Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

Extração Automática de Termos de Textos em Português: Aplicação e Avaliação de Medidas Estatísticas de Associação de Palavras.

Maria Fernanda Teline
Aline Maria Pacífico Manfrin
Sandra Maria Aluísio

NILC-TR-03--10

Outubro, 2003

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

Resumo

Neste trabalho, são descritos os passos para a montagem do *corpus* composto por textos da área de Revestimentos Cerâmicos, da revista eletrônica Cerâmica Industrial, que foi utilizado para a avaliação de medidas estatísticas para a extração de candidatos a termos da área de Revestimentos Cerâmicos. O propósito do trabalho foi avaliar o desempenho das medidas, que são utilizadas primariamente para a extração de n-gramas a partir de um *corpus*, na tarefa de extração de terminologia. Para a comparação do desempenho delas, utilizamos como lista de referência os termos contidos no Dicionário de Revestimentos (DiRC)[1] que estavam presentes no *corpus* em questão. Para bigramas não foi possível escolher um dos métodos estatísticos dentre Freqüência, Informação Mútua, *Log-likelihood* e Dice, pois seus resultados apresentaram-se bastante semelhantes. Já para o caso de trigramas, a Freqüência apresentou um resultado melhor do que as medidas Informação Mútua e *Log-likelihood.*

[1] Dicionário em elaboração, que tem como origem o glossário-piloto de revestimentos cerâmicos (Almeida, 2000), coordenado pela Profa. Dra. Gladis Maria de Barcellos Almeida.

Sumário

Índice de Figuras

1. Contextualização

O trabalho terminográfico necessita do auxílio do especialista da área em foco em vários momentos, que vão desde a indicação de fontes de referência de onde se buscam os termos até a validação de candidatos que comporão o dicionário ou glossário e a validação de suas definições. No entanto, nos momentos de extração de candidatos a termos e organização das informações colhidas da área de especialidade, o terminólogo atua sozinho. Um dos momentos mais complexos para o terminólogo é a extração de uma lista de possíveis termos da área por meio das fontes sugeridas pelos profissionais. Espera-se que tal lista seja a mais sucinta possível para que o tempo gasto pelo especialista seja pouco. Por maior que seja a quantidade de material que o terminólogo inclua em suas pesquisas, ele não pode afirmar com precisão se o que ele seleciona é termo ou não, pois ele não participa da realidade da área, nem interage com ela. Por isso que, na fase de escolha dos termos a serem colocados no dicionário, é fundamental a supervisão do profissional especializado.

Os resultados dos métodos de extração de termos, sejam eles manuais ou automáticos, passam pela análise do especialista. A diferença entre as duas abordagens é o tempo dispendido, e a maneira como os métodos são realizados dentro do processo de elaboração de um dicionário ou glossário. Na extração manual, o terminólogo lê todo o *corpus*, e essa leitura está direcionada para a seleção dos possíveis candidatos a termos, sendo que o critério utilizado é o semântico. Na extração automática, a medida escolhida, que raramente é a semântica dada a pouca disponibilidade de ontologias, gera uma lista de candidatos a termos, na maioria das vezes, menos exata do que a do terminólogo, porém, muito mais rapidamente e de forma mais consistente, pois evita-se a falha humana. Pode ser necessária a limpeza da lista de candidatos pelo terminólogo antes da avaliação final do especialista para excluir colocações e siglas da língua geral, nomes próprios e símbolos especiais eventualmente colhidos pela medida.

Segundo Teline et al (2003), a extração automática, além de otimizar o tempo de trabalho na seleção de possíveis candidatos, porque rapidamente elege e classifica os potenciais termos, permite a recuperação fácil do contexto em que os termos foram encontrados, o que não ocorre na manual. Essa característica resolve um outro problema encontrado na elaboração de dicionários especializados, que é a escolha de um melhor

contexto para ser adicionado ao verbete do termo. O trabalho do terminólogo pode também ser facilitado pelo uso de concordanceadores, que são ferramentas que trazem o contexto esquerdo e direito de aparição no *corpus* de uma palavra ou expressão em foco. Porém, diferentemente dos métodos totalmente automáticos, essas ferramentas pressupõem o julgamento do terminólogo na caracterização dos termos.

Nesse trabalho, focalizaremos nossa atenção nos métodos totalmente automatizados para a extração de candidatos e não nas ferramentas de suporte ao trabalho de extração.

2. Objetivo

Nosso trabalho tem como finalidade relatar os procedimentos utilizados para eleger a medida estatística mais eficiente para a extração automática de termos. Entendemos como mais eficiente a medida estatística que elencar um maior número de termos na relação dos candidatos com maiores escores e que dentre os elementos da lista gerada que possuam menores escores poucos ou, de preferência, nenhum sejam termos da área. Focalizamos nossa atenção em quatro medidas: Freqüência, *Log-likelihood*, Informação Mútua e Dice implementados no pacote para a extração de n-gramas chamado NSP (N-gram Statistics Package) e disponível no *site* http://www.d.umn.edu/~tpederse/nsp.html. Vale ressaltar que o domínio específico o qual estamos trabalhando (Revestimentos Cerâmicos), pertence a uma subdivisão de uma área maior, denominada Materiais Cerâmicos, que por sua vez, está inserida em uma macro-área, denominada Engenharia de Materiais (Almeida, 2000).

3. Metodologia

Para nortear a execução dos procedimentos necessários ao alcance de nosso objetivo, consideramos os estudos de Daille (1994; 1996), porém, com algumas especificidades, considerando que o tamanho do *corpus* e a língua são distintos. Também, a definição do tamanho de um termo difere nos dois trabalhos. Daille (1994; 1996) considera o tamanho de um termo como o número de itens principais que ele contém, sendo que itens principais são: substantivos, adjetivos, advérbios, etc. Preposições e determinantes não são considerados itens principais. Dessa forma, os padrões "Substantivo1 PREP/DET Substantivo2" são considerados bigramas. No nosso trabalho, o tamanho de um termo é definido como o número de itens que ele contém. Os padrões acima são considerados trigramas por nós. Os estudos de Daille (1994; 1996) propõem etapas que consistem em:

1. Seleção do *corpus*;
2. Obtenção de uma lista de termos validados pelos profissionais da área de especialidade. Essa lista servirá como referência para a comparação com os termos candidatos gerados automaticamente;
3. Aplicação de regras para formação de *token* e uso de uma *stoplist* no *corpus* escolhido;
4. Geração das listas de n-gramas (aqui trabalhamos com unigramas, bigramas e trigramas), juntamente com suas respectivas freqüências. Utilizamos aqui o pacote NSP, que será melhor explicitado adiante;
5. Elaboração de tabelas em ordem alfabética com os unigramas e bigramas gerados. Para os trigramas essas tabelas não foram elaboradas;
6. Subdivisão das tabelas em candidatos a termos, palavras e siglas da língua geral, nomes próprios e símbolos especiais, como -, *, /;
7. Aplicação das três medidas estatísticas (coeficiente *Log-Likehood*, Informação Mútua e coeficiente de *Dice*), na lista de bigramas;
8. Aplicação das duas medidas estatísticas (coeficiente *Log-Likehood* e Informação Mútua), na lista de trigramas;
9. Análise da medida de freqüência para unigramas e da medida mais eficiente tanto para bigramas quanto para trigramas.

As duas primeiras etapas serão detalhadas nessa seção. As etapas 3 a 8 serão mostradas na Seção 4 e a 9 na Seção 5.

3.1. Seleção do *corpus*

O *corpus* utilizado para avaliar as medidas estatísticas foi montado com artigos que se encontram no *site* da Revista Cerâmica Industrial[2]. 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 para os cálculos das medidas estatísticas utilizadas neste trabalho, deveriam estar no formato texto. Por esta razão, nem todos os textos do *site* foram utilizados, visto que ocorreram alguns problemas no processo de transformação delas 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 destes artigos eram de 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, nosso trabalho foi reavaliado, pois, ao trabalharmos com extração de termos em um *corpus* que contém textos em língua portuguesa, variante brasileira, se inseríssemos textos traduzidos, eliminaríamos nosso critério de coerência com a língua em que o *corpus* está escrito. A retirada desses textos, por outro lado, comprometeria o tamanho de nosso *corpus*, uma vez que estamos trabalhando com métodos estatísticos, sendo estes dependentes, significativamente, do tamanho do *corpus*.

