologia orientada à análise de textos.

3.2.1.1 Descrição do algoritmo

A exploração de cópulas de textos é realizada através de procedimentos lingüísticos computacionais, que, em geral, consistem de 2 fases importantes:

- pré-processamento lingüístico e anotação automática de texto de dicionários *online*;
- consulta ao cópulas e a extração de informações relevantes para a execução de uma tarefa específica.

Assim, as consultas empreendidas nos textos durante a segunda fase podem apresentar um melhor desempenho, contanto que as informações identificadas e anotadas nos textos no decorrer da primeira fase sejam as mais lingüísticas possíveis.

Pré-análise lingüística

Os passos de análise lingüística e anotação geralmente são constituídos pelas seguintes fases:

- tokenização: identificação de palavras e limites das sentenças;
- análise morfossintática: identificação de categorias gramaticais, características morfossintáticas e distribucionais;
- etiquetagem *part-of-speech*: eliminação da ambigüidade de hipóteses da análise morfossintática e anotação das hipóteses mais prováveis no texto;
- lematização: identificação de candidatos a lema com base nos resultados da análise morfossintática e da etiquetagem *part-of-speech*.

Recuperação de textos para a identificação de termos - as consultas

A extração de material terminológico relevante depende intensamente de consultas em corpús complexos. Os variados tipos de candidatos a termo a serem buscados permitem diferenciação entre as consultas.

Alguns exemplos de candidatos a termos são:

- abreviações: utilizam expressões regulares sobre caracteres;
- termos simples: neste caso, as consultas dependem da pressuposição de que muitos termos nominais contêm prefixos e/ou sufixos;
- colocações: um conjunto de consultas da forma *part-of-speech* designado no projeto DECIDE²³ poderia ser aplicado para colocações verbais;
- grupos nominais: a extração de grupos nominais e preposicionais relevantes terminologicamente também se baseia nas formas *part-of-speech*.

3.2.1.2 Aplicação

Setup

O corpús utilizado pelas fases descritas na subseção anterior é um dicionário *online* de Alemão e tradução Alemão-Francês e Alemão-Inglês no campo de engenharia automotiva, onde são realizadas as consultas, extraindo-se informações relevantes para a execução de uma tarefa particular. As ferramentas usadas para a extração de termos abrangem essas 3 línguas, com ênfase no Alemão.

Avaliação

Os melhores resultados são alcançados pelas consultas na identificação de substantivos, visto que eles formam o maior conjunto de termos candidatos. No entanto, muitos ruídos foram encontrados nos resultados das consultas por adjetivos e verbos. De acordo com (Heid et al, 1996), esse sistema é o único que extrai tanto adjetivos como verbos, dando suporte, então, ao trabalho de fraseologia de língua especializada. Porém, consultas por adjetivos e verbos resultam em maior quantidade de ruídos. Essa quantidade se explica pelo fato de que poucos afixos em adjetivos são típicos de uma linguagem específica.

Também são encontradas quantidades relevantes de ruído na extração de grupos nominais e preposicionais, em razão de problemas típicos de gramáticas regulares. Isso ocorre porque formas

²³ DECIDE (*Designing and Evaluating Extraction Tools for Collocations in Dictionaries and Corpora*), um projeto parcialmente fundado pela Comissão Européia, DG XIII E, Luxemburgo, sob seu Plano de Ação Multilingual (projeto no MLAP-93/19), que decorreu entre 02/94 e 01/96, foi dedicado à língua geral. Ele proporcionou quantidades suficientes de materiais do tipo texto disponíveis - os procedimentos de extração DECIDE podem ser aplicados com sucesso à língua especializada. Maiores detalhes sobre as ferramentas DECIDE, ver (Gérardy, 1996).

de *part-of-speech* não limitam os contextos candidatos suficientemente e acabam extraíndo seqüências de um material anotado categoricamente e morfossintaticamente, sem controle se a seqüência extraída tem a origem ou não em uma construção frasal.

A extração de candidatos à colocação também produz ruído, mas não silêncio.

A comparação desses resultados com aqueles obtidos no mesmo cópús, ao se utilizar a medida estatística de frequência relativa em cópús de língua especializada e geral AHMAD ((Ahmad et al, 1992) apud (Heid et al, 1996)), mostra que os resultados produzidos pela consulta no cópús lingüístico estão todos contidos na saída dos procedimentos estatísticos. Entretanto, a consulta lingüística é muito mais seletiva, pois os métodos estatísticos produzem muito mais ruído que os lingüísticos.

3.2.1.3 Extensão

O problema da alta porcentagem de ruídos encontrados quando são realizadas consultas para adjetivos pode ser amenizado se morfemas e componentes do domínio específico forem usados como um filtro adicional.

No caso da extração de grupos nominais complexos e preposicionais, que também produz quantidade considerável de ruído, uma solução seria um *parsing* (análise sintática) em nível de frase, que ajudaria reduzir essa quantidade.

Para melhorar a precisão dos resultados, um Chunk-Parsing para o Inglês será utilizado, identificando construções de frases não recursivas, melhorando, assim, a precisão da recuperação de termos compostos.

É importante ressaltar que o sistema pode ser ajustado facilmente a um outro domínio, bastando apenas alterar componentes de formação de palavras de um domínio específico.

3.2.1.4 Recursos e ferramentas utilizados

O sistema de ferramentas que realiza a tarefa de extração de termos é constituído por:

- Um processador de consulta a cópús geral, (CQP), que opera sobre expressões de consulta complexas, sendo que a consulta ao cópús pode incluir qualquer número e combinação de tipos de informação anotada;
- Um macroprocessador para a linguagem de consulta ao CQP. Os procedimentos para a extração de termos dependem de listas de afixos, bem como da checagem de contextos típicos de termos candidatos. Assim, essa atividade pode ser auxiliada por meio de uma dada consulta que deve ser executada com um número possivelmente grande de palavras diferentes e todos os outros parâmetros de consulta permanecendo os mesmos. O

macroprocessador executa o mesmo tipo de consulta nos elementos de listas candidatas, além de incluir uma biblioteca macro, permitindo que expressões de consulta que podem ser usadas em diferentes contextos possam ser nomeadas, armazenadas e reutilizadas;

- XKWIC²⁴, uma interface de usuário gráfica baseada no XWindows/ MOTIF, para linguagem de consulta ao corpus CQP, que proporciona palavras-chave em concordâncias no contexto (isto é, um concordanceador) e permite a classificação automática do material extraído conforme parâmetros do contexto, definidos pelo usuário, além da possibilidade de compilação de listas de frequência absoluta e relativa de itens de busca.

Assim, diferentes tipos de informação podem ser recuperados da saída da ferramenta, entre eles: listas de palavras, listas de itens compostos, sentenças de uma dada língua, amostras paralelas e dados de frequência.

3.2.2 Sistema DEFINDER de Klavans and Muresan (2000; 2001a; 2001b)

DEFINDER (Definition Finder) é um sistema baseado em regras que explora artigos orientados ao consumidor a fim de extrair definições e os termos que eles definem. É importante deixar claro que essa pesquisa pertence ao Projeto de Biblioteca Digital na Universidade de Colúmbia, que apresenta a denominação PERSIVAL²⁵ (PErsonalized Retrieval and Summarization of Image, Video And Language Resource). Uma meta desse projeto é apresentar definições precisas e legíveis de termos técnicos, que podem ser encontrados em artigos que apresentam complexidade intermediária, através do uso de uma linguagem que seja compreensível aos pacientes (usuários).

3.2.2.1 Descrição do algoritmo

A abordagem presente no sistema DEFINDER consiste na combinação de técnicas superficiais de processamento de língua natural com análise gramatical, com o intuito de explorar textos eficientemente.

Assim, com a análise de um conjunto de artigos médicos orientados ao consumidor, foi possível identificar expressões lingüísticas típicas e indicadores estruturais que introduzem definições e os termos definidos. O sistema DEFINDER é baseado em dois módulos funcionais principais: 1) um módulo de processamento de texto superficial que realiza análise de padrões com o uso de uma gramática de estado finito guiada por expressões lingüísticas, tais como, “é chamado”,

²⁴ Os autores utilizaram o concordanceador de (Christ, 1994b).

²⁵ <http://persival.cs.columbia.edu/>

“é o termo utilizado para descrever”, “é definido como”, entre outras, e um conjunto limitado de marcadores de texto, representado, por exemplo, por indicadores estruturais, tais como, ((), --), e 2) um módulo de análise gramatical que utiliza uma rica gramática lexicalista orientada à dependência - Slot Grammar do Inglês (McCord, 1991 apud Klavans and Muresan, 2001) para analisar fenômenos lingüísticos mais complexos, como aposição e anáfora.

O módulo de análise de padrões se baseia em um etiquetador morfológico (*part-of-speech*) com uma gramática de estado finito para identificar termos médicos e extrair definições. Na identificação de frases nominais simples (núcleo do substantivo + pré-modificadores) foram utilizados o etiquetador de Brill (1995) e o analisador sintático parcial (*chunker*) de sintagmas nominais de Ramshaw and Marcus (1995). A fim de eliminar alguns padrões errôneos introduzidos pelos marcadores, tais como explicação e enumeração, um módulo de filtragem foi, então, acrescentado.

Já o módulo de análise gramatical é baseado no formalismo Slot Grammar do Inglês (ESG), que permite a identificação de definições introduzidas por fenômenos lingüísticos mais complexos (anáforas, apostos, por exemplo) que não são encontradas de forma fácil pelo processamento superficial. Esta vantagem que o módulo de análise gramatical apresenta se deve ao fato de a ESG possuir uma representação rica.

3.2.2.2 Aplicação

Setup

O MEDLINEplus²⁶ foi, a princípio, considerado, e, posteriormente o Cardiovascular Institute of the South²⁷ foi selecionado, em razão desse último ser bem editado e estruturado, sendo, então, apropriado para técnicas de extração de padrões baseadas em regras. Assim, um cópuz foi, então, construído e dividido randomicamente, sendo que 75% foi utilizado para desenvolvimento e 25% para a realização de testes. Aproximadamente 60% das definições pertencem ao módulo de análise padrão, enquanto que as 40% restantes se enquadram no módulo de análise gramatical.

O conjunto manualmente marcado consistindo de 53 definições (*gold standard*) foi determinado pelo conjunto de definições marcadas por pelo menos 3 das 4 pessoas.

Dois dicionários *online* e um glossário foram utilizados para realizar comparações com o DEFINDER. Eles são: Sistema de Linguagem Médica Unificada (UMLS), Dicionário Médico *Online* (OMD)²⁸ e Glossário de Termos Populares e Médicos Técnicos (GPTMT)²⁹.

²⁶ <http://medlineplus.gov/>

²⁷ <http://www.cardio.com/articles.html>

²⁸ <http://www.graylab.ac.uk/omd/>

²⁹ <http://allserv.rug.ac.be/%7Ervdstich/eugloss/welcome.html>

Avaliação

Nesta seção será abordada uma avaliação quantitativa e qualitativa de DEFINDER. Duas avaliações quantitativas realizadas mostram que DEFINDER alcançou 87% e 75%, respectivamente, para precisão e revocação junto ao desempenho humano, o que revelou a incompletude de recursos existentes (por exemplo, dicionários *online*) e a habilidade de DEFINDER em suprir as carências destes recursos.

A avaliação qualitativa mostra que as definições extraídas pelo sistema DEFINDER (Klavans and Muresan, 2001) são categorizadas como melhores em termos dos critérios utilidade e legibilidade do que definições de dicionários especializados *online*, ou seja, definições encontradas nos textos se apresentam mais legíveis e úteis do que aquelas presentes em *thesauri* e glossários existentes. Dessa forma, a saída do DEFINDER pode ser usada para aumentar tais dicionários.

Métodos estatísticos

Três métodos foram utilizados para avaliar o sistema DEFINDER: 1) o desempenho em termos de precisão e revocação contra um padrão manualmente marcado (*gold standard*), 2) a qualidade de definições extraídas em termos dos critérios legibilidade, utilidade e completude, 3) a cobertura da saída do DEFINDER versus dicionários *online* existentes. Assim, para os dois primeiros métodos foi realizada a avaliação com usuários, utilizando pessoas não especialistas, enquanto que, para o terceiro, um conjunto de termos definidos extraído pelo sistema em questão foi escolhido e comparado com os dicionários UMLS, OMD e GPTMT.

O primeiro método de avaliação está relacionado à comparação da saída do DEFINDER com o padrão manualmente marcado (*gold standard*). Quatro pessoas foram escolhidas para esse experimento, mas elas não auxiliaram no desenvolvimento do sistema. Para cada pessoa foi, então, atribuído um conjunto de nove artigos escolhidos de diferentes gêneros de textos que são direcionados a pessoas leigas, sendo que suas definições e seus termos associados foram marcados manualmente. O padrão manualmente marcado consistiu de 53 definições identificadas por pelo menos três das quatro pessoas. A partir da comparação com esse padrão, o desempenho do DEFINDER pôde ser medido em termos de precisão e revocação.

O segundo método diz respeito à qualidade de definições, sendo que a qualidade da saída do DEFINDER é avaliada através da comparação com os dicionários UMLS e OMD. Oito pessoas não especialistas receberam uma lista de 15 termos médicos escolhidos aleatoriamente junto com suas definições de cada um dos três recursos (UMLS, OMD e DEFINDER). A cada definição foi atribuído um valor especificando a qualidade para utilidade (U), legibilidade (L) e completude (C) de uma escala de 1 a 7, sendo que 1 indica “pior” e 7 indica “excelente”. Utilidade significa que a definição deveria auxiliar o usuário no entendimento de um termo técnico no contexto de um artigo

técnico. Já legibilidade indica facilidade de leitura e entendimento, enquanto que completude significa que a definição deve conter informação completa sobre o termo.

Para a qualidade foram realizados dois estudos, sendo que no primeiro foi medida a Taxa de Qualidade Média (AQR) para cada uma das três fontes de definição nos três critérios. O DEFINDER apresenta um desempenho melhor do que o UMLS e o OMD em termos de utilidade e legibilidade, e tais dicionários são melhores em termos de completude.

Ao computar a AQR, surge uma dúvida se os altos escores dados por uma pessoa podem compensar os baixos valores dados por outras pessoas, introduzindo ruído. A fim de validar essa questão, um segundo estudo foi realizado, avaliando, assim, a classificação relativa das três fontes de definição, sendo que 1 é considerado “melhor” e 3 é tido como “pior”. Foi utilizado o coeficiente de correlação de Kendall (W) para medir a confiabilidade entre avaliadores sobre cada termo. Para os termos com concordância significativa foi computado o nível de correlação entre eles. Se W foi significativo, foi comparada a classificação média global das 3 fontes.

O método para determinar a cobertura baseia-se na comparação da saída do DEFINDER com dicionários *online* existentes. Dessa forma, os três dicionários *online* existentes selecionados para avaliar a cobertura de forma quantitativa foram UMLS, OMD e GPTMT. Para a realização desse experimento foi escolhido um conjunto de testes de 93 termos com suas definições associadas, extraídos pelo sistema DEFINDER. Assim, três casos foram encontrados: 1) o termo é listado em um dos dicionários *online* e é definido em tal dicionário (definido); 2) o termo é listado em um dos dicionários *online* mas não apresenta uma definição associada (indefinido); 3) o termo não é listado em nenhum dos dicionários *online* (ausente).

Resultados

Para o primeiro experimento, DEFINDER identificou 40 das 53 definições, obtendo, assim, 86.95% e 75.47% de precisão e revocação, respectivamente. Além das definições corretas (40), DEFINDER extraiu 6 falsos positivos, decrescendo, assim, a precisão.

Resultados da comparação dos valores da taxa de qualidade média mostram que as definições extraídas pelo DEFINDER são julgadas melhores do que as definições dos outros dois dicionários em termos de utilidade e legibilidade. Já em termos de completude, tanto UMLS quanto OMD apresentam resultados melhores.

No segundo estudo para medir a qualidade, a categorização relativa das três fontes de definição foi analisada, com a utilização do coeficiente de correlação de Kendall para validar os resultados estatisticamente.

Em termos de utilidade, usuários concordaram em 13 dos 15 casos, com valores significativos de W variando de 0.47 a 0.92 (Siegal and Castellan, 1988 apud Klavans and Muresan,

2001). Para estes 13 termos foi medido o nível de correlação e, a média de categorias para as três fontes de definição foi, então, computada. Considerando $W=0.45$ significativo, é possível ter como o valor “verdadeiro” de categorização relativa à ordem proporcionada pelas categorias médias.

Já em termos de legibilidade, a aceitação entre os juízes³⁰ foi significativa em 14 dos 15 casos, W variando de 0.56 a 1.00 (os valores se apresentaram acima de 0.85 para 50% dos termos). A correlação encontrada entre as categorizações nestes 14 termos foi significativa ($W=0.54$), tornando a ordenação relevante.

Considerando a análise de aceitação para completude, em apenas 11 dos 15 casos foi obtida uma aceitação significativa entre as pessoas (valores de W variando de 0.38 a 0.92). O teste de significância para correlação entre as categorizações desses 11 termos falhou ao proporcionar um valor representativo estatisticamente para W ($W=0.26$). Por essa razão, nenhuma inferência pode ser feita considerando a ordem de categorizações entre as três fontes de definições.

Para o terceiro experimento (análise da cobertura), os resultados da Tabela 3.1 mostram que dicionários médicos *online* são incompletos comparados à saída do DEFINDER.

Tabela 3.1: Cobertura de Dicionários *Online*

Termo	UMLS	OMD	GPTMT
Definido	60%(56)	76%(71)	21.5%(20)
Indefinido	24%(22)	-	-
Ausente	16%(15)	24%(22)	78.5%(73)

Observe que a coluna 3 mostra que em OMD somente 71 definições das 93 definições do teste são encontradas, proporcionando 76% de completude, enquanto que o GPTMT, um glossário dedicado especificamente a usuários leigos, ainda está distante de ser completo, sendo que somente 20 dos 93 termos estavam presentes, apresentando a cobertura de 21.5%. Todos esses resultados mostram que DEFINDER identifica muitos termos e definições que estão faltando em recursos existentes, auxiliando no aprimoramento dos dicionários *online*, empregando-se definições úteis e legíveis.

Discussão

Através das avaliações realizadas no decorrer dessa seção, foi possível concluir que a saída do sistema DEFINDER pode ser utilizada, pelo menos, de duas maneiras: 1) buscando aumentar dicionários *online* e 2) tornar a terminologia compreensível para usuários não especialistas, proporcionando definições legíveis e úteis.

Uma análise cuidadosa do desempenho humano na identificação de definições juntamente com os seus termos associados de texto mostra que essa é uma tarefa muito difícil. Além das 53

³⁰ Não especialistas para a realização de uma avaliação do sistema centrada no usuário.

definições constituindo o padrão manualmente marcado, 8 definições foram identificadas por uma pessoa apenas e 10 definições por duas pessoas. O decréscimo na precisão se deve ao fato de que o sistema identificou 6 falsos positivos. No entanto, quatro dessas seis definições foram marcadas por uma pessoa. Já o decréscimo da revocação se justifica em razão de que várias definições identificadas por julgamentos humanos contêm fenômenos lingüísticos complexos (anáfora ou definições paralelas), que não são manipuladas pelo sistema atualmente.

A avaliação qualitativa do DEFINDER, considerando critérios de utilidade, legibilidade e completude baseados no usuário, mostra que DEFINDER reconhece definições de alta qualidade para usuários leigos.

Os resultados da cobertura de dicionários mostraram que no UMLS, 24% dos termos apresentados pertenciam ao vocabulário axiomático, que é altamente técnico, sendo de uso limitado para pessoas leigas e 15 termos dos 93 estavam ausentes no UMLS. Os termos ausentes foram analisados e concluiu-se que, em alguns casos, modificadores exercem uma função importante na decisão de quais termos são “reais”, por exemplo: “cardiac defibrillator” foi o termo extraído definido pelo sistema DEFINDER, enquanto que em UMLS foi encontrado apenas o termo “defibrillator”.

3.2.2.3 Extensão

Futuramente, os autores pretendem endereçar a questão de unir múltiplas definições de fontes diferentes, desenvolver um método para avaliar a exatidão e completude e integrar DEFINDER no sistema de informação médica PERSIVAL.

3.2.2.4 Recursos e ferramentas utilizados

O dicionário *online* MEDLINEplus foi utilizado no processo de extração de definições para usuários leigos. Na identificação de sintagmas nominais simples foram utilizados o etiquetador de Brill (1995) e o analisador sintático parcial de sintagmas nominais de Ramshaw and Marcus (1995). Já o módulo de análise gramatical baseou-se na Slot Grammar do Inglês (ESG) de McCord (1991). Os recursos utilizados na avaliação do sistema DEFINDER foram os dicionários *online*, UMLS (Sistema de Linguagem Médica Unificada) e OMD (Dicionário Médico *Online*), e o glossário GPTMT (Glossário de Termos Populares e Médicos Técnicos).

3.3 Abordagem híbrida

3.3.1 Extrator de termos de Frantzy and Ananiadou (1997)

A abordagem aqui apresentada dedica-se à extração de termos compostos de cópulas de língua especial.

Informação do contexto para termos

Informação do Contexto exerce um papel importante na extração de termos, em razão de unidades fraseológicas/expressões padrões poderem ser modificadas livremente, enquanto que termos compostos não apresentam tal característica (Sager, 1978) apud (Frantzy and Ananiadou, 1997). Um exemplo de unidades fraseológicas/expressões padrões que podem ser modificadas livremente para o domínio da bolsa de valores, retirado de Smadja (1991), é mostrado abaixo e segue o padrão "Advancers DEGREE (declining/losing) issues (by/with) a margin of NUMBER1 to NUMBER2", em que DEGREE é a intensidade da proporção entre "advancers" e "decliners" e NUMBER1 e NUMBER2 são os valores entre alta/baixa.

"Advancers outnumbered declining issues by a margin of 3 to 1"

"Advancers had a slim lead over losing issues with a margin of 2 to 1"

Dessa forma, uma informação que poderia ser usada no procedimento para atribuição de um valor a termos candidatos poderia vir acompanhada de seus modificadores. Isso poderia ser estendido a verbos que pertencem ao contexto de termos candidatos, além de adjetivos ou modificadores de substantivo. Por exemplo, a forma "shows" do verbo "to show" em domínios médicos, é, na maioria das vezes, seguida por um termo, como ocorre em *shows a basal cell carcinoma*. Existem casos em que o contexto que aparece com termos pode até ser independente do domínio, como a forma "called" do verbo "to call", ou a forma "known" do verbo "to know", que frequentemente está envolvida em definições em diversas áreas, por exemplo *is known as the singular existential quantifier, is called the Cartesian product*.

Visto que o contexto apresenta informação sobre termos, ele deveria estar envolvido no procedimento para a extração de termos.

A seguir é apresentada uma rápida descrição do procedimento de extração de termos:

1. Produção de uma lista de termos candidatos com a utilização da abordagem *C-value*;
2. Seleção de algumas das 'primeiras' cadeias obtidas da lista produzida. Estas 'primeiras' cadeias possuem a maior densidade em termos da lista total produzida;
3. Extração do contexto para os 'primeiros' termos candidatos do cópula. Em tal contexto são considerados os verbos, adjetivos e substantivos ao redor do termo candidato;

4. Atribuição de um peso a cada um dos verbos, adjetivos e substantivos, de acordo com algumas características estatísticas, que serão posteriormente discutidas.

A maneira como os pesos são atribuídos ao contexto é totalmente automática.

O filtro lingüístico

Para cópús etiquetado morfossintaticamente, um filtro lingüístico vai permitir que apenas cadeias com etiquetas específicas sejam consideradas. O filtro lingüístico pode ser ‘fechado’ ou ‘aberto’ e gerar resultados positivos ou negativos na precisão e revocação. Assim, um filtro ‘fechado’, que não permite ‘muitas’ seqüências de etiquetas, melhora a precisão mas apresenta efeito negativo sobre a revocação, enquanto que um filtro ‘aberto’, que permite mais seqüências de etiquetas, por exemplo, preposições, adjetivos e substantivos, apresenta o resultado oposto.

O filtro lingüístico escolhido encontra-se no centro, e permite cadeias consistindo de adjetivos e substantivos:

(Substantivo|Adjetivo)⁺Substantivo

No entanto, este filtro específico não deveria ser usado em todos os casos, e, sua escolha poderia também ser mais ‘fechada’ ou ‘aberta’, dependendo da aplicação.

3.3.1.1 Descrição do Algoritmo

O algoritmo de extração de termos consiste dos seguintes passos:

1. O *cópus* é etiquetado com o etiquetador morfossintático de Brill (1995). Os n-gramas do *cópus* etiquetado que obedecem à expressão $(Substantivo|Adjetivo)^+Substantivo$ são extraídos.
2. O *C-value* é calculado para esses n-gramas, resultando em uma lista de termos potenciais, que são classificados pelo *C-value*. O parâmetro do comprimento do n-grama é então incorporado para esse uso de *C-value*. No caso presente, o peso comprimento foi atenuado, obtendo-se, assim, o *C-value'*:

$$C\text{-value}'(a) = \begin{cases} \log_2 |a| \cdot f(a) & \text{se } |a| = \max, \\ \log_2 |a| \cdot (f(a) - \frac{1}{c(a)} \sum_{i=1}^{c(a)} f(b_i)) & \text{caso contrário} \end{cases}$$

onde

a é o n-grama examinado;

$|a|$ o comprimento, em termos de número de palavras, de a ;

$f(a)$ a frequência de a no *cópus*;

b_i os termos candidatos extraídos que contêm a ;

$c(a)$ o número desses termos candidatos (b_i)

A incorporação do contexto terá lugar neste ponto.

3. Visto que *C-value* é uma medida para extrair termos, o topo da lista previamente construída apresenta a maior densidade em termos entre qualquer outra parte da lista. Esse topo da lista, ou senão os 'primeiros' desses termos candidatos classificados darão os pesos ao contexto. As cadeias candidatas classificadas no topo são, então, tomadas, extraíndo seus contextos do *cópus* inicial (isto é, suas concordâncias) que são: verbos, adjetivos e substantivos ao redor do termo potencial. Assim, três parâmetros são considerados para cada um desses verbos, adjetivos e substantivos:
 - a) sua frequência total no *cópus*;
 - b) sua frequência como palavra no contexto (daqueles primeiros n-gramas);
 - c) o número daqueles n-gramas com quem ele aparece.

Essas características são combinadas da seguinte forma, atribuindo um peso à palavra do contexto:

$$Peso(w) = 0.5 * \left(\frac{t(w)}{n} + \frac{ff(w)}{f(w)} \right)$$

onde

w é o substantivo/verbo/adjetivo a quem será atribuído um peso;

n é o número total de termos candidatos considerados;

$t(w)$ o número de termos candidatos com quem a palavra w parece;

$ft(w)$ frequência total de w 's aparecendo com termos candidatos;

$f(w)$ frequência total de w 's no cópús.

A fim de melhorar os resultados, uma variação envolvendo interação humana é considerada, da seguinte forma: os termos candidatos que estão envolvidos para a extração de contexto são, primeiro, avaliados, e apenas os 'termos reais' procederão para a extração do contexto e a atribuição de peso a eles.

Neste ponto, uma lista de palavras do contexto, junto com seus pesos, foi criada.

4. A lista *C-value*', criada anteriormente, será agora re-classificada de acordo com os pesos obtidos no passo 3. Para cada um daqueles n-gramas, seus contextos (verbos, adjetivos e substantivos ao redor dele) são extraídos do cópús. As palavras do contexto que foram encontradas no passo 3 já ganharam seu peso. As que não foram, o peso igual a 0 lhes é atribuído.

