[http://www.inventerm.com/]

RECOLHA E SISTEMATIZAÇÃO DE CORPORA PARA ELABORAÇÃO DO PRIMEIRO DICIONÁRIO-PILOTO EM NANOCIÊNCIA E NANOTECNOLOGIA EM LÍNGUA PORTUGUESA*

** Trabalho apresentado na Jornada Realiter – "Metodologia para a recolha e sistematização de corpora para fins dicionarísticos", realizada em 01 de junho de 2006, no Rio de Janeiro, RJ, Brasil.*

Gladis Maria de Barcelos ALMEIDA [1 [#nb1]]; Sandra Maria ALUÍSIO [2 [#nb2]]; Osvaldo Novais de OLIVEIRA JR. [3 [#nb3]]; Leandro Henrique Mendonça de OLIVEIRA [4 [#nb4]]; Ariani DI FELIPPO [5 [#nb5]]; Luiz Carlos GENOVES JR. [6 [#nb6]]; Leila Garbelini SOARES [7 [#nb7]]; Daniela Ferreira de MATTOS ; Joel Sossai COLETI [8 [#nb8]]

Resumo

Nanociência e Nanotecnologia (doravante N&N) são atualmente áreas centrais das atividades de pesquisa, desenvolvimento e inovação nos países industrializados. Investimentos aplicados nessa área de conhecimento por esses países têm sido crescentes. No Brasil, o cenário para pesquisas em N&N já é promissor, entretanto, ainda há uma grande defasagem dos países do Hemisfério Sul em relação aos países desenvolvidos. Para acompanhar esse desenvolvimento científico e tecnológico que se deseja, além de investimentos financeiros expressivos e formação de recursos humanos especializados, é preponderante a sistematização de repertórios vocabulares em língua portuguesa (doravante LP), posto que não há ainda qualquer glossário e/ou dicionário de N&N em LP. A partir da recolha e sistematização de um corpus em LP, pretendemos criar condições para desenvolver o primeiro Dicionário-Piloto de N&N em LP. Nosso foco aqui estará voltado para as questões relacionadas à elaboração e à manipulação do corpus.

Palavras-chave

lingüística de corpus, processamento de língua natural, terminologia, terminografia, nanociência, nanotecnologia

1. Introdução

Segundo documento elaborado pelo Grupo de Trabalho criado pela portaria do Ministério da Ciência e Tecnologia do Brasil (MCT) n° 252, de 16/05/2003, intitulado "Desenvolvimento da Nanociência e da Nanotecnologia" (2003), a Nanotecnologia é atualmente uma das áreas centrais das atividades de pesquisa, desenvolvimento e inovação nos países industrializados. De acordo com o mesmo documento, os investimentos aplicados nessa área de conhecimento por esses países têm sido crescentes e atingiram, em 2002, cerca de cinco bilhões de dólares. A previsão é de que, entre 2010 e 2015, o mercado mundial envolvendo a Nanotecnologia será de um trilhão de dólares.
No Brasil, o cenário para pesquisas em N&N já é promissor, sobretudo nos segmentos de "manipulação de nano-objetos, nanoeletrônica, nanomagnetismo, nanoquímica e nanobiotecnologia, incluindo os nanofármacos, a nanocatálise e as estruturas nanopoliméricas" ("Desenvolvimento da Nanociência e da Nanotecnologia", 2003). Entretanto, ainda há uma grande defasagem dos países do Hemisfério Sul em relação aos países desenvolvidos, como mostra documento da Organização dos Estados Americanos (OEA), intitulado "Ciência, Tecnologia, Engenharia e Inovação para o Desenvolvimento: uma visão para as Américas no Século XXI" (2005).

Se a defasagem técnico-científica é um fato, maior lacuna se nota no léxico especializado que nomeia esse saber, haja vista que, como costuma acontecer nessas áreas de conhecimento, os termos pertencem à língua

inglesa. Por isso, para acompanhar esse desenvolvimento científico e tecnológico que se deseja, é preponderante a sistematização de repertórios vocabulares em língua portuguesa. Sistematizar terminologias na língua materna significa criar termos fiáveis de forma a facilitar a comunicação especializada, além de demonstrar que a língua portuguesa está apta para nomear conceitos técnicos e científicos. Em outras palavras, ao mesmo tempo em que se promove a disseminação e a divulgação de conhecimentos e de tecnologias, fomenta-se a língua nacional, posto que não há ainda qualquer glossário e/ou dicionário de N&N em língua materna. O que há é um número significativo de produtos terminológicos em língua inglesa (doravante LI), mas, ainda assim, limitados em abrangência e profundidade.
Esta proposta parte do projeto intitulado *Desenvolvimento de uma Estrutura Conceitual (Ontologia) para a Área de Nanociência* e Nanotecnologia, realizado por uma equipe coordenada por Sandra Maria Aluísio, do Núcleo Interinstitucional de Lingüística Computacional (NILC), sediado no Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP), Campus de São Carlos, SP, Brasil (ALUÍSIO et al., 2006). Nesse projeto, totalmente baseado em LI, foram elaborados um corpus, cuja extensão é de 2.570.792 de palavras, e uma ontologia contendo cerca de 1.900 termos. Este projeto foi desenvolvido com o objetivo de organizar o *Portal da Rede de Nanotecnologia da USP* [9 [#nb9]].