Por esta razão, decidimos entrar em contato com o responsável pela revista para esclarecer se estes textos, depois de traduzidos, eram posteriormente analisados por um especialista da área, e, caso isso ocorresse, não haveria problemas em deixá-los no *corpus*. Como a resposta foi afirmativa, os textos estrangeiros, híbridos e de autoria duvidosa foram novamente incluídos, e além desses já existentes, conseguimos inserir mais 23 artigos (dado

[2] http://www.ceramicaindustrial.org.br/

que o problema de transformação foi parcialmente resolvido), totalizando 164 textos para compor o *corpus* de trabalho.

Para a transformação desses textos para o formato texto, foi utilizada a ferramenta denominada ERTEX (Extracção de Texto de Ficheiros Formatados)[3]. 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 *corpus* 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 *corpus* em palavras é 448352.

Também foi encontrada grande quantidade de erros de digitação e gramaticais, dentre eles, de concordância em gênero e em número e de acentuação. Ainda foi possível perceber que alguns termos encontrados no *corpus* apresentavam hífen e não estavam lematizados, enquanto que na lista de referência obtida pela extração manual esses termos se encontravam não hifenizados e lematizados. Para minimizar os erros gramaticais, foi realizada uma varredura no *corpus* com o auxílio de um processador de textos, buscando corrigir os erros encontrados, podendo, dessa forma, analisar os dados de forma mais precisa. Trabalhamos assim, com um *corpus* pré-processado.

3.2. Obtenção de uma lista de termos de referência

A lista foi gerada a partir de um *corpus* 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 termos (unigramas (354), bigramas (169) e trigramas (151)), que totaliza 748. Os termos que apresentam um número de *tokens* superior a 3 do trabalho de Almeida

[3] http://poloclup.linguateca.pt/ferramentas/extex/

(2000), não serão utilizados neste trabalho. Esses termos se transformaram ou se transformarão em verbetes do DiRC, portanto, já avaliados pelo especialista da área[4].

[4] Todas as nossas análises sempre tomarão como comparação essa lista de referência.

4. Resultados

Antes de apresentar os resultados, faremos uma descrição de um pacote computacional utilizado para a extração de candidatos a termos. Os resultados serão apresentados de acordo com as etapas apresentadas na Seção 3:

4.1. Descrição do pacote NSP

As medidas estatísticas utilizadas estão incorporadas no pacote NSP (N-gram Statistics Package)[5], escrito em Perl. O pacote NSP foi implementado por Ted Pedersen, Satanjeev Banerjee e Amruta Purandare na Universidade de Minnesota, Duluth[6]. 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.

Este pacote é encontrado em diversas versões, e a versão utilizada neste 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, visto que para bigramas foram utilizadas a Informação Mútua, *Log-likelihood* e Dice e para trigramas foram utilizadas a Informação Mútua e *Log-likelihood*.

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 deste 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 destas opções, existem outras que não serão aqui descritas por não terem sido utilizadas, mas podem ser encontradas no pacote NSP.

[5] http://www.d.umn.edu/~tpederse/nsp.html
[6] http://www.d.umn.edu/

Considere que o conteúdo apresentado a seguir pertence ao arquivo texto de entrada (entrada.txt, por exemplo) para “count.pl”

```
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):

```
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, estes bigramas foram listados, considerando que cada *token* é separado pelo sinal “<>”. Depois do último “<>” 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 de dice para bigramas, é utilizado o comando "statistic.pl nome_do_arquivo_da_medida nome_do_arquivo_de_saída.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 de dice, é "statistic.pl dice teste.dice bigrama.txt". O resultado gerado ao se executar esta linha de comando é uma saída similar àquela apresentada anteriormente, acrescentando-se o rank e o escore dos bigramas antes dos 3 outros números. Dessa forma, os bigramas são classificados de acordo com os escores que apresentam. Considerando como entrada (arquivo_de_bigramas.txt) o arquivo de saída gerado anteriormente 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 este arquivo com o anterior, é possível notar que existem dois números adicionais, sendo que o primeiro representa o rank do bigrama, que é obtido a partir do segundo número, que representa o escore do bigrama, que é calculado utilizando-se, neste caso, a medida estatística coeficiente de dice. Dessa forma, os bigramas foram classificados em ordem crescente de seus ranks. Os três números restantes são os mesmos apresentados no arquivo anterior.

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

No momento da geração de unigramas, bigramas e trigramas, foi encontrada uma dúvida quanto à entrada para o programa "count.pl", responsável por gerar a lista de unigramas, bigramas e trigramas juntamente com suas freqüências, que, a princípio, deveria apresentar o formato txt, sabendo que os arquivos do *corpus* se localizavam em pastas dentro de um determinado diretório. No entanto, percebemos que ao se utilizar a opção "--recurse", seria possível acessar os arquivos mesmo dentro de suas pastas.

O pacote NSP, além de produzir todos os unigramas, bigramas e trigramas encontrados no *corpus*, 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 *corpus* 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. Esta 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. Ela será descrita na Seção 4.2.

4.2. Regras de formação de *token* e *stoplist*

Na construção da regra de formação de *tokens*, foi necessário utilizarmos a tabela ASCII estendida, já que o pacote apenas reconhece padrões de formação de *tokens* neste formato. A princípio, palavras hifenizadas também não eram geradas como se encontravam no *corpus* e sim separadas por meio do hífen. Neste 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
[\\w\xb0]|	→ representa o " º " (grau)
[\\w\xc0-\xc5]|	→ representa a letra "a" maiúscula com as acentuações possíveis

[\\w\xc7-\xcf]|	→ representa o “ç”, as letras “e” e “i” com acentuações (maiúsculos)
[\\w\xd1-\xd6]|	→ representa o “ñ” e a letra “o” com acentuações (maiúsculos)
[\\w\xd9-\xdc]|	→ representa a letra “u” maiúscula com acentuações
[\\w\xdf-\xe5]|	→ representa a letra “ß” e a letra “a” minúscula com acentuações
[\\w\xe7-\xef]|	→ representa o “ç”, as letras “e” e “i” com acentuações (minúsculos)
[\\w\xf1-\xf6]|	→ representa o “ñ” e a letra “o” com acentuações (minúsculos)
[\\w\xf9-\xfc])+/	→ representa a letra “u” minúscula com acentuações

Após a execução destas tarefas, o resultado produzido pelas listas de unigramas, bigramas e trigramas ainda não foram satisfatórios, visto que as palavras que apareciam com maior freqüência representava 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 este problema, foi construída uma *stoplist* com estas palavras, a fim de obter uma lista menor e com maior probabilidade de serem termos.

A princípio, havia sido construída uma *stoplist* com preposições, artigos, conjunções e alguns advérbios. No entanto, esta *stoplist* encontrava-se incompleta, visto que não apresentava uma quantidade significativa de preposições e advérbios que apareciam no *corpus*. Também foram incluídas 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 B.

Na primeira linha foi utilizado o operador booleano “or”, pois se apenas um dos *tokens* contidos em algum bigrama presente no *corpus* se encontra especificado nesta 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 caracter ou caracteres em questão como opcionais. Assim, na linha quatro pode-se formar os *tokens* “à” e “às” (neste caso o “s” é opcional). Na linha cinco pode-se formar os *tokens* “a”, “o”, “ao”, “aos” e “os” (neste 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 *corpus* ter sido todo processado para caixa alta.

Para os unigramas, esta 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 esta 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”.

4.3. Geração das listas de unigramas, bigramas e trigramas

Dentre as medidas de associação de palavras encontradas no pacote NSP, estamos utilizando a Informação Mútua, o *Log-Likelihood*, e o coeficiente de Dice, bem como a Freqüência para realizar o levantamento dos candidatos a termos no *corpus*. A Freqüência pode ser calculada para n-gramas, e, neste 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 são:

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

para a geração da lista de unigramas (arquivo unigrama.txt), a partir do diretório corpus~1, utilizando-se a regra de formação de *tokens* (arquivo tk3.pl) e a *stoplist* (arquivo FILEOR.pl)

```
count.pl --token tk3.pl --stop FILEOR.pl --recurse bigrama.txt corpus~1
```

para a geração da lista de bigramas, utilizando-se a regra de formação de *tokens* e a *stoplist*

```
count.pl --ngram 3 --token tk3.pl --stop FILEAND.pl --recurse trigrama.txt corpus~1
```

para a geração da lista de trigramas, utilizando-se a regra de formação de *tokens* e a *stoplist*

4.4. Elaboração e análise de tabelas

Para analisarmos o conteúdo das listas geradas de unigramas e bigramas referentes ao nosso *corpus*, ou seja, se ainda continha nelas candidatos a termos com grande freqüência, dividimos as listas de unigramas e bigramas, separadamente, em tabelas de 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 este 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, fizemos classificações do tipo *palavras e siglas da língua geral, candidatos a termos, marcas publicitárias, nomes próprios,* e *símbolos especiais*, para facilitar nossas análises dos uni e bigramas presentes.

