7. Robertson, T. *Dictionnaireidéologique : re cueil des mots, des phra ses, des idiotismes et des proverbes de la langue française classés selon l'ordre idées*. Paris : A. Derache. 1859.
8. Tollenaere, F. de. Lexicographie alphabétique ou idéologique. *Cahiers de lexicologie*. Nº2, p. 19-29. 1960.
9. Baldinger, K. *Te oría semántica : hacia una semántica moderna*. Madrid : Alcalá, 1970.
10. Pottier, B. *Sémantique générale*. Paris : PUF, 1992.
11. Rey-Debove, J. *Étude linguistique et sémiotique des dictionnaires français contemporains*. La Haye/Paris : Mouton. 1971.
12. Spitzer, C. S. J. *Dicion ário analógico da língua portu g u esa*. 2e edição Po rto Alegre (BR) : Livraria do Globo. 1952.
13. Boissiere, P. *Dictionnaire analogique de la langue française : répertoire complet des mots par les idées, des idées par les mots*. 3e éd. Paris : Aug. BoyerM (s.d.)
14. *Le petit Robert : dictionnaire de la langue française*. Paris : Dictionnaires Le Robert, 1996. (CD-Rom)
15. Vide nota 13.
16. Embora o *Dictionnaire de la machine-outil*, de Eugen Wüster, tenha sido publicado em 1935, a edição de maior referência é a que segue: Wuster, E. *Dictionnairemultilingue de la machine-outil. Notions fondamentales, définies et illustrées, présentées dans l'ordre systématique et l'ordre alphabétique. Volume de ba se anglais-français = The machine tool. An interlingual dictionary of basic concepts comprising an alphabetical dictionary and a classified vocabulary with definitions and illustrations*. English-French Master Volume, London : Technical Press. 1968.
17. Vernay, H. *Dictionnaire onomasiologique des langues romanes - DOLR*. Vol. 1 (Monde abstrait – monde concret; Le monde humain; La vie humaine ; L'anatomie humaine). Tübingen: Niemeyer. 1991.
18. Vernay, H. *Dictionnaire onomasiologique des langues romanes : DOLR*. Tübingen: Niemeyer. Volumes : 2 Domaine psycho-physique; Domaine moral et intellectuel. 1992; 3 Adhésion / refus ; Vérité – contre-vérité ; Le domaine cusal ; États et changements d'états ; Organisaiton de l'espace. 1993; 4 Relationsinterhumaines ; Activité-travail. 1994; 5 Monde professionel; Monde agricole ; Eaux et forets ; Viticulture. 1995; 6 Subsistance ; cuisine et repas ; Domaine vestimentaire ; Ameublement. 1996.


# A TERMINOLOGIA NA ERA DA INFORMÁTICA

Gladis Maria de Barcellos Almeida,
Leandro Henrique Mendonça de Oliveira e
Sandra Maria Aluísio

A terminologia e a informática não são independentes uma da outra. Desde 1960, em países desenvolvidos e com grande tradição em pesquisa terminológica, os dois campos de estudo estão ligados de forma a facilitar o armazenamento e a difusão de dados terminológicos na elaboração de grandes bases de dados especializados, denominados bancos de terminologia (1). A integração entre as duas áreas é tal que se cunhou o termo terminótica, dando origem a um novo conceito e inaugurando um novo paradigma metodológico nas pesquisas terminológicas.

A contribuição advinda da informática começa de fato a aparecer nas pesquisas terminológicas no Brasil nos últimos dez anos. O crescimento de estudos e pesquisas na área de lingüística computacional e lingüística de *corpus* e o conseqüente aprimoramento e desenvolvimento de ferramentas computacionais voltadas para o processamento de língua natural (PLN) do português passam a interferir diretamente na prática terminográfica.

Nesse sentido, a terminologia descritiva de viés lingüístico começa a ter um amplo desenvolvimento, já que a possibilidade de lidar com grandes *corpora* permite a observação e descrição de fenômenos lingüísticos recorrentes antes impossível de perceber, dado que os procedimentos de observação e descrição contavam apenas com recursos manuais.

