Compilação de *Corpus* em Língua Portuguesa na Área de Nanociência/Nanotecnologia: Problemas e Soluções

Joel Sossai Coleti[*]
Daniela Ferreira de Mattos[*]
Luiz Carlos Genoves Jr.[**]
Arnaldo Candido Jr.[****]
Ariani Di Felippo[***]
Gladis Maria de Barcelos Almeida[*]
Sandra Maria Aluísio[****]
Osvaldo Novais de Oliveira Jr.[****]

Resumo: A compilação de um *corpus* em língua portuguesa de Nanociência e Nanotecnologia (N&N) é parte integrante de um projeto maior intitulado "Terminologia em Língua Portuguesa da Nanociência e Nanotecnologia: Sistematização do Repertório Vocabular e Elaboração de Dicionário-Piloto – NANOTERM" (CNPq, processo 400506/2006-8), que tem como objetivos: 1) a constituição de um *corpus* em língua portuguesa (LP) da N&N; 2) a elaboração de uma terminologia, também em LP; 3) a construção de uma ontologia; 4) a elaboração do primeiro dicionário-piloto de N&N em LP. De forma a garantir o cumprimento desses objetivos, foi compilado um *corpus* com mais de dois milhões de palavras em LP da área de N&N. Para a efetivação do trabalho, a equipe baseou-se nos fundamentos metodológicos da Lingüística de *Corpus*, que prevêem uma série de etapas que compõem a elaboração de *corpus*, a saber: seleção, compilação e manipulação dos textos, nomeação dos arquivos compilados, anotação (estrutural) e pedidos de permissão de uso. Ainda dentro dos princípios da Lingüística de *Corpus*, consideraram-se os seguintes requisitos: autenticidade, representatividade, balanceamento e diversidade. O *corpus*, depois de pronto, foi processado por ferramentas computacionais, tais como: extrator semi-automático de termos, concordanciador e contador de freqüência. Durante a elaboração do *corpus*, alguns problemas se apresentaram no que se refere à falta de balanceamento em relação aos distintos gêneros textuais, à adequação de ferramentas computacionais ao *corpus* compilado e à dificuldade de obtenção de textos e de permissões de uso. Neste artigo, pretende-se expor todas as etapas percorridas, os problemas enfrentados e as soluções adotadas para saná-los, de forma a auxiliar outras equipes de pesquisa que têm trabalhado com compilação de *corpus*.

Palavras-chave: compilação de *corpus*; Nanociência; Nanotecnologia.

Abstract: The compilation of a Portuguese corpus for Nanoscience and Nanotechnology (N&N) is part of a larger project entitled "Terminology of Nanoscience and Nanotechnology in the Portuguese Language: Systematization and Development of a Pilot Dictionary – NANOTERM" (CNPq proc. 400506 / 2006-8), which has as the main objectives: 1) compilation of a Portuguese corpus for N&N; 2) development of a terminology, also in Portuguese, 3) construction of an ontology; 4) preparation of the first pilot dictionary of N&N in Portuguese. In order to achieve these goals, a corpus of over two million words was compiled for the area of N&N, following the principles of Corpus Linguistics, a research area that provides steps to guide corpus compilation, namely: text selection, text gathering,

[*] Universidade Federal de São Carlos – UFSCar, São Carlos, SP.
[**] Google.
[***] Universidade Estadual Paulista – UNESP, Araraquara, SP.
[****] Universidade de São Paulo – USP, São Carlos, SP.

undesirable metadata cleaning, proper file naming, structural text annotation and requests for permission to use. We also considered the following requirements to guide corpus compilation: use of authentic texts, corpus representativeness and balancing, and text diversity. The corpus has been processed with computational tools, including semi-automatic term extraction, concordancing tools and frequency counters. While preparing the corpus, the main problems encountered were the lack of balance with regard to different text genres, inadequacy of computational tools to process the corpus and the difficulty in obtaining documents and their permissions for use. In this article we discuss these problems and the solutions adopted, with a detailed description of procedures for the compilation of the corpus, which may serve as guidance to other corpus compilation projects.

Key-words: corpus compilation; Nanoscience; Nanotechnology.

1. Introdução

Nano- é um prefixo "que, anteposto ao nome de uma unidade de medida, indica uma unidade derivada igual a 10^{-9} vezes a primeira" (Ferreira, 2004). É comumente utilizado na notação de tempo e comprimento, como 30 nano-segundos e 100 nanometros, por exemplo. Quando utilizado como um prefixo anteposto a algo diferente de uma unidade de medida, como em *nanociência e nanotecnologia*, *nano-* pode ser relacionado à escala dos nanômetros. Um nanômetro vale $1{,}0\times10^{-9}$ metros (ou um milionésimo de milímetro) e tem como símbolo *nm*.

Apenas para que se tenha uma idéia dessa pequeneza, o diâmetro de um fio de cabelo humano mede cerca de 30.000 nanômetros, já "um minúsculo vírus, invisível a olho nu, se apresenta como uma incrível entidade com cerca de 200 nm." (Toma & Araki, 2005). É desse mundo "do muito pequeno" que trata o domínio da N&N, um dos mais promissores campos de pesquisa da ciência, que prometem uma verdadeira revolução tecnológica.