Nossas análises permitem verificar que, tanto no caso de unigramas quanto de bigramas, as palavras de língua geral aparecem em maior número. Em seguida encontramos 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 de língua geral, no que se refere 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*".

4.5. Aplicação das medidas estatísticas de associação de palavras

Segundo Daille (1994), para que possamos comparar os métodos estatísticos, devemos utilizar o critério de separação da lista de unigramas, bigramas e trigramas em classes, sendo que para cada classe foi realizada a intersecção com a lista de referência previamente dividida em unigramas, bigramas e trigramas, gerando histogramas normalizados para unigramas e bigramas.

Para este trabalho, utilizamos os seguintes passos:

1. Utilizar **freqüência** pura para as listas de unigramas, bigramas e trigramas, geradas anteriormente;
2. Organizar os unigramas e bigramas candidatos especificados nas tabelas em ordem decrescente de seus escores;
3. Dividir a classificação dos unigramas em grupos de 10, para formar classes;
4. Dividir a classificação dos bigramas em grupos de aproximadamente 12, para formar classes. Por exemplo: os 12 primeiros da lista formam 1 classe e assim por diante;
5. Fazer a intersecção dos termos da lista de referência (separada em unigramas, bigramas e trigramas), para cada classe dos unigramas e bigramas, para que possamos comparar somente com os termos da lista de referência que aparecem no *corpus*;
6. Montar um histograma para os unigramas (Figura 1) e outro para os bigramas (Figura 2). No caso dos bigramas, efetuamos as divisões da seguinte forma: em cada classe, dividir o número da intersecção por 12 (por exemplo: se uma classe apresenta 6 bigramas em comum com os da lista de referência, e, como a classe apresenta tamanho 12, divide 6 por 12, e por esta razão, os valores das classes no histograma variam entre 0 e 1);
7. Aplicar na lista de bigramas a medida ***log-likelihood***:

```
statistic.pl ll.pm log.ll bigrama.txt
```

para a geração da lista de bigramas (arquivo bigrama.txt) com seus ranks e escores utilizando o coeficiente *log-likelihood*, e repetir os mesmos passos executados para a freqüência (passos 2 a 6). O histograma gerado pode ser visto na Figura 3;

8. Aplicar na lista de bigramas a medida **informação mútua**:

```
statistic.pl tmi.pm infmut.tmi bigrama.txt
```

para a geração da lista de bigramas com seus ranks e escores utilizando a informação mútua, e repetir os mesmos passos executados para a freqüência (passos 2 a 6). O histograma gerado pode ser visto na Figura 4;

9. Aplicar na lista de bigramas o **coeficiente de Dice**:

```
statistic.pl dice.pm dc.dice bigrama.txt
```

para a geração da lista de bigramas com seus ranks e escores utilizando o coeficiente de dice, e repetir os mesmos passos executados para a freqüência (passos 2 a 6). O histograma gerado pode ser visto na Figura 5;

10. Aplicar na lista de trigramas as medidas de ***log-likelihood*** e **informação mútua**:

```
statistic.pl --ngram 3 tmi3.pm infmut3.tmi3 trigrama.txt
```

para a geração da lista de trigramas (no arquivo trigrama.txt) com seus ranks e escores utilizando a informação mútua

```
statistic.pl --ngram 3 ll3.pm log3.ll3 trigrama.txt
```

para a geração da lista de trigramas com seus ranks e escores utilizando o coeficiente *log-likelihood*

11. Dividir a lista de trigramas gerada para estas três medidas em oito classes de 827, e realizar a intersecção entre a lista de referência com cada classe e contar a quantidade de termos que cada medida apresenta em cada classe, e gerar um histograma com os resultados obtidos (Figuras 6, 7 e 8).

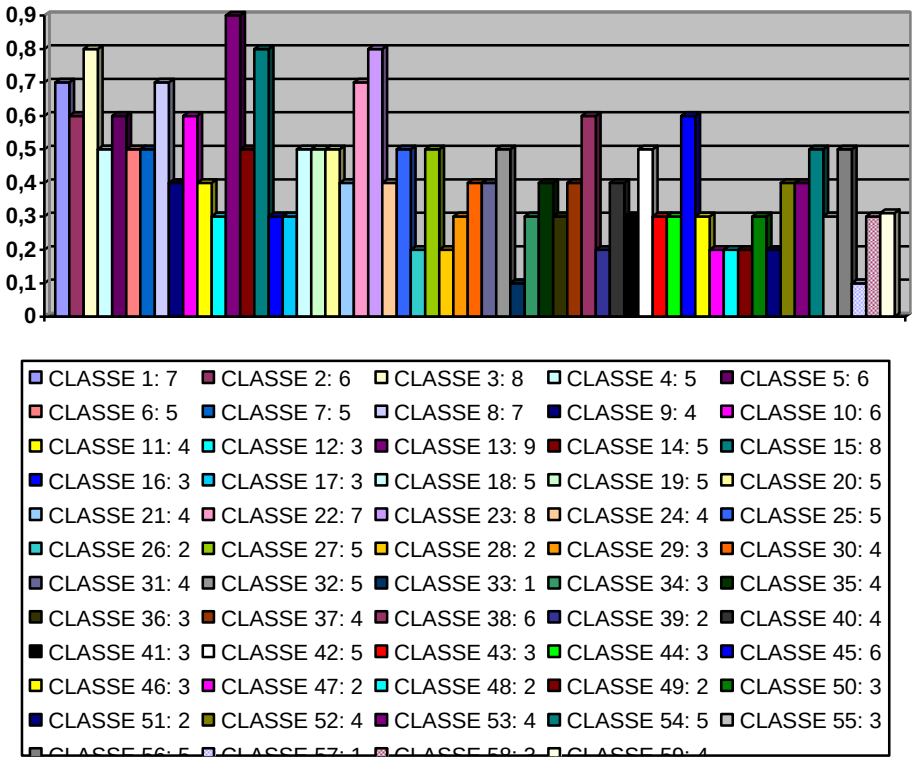

Figura 1. Classes de Candidatos para Unigramas - Freqüência.

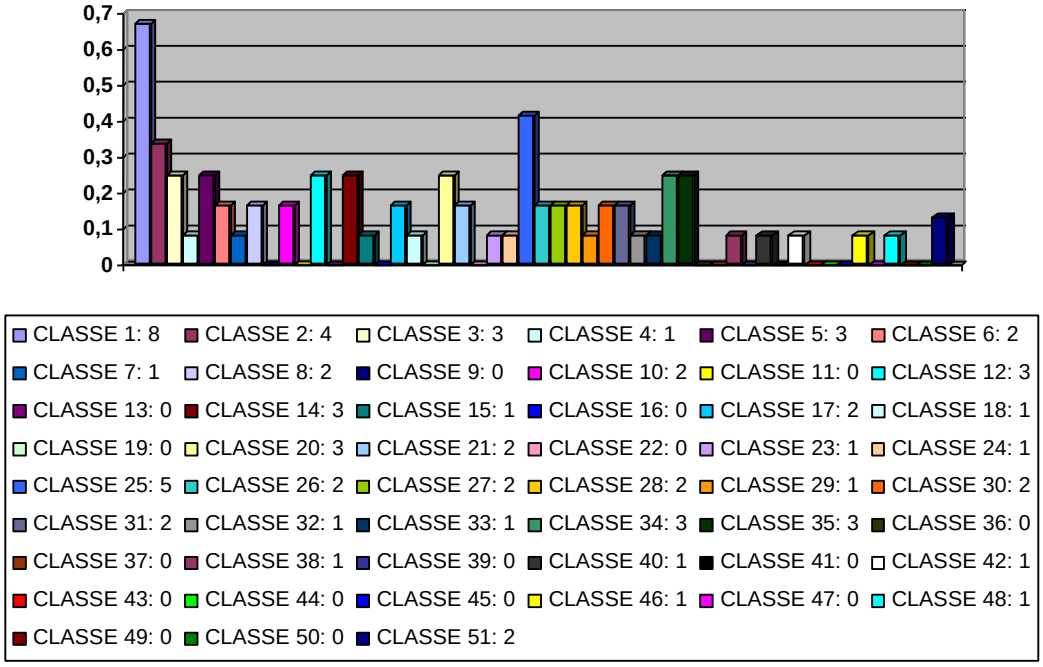

Figura 2. Classes de candidatos para Bigramas - Freqüência.