Esse projeto serviu de motivação para elaborarmos também um corpus e uma ontologia, mas agora em LP, de forma a criar condições para o desenvolvimento do primeiro Dicionário-Piloto de N&N em língua materna. Submetemos, então, um projeto ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, intitulado *Terminologia em Língua Portuguesa da Nanociência e Nanotecnologia: Sistematização do Repertório Vocabular e Elaboração de Dicionário-Piloto*. Esse projeto (processo 400506/2006-8) foi aprovado em maio/2006, com vigência de dois anos.
Nosso projeto tem como objetivos: 1) a constituição de um corpus em língua portuguesa da N&N; 2) a busca de equivalentes em português (língua de chegada) a partir de uma nomenclatura em inglês (língua de partida); 3) uma ontologia em língua portuguesa da área de N&N; 4) a elaboração do primeiro dicionário-piloto de N&N em língua materna.
No que concerne aos aspectos da língua/linguagem, nosso projeto abordará os seguintes pontos: 1) temas pertinentes à Lingüística de *Corpus*, tais como: criação e compilação de corpus, ferramentas de extração automática de terminologias e demais ferramentas de PLN (Processamento de Língua Natural); 2) equivalência terminológica e todas as questões semânticas que suscitam quando se pretende fazer um novo recorte da realidade; e, finalmente, 3) Terminologia Aplicada, envolvendo todas as etapas de elaboração de um produto terminológico.

Com relação aos aspectos metodológicos voltados para Lingüística de Corpus, seguiremos procedimentos de trabalho já empregados no NILC [10 [#nb10]].
No que se refere aos aspectos metodológicos voltados para a Terminologia/Terminografia, esta proposta seguirá procedimentos terminológicos já desenvolvidos e testados no Grupo de Estudos e Pesquisas em Terminologia (GETerm), sediado na UFSCar. Esses procedimentos estão em conformidade com a Terminologia de orientação descritiva, fundamentada em princípios da Lingüística, cujo exemplo mais consolidado é a Teoria Comunicativa da Terminologia (TCT) (CABRÉ, 1999, 2003). Filiar-se teoricamente à TCT significa, metodologicamente, abandonar "o tratamento prescritivo das terminologias em favor de enfoques descritivos capazes de entender o léxico especializado como um elemento natural das línguas naturais" (KRIEGER e BEVILACQUA, 2005).

2. Etapas metodológicas para gerar o corpus em LP

As mesmas diretrizes que foram levadas em consideração para a constituição do *corpus* em língua inglesa (LI) foram repetidas neste projeto que prevê a constituição de um *corpus* em língua portuguesa (LP). A seguir, apresentaremos detalhadamente as etapas até a geração desse corpus.

2.1. Seleção dos textos

Inicialmente, foi realizado um estudo exploratório dos textos existentes em LP bem como dos gêneros aos quais eles pertencem. Embora tivéssemos tentado reproduzir os mesmos procedimentos do corpus em LI, não obtivemos o mesmo sucesso no que se refere à quantidade de textos do gênero "discurso científico", sobretudo provenientes de revistas. Entendemos que isso se deve ao seguinte fator: como a área de N&N é relativamente nova no Brasil, os pesquisadores, fundamentalmente das áreas de Exatas e Biomédicas que atuam em N&N, publicam seus resultados de pesquisa em LI, em razão disso, não há ainda uma grande produção em LP.

2.2. Compilação e manipulação do corpus

A **compilação** consiste no armazenamento em arquivos predeterminados de todos os textos pertinentes e relevantes para a pesquisa. Para essa compilação, estão sendo utilizados os seguintes itens de busca: *nanosciência*, *nanotecnologia*, *genômica*, os mesmo utilizados para formar o corpus em LI. Todavia, após realizarmos buscas, decidimos incluir e o prefixo nano- para abarcar termos como: *nanotubo*, *nanorrede*/*nano-rede*, *nanocápsula*, *nanoesfera*, *nanobiotecnologia*, etc.
A **manipulação** do corpus constitui na (i) conversão manual e automática (Pacote XPDF) de formatos "doc", "html" e "pdf" para "txt" e (ii) limpeza e formatação, de maneira a preparar o corpus para o processamento computacional. A tarefa descrita em (ii) possibilita, por exemplo, a obtenção de uma lista de freqüência, a utilização de um concordanceador e a extração automática (e demais ferramentas de – PLN que se mostrarem pertinentes).

2.3. Nomeação de arquivos e geração de cabeçalhos

Depois que todos os textos forem convertidos em formato "txt", eles devem receber um nome. Ressalte-se que essa nomeação deve seguir determinado padrão de forma a facilitar a recuperação posterior de cada texto. Após a nomeação dos arquivos, é gerado (de forma semi-automática) um cabeçalho para cada texto. A geração semi-automática desse cabeçalho será feita por meio de um editor (programa computacional "com interface gráfica" para criar ou modificar arquivos) que auxilia o lingüista a especificar diversas informações sobre os textos. Ressaltamos que esse programa é uma versão adaptada no Editor de Cabeçalho utilizado no Projeto Lacio-Web [11 [#nb11]] e contém os seguintes campos de informação: título, subtítulo, fonte, editor, local de publicação, data, assunto, autoria, tipo de autoria (individual ou coletiva), sexo do autor, tipo de texto, meio de distribuição e comentários (introduzem-se nesse campo informações adicionais sobre o texto). O preenchimento de todos esses campos é útil para esta pesquisa porque a partir desses dados será possível fazer constatações tais como: o repertório vocabular tem alguma relação com a temática do texto, com o gênero, com a autoria ou com o meio de distribuição? Dependendo do tema tratado em determinado texto, é possível recuperar os descritores desse texto por meio da freqüência? Em outras palavras: num texto cujo tema seja *Nanociência*, o item léxico *nanociência* ocorre quantas vezes? Enfim, além das buscas que poderão ser empreendidas por cada campo constitutivo do cabeçalho, é possível fazer constatações relevantes sobre o léxico.