Para lidar com objetos tão pequenos, além de equipamentos modernos caros e complexos, são necessárias metodologias específicas para manipular partículas de tamanhos tão reduzidos (Knobel, 2005). Esse universo de tamanhos reduzidos perpassa distintos saberes, por isso que a N&N firma-se como um campo amplo e interdisciplinar que envolve a Química, a Física, a Bioquímica, a Biofísica, a Engenharia de Materiais, a Ciência da Computação e a Medicina.

Segundo documento elaborado pelo Grupo de Trabalho criado pela portaria Ministério da Ciência e Tecnologia (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.

Assim, a escolha do domínio da N&N justifica-se por constituir-se num conjunto de saberes e tecnologias relativamente recentes e, por isso, sua terminologia ainda estar em fase de construção, sobretudo no que diz respeito à língua portuguesa, variante brasileira.

2. Compilação do *corpus*

A descrição da linguagem a partir da análise de um c*orpus* com vistas à elaboração de dicionário é uma atividade bastante comum nas pesquisas em Lingüística de *Corpus*. As pesquisas lexicográficas e terminográficas dedicam-se ao estudo e descrição do comportamento morfossintático, léxico e semântico de lexias no meio lingüístico em que ocorrem. São essas evidências empíricas que, resgatadas com recursos de Processamento de Língua Natural (PLN), auxiliarão na composição de dicionários.

O *corpus* desenvolvido é, portanto, essencial para diversas etapas constitutivas do processo de criação do dicionário-piloto de N&N, tais como: a extração dos candidatos a termos, a elaboração da ontologia, o preenchimento das fichas terminológicas e a redação das definições.

Embora existam muitos *corpora* disponíveis tanto livremente como mediante pagamento (a partir dos quais se pode gerar um sub*corpus* de estudo ou mesmo tomar o *corpus* todo como uma unidade, dependendo da questão de pesquisa), ainda pode ser necessário compilar[1] um *corpus* próprio. Para a compilação de tal *corpus*, existem três estágios principais a seguir: 1) projeto do *corpus*, que inclui a seleção dos textos e os cuidados com os requisitos estabelecidos pela Lingüística de *Corpus*, no caso desta pesquisa, consideraram-se: autenticidade, representatividade, balanceamento e diversidade (Biber *et al.*, 1998; Berber Sardinha, 2000; Sinclair, 2005); 2) compilação, manipulação, nomeação dos arquivos de textos e pedidos de permissão de uso; e 3) anotação. Esses estágios foram realizados, obedecendo à seguinte seqüência:

1. seleção de textos (a partir da Web e a partir de textos impressos) e pedidos de permissão de uso;
2. manipulação e limpeza do *corpus*;
3. etiquetagem (anotação estrutural, cabeçalhos) e nomeação dos arquivos;

[1] A compilação é o armazenamento em arquivos dos textos selecionados para a composição do *corpus.* (Aluísio & Almeida, 2006)

4. exclusão das etiquetas XML[2] para processamento do *corpus* e contagem de palavras.

A seguir, será explicitado cada um dos itens indicados acima.

2.1. Seleção dos textos

A seleção dos textos consiste basicamente em definir quais os textos são pertinentes e relevantes para a pesquisa, sempre levando em conta os requisitos autenticidade, representatividade, balanceamento e diversidade.

Existem vários tipos de *corpus*, dentre eles, nesse momento vale ressaltar dois: o falado, composto de porções de fala transcritas; e o escrito, composto de textos escritos, falados ou não (Berber Sardinha, 2000). Para o *corpus* da N&N, foram escolhidos textos escritos.

A seleção de textos foi feita em meios digitais e impressos. Optou-se preferencialmente por fontes disponibilizadas na *Web* devido ao demorado e custoso trabalho de digitalização de material impresso. Os procedimentos são descritos a seguir.

2.1.1 Seleção de textos a partir da Web

Por meio de motores de busca, realizaram-se as pesquisas orientadas por palavras-chaves previamente definidas e posteriormente alteradas almejando-se melhor adequação aos objetivos (cf. item 3 "Problemas e soluções"). Como motor de busca, adotou-se o *Google*[3].

Os resultados gerados pelo *Google* foram previamente selecionados. Descartaram-se os resultados sem domínio próprio. Aqueles com servidores próprios passaram por uma nova seleção, na qual se verificou a procedência e a confiabilidade das fontes. Privilegiaram-se *sites* de instituições públicas, de grandes instituições privadas, de empresas de comunicações conceituadas, por serem considerados confiáveis.

---

[2] E**X**tensible **M**arkup **L**anguage. Trata-se de uma linguagem de marcadores que serve para descrever dados. A sua grande vantagem é que ela é extensível, ou seja, não há um número limitado de etiquetas de marcação, é possível criar outras etiquetas para anotar o que for necessário, daí ela ser considerada uma linguagem autodefinível. (Moacir Casemiro, 2004. Disponível em: http://www.codificando.net/. Acesso em: 13/04/2007)

[3] www.google.com.br.