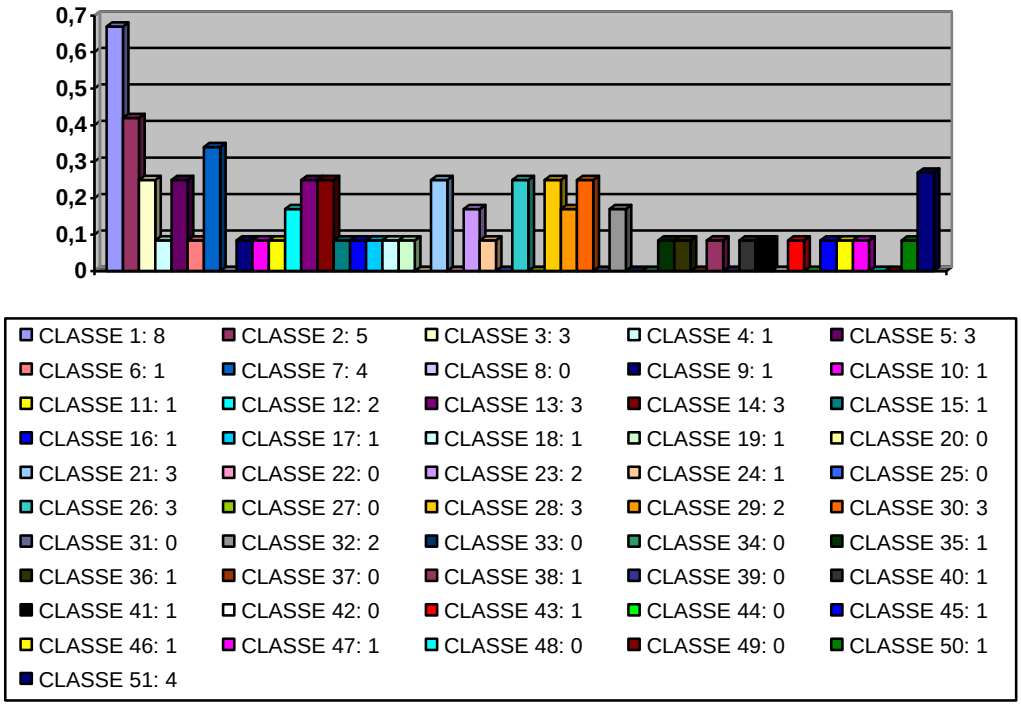

Figura 3. Classes de Candidatos para Bigramas – *Log-Likelihood.*

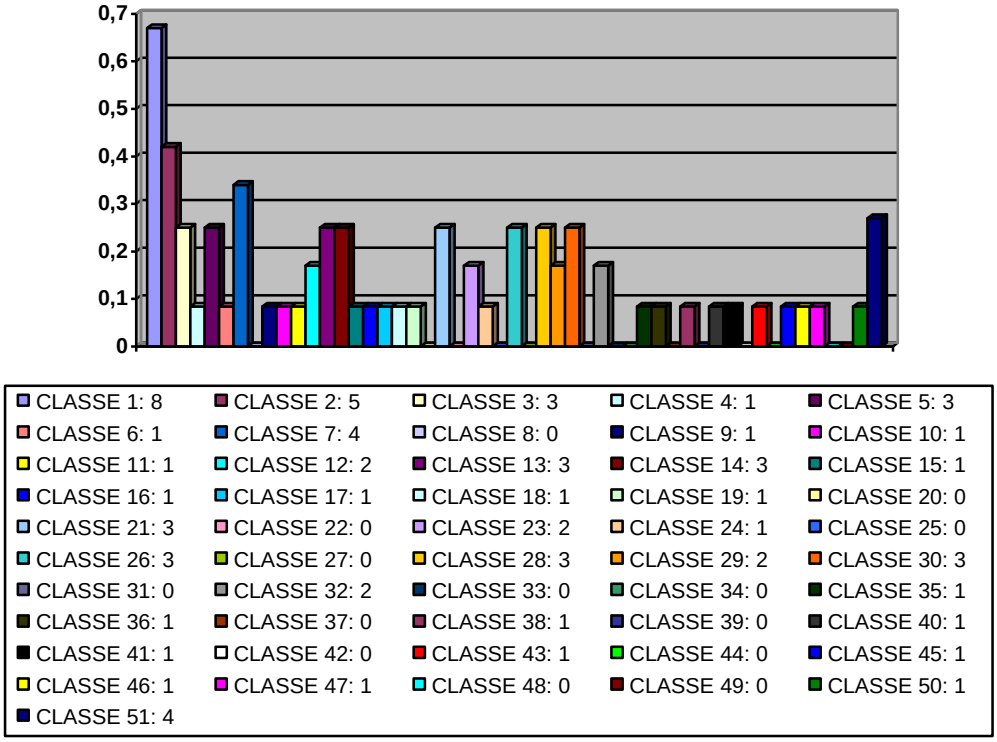

Figura 4. Classes de Candidatos para Bigramas – Informação Mútua

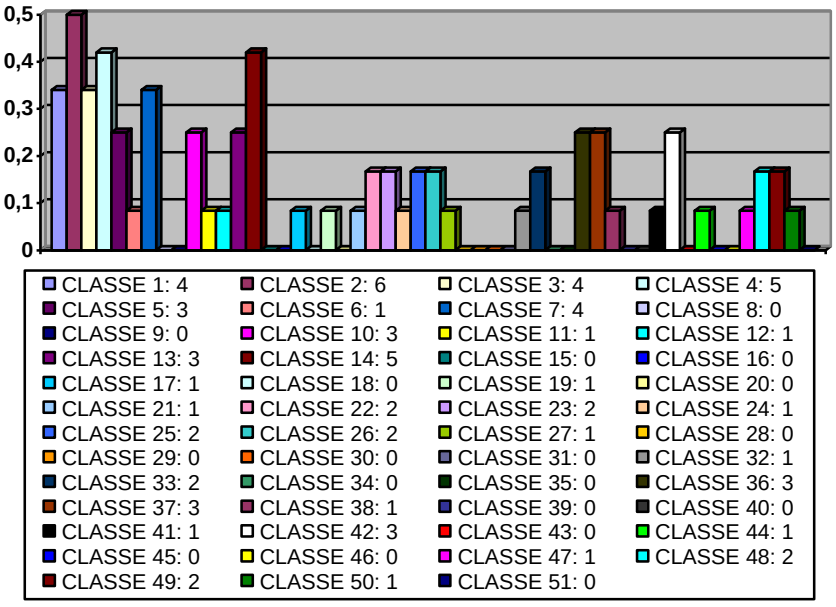

Figura 5. Classes de Candidatos para Bigramas – Coeficiente de Dice

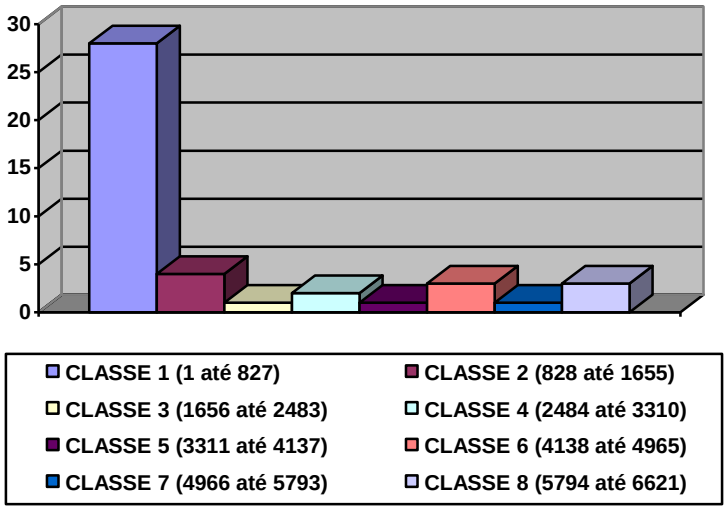

Figura 6. Classes de Candidatos para Trigramas – Freqüência

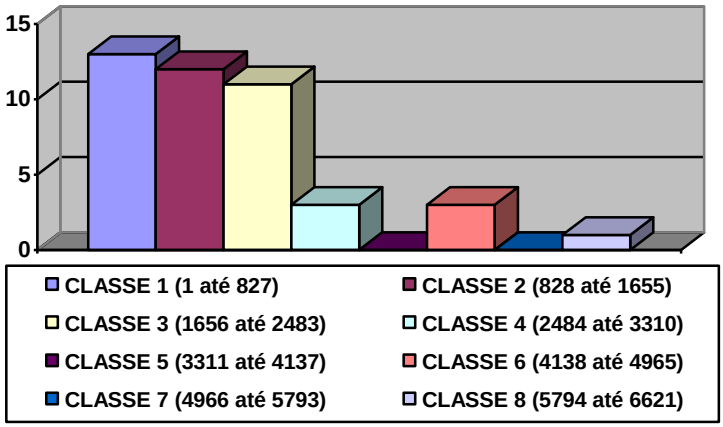

Figura 7. Classes de Candidatos para Trigramas – *Log-likelihood*

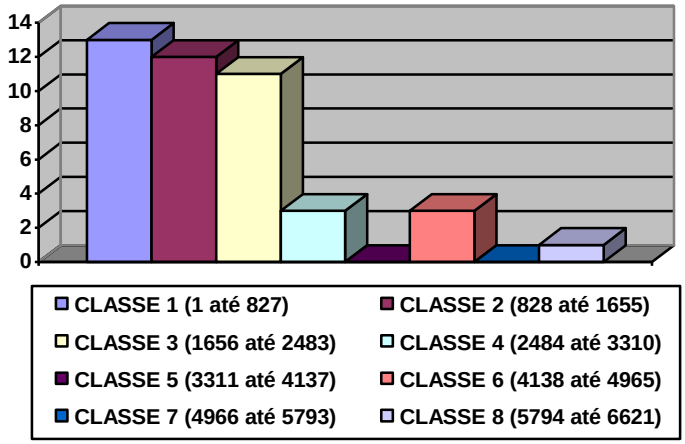

Figura 8. Classes de Candidatos para Trigramas – Informação Mútua

5. Análise das medidas estatísticas

5.1. Medidas estatísticas

5.1.1. Freqüência

Muitos sistemas se utilizam de freqüência, pois certamente ela é a medida mais simples e popular de se encontrar termos em um *corpus*. Se duas palavras ocorrerem muitas vezes juntas, existe, então, uma evidência de que elas apresentam uma função especial. Mesmo para palavras simples, a ocorrência delas com alta freqüência pode indicar que sejam termos de uma área de especialidade.

Nota-se que, apenas selecionar, por exemplo, os bigramas que ocorrem mais freqüentemente em um *corpus* não parece ser muito interessante, pois a maioria deles são pares de palavras funcionais, como artigos e preposições (Manning; Schütze, 1999).

Este método, porém, apresenta mais uma restrição considerando que palavras com baixa freqüência podem também ser termos válidos (Daille, 1996).

5.1.2. Informação mútua

Informação Mútua é uma medida da quantidade de informação que uma variável contém sobre uma outra. A definição de informação mútua é (Pantel; Lin, 2001):

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

onde x e y são palavras ou termos, P(x) e P(y) são, respectivamente, probabilidades de x e y, que correspondem às freqüências das palavras x e y em um *corpus* de tamanho N, e P(x,y) é a probabilidade que as palavras x e y ocorram juntas adjacentemente. Esta medida foi usada inicialmente para extração de colocações. Quando todas as ocorrências de x e y são adjacentes umas às outras, a informação mútua é a maior, deteriorando-se, portanto, em contas de baixa freqüência.

A Informação Mútua depende apenas do tamanho da amostra, da freqüência do bigrama e das palavras individuais no bigrama. Tal medida não apresenta um limite superior para seus escores.

5.1.3. *Log-likelihood*

A medida *log-likelihood*, por se apresentar mais robusta para eventos de baixa freqüência, é utilizada a fim de amenizar o problema da informação mútua quando esta apresenta contagens de baixa freqüência. Considerando que C(x, y) é a freqüência 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 (Pantel; Lin, 2001):

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

onde $k_1 = C(x, y)$, $n_1 = C(x, *)$, $k_2 = C(\neg x, y)$, $n_2 = C(\neg x, *)$, 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 freqüentes que são raramente adjacentes. Este fato, associado ao resultado de que Informação Mútua e *Log-likelihood* deram, no nosso caso, praticamente o mesmo resultado implica que o comportamento dos termos é diferente das colocações que podem aparecer separadas por uma palavra.

Por esta razão, essa característica adicional do *Log-likelihood* (sua razão é alta para 2 termos raramente adjacentes) em relação à Informação Mútua não é útil para o caso de levantamento de termos, já que os termos compostos aparecem juntos. Este fato explica o por quê do resultado das 2 medidas ter sido praticamente o mesmo.

5.1.4. Dice

A medida de associação coeficiente de Dice apresenta uma interpretação similar a Informação Mútua, visto que ele é definido como:

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

onde, assim como acontece com a informação mútua, x e y são palavras ou termos, freq(x,y) representa a freqüência em que as palavras x e y ocorrem juntas adjacentemente, e freq(x) e freq(y) são, respectivamente, freqüências de x e y em um *corpus* de tamanho N.

