Extração Manual e Automática de Terminologia: Comparando Abordagens e Critérios

Maria Fernanda Teline[1], Gladis Maria de Barcellos Almeida[2] e Sandra Maria Aluísio[1]

[1] Núcleo Interinstitucional de Lingüística Computacional, ICMC-USP, CP 668, 13560-970 – São Carlos – SP – Brasil

[2] Departamento de Letras, UFSCAR, Rodovia Washington Luís (SP-310), Km 235 13565-905 – São Carlos – SP – Brasil

{mteline,sandra}@icmc.usp.br, gladis_maria@uol.com.br

Abstract. *This work shows initial results of a project focus in linguistic, statistical and hybrid evaluation of automatic extraction method in Ceramic domain. Three points are developed in this paper: 1) the comparison between manual and statistical process of terms extraction; 2) the comparison between automatic and manual procedures and 3) the comparison among lexical measures employed in the statistical process.*

Resumo. *Neste artigo apresentamos os resultados iniciais de um projeto de avaliação de métodos de extração automática das abordagens lingüística, estatística e híbrida, em textos do domínio de Revestimento Cerâmico. Centramos em três pontos: 1) a comparação entre os processos manual e automático de extração de termos; 2) a comparação entre os procedimentos manual e automático e, 3) a comparação entre medidas estatísticas.*

1. Introdução

A Informática e a Terminologia não são independentes uma da outra. Desde 1960, em países desenvolvidos e com grande tradição em pesquisa terminológica, a Informática e a Terminologia estão ligadas de forma a facilitar o armazenamento e difusão de dados terminológicos na elaboração de grandes bases de dados especializados, denominados bancos de terminologia (Dubuc, 1999). A integração entre as duas áreas é tal que se cunhou o termo Terminótica para marcar essa integração: "...entendendo a Terminologia como a ciência que estuda a formação dos termos, e, ao mesmo tempo, vendo a Terminografia como a atividade de recenseamento, constituição, gestão e difusão dos termos, pode-se compreender a Terminótica como o conjunto de operações automatizadas de tratamento de termos." (Maciel, 2001) É fato que, no Brasil, tal realidade vai se dar muito tardiamente e, ainda assim, o que temos hoje na pesquisa terminológica ainda é muito incipiente.

A Terminologia cumpre um importante papel no mundo moderno, repleto de inovações científico-tecnológicas, posto que esses avanços científicos e tecnológicos precisam ter nomes, e nomes apropriados. Dessa forma, o uso de repertórios terminológicos sistematizados ou harmonizados – por meio da Terminologia – contribui para tornar mais eficaz a comunicação entre especialistas, comunicação essa que se propõe, acima de tudo, a ser concisa, precisa e adequada (Cabré, 1996).

Entretanto, para se empreender a tarefa de sistematizar/harmonizar repertórios terminológicos, é fundamental que haja ferramentas computacionais compatíveis com

esse tipo de empreendimento. O que vemos hoje no Brasil é uma carência muito grande desses recursos. Isso mostra a grande necessidade de a Terminologia se aliar à Informática para gerar produtos terminológicos mais fiáveis.

O artigo que ora apresentamos tem por objetivo explicitar resultados parciais do projeto intitulado ExPorTer (http://www.nilc.icmc.usp.br/nilc/projects/termextract.htm) que pretende avaliar métodos de extração automática das abordagens lingüística, estatística e híbrida, em textos do domínio de Revestimento Cerâmico.

A partir desses resultados, foi possível fazer uma análise comparativa com o método manual de extração de termos que vem sendo implementado desde 1998, quando se iniciou a elaboração do Dicionário de Revestimento Cerâmico (doravante DiRC).

Interessa-nos chamar atenção aqui para dois pontos: primeiro, para a comparação entre os processos manual e automático de extração de termos; e, segundo, para a comparação de critérios utilizados nesses processos. Isso porque em cada um dos processos empregados um critério foi privilegiado. No processo manual de extração de termos, foi privilegiado o critério semântico; no processo automático, o critério de freqüência. Importa, pois, fazer uma análise comparativa de ambos os critérios, demonstrando em que medida são critérios pertinentes, operacionais e complementares numa pesquisa terminológica com fins terminográficos.