Pelo procedimento "copia-e-cola", os textos foram capturados das fontes na *Web* e salvos diretamente em arquivos do programa *Bloco de Notas*[4]. Ressalte-se que o uso mais comum do *Bloco de Notas* é exibir ou editar arquivos de formato texto (.txt). Todos os arquivos compilados foram salvos em servidor local do laboratório do Grupo de Estudos e Pesquisas em Terminologia (GETerm)[5].

2.1.2 Seleção e digitalização de textos impressos

Fontes impressas consideradas relevantes para as pesquisa foram digitalizadas, compiladas e integradas ao *corpus*. Todos os livros originalmente escritos em língua portuguesa de que se tinha conhecimento foram digitalizados, totalizando 4 obras, a saber:

1. DURAN, N; MATTOSO, L.H.C; MORAIS, P.C. *Nanotecnologia: introdução, preparação e caracterização de nanomateriais e exemplos de aplicação.* São Paulo: Artliber, 2006.
2. ALVES, E. G.; CHAVES, A. S.; VALADARES, E. C. *Aplicações da física quântica do transistor à nanotecnologia*. São Paulo: Editora Livraria da Física. 2005.
3. TOMA H. E. *O Mundo Nanométrico: A Dimensão do Novo Século.* São Paulo: Oficina de Textos. 2004.
4. CNI/SENAI. *Nanotecnologias. Série ocupações emergentes.* nº 1. Brasília, 2004.

Para a digitalização dessas obras, utilizou-se o software *ABBYY FineReader 8.0 Professional Edition* (doravante *ABBYY*), sistema de reconhecimento óptico de caracteres (*OCR*, na sigla em inglês para *Optical Character Recognition*) baseado na premiada tecnologia da *ABBY Software*[6], cujos procedimentos serão descritos a seguir.

---

[4] Bloco de Notas (ou *Notepad*, originalmente, em inglês) é um editor de textos disponível em todas as versões do sistema operacional *Microsoft Windows* desde a versão 1.0 em 1985. Pode ser acessado a partir do menu "Iniciar", opção "Executar", digitando-se "*notepad*" (com ou sem aspas) e selecionando o botão "ok".

[5] O GETerm está localizado no Departamento de Letras da Universidade Federal de São Carlos (UFSCar, SP) e tem como objetivos: 1) estudar conteúdos pertinentes à Terminologia/Terminografia; 2) desenvolver pesquisas que gerem produtos terminológicos em língua portuguesa, tais como: glossários, dicionários, enciclopédias e assemelhados, que satisfaçam demandas reais.

[6] Mais informações podem ser obtidas em: http://www.abbyy.com/finereader8/?param=44890

O procedimento inicial é chamado pelo *ABBYY* de *Scan&Read* e consiste em: "Escanear" (digitalizar a impressão transformando-a em imagem no computador); "Ler Tudo" (reconhecimento óptico de caracteres); "Corrigir Ortografia" ou "Verificar" (momento em que o *ABBYY FineReader* solicita também que se corrijam caracteres sobre os quais haja dúvida, incerteza, ou ainda, imprecisão, no reconhecimento óptico dos caracteres) e, por fim, "Salvar" (salva o arquivo em forma de texto no formato definido pelo usuário), como não é possível salvar em *Bloco de Notas* (.txt) diretamente do *ABBYY*, opta-se por salvar em *Microsoft Office Word* (.doc) e, posteriormente, faz-se a conversão de formatos, pelo procedimento "copia-e-cola".

Para o escaneamento ou digitalização da material impresso, o *ABBYY* utiliza-se de software do próprio escâner. Para o processo de digitalização, utilizou-se o escâner *HP scanjet 3670*.

2.2. Manipulação e limpeza do *corpus*

Para que os textos, em seus formatos originais de disponibilização, pudessem ser corretamente processados por ferramentas computacionais de PLN, fez-se necessário converter todos os arquivos de texto de seus formatos originais (*Microsoft Word* de extensão ".doc", *HyperText Markup Language* de extensão ".html"[7], *Portable Document Format* de extensão ".pdf"[8] e outros) para um único formato padrão. Optou-se pelo formato de destino *Bloco de Notas* de extensão ".txt", pois este não possui códigos de formatação, mas apenas caracteres do teclado (letras, números e símbolos ortográficos).

O procedimento de conversão de formatos é manual, ou seja, vale-se do procedimento "copia-e-cola", exceto para a conversão dos arquivos *Portable Document Format* de extensão ".pdf",

[7] Linguagem de programação utilizada para produzir páginas na *Web*. Documentos *HTML* podem ser interpretados por navegadores de Internet.

[8] Este formato de arquivo tem como principal característica representar um documento com integridade e fidelidade ao seu formato original independentemente do aplicativo, do computador e do sistema operacional usados para criá-lo. Para visualizar esses arquivos é necessária a utilização do programa *Adobe Acrobat Reader* de distribuição gratuita pelo fabricante.

para os quais se utilizou um procedimento automático por meio do programa XPDF[9] (disponível apenas em ambiente *Linux*, ou em ambientes *Linux* emulado[10] em *Windows*) que realiza a conversão dos formatos automaticamente[11].

Os textos, já em formato ".txt", foram submetidos à limpeza, ou seja, foram excluídos tabelas, gráficos, fórmulas, cálculos, fotos e toda informação que não estivesse em forma de texto. Foi feita também a formatação dos textos, conferindo padronização ao *corpus*.