Esta 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 de Dice depende apenas da freqüência do bigrama e das palavras do bigrama. Diferentemente do que ocorre com a informação mútua, esta medida não depende do tamanho da amostra[7].

5.2. Análise dos resultados das medidas estatísticas

Para unigramas foi realizado apenas o cálculo de freqüência, visto que o pacote NSP apresenta somente esta opção. A lista de unigramas gerada a partir do *corpus*, tendo sido aplicada a regra de formação de *token* e a *stoplist*, apresenta 235554 palavras, sendo 18320 não repetidas. Com a utilização desta lista foi feito um levantamento dos candidatos a termos (594) nas tabelas, como comentado na Seção 3, e então foi feita a intersecção com a lista de referência (354), obtendo um total de 252 termos. A lista de candidatos, ordenada em ordem decrescente de freqüência, foi dividida em 58 classes de 10 unigramas e 1 classe com 14 (veja Figura 1). Dentro de cada classe foram marcados os termos em comum com a lista de referência, para descobrirmos se a maioria dos termos se concentrava nas primeiras classes (com maior freqüência). No entanto, notou-se que os termos se encontram bem espalhados pelas classes, não sendo possível afirmar que quanto maior a freqüência, maior a probabilidade de termos aparecerem, pelo menos no domínio em análise, e no *corpus* específico dele.

Já para bigramas, foram calculadas as seguintes medidas: freqüência, informação mútua, *log-likelihood* e coeficiente de dice. A lista de bigramas gerada a partir do *corpus*, tendo sido aplicada a regra de formação de *token* e a *stoplist*, é constituída por 81695 pares de palavras, sendo 52313 não repetidos. Com a utilização desta lista foi feito um levantamento dos candidatos a termos (615) nas tabelas (Seção 3) e então foi feita a intersecção com a lista de referência (169), obtendo um total de 74 termos. A lista de candidatos, ordenada em ordem decrescente do escore de cada medida, foi dividida em 50 classes de 12 bigramas e 1 classe de 15 (veja Figuras 2, 3, 4 e 5). Dentro de cada classe foram marcados os termos em comum com a lista de referência, para descobrirmos se a

[7] http://www.d.umn.edu/~tpederse/Group01/bsp.txt

maioria dos termos se concentrava nas primeiras classes (com maior escore). No entanto, notou-se, para as quatro medidas aplicadas aos bigramas que os termos se encontram bem espalhados pelas classes, não implicando na ocorrência da maior concentração de termos nas primeiras classes (que apresentam maior escore), com exceção das duas primeiras classes nas medidas Freqüência, Informação Mútua e *Log-likelihood*. A fim de comparar essas 4 medidas, a lista com as 51 classes de cada uma dessas medidas foi primeiramente dividida ao meio, obtendo na parte superior da divisão (até a classe 25), em que os termos aparecem com maior escore, 47, 47, 50 e 48 termos para *log-likelihood*, informação mútua, dice e freqüência. Considerando que a diferença encontrada entre as medidas é mínima, a lista com as classes foi dividida ainda em quatro partes. Através dessas divisões, percebeu-se que na primeira parte (até a classe 12) foram obtidos 30, 30, 32 e 29 termos para *log-likelihood*, informação mútua, dice e freqüência. Novamente, o resultado obtido não foi satisfatório para eleger uma das 4 medidas como a mais eficiente para a extração de termos, o que levou à realização de mais uma divisão da lista em oito partes, sendo que na primeira parte foram obtidos 21, 21, 23 e 21 termos para *log-likelihood*, informação mútua, dice e freqüência. Conclui-se com isso que, não foi possível eleger nenhuma dessas medidas, já que apresentaram resultados semelhantes, diferentemente do que ocorre no trabalho de Daille (1994) para o domínio de telecomunicações, em que a medida eleita foi o *log-likelihood*.