2. Método manual para a extração de terminologia e o critério semântico

Com relação especificamente ao que se refere à extração de termos, cumpre registrar a precariedade dessa atividade sem os recursos oferecidos pela Informática. No DiRC que estamos elaborando desde 1998, temos trabalhado com extração manual de termos. Extrair manualmente do corpus (cf. Seção 4.1 a seguir) os candidatos a termo faz com que o terminólogo enfrente uma das maiores dificuldades na pesquisa terminológica, qual seja, o terreno movediço que há entre palavra (unidade da língua geral) e termo (unidade das comunicações especializadas). Uma das etapas fundamentais de qualquer pesquisa desse tipo é a coleta de termos nos textos especializados. Ora, que critérios deveremos utilizar para efetuar essa tarefa a contento? Dito de outro modo: como saber, ao certo, se aquela unidade selecionada é termo ou palavra, já que o terminólogo, na maioria das vezes, não é um especialista da área que está sendo objeto de investigação?

É fato que o especialista da referida área deve acompanhar a pesquisa; entretanto, a coleta dos termos é feita pelo terminólogo e não pelo especialista. A este último cabe sugerir as fontes relevantes e mais representativas para servir de base para a constituição do corpus, como também apontar, nas listas de candidatos a termo elaboradas pelo terminólogo, os termos que devem ser incluídos e os que devem ser rechaçados.

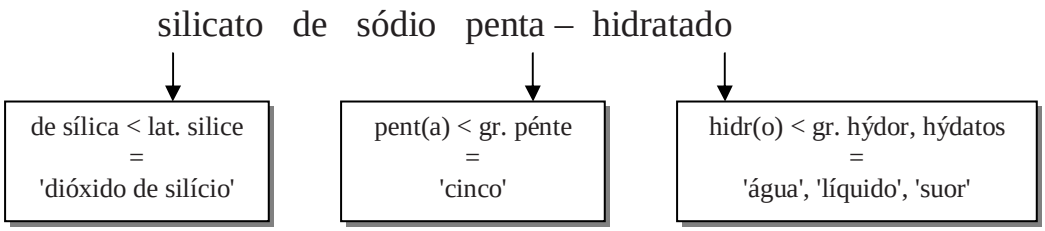
Essa dificuldade foi enfrentada por nós durante a coleta de termos em revistas especializadas para a elaboração do dicionário (Almeida, 2000). Essa dificuldade dizia respeito sobretudo àqueles termos que eram utilizados também na língua geral por um não-especialista, como também às supostas ocorrências de lexias complexas que, muitas vezes, eram tão-somente uma combinatória de palavras, ou um sintagma discursivo.

A pergunta que se nos apresenta é: como saber o que realmente é termo num texto especializado, para elaborarmos o inventário terminológico acerca de determinado domínio? Que critérios deveremos utilizar para apontar o que de fato são termos?

É claro que o terminólogo pode compreender perfeitamente qual o escopo da

Terminologia, como também tem condições de diferenciar um texto de especialidade de um texto da língua geral, pois há uma série de indicadores lingüísticos e textuais para isso; entretanto, ainda não fica claro como distinguir, num texto de especialidade, termo de palavra.

O aspecto mais perceptível, evidentemente, é aspecto formal, entretanto, se esse aspecto fosse suficiente, a extração dos termos nos textos não seria tão penosa. Se estamos diante de uma formação marcadamente técnico-científica, como por exemplo aquelas que utilizam morfemas greco-latinos, não encontramos tanta dificuldade, posto que o nível morfológico já é suficiente para indicar que se trata de um termo e não de uma palavra. Observe-se o exemplo a seguir:



No exemplo exposto acima, o critério formal (nível morfológico) é válido para distinguir termos de palavras. Outra possibilidade é utilizar o nível lexical e constatar os paradigmas derivacionais, como por exemplo: ***floculante, deflocular, defloculação***. A partir de uma ocorrência de um dos termos acima, podemos inferir que as demais ocorrências constituam termos também.