Cada um desses n-gramas está agora pronto para que lhe seja atribuído um peso do contexto, que é a soma dos pesos de suas palavras do contexto:

$$P(a) = \sum_{b \in C_a} \text{Peso}(b) + 1$$

onde

a é o n-grama examinado;

C_a o contexto de a ;

$Peso(b)$ o peso anteriormente calculado para a palavra b .

Os n-gramas agora serão re-classificados de acordo com:

$$NC - value(a) = \frac{1}{\log(N)} * C - value'(a) * P(a)$$

onde

a é o n-grama examinado;

$C-value'(a)$, o $C-value'(a)$ calculado anteriormente;

$P(a)$ é a soma previamente calculada dos pesos do contexto para a ;

N é o tamanho do *cópus* em termos de número de palavras.

3.3.1.2 Aplicação

Setup

Um *cópus* utilizado para a realização dos testes apresenta textos em Inglês do domínio médico.

Avaliação

Nesse trabalho, a informação do contexto foi usada para a extração de palavras sinônimas, enquanto que para a extração de termos, a informação usada foi a 'interna', ou seja, a lingüística e estatística, que caracterizou o termo candidato e não seu ambiente. A investigação, para a implementação atual, não foi completa para o tipo de contexto a ser considerado. Foi assumido que os verbos, adjetivos e substantivos ao redor do termo candidato possuem a mesma quantidade de informação.

3.3.1.3 Extensão

São três trabalhos futuros apresentados:

- A investigação do contexto usado para a avaliação das cadeias candidatas e a quantidade de informação que possuem;
- A investigação da atribuição de pesos aos parâmetros usados para as medidas;
- A comparação desse método com outras abordagens automáticas de reconhecimento de termo, aplicando-as aos mesmos dados, em vários domínios.

3.3.1.4 Recursos e ferramentas utilizados

O etiquetador morfossintático de Brill (1995) é um recurso utilizado no primeiro passo do algoritmo para etiquetar o córpus.

3.3.2. Extrator de termos de Dias et al (2000)

O objetivo do trabalho de Dias et al (2000) é mostrar os resultados obtidos com a aplicação de uma metodologia criada pelos autores para extrair termos compostos de 2 formas:

- 1) considerando apenas estatísticas léxicas;
- 2) combinando estatísticas léxicas com informação lingüística adquirida internamente de um córpus previamente etiquetado.

Ao final da descrição do trabalho são apresentadas as vantagens e desvantagens das 2 formas acima. Os autores advogam para a não-utilização de uma metodologia única para se extrair termos compostos e sim para o uso de métodos diferentes dependendo do tamanho do n-grama.

A metodologia aqui utilizada baseia-se em dois princípios principais: princípio da rigidez e princípio da integridade do córpus. O primeiro princípio mostra que quanto menos uma seqüência de palavras aceita transformações morfológicas e sintáticas (isto é, quanto mais fixa ela é), maior é a probabilidade de ela ser um termo composto. Assim, a informação que aparece em córpus etiquetado morfossintaticamente deveria ser suficiente para extrair unidades textuais sem a necessidade de se aplicar heurísticas dependentes do domínio ou dependentes da língua. Já o segundo princípio implica que o córpus de entrada não deveria ser modificado, ou seja, a lematização e a poda de textos através de listas de *stop words* (artigos, pronomes, preposições e conjunções), por exemplo, deveriam ser evitados em razão de introduzirem restrições que não estão contidas no texto original de entrada. Dessa forma, menos restrições externas são introduzidas no processo de extração, fazendo com que o texto de entrada não sofra modificações.

3.3.2.1 Descrição do algoritmo

Primeiro passo: consiste na transformação do texto de entrada em um conjunto de n-gramas³¹. Por exemplo, para o caso a seguir em que a palavra pivô é *Lei*, dada a sentença de entrada (1), temos que *Lei de Imprensa* é um termo composto específico.

(1) *O artigo 35 da Lei de Imprensa prevê esse procedimento em caso de burla agravada.*

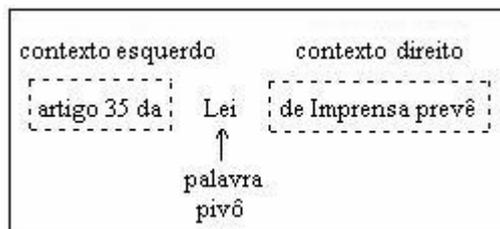


Figura 3.1 - Contexto do *Span* para a palavra pivô *Lei*.

Um n-grama é um vetor de n palavras em que cada palavra é indexada pela distância de sinal que a separa da palavra pivô. Um n-grama pode, então, ser contíguo ou não contíguo caso as palavras envolvidas no n-grama representem ou não uma seqüência contínua de palavras no córpus. Considerando a sentença (1) como sendo o texto de entrada atual e “Lei” a palavra pivô, um 3-grama de palavra contíguo e um não contíguo são respectivamente mostrados nas duas últimas linhas da Tabela 3.2.

³¹ Trabalhos aplicados de lexicografia evidenciam que a maioria das relações lexicais associa palavras separadas por no máximo cinco outras palavras e que termos compostos compartilham tal propriedade. Conseqüentemente, um termo composto pode ser definido em termos de estrutura como um n-grama de palavra específico calculado no contexto imediato de três palavras à esquerda e três palavras à direita de uma palavra pivô.

Tabela 3.2: Amostra de 3-gramas calculados a partir da palavra pivô *Lei*

w_1	Posição ₁₂	w_2	Posição ₁₃	w_3
Lei	+1	de	+2	Imprensa
Lei	-3	artigo	+3	prevê

De forma geral, um n-grama é um vetor com n unidades textuais onde cada unidade textual é indexada pela distância com sinal que a separa da unidade textual pivô associada. A unidade textual pivô é, por convenção, sempre o primeiro elemento do vetor e sua distância com sinal é zero. Um n-grama é representado por um vetor $[p_{11} u_1 p_{12} u_2 p_{13} u_3 \dots p_{1i} u_i \dots p_{1n} u_n]$ onde p_{11} é equivalente a zero e p_{1i} (para $i=2$ até n) denota a distância com sinal que separa a unidade textual u_i da unidade textual pivô u_1 .

Cada n-grama de palavras pode ser ligado ao seu n-grama de etiquetas correspondente. Um n-grama de etiquetas é um vetor de etiquetas morfossintáticas onde cada etiqueta é indexada pela distância com sinal que a separa da etiqueta da palavra pivô associada. Então, se a sentença (2) etiquetada, que é resultado do processo de etiquetagem sobre a sentença (1), for considerada como a entrada atual e $[N]$ a etiqueta associada à palavra pivô *Lei*, os 3-gramas de etiquetas correspondentes aos 3-gramas de palavras apresentados na Tabela 3.2 são respectivamente mostrados nas últimas duas linhas da Tabela 3.3.

(2) O[ART] artigo[N] 35[NUM] da[PREP] Lei[N] de[PREP] Imprensa[N] prevê[V] esse[ART] procedimento[N] em[PREP] caso[N] de[PREP] burla[N] agravada[ADJ].

Tabela 3.3: 3-gramas de etiquetas correspondentes aos 3-gramas de palavras da Tabela 3.2

t_1	Posição ₁₂	t_2	Posição ₁₃	t_3
[N]	+1	[PREP]	+2	[N]
[N]	-3	[N]	+3	[V]

O estágio de preparação de dados termina, então, com a definição de sete tabelas utilizadas no armazenamento do conjunto de todos os n-gramas obtidos por processar seqüencialmente o cópulus de entrada, cada palavra sendo sucessivamente o pivô.

Segundo e terceiro passos: a frequência e a medida de associação de cada n-grama único são calculadas respectivamente.

Quarto passo: o algoritmo LocalMaxs é aplicado com o intuito de eleger termos candidatos compostos do conjunto de todos os n-gramas valorados. A seguir será apresentada a medida de Esperança Mútua, que será aplicada a cada n-grama de palavras a fim avaliar seu grau de coesão.

Esperança normalizada e esperança mútua

A maioria dos modelos matemáticos que visam avaliar o grau de coesão entre unidades textuais avalia somente o grau de coesão entre duas unidades textuais (bigramas) e não generalizam para o caso de n unidades textuais individuais. Para o caso de unidades textuais com mais de 2 palavras, técnicas de atração são utilizadas para adquirir associações. Entretanto, de acordo com os autores, essas técnicas apresentam limitações pois dependem de bigramas adequados para se iniciar o processo iterativo.

Buscando resolver tal problema uma nova medida de associação, a Esperança Mútua (ME), é introduzida pelos autores. A Esperança Mútua avalia o grau de rigidez que liga todas as unidades textuais contidas em um n-grama ($\forall n, n \geq 2$), baseando-se no conceito de Esperança Normalizada (NE).

Background estatístico

Por definição, uma associação textual é caracterizada por algum tipo de atração existente entre seus componentes. Uma tabela de contingência n-ária é construída para cada n-grama com o objetivo de investigar esse relacionamento particular entre palavras ou etiquetas morfossintáticas. A tabela de contingência fornece uma boa exposição dos dados a serem analisados. Para começar construir tal tabela é necessário definir um Espaço de Probabilidade. O Espaço de Probabilidade $(\Omega, A, P[.])$ é introduzido na Tabela 3.4, onde Ω é o espaço Domínio, A o espaço Evento e $P[.]$ a função Probabilidade.

Tabela 3.4: O Espaço de Probabilidade $(\Omega, A, P[.])$

A	O evento espaço A mapeia uma variável randômica discreta X_{ip} para cada unidade textual u_i , sendo que essa variável recebe o valor “1” se a unidade textual u_i aparece em um n-grama na posição ³² p e “0” caso contrário.
Ω	O espaço Domínio Ω é a coleção de todos os resultados possíveis de um experimento conceitual sobre o espaço de instância, sendo definido como $\Omega = \{0,1\}$.
$P[.]$	Uma boa aproximação para a função de probabilidade $P[.]$ é definida como o número de sucessos para um resultado particular dividido pelo número de instâncias.

³² A posição p é a posição da unidade textual u_i na relação com a primeira unidade textual do n-grama.

O espaço das instâncias sobre o qual o Espaço de Probabilidade pode ser aplicado é o conjunto de todos os n-gramas construídos do texto de entrada.

A fim de proporcionar uma melhor compreensão do método, somente será detalhado o caso de bigramas envolvendo a definição de uma tabela de contingência com duas dimensões para cada bigrama. Um bigrama pode ser definido como uma quádrupla $[p_{11} u_1 p_{12} u_2]$ onde u_1 e u_2 são duas unidades textuais e p_{12} denota a distância com sinal que separa ambas palavras e p_{11} é equivalente a zero. Assim como foi definido na Tabela 3.4, u_1 e u_2 são mapeadas, respectivamente, para duas variáveis randômicas discretas X_{1p} e X_{2k} cuja coesão tem de ser testada com o intuito de medir a atração entre elas. A tabela de contingência é definida como na Tabela 3.5 para cada quádrupla $[p_{11} u_1 p_{12} u_2]$ do espaço das instâncias.

Tabela 3.5: Uma tabela de contingência para bigramas

	X_{2k}	$\neg X_{2k}$	Total da Linha
X_{1p}	$f(p_{11}, u_1, p_{12}, u_2)$	$f(p_{11}, u_1, p_{12}, \neg u_2)$	$f(u_1)$
$\neg X_{1p}$	$f(p_{11}, \neg u_1, p_{12}, u_2)$	$f(p_{11}, \neg u_1, p_{12}, \neg u_2)$	$f(\neg u_1)$
Total da Coluna	$f(u_2)$	$f(\neg u_2)$	N

onde

N é o número de palavras presentes no texto de entrada;

$f(p_{11}, u_1, p_{12}, u_2)$ é a frequência de u_1, u_2 ocorrerem juntas na distância com sinal p_{12} ;

$f(p_{11}, u_1, p_{12}, \neg u_2)$ é a frequência de u_1 ocorrer com palavras diferentes de u_2 na distância com sinal p_{12} ;

$f(p_{11}, \neg u_1, p_{12}, u_2)$ é a frequência de u_2 ocorrer com palavras diferentes de u_1 na distância com sinal p_{12} ³³;

$f(p_{11}, \neg u_1, p_{12}, \neg u_2)$ é a frequência de u_1, u_2 nunca ocorrerem na distância com sinal p_{12} , $f(u_1)$ e $f(u_2)$ são as respectivas frequências marginais de u_1 e u_2 ;

$f(\neg u_1)$ e $f(\neg u_2)$ são respectivamente iguais a $N - f(u_1)$ e $N - f(u_2)$.

Esperança normalizada

A Esperança Normalizada entre n unidades textuais é definida como a esperança média de uma unidade textual ocorrer em uma determinada posição, dada a ocorrência das outras $n-1$ unidades textuais também restritas às suas posições. A idéia básica da Esperança Normalizada é avaliar o custo, em termos de coesão, da perda de uma unidade textual em um n-grama. Portanto, quanto mais coeso um grupo de unidades textuais é, menos ele aceita a perda de um de seus componentes,

³³ p_{21} corresponde à distância com sinal entre u_2 e u_1 .

fazendo com que seja maior sua Esperança Normalizada. O conceito fundamental da Esperança Normalizada baseia-se na probabilidade condicional definida na próxima equação.

$$p(X = x | Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}$$

Cada unidade textual do corpus é mapeada em uma variável randômica discreta no Espaço de Probabilidade $(\Omega, A, P[.])$, por esta razão, a definição da probabilidade condicional pode ser aplicada com o intuito de medir a esperança de ocorrência de uma unidade textual em uma determinada posição, dada a ocorrência das outras $n-1$ unidades textuais também restritas por suas posições. No entanto, essa definição não acomoda o fator tamanho do n -grama. Um n -grama é associado a n probabilidades condicionais possíveis. Baseando-se em uma normalização da probabilidade condicional, a Esperança Normalizada propõe uma boa solução para representar todas as n probabilidades condicionais envolvidas por um n -grama em uma única fórmula. Para esse propósito, o conceito de Ponto Médio de Esperança (FPE) é introduzido pelos autores.

Em um n -grama genérico $[p_{11} u_1 p_{12} u_2 p_{13} u_3 \dots p_{1i} u_i \dots p_{1n} u_n]$, p_{11} é igual a zero e p_{1i} (para $i=2$ até n) denota a distância com sinal que separa a unidade textual u_i do seu pivô u_1 ³⁴. A extração de uma unidade textual por vez do n -grama genérico causa a ocorrência de qualquer um dos n eventos apresentados na Tabela 3.6, considerando que o sublinhado (“ ”) representa a unidade textual que está faltando no n -grama.

Tabela 3.6: $(n-1)$ -gramas e unidades textuais que estão faltando

(n-1)-grama	Unidade Textual Faltando
[<u> </u> $p_{12} u_2 p_{23} u_3 \dots p_{2i} u_i \dots p_{2n} u_n$]	$p_{11} u_1$
[$p_{11} u_1$ <u> </u> $p_{13} u_3 \dots p_{1i} u_i \dots p_{1n} u_n$]	$p_{12} u_2$
...	...
[$p_{11} u_1 p_{12} u_2 p_{13} u_3 \dots p_{1(i-1)} u_{(i-1)}$ <u> </u> $p_{1(i+1)} u_{(i+1)} \dots p_{1n} u_n$]	$p_{1i} u_i$
...	...
[$p_{11} u_1 p_{12} u_2 p_{13} u_3 \dots p_{1i} u_i \dots p_{1(n-1)} u_{(n-1)}$ <u> </u>]	$p_{1n} u_n$

Portanto, cada evento pode estar associado a uma probabilidade condicional que avalia a esperança de ocorrer a unidade textual que está faltando, dado o seu $(n-1)$ -grama. As n probabilidades condicionais são apresentadas nas seguintes equações, sendo que a primeira representa a extração do pivô do n -grama, enquanto que a segunda representa a extração de todas as unidades textuais exceto o pivô.

$$\frac{p(p_{11}u_1 | [p_{12}u_2 \dots p_{2i}u_i \dots p_{2n}u_n])}{p([p_{12}u_2 \dots p_{2i}u_i \dots p_{2n}u_n])} = \frac{p([p_{11}u_1 p_{12}u_2 \dots p_{2i}u_i \dots p_{2n}u_n])}{p([p_{12}u_2 \dots p_{2i}u_i \dots p_{2n}u_n])}$$

³⁴ Este n -grama é equivalente ao vetor $[p_{11} u_1 p_{12} u_2 p_{23} u_3 \dots p_{2i} u_i \dots p_{2n} u_n]$ onde p_{2i} denota a distância com sinal que separa a unidade textual u_i de u_2 e $p_{2i} = p_{1i} - p_{12}$ (para $i=3$ até n).

$$\forall i, i = 2..n, \quad p(p_{1i}u_i | [p_{11}u_1 \dots p_{1(i-1)}u_{(i-1)} p_{1(i+1)}u_{(i+1)} \dots p_{1n}u_n]) = \frac{p([p_{11}u_1 p_{12}u_2 \dots p_{1i}u_i \dots p_{1n}u_n])}{p([p_{11}u_1 \dots p_{1(i-1)}u_{(i-1)} p_{1(i+1)}u_{(i+1)} \dots p_{1n}u_n])}$$

Analisando-se essas equações, é possível notar que os denominadores se alteram de uma probabilidade para outra, enquanto que os numeradores permanecem inalterados. Assim, para se obter uma normalização precisa é importante avaliar o centro de gravidade dos denominadores por meio da definição de um evento médio denominado Ponto Médio de Esperança, que consiste na média aritmética dos denominadores de todas as probabilidades condicionais incorporadas pelas duas equações anteriores. Teoricamente, o Ponto Médio de Esperança é a média aritmética das n probabilidades conjuntas (isto é, a probabilidade de dois eventos ocorrerem juntos) dos (n-1)-gramas contidos em um n-grama, sendo então definido na seguinte equação:

$$FPE([p_{11}u_1 p_{12}u_2 \dots p_{1i}u_i \dots p_{1n}u_n]) = \frac{1}{n} (p([p_{12}u_2 \dots p_{2i}u_i \dots p_{2n}u_n]) + \sum_{i=2}^n p([p_{11}u_1 \dots \hat{p}_{1i}u_i \dots p_{1n}u_n]))$$

onde “^” corresponde a uma convenção utilizada em Álgebra e tem a função de escrever “^” sobre o termo omitido de uma dada sucessão que apresenta índices que variam de 2 até n.

Portanto, a normalização da probabilidade condicional se realiza através da introdução do Ponto Médio de Esperança em uma definição geral da probabilidade condicional, resultando em uma medida simétrica denominada Esperança Normalizada, sendo proposta como uma probabilidade condicional “justa”. Tal medida encontra-se na próxima equação:

$$NE([p_{11}u_1 \dots p_{1i}u_i \dots p_{1n}u_n]) = \frac{p([p_{11}u_1 \dots p_{1i}u_i \dots p_{1n}u_n])}{FPE([p_{11}u_1 \dots p_{1i}u_i \dots p_{1n}u_n])}$$

Esperança mútua

Vários estudos mostram que freqüência é uma das estatísticas mais relevantes para identificar termos compostos com padrões sintáticos específicos e que também é importante no contexto da extração de colocações interruptas. Assim, os autores deduzem que entre dois n-gramas com a mesma Esperança Normalizada, o n-grama que apresenta a maior freqüência é mais passível de ser uma unidade composta relevante. Portanto, a Esperança Mútua (ME) entre n palavras é definida baseando-se na Esperança Normalizada e na freqüência relativa na equação a seguir:

$$ME([p_{11}u_1 \dots p_{1i}u_i \dots p_{1n}u_n]) = p([p_{11}u_1 \dots p_{1i}u_i \dots p_{1n}u_n]) \times NE([p_{11}u_1 \dots p_{1i}u_i \dots p_{1n}u_n])$$

A Esperança Normalizada permite a avaliação do grau de coesão que liga todas as unidades textuais contidas em um n-grama ($\forall n, n \geq 2$) já que ela acomoda o fator tamanho do n-grama.

Processo de aquisição

O LocalMaxs³⁵ apresenta uma abordagem flexível e reajustável para o processo de seleção, pois ele se concentra na identificação do máximo local de valores da medida de associação. O LocalMaxs se baseia em duas suposições para a eleição de termos compostos do conjunto de todos os n-gramas de palavras valoradas, que são:

- 1) as medidas de associação mostram que quanto mais coeso um grupo de unidades é, maior será o seu escore³⁶;
- 2) termos compostos são grupos de palavras associadas localizadas.

A partir dessas suposições é possível deduzir que um n-grama é um termo composto se seu valor de medida de associação é maior ou igual aos valores de medidas de associação de todos os seus sub grupos de (n-1) palavras e, se ele é estritamente maior que os valores de medida de associação de todos os seus super-grupos de (n+1) palavras.

O LocalMaxs é definido da seguinte forma:

$$\forall x \in \Omega_{n-1}, \forall y \in \Omega_{n+1} \quad W \text{ é um termo multipalavra} \\ \text{se } (tamanho(W) = 2 \wedge assoc(W) > y) \vee \\ (tamanho(W) \neq 2 \wedge assoc(W) \geq x \wedge assoc(W) > y)$$

onde

assoc é uma medida de associação;

W é um n-grama;

Ω_{n-1} é o conjunto de todos os (n-1)-gramas contidos em *W*;

Ω_{n+1} é o conjunto de todos os (n+1)-gramas contendo *W*;

tamanho é a função que retorna o número de palavras de um n-grama de palavras.

³⁵ O algoritmo LocalMaxs foi descrito detalhadamente no artigo de Silva et al (1999).

³⁶ A medida de entropia condicional é uma das exceções.

3.3.2.2 Aplicação

Setup

O córpus de entrada foi extraído de uma coleção de informativos formulados pela *Procuradoria Geral da República* de Portugal no contexto do projeto “*PGR – Acesso Selectivo aos Pareceres da Procuradoria Geral da República*” fundado pela *Fundação para a Ciência e Tecnologia*³⁷. O córpus total contém mais de 3 milhões de palavras, mas apenas 1.5 milhão de palavras foram etiquetadas com o uso do etiquetador de rede neural desenvolvido por Marques (2000). Os experimentos realizados foram testados em um sub córpus contendo 500000 palavras anteriormente etiquetadas.

Avaliação

O LocalMaxs permite o teste de várias medidas de associação que respeitam a primeira suposição anteriormente apresentada (isto é, quanto mais coesa for a seqüência de palavras maior será a sua medida de associação³⁸) e a extração de termos compostos obtidos por composição. Por exemplo, o LocalMaxs conjugado com a Esperança Mútua, elege o termo composto *Presidente da República Jorge Sampaio*, construído a partir da composição dos termos extraídos *Presidente da República* e *Jorge Sampaio*. Isso é explicado pela seguinte análise: podemos esperar que existam muitos “Presidente da República”, portanto o valor de medida de associação de *Presidente da República Jorge* deveria ser menor que um para *Presidente da República*, já que existem muitas palavras possíveis além de *Jorge* que podem ocorrer depois de *Presidente da República*. Assim, a medida de associação de qualquer super grupo contendo a unidade *Presidente da República* teoricamente deveria ser menor que a medida de associação para *Presidente da República*. No entanto, se o primeiro nome do presidente é *Jorge*, a esperança de aparecer *Sampaio* é muito maior e o valor de medida de associação de *Presidente da República Jorge Sampaio* deveria ser maior que os valores de medida de associação de todos os seus sub grupos e super grupos, já que no segundo caso não se espera que alguma palavra possa fortalecer a unidade total *Presidente da República Jorge Sampaio*.

A vantagem do LocalMaxs está no fato dele evitar a definição de freqüência global e/ou limiares de medidas de associação baseados em experimentos.

³⁷ Mais informações sobre este projeto podem ser encontradas através do acesso do *web site* <http://kholosso.di.fct.unl.pt/~di/people.phmtl?it=CENTRIA&ch=gpl>.

³⁸ Os autores realizaram vários experimentos com medidas de associação diferentes: o coeficiente de associação (Church and Hanks, 1990); o coeficiente *dice* (Smadja, 1996); o ϕ^2 (Galé and Church, 1991) e o coeficiente *log-likelihood* (Dunning, 1993). A Esperança Mútua obteve os melhores resultados. O coeficiente ϕ^2 se traduz pela seguinte fórmula:

$$\phi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

Através da análise dos resultados é possível observar que o método puramente estatístico apresenta melhor desempenho que a metodologia híbrida para o caso da extração de termos compostos contendo duas palavras. Por outro lado, o método híbrido apresenta um desempenho muito melhor do que a metodologia puramente estatística para a extração de termos compostos contendo três a quatro palavras. E, finalmente, ambas as metodologias tendem a eleger o mesmo conjunto de termos compostos que contêm mais de cinco palavras. Esses resultados mostram que eleger o paradigma “uma metodologia” para a extração de multitermos deve ser questionado, pois diferentes métodos devem ser levados em conta dependendo do tamanho do termo a ser extraído.

Abaixo são apresentadas as vantagens e desvantagens do método híbrido.

Vantagens. Alguns dos problemas apresentados por métodos puramente estatísticos são parcialmente resolvidos. Por um lado, a introdução de informação lingüística internamente adquirida no processo de aquisição garante a extração de termos compostos com estruturas lingüísticas adequadas que permitem a inserção das mesmas em bancos de dados lexicais. Por exemplo, o sistema extrai o termo composto *concessão de bolsas* embora a seqüência sempre ocorra no cópús com a preposição *de* concatenada no final, ou seja, *concessão de bolsas de*. Já a abordagem puramente estatística elege toda a unidade *concessão de bolsas de*. Por outro lado, unidades compostas corretas que não foram extraídas com o uso de apenas estatísticas lexicais foram recuperadas beneficiando-se da identificação de padrões sintáticos relevantes. Por exemplo, *concessão de auxílios*, *jogo de futebol* e *estabelecimentos de ensino* foram recuperados já que eles incorporam o padrão sintático idiossincrático Substantivo+Preposição+Substantivo. Outros resultados são mostrados na Tabela 3.7.

Tabela 3.7: Resultados comparativos entre ambos os experimentos

Termos obtidos pelo experimento híbrido	Termo correspondente obtido pelo experimento estatístico
BAPTISTA MACHADO	BAPTISTA MACHADO
Direitos Conexos	De autor e Direitos Conexos
imagens recolhidas	de imagens recolhidas
espectáculos cinematográficos	espectáculos cinematográficos,
Direitos do homem	dos Direitos do homem
Ministério da Justiça	Boletim do Ministério da Justiça
Direito a Informação Desportiva	Direito a Informação
Direito de crónica	O Direito de crónica e
Ministro das Obras Públicas	das Obras Públicas
elaborar projectos	--
ensino primário	--
proceder a	devem proceder a
Campo de aplicação	seu campo de aplicação

No entanto, a introdução de informação lingüística também leva à incoerência e ruído no processo de recuperação, que serão descritos a seguir.

Desvantagens. A introdução de informação lingüística no processo de aquisição apresenta três maiores desvantagens. Com o intuito de definir algumas regras “gerais” para medir a precisão do sistema, propõe-se que um n-grama seja um termo composto se ele é gramaticalmente apropriado, ou seja, se ele é um substantivo/nome composto ou verbo composto. Conseqüentemente, os resultados de precisão são calculados com o uso do quociente entre o número de unidades gramaticalmente corretas e o número de todas as unidades extraídas.