2.3. Anotação estrutural, geração de cabeçalhos e nomeação de arquivos

Por meio da versão adaptada do Editor de Cabeçalhos[12] do Projeto Lacio-Web[13], foi possível realizar a anotação estrutural dos textos, adicionando-se em cada texto um cabeçalho e uma nomeação específica e padronizada. Esses procedimentos são descritos a seguir.

2.3.1. Anotação estrutural

De acordo com Aluísio & Almeida,

> A *anotação estrutural* compreende a marcação de dados externos e internos dos textos. Como dados externos entendemos a documentação do *corpus* na forma de um cabeçalho que inclui os metadados textuais (ou dados estruturados sobre dados), isto é, dados bibliográficos comuns, dados de catalogação como tamanho do arquivo, tipo da autoria, a tipologia textual e informação sobre a distribuição do *corpus*. Como dados internos temos a anotação de segmentação do texto cru, que envolve: a) marcação da estrutura geral – capítulos, parágrafos, títulos e subtítulos, notas de rodapé e elementos gráficos como tabelas e figuras, e b) marcação da estrutura de subparágrafos – elementos que são de interesse lingüístico, tais como sentenças, citações, palavras,

[9] Disponível na Internet em: http://www.foolabs.com/xpdf/index.html .

[10] Segundo o dicionário Ferreira (2004), comportar-se (programa ou equipamento) como (outro), aceitando as mesmas entradas e produzindo as mesmas saídas, ainda que não com a mesma velocidade ou pelos mesmos processos.

[11] Este procedimento é possível, entretanto, apenas com "arquivos *lives*", não podendo ser aplicado a arquivos criptografados.

[12] Para o projeto NANOTERM, utilizou-se a versão do Editor de Cabeçalhos do Projeto Lácio-*Web* adaptado por Luiz Carlos Genoves Jr.

[13] O projeto Lacio-Web tem como objetivo divulgar e disponibilizar livremente na Web *corpus* do português brasileiro e ferramentas lingüístico-computacionais. O *corpus* do Lacio-Web é formado por 6 diferentes sub*corpora* e abrange os gêneros: Informativo, Cientifico Instrucional, Jurídico e Literário. O público-alvo do projeto é heterogêneo: de um lado lingüistas, cientistas da computação, lexicógrafos, e de outro lado o público em geral. Maiores informações podem ser obtidas em: http://www.nilc.icmc.usp.br/lacioweb/index.htm

> abreviações, nomes, referências, datas e ênfases tipográficas do tipo negrito, itálico, sublinhado, etc. (Aluísio & Almeida, 2006)

A anotação estrutural do texto é um processo pelo qual parcelas de texto que constituem informações diferenciadas por sua relevância (ou pela falta de relevância)[14] são marcadas por etiquetas[15] por meio do Editor de Cabeçalhos.

A seguir serão reproduzidas 2 telas do Editor, para que se possa ter idéia do seu aspecto visual. As etapas para a anotação dos textos utilizando o Editor são as seguintes:

1) Abre-se o texto no editor de cabeçalho (figura 1) e localiza-se o trecho a ser etiquetado (em destaque na figura 2).

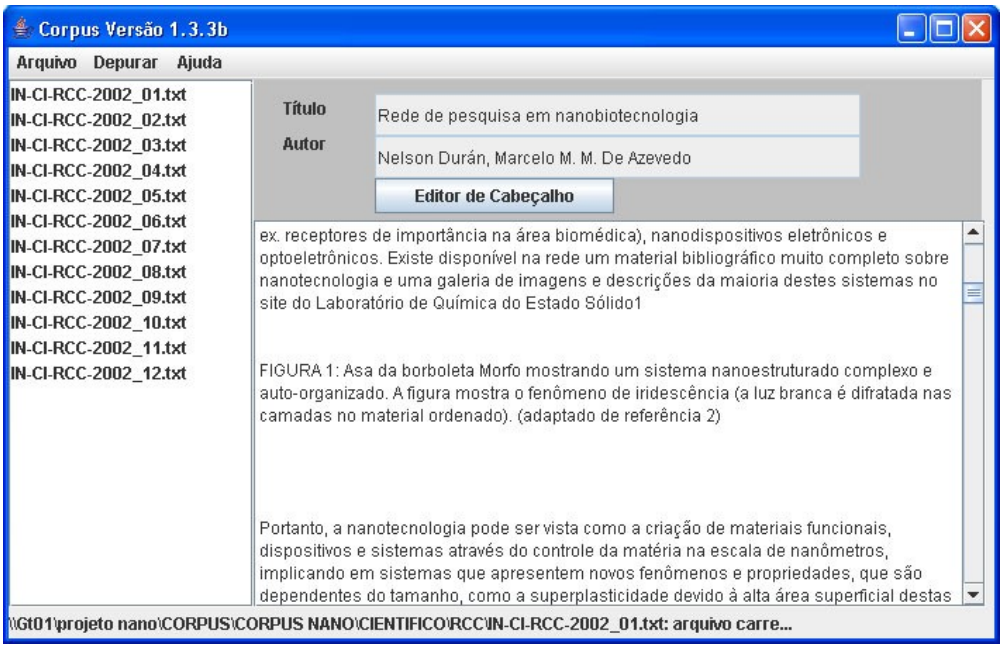

Figura 1: Visão geral do editor

[14] Por exemplo: título, resumo e legendas constituem informações relevantes; enquanto que *abstract* e referências bibliográficas constituem informações irrelevantes para o projeto NANOTERM.