Infelizmente isso não é possível com a grande maioria dos termos originários da língua geral, termos esses que não têm marcas formais para facilitar a sua recolha em textos especializados, como por exemplo: *forno, peneira, secador, biscoito, argila magra, suporte queimado*, etc. Contextos como:

a. *"Para solucionar estes problemas, é necessário evitar a introdução de material úmido no **forno**", diz Quintanilla. "A umidade residual do **secador** deve ser de 4% e, sobretudo, deve-se evitar a entrada de ar desnecessário", completa*. (Revista *Mundo Cerâmico*, no. 51, p. 16)

b. *As **peneiras** são equipamentos utilizados para separar materiais de granulometria diferenciada.* (Revista *Guia de Compras*, no. 53, p.44)

c. *Empregando o critério de cor do produto queimado, é possível dividir os revestimentos cerâmicos em dois grandes grupos: produtos de queima vermelha e produtos de queima branca. A cor do **suporte queimado** depende quase exclusivamente do conteúdo em óxidos colorantes presentes na composição, principalmente de óxidos de ferro.* (Revista *Cerâmica e Informação*, 4, p.15)

d. *Nestes casos, Quintinilla, da Verdés, recomenda uma secagem mais apurada, o aumento da porosidade das peças, através de adição de **argilas magras** durante o preparo e a diminuição da velocidade de pré-aquecimento.* (Revista *Mundo Cerâmico*, no. 51, p. 16)

Nos contextos A e B, a dúvida do leitor vai repousar sobre o fato de *forno*, *secador* e *peneira* constituírem (ou não) termos, já que são signos da língua geral e, se não houver nenhuma ilustração no texto, não há meios de saber se esses dois signos se reportam de fato a conceitos individualizados no domínio específico, ou seja, se existe

alguma característica particularizante em *forno, secador* e *peneira* para que o leitor não se reporte aos referentes do mundo objetivo que já tem interiorizado pela cultura.

Com relação ao contexto C, não há como saber se *queimado* é apenas um adjetivo modificando a base *suporte* ou se faz parte do termo formando uma única unidade terminológica: *suporte queimado*. A dúvida é reforçada pela expressão *produto queimado* que aparece no início do texto. O leitor leigo pode inferir, a partir dessa expressão, que *queimado* é apenas um adjetivo. O mesmo acontece com a expressão *argila magra*, no contexto D: trata-se de um adjetivo apenas ou compõe a unidade terminológica?

Situações desse tipo são habituais durante a recolha dos termos, fazendo com que essa fase torne-se muito mais lenta do que o esperado. Entretanto, essas dificuldades podem ser minimizadas com o auxílio de uma ferramenta que realiza a extração de termos de forma automática, isso porque o critério **freqüência** pode ser considerado para a seleção dos candidatos a termo. Retomando um dos exemplos mencionados acima, podemos afirmar que o item *forno* não teria uma freqüência tão alta se não fosse realmente termo da área de Revestimento Cerâmico. Ainda que o aspecto formal não possa ser considerado para a seleção dos termos nos textos especializados, podemos contar com o critério **freqüência** para selecionar os candidatos.

Além disso, a coleta manual de terminologias não permite que sejam armazenados todos os contextos relevantes sobre cada termo, dificultando a seleção do melhor contexto para integrar cada verbete. Desnecessário dizer que num produto terminológico é de extrema relevância o acesso a contextos para a elaboração da definição.

No caso da extração manual, importa explicitar os procedimentos que se seguiam à extração dos candidatos a termo. Em outras palavras: qual o critério utilizado para saber ao certo se os candidatos poderiam constituir-se em termos?

Todos os candidatos a termo selecionados iam sendo armazenados no que denominamos estrutura conceitual. A estrutura conceitual constitui uma representação da realidade no âmbito do domínio que se toma como objeto de estudo. Essa representação procura recolher e organizar todas as ramificações que são próprias do referido domínio, de modo a refletir, em forma de esquema, a realidade da área em questão, semelhante, portanto, a uma ontologia. Nessa estrutura, são indicados todos os campos e subcampos nocionais contemplados no trabalho terminológico, o que permite que cada candidato a termo seja armazenado no campo nocional correspondente. Isso permite que cada termo seja determinado pela posição que ocupa no sistema nocional representado, explicitando, dessa maneira, as relações hierárquicas e não-hierárquicas entre eles. Assim, esses campos ou subcampos nocionais constituíam listas que eram submetidas aos profissionais da área para que apontassem, dentre todos os termos constantes de determinado campo, aqueles que fossem relevantes. Adotando esse procedimento, estávamos na verdade considerando o critério **semântico**, que significa selecionar, num campo nocional dado, os termos/conceitos relevantes.