3. Etapas para gerar o dicionário

Com o *corpus* elaborado, é possível iniciar as atividades concernentes à parte terminológica/terminográfica. Estabelecemos, então, como método de trabalho, uma seqüência de etapas que devem integrar um projeto terminológico com fins terminográficos e que tenham a filiação teórica mencionada anteriormente, qual seja: a Terminologia de orientação descritiva, fundamentada em princípios da Lingüística.
As etapas são as seguintes: 1) extração (automática) de candidatos a termos; 2) elaboração da ontologia (também denominado mapa conceitual); 3) inserção dos termos na ontologia e sua validação pelos

especialistas; 4) elaboração e preenchimento das fichas terminológicas; 5) elaboração e incremento da base definicional; 6) elaboração das definições e informações enciclopédicas (quando for o caso); 7) edição dos verbetes. Faremos, a seguir, uma breve explanação de cada uma dessas etapas [12 [#nb12]].

3.1. Extração de candidatos a termos

A extração de termos diz respeito à obtenção do conjunto terminológico que comporá a nomenclatura do glossário ou dicionário. Ressaltamos que há três abordagens por meio das quais se pode fazer a extração: 1) a primeira, denominada estatística, utiliza-se de sistemas baseados em estatística; 2) outra abordagem existente é a lingüística, em que os sistemas detectam padrões recorrentes de unidades terminológicas complexas, tais como "substantivo–adjetivo" e "substantivo–preposição–substantivo", por exemplo; 3) o terceiro tipo é a híbrida, em que os sistemas começam a detectar algumas estruturas lingüísticas básicas, tal como expressões nominais, e depois de os termos candidatos terem sido identificados, uma estatística relevante é usada para decidir se eles correspondem a um termo. O inverso também é possível, começando-se com uma lista de candidatos levantados estatisticamente, sendo que a informação lingüística, nesse caso, é usada para filtrar termos válidos da lista.
Antes de optarmos pela abordagem mais adequada, faremos a avaliação do processo de extração de termos, utilizando métricas clássicas da área de processamento, como a Precisão e a Revocação (Recall) [13 [#nb13]] . A partir dos resultados obtidos, optaremos pela melhor abordagem.

3.2. Inserção dos termos na ontologia

Os termos obtidos devem ser inseridos na ontologia, por isso ela deve ser organizada preliminarmente, ou concomitantemente à extração dos termos, já que, à medida que os termos vão sendo obtidos, é que se pode ter uma visão real de quais serão os campos nocionais que deverão integrar a ontologia.
A ontologia é uma organização semântica da área-objeto, semelhante ao que se entende por árvore de domínio, a diferença é que os conceitos/termos estão ali armazenados. Organiza-se uma estrutura constituída de campos nocionais, de forma que essa estrutura reflita os conceitos da área-objeto bem como as relações entre eles. A ontologia deve ser elaborada pelos terminólogos com assessoria dos profissionais da área-objeto. Na pesquisa terminológica, a ontologia é fundamental para: 1) possibilitar uma abordagem mais sistemática de um campo de especialidade; 2) circunscrever a pesquisa, já que todas as ramificações da área-objeto, com seus campos, foram previamente consideradas; 3) delimitar o conjunto terminológico; 4) determinar a pertinência dos termos, pois separando cada grupo de termos pertencente a um determinado campo, poder-se-á apontar quais termos são relevantes para o trabalho e quais não são; 5) prever os grupos de termos pertencentes à área-objeto, como também os que fazem parte de matérias conexas; 6) definir as unidades terminológicas de maneira sistemática e, finalmente; 7) controlar a rede de remissivas (ALMEIDA, 2000).
A partir do momento em que os termos estão alocados na ontologia, pode-se proceder à sua validação pelos especialistas. A validação de termos pelos especialistas é feita da seguinte maneira: selecionam-se da ontologia determinados campos nocionais e pede-se que cada assessor assinale os termos considerados semanticamente relevantes em cada campo.

3.3. Preenchimento das fichas terminológicas

À medida que os termos forem sendo validados pelos especialistas, inicia-se o preenchimento das fichas terminológicas. O preenchimento das fichas é uma etapa imprescindível numa pesquisa terminológica. A ficha constitui-se num verdadeiro dossiê do termo, contendo toda a sorte de informações que se mostrarem pertinentes para a pesquisa em foco. Daí a razão de ela ser planejada logo no início do trabalho. Importa mencionar que não há um modelo ideal de ficha terminológica, cada ficha deve refletir as necessidades do projeto, isto é: "para quê" e "para quem" se faz determinado glossário ou dicionário. Isso auxilia o

terminólogo a prever quais campos deverão constar do protocolo de preenchimento da ficha terminológica.