Para trigramas, foi realizado o cálculo da freqüência, informação mútua e *log-likelihood* (Figuras 6, 7 e 8). Como o *corpus* apresenta o tamanho de 448352 palavras, foi decidido realizar um corte, seguindo o mesmo critério que Daille (1996), para os trigramas que apresentassem freqüência igual ou inferior a 4. Tendo sido gerada a lista de trigramas para a freqüência, informação mútua e *log-likelihood*, foi realizada a intersecção da mesma (6621 trigramas diferentes) com a lista de referência de 151 termos (trigramas), obtendo-se um total de 43 termos em comum. Para realizarmos a comparação entre as 3 medidas acima, a lista de cada uma dessas medidas foi dividida em oito classes de 827. Embora o número de termos em comum seja pequeno, a freqüência apresentou uma maior concentração de termos nas duas primeiras classes, podendo ser eleita a melhor medida.

É importante notar que o processo de análise das medidas para unigramas e bigramas é diferente daquele utilizado para trigramas. Para o primeiro caso, as listas de

unigramas e bigramas foram analisadas por um pesquisador lingüista que trabalhou durante três anos na área de especialidade de Revestimentos Cerâmicos, para a construção do glossário-piloto de revestimentos cerâmicos de Almeida (2000). Com a análise das listas geradas, os candidatos a termos foram separados de outras palavras com o auxílio das tabelas comentadas na Seção 3. As classes mencionadas anteriormente foram, então, construídas a partir dessa separação, utilizando-se somente os unigramas e bigramas candidatos. Já para o segundo caso, análise das medidas estatísticas para trigramas, não foi feito o levantamento dos candidatos a termos, sendo que as classes foram construídas a partir dos trigramas encontrados no *corpus* com freqüência superior a 4.

6. Considerações Finais

Tendo realizado a comparação entre os métodos estatísticos *Log-Likelihood*, Informação Mútua, Coeficiente de Dice e Freqüência para bigramas, foi possível perceber que os resultados apresentados por eles mostraram-se bastante semelhantes, não sendo possível eleger um deles para a extração automática dos termos a partir de textos no domínio de Revestimentos Cerâmicos.

Essa semelhança apresentada talvez se explique pela escolha domínio na qual os termos de referência estão contidos, pois este trabalho utilizou, para a análise das medidas estatísticas aplicadas a bigramas, os mesmos critérios adotados por Daille (1994), com algumas adaptações em relação ao tamanho do *corpus*, do número de candidatos e da lista de referência. No entanto, Daille (1994), utilizando um *corpus* formado por textos do domínio de telecomunicações, conseguiu eleger a medida estatística *Log-Likelihood*.

É possível dizer que não conseguimos eleger nenhuma medida estatística, diferentemente de Daille (1994), em razão de esta ter definido o tamanho do termo de forma diferente da definição utilizada por nós. Para Daille (1994), os padrões "Substantivo1 PREP/DET Substantivo2" são considerados bigramas, enquanto que em nosso trabalho, esses padrões são considerados como trigramas. Por este motivo, os resultados obtidos em nosso trabalho se apresentaram diferentes daqueles encontrados em Daille (1994).

Por outro lado, considerando a análise do comportamento das medidas *Log-Likelihood*, Informação Mútua e Freqüência para a extração de termos - trigramas, a Freqüência apresentou um resultado melhor do que as outras duas medidas, já que apresentou um maior número de termos em grupo onde os escores encontrados são mais significativos.

Com relação à análise das medidas estatísticas aplicadas a bigramas e trigramas, é importante notar que para o primeiro caso houve uma interferência humana, com um certo conhecimento no domínio de Revestimentos Cerâmicos, no levantamento dos candidatos a termos. Porém, para os trigramas o processamento foi totalmente automático, e o levantamento de candidatos não foi realizado, mas somente foram construídas classes a partir dos trigramas encontrados no *corpus* e, nestas classes, foi realizada a intersecção com a lista de referência.

Referências Bibliográficas

Almeida, G.M.B. (2000) Teoria Comunicativa da Terminologia: uma aplicação. Araraquara, vol I. Tese (Doutorado em Lingüística e Língua Portuguesa) – Faculdade de Ciências e Letras, Campus de Araraquara, Universidade Estadual Paulista.

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.

Manning, C.; Schütze, H. (1999) Collocations, In: Foundations of Statistical Natural Language Processing, p. 141-77, MIT Press, Cambrige.

Pantel, P.; Lin, D. (2001) A statistical corpus-based term extractor, In: Stroulia, E. and Matwin, S. (Ed.), AI 2001, Lecture Notes in Artificial Intelligence, Springer-Verlag, p. 36–46.

Teline, M.F; Almeida, G.M.B; Aluísio, S.M. (2003) Extração Manual e Automática de Terminologia: Comparando abordagens e Critérios, a ser publicado em: I Workshop em Tecnologia da Informação e da Linguagem Humana (I Til), São Carlos, SP.

Tiedemann, J. (1997) Automatical Lexicon Extraction from Aligned Bilingual Corpora, Diploma thesis, Magdeburg.

Apêndice A

A. Lista de Referência

abóbada	abrasão profunda
abrasão superficial	absorção de água
acabamento	ácido húmico
ácido tânico	aderência
aditivo	aerógrafo
agalmatolito	agitador
aglomeração	aglomerado
aglomerante	aglutinante
agregado	albita
álcali	alcalinidade em defloculantes
alimentação	alimentador
alumina	alumina alfa
alumina calcinada Al0	amarelo
amarelo marron	amolecimento
amostra	amostra compactada
amostra compactada e queimada	amostra granulada
amostra in natura	análise de fase cristalina
análise de frita e esmalte	análise granulométrica
análise granulométrica por peneiramento	análise granulométrica por sedimentação
análise granulométrica por sedimentação a laser	análise granulométrica por sedimentação por raios-X
análise química	análise química de argila
análise química semiquantitativa em matérias-primas diversas	análise racional quantitativa
análise térmica simultânea	anéis de retenção
anortita	antiderrapante
antiespumante	antifungo
anverso	aplicação a seco de pó de vidro
aplicação de pasta com grande espessura com tela rígida	aplicação especial de corante solúvel
aplicador de granilha	aplicador e aspirador de granilha
aquecimento	areia
areia de zircônio	areia feldspática
argila	argila branca
argila caulinítica	argila clorítica
argila de baixo conteúdo de carbonato	argila de cor branca
argila de cor vermelha	argila de elevado conteúdo de carbonato
argila de médio conteúdo de carbonato	argila esmectítica
argila fundente	argila gorda
argila haloisítica	argila ilítica
argila magra	argila montmorilonítica
argila paligorsquita	argila plástica
argila refratária	argila refratária aluminosa
argila refratária sílico-luminosa	argila sepiolita
argilito fundente	argilomineral
armazenagem	arqueamento
aspecto superficial	atomização
atomizador	avaliação do acoplamento cerâmica-esmalte
avaliação do acoplamento massa-esmalte e variação linear com temperatura	avaliação microscópio óptico

azul	azul cobalto
azulejo	baixa resistência mecânica
balancim	ball clay
barita	barra
basalto	bege
bentonita	bentonita cálcica
bentonita sódica	bico
bico pulverizador	biqueima
biqueima rápida	biqueima tradicional
biscoito	bissilicato de chumbo
bitola	bola
bolha negra	bomba
bomba de barbotina	bombatura
boquilha	bordura
borossilicato de chumbo	brancura
brilhante	brilho
britador	britador de martelo
britagem	cabeçote
cabine de esmaltação	cabine de discos
caco	calcário
calcário dolomítico	calcimetria
calcinação	calcita
calcografia	calibre
camada de aplicação	câmara de entrada
câmara móvel	câmaras de pré-secagem
campana	cantoneira
caoboxi-metil-celulose	caolin
capacidade bactericida	caracol
característica dimensional	carbonato
carbonato de cálcio	carbonato de magnésio
carbono total em argilas	carboxi-metil-amido
carga de ruptura	carregamento
carro de prensagem	carro de queima
casagrande	cascata
caulim	caulinita
china clay	ciclo de impressão
ciclo de queima	ciclo de secagem
classes de abrasão superficial	classificação
classificador	clorita
coeficiente de atrito	coeficiente de dilatação
cola para granilha	colagem
colorido	colorifício cerâmico
cominuição	compactação
composição	composição granulométrica
concavidade	concentração de sólidos
conformação	conformação a úmido
contração de queima	controlador de calibre
coordenada cromática	cor de queima
cor in natura	coração negro
corante	corante micronizado de alta dispersão
córindon	córindon artificial
corpo-de-prova	cortador
cortina contínua	cotto
Creta Print	cristalina