3. Métodos automáticos de extração de terminologia e o critério freqüência

Dado o grande volume de informação técnica disponível nesta última década e com o crescente uso da WWW como fonte de pesquisa e depósito de textos técnicos e

científicos, esforços manuais para a extração de terminologia de *corpora*[1], como o reportado acima, se tornaram ineficazes.

Os sistemas de extração automática de terminologia são de extrema importância para aplicações tais como: tradução humana ou automática, indexação, construção de *thesaurus*, organização do conhecimento, entre outras. Um sistema de extração automática de candidatos a termo[2] (SEACAT) é formado por um conjunto de programas para o reconhecimento de unidades terminológicas de *corpora* (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. Os SEACAT são tradicionalmente classificados conforme a metodologia que utilizam para reconhecer as unidades terminológicas, em sistemas que: a) utilizam apenas métodos baseados em conhecimento estatístico; b) utilizam apenas métodos baseados em conhecimento lingüístico; c) utilizam métodos baseados em conhecimento estatístico e lingüístico.

Os métodos baseados em conhecimento estatístico geralmente detectam as unidades terminológicas de acordo com a freqüência com que elas ocorrem em um *corpus*. Existem métodos estatísticos que utilizam desde simples freqüências (Daille, 1996) até estatísticas mais complexas, como informação mútua e coeficiente *log-likelihood* (Pantel and Lin, 2001) e *c-value* (Frantzi and Ananiadou, 1997). A função é, em todos os métodos, identificar os candidatos a termo.

Os métodos estatísticos são dependentes do tamanho do *corpus* que utilizam, diferentemente dos métodos lingüísticos. Dessa forma, se o *corpus* 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 *corpus* apresenta milhões de ocorrências, há sempre uma porcentagem de palavras que não pode ser recuperada em razão de sua baixa freqüência de uso no *corpus*. Os métodos estatísticos também são responsáveis por gerar bastante ruído, que vem a ser o número de candidatos a termo que não apresenta valor terminológico, isto é, aquelas palavras que não apresentam significado especializado, sendo pertencentes, portanto, à língua geral (Estopà Bagot, 2001). Embora possuam os problemas levantados acima, os métodos estatísticos são independentes da língua, sendo esta mais uma característica que os diferencia dos métodos lingüísticos.

Os sistemas baseados em conhecimento lingüístico (Heid et al., 1996; Klavans and Muresan, 2000, 2001a, 2001b) utilizam diferentes recursos que contêm diferentes informações lingüísticas para a extração dos termos. Essas informações lingüísticas dizem respeito a: informações lexicográficas – dicionários de termos e lista de palavras auxiliares ("*stopwords*"); informações morfológicas – padrões de estrutura interna da palavra; informações morfossintáticas – categorias morfossintáticas e funções sintáticas; informações semânticas – classificações semânticas; informações 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. Ressalte-se que a sua utilização em textos em uma língua diferente

[1] Sejam eles construídos de textos da WWW ou textos impressos.
[2] Os termos candidatos devem ser, posteriormente, validados por humanos.

exige um estudo lingüístico prévio e necessita de um novo projeto para alguns dos módulos do sistema.

De acordo com Estopà Bagot (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 sintagmas discursivos próprios do discurso científico, mas não unidades terminológicas, como "o objetivo deste trabalho", "nesta seção", sem caráter especializado.

Por esta 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 (Frantzi and Ananiadou, 1997; Dias et al., 2000) 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 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 de a estatística auxiliar no momento do processo de detecção, reafirmando ou recusando a condição de termo de uma unidade lingüística.

Neste artigo exploramos as estatísticas léxicas, informação mútua e o coeficiente *log-likelihood* para a extração de candidatos a termos. Informação Mútua é uma medida da quantidade de informação que uma variável contém sobre uma 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 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 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:

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

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.