[15] Código de formatação próprio da Computação que tem como função inserir uma marcação específica para cada trecho do texto de forma que a especificidade desse trecho possa ser reconhecida por outros programas de PLN.

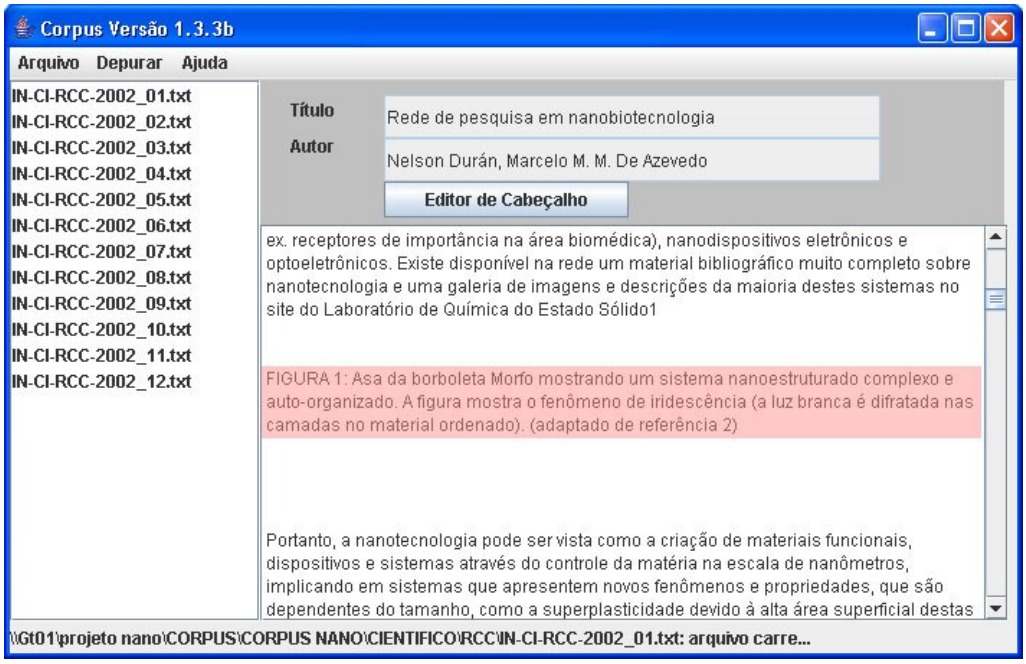

Figura 2: Seleção de parcela de texto

2) Seleciona-se o trecho a ser etiquetado. Sobre a seleção, clica-se com o botão direito do *mouse*. Um menu será aberto. Deve-se escolher "Marcar com a tag...". Uma lista de opções será aberta, clica-se então na opção desejada, que pode ser: gráfico, figura, tabela, fórmula, legenda, etc.
3) O editor de cabeçalho automaticamente gerará uma etiqueta (em linguagem XML de programação) capaz de ser reconhecida por outros programas de PLN.
4) Esta etiqueta será salva no próprio arquivo ".txt" junto com o texto original, sendo reconhecida por leitores de textos comuns como se fosse uma nova palavra.

2.3.2. Geração de cabeçalhos

O mesmo editor foi usado para a criação dos cabeçalhos. Fazem parte do cabeçalho as seguintes informações: Título, Subtítulo, Editor, Fonte, Local de Publicação, Data, Status (revelando se o texto foi escrito originalmente em língua vernácula ou se é tradução), Comentários (informações adicionais), Autoria (Individual, Múltipla, Institucional, Desconhecida, Orientador, Sexo do Autor) Gênero Textual, Tipo de Texto e Meio de

Distribuição. Essas informações mostram-se relevantes para o NANOTERM bem como para outros projetos que pretendam reutilizar o *corpus*.

Reproduzem-se, abaixo, a título de exemplo, todas as informações que podem ser recuperadas do texto "Rumo a Nanotecnologia Global" a partir de seu cabeçalho:

Texto: **Rumo a Nanotecnologia Global**
Nome do arquivo: **IN-IF-AF-not-07nov06**
Numero de Palavras: **206**
Amostra: **Íntegra**
Língua: **Português do Brasil (PB)**
Fonte: **Agência Fapesp**
Local de Publicação: **São Paulo**
Data: **07 de novembro de 2006**
Status: **Original**
Comentários: **Caderno "Notícias"**
Data de Acesso: **16 de novembro de 2006**
Endereço Eletrônico: **http://www.agencia.fapesp.br/boletim_dentro.php?id=6307**
Tipo de Autoria: **Individual**
Nome do Autor do Texto: **Thiago Romero**
Sexo do autor: **Masculino**
Gênero: **Informativo**
Subgênero: **Jornalístico**
Tipo Textual: **Reportagem**
Domínio Geral: **Generalidades**
Domínio Específico: **Ciência & Tecnologia**
Definição: **Anotador**
Distribuição: **Internet (IN)**