cristalização	cromo
curva de compactação	curva de defloculação
curva de queima	curvatura
curvatura central	curvatura lateral
decalcomania	decoração
decoração com sais solúveis	decorador
decoradora	decoradora serigráfica rotativa
defloculação	defloculante
deflocular	densidade
densidade aparente	densidade aparente versus pressão
densidade real de sólidos	densímetro
desaeração	descansar a massa
descarga de massa	descarga do ciclone
descarga do filtro	descarte
desconchamento	desferrização
desintegrador	despoeiramento
dessecador	destacamento por estufamento
destorroador	destorroamento
desvitrificação	desvitrificar
dextrina	difração de raios-X
dilatação higroscópica	dilatação térmica
dilatometria	dimensional
diopsídio	disco
dispersante	dispersar
distribuição granulométrica	dolomita
dosador	dosagem
dureza ao risco	duro
efeito oxidado	eflorescência
eletrofusão	elevador de caneca
empeno	empilhador
empilhamento	emulsão
encaixotamento	engobadeira
engobe	engobe de cobertura
engobe de muratura	engobe de proteção
engobe impermeabilizante	engobe refratário
engobe selador	ensilagem
envelhecer a massa	envelhecimento
enxofre	EPU
escorrimento	esfoliação
esfumadora	esmaltação
esmaltação	esmaltação centrada
esmaltação em baixo relevo	esmaltação estruturada
esmaltação por campana	esmalte
esmalte acetinado	esmalte brilhante
esmalte ceroso	esmalte colorido
esmalte cristalina	esmalte cristalizado
esmalte cru	esmalte de monoporosa
esmalte fritado	esmalte mate
esmalte mole	esmalte opaco
esmalte para biqueima rápida	esmalte para biqueima tradicional
esmalte para monoqueima	esmalte polido
esmalte semibrilhante	esmalte transparente
esmectita	espalhamento
espátula	espatuladora

espatulatriz	espessante
espessura de camada	espodumênio
esquadro	estado cru
estado verde	estampo
estampo espelho	estampo penetrante
esteira com pressão lateral	esteira de transporte
esteira transportadora	estiramento
expansão de queima	expansão linear
expansão por umidade	expansão térmica
extrusão	extrusora
faixa	falta de acoplamento do esmalte
fase vítrea	feldspato
feldspato flotado	feldspato potássico
feldspato sódico	feldspato-anortita
fervura 05 peças	festone
fileira	filete
filito	filtro de mangas
fire clay	fissura
fixador	flexão a três pontos
flexografia	flint clay
floculante	fluidificante
fluorescência de raios-X	fluorita
fonolito	força de impressão
forno	forno a rolos
forno de queima rápida	forno elétrico
forno monoestrado de rolos	forno túnel
fotolito	friso
frita	frita alcalina
frita branca	frita cristalina
frita de biqueima	frita monoporosa
frita opaca	fundente
granilha	granulação
granulação a seco	granulação a úmido
granulador	grânulo
granulometria	grão
grau de compactação	grau de moagem
gravimétrico	grelha
grês	grês não esmaltado não polido
grês polido	grês porcelanato
grês porcelanato esmaltado	gretamento
grupo das peças estruturadas	GST
gunitagem	haloisita
hematita	hidratação
hidróxido de alumínio	homogeinezação
ilita	ímã-diferença de peso
imantável	imersão em mercúrio
impermeabilizante	impressora flexográfica
índice de plasticidade e limite de liquidez	insert
ISO 10545	jateamento
jato de esmalte	ladrilho granito
lança	laranja
lasca	lascamento
ligante	lignossulfonato
limpabilidade	linha de escolha

linha de esmaltação	linha e ondulação
lubrificante	luneta
lustre	maçarico
magnesita	mangote
manta refratária	máquina de escolha
máquina de espatular	máquina serigráfica
máquina serigráfica circular	máquina serigráfica plana
máquina serigráfica rotativa	máquina serigráfica rotativa com limpador
marca d'água	mármore
mármore travertino	marmorizado
maromba	maromba a vácuo
marron avermelhado	marron escuro
martelo recambiável	massa
massa cerâmica	matéria orgânica
mate	matização
matriz	matriz serigráfica
maturação	medida dimensional
meio de moagem	mesa de escolha visual
mesa de serigrafia	metamerismo de cor
mica	micra
microfissura	micronização
micronizador	microssílica
mineral	mistura
moagem	moagem a seco
moagem a úmido	moagem primária
módulo de resistência à flexão	mohs
moinho	moinho contínuo
moinho de bolas	moinho de martelo
moinho de martelos fixos	moinho excêntrico
moinho intermitente	moinho pendular
molde	mole
monocalibre	monoporosa
monoqueima	montmorilonita
mosaico cerâmico	mosaico esmaltado
moscovita	muratura
nefelina sienito	nitrato de bismuto
nitrato de cobalto	ocografia
ocre	opacidade
opacificante	opacificante de zircônio
opaco	ortoclásio
ortogonalidade	óxido cromóforo
óxido de alumínio	óxido de cobalto
óxido de ferro	óxido de magnésio
óxido de silício	óxido de zinco
painel lateral	paletizador
palets	paligorsquita
paralelismo	partícula
pasta serigráfica	pastilha
pastilha de porcelana	pastilha de vidro
pastilha esmaltada	peça
PEI	pêndulo
peneira	peneira circular
peneira malha	peneira vibratória
peneira vibratória circular	peneiramento

peneiramento a seco	peneiramento a úmido
peneiramento via seca	peneiramento via úmida
perda ao fogo	perolizador
pesagem	PH
picnometria	pigmento cerâmico
pirômetro	piso
piso extrudado	placa cerâmica
placa cerâmica para revestimento	placa cerâmica polida
placa cerâmica retificada	planaridade
plasticidade	plastificante
pó atomizado	polido
polimento	ponto de orvalho
porosidade	potenciométrico
pré-misturador	prensa
prensa hidráulica	prensa isostática
prensa isostática de dupla ação	prensa rotativa
prensagem	prensagem uniaxial
pressão de prensagem	preto
produto acabado	pseudoplasticidade
pulverização	punção inferior
punção isostático	punção isostático de dupla ação
punção superior	quarta queima
quartzito	quartzo
queima	queimador
rampa	rampa de aquecimento
rampa de resfriamento	raspa correia
refratariedade	relevo
relevo de prensa	reologia
requeima	resfriamento
resíduo bruto	resíduo em malha
resistência à abrasão	resistência à abrasão profunda
resistência à abrasão superficial	resistência à flexão
resistência à flexão após queima	resistência ao ataque de escória
resistência ao ataque químico	resistência ao choque térmico
resistência ao desgaste mecânico	resistência ao gelo
resistência ao gretamento	resistência ao manchamento
resistência ao rejunte colorido	resistência mecânica
resistência mecânica a verde	resistência química esmaltados
resistência química não esmaltados	retentor mecânico
retificação	retitude lateral
retração de queima	retração de queima linear
retração de queima volumétrica	retração de secagem
revestimento bactericida	revestimento cerâmico
revestimento de parede	revestimento para fachada e piscina
rodapé	rosa
Rotocolor	rugosidade superficial
sal solúvel	secador
secador contínuo	secador estático
secador horizontal	secador modular
secador rápido a rolos	secador resfriador de argila
secador semicontínuo	secador vertical a rolos
secagem	secagem prévia artificial
secagem prévia natural	semigrês
semipolido	semiporoso

sepiolita	sericita
serigrafia	silagem
sílex	sílica
silicato	silicato de sódio
silicato de sódio anidro	silicato de sódio hidratado
silicato de sódio penta-hidratado	silicato de zircônio
silo	sincronizada
sistema de aplicação de água e cola	sistema de captação de pó
soda	sulfato de bário
superfície	suporte
suporte queimado	suspensão
taguá	talco
tamização	tanino
tanque	tanque agitador
tanque de alimentação	tanque de diluição
tardoz	tela serigráfica
telado	telado de pastilha
telado de pequenos formatos	temperatura de acoplamento
temperatura de amolecimento	temperatura de maturação
temperatura de queima	temperatura de transição vítrea
tempo de secagem	teor de carbonato em argila
teor de quartzo livre em argila	teor de umidade
terceira queima	termopar
TG/DTA	tinta serigráfica
tiragem	titânia
tixotropia	tonalidade
torello	torres de lavagem
TOT	tozeto
trabalhabilidade	trabalhável
trabalho térmico	trança
transparência	transparente
transportador de correia	transportador helicoidal
trapézio	triagem
trinca	trinca de resfriamento
trituração	trituração primária
triturador	turbo-secador
turquesa	vagoneta de queima
variação de espessura	variação de tonalidade
vasca	veículo auto-fixante
veículo de terceira queima	veículo especial
veículo hidrossolúvel	veículo oleoso
veículo resinado	veículo serigráfico
vela	venturi scrubber
verde	vermelho
vermiculita	verso
véu	via a laser inferior a 75 mícrons
via seca	via úmida
via úmida capacidade 5 litros	via úmida com barbotina defloculada
vibratório	vidrado
vidrado composto	vidro
viscosidade	viscosímetro
viscosímetro com torque controlado	visual
visual com foto microscópio óptico	vitrificar
vitro-cerâmico	volumetria

wollastonita	zircônia
zirconita	zona de aquecimento
zona de queima	zona de resfriamento
zona de sinterização	zona não esmaltada