4. Os corpora utilizados

4.1 Material a partir do qual foi constituído o corpus para a extração manual

Foram selecionados somente textos em língua portuguesa, tanto para as fontes escritas quanto para as fontes orais. Isso porque nosso intuito foi organizar a terminologia de revestimentos cerâmicos segundo o recorte de mundo feito pela cultura brasileira, para isso, é fundamental considerar o universo conceptual e terminológico da língua do usuário e eventual consulente. Segue, abaixo, a relação de todo o material a partir do qual constituímos o corpus para a extração manual.

4.1.1. Fontes escritas

a) Documentos da ABNT: incluímos como fontes todos os termos já normalizados ou em processo de normalização. Para tanto, consultamos todos os comitês da ABNT cujo domínio fosse conexo com a área de revestimentos cerâmicos;

b) Revistas científicas e/ou de divulgação: ainda como fontes escritas, consideramos as revistas científicas e/ou de divulgação. Sendo assim, nosso repertório consta de textos científicos, informativos e publicitários, todos versando sobre o domínio dos revestimentos cerâmicos;

c) lista de termos em obras especializadas: consideramos também glossários apresentados como anexos de obras especializadas em Materiais Cerâmicos, que são utilizadas no curso de graduação em Engenharia de Materiais da Universidade Federal de São Carlos (UFSCar).

4.1.2. Fontes de língua oral

Foram igualmente considerados os termos obtidos em entrevistas e outras situações de interação oral, conforme explicitado a seguir.

a) entrevistas: entrevistas com pesquisadores do Núcleo de Informação Tecnológia em Materiais (NIT/Materiais-UFSCar) e do Centro de Caracterização e Desenvolvimento de Materiais (CCDM-UFSCar) e docentes da área de Materiais Cerâmicos do Departamento de Engenharia de Materiais (DEMa-UFSCar);

b) entrevistas com profissionais de indústrias realizadas durante as visitas às fábricas situadas nos municípios de Rio Claro, Santa Gertrudes, Piracicaba e Porto Ferreira;

c) outras situações de interação oral: foram igualmente considerados dados e informações recolhidos em congressos, *workshops*, seminários ou palestras. Nessas ocasiões, tivemos a oportunidade de encontrar vários representantes do domínio de revestimentos cerâmicos, tanto dos setores de ensino/P&D quanto do setor industrial.

4.2 Material a partir do qual foi constituído o corpus para a extração a partir do método estatístico

O *corpus* utilizado para avaliar as medidas estatísticas foi extraído de textos que se encontram no *site* de Cerâmica Industrial[3]. Estes textos estão agrupados por anos que vão desde 1996 até 2003, totalizando 196 artigos, possuindo cada uma média de 7 a 8 páginas (aproximadamente 4000 palavras).

Todos os textos presentes no *site* estão no formato pdf. No entanto, para os textos serem processados para os cálculos das medidas estatísticas utilizadas neste trabalho, eles devem estar no formato txt, e, por esta razão, nem todos os textos do *site* foram utilizados, visto que ocorreram alguns problemas na transformação de alguns desses textos do formato pdf para o formato txt. Destes 196, 141 foram utilizados para compor o *corpus* de trabalho.

Para transformar esses textos para o formato txt, foi utilizada uma ferramenta denominada ERTEX (Extracção de Texto de Ficheiros Formatados)[4]. No entanto, essa ferramenta, ao realizar a transformação, faz a junção de algumas palavras em apenas uma, preserva os índices de referência bibliográfica e as notas de rodapé anexados às palavras e também a hifenização dos textos no formato pdf. Para resolver esses problemas, esses textos sofreram um processo cuidadoso de correção manual.

Todos os arquivos do *corpus* foram também 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 388.378.