Por um lado, resultados pobres de precisão são obtidos para o caso de bigramas, sendo que muitos padrões sintáticos irrelevantes evidenciam altos escores de associação que introduzem ruído no processo de aquisição. Por exemplo, altos escores associados aos padrões sintáticos Determinante+Substantivo e Preposição+Substantivo resultam na eleição de n-gramas não interessantes tais como *o concurso, da carreira*, que podem ser observados na Tabela 3.8.

Por outro lado, a maioria das unidades não contíguas eleitas revelam interrupções que raramente correspondem à ocorrência de diferentes modificadores. Por exemplo, a unidade léxica composto *Medalha do ___ militar* é eleita pelo LocalMaxs, apesar da probabilidade de ocorrer a palavra *mérito* (na lacuna) ser igual a um, ou seja, *mérito* é o único *token* que satisfaz a interrupção no cópuz.

E, por fim, o padrão sintático Substantivo+Preposição+Substantivo que recebe um alto escore de associação, e como ele corresponde a uma seqüência morfossintática bem conhecida no contexto de termos compostos, introduz ruído no processo de eleição. Por exemplo, a unidade léxica composta *Ministro dos Negócios* é de preferência eleita em comparação ao termo composto correto *Ministro dos Negócios Estrangeiros*. O mesmo ocorre com a seqüência *Direito de Informação Jornalística* que não é extraída enquanto o sistema elege a unidade insatisfatória *Direito de Informação*.

Tabela 3.8: Resultados comparativos entre ambos os experimentos

Termos obtidos pelo experimento estatístico	Termo correspondente obtido pelo experimento híbrido
Isenção de	isenção de propinas
Acesso a	acesso a categoria
anos de serviço	anos de serviço efectivo
Estatuto social do Bombeiro	“Estatuto social do Bombeiro”
Licenciamento municipal de obras	Licenciamento municipal de obras públicas
A primeira	--

3.3.2.3 Extensão

No contexto de Recuperação de Informação, os autores acreditam que experimentos deverão ser realizados a fim de decidir se a introdução de informação lingüística supera ou não a metodologia puramente estatística para o processo de recuperação. No entanto, não acreditam que mais experimentos tenham de ser realizados da mesma forma a fim de avaliar a interdependência real entre o estágio de filtragem e o processo de aquisição. A metodologia foi testada no mesmo *cópus* etiquetado anteriormente com um conjunto de etiquetas mais completo e os resultados obtidos não apresentaram melhorias. Ao contrário, eles evidenciaram resultados piores para precisão e revocação mostrando que informação mais rica não necessariamente beneficia o processo de aquisição.

3.3.2.4 Recursos e ferramentas utilizados

O etiquetador de redes neurais desenvolvido por Marques (2000) foi utilizado para etiquetar 1.5 milhão de palavras do *cópus* de entrada.

O sistema foi, a princípio, desenvolvido para propósitos experimentais no *script shell* para Linux. A linguagem *gawk*, apesar de fornecer características interessantes para manipulação de texto, apresenta certa deficiência no poder de representação, o que leva a uma incapacidade de manipular programação otimizada. Buscando melhorar sua eficiência, o sistema está sendo desenvolvido utilizando-se a linguagem C, utilizando estruturas de dados estruturados e algoritmos de busca.

Além disso, um algoritmo otimizado tem sido projetado a fim de poder trabalhar com grande quantidade de dados resultantes do cálculo de todas as possíveis combinações de n-gramas, em razão de ele evitar o cálculo duplicado de n-gramas quando o texto está sendo processado³⁹. Finalmente, com o intuito de reduzir o espaço de memória necessário, é utilizado um compressor Huffman-like (menores códigos são atribuídos às palavras mais freqüentes) que alcança uma taxa de compressão de 52.8%.

3.4 Experimentos: estatístico e híbrido no trabalho de Dias et al (2000)

Nesta seção, serão apresentados os resultados obtidos através da realização de dois experimentos distintos. O primeiro experimento realiza a extração de termos compostos de um texto em Português considerando apenas estatísticas léxicas. Neste caso, cada n-grama é associado a seu

³⁹ Duplicações são n-gramas equivalentes que incorporam diferentes representações, por exemplo, [*United +1 States*] e [*States -1 United*].

valor de Esperança Mútua e o LocalMaxs extrai todos os termos possíveis do conjunto de todos os n-gramas valorados. Já no segundo experimento, o sistema extrai unidades léxicas compostas relevantes terminologicamente, através da combinação de estatísticas léxicas com informação lingüística adquirida do mesmo cópús em Português anteriormente etiquetado. Nesse caso, cada n-grama é ligado ao seu n-grama de etiqueta correspondente e o valor da medida de associação final do n-grama é o produto entre seu valor de Esperança Mútua e o valor de Esperança Normalizada de seu n-grama de etiqueta associado.

3.4.1 Abordagem puramente estatística

Este experimento realiza a extração de termos compostos através da utilização exclusiva de regularidades de palavras estatísticas. Dessa forma, etiquetas *part-of-speech* são consideradas menos significativas para o processo de aquisição e todas as etiquetas são excluídas do cópús etiquetado de entrada original. O conjunto de todos os n-gramas de palavras é calculado do “texto de palavras” e cada n-grama de palavras é associado ao seu valor de Esperança Mútua. Para finalizar, o LocalMaxs extrai os termos compostos candidatos do conjunto de todos os n-gramas de palavras com valor. Os resultados mostram que três categorias de termos compostos foram extraídas: termos base, termos obtidos por composição e termos obtidos por modificação.

Termos base: correspondem a n-gramas de palavras não contíguos que não contêm qualquer outro n-grama de palavra extraído. Alguns termos base são mostrados na Tabela 3.9.

Tabela 3.9: Termos Base

ME	Freq.	Termos Base
0.000107062	12	Comunicação Social
0.000105067	3	Oliveira Ascensão
0.000105067	3	cobrar taxas
0.000105519	5	Liberdade de Informação
0.000100499	2	oferta e procura
0.000100499	2	entrar em vigor
0.000100499	2	Associação dos Arquitectos Portugueses
0.000102732	4	Regulamento Municipal de Edificações Urbanas
0.000100499	2	rendimento familiar anual ilíquido per capita

Termos obtidos por composição: correspondem a termos compostos que são construídos de um ou mais termos base. Esta categoria incorpora construções específicas de justaposição, substituição, posposição e coordenação. Alguns exemplos de termos extraídos por composição são mostrados na Tabela 3.10⁴⁰.

⁴⁰ Os termos base são identificados por colchetes.

Tabela 3.10: Termos obtidos por composição

ME	Freq.	Termos obtidos por composição
0.000102	3	[Direcção Geral] dos Desportos
0.000102	4	Teoria Geral do [Direito Civil]
0.000128	2	[licenciamento municipal] de obras particulares
0.000102	2	[estabelecimentos oficiais] não militares
0.000102	2	[pessoas colectivas] de [utilidade pública]
0.000102	2	[artigo 35°] da [Lei de Imprensa]
0.000102	2	[Liga Nacional] de [Futebol Profissional]

Termos obtidos por modificação: correspondem a n-gramas de palavras não contíguos que contêm exatamente uma interrupção⁴¹. De fato, a inserção de modificadores em um termo implica a introdução de um fator de flexibilidade que corresponde a uma interrupção. Assim, várias ocorrências de palavras podem modificar uma seqüência complexa de palavras. Por exemplo, as palavras *Europeu* e *Mundial* podem modificar a seqüência complexa *Conselho das Telecomunicações* através da introdução de algum fator de especificação, resultando nas seguintes seqüências de palavras modificadas: *Conselho Europeu das Telecomunicações* e *Conselho Mundial das Telecomunicações*. A análise de n-gramas de palavras não contíguos permite representar os termos obtidos por modificação através da identificação de cada conjunto de modificadores como uma interrupção na seqüência de palavras. Dessa forma, o LocalMaxs associado à Esperança Mútua elegeria o termo composto *Conselho ___ das Telecomunicações* que incorpora ambas seqüências complexas modificadas. A Tabela 3.11 mostra que modificadores podem incorporar uma grande quantidade de categorias morfosintáticas.

Tabela 3.11: Termos obtidos por modificação

ME	Freq.	Termos obtidos por modificação	Modificadores
0.000105	2	Controle ___ fronteiras	De Das
2.708e-05	2	Transporte de ___ perigosas	Matérias Substâncias
2.708e-05	4	Artigo ___ do regulamento	3° 6° 32° 45°
0.000102	2	Proposta de ___ do Conselho	Directiva

⁴¹ Os termos caracterizados por várias interrupções exibem unidades relevantes lexicograficamente que são, na maioria das vezes, terminologicamente irrelevantes.

			Regulamento
--	--	--	-------------

Limites do processo estatístico. Apesar do interesse na extração de termos compostos no contexto da Recuperação de Informação, metodologias estatísticas extraem unidades léxicas compostas que podem não ser consideradas termos (Habert and Jacquemin, 1993 apud Dias et al, 2000) e a metodologia aqui utilizada não evita esse tipo de problema. Uma análise detalhada do resultado mostra que locuções adverbiais, adjetivas, preposicionais e conjuntivas também são recuperadas. O LocalMaxs associado à Esperança Mútua extrai a unidade léxica composta *a assembleia municipal*, sendo que o determinante *a* ocorre sempre que a unidade composta *assembleia municipal* ocorre no texto. Nesse caso só deveria ser considerado o termo composto *assembleia municipal* como sendo a única unidade relevante. Um outro problema encontrado durante a extração de unidades léxicas compostas baseada somente em medidas estatísticas em um ambiente não contínuo é a eleição de unidades léxicas não contíguas incorretas. Por exemplo, a unidade *a ___ técnicos responsáveis* é eleita pelo sistema, sendo que o determinante/preposição *a* ocorre sempre que a unidade *técnicos responsáveis* ocorre, como pode ser observado na Tabela 3.12.

Tabela 3.12: Concordâncias para técnicos responsáveis

oficial	a	exigir	aos	técnicos	responsáveis	pelos
cancelada	a	inscrição	dos	técnicos	responsáveis	pelo
obdecer	a	qualificação	dos	técnicos	responsáveis	por

Uma forma de resolver esse problema é a introdução de informação lingüística no processo de aquisição, além de regularidades sintáticas que também devem ser consideradas a fim de filtrar unidades lexicais incorretas. Essa é a meta do segundo experimento.

3.4.2 Abordagem híbrida

Buscando superar os problemas apresentados pela maioria das abordagens estatísticas, métodos híbridos (lingüístico-estatísticos) definem co-ocorrências de interesse em termos de padrões sintáticos e regularidades estatísticas. A metodologia aqui utilizada combina estatísticas de palavras com informação lingüística adquirida internamente. Em paralelo à avaliação do valor da Esperança Mútua de cada n-grama de palavra, todas as palavras do córpus de entrada são podadas e a Esperança Normalizada é aplicada a cada n-grama de etiqueta calculado anteriormente do “texto de etiquetas”. Por fim, o LocalMaxs é aplicado sobre o conjunto de todos os n-gramas de palavras associados aos seus novos valores de medida de associação, obtidos a partir do produto de seus valores de Esperança Mútua com os valores de Esperança Normalizada de seus n-gramas de etiquetas associados.

3.5 Considerações finais

Através da análise das três abordagens apresentadas nesse capítulo, nota-se que nenhum dos sistemas de extração automática de candidatos a termo é totalmente satisfatório, visto que todos eles produzem grande quantidade de silêncio principalmente os que se baseiam em estatística; e todos geram uma grande quantidade de ruído, especialmente aqueles que se baseiam em conhecimento lingüístico e que utilizam padrões morfossintáticos (são a maioria) para identificar os termos compostos, tomando como referência apenas os aspectos formais da unidade terminológica.

A fim de melhorar os resultados de tais sistemas, obtendo-se, assim, a redução do silêncio e ruído gerados, Estopà Bagot (1999) propõe um aprofundamento em dois tipos de estudos. Por um lado, seriam necessários mais estudos lingüísticos, por exemplo, sobre:

- As categorias gramaticais que são prováveis de serem termos nos diferentes campos de especialidade
- As relações semânticas entre os diferentes constituintes de uma unidade terminológica
- A disposição dos termos nos textos
- As relações em diferentes línguas dos termos de uma mesma temática.

Por outro lado seria importante trabalhar com sistemas computacionais que:

- Alternassem de forma mais ativa os métodos estatísticos com os lingüísticos
- Melhorassem os cálculos estatísticos
- Combinassem mais de uma estratégia
- Melhorassem as interfaces para favorecer a interação usuário-computador.

Por fim, para que se obtenham bons resultados no âmbito da extração automática de termos, é necessário que haja uma forte interação dos métodos estatísticos com os lingüísticos, fazendo com que os extratores consigam reduzir grande parte de silêncio e ruído, reconhecendo, assim, a maioria dos termos (monoléxicos ou poliléxicos) do texto.

Para a implementação dos métodos anteriormente descritos, as seguintes ferramentas foram utilizadas, agrupadas em classes genéricas mostradas abaixo:

- 1) Ferramentas que fazem a etiquetagem morfossintática (dão a categoria gramatical para cada palavra de uma oração):
 - 1-a) O etiquetador de redes neurais desenvolvido por Marques (2000)
 - 1-b) analisador morfossintático

- 1-c) etiquetador morfológico (*part-of-speech*)
- 1-d) etiquetador morfossintático
- 1-e) etiquetador de Brill
- 2) Tokenizador (analisador léxico de compiladores):
 - 2-a) tokenizador
- 3) Lematizador (dão a canônica/lema de cada palavra):
 - 3-a) lematizador
- 4) Ferramenta de busca em córpus (aceita consultas em expressão regular):
 - 4-a) um processador de consulta a córpus geral (CQP)
 - 4-b) um macroprocessador para a linguagem de consulta ao CQP
- 5) Concordanceador (apresenta todas as instâncias de uma palavra consultada mostrando o seu contexto à esquerda e direita):
 - 5-a) XKWIC -- Os autores utilizaram o concordanceador de (Christ 1994b)
- 6) Parsers (análise sintática):
 - 6-a) Chunk Parsing
 - 6-b) módulo de análise gramatical baseado no formalismo Slot Grammar do Inglês (ESG) de McCord (1991)
 - 6-c) analisador sintático parcial de sintagmas nominais de Ramshaw and Marcus (1995)
- 7) Compressor:
 - 7-a) compressor Huffman-like

A Tabela 3.13 mostra as ferramentas que foram utilizadas em cada método das abordagens estatística, lingüística e híbrida apresentadas.

Tabela 3.13: Ferramentas utilizadas em cada método

		MÉTODOS/SEÇÕES								
		3.1.1	3.1.2	3.1.3	3.1.4	3.1.5	3.2.1	3.2.2	3.3.1	3.3.2
FERRAMENTAS	1-a)									X
	1-b)						X			
	1-c)						X	X		
	1-d)								X	X
	1-e)							X	X	
	2-a)						X			
	3-a)						X			
	4-a)						X			
	4-b)						X			
	5-a)						X			
	6-a)						X			
	6-b)							X		
	6-c)							X		
	7-a)									X

Para adaptar os métodos revisados para o Português, as ferramentas utilizadas também devem ser treinadas para essa língua (quando forem estatísticas) ou utilizadas ferramentas similares (quando baseadas em conhecimento lingüístico). Como apresentado na introdução, o NILC possui as ferramentas 1, 2, 3 e 6, sendo que concordanceadores, processadores de consulta e compressores são independentes de língua e facilmente encontrados na *Web*.

Capítulo 4

Recursos e ferramentas utilizados

Tendo definido o objetivo desse trabalho de mestrado, que consiste na avaliação de métodos para a extração automática de terminologia, esse capítulo foi elaborado com o propósito de apresentar os meios que permitiram a realização das atividades planejadas para a avaliação das abordagens estatística, lingüística e híbrida descritas no Capítulo 3, tomando-se um *cópus* e uma lista de termos da área de Revestimentos Cerâmicos. Após uma análise dos métodos utilizados em cada abordagem, foi possível fazer um levantamento de recursos e ferramentas que seriam necessários para a implementação de cada um desses métodos ou adaptações deles. Na Seção 4.1 será mostrado o processo de seleção e preparação do *cópus*, que foi utilizado na avaliação dos métodos. A lista de referência, descrita na Seção 4.2, foi o resultado da intersecção da Lista de Termos resultante do Trabalho de Almeida (2000) com o *cópus* desse trabalho de mestrado, sendo um elemento indispensável no processo de avaliação de cada método implementado. As Seções 4.3 e 4.4 são dedicadas, respectivamente ao tokenizador desenvolvido no NILC e ao etiquetador MXPOST, que foram utilizados com o intuito de realizar um pré-processamento no *cópus* para as abordagens lingüística e híbrida. A idéia de se usar o concordanceador AntConc (Seção 4.5) foi de visualizar as concordâncias produzidas para as expressões lingüísticas utilizadas no método da abordagem lingüística. Na Seção 4.6 é descrito o pacote NSP, utilizado para a implementação dos métodos da abordagem estatística. E, na última seção (Seção 4.7), são apresentadas as características do *CópusEco*, que foi utilizado com a finalidade de se realizar buscas por expressões lingüísticas.

4.1 Seleção e preparação do *cópus* alvo

O *cópus* utilizado nesse trabalho foi criado com artigos que se encontram no *site* da Revista Cerâmica Industrial⁴², autorizados por uso no Projeto Lácio-Web (Aluísio et al, 2003). Os textos estão agrupados pelos anos em que foram publicados, 1996-2003, e totalizam 196, possuindo, cada texto, uma média de 7 a 8 páginas (aproximadamente 4000 palavras).

Todos os textos presentes no *site* acima estão no formato pdf. Porém, para que eles pudessem ser processados pelos métodos propostos nesse trabalho, deveriam estar no formato texto. Por essa razão, nem todos os textos do *site* foram utilizados, visto que ocorreram alguns problemas no processo de transformação deles para o formato txt.

A princípio, dos 196 artigos do *site*, 141 estavam sendo utilizados, já que o restante apresentava problemas na transformação. No entanto, percebeu-se que 55 desses artigos eram de

⁴² <http://www.ceramicaindustrial.org.br/>

autores estrangeiros, 4 de autoria híbrida (escritos por autores estrangeiros e nacionais), e 4 de autoria duvidosa (não era mencionada a nacionalidade do autor).

Diante dessas constatações, a montagem do *córpus* foi reavaliada, pois, quando se trabalha com extração de termos em um *córpus* que contém textos em língua portuguesa, variante brasileira, se textos traduzidos fossem inseridos, o critério de coerência com a língua em que os textos do *córpus* estão escritos seria desrespeitado. A retirada desses textos, por outro lado, comprometeria o tamanho do *córpus*, uma vez que uma das abordagens utilizada é a estatística, sendo essa dependente, significativamente, do tamanho do *córpus*.

Por essa razão, decidiu-se contactar o responsável pela revista para esclarecer se esses textos, depois de traduzidos, eram analisados por um especialista da área, e, caso isso ocorresse, não haveria problemas em deixá-los no *córpus*. Como a resposta foi afirmativa, os textos estrangeiros, híbridos e de autoria duvidosa foram novamente incluídos, e além desses já existentes, foram ainda inseridos mais 23 artigos (dado que o problema de transformação foi parcialmente resolvido), totalizando 164 textos para compor o *córpus* de trabalho.

Para a transformação desses textos para o formato texto, foi utilizada a ferramenta denominada EXTEX (Extração de Texto de Ficheiros Formatados)⁴³. Uma característica dessa ferramenta, ao realizar a transformação, é a de que o texto transformado não é totalmente igual ao texto original. Ele se apresenta com junção de algumas palavras, preserva os índices de referência bibliográfica e as notas de rodapé anexadas às palavras, e a hifenização dos textos no formato pdf. Para resolver esses problemas, esses textos foram submetidos a um processo cuidadoso de correção manual.

Vale ressaltar também que todos os arquivos do *córpus* foram pré-processados para a retirada de informações de autoria e filiação, referências bibliográficas, figuras, tabelas e quadros, fazendo com que o tamanho médio dos artigos diminuísse de 8 para 5 páginas. O tamanho total do *córpus* em palavras é 448352.

Também foi encontrada grande quantidade de erros de digitação e gramaticais, dentre eles, erros de concordância em gênero e em número e erros de acentuação. Ainda foi possível perceber que alguns termos encontrados no *córpus* apresentavam hífen e não estavam lematizados, enquanto que na lista de termos do trabalho de Almeida (2000), obtida pela extração manual, esses termos se encontravam não hifenizados e lematizados (veja na Seção 4.2). Para minimizar os erros gramaticais, foi realizada uma varredura no *córpus* com o auxílio de um processador de textos, buscando corrigir os erros encontrados, podendo-se, dessa forma, analisar os dados de forma mais precisa. Assim, a avaliação das três abordagens para a extração de termos foi realizada com um *córpus* pré-processado. Esse *córpus* será referenciado nesse trabalho como *córpus* alvo.

⁴³ <http://poloclup.linguatca.pt/ferramentas/extex/>

4.2 Lista de termos de referência (da área de Revestimentos Cerâmicos)

Uma lista de termos foi gerada a partir de um *cópus* formado de fontes escritas (documentos da ABNT, revistas científicas e/ou de divulgação, lista de termos em obras especializadas) e fontes de língua oral (entrevistas e outros tipos de interação oral, como palestras e seminários) compiladas em uma estrutura conceitual da área de Revestimentos Cerâmicos, elaborada e alimentada por extração manual de termos, no trabalho de Almeida (2000). Ela consiste de 351 unigramas, 169 bigramas e 151 trigramas, que totalizam 747, contando com os termos que apresentam mais de 3 *tokens*. Os termos que apresentam um número de *tokens* superior a 3 no trabalho de Almeida (2000) não foram utilizados nesse trabalho. Todos os termos acima se transformaram ou se transformarão em verbetes do Dicionário de Revestimentos Cerâmicos (DiRC), portanto, já foram avaliados por especialistas da área.

Porém, nem todos os termos encontrados na lista acima estão presentes no *cópus* e, por essa razão, foi realizada a intersecção da mesma com o *cópus* alvo, obtendo uma lista com um total de 264 unigramas, 74 bigramas e 43 trigramas, totalizando 381. Essa última lista que será referenciada como lista de referência no decorrer desse trabalho, e ela pode ser encontrada no Apêndice B.

4.3 O tokenizador desenvolvido no NILC

O *cópus* foi pré-processado utilizando-se um tokenizador desenvolvido no NILC chamado *Sentencer*, que é um tokenizador e segmentador sentencial para português, que tokeniza um texto de entrada, inserindo um caractere de fim de linha ao fim de cada sentença. Linhas em branco marcam fronteiras de parágrafo. Apenas caracteres de fim de linha, como ponto final/de interrogação/exclamação e reticências, são considerados possíveis finais de sentença.

O programa *Sentencer* trata de abreviações como “Dr.”, “Prof.”, não considerando, nesse caso, o ponto final como um caractere de fim de linha, ao contrário, o ponto é desconsiderado. Além disso, o programa *Sentencer* também apresenta a função de separar os caracteres (como aspas, vírgulas, pontuações, entre outros) dos *tokens*.

A linha de comando utilizada para tokenizar o arquivo de entrada é:

```
sentencer sentencer.lex <arquivo de entrada.txt >arquivo de saída.txt
```

O arquivo *sentencer.lex* traz um léxico do português, com artigos, substantivos, adjetivos, verbos, etc., e seu uso permite a identificação de fronteiras de sentenças, o tratamento de abreviações e a separação de caracteres como aspas, vírgulas, pontuações, entre outros.

4.4 O etiquetador MXPOST treinado com um corpus de textos em português

Após o corpus alvo ter sido tokenizado pelo *Sentencer*, ele foi etiquetado utilizando-se o tagger MXPOST (Ratnaparkhi, 1996), que foi treinado com um corpus manualmente etiquetado de 104.963 palavras selecionadas do corpus do NILC. Esse corpus de 104.963 palavras contempla três gêneros diferentes: textos jornalísticos (56.653 palavras), didáticos (16.256 palavras), e literários (32.054 palavras). O tagset inicial utilizado contém 36 etiquetas além das de pontuação, e sua precisão foi de 89,6%.

O tagger MXPOST foi retreinado no NILC, utilizando-se um tagset simplificado que possui 15 etiquetas⁴⁴ e sua precisão foi de 97%. A maioria do corpus do NILC encontra-se hoje marcada com esse tagset simplificado.

Para usar o MXPOST no arquivo de entrada, cada *token* deve estar separado por um espaço em branco, ou seja, nenhum caractere, incluindo pontuação, deve estar anexo às palavras; essa foi uma das razões do uso do programa *Sentencer*. O arquivo de saída gerado apresenta o seguinte formato: “token1_TAG1 token2_TAG2 ...”.

4.5 O concordanceador AntConc

Com o intuito de fazer buscas por palavras no corpus e visualizar o contexto das mesmas para a escrita de programas da abordagem lingüística, foi utilizado o concordanceador AntConc 2.5.1⁴⁵. Essa ferramenta produz um conjunto de linhas de concordância de texto, tendo o usuário selecionado um arquivo ou um diretório, onde a busca será realizada, e tendo ele informado a palavra que se deseja buscar no arquivo ou diretório escolhido. As linhas de concordância então geradas trazem em destaque a palavra que foi buscada, acompanhada de seus contextos à esquerda e à direita. Além dessa ferramenta ter a opção de busca por uma determinada palavra, ela também permite a busca por expressões regulares e as opções de sensibilidade à caixa e subcadeias.

⁴⁴ I – interjeição, LOCU – locução, PREP – preposição, N – substantivo, NP – nome próprio, VERB – verbo, ADJ – adjetivo, AUX – verbo auxiliar, ADV – advérbio, PRON – pronome, CONJ – conjunção, NUME – numeral, ART – artigo, RES – resíduo, PDEN – palavra denotativa e mais 4 tipos de contrações: PREP+ART, para palavras como “da”, “na”, PREP+PD, para palavras como “nesta”, “naquela”, “nessa”, PREP+PPR, para palavras como “dela”, “nela” e PREP+N, para palavras como “d’alma”, “d’água”, “d’arte”)

⁴⁵ <http://antpcl.ice.ous.ac.jp/>

4.6 O pacote estatístico NSP

As medidas utilizadas pela abordagem estatística estão incorporadas no pacote NSP (N-gram Statistics Package)⁴⁶, escrito em Perl. O pacote NSP foi implementado por Ted Pedersen, Satanjeev Banerjee e Amruta Purandare na Universidade de Minnesota, Duluth. Ele é constituído por um conjunto de programas que auxilia na análise de n-gramas em arquivos texto. No pacote, um n-grama é definido como uma seqüência de ‘n’ *tokens* que ocorrem dentro de uma janela de pelo menos ‘n’ *tokens* no texto.

Esse pacote é encontrado em diversas versões, e a versão utilizada nesse trabalho foi a 0.57. Essa versão apresenta dois programas principais que são o ‘count.pl’ e o ‘statistic.pl’, cujas funções serão apresentadas nesta seção. Essa versão proporciona dez medidas de associação para bigramas e 2 para trigramas. Para bigramas foram utilizadas a Informação Mútua, *Log-likelihood* e Coeficiente *Dice* e para trigramas foram utilizadas a Informação Mútua e *Log-likelihood*. Para unigramas, a única medida proporcionada por essa versão do pacote NSP é a freqüência.