2.3.3. Nomeação dos arquivos

A nomeação dos arquivos (gerada de forma manual, semi-automática ou automática) pelo pesquisador ou pelo Editor de Cabeçalhos atende à padronização estabelecida pelo Projeto Lácio-*Web*. Essa padronização prevê a nomeação dos arquivos por meio de siglas, que remetem às informações de Meio de Divulgação, Gênero Textual, Fonte, Data de Publicação, nesta ordem. As siglas são previamente definidas e inseridas no banco de dados do Editor de Cabeçalhos. A seguir, apresentar-se-á um exemplo de nomeação para cada gênero que compõe o *corpus*.

→ para textos científicos:

IN-CI-Gomes-01abr03_17
IN: Texto divulgado pela Internet
CI: Gênero textual Científico
Gomes: Sobrenome do Autor
01abr03: Data de publicação (01 de abril de 2003)
_17: 17° texto obtido da mesma fonte (Banco de Teses da Capes)

→ para textos científicos de divulgação:

IN-CD-INOVATEC-nanotec-05jul06
IN: Texto divulgado pela Internet
CD: Gênero textual Científico de Divulgação
INOVATEC: Sigla que representa a Fonte (Inovação Tecnológica)
nanotec: Sigla que representa o caderno/seção em que o texto foi publicado na fonte (Seção Nanotecnologia)
05jul06: Data de publicação (05 de julho de 2006)

→ para textos informativos:

IN-IF-REVGAL-cadtec-01set05_02
IN: Texto divulgado pela Internet
IF: Gênero textual Informativo
REVGAL: Sigla que representa a Fonte (Revista Galileu)
cadtec: Sigla que representa o caderno/seção em que o texto foi publicado na fonte (Caderno Tecnologia)
01set05: Data de publicação (01 de setembro de 2005)
_02: 2° texto obtido da mesma fonte

→ para textos técnico-administrativos:
IN-TA-MCT-2004_03
IN: Texto divulgado pela Internet
TA: Gênero textual Técnico-Administrativo
MCT: Sigla que representa a fonte (Ministério da Ciência e Tecnologia)
2004: Data de Publicação
_03: 3° texto obtido da mesma fonte

2.4. Exclusão das etiquetas XML para contagem e processamento do *corpus*

Com a finalização do *corpus,* era preciso contar o número de palavras que ele continha. Para isso, não era recomendável considerar as etiquetas XML inseridas pelo Editor de Cabeçalhos, pois o programa contaria todas as palavras que estão entre as etiquetas como sendo palavras do *corpus*. Dessa forma, foi preciso remover tais etiquetas a fim de deixar o *corpus* enxuto. Uma outra razão para a exclusão das etiquetas é o fato de que muitos programas de PLN não

processam com facilidade (e alguns simplesmente não processam) *corpus* com etiquetas XML. Então, para se excluir automaticamente todas as etiquetas XML assim como seus conteúdos (quando indesejados, por exemplo: *abstract*, referências bibliográficas, etc.), valeu-se do programa *PROcessamento de TExtos históricos em Java* – PROTEJ[16], adaptado para o NANOTERM por Arnaldo Cândido Júnior (NILC)[17]. Criou-se, então, uma "versão espelho" do *corpus*, ou seja, um *corpus* idêntico ao concluído, mas sem as etiquetas XML. Essa versão espelho é que foi utilizada para contar[18] as palavras bem como para efetuar as primeiras experiências com os programas de extração semi-automática de candidatos a termo.

3. Problemas e soluções

Durante a elaboração do *corpus*, alguns problemas se apresentaram no que se refere à falta de balanceamento em relação aos distintos gêneros textuais, à adequação/criação de ferramentas computacionais ao *corpus* compilado e à dificuldade de obtenção de textos e de permissões de uso. A seguir, serão explicitados cada um dos problemas seguidos das soluções adotadas.

3.1. Falta de balanceamento em relação aos distintos gêneros textuais

As buscas de textos na *web,* iniciadas pelas palavras-chaves: nanociência, nanotecnologia e genômica produziam resultados indesejados, principalmente aqueles obtidos nos resultados pela busca da palavra-chave "gênomica", já que os resultados não possuíam relação com a área de N&N, o que levou ao abandono desta palavra-chave para as buscas. Já os resultados obtidos pelas buscas das palavras-chaves nanociência e nanotecnologia remetiam majoritariamente a textos classificados em gênero como "Informativos".

---

[16] CANDIDO JR, A. Criação de um Ambiente para o Processamento de Córpus de Português Histórico. Qualificação (Mestrado) — Instituto de Ciências Matemáticas e de Computação de São Carlos, USP, abril 2007.
[17] Mestrando ICMC-USP. Página pessoal: http://www.nilc.icmc.usp.br/nilc/pessoas/arnaldo.htm
[18] A contagem de palavras foi feita em ambiente Linux emulado pelo Cywgin através de programa especifico para esse fim. Foi possível obter automaticamente a contagem de palavra em cada texto e o valor total da soma de todos os textos.