As medidas estatísticas utilizadas estão incorporadas no pacote NSP (N-gram Statistics Package)[5], escrito em Perl. Dentre as medidas de associação encontradas nesse pacote, estão sendo utilizadas a Informação Mútua e o *Log-Likelihood*, bem como a Freqüência para realizar um levantamento dos termos encontrados no *corpus*. A Freqüência pode ser calculada para n-gramas, e, para este trabalho, esse n está limitado aos valores 1, 2 (unigramas e bigramas) para serem comparados com a lista de termos candidatos gerada na extração manual realizada no *corpus* descrito acima (cf. Seção 4.1). A partir da lista gerada na extração manual, constataram-se como mais produtivos os sintagmas nominais assim realizados: [subst+adj], [subst+prep+subst], [subst+adj+adj], como por exemplo *fase vítrea, ciclo de queima, argila refratária aluminosa*, respectivamente.

A medida Informação Mútua e *Log-Likelihood* foram utilizadas para o cálculo do escore de bigramas. Após a geração desses unigramas e bigramas e o cálculo das medidas apresentadas, foi realizado um levantamento dos termos encontrados na lista de termos obtida na extração manual. Por meio desse levantamento, foi encontrada uma grande quantidade de erros gramaticais, dentre eles, erros de digitação, de concordância em gênero e em número e de acentuação. Também foi possível perceber que alguns termos encontrados no *corpus* apresentavam hífen, enquanto que na lista obtida pela extração manual esses termos se encontravam não hifenizados. Para minimizar os erros gramaticais, foi realizada uma varredura no *corpus*, buscando corrigir os erros

[3] http://www.ceramicaindustrial.org.br/
[4] http://poloclup.linguateca.pt/ferramentas/extex/
[5] http://www.d.umn.edu/~tpederse/nsp.html

encontrados pelo levantamento, podendo, dessa forma, analisar os dados de forma mais precisa.

A partir da lista de palavras geradas pelo pacote NSP, foi possível perceber que as palavras que apareciam com maior freqüência eram palavras funcionais (preposições, artigos, conjunções) que não apresentam nenhum valor terminológico. Por essa razão, foi construída uma *stoplist* com essas palavras e alguns advérbios bastante produtivos, a fim de obter uma lista menor e "mais limpa" de palavras para ser analisada pelo especialista. Estes experimentos estão descritos detalhadamente abaixo.

5. Experimentos com medidas estatísticas

Após a execução das tarefas descritas acima, foram realizados 3 experimentos, sendo que em cada um destes foi gerada uma lista para unigramas e para bigramas com o uso das medidas mencionadas acima. Para o primeiro teste, os cálculos das medidas foram realizados com o *corpus* não corrigido e sem o uso da *stoplist*. Foram recuperados 388.378 unigramas e 388.377 bigramas. Os 10 unigramas mais freqüentes são as palavras de classe fechada (preposição, artigos e pronomes) como era de se esperar. Já para o segundo teste, a *stoplist* construída foi, então, utilizada. Realizando uma comparação entre os dois testes, foi possível notar uma grande diferença na quantidade de palavras geradas, visto que as palavras funcionais foram excluídas. Neste experimento foram recuperados 210.795 unigramas e 78.099 bigramas. O primeiro item mais freqüente é termo (unidade especializada), pertencente à lista de Referência.

Para o terceiro teste, os resultados foram gerados utilizando-se também a *stoplist* e o *corpus* revisado. A comparação deste último com o segundo não apresentou diferenças discrepantes na quantidade de palavras, mas apenas produziu uma lista de termos corrigidos quanto à ortografia e concordância em número e gênero. A Tabela 1 apresenta os 10 unigramas mais freqüentes e 10 exemplos de unigramas menos freqüentes, mostrados com sua freqüência entre parênteses. Por exemplo, “alargando” é um dos 7309 itens com freqüência 1, freqüência essa que colabora com 3.47% do total de itens. O número de palavras com freqüência 2 corresponde a 2.38% do total de itens; o número de palavras com freqüência 3 a 1.87% e o de 4 a 1.68%. A Tabela 2 apresenta os maiores escores para as três medidas estatísticas consideradas. A fim realizar uma comparação entre os escores das medidas de associação utilizadas (Informação Mútua e *Log-likelihood*), foi calculado o coeficiente de correlação de Spearman, também presente no pacote NSP. Este coeficiente produz uma saída que varia entre -1 e 1. Caso este resultado seja 1, significa que as duas medidas apresentam uma coincidência perfeita entre os escores. Já um resultado equivalente a -1 indica que os escores das medidas são completamente reversos, enquanto que um resultado igual a 0 indica um par de escores completamente não relacionados. A informação mútua e o *log-likelihood* apresentam um coeficiente igual a 0.5463. Ainda precisam ser realizadas mais análises para a escolha da melhor medida estatística, isto é, aquela que produza uma lista de termos com as menores taxas de silêncio e ruído.