Entendemos que fazer terminologia na era da informática significa criar um conjunto de procedimento automatizados ou semi-automatizados que dêem suporte às tarefas envolvidas no trabalho terminológico, quais sejam: 1. criação de *corpora* descartáveis; 2. extração automática de candidatos a termos desses *corpora*; 3. inserção dos termos numa ontologia (mapa conceitual); 4. elaboração e edição de fichas terminológicas; 5. elaboração e constante atualização da base definicional (2); 6. elaboração de definições; 7. edição de verbetes, 8. difusão dos dados para intercâmbio com outras aplicações ou usuários.

Nosso objetivo aqui é apresentar ambientes e/ou ferramentas que dão suporte às tarefas citadas acima, se não a todas, pelo menos a algumas delas e, sempre que possível, apontaremos os trabalhos referentes ao português.

Desenvolvido pela Faculdade de Letras da Universidade do Porto (FLUP), o Corpógrafo (3, 4, 5) é um gestor de *corpus* que se encontra, atualmente, na versão 3.0 e é otimizado para pesquisas terminológicas, isto é, para a extração de termos e sua organização em bases de dados. Fornece um ambiente web integrado para o manejo de *corpus*, disponibilizando ferramentas para processamento de *corpus*. Dentre as ferramentas que possui, estão concordanciadores, contadores de freqüência e, também, ferramentas de pré-processamento de *corpus*, como as de limpeza e sentenciadores. Toda funcionalidade do corpógrafo está associada a um dos quatro ambientes de trabalho ou módulos: gestor de arquivos e de *corpus*, pesquisa em *corpus*, centro de conhecimento, centro de documentação, o que diminui a sobrecarga ao se

trabalhar no ambiente. Nessa nova versão, o corpógrafo dá um bom suporte para a difusão dos produtos terminológicos, permitindo que se exporte bases terminológicas no formato XML para serem usadas em outras aplicações; que se gerem glossários em HTML usando uma base terminológica, além de se exportar o próprio *corpus* em XML para ser usado em outras aplicações ou por outros usuários.

**CORPORA DESCARTÁVEIS PARA PESQUISA TERMINOLÓGICA** Embora já exista um número razoável de *corpora* genéricos (ou de referência, como são chamados) para várias línguas, o número de *corpora* específicos disponíveis para suporte à pesquisa terminológica ainda é deficiente. Essa deficiência dá-se pela própria especificidade de tais *corpora* que são muitas vezes construídos para serem utilizados por um período curto de tempo e somente em um projeto, daí se questionar o investimento de grandes esforços na sua compilação e anotação que visam a sua reutilização (ou reuso, como se costuma referir em lingüística computacional).

Ainda que seja possível construir tais *corpora* pela busca manual na web, esse processo consome muito tempo, se levarmos em conta os benefícios para pesquisas tão pontuais. Para atender a essa necessidade específica de criação e de pesquisa terminológica nesses tipos de *corpora*, existe pelo menos um projeto de nosso conhecimento para criação de *corpus* a partir da web e para extração automática de candidatos a termos. O BootCaT (6, 7) – extrator automático de *corpus* e de termos (do inglês "Bootstrapping Corpora and Terms") – propõe a montagem de *corpus*, de modo iterativo, a partir de textos obtidos na web. É composto por várias ferramentas escritas em Perl (8), que foram projetadas para executar pequenas partes do processo de montagem de *corpus*. Basicamente, o processo de montagem de *corpus* do BootCaT é composto de quatro passos: 1. construir um *corpus* automaticamente a partir de buscas ao Google (9), utilizando um pequeno conjunto de sementes (*seed*) (10); 2. extrair novas sementes desse *corpus*; 3. utilizar essas novas sementes para novas buscas ao Google, cujos textos recuperados serão concatenados ao *corpus*, aumentando-o; 4. extrair novas sementes desse *corpus* complementado, e assim por diante.

O BootCaT também dispõe de ferramentas para extração de termos com mais de uma palavra, ou termos multipalavra.