3.4. Elaboração da base definicional

Concomitante ao preenchimento da ficha, incrementa-se a base definicional. A base definicional tem como função armazenar todos os excertos definitórios ou quaisquer contextos explicativos referentes aos termos, de forma a facilitar a redação da definição. Esses excertos são extraídos tanto do *corpus* em LP como da bibliografia especializada disponível [14 [#nb14]]. É imprescindível armazenar essas informações, uma vez que: 1) somente com o preenchimento de um número suficiente de excertos definitórios é que a redação de uma definição pode ser iniciada; 2) a quantidade e qualidade de excertos devem ser suficientes para elucidar o redator das definições, uma vez que este não é um especialista da área-objeto; 3) as definições, depois de elaboradas, são submetidas à apreciação dos especialistas, caso eles encontrem algum problema conceitual, questionem as fontes bibliográficas ou peçam que o trabalho seja refeito, é possível um retorno a essas informações constantes da base definicional, não sendo necessária uma volta aos textos originais, que nem sempre estão à disposição do terminólogo. Em vista disso, a base deve ser freqüentemente atualizada.

3.5. Redação da definição terminológica

A etapa de redação da definição terminológica é a mais complexa e custosa numa pesquisa terminológica que objetiva a elaboração de dicionários especializados, já que um bom dicionário se avalia, principalmente, pela qualidade das suas definições.
No âmbito do nosso método de trabalho, as definições são geradas levando-se em consideração algumas orientações. São elas: 1) o dicionário terminológico tem a função precípua de facilitar a comunicação, para tanto, o texto definitório deve ser suficientemente claro e completo para que o consulente entenda. Assim, ainda que os tipos de definição sejam utilizados como orientação, eles não devem subjugar o texto. Ao contrário, se tivermos de fazer concessões para que se dê o entendimento do termo-entrada, essas concessões serão feitas; 2) não estabelecemos com exatidão que a(s) característica(s) intrínseca(s) deve(m) figurar na definição e a(s) extrínseca(s) na informação enciclopédica, pois nem sempre se pode classificar com segurança o que são características (ou traços distintivos) intrínsecas e extrínsecas do conceito cujo termo está sendo definido.
Além dessas orientações, temos considerado características que já são consenso em Terminologia, mas que não explicitaremos aqui. Gostaríamos de salientar, entretanto, que uma série de procedimentos no tocante à redação da definição terminológica podem ser sistematizados, de modo a facilitar tanto a elaboração quanto a validação do texto definitório.
Juntamente com a redação da definição, costumamos redigir também as informações enciclopédicas. Ambas são tratadas de modo diferente porque normalmente a definição é um campo obrigatório do verbete, e a informação enciclopédica não. Ressaltamos, ainda, que tanto definições quanto informações enciclopédicas são validadas pelos especialistas. Depois de elaboradas e validadas, as definições e as informações enciclopédicas devem ser inseridas nos campos correspondentes da ficha terminológica, o que permitirá a edição final dos verbetes.

3.6. Edição do verbete

A edição dos verbetes nada mais é do que a seleção de alguns campos da ficha para constarem do modelo do verbete final. Via de regra, há nos verbetes informações sistemáticas (obrigatórias em todos os verbetes) e não-sistemáticas (informações não recorrentes). Um modelo de verbete que nos pareceu bastante satisfatório é o apresentado no Glossário de termos neológicos da economia, projeto coordenado pela Profa. Dra. Ieda Maria Alves, USP-SP, 1998.

4. Ambiente Colaborativo Web *e-Termos*

Importa mencionar que todas as etapas descritas acima serão implementadas num ambiente computacional denominado *e-Termos* [15 [#nb15]], um ambiente para auxiliar a pesquisa terminológica/terminográfica. Entendemos que fazer Terminologia na era da Informática significa criar um conjunto de procedimentos automatizados ou semi-automatizados que dêem suporte àquelas tarefas do trabalho terminológico citadas. (ALMEIDA, *et al*., 2006b)
Amparado pelos pressupostos da Terminologia de orientação descritiva, o e-Termos é um ambiente computacional que contempla as atividades de desenvolvimento de terminologias. Como uma aplicação *Computer Supported Collaborative Work* (CSCW), o *e-Termos* é um Ambiente Web Colaborativo, composto por seis módulos de trabalho independentes, mas inter-relacionados, cujo propósito é automatizar ou semi-automatizar as tarefas de criação e gerenciamento do trabalho terminológico.

Perfazendo desde a criação automática de corpora especializados (Módulo 0) até a distribuição e intercâmbio do conjunto de verbetes (Módulo 5), o principal diferencial do *e-Termos* está na característica colaborativa que este ambiente computacional implementa. Baseado nos processos de apoio e cooperação de um conjunto diferenciado de profissionais, o e-Termos possibilita o trabalho, a produção em conjunto e a troca de informações para melhorar as atividades de grupos de usuários com interesses e propósitos comuns. Além disso, o aspecto colaborativo permite a sistematização e o mapeamento do fluxo de atividades dos diversos profissionais envolvidos na criação de produtos terminológicos, produzindo resultados mais rápidos e fiáveis. Em outras palavras, o e-Termos permite que os diferentes integrantes de uma mesma equipe de pesquisa possam acessar, editar, atualizar, inserir e retirar informações de todos os Módulos (corpus, ontologia, fichas terminológicas, base definicional, redação de definições, edição de verbetes), bastando conectar-se à Internet, buscar a URL [16 [#nb16]] e utilizar uma senha de acesso. O mesmo procedimento pode ser utilizado para a interação com os especialistas da área-objeto, ou seja, especialistas previamente selecionados podem opinar, criticar, sugerir alterações, ratificar os dados (lista de termos, definições, etc.) por meio de acesso à Internet, num processo incremental de construção do produto terminológico.
Outras vantagens do *e-Termos* são: 1) a criação rápida e automática de corpus; 2) a possibilidade de análise qualitativa do corpus; 3) a categorização e visualização dos termos em uma ontologia; 4) a criação customizada das fichas terminológicas; 5) o gerenciamento da base definicional; 6) a redação assistida da definição terminológica; e, finalmente, 7) a edição de verbetes.