Tabela 1: Unigramas mais e menos freqüentes

10 Unigramas Mais Freqüentes	10 Unigramas Menos Freqüentes - Exemplos
Queima (960)	Escolhida (10)
Produção (952)	Vácuo (9)
C (932)[6]	Utilidade (8)
Processo (914)	Fundidos (7)
Maior (879)	Categoria (6)
Foi (866)	Maximizar (5)
Cerâmicos (835)	Distribuído (4)
Temperatura (823)	Fomentar (3)
Foram (811)	Leucita (2)
Revestimentos (807)	Alargando (1)

Tabela 2: Bigramas de maior escore gerados pelas medidas estatísticas

Log-Likelihood	**Freqüência**	**Informação Mútua**
Revestimentos cerâmicos	Revestimentos cerâmicos (496)	Revestimentos cerâmicos
Resistência mecânica	Resistência mecânica (179)	Resistência mecânica
Grês porcelanato	Expansão térmica (145)	Grês porcelanato
Matérias primas	Fase vítrea (136)	Matérias primas
Expansão térmica	Resultados obtidos (134)	Expansão térmica
Fases cristalinas	Grês porcelanato (131)	Fase vítrea
Fase vítrea	Fases cristalinas (119)	Fases cristalinas
Via úmida	Indústria cerâmica (109)	Via úmida
Resultados obtidos	Matérias primas (108)	Resultados obtidos
Placas cerâmicas	Placas cerâmicas (108)	Via seca
Via seca	Via Úmida (104)	Placas cerâmicas
Presente trabalho	Via Seca (102)	Presente trabalho
Indústria cerâmica	Revestimento Cerâmico (97)	Indústria cerâmica
Densidade aparente	Setor Cerâmico (82)	Densidade aparente
Coração negro	Processo Produtivo (79)	Coração negro
Processo produtivo	Presente Trabalho (77)	Processo produtivo
Santa Gertrudes	Densidade Aparente (76)	Revestimento cerâmico
Revestimento cerâmico	Produtos Cerâmicos (68)	Santa Gertrudes
Construção civil	Produto Final (63)	Construção civil
Distribuição granulométrica	Área Específica (57)	Distribuição granulométrica

Considerando que se utilizou a *stoplist* mencionada anteriormente e tendo o *corpus* sido corrigido gramaticalmente, esse passou a apresentar 210.797 unigramas (total de palavras), variando de freqüência 1 a 960 (termo "queima" na Tabela 1) e 78.096 bigramas, variando de freqüência 1 a 496 (termo "Revestimentos cerâmicos" na Tabela 2). No entanto, foram considerados apenas unigramas e bigramas com freqüência maior que 4 para efeito de comparação com uma lista de candidatos a termos gerada pelo processo manual de extração (chamada de Lista da Extração Manual), que engloba a Lista de Referência, provocando uma diminuição para 5.359 unigramas não repetidos e 1.563 bigramas não repetidos. A Lista da Extração Manual apresenta um total de 782 candidatos, considerando que 351 correspondem a unigramas e 175 a bigramas. Foi, então, realizada a intersecção, para unigramas e bigramas respectivamente, entre a Lista da Extração Manual e a lista gerada pelo *corpus*

[6] Este item vem da notação de graus Celsius que perdeu o símbolo de grau na formação do *token* pelo pacote NSP.

(primeiramente para unigramas e posteriormente para bigramas), produzindo 170 unigramas (48.4 %) e 38 bigramas (21.7%) que são comuns às duas listas.

Apesar das baixas porcentagens apresentadas, deve ser considerado que a Lista da Extração Manual apresenta-se lematizada, enquanto que as palavras que aparecem no *corpus* não apresentam tal característica. Isso implica que muitas palavras podem não ter sido consideradas para este total de 170 unigramas e 38 bigramas em razão de elas se apresentarem flexionadas no decorrer do *corpus*. Isso sem considerar os erros gramaticais que o *corpus* ainda apresenta e que não foram totalmente corrigidos.