Optou-se então por ampliar a lista de palavras-chaves com traduções livres (acrescidas de suas flexões em gênero e número) de palavras-chaves de busca adotadas pela Scielo[19], *site* reconhecido internacionalmente por excelência em distribuição de conteúdo científico. Seguem as palavras-chaves adicionadas: nanomateriais, nanocápsulas, nanocápsula, nanocompósitos, nanocristalinos, nanocristalino, nanocristalinas, nanocristalina, nanocristais, nanocristal, nanodiamantes, nanodiamante, nanoestruturados, nanoestruturado, nanoestruturadas, nanoestruturada, nanafabricação, nanofibras, nanofibra, nanofiltração, nanofitas, nanofita, nanoflagelados, nanoflagelado, nanométricos, nanométrico, nanopartículas, nanopartícula, nanopolímeros, nanopolímero, nanopós, nanosílica, nanotubos, nanotubo.

Dessa forma, os novos resultados das buscas passaram a indicar diversos textos classificados como "Científico" ou "Científico de Divulgação", o que colaborou para a diversidade do *corpus,* além da representatividade dos diversos gêneros textuais que compõem esta área de conhecimento, atendendo-se assim a 2 importantes requisitos da Lingüística de *Corpus*.

Há que se admitir, entretanto, que não há um balanceamento entre os gêneros, como se pode observar na tabela abaixo:

Gêneros	Número de palavras
Científico	1.846.763
Científico de Divulgação	310.018
Técnico-Administrativo	26.877
Informativo	361.307
Outros (prospectos de empresas e instituições de pesquisas, *slides* de palestras, etc.)	20.525
Total:	***2.565.490***

Tabela 1: Total de palavras do *corpus*

Essa discrepância entre o número de palavras por gênero reflete, na verdade, o estágio atual das produções de textos de uma área emergente, e não as falhas na seleção dos textos pelos

[19] Site da Scielo: http://www.scielo.br

motores de busca, haja vista que as pesquisas foram orientadas por palavras-chave e não por tipos de fonte.

3.2. Adequação/criação de ferramentas computacionais ao *corpus* compilado

Inicialmente, começou-se a utilizar o Editor de Cabeçalho, originalmente elaborado para o projeto Lácio-Web, conforme mencionado anteriormente, para anotar estruturalmente os textos do *corpus* do projeto NANOTERM. Conforme a atividade de anotação assistida pelo Editor de Cabeçalho desenvolvia-se, percebia-se que as etiquetas disponíveis no Editor eram insuficientes para anotar os textos selecionados para o *corpus*. Foi necessário, então, adaptar o Editor de Cabeçalho de forma a contemplar as especificidades do *corpus*. As adaptações dizem respeito à inserção de novas etiquetas.

Outra ferramenta que precisou ser adaptada foi o programa PROTEJ, mencionado acima. Originalmente, esse programa foi desenvolvido no âmbito do projeto *Dicionário Histórico do Português do Brasil dos séculos XVI, XVII e XVIII*[20], portanto, estava preparado para processar *corpus* histórico. Para lidar com *corpus* contemporâneo, como é o caso deste *corpus* do projeto NANOTERM, o PROTEJ precisou passar por adaptações.

Há que se mencionar ainda a necessidade de criação de um programa em PEARL para agrupar os arquivos. Isso porque o *corpus* acabou totalizando 1.057 arquivos distribuídos em 5 gêneros diferentes. Essa quantidade de textos estava inviabilizando a tarefa de extração dos candidatos a termos, sobretudo porque era preciso estabelecer listas de freqüência por gênero. Assim, os arquivos divididos por pastas (cada pasta constituía um gênero) foram agrupados

[20] O projeto é coordenado por Maria Tereza Camargo Biderman (FCL,UNESP, *Campus* de Araraquara) e tem como objetivo elaborar um dicionário do português do Brasil dos séculos XVI, XVII e XVIII a partir de *corpora*. Instituições parceiras: Universidade de Évora (Portugal); Universidade de São Paulo, Campus de São Paulo e Campus de São Carlos; Universidade Federal de São Carlos; Universidade Federal do Rio Grande do Sul; Universidade Federal de Minas Gerais; Universidade Federal do Mato Grosso do Sul; Universidade Federal da Bahia.

automaticamente por meio deste programa, facilitando o processamento, pois se passou a processar 5 arquivos e não mais 1.057.

3.3. Dificuldade de obtenção de textos e de permissões de uso

Durante as buscas na *Web*, encontraram-se diversas bibliotecas digitais de universidades e instituições de pesquisa, entretanto, poucas disponibilizavam digitalmente seu conteúdo na *Web.* Estas "bibliotecas virtuais" resumiam-se em motores de buscas para localização dos itens de seu acervo em suas estruturas físicas. Dentre as bibliotecas que disponibilizam digitalmente seu conteúdo (como a da UNICAMP), muitas (como a da UFSCar) protegem os arquivos (em geral no formato "pdf") com criptografia[21], o que impossibilita sua manipulação/ conversão para o formato de destino (extensão ".txt"), processável pelos programas de PLN. Assim, os arquivos livres foram imediatamente adicionados ao *corpus,* enquanto solicitações, por *e-mail*, foram enviadas aos autores dos demais textos, solicitando contribuições em Língua Portuguesa como teses, artigos, monografias, juntamente com o pedido de permissão de uso. Somente 3 responderam e contribuíram com a tese[22] já no formato desejado (.doc*).* De mais de 100 *e-mails* enviados, apenas 5 responderam e contribuíram com demais tipos de texto para a pesquisa. Assim, a solução encontrada foi a não disponibilização do *corpus* na *Web*, como era a intenção inicial, mas apenas para pesquisa de pequenos grupos por meio de solicitação ao GETerm.