O comando necessário para produzir unigramas, bigramas e trigramas junto com suas freqüências é:

```
count.pl [opções] arquivo_de_saída.txt arquivo_de_entrada.txt
```

Em “opções” pode-se especificar o n-grama, caso ele seja diferente de 2 em razão desse ser o padrão. Por exemplo, para produzir unigramas utiliza-se “--ngram 1”. Também em “opções” pode-se especificar o arquivo com a regra de formação de *tokens* (“--token nome_do_arquivo.pl”), o arquivo que contém a *stoplist* (“--stop nome_do_arquivo.pl”), limitar a lista de n-gramas utilizando-se somente aqueles que apresentam freqüência equivalente ou superior a um determinado valor especificado (“--remove N”). Além dessas opções, existem outras que não serão aqui descritas por não terem sido utilizadas, mas podem ser encontradas no pacote NSP.

Considere que o conteúdo apresentado a seguir pertença ao arquivo texto de entrada “entrada.txt”, por exemplo:

```
primeira linha de texto  
segunda linha  
e uma terceira linha de texto
```

Ao utilizar a linha de comando “count.pl saída.txt entrada.txt”, a seguinte saída é produzida (arquivo “saída.txt”):

⁴⁶ <http://www.d.umn.edu/~tpederse/nsp.html>

Quadro 4.1: Saída do programa count.pl

```
11
linha<de>2 3 2
de<texto>2 2 2
terceira<linha>1 1 3
linha<e>1 3 1
texto<segunda>1 1 1
primeira<linha>1 1 3
e<uma>1 1 1
uma<terceira>1 1 1
segunda<linha>1 1 3
```

O número 11 na primeira linha indica que o arquivo de entrada “entrada.txt” apresenta um total de 11 bigramas. Nas próximas linhas, esses bigramas foram listados, considerando que cada *token* é separado pelo sinal “<”. Depois do último “<” em cada linha encontram-se 3 números, sendo que o primeiro representa o número de vezes que o bigrama ocorre no arquivo texto de entrada. Dessa forma, o bigrama “linha<de>” ocorre 2 vezes no texto de entrada. O segundo número está relacionado ao número de bigramas em que o *token* “linha” ocorre do lado esquerdo. Assim, “linha” ocorre no lado esquerdo de 3 bigramas. E, finalmente, o terceiro número representa o número de bigramas em que o *token* “de” ocorre do lado direito.

Já para realizar o cálculo das medidas coeficiente *log-likelihood*, informação mútua e coeficiente *dice* para bigramas, é utilizado o comando:

```
statistic.pl nome_do_arquivo_da_medida
nome_do_arquivo_de_saida.nome_do_arquivo_da_medida arquivo_de_bigramas.txt
```

O mesmo comando é utilizado para o cálculo das medidas coeficiente *log-likelihood* e informação mútua para trigramas, adicionando-se a opção “--ngram 3” depois de “statistic.pl”. Um exemplo, para o cálculo do coeficiente *dice*, é “statistic.pl dice teste.dice bigrama.txt”. O resultado gerado ao se executar essa linha de comando é uma saída similar àquela apresentada anteriormente, acrescentando-se o ranking e o score dos bigramas antes dos 3 outros números. Dessa forma, os bigramas são classificados de acordo com os scores que apresentam. Considerando como entrada o arquivo de saída gerado no Quadro 4.1 e utilizando-se a linha de comando

```
statistic.pl dice saida2.dice saida.txt
```

obtém-se o arquivo “saida2.dice”:

```
11
```

```
de<=>texto<=>1 1.0000 2 2 2
e<=>uma<=>1 1.0000 1 1 1
uma<=>terceira<=>1 1.0000 1 1 1
texto<=>segunda<=>1 1.0000 1 1 1
linha<=>de<=>2 0.8000 2 3 2
terceira<=>linha<=>3 0.5000 1 1 3
linha<=>e<=>3 0.5000 1 3 1
primeira<=>linha<=>3 0.5000 1 1 3
segunda<=>linha<=>3 0.5000 1 1 3
```

Comparando-se esse arquivo com o anterior, é possível notar que existem dois números adicionais, sendo que o primeiro representa o ranking do bigrama, que é obtido a partir do segundo número, que representa o escore do bigrama e é calculado utilizando-se, nesse caso, a medida estatística coeficiente *dice*. Dessa forma, os bigramas foram classificados em ordem crescente de seus rankings. Os três números restantes são os mesmos apresentados anteriormente.

O resultado do cálculo do escore das medidas estatísticas é apresentado com apenas 4 casas decimais, que é o número padrão. Para alterar esse número, utiliza-se a opção “--precision n”, modificando-se a precisão para um determinado número n de casas decimais. Além dessa opção, o pacote apresenta algumas outras para serem utilizadas com o programa “statistic.pl”, que não serão aqui descritas.

Como os arquivos do cópús foram agrupados em pastas dentro de um determinado diretório, foi utilizada a opção “--recurse” para acessá-los.

O pacote NSP, além de produzir todos os unigramas, bigramas e trigramas encontrados no cópús, permite que se faça algumas limitações e incrementos quanto ao que se deseja. Por exemplo, quando se geraram as listas de unigramas, bigramas e trigramas utilizando-se apenas a função “count.pl”, as acentuações encontradas no cópús não foram reconhecidas, já que a língua padrão do pacote é a língua inglesa, sendo então necessário construir uma regra de formação de *token* que aceitasse acentuação. Essa regra também foi essencial para a eliminação de alguns caracteres que não seriam importantes na busca por termos, tais como aspas, números, pontuações, entre outros.

Na construção da regra de formação de *tokens*, foi necessário utilizar a tabela ASCII estendida, já que o pacote apenas reconhece padrões de formação de *tokens* nesse formato. A princípio, palavras hifenizadas também não eram geradas como se encontravam no cópús e sim separadas por meio do hífen. Nesse caso, a regra de formação de *tokens* também foi aplicada. A regra de formação do *token* utilizada é:

```
/([a-zA-Z-]) → representa caracteres alfabéticos que podem apresentar hífen
```

<code>[\\w\\xb0]</code>	→ representa o “°” (grau)
<code>[\\w\\xc0-\\xc5]</code>	→ representa a letra “á” maiúscula com as acentuações possíveis
<code>[\\w\\xc7-\\xcf]</code>	→ representa o “ç”, as letras “é” e “í” com acentuações (maiúsculos)
<code>[\\w\\xd1-\\xd6]</code>	→ representa o “ñ” e a letra “ó” com acentuações (maiúsculos)
<code>[\\w\\xd9-\\xdc]</code>	→ representa a letra “ú” maiúscula com acentuações
<code>[\\w\\xdf-\\xe5]</code>	→ representa a letra “ß” e a letra “á” minúscula com acentuações
<code>[\\w\\xe7-\\xef]</code>	→ representa o “ç”, as letras “é” e “í” com acentuações (minúsculos)
<code>[\\w\\xf1-\\xf6]</code>	→ representa o “ñ” e a letra “ó” com acentuações (minúsculos)
<code>[\\w\\xf9-\\xfc]+/</code>	→ representa a letra “ú” minúscula com acentuações

Note que o caractere “[”, presente na regra de formação de *tokens*, indica alternância, e o caractere “+” indica que a expressão em questão pode ocorrer uma ou mais vezes.

Após a execução dessas tarefas, o resultado produzido pelas listas de unigramas, bigramas e trigramas ainda não foram satisfatórios, visto que as palavras que apareciam com maior frequência representavam um grupo de palavras funcionais, tais como preposições, artigos, conjunções, e também uma quantidade significativa de advérbios que não apresentam nenhum valor terminológico. Para resolver esse problema, foi construída uma *stoplist* com essas palavras, a fim de obter uma lista menor, apresentando candidatos com maior probabilidade de serem termos.

A princípio, havia sido construída uma *stoplist* com preposições, artigos, conjunções e alguns advérbios. No entanto, essa *stoplist* encontrava-se incompleta, visto que não apresentava uma quantidade significativa de preposições e advérbios que apareciam no cópús. Também foram incluídos algumas palavras-chave (como “Conclusões”), alguns numerais e alguns símbolos (como “-”). A seguir encontra-se um trecho da *stoplist* construída:

```
@stop.mode=OR
/^E$/
/^É$/
/^À(S)?$/
/^A?O?S?$/
```

A *stoplist* inteira encontra-se no Apêndice C. Na primeira linha foi utilizado o operador booleano “or”, pois se apenas um dos *tokens* contidos em algum bigrama presente no cópús se encontra especificado nessa lista, o bigrama já é eliminado da lista de bigramas produzida. Nas linhas posteriores é possível notar que é utilizada a marcação “/^ \$/” para delimitar o *token*. Dessa

forma, é fácil perceber que a segunda e terceira linhas foram utilizadas para eliminar as letras “e” e “é”, respectivamente. Na quarta e quinta linhas aparece o ponto “?”, que é utilizado para tornar o caractere ou caracteres em questão como opcionais. Assim, na linha quatro podem ser formados os *tokens* “à” e “às” (nesse caso o “s” é opcional). Na linha cinco podem ser formados os *tokens* “a”, “o”, “s”, “ao”, “aos”, “as” e “os” (nesse caso as 3 letras, “a”, “o” e “s”, são opcionais).

É importante notar que, na *stoplist*, as letras são apresentadas em caixa alta, em razão de o cópús ter sido todo transformado para caixa alta.

Para os unigramas, essa mesma *stoplist* foi utilizada, enquanto que para trigramas, ao invés do operador “or” ser utilizado na primeira linha, foi empregado o operador “and”, a fim de evitar que fossem eliminados trigramas como “absorção de água” (apresenta a preposição “de”), já que essa opção (“and”) elimina um trigrama somente quando os 3 componentes do mesmo se encontram na *stoplist*, ao contrário do que ocorre ao se utilizar o operador “or”.

4.7 O cópús CópúsEco da área de Ecologia e sua lista de termos

O cópús CópúsEco foi utilizado nesse trabalho com a finalidade de se realizar buscas por expressões lingüísticas que aparecem em orações concentradoras de termos para serem utilizadas na abordagem lingüística. Esse cópús, composto por textos do gênero científico, domínio da Ecologia, Botânica, Biologia Geral, Zootecnia, Recursos Florestais e Engenharia Florestal, apresenta um total de 260.921 palavras. Os textos foram extraídos de partes dos livros “A Economia da natureza” e “Ecologia”, da editora Guanabara Koogan e de revistas presentes na primeira disponibilização pública do Projeto Lácio-Web⁴⁷. Esse cópús, desenvolvido no NILC, foi construído para o Projeto Bloc-Eco⁴⁸ e apresenta uma lista de referência com 96 termos.

⁴⁷ <http://www.nilc.icmc.usp.br/lacioweb/index.htm>

⁴⁸ <http://nilc.icmc.usp.br/nilc/projects/bloc-eco.htm>

Capítulo 5

Implementação e avaliação de quatro métodos estatísticos

Esse capítulo tem como finalidade relatar os procedimentos utilizados para eleger a medida estatística mais eficiente para a extração automática de termos, tomando-se como base o *córpus* de Revestimentos Cerâmicos criado nessa pesquisa. Entende-se como mais eficiente a medida estatística que apresentar a maior precisão, embora tenham sido calculadas a revocação e a medida F. As medidas estatísticas utilizadas nesse trabalho são quatro: Frequência, *Log-likelihood*, Informação Mútua e Coeficiente *Dice*, implementadas no pacote para a extração de n-gramas NSP⁴⁹ (N-gram Statistics Package) apresentado na Seção 4.6. Além da eleição de uma medida estatística para a extração automática de unigramas, bigramas e trigramas (termos que apresentam, respectivamente, o número de *tokens* igual a 1, 2 e 3), nesse capítulo também é apresentado um método de extração de termos que depende fortemente do esforço humano, garantindo, dessa forma, uma melhoria nos resultados produzidos pelas medidas estatísticas aplicadas sobre o *córpus* alvo, visto que é produzida uma lista de candidatos mais próxima da lista real de termos do domínio escolhido.

5.1 Geração das listas de unigramas, bigramas e trigramas

Dentre as medidas de associação de palavras encontradas no pacote NSP, foram utilizados a Informação Mútua, o *Log-likelihood*, e o Coeficiente *Dice*, bem como a Frequência para realizar o levantamento dos candidatos a termos no *córpus* alvo. A Frequência pode ser calculada para n-gramas, e, nesse trabalho, esse n está limitado aos valores 1, 2 e 3 (unigramas, bigramas e trigramas) para serem comparados com a lista de termos de referência.

Os comandos utilizados para a produção das listas de n-gramas, considerando que n assume os valores 1, 2 e 3, são apresentados a seguir.

```
count.pl --ngram 1 --token tk3.pl --stop FILEOR.pl --remove 5 --recurse unigrama.txt corpus
```

para a geração da lista de unigramas (arquivo unigrama.txt), a partir do diretório corpus, utilizando-se a regra de formação de *tokens* (arquivo tk3.pl) e a *stoplist* (arquivo FILEOR.pl), eliminando os unigramas com frequência inferior a 5 (--remove 5), pois o *córpus* apresenta 448.352 palavras.

```
count.pl --token tk3.pl --stop FILEOR.pl --remove 5 --recurse bigrama.txt corpus
```

⁴⁹ <http://www.d.umn.edu/~tpederse/nsp.html>

para a geração da lista de bigramas, utilizando-se a regra de formação de *tokens* e a *stoplist*, e eliminado os bigramas com frequência inferior a 5 (--remove 5)

```
count.pl --ngram 3 --token tk3.pl --stop FILEAND.pl --remove 5 --recurse trigram.txt corpus
```

para a geração da lista de trigramas, utilizando-se a regra de formação de *tokens* e a *stoplist*, e eliminando os trigramas com frequência inferior a 5 (--remove 5)

5.2 Os métodos estatísticos implementados

Após a geração das listas de frequência para unigramas, bigramas e trigramas, foram realizados os cálculos da informação mútua, do *log-likelihood* e do coeficiente *dice* para bigramas, que utilizam como entrada a lista de frequência gerada para os bigramas do corpus. Em seguida, foram realizados os cálculos da informação mútua e do *log-likelihood* para trigramas, que utilizam como entrada a lista de frequência gerada para os trigramas encontrados no corpus. Para unigramas somente foi realizado o cálculo da frequência, pois é a única medida para unigramas disponível no pacote NSP. Tendo realizado o cálculo de frequência para unigramas, bigramas e trigramas, e o cálculo das medidas utilizadas para bigramas e trigramas, a lista gerada por cada um deles foi dividida em oito classes, sendo que em cada classe foi especificada a quantidade de termos, obtida com a intersecção das palavras da classe com a lista de referência. Para cada uma dessas medidas foi construído um histograma (Figuras 5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7 e 5.8) para uma melhor visualização da distribuição dos termos pelas classes.

A idéia de divisão das listas de unigramas, bigramas e trigramas em várias classes foi retirada do trabalho de Daille (1994; 1996).

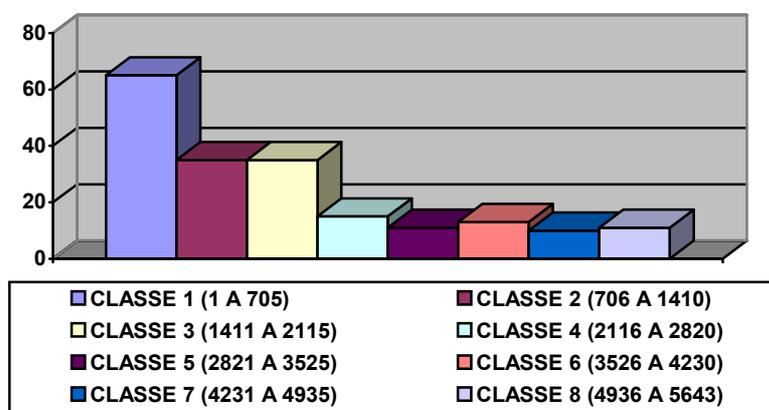


Figura 5.1 - Classes de Palavras para Unigramas – Frequência.

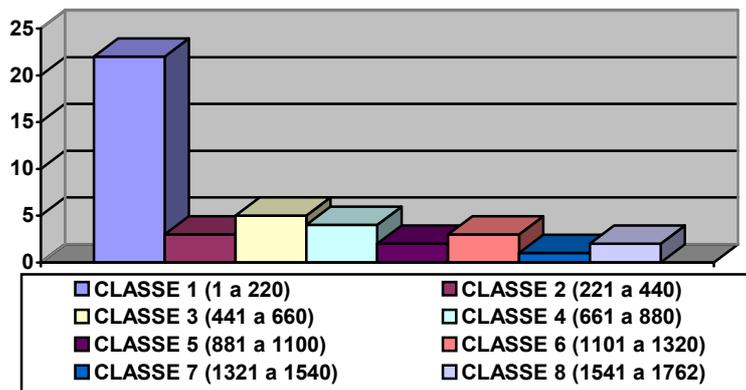


Figura 5.2 - Classes de Palavras para Bigramas – Frequência.

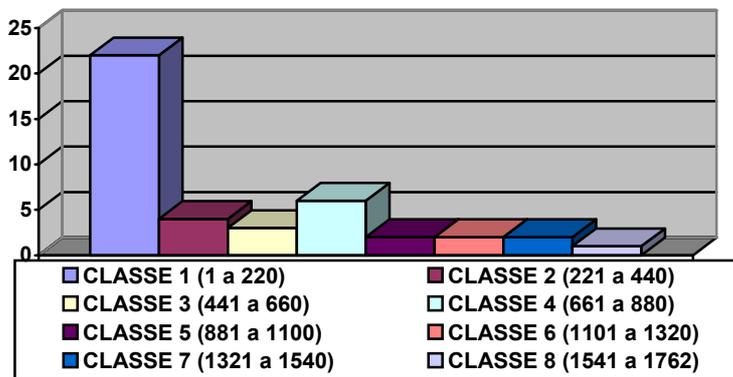


Figura 5.3 - Classes de Palavras para Bigramas – Informação Mútua.

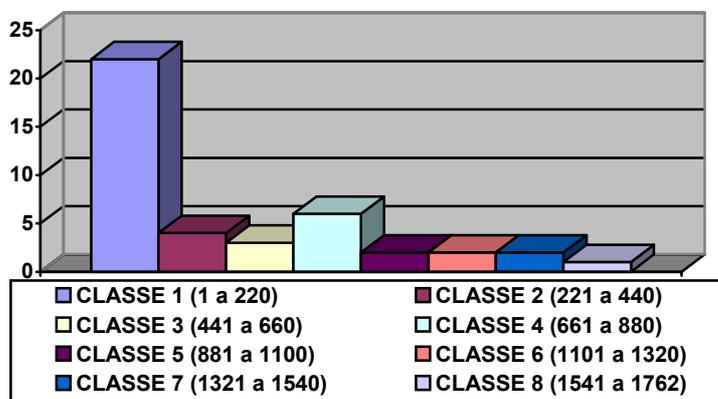


Figura 5.4 - Classes de Palavras para Bigramas – Log-likelihood.

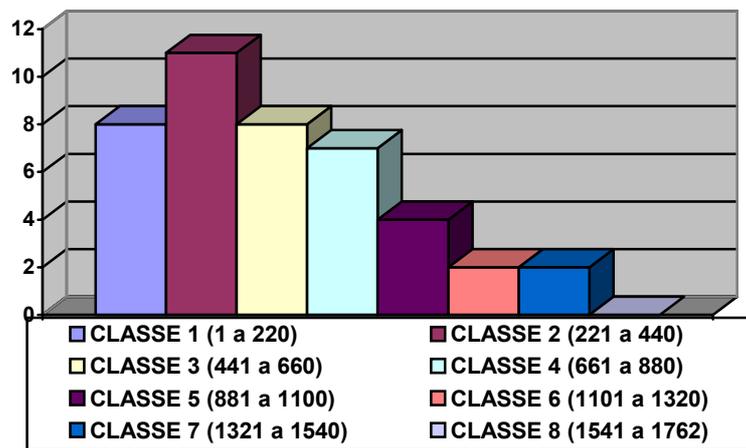


Figura 5.5 - Classes de Palavras para Bigramas – Coeficiente Dice.

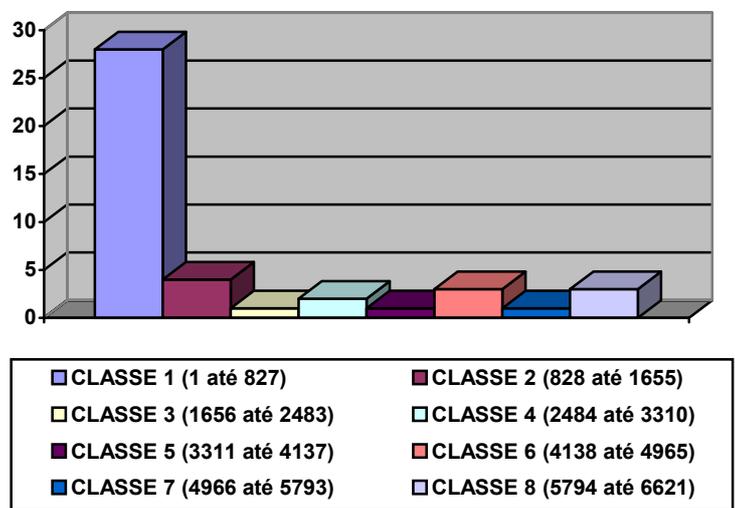


Figura 5.6 - Classes de Palavras para Trigramas – Frequência.

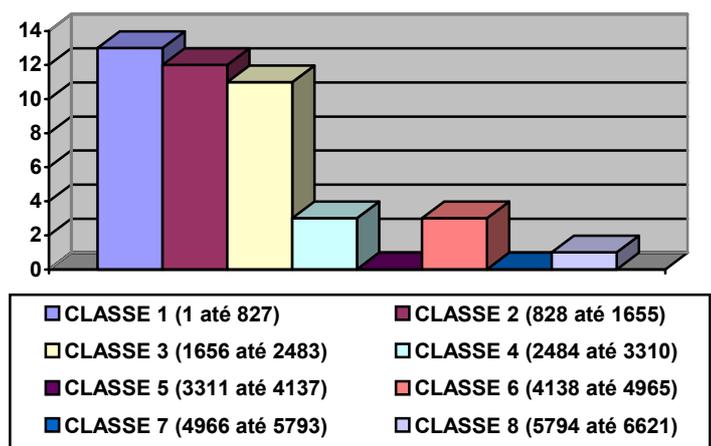


Figura 5.7 - Classes de Palavras para Trigramas – Informação Mútua.

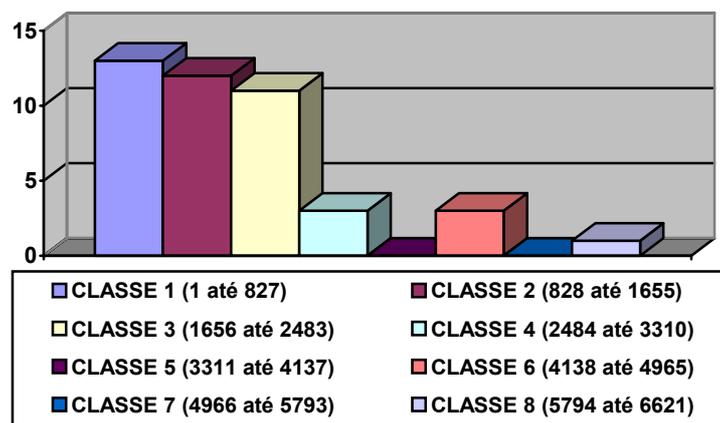


Figura 5.8 - Classes de Palavras para Trigramas – *Log-likelihood*.

Cada histograma foi analisado, buscando determinar em quais classes havia grande concentração de termos, para então poder especificar um corte na frequência ou no escore baseando-se na frequência ou no escore do último termo da classe que apresentou uma quantidade considerável de termos, eliminando, dessa forma, aquelas classes caracterizadas por um número insignificante de termos. Essa foi a abordagem considerada para a criação dos vários métodos estatísticos apresentados a seguir.

5.2.1 Método estatístico para unigramas

Para unigramas o corte foi realizado na frequência 20, ou seja, palavras com frequência inferior a 20 foram excluídas. Portanto, foram considerados 136 termos, que é o número de termos acumulados até a classe 3 (veja Figura 5.1), incluindo 1 termo da classe 4, pois foi a classe que apresentou o último termo com frequência 20. O total de palavras considerado foi 2128, que corresponde ao número da última palavra que apresentou frequência 20. A Precisão, a Revocação e a Medida F, calculadas para a frequência de unigramas, foram respectivamente 6%, 51% e 10%. Na Figura 5.9 é apresentado o método estatístico para unigramas, baseado na frequência.

1. Geração da lista de frequência para unigramas
2. Corte na frequência 20

Figura 5.9 - Método Estatístico para Unigramas – Frequência.

5.2.2 Métodos estatísticos para bigramas

Na lista de frequência para bigramas o corte foi realizado na frequência 18. Dessa forma, foram considerados 22 termos, que correspondem ao total de termos apresentado na classe 1 (veja Figura 5.2), classe que apresentou o último termo com frequência 18. O total de palavras considerado foi 218, número correspondente à última palavra com frequência 18. A Precisão, a Revocação e a

Medida F, calculadas para a frequência de bigramas, foram respectivamente 10%, 29% e 14%. Na Figura 5.10 é apresentado o método estatístico para bigramas, baseado na frequência.

1. Geração da lista de frequência para bigramas
2. Corte na frequência 18

Figura 5.10 - Método Estatístico para Bigramas – Frequência.

Para a informação mútua, calculada para bigramas, o corte foi realizado no escore 0.0066, que é o escore do último termo que apareceu na classe 1 (veja Figura 5.3). O número de termos considerados foi 22, que é o total de termos da primeira classe, e a quantidade de palavras considerada foi 190, que é o número que corresponde ao último termo da classe 1. A Precisão, a Revocação e a Medida F, calculadas para a informação mútua para bigramas, foram respectivamente 11%, 29% e 15%. Na Figura 5.11 é apresentado o método estatístico para bigramas, baseado na informação mútua.

1. Geração da lista de frequência para bigramas
2. Aplicação da medida informação mútua
3. Corte no escore 0.0066

Figura 5.11 - Método Estatístico para Bigramas – Informação Mútua.

Para o *log-likelihood*, calculado para bigramas, o número de termos e o total de palavras considerados, a Precisão, a Revocação e a Medida F calculadas foram exatamente os mesmos da informação mútua, com exceção do escore onde foi realizado o corte no *log-likelihood*, que foi equivalente a 187.8240. Na Figura 5.12 é apresentado o método estatístico para bigramas, baseado no *log-likelihood*.

1. Geração da lista de frequência para bigramas
2. Aplicação da medida *log-likelihood*
3. Corte no escore 187.8240

Figura 5.12 - Método Estatístico para Bigramas – *Log-likelihood*.

O corte para o coeficiente *dice* foi realizado no escore 0.2174, que corresponde ao escore do último termo da classe 4 (veja Figura 5.5). A quantidade de palavras é 879, que é o número do último termo da classe 4, e como o número de termos considerado é a soma dos termos das quatro primeiras classes, esse número equivale a 34. A Precisão, a Revocação e a Medida F, calculadas

para o coeficiente *dice* para bigramas, foram respectivamente 3%, 45% e 5%. Na Figura 5.13 é apresentado o método estatístico para bigramas, baseado no coeficiente *dice*.

1. Geração da lista de frequência para bigramas
2. Aplicação da medida coeficiente *dice*
3. Corte no escore 0.2174

Figura 5.13 - Método Estatístico para Bigramas – Coeficiente *Dice*.