**O CORPÓGRAFO DÁ UM BOM SUPORTE PARA A DIFUSÃO DOS PRODUTOS ...**

**EXTRAÇÃO AUTOMÁTICA DE CANDIDATOS A TERMOS** Como visto na seção acima, o BootCat também extrai candidatos a termos, tanto unigramas (lexias com apenas uma palavra) como multipalavras. Outro projeto de avaliação e implementação de métodos automáticos de extração foi desenvolvido no Núcleo Interinstitucional de Lingüística Computacional – NILC/USP (11), como resultado de um trabalho de mestrado. Denominado ExPortTer (12), esse projeto avaliou especificamente métodos das três abordagens (lingüística, estatística e híbrida) para o português. As medidas estatísticas utilizadas nesse trabalho foram quatro: freqüência, *log-likelihood*, informação mútua e coeficiente *Dice*, implementadas no pacote para a extração de n-gramas denominado *N-gram Statistics Package* – NSP (13), com objetivo de eleger a melhor medida estatística para a extração automática de unigramas, bigramas (lexias com duas palavras) e trigramas (lexias com três palavras). O método lingüístico implementado baseia-se em expressões lingüísticas e indicadores estruturais, bem como em padrões morfossintáticos dos termos de um dado domínio. Para a abordagem híbrida, foi gerado um conjunto de orações do *corpus*, aqui chamado de sub*corpus*, que apresentassem as expressões lingüísticas definidas no método lingüístico, de maneira que cada oração é impressa no sub*corpus* somente uma vez, independentemente do número de expressões que pode apresentar. Este procedimento representou a "parte lingüística" do método híbrido. O sub*corpus* de saída, constituído pelas orações que apresentaram alguma expressão lingüística, é tomado como entrada para o pacote NSP, representando, assim, a parte estatística desse método.

Esses métodos foram utilizados em outros três projetos desenvolvidos no NILC: o Bloc-Eco (14, 15), cujo objetivo foi a criação de uma base de conhecimento com informações ontológicas para termos em português da área de ecologia; o projeto de desenvolvimento de uma estrutura conceitual (ontologia) para a área de nanociência e nanotecnologia (16) e o projeto *e-Termos*, projeto de doutorado em andamento no NILC, que será explorado mais abaixo.

**EDITOR E VISUALIZADOR DE ONTOLOGIAS** Diversas linguagens, técnicas e ferramentas têm sido propostas para a organização do conhecimento e sua visualização na web. No que se refere a editores e visualizadores de ontologias, alguns bons exemplos são: Protégé 2001 (17), o OntoEdit (18), o Inxight StarTree (19), o TreeBolic Generator (20) e o HyperEditor (21). Uma característica comum entre eles é que são aplicações *stand-alone* que executam fora do ambiente web, impedindo sua utilização por vários usuários ao mesmo tempo. Entretanto, apesar dessa característica, os editores StarTree, TreeBolic Generator e HyperEditor foram construídos em Java e, portanto, sua migração para o ambiente web, na forma de *applets*, pode ser possível.

Os dois primeiros editores citados acima são ferramentas que implementam a visualização de ontologias na forma *folder-tree*, aqui chamado de visualização arbórea. Nesse tipo de visualização, quando um nó é selecionado na árvore à esquerda, seu conteúdo é apresentado à direita da seleção, e assim sucessivamente até o ultimo nível da árvore (nó folha). Nesse caso, quando a estrutura da ontologia possui vários níveis de abstração, a visualização fica prejudicada.

Os editores StarTree, TreeBolic Generator e HyperEditor apresentam a visualização da ontologia na forma de árvore hiperbólica (*Hyperbolic Tree*) (22). Segundo Freitas *et al.* (23), esta visualização representa hierarquias através de um *layout* radial disposto em um plano hiperbólico mapeado para um plano de duas dimensões (2D). Além disso, apresenta aspectos de construção – como o efeito *fisheye* (24) – aliados a mecanismo simples de navegação pela indicação de um nó de interesse, que é exibido no centro da representação em detalhe, cujo contexto é mantido pela exibição do restante da estrutura da ontologia com nós diminuindo de tamanho até serem suprimidos na borda do círculo radial. A abordagem do plano hiperbólico pode manipular uma estrutura em árvore usando o conceito de foco e contexto. Ou seja, o plano hiperbólico permite ao usuário navegar através dos nós e visualizar a relação da porção visível do plano com a estrutura inteira sobre uma única tela (25).