5. Estágio atual [**17 [#nb17]**] da pesquisa

As atividades relacionadas à geração do corpus em LP se iniciaram em abril/2006, ocasião em que procedemos à avaliação exploratória dos textos em LP disponíveis da Internet. Nessa primeira avaliação, obtivemos basicamente textos representativos de três gêneros: informativo, científico de divulgação e técnico-científico, com especial destaque para o segundo tipo. Os textos obtidos foram distribuídos conforme abaixo:

1. **Gênero informativo**: notícias, resenhas, reportagens, entrevistas e editoriais (revistas *IstoÉ*, *Época*, *Veja*, *Galileu* e jornal *Folha de S. Paulo*);
2. **Gênero científico de divulgação**: revistas Pesquisa (da FAPESP [18 [#nb18]]), *ComCiência* e *Ciência Hoje* (da SBPC [19 [#nb19]]), Revista da Indústria Brasileira da Confederação Nacional da Indústria (CNI), Revista da Embrapa [20 [#nb20]] ; *Scientific American Brasil*, *Espaço Acadêmico* da Universidade Estadual de Maringá (UEM), *Plástico Moderno* da Editora QD, MINAPIN [21 [#nb21]] da SUFRAMA [22 [#nb22]] , ENVOLVERDE [23 [#nb23]] da Webjornal.net Editora Ltda.
3. **Gênero técnico-científico**: dissertações; teses; resumos (Scielo-CAPES); artigos científicos; manuais; Material das Redes que surgiram a partir do Programa Nacional de Desenvolvimento da Nanociência e Nanotecnologia do MCT [24 [#nb24]] , por exemplo: a Nanoseminat (Rede Cooperativa para Pesquisa em Nanodispositivos Semicondutores e Materiais Nanoestruturados), a

RENAMI (Rede de Nanotecnologia Molecular e de Interfaces), a Rede de Nanobiotecnologia e a Rede Nacional de Pesquisa em Materiais Nanoestruturados, entre outras.

É importante destacar que muitas páginas da Internet, embora se tivessem revelado útil para a pesquisa, estavam acessíveis somente para sócios ou assinantes, inviabilizando, portanto, a obtenção dos textos. Após a avaliação exploratória dos textos em LP disponíveis na Internet, iniciamos a busca de textos impressos, os quais serão posteriormente digitalizados. Depois dessa primeira busca, constatamos o que segue:

DULLEY, R. D. Nanotecnologia e "agricultura inteligente". Informações Econômicas, v. 35, n. 7, p. 53-56, jul. 2005.
DURAN, N; MATTOSO, L.H.C.; MORAIS, P.C. Nanotecnologia: introdução, preparação e caracterização de nanomateriais e exemplos de aplicação. Artliber, 2006.
EMBRAPA: o que vem pela frente. Agroanalysis, Rio de Janeiro, v. 25, n. 4, p. E-1-E16, abr. 2005. Encarte especial: ciência e tecnologia.
GAZZONI, D. L. Nanotecnologia. Cultivar, Pelotas, v. 7, n. 79. p. 40, nov. 2005.
HERRMANN JR., P. S. de P. Relatório de Pós-Doutoramento -1. São Carlos, SP: Embrapa Instrumentação Agropecuária, 2003. Relatório de Pós-Doutoramento. Período: junho/2002 a dezembro/2003. Orientador: Prof. Dr. Alan G. MacDiarmid.
HERRMANN JUNIOR, P. S. de. O desafio de manipular os átomos. Ciência e Tecnologia, n. 6, mar. 2005, p. 10. Edição especial.
SIMPÓSIO BRASILEIRO DE ELETROQUÍMICA E ELETROANALÍTICA, 14., Teresópolis, 2004. A eletroquímica na interface da nanociência e nanotecnologia. Abstracts.Teresópolis, 2004.
TOMA, H.E. O mundo nanométrico: a dimensão do novo século. Oficina de Textos, 2006.

Pela relação acima, é possível perceber que a quantidade é muito pequena. Há apenas 2 livros neste grupo, o restante é composto de 5 artigos e 1 relatório. Evidentemente, será necessário insistir na busca por mais textos impressos. A título de comparação, no corpus em LI foram incluídos 22 livros.
A compilação das fontes provenientes da Internet foi iniciada em maio/2006. Com relação aos gêneros enumerados acima, o que está se revelando menos representativo é o técnico-científico, conforme mencionamos no início deste artigo, ao contrário do corpus em LI, cuja característica era a presença maciça de textos pertences a esse gênero.
Paralelamente à busca por materiais digitalizados e impressos, foram feitos ajustes do editor de cabeçalho de forma que todos os campos pudessem ser preenchidos satisfatoriamente. Apresentamos, nas figura 1, 2 e 3 a seguir, algumas telas do editor.