5.2.3 Métodos estatísticos para trigramas

Na lista de frequência para trigramas o corte foi realizado na frequência 18, frequência do último termo da classe 1 (veja Figura 5.6). A última palavra que apresentou frequência 18 foi a de número 683, e o número de termos considerado foi 28, isto é, o total de termos da primeira classe. A Precisão, a Revocação e a Medida F, calculadas para a frequência de trigramas, foram respectivamente 4%, 65% e 7%. Na Figura 5.14 é apresentado o método estatístico para trigramas, baseado na frequência.

1. Geração da lista de frequência para trigramas
2. Corte na frequência 18

Figura 5.14 - Método Estatístico para Trigramas – Frequência.

Para a informação mútua, calculada para trigramas, o corte foi realizado no escore 0.0051, que é o escore do último termo da classe 3 (veja Figura 5.7), e o número de palavras, 2412, foi contado até esse termo. O total de termos considerado foi 36, que corresponde à soma dos termos das três primeiras classes. A Precisão, a Revocação e a Medida F, calculadas para a informação mútua aplicada aos trigramas, foram respectivamente 1%, 83% e 1%. Na Figura 5.15 é apresentado o método estatístico para trigramas, baseado na informação mútua.

1. Geração da lista de frequência para trigramas
2. Aplicação da medida informação mútua
3. Corte no escore 0.0051

Figura 5.15 - Método Estatístico para Trigramas – Informação Mútua.

O corte foi realizado no escore 483.1120 para o coeficiente *log-likelihood* calculado para trigramas, e a quantidade de palavras, 2427, corresponde ao número do termo que apresenta esse escore. O total de termos considerado é 36, que equivale à soma dos termos das três primeiras

classes (veja Figura 5.8). A Precisão, a Revocação e a Medida F, calculadas para o coeficiente *log-likelihood* aplicado aos trigramas, foram respectivamente 1%, 83% e 1%, as mesmas obtidas para a informação mútua. Na Figura 5.16 é apresentado o método estatístico para trigramas, baseado no *log-likelihood*.

1. Geração da lista de frequência para trigramas
2. Aplicação da medida *log-likelihood*
3. Corte no escore 483.1120

Figura 5.16 - Método Estatístico para Trigramas – *Log-likelihood*.

5.2.4 Discussão dos resultados dos métodos estatísticos

Com as precisões obtidas para as medidas aplicadas nos bigramas e trigramas do cópuz, a informação mútua e o *log-likelihood* foram as melhores medidas para bigramas, visto que apresentaram 11% de precisão, diferença de apenas 1% em relação à frequência, enquanto que a frequência foi a medida que apresentou um melhor desempenho no caso dos trigramas, com 4% de precisão.

Após o cálculo da Precisão, Revocação e Medida F para unigramas, bigramas e trigramas, foi possível notar que os valores apresentados para Precisão e Medida F apresentaram-se muito baixos quando comparados com o valor da medida F para a tarefa de extração, 60%, apresentado na Introdução. Por essa razão, decidiu-se analisar o conteúdo das listas geradas de unigramas e bigramas referentes ao cópuz desse trabalho a fim de verificar que tipo de *tokens* elas estavam trazendo. Na próxima seção apresenta-se um método de análise manual para limpeza das listas de unigramas e bigramas, com o intuito de eliminar *tokens* que não sejam candidatos a termos.

5.3 Levantamento de candidatos a termos

As listas de unigramas e bigramas foram divididas, separadamente, em tabelas de A a Z que continham respectivamente palavras que iniciavam em A a Z. Esse procedimento não foi realizado para trigramas.

Como as listas de unigramas e bigramas geradas pelo método estatístico trazem muitos elementos que não são interessantes para esse trabalho, como palavras e siglas da língua geral, marcas publicitárias, nomes próprios, e símbolos especiais, essas tabelas foram construídas, separando, dessa forma, os candidatos a termos e reduzindo a quantidade de informações a serem validadas pelo especialista.

Dentro de cada tabela, foram feitas classificações do tipo *palavras e siglas da língua geral, candidatos a termos, marcas publicitárias, nomes próprios, e símbolos especiais*, para facilitar as análises dos unigramas e bigramas presentes.

Essas análises permitem verificar que, tanto no caso de unigramas quanto de bigramas, as palavras da língua geral aparecem em maior número. Em seguida aparecem os candidatos a termos, depois os nomes próprios, marcas e siglas, e, finalmente, os símbolos especiais. Já era esperado que o número de palavras pertencentes à língua geral fosse maior, pois os programas utilizados do pacote NSP servem ao propósito primário de levantar colocações e muitas dessas pertencem à língua geral.

Dentre as palavras da língua geral, no que se referem aos unigramas, os substantivos e verbos são os mais freqüentes. Nos bigramas, as palavras se apresentam em sua maioria, com algum sentido, por exemplo, *aquecimento constante e alto grau*, em contraste com *orgânica aumentar*.

Com as separações por palavras e siglas da língua geral, candidatos a termos, marcas publicitárias, nomes próprios, e símbolos especiais, efetuadas nas tabelas, foi possível fazer um levantamento de uma lista de candidatos a termos, tanto para unigramas quanto para bigramas. A freqüência calculada anteriormente para unigramas e bigramas e as medidas informação mútua, *log-likelihood*, e coeficiente *dice*, calculadas anteriormente para bigramas, foram utilizadas para organizar os unigramas e bigramas candidatos especificados nas tabelas em ordem decrescente de seus escores. Dessa forma, a classificação dos unigramas candidatos foi dividida em classes de 10 e a classificação dos bigramas candidatos foi dividida em classes de 12. Para cada uma dessas classes foi realizada a intersecção com a lista de referência (de unigramas e bigramas separadamente), a fim de obter o total de termos encontrado em cada classe. Depois de levantado o número de termos presente em cada classe (de unigramas e bigramas), foi montado um histograma para a freqüência dos unigramas (sem a realização do corte na freqüência 5), construído a partir da divisão da quantidade de termos de cada classe por 10 (veja Figura 5.17).

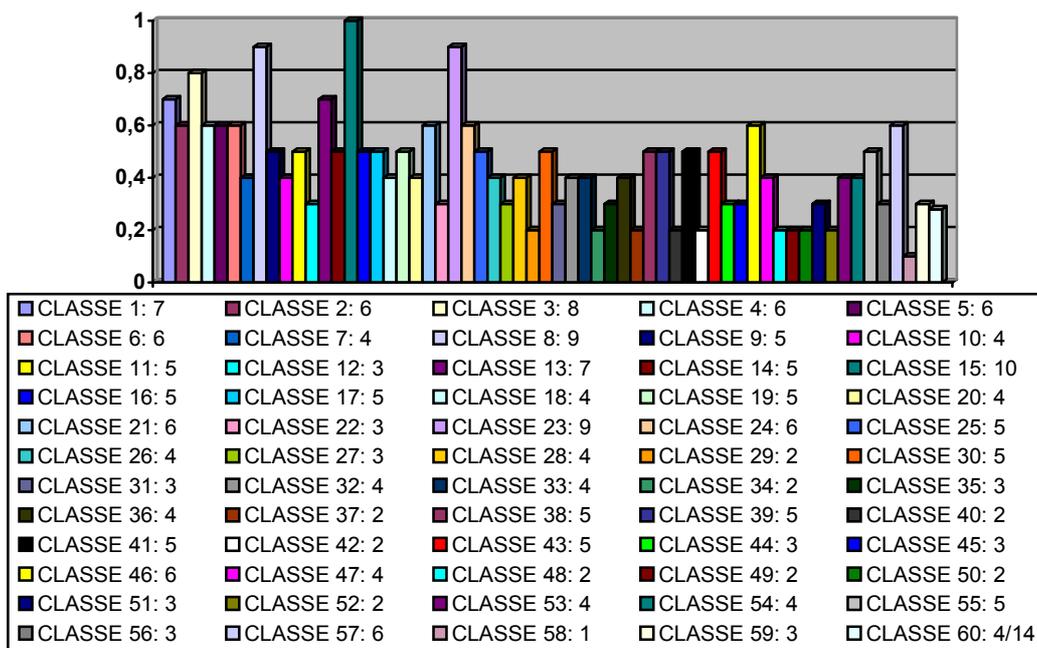
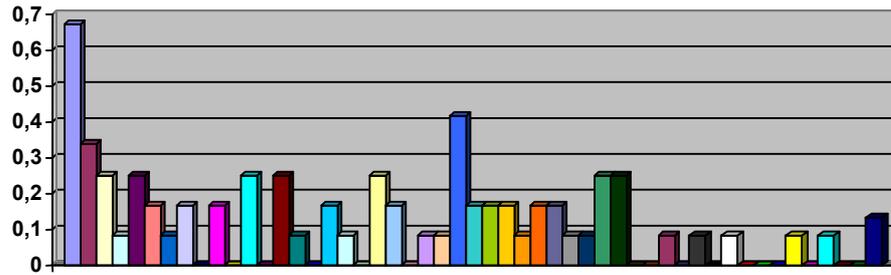


Figura 5.17 - Classes de Candidatos para Unigramas – Frequência.

Após a construção do histograma de frequência para unigramas, foi realizado um corte na frequência 5. Das 60 classes que existiam, restaram 41, sendo que a 41ª apresentou apenas 1 candidato com frequência igual a 5. O total obtido então foram 40 classes com 10 candidatos e uma com 1, ou seja, foram eliminados 203 candidatos, restando 401, e o total de termos recuperados até a 41ª classe foi 196. A Precisão, a Revocação e a Medida F calculadas para a frequência de unigramas foram respectivamente 48%, 74% e 58%.

Para bigramas também foi montado um histograma para a frequência (sem a realização do corte na frequência 5), construído a partir da divisão da quantidade de termos de cada classe por 12 (veja Figura 5.18). Em seguida, foi realizado um corte na frequência 5, fazendo com que, das 51 classes que existiam, restassem somente 23, excluindo 2 candidatos da 23ª classe que apresentaram frequência menor que 5. Dessa forma foram obtidas 22 classes com 12 candidatos e 1 com 10, ou seja, foram eliminados 341 candidatos, restando 274, dos quais 42 são termos. A Precisão, a Revocação e a Medida F calculadas para a frequência de bigramas foram respectivamente 15%, 56% e 23%.



CLASSE 1: 8	CLASSE 2: 4	CLASSE 3: 3	CLASSE 4: 1	CLASSE 5: 3	CLASSE 6: 2
CLASSE 7: 1	CLASSE 8: 2	CLASSE 9: 0	CLASSE 10: 2	CLASSE 11: 0	CLASSE 12: 3
CLASSE 13: 0	CLASSE 14: 3	CLASSE 15: 1	CLASSE 16: 0	CLASSE 17: 2	CLASSE 18: 1
CLASSE 19: 0	CLASSE 20: 3	CLASSE 21: 2	CLASSE 22: 0	CLASSE 23: 1	CLASSE 24: 1
CLASSE 25: 5	CLASSE 26: 2	CLASSE 27: 2	CLASSE 28: 2	CLASSE 29: 1	CLASSE 30: 2
CLASSE 31: 2	CLASSE 32: 1	CLASSE 33: 1	CLASSE 34: 3	CLASSE 35: 3	CLASSE 36: 0
CLASSE 37: 0	CLASSE 38: 1	CLASSE 39: 0	CLASSE 40: 1	CLASSE 41: 0	CLASSE 42: 1
CLASSE 43: 0	CLASSE 44: 0	CLASSE 45: 0	CLASSE 46: 1	CLASSE 47: 0	CLASSE 48: 1
CLASSE 49: 0	CLASSE 50: 0	CLASSE 51: 2			

Figura 5.18 - Classes de Candidatos para Bigramas - Frequência.

Para o cálculo da informação mútua para bigramas foi realizado um corte no escore 0.0001539, na classe 32, em razão das classes posteriores apresentarem uma quantidade de termos irrelevante (com exceção da classe 51). Das 51 classes que existiam, restaram 32, sendo que essa última apresentou apenas 5 candidatos. O total obtido então foram 31 classes com 12 candidatos e uma com 5, ou seja, foram eliminados 238 candidatos, restando 377, dos quais 60 são termos. A Precisão, a Revocação e a Medida F calculadas para a informação mútua (para bigramas) foram respectivamente 15%, 81% e 25%.

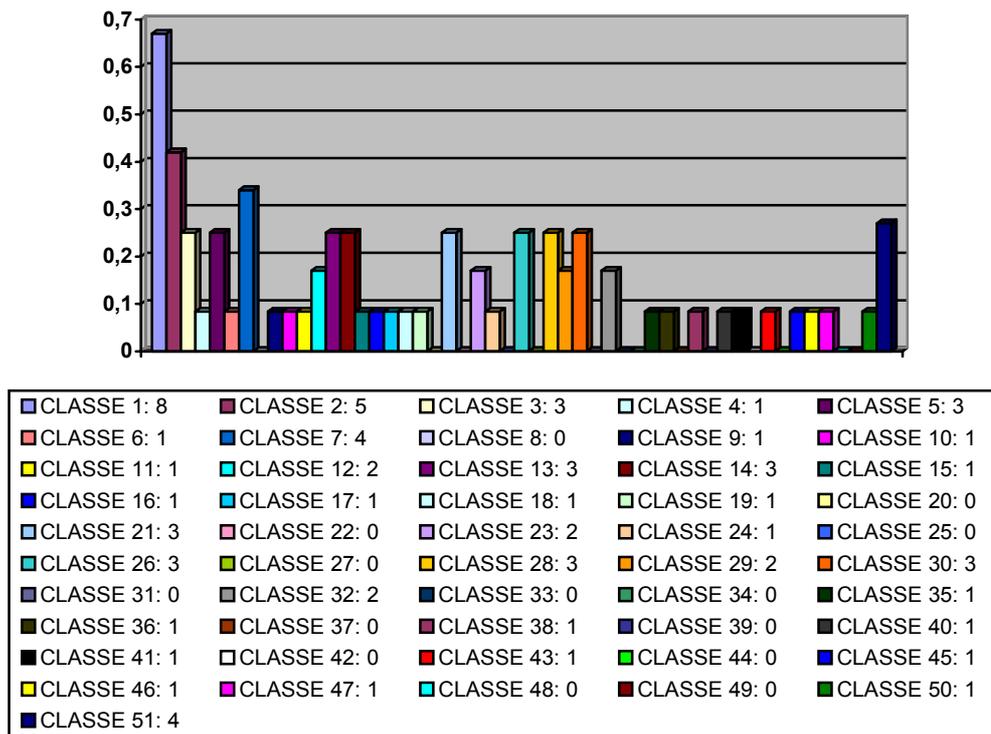


Figura 5.19 -

Classes de Candidatos para Bigramas – Informação Mútua.

O histograma produzido e os resultados obtidos para o coeficiente *log-likelihood* calculado para bigramas foram os mesmos obtidos para a informação mútua, com a diferença que o corte para o *log-likelihood* foi feito no escore 17.4304407.

Para o coeficiente *dice*, o corte foi realizado no escore 0.0595238, na classe 27 (veja Figura 5.20), em razão de um número considerável de classes posteriores não apresentarem termos. Das 51 classes que existiam, restaram 27, sendo que a 27ª apresentou somente 2 candidatos. O total obtido então foi 26 classes com 12 candidatos e uma com 2, ou seja, foram eliminados 301 candidatos, restando 314, dos quais 53 correspondem a termos. A Precisão, a Revocação e a Medida F calculadas para o coeficiente *dice* foram respectivamente 16%, 71% e 26%.

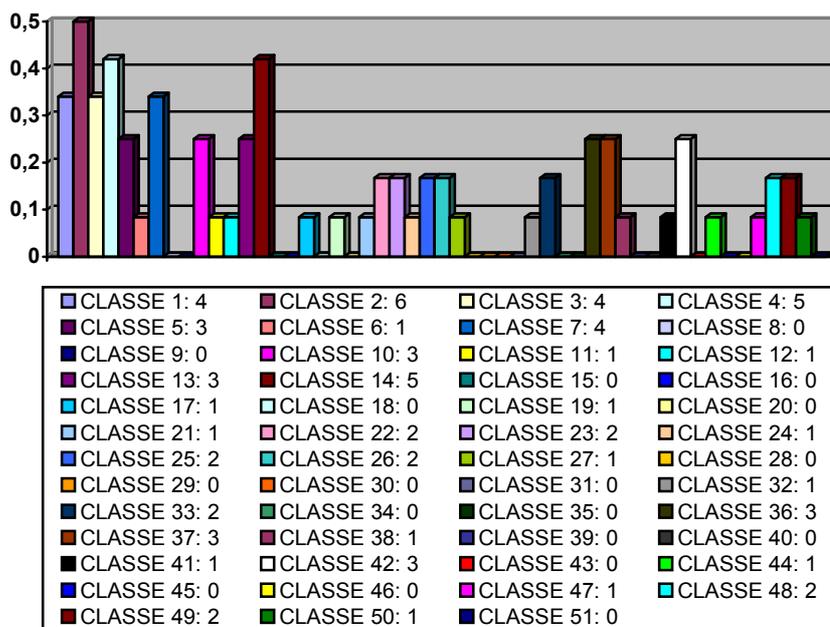


Figura 5.20 - Classes de Candidatos para Bigramas – Coeficiente *Dice*.

É fácil perceber que a divisão por classes realizada somente sobre candidatos a termos, tanto para unigramas quanto para bigramas, apresentou aumentos significativos na Precisão, Revocação e na Medida F, permitindo concluir que, a interferência humana (sem a necessidade de ser especialista do domínio) no levantamento de candidatos a termos é de suma importância para a obtenção de uma lista de candidatos a termos menor e mais precisa. Porém, o trabalho de levantamento de candidatos é bastante dispendioso, tendo durado 2 meses e sendo realizado por duas pessoas (uma lingüista e a própria pesquisadora desse trabalho).

Capítulo 6

Implementação e avaliação de um método lingüístico

Neste capítulo são apresentadas uma comparação e contraste do método lingüístico implementado com os trabalhos no qual se baseou. Na Seção 6.1 é apresentado um método para a extração de candidatos a termos que se baseia em expressões lingüísticas e indicadores estruturais, bem como nos padrões morfossintáticos dos termos do domínio de Revestimentos Cerâmicos. A Seção 6.2 tem como objetivo melhorar a precisão alcançada pelo método descrito na Seção 6.1. Considera cada expressão separadamente e, com a ajuda de um subcórpus identifica as palavras que acompanham tais expressões, definindo, dessa forma, a posição em que o termo aparece, para a criação de scripts específicos para cada uma delas. Para esse trabalho foi mostrado apenas o desempenho dos scripts construídos para cada expressão sobre o seu subcórpus com uma avaliação qualitativa deles.

6.1 O método lingüístico implementado e sua avaliação quantitativa

O método lingüístico implementado baseou-se tanto no trabalho de Heid et al (1996), no sentido de realizar um pré-processamento lingüístico no cópuz utilizado e posteriormente a realização de consultas sobre o mesmo, quanto no trabalho de Klavans e Muresan (2000; 2001a; 2001b), no sentido de realizar uma busca por expressões lingüísticas e indicadores estruturais que introduzem definições e os termos definidos. Esses dois trabalhos foram descritos detalhadamente no Capítulo 3. O trabalho em questão não se assemelha totalmente ao método proposto por Heid et al (1996) em razão do cópuz desse trabalho de mestrado não ter sofrido o processo de lematização. Por outro lado, o método aqui implementado fugiu um pouco da proposta feita por Klavans e Muresan (2000; 2001a; 2001b), em razão de não terem sido realizadas buscas somente de expressões de definições, mas também de classificações, descrições e outras que concentram termos, além de não ter sido utilizado um módulo de análise gramatical, responsável por identificar definições introduzidas por fenômenos lingüísticos mais complexos, tais como anáforas e apostos.

Baseando-se, a princípio, no trabalho de Klavans e Muresan (2000; 2001a; 2001b), foi realizado um levantamento de expressões lingüísticas e indicadores estruturais que geralmente vem acompanhados de definições, descrições, classificações e de outros tipos de orações que concentram termos, para identificar o termo ou termos que aparecem nelas.

Essas expressões e indicadores estruturais foram levantados de várias fontes: Aluísio (1995), Sager (1993), Klavans e Muresan (2000; 2001a; 2001b), ISO/TC 37/SC 1 Terminologia –

Princípios e Métodos (1996)⁵⁰, da lista conceitual do domínio de Revestimentos Cerâmicos do trabalho de Almeida (2000) e do córpus *CórpusEco*, com sua lista de referência, desenvolvido no Projeto Bloc-Eco.

O total de expressões e indicadores estruturais levantado foi 56, sendo dividido em três classes:

1. **Uso Geral (UG):** as expressões apresentadas nessa classe podem ser utilizadas em qualquer domínio de especialidade. Nela encontram-se expressões extraídas de Aluísio (1995), Sager (1993), Klavans e Muresan (2000; 2001a; 2001b), ISO/TC 37/SC 1 Terminologia – Princípios e Métodos (1996), e do *CórpusEco*. Para essa classe foram levantadas 43 expressões, sendo 35 implementadas;
2. **Conceitual do domínio de Revestimentos Cerâmicos (CD):** as expressões encontradas nessa classe podem ser aplicadas somente para córpus do domínio de Revestimentos Cerâmicos (domínio de pesquisa desse trabalho), já que foram levantadas da lista conceitual do domínio de revestimentos cerâmicos. Para essa classe foram selecionadas 10 expressões, das quais 7 foram implementadas;
3. **Sinais gráficos (SG):** nessa classe considera-se “()”, “:” e “-”, que foram levantados do trabalho de Klavans e Muresan (2000; 2001a; 2001b). Esses três indicadores estruturais foram implementados.

As expressões levantadas, acompanhadas de suas fontes são apresentadas na Figura 6.1. As que foram implementadas aparecem com uma indicação de sua classe na frente das expressões.

Aluísio (1995)		CórpusEco	
apresenta	(UG)	adição de	
atua		chamamos	(UG)
caracterizado	(UG)	constitui	(UG)
classe de	(UG)	constituído	(UG)
compreendendo		depende	(UG)
compreendido	(UG)	desenvolvido	(UG)
conhecido como	(UG)	determinado	(UG)
consiste	(UG)	empregado	(UG)
contém, contém	(UG)	expresso	(UG)
em outras palavras		formado	(UG)
implica	(UG)	obtido	(UG)
isto é	(UG)	palavra	(UG)
ou seja	(UG)	relacionado	(UG)
por exemplo	(UG)		
tal como	(UG)	Klavans e Muresan	

⁵⁰ <http://www.tc37sc4.org/index.html>

Sager (1993)		(2000;2001a;2001b)	
é	(UG)	chamado	(UG)
são	(UG)	definido como	(UG)
utilizado	(UG)	expressão	(UG)
		(se) entende	
		significa	
Almeida (2000)		termo	(UG)
característica do	(CD)	()	(SG)
composição do	(CD)	-	(SG)
composto	(CD)	:	(SG)
estado de			
matéria-prima	(CD)	ISO/TC 37/SC 1	
método		conceito	(UG)
parte de		corresponde	(UG)
processo	(CD)	define	
propriedade de	(CD)	denominado	(UG)
tipo de	(CD)	feito de	
		usado	(UG)

Figura 6.1 - Expressões lingüísticas no singular, masculino e marcadores estruturais utilizados.

Tendo apresentado as expressões (sem indicar suas variações de gênero e número) que aparecem acompanhadas de definições, descrições, classificações, ou simplesmente de termos, o outro recurso utilizado no método lingüístico pode ser apresentado: os padrões de termos. É importante notar que o cópuz foi pré-processado com o tokenizador e o etiquetador descritos nas Seções 4.3 e 4.4.

O uso do etiquetador se justifica pela busca por padrões de termos (Figura 6.2) que podem ser unigramas, bigramas e trigramas. A Figura 6.2 apresenta o padrão morfossintático acompanhado de um exemplo de termo da lista de referência que “casa” com ele.

TRIGRAMAS:	substantivo preposição(+artigo) substantivo -- absorção de água
	substantivo adjetivo adjetivo -- grês porcelanato esmaltado
	substantivo preposição adjetivo -- moagem a seco
BIGRAMAS:	substantivo adjetivo -- argila gorda
	substantivo substantivo -- nefelina sienito
	adjetivo substantivo -- ball clay
UNIGRAMAS:	substantivo -- massa
	nome próprio -- Rotocolor
	adjetivo -- vitro-cerâmico
	verbo -- deflocular

Figura 6.2 - Padrões da lista de referência.

Foi desenvolvido um programa genérico em PERL (um script) que toma como entradas um arquivo com as 45 expressões e marcadores estruturais da Figura 6.1 e outro com os padrões para unigramas, bigramas e trigramas mostrados na Figura 6.2, e apresenta como saída todos os candidatos a termos pertencentes às orações que casaram com os padrões de termos especificados no arquivo de entrada. A saída gerada primeiramente mostra o arquivo que está sendo pesquisado, verificando, dessa forma, se ele apresenta orações que contêm as características especificadas no script. Posteriormente são impressos a oração que apresenta a expressão buscada, a expressão buscada naquela oração, bem como os termos à esquerda e à direita da expressão, isto é, são apresentados os candidatos que casaram com os padrões de termos do arquivo de entrada. A impressão de todas essas informações serviu para avaliar a corretude do script. É importante elucidar que a grande maioria dos candidatos apresentados em TERMOS À ESQUERDA e em TERMOS À DIREITA (veja o Quadro 6.1) não corresponde realmente a termos, visto que o script está considerando como termos todas as palavras que apresentaram os padrões presentes no arquivo de entrada (padrões de termos) e, por essa razão, muito ruído é mostrado no arquivo de saída. Um exemplo de uma oração presente na saída desse script é mostrado no Quadro 6.1.

Quadro 6.1: Trecho da saída do método lingüístico para a expressão “constituído”

```
*****
ARQUIVO: c:/perl/analise/tercei~1/etique~1/nacionais/abordagem_industria.sent.txt.tagged.txt
*****
O_ART segmento_N de_PREP vidros_N é_VERB constituído_VERB por_PREP cerca_LOCU de_LOCU 30_NUME
grandes_ADJ empresas_N ,_, com_PREP predominância_N de_PREP capital_N estrangeiro_ADJ
EXPRESSÃO -> constituído_
TERMOS À ESQUERDA -> segmento_N de_PREP vidros_N / segmento_N / vidros_N / é_VERB /
TERMOS À DIREITA -> predominância_N de_PREP capital_N / capital_N estrangeiro_ADJ / grandes_ADJ
empresas_N / 30_N / empresas_N / predominância_N / capital_N / grandes_ADJ / estrangeiro_ADJ /
```

Observe que dentre todos os termos considerados pelo script, somente vidros (terceiro termo de TERMOS À ESQUERDA) realmente corresponde a um termo no domínio em questão.

A fim de avaliar quantitativamente esse método, esse script sofreu algumas alterações, de forma a produzir em sua saída somente os termos considerados por ele, excluindo, dessa forma, o arquivo pesquisado, a oração e a expressão extraídas, bem como os rótulos “EXPRESSÃO”, “TERMOS À ESQUERDA” e “TERMOS À DIREITA”.