O efeito *fisheye* fornece um esquema que, em geral, é suficiente para lidar com a navegação e orientação de grandes redes de informação. Na movi-

mentação dessa interface, os nós da ontologia aumentam e diminuem de tamanho, saindo e entrando em foco, podendo ser expandidos quando o usuário arrastar os nós com o *mouse* ou, ainda, por meio de pesquisa direta pelo nó. Esse recurso permite grande flexibilidade e agilidade na tela.

Na visualização hiperbólica, a expansão e poda dos nós na estrutura são operações que mantêm sempre uma subárvore visível reduzindo, para o usuário, a sensação de perda de contexto. Assim, as árvores hiperbólicas são uma representação dinâmica da estrutura hierárquica de uma ontologia e representa uma maneira eficiente de exibir árvores complexas com exatidão.

No contexto do projeto de desenvolvimento de uma ontologia para a área de nanociência e nanotecnologia (16) citado acima, foi desenvolvido o OntoEditor, uma ferramenta para edição de ontologias via internet, que implementa tanto a visualização arbórea (*folder-tree*) quanto a visualização hiperbólica (*Hyperbolic Tree*). As principais contribuições desse trabalho são: 1. a possibilidade de visualizar a estrutura de uma ontologia no formato de árvore hiperbólica; 2. a capacidade de converter a estrutura de uma ontologia no formato texto em estruturas de visualização (arbórea e hiperbólica) a partir de uma operação de *upload*; e 3. a flexibilidade de estar acessível, via internet, para o público geral e especializado, possibilitando que qualquer usuário crie e visualize suas ontologias a qualquer tempo. Ressalte-se que o OntoEditor foi incorporado pelo projeto *e-Termos* já citado e que será apresentado a seguir.

Amparado pelos pressupostos da terminologia de orientação descritiva de viés lingüístico, 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 esse 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 o trabalho de grupos de usuários com interesses e propósitos comuns. Além disso, o aspecto colaborativo permite a harmonização 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 e utilizar uma senha de acesso. O mesmo procedimento pode ser utilizado para a interação com os especialistas do domínio, ou seja, especialistas previamente selecionados podem opinar, criticar, sugerir alterações, ratificar os dados (lista de termos, definições, etc) por meio de acesso à internet.

Outras vantagens do *e-Termos* são: 1. a possibilidade de análise qualitativa do *corpus*; 2. a categorização e visualização dos termos em uma ontologia; 3. a criação customizada das fichas terminológicas; 4. o gerenciamento da base definicional; 5. a redação assistida da definição terminológica; e, finalmente, 7. a edição de verbetes a partir dos campos previamente selecionados nas fichas terminológicas.

Como o *e-Termos* ainda está em fase de elaboração, nem todos os módulos estão implementados. O primeiro lançamento público deve acontecer em abril de 2006. A previsão é de que a versão final e oficial esteja pronta para utilização a partir de 2009.

Em vista do que foi apresentado, percebe-se que a associação entre terminologia e informática é viável e, sobretudo, necessária para as ações e pesquisa de soluções terminológicas assistidas por computador. Soma-se a isso a existência de ferramentas e/ou ambientes para o português, o que confere às pesquisas terminológicas em língua portuguesa um avanço evidente.

***Gladis Maria de Barcellos Almeida*** *é lingüista e professora adjunta do Departamento de Letras da Universidade Federal de São Carlos (UFSCar). Coordena o projeto "Extração automática de termos e elaboração colaborativa de terminologias para intercâmbio e difusão de conhecimento especializado – TermEx". É líder do Grupo de Estudos e Pesquisas em Terminologia (GETerm) e orienta projetos de pesquisa em lexicologia e terminologia na linha de pesquisa "Linguagem humana e tecnologia".*

***Leandro Henrique Mendonça de Oliveira*** *é mestre e doutorando do Instituto de Ciências Matemáticas e de Computação do (ICMC), USP-São Carlos, onde atua em pesquisas em terminologia na área de processamento de língua natural (PLN). É funcionário da Empresa Brasileira de Pesquisa Agropecuária (Embrapa).*