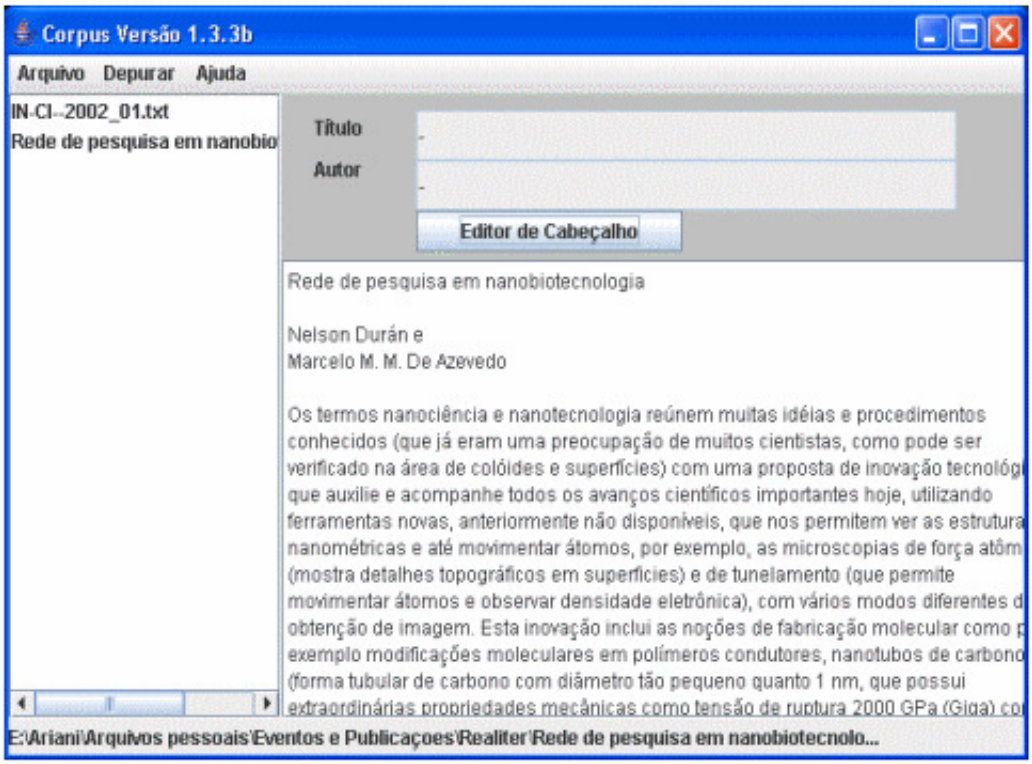

Figura 1: Editor de cabeçalho adaptado do projeto Lacio-Web

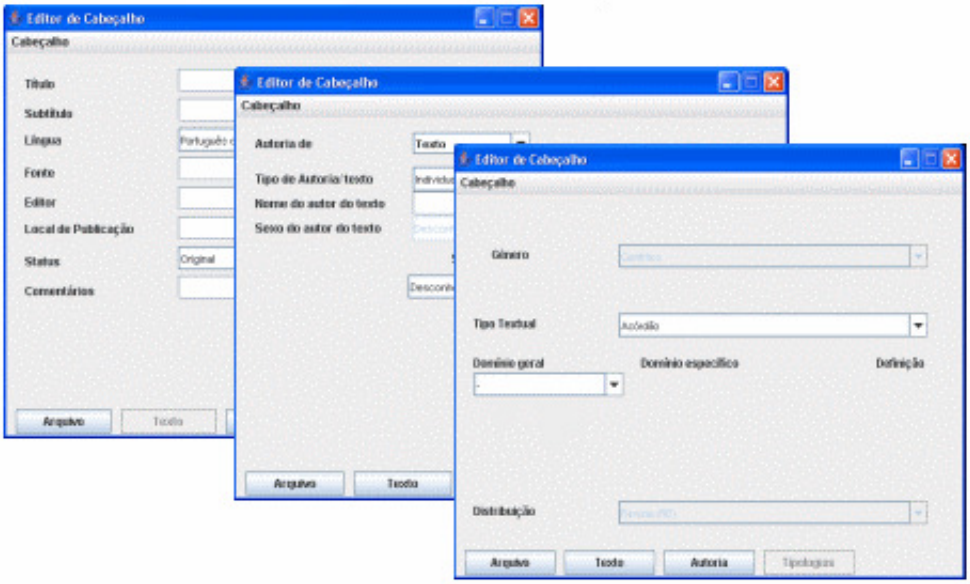

Figura 2: Janelas do editor para a especificação de diversas informações

Para cada texto, é gerado um cabeçalho. É possível ver na figura 3 como ficam as informações anotadas em XML:

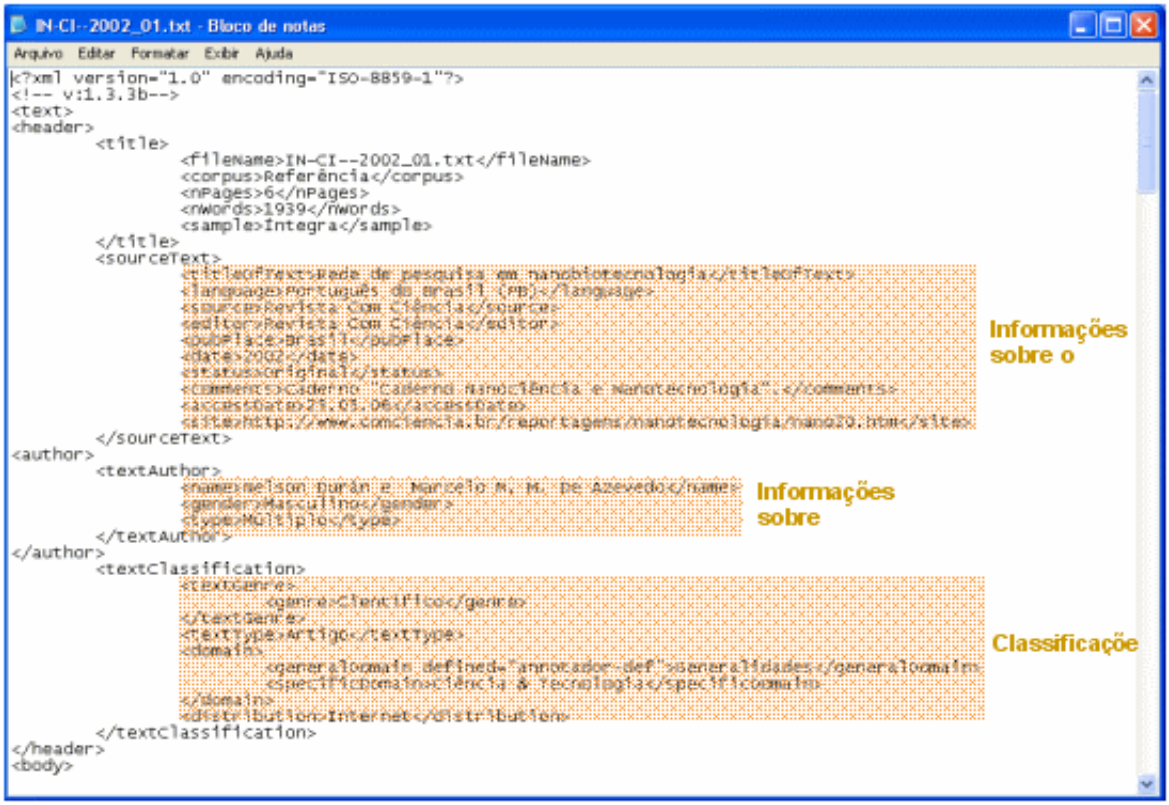

Figura 3: Texto em "txt" gerado pelo Editor de Cabeçalho

6. Resultados esperados

Em nosso cronograma de atividades, reservamos três meses para a compilação e manipulação do corpus. Em vista da pequena amostra de textos que já compilamos, nossa hipótese é de que o corpus em LP não chegará a um milhão de palavras, consistindo, portanto, em menos da metade do tamanho do corpus em LI. Ainda assim, esperamos: 1) constituir um corpus em LP de N&N, cujo tamanho e representatividade permita a obtenção de equivalentes para os termos obtidos no corpus em inglês, os quais compõem a ontologia em inglês da N&N; 2) elaborar a ontologia em língua portuguesa de N&N; 3) elaborar o primeiro dicionário-piloto de N&N em língua materna.

7. Referências bibliográficas

ALMEIDA, G. M. B. *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, Campus de Araraquara, Universidade Estadual Paulista, 2000.

ALMEIDA, G.M.B.; ALUÍSIO, S.M.; OLIVEIRA, L.H.M. O *método em Terminologia: revendo alguns procedimentos*. In: ISQUERDO, A.N.; ALVES, I.M. *Ciências do léxico: lexicologia, lexicografia, terminologia*, vol. III, Campo Grande: Editora da UFMS, 2006a (no prelo).

ALMEIDA, G.M.B, OLIVEIRA, L.H.M. e ALUISIO, S.M. *A terminologia na era da informática. Cienc. Cult*. [online]. abr./jun. 2006b, vol.58, no.2, p.42-45. Disponível em: http://cienciaecultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252006000200016&lng=pt&nrm=iso [http://cienciaecultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252006000200016&lng=pt&nrm=iso].

ALUISIO, S.M., OLIVEIRA JR, O.N., ALMEIDA, G.M.B., NUNES, M.G.V., OLIVEIRA, L.H.M., FELIPPO, A.D., ANTIQUEIRA, L., GENOVES JR., L.C., CASELI, L., ZUCOLOTTO, L. E SANTOS JR., D.S. *Desenvolvimento de uma estrutura conceitual (ontologia) para a área de Nanociência e*

Nanotecnologia, Relatório Técnico do ICMC n.o 276, 182 p., 2006.

ALVES, I. M. *Glossário de termos neológicos da economia*. São Paulo: Humanitas, FFLCH, USP, 1998.

CABRÉ, M.T. *La terminología: representación y comunicación – elementos para una teoría de base comunicativa y outros artículos*. Barcelona: Institut Universitari de Lingüística Aplicada, 1999.

CABRÉ, M.T. *Theories of Terminology: their description, prescription and explanation*. Terminology, v. 9, n. 2, 2003, p. 163-200.