Dois scripts foram criados para a avaliação quantitativa. O primeiro script, que teve como entradas a lista de padrões mostrada na Figura 6.2 e a lista das 45 expressões da Figura 6.1, produziu uma saída com todos os possíveis termos extraídos (unigramas, bigramas e trigramas que

casaram com os padrões encontrados na lista de referência), já com a transformação dos mesmos para minúsculo e com a eliminação de suas etiquetas. A princípio pretendia-se considerar todas as 56 expressões da Figura 6.1, porém, como existia a idéia de se implementar cada expressão separadamente, decidiu-se escolher algumas delas, dada a escassez de tempo. O número de possíveis termos extraídos pelo primeiro script foi 519.956, incluindo candidatos a termos repetidos. Esses possíveis termos extraídos correspondem àquelas palavras impressas em TERMOS À ESQUERDA e TERMOS À DIREITA pelo script descrito anteriormente. O segundo script tomou a saída do primeiro como entrada, e dessa forma, realizou algumas operações sobre a mesma. Primeiramente foram eliminados os candidatos a termos repetidos, encontrados na saída do primeiro script. Posteriormente, foi realizada a diferença entre a lista produzida com a eliminação dos candidatos a termos repetidos e a *stoplist* (Apêndice C), que também foi passada como um parâmetro de entrada para esse segundo script. Tendo realizado essa operação, a lista obtida foi dividida em três outras listas – uma contendo os unigramas, a outra os bigramas e a última contendo os trigramas. A Figura 6.3 mostra o método lingüístico implementado.

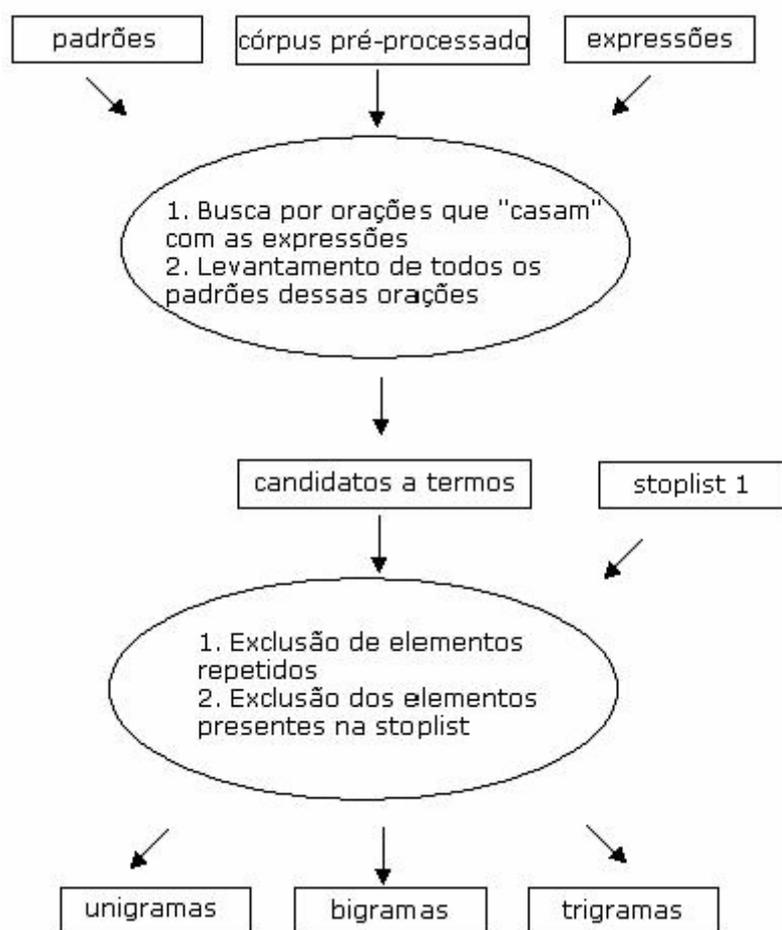


Figura 6.3 – O método lingüístico implementado.

Finalmente, as listas de unigramas, bigramas e trigramas foram interseccionadas com suas respectivas partes na lista de referência (Apêndice B), sendo os resultados de cada uma dessas operações impressos em arquivos separados.

Com a eliminação de candidatos a termos repetidos, obtidos a partir da saída do primeiro script, o número de candidatos a termos resultante passou a ser 48837 (redução de 90,60%). A diferença entre esse número e a *stoplist* produziu um total de 48594 candidatos a termos (diminuição de 0,49%), separados em três listas de 15932 unigramas, 15016 bigramas e 17646 trigramas, e a intersecção dessas listas com suas respectivas listas de referência resultou em 251 unigramas, 67 bigramas e 43 trigramas.

Tendo realizado essas operações, os resultados obtidos para Precisão, Revocação e Medida F, calculados para unigramas, bigramas e trigramas são apresentados no Quadro 6.2.

Quadro 6.2: Precisão, Revocação e Medida F do método lingüístico

<i>Unigramas</i>	
Precisão = $251/15932 = 0,01 = 1\%$	
Revocação = $251/264 = 0,95 = 95\%$	
Medida F = $0,019/0,96 = 0,01 = 1\%$	
Bigramas	
Precisão = $67/15016 = 0,004 = 0,4\%$	
Revocação = $67/74 = 0,90 = 90\%$	
Medida F = $0,0072/0,904 = 0,007 = 0,7\%$	
<i>Trigramas</i>	
Precisão = $43/17646 = 0,002 = 0,2\%$	
Revocação = $43/43 = 1 = 100\%$	
Medida F = $0,004/1,002 = 0,003 = 0,3\%$	

A revocação apresentou porcentagens elevadas, não apenas para unigramas, mas também para bigramas e trigramas, o que permite concluir que o método apresentou uma ótima cobertura, que corresponde à capacidade de extrair o maior número de termos corretos da lista de referência. Por outro lado, a precisão, que corresponde à consistência do método de extração, apresentou porcentagens insignificantes, permitindo concluir que grande parte das informações extraídas por esse método foi irrelevante.

É necessário esclarecer que o cópús não foi lematizado, influenciando, conseqüentemente, no desempenho do método, pois alguns termos são encontrados no cópús em sua forma flexionada quanto ao gênero e número, enquanto que os termos presentes na lista de referência se encontram

não flexionados, e dessa forma, quando é realizada a interseção entre os dois, aqueles termos que se encontram flexionados não são considerados. No entanto, a lematização do *córpus* melhoraria apenas o resultado da revocação.

6.2 Variações do método lingüístico e suas avaliações qualitativas

Para tentar aumentar a precisão do método lingüístico, foram criados sub*córpus* a partir do *córpus* original para preparar scripts dedicados a encontrar o lugar exato do termo nas orações que trazem expressões e indicadores estruturais de definições, classificações, descrições ou outros tipos concentradores de termos.

Assim, dentro de cada script foi especificada a expressão a ser buscada no *córpus*, considerando uma possível seqüência em que ela pode aparecer, baseando-se nas orações do sub*córpus* respectivo à expressão. Para cada script é produzida uma saída com as orações extraídas, e, para cada oração, são indicados o termo (ou termos) que ela apresenta e o arquivo de origem da mesma. Para cada expressão, foi utilizado o seu sub*córpus* de análise a fim de verificar a corretude do script.

É importante salientar que, nos 45 scripts implementados, foram especificados todos os possíveis padrões (veja Figura 6.2) encontrados na lista de referência, e esses padrões foram definidos nos scripts em ordem decrescente de tamanho, ou seja, primeiramente os trigramas, depois os bigramas e, por último, os unigramas. Durante a execução do programa, foi realizada uma varredura nesses padrões, visto que eles correspondem aos termos pertencentes à seqüência de etiquetas, para a extração de orações, presente em cada script, a fim de verificar se um desses padrões pertencia às orações analisadas pelo script, ou seja, para verificar se uma dada oração satisfazia a condição do programa, isto é, se ela apresentasse um determinado padrão dentro de uma seqüência de etiquetas definida baseando-se no sub*córpus* de cada expressão. Dessa forma, os padrões encontrados em uma oração extraída correspondem aos termos daquela oração.

A especificação de uma ordem de prioridade para os padrões prejudicou o desempenho dos scripts, pois se na oração em análise pelo script existir um trígama (na posição especificada pela lógica do programa), mas na verdade o termo corresponder a um bigrama ou unígama, o script considera como termo o trígama, pois ele apresenta maior prioridade na lista de padrões para termos, e depois que é encontrado um termo, ou seja, se houve um casamento com um padrão, o script o considera como sendo termo e pára a busca por termos. Um exemplo desse problema é encontrado na oração a seguir.

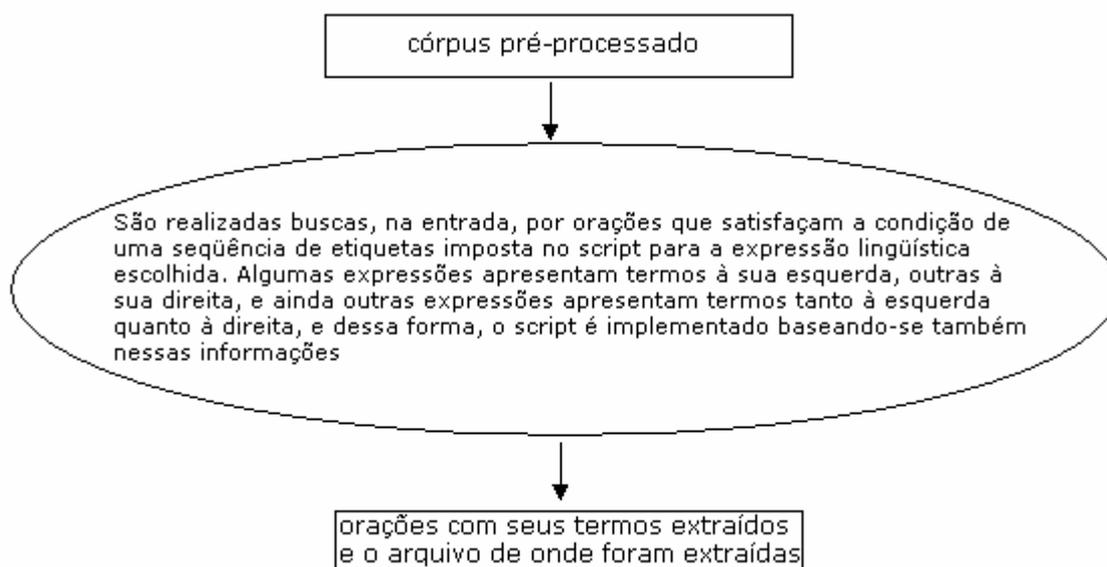
Com base nessas informações e na literatura sobre defeitos em esmaltes , pode-se concluir que os defeitos " furos no esmalte " estudados estão diretamente ligados à presença de carbonato de cálcio nas argilas empregadas na massa .

Termo -> cálcio nas argilas

Arquivo: c:/perl/analise/sent_c~1/nacionais/v6n1_1.sent.txt

O termo que deveria ser extraído é argilas (unigrama), porém, como o script busca primeiramente por trigramas na lista de padrões, ele casa a seqüência “cálcio nas argilas” com o padrão “substantivo preposição+artigo substantivo”, e dessa forma, considera tal seqüência como o termo na oração.

Uma forma de possibilitar um melhor entendimento dos scripts implementados é traduzindo-os em um único algoritmo:



A expressão “obtid” será utilizada para exemplificar esse algoritmo.

Para escrever o script para as expressões “obtido/obtida/obtidos/obtidas” foram retiradas do córpus 23 orações contendo a expressão "obtid". Porém, foram utilizadas 22 orações para a escrita do script da expressão “obtid”, pois uma das orações não apresentou uma forma padronizada de repetição de termos que pudesse ser incluída na seqüência de termos generalizada, sendo, dessa forma, eliminada.

Apesar de ter sido construído um subcórpus composto somente por 22 orações, é importante ressaltar que todo o córpus contém 658 orações que apresentam a expressão "obtid". Em todas as 22 orações os termos aparecem do lado esquerdo, portanto o script resultante é:

```
(artigo? termo(( , artigo? termo)*(( , )?(( e|ou ) | e / ou ) artigo? termo))?)(( advérbio )?(verbo verbo ))|( preposição )? verbo )( : )|( advérbio )|( )obtid(o)?(a)?(s)?_
```

onde

? indica que uma certa palavra ou caractere, ou uma seqüência de palavras e caracteres podem aparecer ou não em uma determinada oração;

* indica que uma seqüência de caracteres poderá aparecer zero ou mais vezes em uma determinada oração;

| indica o conhecido “ou”, que fornece a opção de aparecer alternativamente uma palavra ou outra na oração;

termo indica padrões para unigramas, bigramas ou trigramas (Figura 6.2).

Considerando o esquema acima escrito na forma de uma expressão regular, no caso em que o artigo aparece acompanhado de “?” significa que o artigo pode aparecer ou não na oração. Já para a seqüência “(, artigo? termo)*” o caractere “*” tem a função de permitir que a seqüência delimitada por ele (que se encontra dentro dos parênteses) apareça zero ou mais vezes na oração. Para o caso da seqüência “e|ou”, o caractere “|” faz com que somente a palavra “e” ou somente a palavra “ou” apareça na oração. No esquema de repetição também aparece a seqüência “e / ou”, indicando que tal seqüência pode ser encontrada nas orações do córpus. Um exemplo dessa seqüência é : “Calcitas e / ou dolomitas : são matérias-primas...”. O “e” fica separado do “/” e do “ou” pelo fato de o córpus estar tokenizado.

Um exemplo de oração produzida pelo script implementado para a expressão “obtid” é mostrado na Figura 6.4.

Na tabela 1 , são apresentadas as distribuições granulométricas <u>obtidas após moagem das 4 argilas , massa padrão</u> . Termo -> distribuições granulométricas Arquivo: c:/perl/analise/sent_c~1/nacionais/variacao_prop.sent.txt
--

Figura 6.4 - Saída do script para “obtid”.

A expressão “obtid” só contempla a busca por termos do lado esquerdo, embora possa existir em todo o córpus ocorrências de "obtid" com termos do lado direito. Essas ocorrências não serão recuperadas.

6.2.1 Experimento 1: avaliando o etiquetador utilizado

A fim de avaliar a precisão do tagger aplicado ao córpus de Cerâmica Industrial, foi escolhido um artigo do córpus (v6n1_5.txt) de 238 palavras, sendo 137 não repetidas. Para o mesmo artigo foram realizados 2 tipos de etiquetações com o mesmo conjunto de etiquetas: manual e automática (utilizando-se o etiquetador MXPOST descrito na Seção 4.4).

Com a etiquetagem manual foi possível encontrar 12 etiquetagens efetuadas de forma errônea pelo tagger, obtendo uma porcentagem de erro de 5%, ou seja, uma precisão de 95%, que se aproxima da precisão geral do tagger (97%) quando calculada no mesmo corpus em que o tagger foi treinado. Entretanto, 2% de diferença importa bastante para essa tarefa básica, pois pode causar erros severos em processamentos que dependem de uma alta precisão.

Em razão dessa taxa de erros produzidos pelo etiquetador no corpus de pesquisa desse mestrado, decidiu-se realizar um experimento com 10 dos 45 scripts produzidos para extração de termos, a fim de avaliar a influência que erros produzidos pelo etiquetador podem apresentar no resultado da extração de termos.

O primeiro experimento proposto consistiu na comparação dos resultados produzidos pelos scripts considerando a expressão em questão com e sem a especificação da etiqueta, já que todo o corpus foi anteriormente etiquetado.

A Tabela 6.1 apresenta as dez expressões escolhidas, comparando os resultados que os scripts elaborados para elas apresentaram com e sem etiqueta.

Tabela 6.1: Resultados do primeiro experimento

	Expressão	Total de orações a serem geradas	Orações extraídas	Orações extraídas com o termo correto
Com etiqueta	denominado	21	16	13
Sem etiqueta			21	15
Com etiqueta	desenvolvido	7	5	1
Sem etiqueta			7	2
Com etiqueta	determinado	20	7	4
Sem etiqueta			20	10
Com etiqueta	empregado	36	10	4
Sem etiqueta			36	17
Com etiqueta	expresso	8	7	5
Sem etiqueta			8	6
Com etiqueta	formado	16	5	4
Sem etiqueta			16	14
Com etiqueta	obtido	23	22	17
Sem etiqueta			23	17
Com etiqueta	ou seja	4	2	2
Sem etiqueta			4	3
Com etiqueta	relacionado	13	12	10
Sem etiqueta			13	10
Com etiqueta	tal como	9	5	1
Sem etiqueta			9	2

A primeira coluna da Tabela 6.1 indica se a etiqueta para a expressão (de definição, descrição ou classificação), apresentada na coluna 2, foi ou não especificada em seu script. Na

coluna seguinte são listadas as dez expressões escolhidas dentre as 45. A coluna 3 mostra o total de orações (subcórpus de análise) que deveria ser produzido pelos scripts, e, na coluna 4, o número de orações que os scripts realmente conseguiram extrair. Na quinta coluna é mostrada a quantidade de orações em que o script consegue extrair corretamente o termo relacionado.

Às expressões “denominado”, “determinado”, “empregado” e “formado” foi atribuída a etiqueta “verbo” no script. Porém, no cópulus, essas expressões algumas vezes aparecem etiquetadas como “adjetivo”, às vezes como “substantivo”.

A seguir são mostrados alguns exemplos de orações extraídas pelos scripts, e, para cada uma delas, são apresentados o termo considerado (correto ou não) e o arquivo na qual ela se encontra.

Oração 1:

Os **feldspatos potássicos**, denominados ortoclási e microclina, são os minerais dominantes nos granitos e sienitos, bem como em rochas de grana fina equivalentes, como os riolitos e traquitos. termo -> feldspatos potássicos
Arquivo: c:/perl/analise/sent_c~1/nacionais/pilhas_homogenev2.sent.txt

Esta primeira oração foi produzida pelo script construído para a expressão “denominado” com suas variações gramaticais de gênero e número. Considerando que a expressão *denominados* aparece etiquetada como verbo no seu script, essa oração não foi extraída, pois, no cópulus, a expressão *denominados* está etiquetada como adjetivo. Porém, quando não é considerada a etiqueta para *denominados*, a oração é, naturalmente, extraída.

Oração 2:

Onde, pa é a porosidade aberta, da e aa são a **densidade aparente** e a **absorção de água** determinadas pelo método baseado no princípio de Arquimedes, e r é a densidade real do compósito, calculada a partir da densidade real de cada uma das fases cristalinas adicionadas (determinadas por picnometria de gás hélio) e aplicando-se a regra linear das misturas segundo a fração conhecida de cada fase neste compósito. termos -> densidade aparente / absorção de água
Arquivo: c:/perl/analise/sent_c~1/nacionais/v5n3_3.sent.txt

Esta segunda oração foi gerada pelo script implementado para a expressão “determinado” com suas variações gramaticais de gênero e número. Para esse exemplo ocorreu o mesmo problema descrito no exemplo anterior. No cópulus, a expressão *determinadas* apresenta a etiqueta adjetivo, diferentemente do que ocorre no script, considerando que *determinadas* está inserida na classe gramatical verbo.

Oração 3:

A qualidade de que necessitam as **fritas** atualmente **empregadas nos novos processos de fabricação de azulejos** reduziu ainda mais os intervalos de variação admissíveis , tanto com respeito às matérias-primas , como com respeito às variáveis de operação do processo .

termo -> fritas

Arquivo: c:/perl/analise/sent_c~1/estrangeiros/v6n6_3.sent.txt

Para a oração 3, ocorreu o mesmo problema descrito nos dois exemplos anteriores.

Oração 4:

Este efeito está relacionado às deformações da rede e **fissuras** **formadas por alterações de volume no ponto de transição** ; estas deformações da rede e fissuras ocorrem extensivamente no quartzo , por exemplo , no qual a mudança de volume é grande .

termo -> fissuras

Arquivo: c:/perl/analise/sent_c~1/nacionais/PNOVAESv2.sent.txt

A oração 4 também apresentou as mesmas características abordadas nos exemplos anteriores.

Oração 5:

Assim , o termo “ **variação de tonalidade** ” , amplamente **empregado nos meios cerâmicos** , é um termo mal utilizado , visto que não engloba todos os tipo de variação associados à cor de um objeto , que podem ser detectadas pelo olho humano .

termo -> amplamente

Arquivo: c:/perl/analise/sent_c~1/nacionais/cores_tonal.sent.txt

Para ilustrar mais uma falha do etiquetador utilizado, a oração 5 apresenta um caso em que o termo considerado não é o termo real. A princípio, essa oração não havia sido extraída, pois no script a expressão “empregado” é etiquetada como verbo, e, no entanto, no cópús essa expressão apresenta substantivo como etiqueta. Sem especificar a etiqueta no script, a oração passa a ser extraída. Porém, o termo considerado não corresponde realmente ao verdadeiro termo da oração. Este erro se deve ao fato de “amplamente” ter sido etiquetado no cópús como adjetivo, e o script escrito para a expressão “empregado” apresenta a opção da ocorrência de um advérbio entre o termo e a expressão, mas não um adjetivo, como acontece nesse caso. Como um dos padrões para unigramas é ADJETIVO, o suposto termo “amplamente” foi extraído.

6.2.2 Experimento 2: extraindo listas de termos

O segundo experimento projetado serviu para avaliar a extração de listas de termos nos scripts em que o subcópús de análise trouxesse tal característica. Tal experimento exigiu a troca, nos scripts, de uma variável correspondente a um termo por um esquema de repetição de variáveis que permite

uma seqüência de termos, generalizando, dessa forma, um padrão de seqüência de termos. Um exemplo de seqüência de termos é: “Os vidrados, as granilhas e os grãos são...”. O esquema de repetição de variáveis utilizado nos scripts foi:

artigo? termo((, artigo? termo)*((,)?((e|ou) | e / ou) artigo? termo))?

Esse esquema de repetição de variáveis foi implementado em 20 scripts, pois dentre as orações escolhidas para cada uma das 20 expressões havia ao menos um caso de seqüência de termos. Essa reescrita dos scripts foi proposta com o intuito de padronizar os 20 scripts com o esquema de repetição de variáveis correspondentes a termos.

As expressões que apresentaram essa seqüência de termos foram: apresenta(m), composiç(ão)(ões) d(o)(a)(s), constitu(i)(em), constituíd(o)(a)(s), denominad(o)(a)(s), desenvolvid(o)(a)(s), determinad(o)(a)(s), : (dois pontos), é, empregad(o)(a)(s), matéria(s)-prima(s), () (parênteses), processo, relacionad(o)(a)(s), são, ta(l)(is) como, tipo(s) de, - (traço), usad(o)(a)(s), utilizad(o)(a)(s). Para cada uma dessas expressões foi realizada a reescrita dos seus scripts.

Uma tela, retirada do AntConc, apresentando a saída de “composiç(ão)(ões) d(o)(a)(s)”, é apresentada na Figura 6.5.

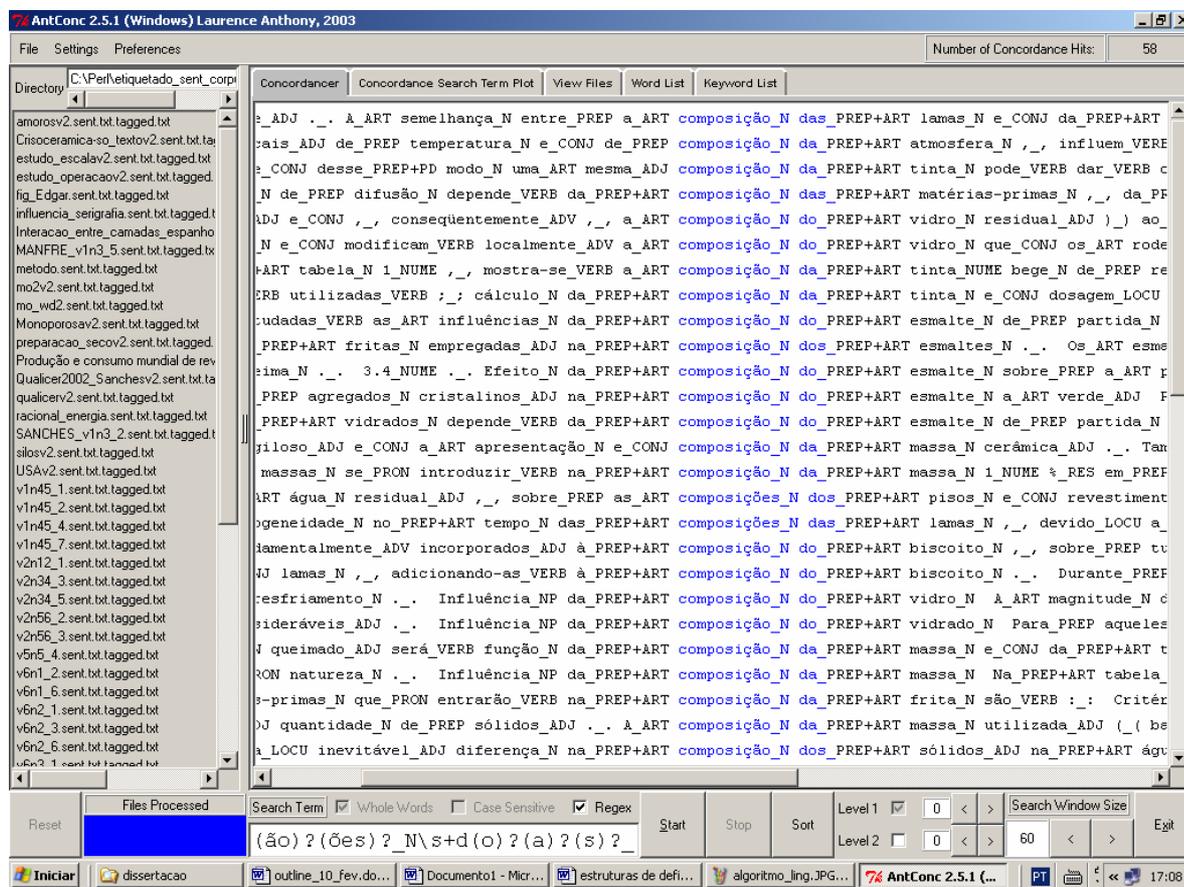


Figura 6.5 - Tela com as concordâncias da expressão “composiç(ão)(ões) d(o)(a)(s)”.

Note que a opção selecionada no concordanceador AntConc foi “Regex”, permitindo que a expressão seja procurada na forma de uma expressão regular, ou seja, a busca, para o exemplo dado na Figura 6.5, foi realizada utilizando-se a expressão regular “\s+composiç(ão)?(ões)?_N\s+d(o)?(a)?(s)?_”, onde “\s+” indica um ou mais espaços, “?” indica que os caracteres circundados por ele são opcionais e “_N” indica que a palavra foi etiquetada como substantivo.

Telas com exemplos das expressões “são” e “denominad” no cópús dessa pesquisa são mostradas no Apêndice A.

Tabela 6.2: Resultados do segundo experimento

Expressão	Total de orações a serem geradas	Orações extraídas	Orações extraídas com o termo correto
denominado	20	20	15
desenvolvido	7	7	3
determinado	20	20	10
empregado	35	35	17
relacionado	13	13	10
tal como	8	8	2

Na Tabela 6.2 são apresentados os resultados produzidos pelas seis expressões escolhidas para avaliar a corretude dos scripts, reescritos no segundo experimento, sobre seus respectivos subcórpus. É importante deixar claro que essas seis expressões foram somente as expressões utilizadas no primeiro experimento que apresentaram seqüência de termos em suas orações.

Analisando o resultado produzido por cada um dos scripts, considerando o mesmo subcórpus do primeiro experimento, foi possível perceber que o resultado obtido nesse segundo experimento foi bastante semelhante, ou seja, os termos extraídos foram, na maioria, os mesmos, e, as diferenças percebidas consistem em informações adicionais que, em alguns casos não foram produtivas (informações não correspondentes a termos).

Porém, na maioria dos casos, as seqüências de termos encontradas nas orações, antes não padronizadas para todos os scripts, continuam sendo corretamente extraídas.