[21] O termo **criptografia** surgiu da fusão das palavras gregas "Kryptós" e "gráphein", que significam "oculto" e "escrever", respectivamente. Na computação, as técnicas mais conhecidas envolvem o conceito de chaves, as chamadas "chaves criptográficas". Trata-se de um conjunto de bits baseado em um determinado algoritmo capaz de codificar e de decodificar informações. Se o receptor da mensagem usar uma chave incompatível com a chave do emissor, não conseguirá extrair a informação. (Informações da Info Wester, disponivel em http://www.infowester.com/criptografia.php)

[22] As demais teses foram disponibilizadas pela Biblioteca Comunitária da UFSCar, que autorizou o uso dos arquivos em formato *.doc*. Essas teses estão convertidas em ".txt", devidamente nomeadas, com cabeçalhos e integram o *corpus.*

4. Considerações finais

A N&N configura-se como uma área cujo desenvolvimento científico promete ser promissor na próxima década, e o léxico, obviamente, acompanha esse desenvolvimento, já que para cada novo referente criado (técnica, produto, processo, etc.) há uma nova denominação. Nesse sentido, sistematizar a terminologia em língua portuguesa dessa área revela-se preponderante.

Para ter acesso à terminologia de qualquer domínio de especialidade, é necessário antes ter acesso aos textos, já que cada termo está inserido em um contexto. Por essa razão, compilou-se um *corpus* da N&N, de forma a dar subsídios para a posterior elaboração do dicionário-piloto de N&N, haja vista que o *corpus* desenvolvido é essencial para diversas etapas constitutivas do processo de criação de um dicionário especializado, tais como: a extração dos candidatos a termos, a elaboração da ontologia, o preenchimento das fichas terminológicas e a redação das definições.

Para a compilação do *corpus*, percorreram-se as seguintes etapas: seleção de textos e pedidos de permissão de uso; manipulação e limpeza do *corpus*; etiquetagem (anotação estrutural, cabeçalhos) e nomeação dos arquivos; exclusão das etiquetas XML para processamento do *corpus* e contagem de palavras.

Durante a elaboração do *corpus*, alguns problemas surgiram no que diz respeito à falta de balanceamento em relação aos distintos gêneros textuais, à adequação/criação de ferramentas computacionais ao *corpus* compilado e à dificuldade de obtenção de textos e de permissões de uso. No entanto, alguns procedimentos foram adotados para sanar ou, ao menos, minimizar os problemas, de forma que o *corpus* pudesse ser finalizado, cumprindo, assim, as metas inicias.

O *corpus*, dividido em 5 gêneros (Científico, Científico de Divulgação, Técnico-Administrativo, Informativo e Outros), foi concluído em 12 de julho de 2007, contendo 1.057 textos de 57 fontes diferentes, e totalizou 2.565.490 palavras.

Como a compilação de *corpus* tem sido freqüente nas pesquisas terminológicas e lexicológicas, espera-se, com esta descrição, ter colaborado com demais grupos que desenvolvem projetos lingüísticos baseados em *corpora*.

5. Referências bibliográficas

ALUÍSIO, S.M.; ALMEIDA, G. M. B. .O que é e como se constrói um *Corpus*? Lições aprendidas na compilação de vários corpora para pesquisa lingüística. *Calidoscópio* (UNISINOS), v. 4, p. 156-178, 2006. Disponível em: http://www.unisinos.br/publicacoes_cientificas/images/stories/pdfs_calidoscopio/vol4n3/art04_aluisio.pdf

BERBER SARDINHA, T. Lingüística de *Corpus*: Histórico e Problemática. *D.E.L.T.A*., vol. 16, n° 2, 2000 (323-367)

BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press, Cambridge, 1998

KNOBEL, M. O futuro da nanotecnologia no Brasil: vinte anos não são nada?. *Cienc. Cult.* [online]. Jan./Mar. 2005, vol.57, no.1 [Acessado em 18 de Abril de 2007], p.4-5. Disponível em: http://cienciaecultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252005000100002&lng=en&nrm=iso ISSN 0009-6725.

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

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

SINCLAIR, J. 2005. Corpus and Text - Basic Principles. In: *Developing Linguistic Corpora: a Guide to Good Practice*, ed. M. Wynne. Oxford: Oxbow Books: 1-16. Disponível em: http://ahds.ac.uk/linguistic-corpora/. Acesso em: 30/10/2006.

TOMA, H.E.; ARAKI K. O gigantesco e promissor mundo do muito pequeno, 2005. *Ciência Hoje*. Disponível em: http://cienciahoje.uol.com.br/materia/view/3440. Acessado em 18 de abril de 2007.