KRIEGER, M.G.; BEVILACQUA, C.R. A pesquisa terminológica no Brasil: uma contribuição para a consolidação da área. *Debate terminológico*, no. 1, 03/2005. Disponível em http://www.riterm.net/ [http://www.riterm.net/] revista.

Ministério da Ciência e Tecnologia - Grupo de Trabalho criado pela portaria MCT n° 252, de 16.05.2003. *Desenvolvimento da Nanociência e da Nanotecnologia*. Disponível em: http://www.mct.gov.br/temas/nano/prog_nanotec.pdf [http://www.mct.gov.br/temas/nano/prog_nanotec.pdf]

Organização dos Estados Americanos - Secretaria Executiva para o Desenvolvimento Integral - Escritório de Educação, Ciência e Tecnologia. Ciência, Tecnologia, Engenharia e Inovação para o Desenvolvimento: uma visão para as Américas no Século XXI. 2ª. ed., nov/2005. Disponível em: http://www.science.oas.org/Ministerial/ingles/documentos/portugues_web.pdf. [http://www.science.oas.org/Ministerial/ingles/documentos/portugues_web.pdf.]

[1 [#nh1]] Grupo de Estudos e Pesquisas em Terminologia (GETerm), Universidade Federal de São Carlos (UFSCar), São Carlos, SP, Brasil

[2 [#nh2]] Núcleo Interinstitucional de Lingüística Computacional (NILC), Universidade de São Paulo (USP), São Carlos, SP, Brasil

[3 [#nh3]] Núcleo Interinstitucional de Lingüística Computacional (NILC), Universidade de São Paulo (USP), São Carlos, SP, Brasil

[4 [#nh4]] Núcleo Interinstitucional de Lingüística Computacional (NILC), Universidade de São Paulo (USP), São Carlos, SP, Brasil

[5 [#nh5]] Núcleo Interinstitucional de Lingüística Computacional (NILC), Universidade Estadual Paulista (UNESP), Araraquara, SP, Brasil

[6 [#nh6]] Núcleo Interinstitucional de Lingüística Computacional (NILC), Universidade de São Paulo (USP), São Carlos, SP, Brasil

[7 [#nh7]] Grupo de Estudos e Pesquisas em Terminologia (GETerm), Universidade Federal de São Carlos (UFSCar), São Carlos, SP, Brasil

[8 [#nh8]] Grupo de Estudos e Pesquisas em Terminologia (GETerm), Universidade Federal de São Carlos (UFSCar), São Carlos, SP, Brasil

[9 [#nh9]] http://www.usp.br/prp/nanotecnologia/ [http://www.usp.br/prp/nanotecnologia/]

[10 [#nh10]] http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm

[http://www.nilc.icmc.usp.br/nilc/tools/corpora.htm]

[11 [#nh11]] http://www.nilc.icmc.usp.br/lacioweb/ [http://www.nilc.icmc.usp.br/lacioweb/]

[12 [#nh12]] A descrição de todas essas etapas baseou-se em Almeida et al., 2006a.

[13 [#nh13]] **Precisão** é a razão das respostas corretas recuperadas pelo sistema e todas as respostas recuperadas e **Revocação** é a razão de respostas corretas e todas as respostas corretas possíveis.

[14 [#nh14]] Nesse caso, são considerados também como fonte de excerto obras e/ou artigos em outras línguas: inglês, espanhol, francês, por exemplo. Ressaltamos que o corpus em LI será muito útil para a obtenção de excertos (definitórios e/ou explicativos).

[15 [#nh15]] O projeto está sendo desenvolvido por Leandro Henrique Mendonça de Oliveira, como tese de doutorado em Ciências de Computação e Matemática Computacional, com orientação de Sandra Maria Aluísio na Universidade de São Paulo, campus de São Carlos, SP, Brasil. O e-Termos resultou de um Projeto FAPESP (proc. no. 2003/06569-3) coordenado por Gladis Maria de Barcellos Almeida e intitulado Extração automática de termos e elaboração colaborativa de terminologias para o intercâmbio de conhecimento especializado – TermEx. Para mais informações sobre o e-Termos, consultar: http://www.nilc.icmc.usp.br/etermos/ [http://www.nilc.icmc.usp.br/etermos/]

[16 [#nh16]] "[Sigla do ingl. u(niform) (ou, originalmente, universal) r(esource) l(ocator), 'localizador uniforme (ou universal) de recursos'.] Sigla que designa a localização de um objeto na internet, segundo determinado padrão de atribuição de endereços em redes." (Novo Dicionário Eletrônico Aurélio, versão 5.0)

[17 [#nh17]] Correspondendo a maio de 2006

[18 [#nh18]] Fundação de Amparo à Pesquisa do Estado de São Paulo.

[19 [#nh19]] Sociedade Brasileira para o Progresso da Ciência.

[20 [#nh20]] Empresa Brasileira de Pesquisa Agropecuária.

[21 [#nh21]] Revista de Micro e Nanotecnologia do Pólo Industrial de Manaus.

[22 [#nh22]] Superintendência da Zona Franca de Manaus.

[23 [#nh23]] Revista Digital de Meio Ambiente e Desenvolvimento.

[24 [#nh24]] Ministério da Ciência e Tecnologia. [Ministério da Ciência e Tecnologia.]