Para a oração 1, os termos extraídos pelo script, antes dele sofrer a reescrita do segundo experimento, foram caulins, argilas, feldspatos e quartzo, que correspondem aos termos reais.

Oração 1:

A proporção entre cada componente deve ser testada caso a caso, pois as características particulares das outras [matérias-primas](#): **caulins, argilas, feldspatos, quartzo** e etc mudam esta relação entre adição de alumina e aumento do teor de frita.

Termos -> caulins / argilas / feldspatos / quartzo / etc

Arquivo: c:/perl/analise/sent_c~1/nacionais/reformul_engobes.sent.txt

Depois de realizada a reescrita, o script passou a extrair como termos não somente caulins, argilas, feldspatos e quartzo, mas também etc, que foi considerado como tal em razão de apresentar padrão de um termo (substantivo) e aparecer depois da conjunção “e”, casando, dessa forma, com o esquema de repetição de termos, ou seja, o esquema aceita a seqüência “termo, termo, termo, termo e termo”.

A oração 1 mostra um erro gramatical presente no cópulus, que consiste na ocorrência dessa abreviatura (“etc”) após uma conjunção. A abreviatura “etc” deve aparecer após uma vírgula, depois de uma repetição de palavras, por exemplo: caulins, argilas, feldspatos, quartzo, etc.

Porém, tendo sido realizada a reescrita, os scripts também extraem uma seqüência verdadeira de termos, assim como pode ser observado na oração 2, em que os termos extraídos, filito e talco, correspondem aos termos reais.

Oração 2:

Dentre as matérias-primas não-plásticas destacam-se os filitos , fundentes feldspáticos (feldspato , granito , sienito etc) , talco e carbonatos (calcário e dolomito) , sendo que o **filito** e o **talco** apresentam também características plásticas .

Termos -> filito / talco

Arquivo: c:/perl/analise/sent_c~1/nacionais/panorama.sent.txt

Um exemplo em que houve a interferência do etiquetador no resultado dos termos produzidos encontra-se na oração 3.

Oração 3:

1 -> Neste trabalho foi aplicada a teoria de Jenike no desenho de silos para três tipos de materiais na forma de pó : massa atomizada , empregada na fabricação de **grês porcelanato** , **óxido de zinco** e **quartzo** , empregados na fabricação de fritas cerâmicas .

Termos -> zinco / quartzo

Arquivo: c:/perl/analise/sent_c~1/estrangeiros/silosv2.sent.txt

A seqüência correta de termos a ser produzida seria “grês porcelanato, óxido de zinco e quartzo”. Porém, a palavra óxido apresenta a etiqueta de um adjetivo, e dessa forma, óxido de zinco não é considerado um termo, mas somente zinco, fazendo com que o script considere como termos as palavras zinco e quartzo.

A fim de avaliar se o esquema de repetição de termos, inserido em scripts que foram baseados em orações que não apresentam essa característica, prejudicaria o resultado produzido pelos mesmos, tal esquema foi implementado para três desses scripts: “formad(o)(a)(s)”, “ou seja” e “obtid(o)(a)(s)”. O resultado produzido foi semelhante ao dos scripts que apresentaram seqüência

de termos nos casos em que as orações apresentaram somente um termo para ser extraído e não uma seqüência.

Na oração 4, o script extrai “deformações da rede” e “fissuras” como termos, ao contrário do que ocorre na saída original (sem esquema de repetição de termos), em que “fissuras” (o termo real) é considerado como termo.

Oração 4:

Este efeito está relacionado às deformações da rede e **fissuras** formadas por alterações de volume no ponto de transição; estas deformações da rede e fissuras ocorrem extensivamente no quartzo, por exemplo, no qual a mudança de volume é grande.

Termos -> deformações da rede / fissuras

Arquivo: c:/perl/analise/sent_c~1/nacionais/PNOVAESv2.sent.txt

6.2.3 Discussão dos resultados dos experimentos 1 e 2

Tendo analisado os resultados produzidos pelos experimentos 1 e 2, que utilizaram um subcórpus para a construção de um script para cada expressão, é possível perceber que eles atenderam aos requisitos inicialmente propostos, ou seja, o objetivo do primeiro experimento consistiu na extração de todas as orações selecionadas do subcórpus, e dessa forma, extrair uma maior quantidade de termos em cada oração através da não-especificação de etiquetas para as expressões presentes nos scripts, visto que tais expressões algumas vezes apareciam etiquetadas erroneamente no córpus, prejudicando, assim, o desempenho dos scripts. A Tabela 6.1, construída para um melhor entendimento do processo ocorrido no primeiro experimento, considerou a quantidade de orações e termos extraídos com e sem a especificação de etiquetas para as 10 expressões escolhidas, e os números obtidos indicaram que todas as orações do subcórpus passaram a ser extraídas quando da não-especificação das etiquetas das expressões, e dessa forma, na maioria dos casos, o número de termos selecionados nessas orações aumentou. Por outro lado, o objetivo do segundo experimento era a padronização dos scripts construídos a partir de orações que apresentaram repetição de termos, isto é, conseguir uma forma de generalizar a parte dos scripts em que havia termos se repetindo. Esse experimento conseguiu atingir seus objetivos inicialmente propostos, dado que foi utilizado um esquema de repetição de variáveis correspondentes a termos em todos os scripts que apresentavam tal característica, e a saída produzida por eles, considerando o subcórpus selecionado, foi semelhante àquela produzida antes da padronização dos scripts. Porém, algumas seqüências de termos presentes nas orações das expressões em questão não foram consideradas, considerando que elas apresentaram-se muito específicas, fugindo da proposta desse experimento. Notou-se também erros de português influenciando na extração de candidatos a termo, como nos casos a seguir com a abreviatura “etc”.

AntConc 2.5.1 (Windows) Laurence Anthony, 2003

File Settings Preferences Number of Concordance Hits: 25

Directory: C:\Perif\etiquetado_sent_copi

Concordancer Concordance Search Term Plot View Files Word List Keyword List

[1] _ADJ , , pvc_NP , , borracha_N e_CONJ etc_N ((linguagem_N comercial_A analise_critica.sent.t

[2] istribuidores_N , , técnicos_N e_CONJ etc_N)) usa-se_VERB correntemen analise_critica.sent.t

[3] _PREP chuveiro_N , , piscinas_N etc_ADJ , , a_ART adsorção_N de_P analise_critica.sent.txt.tag

[4] institutos_N de_PREP pesquisa_N etc_ADJ . . O_ART estudo_N da_PRE edmilsonv2.sent.txt.tagged.t

[5] ras_N , , pisos_N plásticos_ADJ etc_ADJ . . As_ART empresas_N que edmilsonv2.sent.txt.tagged.t

[6] onalização_N de_PREP processo_N etc_ADJ , , que_PRON terão_VERB i edmilsonv2.sent.txt.tagged.t

[7] s_PREP o_ART assentamento_N ; ; etc_ART . . N Como_CONJ se_PRON po mancha_daguav2.sent.txt.tag

[8] ADJ , , granito_N , , sienito_N etc_ADJ)) , , talco_N e_CONJ ca panorama.sent.txt.tagged.txt

[9] , , feldspatos_N , , quartzo_N e_CONJ etc_N mudam_VERB esta_PRON relação reformul_engobes.sent.

[10] , , feldspatos_N , , quartzo_N e_CONJ etc_N mudam_VERB esta_PRON relação reformul_engobes.sent.

[11] N , , granitos_N , , arenitos_N etc_ADJ . . A_ART decoração_N pod tecnologia_avancada.sent.txt

[12] ribuição_N das_PREP+ART peças_N etc_ADJ . . É_VERB importante_AD v1n45_5.sent.txt.tagged.txt

[13] , , Indonésia_NP , , Tailândia_NP etc_NP , , estão_VERB abastecidos v2n34_2.sent.txt.tagged.txt

[14] nal_ADJ , , matérias-primas_ADJ etc_N , , que_PRON elencamos_VERB v2n34_2.sent.txt.tagged.txt

[15] ima_N , , peças_N especiais_ADJ etc_ADJ)) . . o_ART estímulo_N v2n34_2.sent.txt.tagged.txt

[16] DJ , , hotéis_N , , hospitais_N etc_ADJ , , aproveitando_VERB as_ v2n34_2.sent.txt.tagged.txt

[17] N , , feldspatos_N , , micas_N etc_ADJ . . Em_PREP alguns_ADJ ca v2n56_6.sent.txt.tagged.txt

[18] ADJ , , granito_N , , sienito_N etc_ADJ)) , , talco_N e_CONJ ca v5n6_5.sent.txt.tagged.txt

[19] ras_N , , jogos_N de_PREP chá_N etc_ADJ)) ; ; porcelana_N elétr v6n2_4.sent.txt.tagged.txt

[20] s_N , , peças_N decorativas_ADJ etc_ADJ . . Para_PREP abordar_VE v6n2_4.sent.txt.tagged.txt

[21] _ADJ , , talco_ADJ carbonatos_N etc_ADJ)) ; ; e_CONJ até_PREP 5 v6n2_4.sent.txt.tagged.txt

[22] es_N , , peças_N artesanais_ADJ etc_ADJ)) . . Observa-se_VERB t v6n2_4.sent.txt.tagged.txt

[23] vil_ADJ de_PREP cada_ADJ país_N etc_ADJ . . Mesmo_ADV com_PREP t v6n3_2.sent.txt.tagged.txt

[24] ras_N , , pisos_N plásticos_ADJ etc_ADJ . . Assim_ADV as_ART empr v6n3_2.sent.txt.tagged.txt

[25] P , , SENAI_NP , , sindicatos_N etc_NP)) E_CONJ , , nesse_LOCU v8n1_5.sent.txt.tagged.txt

Files Processed Search Term Whole Words Case Sensitive Regex Start Stop Sort Level 1 0 < > Search Window Size Level 2 0 < > 30 < > Exit

[^\\, \\,] (\\s+|e_CONJ\\s+)etc

17:46

Capítulo 7

Implementação e avaliação de um método híbrido

Para a abordagem híbrida, foi primeiramente construído um script, semelhante àquele utilizado no método lingüístico (ver Seção 6.1), para geração de orações do corpúsculo que apresentassem as expressões passadas como entrada, com a diferença que agora cada oração é impressa na saída somente uma vez, independente do número de expressões que ela pode apresentar. As orações geradas também foram transformadas para caixa baixa (letra minúscula) por esse script. O subcorpúsculo de saída, constituído pelas orações que apresentaram alguma expressão lingüística (Figura 6.1), é tomado como entrada para o pacote NSP. Como o número de palavras presente no subcorpúsculo é 313019 (19578 não repetidas), realizou-se um corte na frequência 3, seguindo, dessa forma, o mesmo critério adotado na abordagem estatística. Em seguida, a frequência, única medida estatística para unigramas encontrada no pacote NSP, foi calculada para os unigramas do subcorpúsculo, utilizando-se o mesmo corte determinado na abordagem estatística, ou seja, 20 (ver Figura 5.9). Para o cálculo da frequência foram utilizadas a *stoplist 2* (construída para o método híbrido), constituída tanto pelos elementos presentes na *stoplist 1* quanto por novos *tokens* mostrados no Apêndice C, e a regra de formação de *tokens*, que sofreu algumas alterações (mostradas nas 3 últimas linhas do quadro a seguir) em relação àquela utilizada na abordagem estatística, a fim de permitir a extração dos *tokens* seguidos de suas etiquetas:

/([a-zA-Z-])	→ representa caracteres alfabéticos que podem apresentar hífen
[\\w\\xb0]	→ representa o “°” (grau)
[\\w\\xc0-\\xc5]	→ representa a letra “á” maiúscula com as acentuações possíveis
[\\w\\xc7-\\xcf]	→ representa o “ç”, as letras “é” e “í” com acentuações (maiúsculos)
[\\w\\xd1-\\xd6]	→ representa o “ñ” e a letra “ó” com acentuações (maiúsculos)
[\\w\\xd9-\\xdc]	→ representa a letra “ú” maiúscula com acentuações
[\\w\\xdf-\\xe5]	→ representa a letra “ß” e a letra “á” minúscula com acentuações
[\\w\\xe7-\\xef]	→ representa o “ç”, as letras “é” e “í” com acentuações (minúsculos)
[\\w\\xf1-\\xf6]	→ representa o “ñ” e a letra “ó” com acentuações (minúsculos)
[\\w\\xf9-\\xfc]+	→ representa a letra “ú” minúscula com acentuações
([\\w\\x5f])	→ representa o “_” (underscore)
([w+])	→ representa caracteres alfanuméricos
[\\w\\x2b]+/	→ representa o “+” (sinal de mais)

O cálculo da frequência também foi efetuado para os bigramas e os trigramas do subcorpúsculo, realizando o corte na frequência 18 tanto para bigramas quanto para trigramas, como estipulado na abordagem estatística (ver Figuras 5.10 e 5.14). Como a informação mútua, quando calculada para bigramas, apresentou a melhor precisão na abordagem estatística, ela também foi utilizada nessa abordagem, considerando como corte o escore 0.0066 (ver Figura 5.11). Após o cálculo das

medidas para unigramas, bigramas e trigramas, as listas resultantes desse cálculo são tomadas como entrada para um outro script (um script para cada uma das listas), que realiza a intersecção das listas de unigramas, bigramas e trigramas com os seus respectivos padrões, produzindo uma lista somente com aqueles unigramas, bigramas e trigramas que apresentaram algum padrão de termo (ver Figura 6.2). Em razão de algumas palavras apresentarem mais de uma etiqueta (falha do etiquetador), elas apareceram mais de uma vez na lista interseccionada com os padrões, mas com etiquetas diferentes. Por esse motivo, foi construído um outro script a fim de eliminar essas palavras repetidas e posteriormente realizar a intersecção dos unigramas, bigramas e trigramas com suas respectivas listas de referência. A Figura 7.1 apresenta o método híbrido implementado.

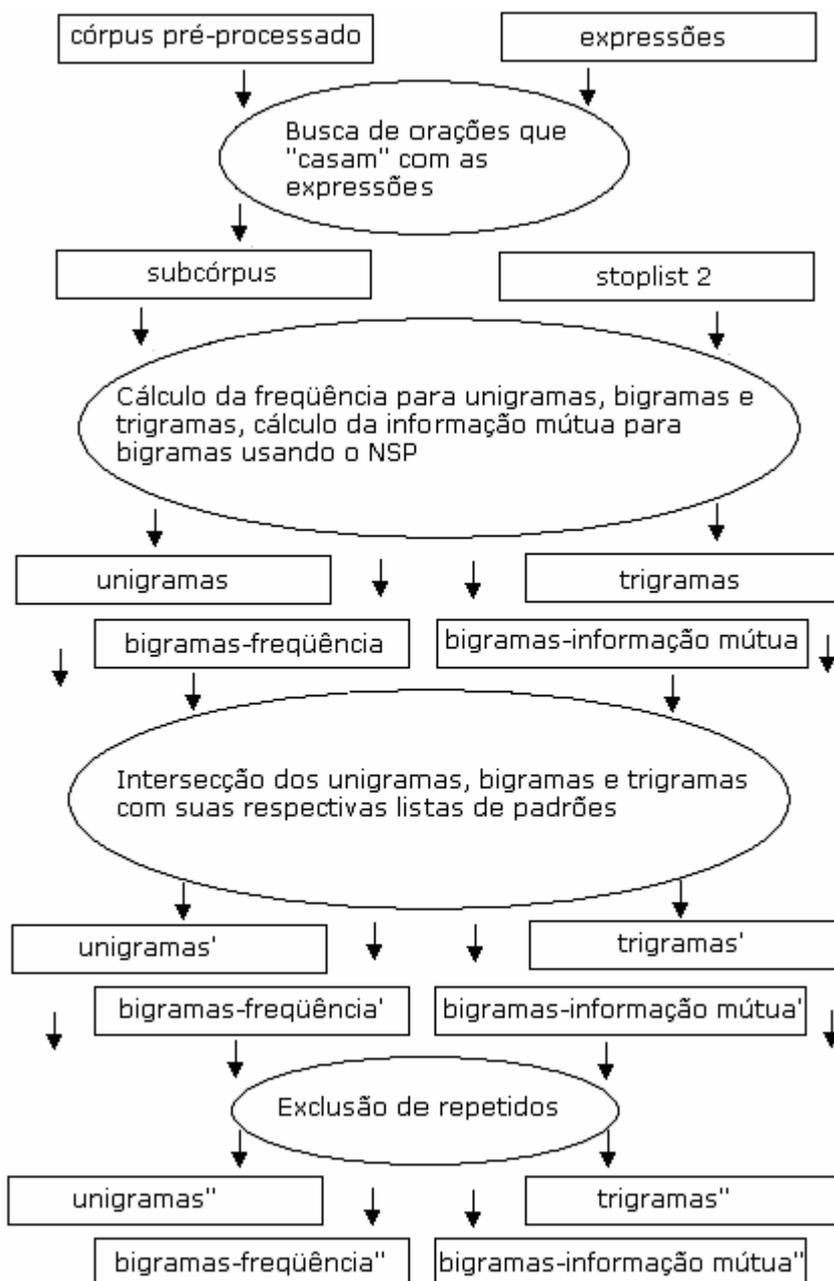


Figura 7.1 – O método híbrido implementado.

Analisando as listas de unigramas, bigramas e trigramas produzidas, é possível perceber que alguns dos termos presentes na abordagem estatística, após a realização do corte, não foram extraídos, pois alguns deles foram etiquetados erroneamente, apresentando, dessa forma, padrões diferentes daqueles passados em um arquivo de entrada para o segundo script. Outros termos não foram extraídos em razão do corte, pois como alguns desses termos apresentaram mais de uma etiqueta, a frequência total correspondente àquele termo (sem a etiqueta) foi repartida entre os termos que apresentaram etiquetas diferentes, não alcançando, assim, o corte estabelecido.

O total de unigramas levantados com o cálculo da frequência foi 1412 (uma redução de 91,1% se comparada com a lista produzida pelo método lingüístico para unigramas), e a intersecção desse número com a lista de referência produziu 105 termos.

Para bigramas, foram calculadas as medidas de frequência e informação mútua. Para a primeira, a quantidade de bigramas obtida foi 78 (redução de 99,4% comparada com o método lingüístico), sendo 14 o resultado da intersecção dessa quantidade com a lista de referência. Já o número de bigramas levantados pela informação mútua foi 78 (apresentando 99,4% de redução comparado com o método lingüístico), dentre os quais, 13 são termos.

A frequência, calculada para trigramas, produziu um total de 76 trigramas (redução de 99,5% comparado com o método lingüístico), sendo 20 o resultado da intersecção desse total com a lista de referência.

Após o cálculo dessas medidas sobre os unigramas, os bigramas e os trigramas, produzidos pelos diversos passos do método híbrido, a obtenção da quantidade desses unigramas, bigramas e trigramas, bem como a obtenção de seus respectivos termos, é possível calcular a Precisão, a Revocação e a Medida F para cada um deles (Quadro 7.1).

Quadro 7.1: Precisão, Revocação e Medida F do método híbrido

Unigramas - Frequência

$$\text{Precisão} = 105/1412 = 0,07 = 7\%$$

$$\text{Revocação} = 105/264 = 0,39 = 39\%$$

$$\text{Medida F} = 0,0546/0,46 = 0,11 = 11\%$$

Bigramas - Frequência

$$\text{Precisão} = 14/78 = 0,17 = 17\%$$

$$\text{Revocação} = 14/74 = 0,18 = 18\%$$

$$\text{Medida F} = 0,0612/0,35 = 0,17 = 17\%$$

Bigramas - Informação Mútua

Precisão = $13/78 = 0,16 = 16\%$

Revocação = $13/74 = 0,17 = 17\%$

Medida F = $0,0544/0,33 = 0,16 = 16\%$

Trigramas - Frequência

Precisão = $20/76 = 0,26 = 26\%$

Revocação = $20/43 = 0,46 = 46\%$

Medida F = $0,2392/0,72 = 0,33 = 33\%$

Analisando os resultados produzidos, apesar de as precisões nessa abordagem não terem alcançado valores elevados, elas foram as melhores, comparadas com as precisões obtidas nas abordagens estatística e lingüística, permitindo concluir que a abordagem híbrida é a melhor escolha para a extração automática de termos se a medida escolhida for a precisão. Se a revocação for a medida escolhida para avaliar os métodos, o método lingüístico se sobressai.

Capítulo 8

Conclusões

Esse projeto de mestrado, denominado Evaluation of Terminology Automatic Extraction Methods for Portuguese Texts (ExPorTer)⁵¹, foi dedicado à implementação e avaliação de métodos de extração automática de termos das abordagens estatística, lingüística e híbrida. Para os métodos propostos em cada abordagem foram realizados o cálculo da precisão, revocação e medida F, e os resultados obtidos para essas medidas mantiveram-se bem distantes daquele normalmente obtido pela tarefa de extração (medida F em torno de 60%). As baixas porcentagens obtidas por essas medidas de avaliação se justificam pelo fato de que o tamanho do córpus, embora grande (448.352 palavras), é de fonte única, a Revista Cerâmica Industrial, o que fez com que a intersecção com a lista original de referência, que continha 747 termos (671 dos quais eram uni, bi e trigramas) e havia sido gerada de vários tipos diferentes de materiais da área de Revestimentos Cerâmicos, gerasse apenas 381 termos. Assim, temos um córpus grande quando comparado com a lista de referência. Das três abordagens, aquela que apresentou um método com a maior revocação foi a lingüística, alcançando valores entre 90 e 100%. Por outro lado, o método proposto na abordagem híbrida foi aquele que apresentou os melhores valores de precisão, apesar de esses valores não terem sido elevados. A fim de solucionar esse problema de baixas precisões, foi proposto, na abordagem estatística, um método semi-automático, isto é, com intervenção humana, que tem como objetivo o levantamento de candidatos mais prováveis de serem termos e exclusão das palavras da língua geral. A precisão obtida para unigramas (48%), considerando esse método, se apresentou bem melhor em relação àquela obtida pelo método híbrido (7%). Já para bigramas, os valores da precisão para ambos os métodos foram bem próximos (15% para frequência e informação mútua do método estatístico com interferência humana, 17% para frequência e 16% para informação mútua do método híbrido). A revocação obtida considerando esse método (estatístico com interferência humana) apresentou bons resultados para frequência aplicada para unigramas (74%) e para a informação mútua aplicada para bigramas (81%), e um resultado razoável para a frequência de bigramas (56%). Considerando essas comparações, entende-se que é de grande importância a aplicação de esforço humano à abordagem estatística, visto que os resultados obtidos por esse método foram os mais notórios. É importante salientar que esse esforço realizado na separação e categorização das palavras da língua geral em *palavras e siglas da língua geral, marcas publicitárias, nomes próprios, e símbolos especiais* levou dois meses de trabalho de duas pessoas (a pesquisadora desse trabalho e uma lingüista), tempo considerado pequeno se comparado com o

⁵¹ <http://www.nilc.icmc.usp.br/nilc/projects/termextract.htm>

benefício. Esse trabalho pode também servir para a criação de uma *stoplist* que eliminaria automaticamente palavras da língua geral aumentando a precisão dos métodos das três abordagens.

8.1 Contribuições

A primeira contribuição do projeto foi a ampla avaliação dos métodos das três abordagens implementados, que foi exposta de uma forma clara para serem entendidos por terminólogos pouco acostumados aos métodos de EAT.

O método proposto na abordagem híbrida, que consiste na integração das medidas apresentadas na abordagem estatística com o método utilizado na abordagem lingüística, foi outra contribuição do projeto ExPorTer.

A proposta da inclusão de esforço humano junto às medidas estatísticas também foi uma idéia que surgiu durante a realização das tarefas desse projeto.

Alguns recursos construídos para a constituição dos métodos apresentados nesse trabalho já estão sendo utilizados em outros projetos, como o Development of the Computational Ontological Database of the Ecology for Brazilian Portuguese (Bloc-Eco)⁵².

8.2 Limitações

O tamanho da lista de referência utilizada nesse trabalho foi muito pequeno em relação ao tamanho do córpus, sendo esse um dos fortes elementos que contribuiu na diminuição dos valores de precisão calculados para os métodos das três abordagens. É também importante salientar que os métodos são simples e que não se aplicou um lematizador no córpus o que influenciou no levantamento de bigramas e trigramas além do cálculo da precisão e revocação.

A falta de uma *stoplist* mais elaborada, apresentando um maior número de palavras da língua geral e verbos improváveis de serem termos, freqüentes em textos científicos, também influenciou nos baixos resultados de precisão, pois uma *stoplist* é responsável por eliminar palavras da lista de candidatos a termos.

A escassez de tempo foi o principal motivo pelo qual não se implementou métodos mais avançados para a extração de termos, ou até mesmo não se buscou a realização de uma melhoria dos métodos utilizados nesse trabalho, o que permitiria uma tentativa de se obter resultados de precisão e revocação melhores e mais próximos daqueles encontrados na literatura.

⁵² <http://www.nilc.icmc.usp.br/nilc/projects/bloc-eco.htm>

8.3 Trabalhos Futuros

Pretende-se futuramente aplicar os métodos aqui descritos em córpus de outros domínios, buscando uma lista de referência que seja proporcional ao tamanho do córpus, a fim de poder avaliar esses métodos de uma forma mais coerente. Inclusive explicitar melhor a relação entre esses tamanhos de uma forma experimental.

Os problemas encontrados nesse trabalho também podem ser facilmente solucionados. A *stoplist*, por exemplo, poderia ser reelaborada, de forma a torná-la mais completa possível e um lematizador poderia ser utilizado.

Os vários scripts implementados para cada uma das expressões lingüísticas, descritos no Capítulo 6, poderiam ser avaliados quantitativamente, aplicando-os somente sobre a parte do córpus que não foi utilizada para a escrita dos mesmos (desconsiderando o subcórpus utilizado).

Métodos de extração de termos mais elaborados também poderiam ser aplicados sobre o domínio de Revestimentos Cerâmicos, com o intuito de realizar comparações com os métodos implementados nesse projeto, buscando, dessa forma, uma melhor técnica de extração automática de terminologia para o português.

Referências Bibliográficas

- ALMEIDA, G.M.B. (2000). Teoria Comunicativa da Terminologia (TCT): uma aplicação. vol.I. Tese (Doutorado em Lingüística e Língua Portuguesa) – Faculdade de Ciências e Letras, Campus de Araraquara, Universidade Estadual Paulista.
- ALUÍSIO, S.M. (1995). Ferramentas de Auxílio à Escrita de Artigos Científicos em Inglês como Língua Estrangeira. Tese de Doutorado, IFSQ-USP, 228p., Agosto de 1995.
- ALUÍSIO, S.M.; PINHEIRO, G.; FINGER, M.; NUNES, M.G.V.; TAGNIN, S.E. (2003). The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation. In: CORPUS LINGUISTICS 2003, 2003, Lancaster, UK. Proceedings of Corpus Linguistics 2003 (Also as UCREL Technical Report, Vol 16 Part). Lancaster: 2003. v. 16, p. 14-21.
- APPELT, D.; ISRAEL, D. (1999). Introduction to Information Extraction Technology. IJCAI'99 Tutorial. Disponível em: www.ai.sri.com/~appelt/ie-tutorial/.
- BOLSHAKOVA, E. (2001). Recognition of Author's Scientific and Technical Terms. LNCS 2004, p. 281-90.
- BOURIGAULT, D. (1992). Surface grammatical analysis for the extraction of terminological noun phrases. In Proceedings of the 14th International Conference on Computational Linguistics, COLING 1992, p. 977-981.
- BRILL, E. (1995). Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging. Computational Linguistics, Dec. 1995.
- DAILLE, B. (1994). Combined approach for terminology extraction: lexical statistics and linguistic filtering, PhD thesis, University of Paris 7.
- DAILLE, B. (1996). Study and Implementation of Combined Techniques for Automatic Extraction of Technology. In: Klavans, J., Resnik, P. The Balancing ACT- Combining Symbolic and Statistical Approaches to Language, The MIT Press, p. 49-66.