6. Considerações finais

Com relação à análise comparativa entre os processos **manual** e **automático** de extração de terminologias, foi possível observar resultados extremamente relevantes não só para o andamento do projeto ExPorTer para a avaliação de métodos de extração automática das abordagens lingüística, estatística e híbrida, em textos do domínio de Revestimento Cerâmico, em desenvolvimento, como também para a pesquisa terminológica em língua portuguesa.

A despeito das baixas porcentagens apresentadas na intersecção entre a lista obtida por meio da extração manual e a lista gerada pela extração automática, acreditamos na eficácia do método automático. Isso porque um grande percentual obtido, que de fato não equivalia à lista manual, era composto por sintagmas discursivos, lexias complexas sem valor terminológico ou mesmo itens lexicais com inadequações ortográficas, morfológicas e/ou sintáticas (falta de concordância nominal e verbal, por exemplo). Além disso, cumpre ressaltar o fato de a lista gerada pela extração automática não estar lematizada, o que, evidentemente, altera os percentuais obtidos.

Acreditamos serem esses resultados preliminares considerando o fato de termos lidado com o método estatístico de extração automática. Se utilizarmos um método extrator híbrido dotado de um lematizador, é possível que os percentuais de intersecção aumentem consideravelmente.

Com relação aos critérios **semântico** e de **freqüência**, importa registrar o fato de ambos serem pertinentes, operacionais e complementares para a pesquisa terminológica com fins terminográficos. Se o critério de freqüência utilizado na extração automática deve ser validado por humanos, por que não fazê-lo considerando a estrutura conceitual (=ontologia) já organizada para o domínio de Revestimento Cerâmico? Ou seja, após a extração automática, listas de candidatos a termo são geradas. Antes de serem enviadas aos profissionais da área, esses candidatos a termo devem ser inseridos nos campos e/ou subcampos nocionais da estrutura conceitual de forma a facilitar a validação pelos especialistas. Ao proceder dessa maneira, estamos considerando o critério **semântico** que, na verdade, complementa e legitima o critério **freqüência**.

Referências bibliográficas

Almeida,G.M.B.(2000) "Teoria Comunicativa da Terminologia: uma aplicação", Araraquara, vol. I, 290 p., vol. II, 86 p. Tese (Doutorado em Lingüística e Língua Portuguesa) – Faculdade de Ciências e Letras, Câmpus de Araraquara, Universidade Estadual Paulista.

Cabré, M.T. (1996) “Importancia de la terminología en la fijación de la lengua”, Revista internacional de língua portuguesa, Núm. 15, jul. 96, Lisboa: Editorial Notícias, p.9-24.

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. and 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.

Dubuc, R. (1999) “Manual de terminologia”, 3ª. ed. Chile: Unión Latina/RiL editors.

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 Terminologia: elementos para la construcción de un extractor”, TradTerm **7,** Revista do Centro Interdepartamental de Tradução e Terminologia FFLCH - USP, p. 225-50.

Frantzy, K. T. and 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.

Heid, U., Jauβ, S., Krüger, K. and Hohmann, A. (1996) “Term extraction with standard tools for corpus exploration”, In: 4th International Congress on Terminology and Knowledge Engineering, Wien, August.

Klavans, J. L. and Muresan, S. (2000) “DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-line Text”, In: Proceedings of AMIA 2000.

Klavans, J. L. and Muresan, S. (2001a) “Evaluation of DEFINDER: A System to Mine Definitions from Consumer-oriented Medical Text”, In: Proceedings of JCDL 2001.

Klavans, J. L. and Muresan, S. (2001b) “Evaluation of the DEFINDER System for Fully Automatic Glossary Construction”, In: Proceedings of AMIA 2001.

Maciel, A.M.B. (2001) “TERMISUL e terminótica”, In: Krieger, G. and Maciel, A.M.B., Temas de terminologia, Porto Alegre: Editora da Universidade Federal do rio Grande do Sul, São Paulo: FFLCH/USP-Humanitas.

Pantel, P. and 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.