Apêndice B

B. *Stoplists*

@stop.mode=OR
/^E$/
/^É$/
/^À(S)?$/
/^A?O?S?$/
/^N?D?E?I?(SS)?(ST)?A?E?O?S?$/
/^N(O)?A?S?$/
/^N(EL)(A)?(E)?S?$/
/^N?D?A?(QUEL)?(QUIL)?A?E?O?S?$/
/^D(E)?A?O?S?$/
/^U(M)?(N)?(A)?(S)?$/
/^NÓS$/
/^EU$/
/^TU$/
/^EL(A)?E?S?$/
/^VOCÊ(S)?$/
/^SOB(RE)?$/
/^SOBRETUDO$/
/^DEPOIS$/
/^DURANTE$/
/^T(O)?U?D(O)?A?S?$/
/^SÃO$/
/^COM$/
/^COMO$/
/^EM$/
/^ATRÁS$/
/^ACERCA$/
/^SE(R)?(EM)?(JA)?(NDO)?(IA)?Á?(ÃO)?M?$/
/^S(IDO)?(ENDO)?$/
/^SE$/
/^EST(AR)?(AVA)?A?N?(DO)?(EVE)?Á?(ÃO)?(E)?(IA)?M?$/
/^(POR QUE)?(POR QUÊ)?(PORQUE)?(PORQUÊ)?$/
/^P(OR)?(ARA)?(EL)?A?O?S?$/
/^QUE(M)?$/
/^QUA(L)?(IS)?$/
/^ANTERIORMENTE$/
/^ENTRE$/
/^ENTRETANTO$/
/^MAS$/
/^MAIS$/
/^EXCETO$/
/^OUTR(O)?A?S?$/
/^(A)?ONDE$/
/^LOGO$/
/^RESUMO$/
/^INTRODUÇÃO$/
/^(PALAVRA-CHAVE)?(PALAVRAS-CHAVE)?(PALAVRAS-CHAVES)?$/
/^CONCLUSÃO$/
/^RESPECTIVAMENTE$/
/^TA(L)?(IS)?$/
/^TANT(O)?A?S?$/
/^ETC(.)?$/
/^CONFORME$/
/^GERALMENTE$/
/^INICIALMENTE$/
/^ADIANTE$/
/^DIANTE$/
/^B(E)?O?A?M?N?S?$/
/^BASTANTE(S)?$/
/^PORTANTO$/
/^CONSE(QUENTEMENTE)?(QÜENTEMENTE)?$/
/^ATRAVÉS$/
/^FINALMENTE$/
/^POIS$/
/^JUNTAMENTE$/
/^JÁ$/
/^MESM(O)?(A)?(S)?$/
/^PRIMEIRAMENTE$/
/^PREFERENCIALMENTE$/
/^S?T?(UA)?(EU)?(S)?$/
/^SIMPLESMENTE$/
/^DEL(A)?E?S?$/
/^DENTRO$/
/^DENTRE$/
/^APENAS$/
/^APESAR$/
/^MUIT(O)?A?S?$/
/^NÃO$/
/^SIM$/

/^NECESSARIAMENTE$/
/^SE$/
/^SI$/
/^AGORA$/
/^ATÉ$/
/^APÓS$/
/^AINDA$/
/^ASSIM$/
/^SOMENTE$/
/^OU$/
/^NOSS(O)?A?S?$/
/^FORA$/
/^CIMA$/
/^E?A?M?BAIXO$/
/^EXEMPLO(S)?$/
/^APROPRIADAMENTE$/
/^PESSOALMENTE$/
/^PESSOA(S)?$/
/^POUC(O)?A?S?$/
/^PRÓXIM(O)?A?S?$/
/^ENTÃO$/
/^ALG(U)?O?É?M?N?A?S?$/
/^L(A)?(HE)?O?S?$/
/^LÁ$/
/^D(OI)?(UA)?S$/
/^SEGUND(A)?O?S?$/
/^TRÊS$/
/^QUATRO$/
/^CINCO$/
/^SEIS$/
/^SETE$/
/^OITO$/
/^NOVE$/
/^DEZ$/
/^ERA(M)?$/
/^FOR$/
/^QUANDO$/
/^QUANT(O)?A?S?$/
/^CUJ(O)?A?S?$/
/^ENQUANTO$/
/^ANO(S)?$/
/^AMB(OS)?(AS)?$/
/^CADA$/
/^TOTALMENTE$/
/^ALÉM$/
/^ACEITÁ(VEL)?(EIS)?$/
/^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$/
/^ESQUEMATICAMENTE$/
/^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$/
/^PARALELAMENTE$/
/^PARCIALMENTE$/
/^PARTICULARMENTE$/
/^PAULATINAMENTE$/
/^PERFEITAMENTE$/
/^PERIODICAMENTE$/
/^PLENAMENTE$/
/^POSSIVELMENTE$/
/^POSTERIORMENTE$/
/^PRATICAMENTE$/
/^PRECISAMENTE$/
/^PREDOMINANTEMENTE$/
/^PREVIAMENTE$/
/^PRINCIPALMENTE$/
/^PRIORITARIAMENTE$/
/^PROFUNDAMENTE$/
/^PROGRESSIVAMENTE$/
/^PROPORCIONALMENTE$/
/^PROPRIAMENTE$/
/^PROVAVELMENTE$/
/^QUALITATIVAMENTE$/
/^QUANTITATIVAMENTE$/
/^QUIMICAMENTE$/
/^RAPIDAMENTE$/
/^RARAMENTE$/
/^REALMENTE$/
/^RECENTEMENTE$/
/^REGULARMENTE$/
/^RELATIVAMENTE$/
/^RESUMIDAMENTE$/
/^RIGOROSAMENTE$/
/^SENSIVELMENTE$/
/^SEPARADAMENTE$/
/^SIGNIFICATIVAMENTE$/
/^SIMULTANEAMENTE$/
/^SISTEMATICAMENTE$/
/^SUBSTANCIALMENTE$/
/^SUCESSIVAMENTE$/
/^SUFICIENTEMENTE$/
/^TECNOLOGICAMENTE$/
/^TEORICAMENTE$/
/^TERMICAMENTE$/
/^TIPICAMENTE$/
/^TRADICIONALMENTE$/
/^UNICAMENTE$/
/^UNIFORMEMENTE$/
/^USUALMENTE$/
/^VISUALMENTE$/
/^AGRADECIMENTO(S)?$/
/^BIBLIOGRAFIA(S)?$/
/^CONCLUS(ÕES)?(ÃO)?$/

/^CONSIDERAÇÕES$/
/^FINAIS$/
/^AFIM$/
/^AÍ$/
/^ALI$/
/^ALIÁS$/
/^AQUI$/
/^CONTUDO$/
/^CONVÉM$/
/^DESDE$/
/^HOJE$/
/^NEM$/
/^NUM(A)?$/
/^NUNCA$/
/^OBSTANTE$/
/^QUASE$/
/^SEGUID(A)?(O)?$/
/^SEGUINTE(S)?$/
/^SÓ$/
/^TALVEZ$/
/^TAMBÉM$/
/^TODAVIA$/
/^TRÁS$/
/^ULTIM(O)?(A)?(S)?$/
/^ANTE$/
/^CONTRA$/
/^DESDE$/
/^PERANTE$/
/^SEM$/
/^TRÁS$/
/^DEZENAS$/
/^DOBRO$/
/^DOZE$/
/^DUPLO$/
/^DUPLA$/
/^MIL$/
/^MILHEIRO$/
/^MILHÕES$/
/^QUARTA(O)?$/
/^VINTE$/
/^-$/
/^°$/
/^°C$/
/^Å$/
/^Á$/
/^a$/