- DIAS, G.; GUILLORÉ, S.; BASSANO, J. C.; LOPES, J.G.P. (2000). Combining Linguistics with Statistics for Multiword term Extraction: A Fruitful Association? In: Proceedings of Recherche d'Informations Assisté par Ordinateur. Paris, France.
- ESTOPÀ BAGOT, R. (1999). Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada). Tese de Doutorado. Universidade Pompeu Fabra.
- ESTOPÀ BAGOT, R. (2001). Extracción de Terminología: elementos para la construcción de un extractor. In TradTerm 7. Revista do Centro Interdepartamental de Tradução e Terminologia FFLCH - USP, p. 225-50.
- FRANTZY, K. T.; ANANIADOU, S. (1997). Automatic Term Recognition using Contextual Cues. Manchester Metropolitan University. THIRD DELOS WORKSHOP Cross-Language Information Retrieval Zurich, 5-7 March 1997 ISBN 2-912335-02-7.
- GEORGANTOPOULOS, B.; PIPERIDIS, S. (1998). Automatic acquisition of terminological resources for Information Extraction Applications. In: Proceedings of the 1st Panhellenic Conference on New Information Technologies, NIT'98.
- HA, L.A. (2004). Co-training applied in automatic term extraction: an experiment. In: 7th Annual CLUK Research Colloquium, University of Birmingham, Jan 2004. Disponível em <http://www.cs.bham.ac.uk/~mgl/cluk/titles.html>.
- HEID, U.; JAUß, S.; KRÜGER, K.; HOHMANN, A. (1996). Term extraction with standard tools for corpus exploration. In: 4th International Congress on Terminology and Knowledge Engineering, Wien. August.
- HOBBS, J. R.; APPELT, D.; BEAR, J.; ISRAEL, D.; KAMEYAMA, M.; STICKEL, M.; TYSON, M. (1997). FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. Disponível em: <http://www.ai.sri.com/~appelt/fastus-schabes.html>.
- KLAVANS, J. L.; MURESAN, S. (2000). DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from Online Text. In: Proceedings of AMIA 2000.

- KLAVANS, J. L.; MURESAN, S. (2001a). Evaluation of DEFINDER: A System to Mine Definitions from Consumer-oriented Medical Text. In: Proceedings of JCDL 2001.
- KLAVANS, J. L.; MURESAN, S. (2001b). Evaluation of the DEFINDER System for Fully Automatic Glossary Construction. In: Proceedings of AMIA 2001.
- KRIEGER, M.G. (2001). O Termo: Questionamentos e Configurações. TradTerm 7, Revista do Centro Interdepartamental de Tradução e Terminologia FFLCH – USP, p. 111-40.
- MACIEL, A. M. B. (1996). Revista Internacional de Língua Portuguesa, nº 15, p. 69-76. Lisboa.
- MANNING, C.; SCHÜTZE, H. (1999). Collocations. In: Foundations of Statistical Natural Language Processing, p. 141-77. MIT Press. Cambridge.
- MARQUES, N. (2000). Metodologia para a Modelação Estatística de Subcategorização verbal. Ph.D. Thesis. Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia, Lisbon, Portugal.
- MORALES, D. F. G. (2001). TermDic: um dicionário eletrônico. In: Temas de Terminologia, Krieger, M. G.; Maciel, A. M. B. (Ed.) Editora de Universidade, p. 364-66.
- OH, J.; CHAE, Y.; CHOI, K. (2000). A Statistical Model for Automatic Recognition of Biological Terminologies, Workshop on Computational Terminology for Medical and Biological Applications, NLP2000. Disponível em <http://nlplab.kaist.ac.kr/~rovellia/english/atr.htm>.
- PANTEL, P.; LIN, D. (2001). A statistical corpus-based term extractor. In: E. Stroulia & S. Matwin (Ed.), AI 2001, Lecture Notes in Artificial Intelligence, Springer-Verlag, p. 36–46.
- PIAO, S. L.; RAYSON, P.; ARCHER, D.; WILSON, A.; MCENERY, T. (2003). Extracting multiword expressions with a semantic tagger. In Proceedings of the Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, at ACL'03, the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, pp. 49-56.

- RAMSHAW, L. A.; MARCUS, M. P. (1995). Text Chunking Using Transformation-Based Learning. In: Proceedings of Third ACL Workshop on Very Large Corpora, MIT.
- RATNAPARKHI, A. (1996). A Maximum Entropy Part-Of-Speech Tagger. In Proceedings of the Empirical Methods in Natural Language Processing Conference, May 17-18, 1996. University of Pennsylvania.
- SAGER, J.C. (1993). Curso práctico sobre el procesamiento de la terminología. La dimensión cognoscitiva. Madrid: Pirámide.
- SILVA, J.; DIAS, G.; GUILLORÉ, S.; LOPES, J.G.P. (1999). Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In: Proceedings of 9th Portuguese Conference in Artificial Intelligence. Springer-Verlag.
- SMADJA, F. (1991). Retrieving Collocational Knowledge from Textual Corpora. An application: Language Generation. PhD Thesis, Computer Science Department, Columbia University.
- TELINE, M.F.; ALMEIDA, G.M.B.; ALUÍSIO, S.M. (2003). Extração Manual e Automática de Terminologia: comparando abordagens e critérios. In I Workshop em Tecnologia da Informação e da Linguagem Humana (I TIL). ICMC-USP, São Carlos, São Paulo, Outubro 2003.
- TIEDEMANN, J. (1997). Automatical Lexicon Extraction from Aligned Bilingual Corpora, Diploma thesis, Magdeburg.
- ZHAO, J. (1999). The impact of cross-entropy on language modelling. Disponível em: www.isip.msstate.edu/publications/courses/ece_7000_speech/lectures/1999/lecture_06/paper/paper_v1.pdf.
- YANGARBER, R.; GRISHMAN, R. (2000). Extraction Pattern Discovery through Corpus Analysis. TR- 00-143, The Proteus Project, New York University. In: Proceedings of the Workshop Information Extraction meets Corpus Linguistics, Second International Conference on Language Resources and evaluation (LREC 2000), Athens, Greece.

Apêndice A

A. Telas com exemplos de expressões linguísticas e seus contextos

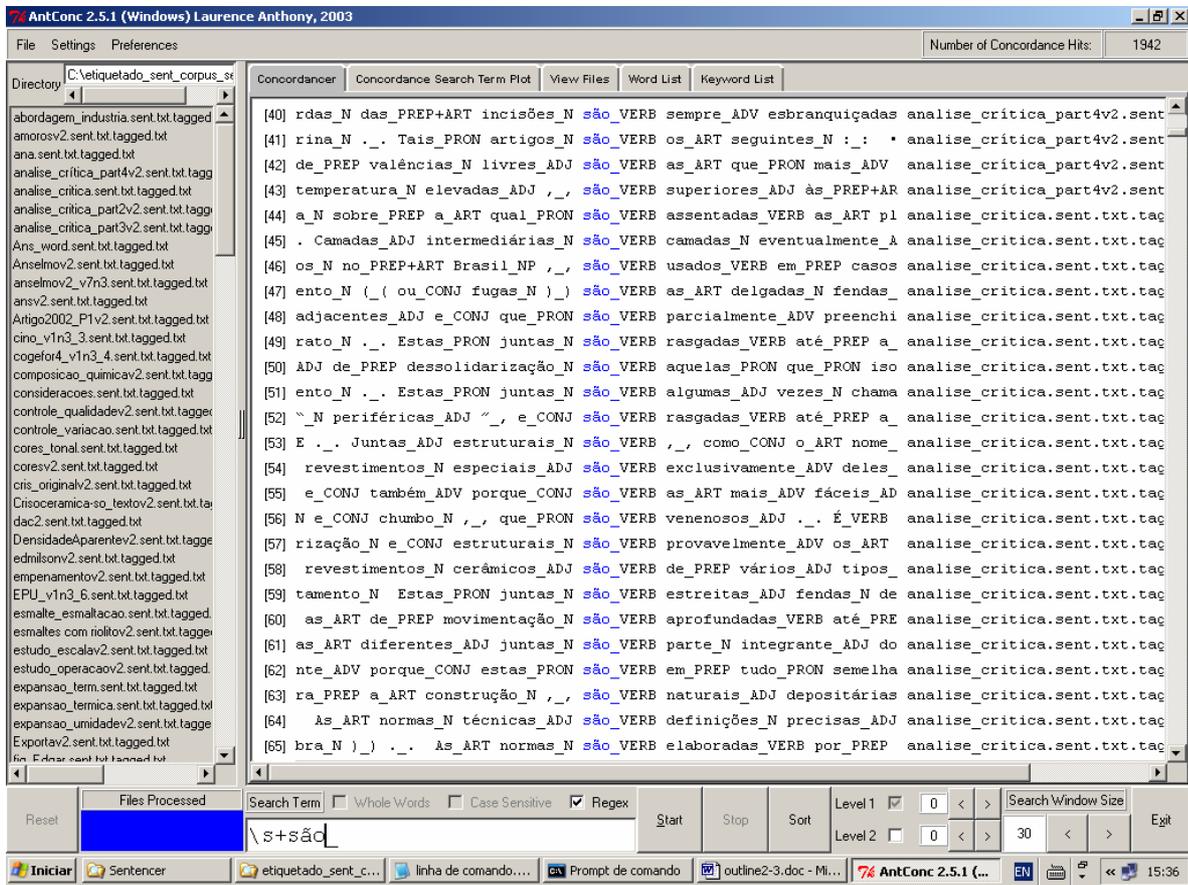


Figura A.1 - Tela com 1942 concordâncias da expressão “são” no corpús.

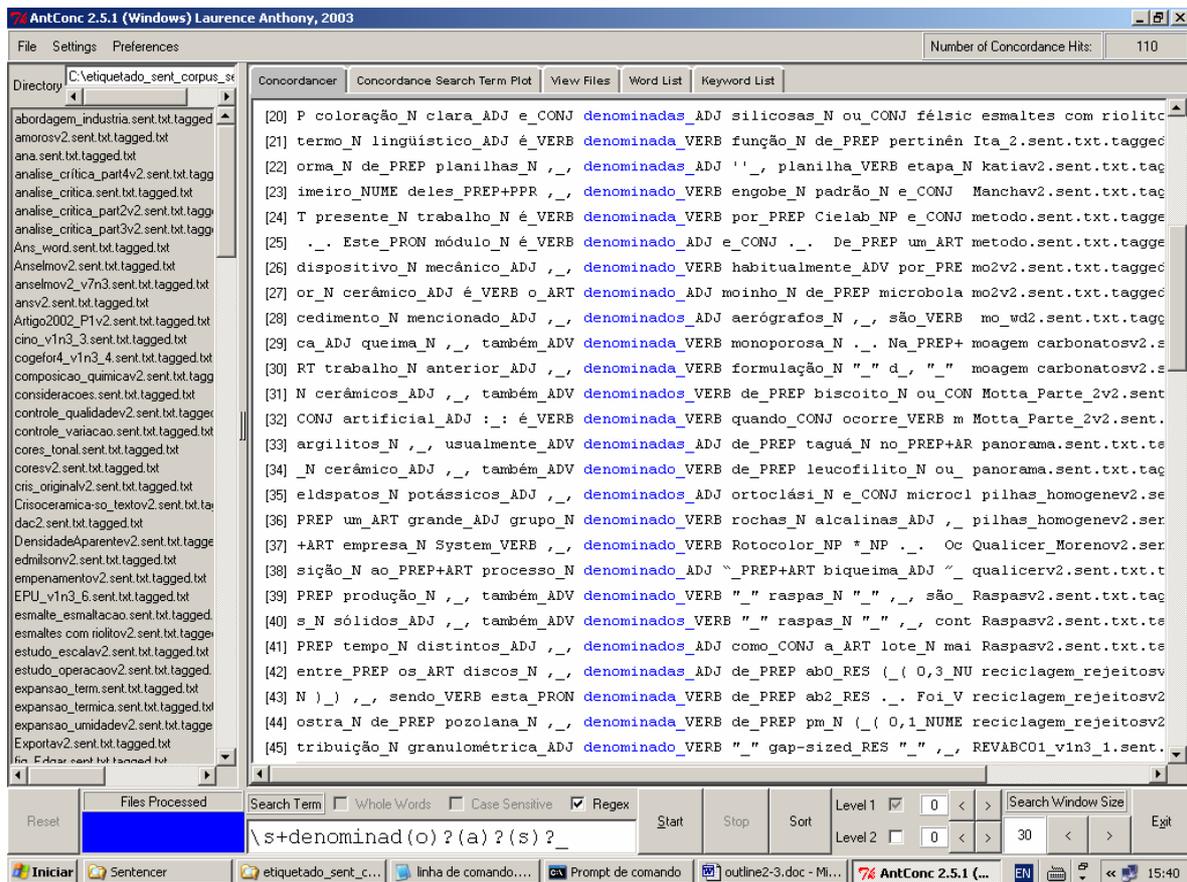


Figura A.2 - Tela com 110 concordâncias da expressão “denominad” no corpus.

Apêndice B

**B. Lista de referência separada em unigramas,
bigramas e trigramas**

Unigramas:

acabamento	calcita	dolomita	grês
aderência	calibre	dosador	gretamento
aditivo	campana	dosagem	haloisita
aerógrafo	cantoneira	duro	hematita
agalmatolito	carbonato	eflorescência	hidratação
agitador	carregamento	eletrofusão	ilita
aglomeração	cascata	empeno	impermeabilizante
aglomerado	caulim	empilhamento	jateamento
aglomerante	caulinita	emulsão	lança
aglutinante	classificação	engobe	laranja
agregado	classificador	envelhecimento	lascamento
albita	CMC	enxofre	ligante
álcali	colagem	EPU	limpabilidade
alimentação	colorido	escorrimento	lubrificante
alumina	cominuição	esmaltação	luneta
amarelo	compactação	esmalte	magnesita
amolecimento	composição	esmectita	mármore
amostra	conformação	espalhamento	marmorizado
anortita	corante	espátula	maromba
anverso	corpo-de-prova	espodumênio	massa
aquecimento	cortador	esquadro	mate
areia	cotto	estampo	matização
argila	crystalina	estiramento	matriz
argilomineral	crystalização	extrusão	maturação
armazenagem	chromo	extrusora	mica
atomização	curvatura	faixa	microfissura
atomizador	decalcomania	feldspato	mineral
azul	decoração	fileira	mistura
azulejo	decorador	filito	moagem
barra	decoradora	fissura	mohs
bege	defloculação	fixador	moinho
bentonita	defloculante	flexografia	molde
bico	deflocular	floculante	mole
biqueima	densidade	fluorita	monoporosa
biscoito	densímetro	fonolito	monoqueima
bitola	desaeração	forno	montmorilonita
bola	descarte	fotolito	muratura
bomba	desferrização	frita	ocografia
boquilha	dessecador	fundente	ocre
brancura	destorroador	granilha	opacidade
brilhante	dextrina	granulação	opacificante
brilho	dilatometria	granulador	opaco
britador	dimensional	grânulo	ortoclásio
caco	diopsídio	granulometria	ortogonalidade
calcário	disco	grão	paralelismo
calcimetria	dispersante	gravimétrico	partícula
calcinação	dispersar	grelha	pastilha

peça	quartzo	silex	transparente
PEI	queima	sílica	trinca
peneira	queimador	silicato	trituração
peneiramento	rampa	silo	turquesa
pesagem	refratariedade	sincronizada	vasca
PH	relevo	superfície	vela
picnometria	reologia	suporte	verde
piso	requeima	suspensão	vermelho
planaridade	resfriamento	taguá	verso
plasticidade	retificação	talco	véu
plastificante	rosa	tanque	vidrado
polido	Rotocolor	tardoz	vidro
polimento	secador	termopar	viscosidade
porosidade	secagem	titânia	viscosímetro
prensa	semigrês	tixotropia	visual
prensagem	semiporoso	tonalidade	vitro-cerâmico
preto	sericita	TOT	wollastonita
pseudoplasticidade	serigrafia	trabalhabilidade	zircônia
pulverização	silagem	transparência	zirconita

Bigramas:

abrasão superficial	dilatação térmica
análise granulométrica	distribuição granulométrica
análise química	engobe refratário
areia feldspática	esmalte cru
argila branca	esmalte fritado
argila caulinitica	esmalte opaco
argila fundente	esmalte transparente
argila gorda	esteira transportadora
argila magra	expansão linear
argila plástica	expansão térmica
argila refratária	fase vítrea
aspecto superficial	feldspato potássico
ball clay	feldspato sódico
bico pulverizador	forno elétrico
biqueima rápida	forno túnel
biqueima tradicional	frita branca
composição granulométrica	frita cristalina
coordenada cromática	frita opaca
coração negro	grês polido
cortina contínua	grês porcelanato
curvatura central	máquina serigráfica
curvatura lateral	massa cerâmica
densidade aparente	matéria orgânica
dilatação higroscópica	matriz serigráfica
	moinho contínuo

moinho pendular
nefelina sienito
pasta serigráfica
peneira malha
pigmento cerâmico
placa cerâmica
pó atomizado
prensa hidráulica
prensagem uniaxial
produto acabado
punção inferior
punção superior
resistência mecânica

retitude lateral
revestimento cerâmico
rugosidade superficial
sal solúvel
suporte queimado
tela serigráfica
terceira queima
tinta serigráfica
trabalho térmico
veículo serigráfico
via seca
via úmida

Trigramas:

absorção de água
carbonato de cálcio
carbonato de magnésio
carga de ruptura
ciclo de queima
coeficiente de atrito
coeficiente de dilatação
concentração de sólidos
cor de queima
curva de defloculação
curva de queima
difração de raios-X
espessura de camada
expansão por umidade
forno a rolos
granulação a seco
grau de compactação
grau de moagem
grês porcelanato esmaltado
linha de esmaltação
moagem a seco
moagem a úmido

moinho de bolas
óxido de ferro
óxido de zinco
perda ao fogo
pressão de prensagem
resíduo em malha
resistência à abrasão
resistência à flexão
resistência ao manchamento
retração de queima
retração de secagem
silicato de zircônio
temperatura de acoplamento
temperatura de amolecimento
temperatura de maturação
temperatura de queima
tempo de secagem
teor de umidade
variação de tonalidade
zona de queima
zona de resfriamento

Apêndice C

C. *StopList*

Alguns erros, como os *tokens* “á”, “palavras-chaves” e “consequentemente”, foram incluídos na *stoplist* com o intuito de desconsiderar a ocorrência dos mesmos em arquivos do *corp*us.

A *stoplist 1* é constituída pelos seguintes *tokens*:

e	da	um
é	das	uma
à	do	uns
às	dos	umas
a	de	nós
ao	essa	eu
o	essas	tu
aos	esse	ela
os	esses	ele
as	esta	elas
na	estas	eles
nas	este	você
no	estes	vocês
nos	isso	sob
nessa	isto	sobre
nessas	nele	sobretudo
nesse	neles	depois
nesses	nela	durante
nisso	nelas	todo
nesta	naquele	todos
nestas	naqueles	toda
neste	naquela	todas
nestes	naquelas	tudo
nisto	naquilo	são
dessa	daquele	com
dessas	daqueles	como
desse	daquela	em
desses	daquelas	atrás
disso	daquilo	acerca
disto	aquele	ser
desta	aqueles	serem
destas	aquela	seja
deste	aquelas	seria
destes	aquilo	seriam

sendo
sejam
será
serão
sido
se
estar
estará
estarão
estaria
estariam
estava
estavam
estando
estado
esteve
estão
por que
por quê
porque
porquê
por
para
pelo
pelos
pela
pelas
que
quem
qual
quais
anteriormente
entre
entretanto
mas
mais
exceto
outro
outros
outra
outras
onde
aonde
logo
resumo
introdução
palavra-chave

palavras-chave
palavras-chaves
conclusão
respectivamente
tal
tais
tanto
tantos
tanta
tantas
etc
etc.
conforme
geralmente
inicialmente
adiante
diante
bem
bom
boa
bons
bens
boas
bastante
bastantes
portanto
consequentemente
conseqüentemente
através
finalmente
pois
juntamente
já
mesmo
mesmos
mesma
mesmas
primeiramente
preferencialmente
sua
suas
seu
seus
tua
tuas
teu
teus

simplesmente
dela
delas
dele
deles
dentro
dentre
apenas
apesar
muito
muitos
muita
muitas
não
sim
necessariamente
se
si
agora
até
após
ainda
assim
somente
ou
nosso
nossos
nossa
nossas
fora
cima
embaixo
abaixo
baixo
exemplo
exemplos
apropriadamente
pessoalmente
pessoa
pessoas
pouco
poucos
pouca
poucas
próximo
próximos
próximas

próxima
então
algo
algum
alguns
alguma
algumas
alguém
lhe
lhes
la
las
lo
los
lá
dois
duas
segundo
segundos
segunda
segundas
três
quatro
cinco
seis
sete
oito
nove
dez
era
eram
for
quando
quanto
quantos
quanta
quantas
cujo
cujos
cuja
cujas
enquanto
ano
anos
ambos
ambas
cada

totalmente
além
aceitável
aceitáveis
aconselhável
atacável
adequadamente
adicionalmente
aleatoriamente
altamente
amplamente
aparentemente
apreciavelmente
aproximadamente
atualmente
basicamente
brevemente
bruscamente
certamente
claramente
comparativamente
completamente
comumente
concomitantemente
conjuntamente
consideravelmente
continuamente
convenientemente
corretamente
cuidadosamente
definitivamente
demasiadamente
detalhadamente
devidamente
diariamente
diferentemente
dificilmente
diretamente
drasticamente
economicamente
efetivamente
especialmente
especificamente
esquemáticamente
essencialmente
esteticamente
eventualmente

evidentemente
exatamente
excessivamente
exclusivamente
experimentalmente
extremamente
facilmente
finamente
fortemente
freqüentemente
frequentemente
fundamentalmente
geograficamente
globalmente
gradualmente
gradativamente
graficamente
habitualmente
historicamente
igualmente
imediatamente
indiretamente
individualmente
inevitavelmente
infelizmente
inteiramente
intimamente
isoladamente
justamente
largamente
lentamente
levemente
ligeiramente
linearmente
localmente
logicamente
majoritariamente
manualmente
matematicamente
meramente
mundialmente
naturalmente
negativamente
nomeadamente
normalmente
notavelmente
novamente

obviamente	sistematicamente	seguida
paralelamente	substancialmente	seguinte
parcialmente	sucessivamente	seguintes
particularmente	suficientemente	só
paulatinamente	tecnologicamente	talvez
perfeitamente	teoricamente	também
periodicamente	termicamente	todavia
plenamente	tipicamente	trás
possivelmente	tradicionalmente	último
posteriormente	unicamente	últimos
praticamente	uniformemente	últimas
precisamente	usualmente	últimas
predominantemente	visualmente	ante
previamente	agradecimentos	contra
principalmente	bibliografia	desde
prioritariamente	bibliografias	perante
profundamente	conclusão	sem
progressivamente	conclusões	trás
proporcionalmente	considerações	dezenas
propriamente	finais	dobro
provavelmente	afim	doze
qualitativamente	ái	duplo
quantitativamente	ali	dupla
quimicamente	aliás	mil
rapidamente	aqui	milheiro
raramente	contudo	milhões
realmente	convém	quarto
recentemente	desde	quarta
regularmente	hoje	vinte
relativamente	nem	-
resumidamente	num	o
rigorosamente	numa	°c
sensivelmente	nunca	å
separadamente	obstante	á
significativamente	quase	a
simultaneamente	seguido	

Para o método híbrido, novos *tokens* (mostrados a seguir) foram inseridos na *stoplist* 1, dando origem à *stoplist* 2.

apresenta	caracterizado	classe
apresentam	caracterizados	classes
atua	caracterizada	compreendendo
atuam	caracterizadas	compreendido

compreendida	constituídos)
compreendidos	depende	-
compreendidas	dependem	:
conhecido	desenvolvido	conceito
conhecida	desenvolvida	conceitos
conhecidos	desenvolvidos	corresponde
conhecidas	desenvolvidas	correspondem
consiste	determinado	define
contém	determinada	definem
contêm	determinados	denominado
em	determinadas	denominada
outras	empregado	denominados
palavras	empregada	denominadas
implica	empregados	feito
implicam	empregadas	feitos
isto	expresso	usado
ou	expressos	usados
seja	formado	figura
por	formada	figuras
exemplo	formados	tabela
tal	formadas	tabelas
utilizado	obtido	quadro
utilizada	obtidos	quadros
utilizados	palavra	fig
utilizadas	palavras	figs
característica	relacionado	metodologia
características	relacionada	metodologias
composto	relacionados	resultados
compostos	relacionadas	resultado
estado	chamado	discussão
matéria-prima	chamada	exemplo
matérias-prima	chamados	exemplos
método	chamadas	ter
métodos	definido	tendo
parte	definida	tido
partes	definidos	tenho
processo	definidas	tens
processos	como	tem
propriedade	expressão	temos
propriedades	expressões	tendes
tipo	se	têm
tipos	entende	tinha
adição	significa	tinhas
chamamos	significam	tinha
constitui	termo	tínhamos
constituem	termos	tínheis
constituído	(tinham

tive
tiveste
teve
tivemos
tivestes
tiveram
tivera
tiveras
tivera
tivéramos
tivéreis
tiveram
teria
terias
teria
teríamos
teríeis
teriam
terei
terás
terá
teremos
tereis
terão
tem
tenha
tenhamos
tende
ser
sendo
sido
sou
és
é
somos
sois
são
era
eras
era
éramos
éreis
eram
fui
foste
foi
fomos

fostes
foram
fora
foras
fora
fôramos
fôreis
foram
seria
serias
seria
seríamos
seríeis
seriam
serei
serás
será
seremos
sereis
serão
haver
havendo
havido
hei
hás
há
havemos
haveis
hão
havia
havia
havia
havíamos
havíeis
havam
houve
houveste
houve
houvemos
houvestes
houveram
houvera
houveras
houvera
houvéramos
houvéreis
houveram

haveria
haverias
haveria
haveríamos
haveríeis
haveriam
haverei
haverás
haverá
haveremos
havereis
haverão
estar
estando
estado
estou
estás
está
estamos
estais
estão
estava
estavas
estava
estávamos
estáveis
estavam
estive
estiveste
estive
estivemos
estivestes
estiveram
estivera
estiveras
estivera
estivéramos
estivéreis
estiveram
estaria
estarias
estaria
estaríamos
estariéis
estariam
estarei
estarás

estará
estaremos
estareis
estarão
existir
existindo
existido
existo
existes
existe
existimos
existis
existem
existia
existias
existia
existíamos
existíeis
existiam
existi
exististe
existiu
existimos
exististes
existiram
existira
existiras
existira
existíramos
existíreis
existiram
existiria
existiriam
existiriam
existirá
existiremos
existireis
existirão
fazer
fazendo
feito
faço

fazes
faz
fazemos
fazeis
fazem
fazia
fazias
fazia
fazíamos
fazíeis
faziam
fiz
fizeste
fiz
fizemos
fizestes
fizeram
fizera
fizeras
fizera
fizéramos
fizéreis
fizeram
faria
farias
faria
fariamos
fariéis
fariam
farei
farás
fará
faremos
fareis
farã