***Sandra Maria Aluísio*** *é professora assistente no Departamento de Ciências de Computação e Estatística do ICMC-USP-São Carlos e uma das coordenadoras do NILC/ICMC-USP, desde sua fundação em 1993. Foi coordenadora do curso de bacharelado em informática do mesmo departamento de 2000-2002. Seu interesse em pesquisa envolve processamento de língua natural (PLN) e inteligência artificial na educação.*

## NOTAS E REFERÊNCIAS BIBLIOGRÁFICAS

1. Dubuc, R. *Manual de terminologia*. 3a. ed. Chile: Unión Latina/RiL editors, 1999.
2. A base definicional consiste num repositório de contextos definitórios compilados de diversas e variadas fontes, a partir das quais o terminólogo redige o texto final da definição que deverá integrar o verbete.
3. *http://www.linguateca.pt/Corpografo/*
4. Sarmento, L.; Maia, B.; Santos, D. "The corpógrafo: a web-based environment for corpora research". *In Proceedings of LREC 2004*. Lisboa, Portugal, 25 May 2004.
5. Sarmento, L.; Maia, B. "CG - An integrated environment for *corpus* linguistics". Pôster apresentado na conferência *CL2003: Corpus linguistics 2003* - Lancaster University (UK., 28 - 31 March 2003.
6. *http://sslmit.unibo.it/~baroni/bootcat.html*
7. Baroni, M.; Bernardini, S. "BootCaT: Bootstrapping corpora and terms from the web". In: *Proceedings of LREC 2004*, 1313-1316.
8. *http://www.perl.com*
9. *http://www.google.com.br/*
10. Sementes são termos típicos em textos do domínio específico a partir do qual busca-se construir a amostragem.
11. O NILC está sediado no Instituto de Ciências Matemáticas e de Computação (ICMC) da USP, campus de São Carlos, SP (*http://www.nilc.icmc.usp.br/nilc/*).

12. Teline, M. F. "Avaliação de métodos para extração automática de terminologia de tex tos em português". ICMC-USP, São Car los, Disse rtação de mestrado, 2004.
13. *http://www.d.umn.edu/~tpederse/nsp.html*
14. *http://www.nilc.icmc.usp.br/nilc/projects/bloc-eco.htm*
15. Zavaglia, C.; Oliveira, L.H.M.; Nunes, M.G.V.; Teline, M.F.; Aluísio, S.M. "Avaliação de métodos de extração automáticade ter mos para a construção de ontologias". *Re latório Técnico do NILC*. NILC-TR-05-01. 13 p., São Car los-SP, 2005.
16. Genovês Jr.; L. C.; Aluisio, S.M. "Avaliação de ambientes de suporte à montagem automática de *corpus* a partir de textos da Web e extração automática de termos", *Relatório Técnico do NILC*. NILC-TR-05-15. *Relatório Técnico do ICMC* No 266, 52 p., São Carlos-SP, 2005.
17. Noy, N.F.; Sintek, M.; Decke r, S.; Crubez y, M.; Fregerson, R. W.; Musen, M. A. "C reating semantic Web contents with protégé-2000". *IEEE Intelligent Systems*, 16 (2.:60-71, 2002.
18. Staab, S.; Maedche, A. "Knowledge portals – ontologiesat work", *AI Magazine*, Summer 2001, pgs. de 63-75.
19. Inxight Software Incorporated. *Inxight Star Tree*. Disponível em: *http://www.-inxight.com/products/oem/star_tree/*
20. Disponível para dow nload em: *http://treebolic.sourceforge.net/en/home. htm*
21. Atualmente, o *HyperEditor* é desenvolvido e distribuído pela Embrapa Informática Agropecuária (*www.cnptia.embrapa.br*).
22. Lamping, J.; Rao, R.; Pirolli, P. "A focus+context technique based on hyperbolic geometry for visualizing large hierarchies". *In: Proceedings of ACM SIGCHI Conf. on Human Factor in Computing System*, 1995, 401-408.
23. Freitas, C. M. dal S.; Chubachi, O. M.; Luzzardi, P. R. G.; Cava, R. A. "Introdução à visualização de informações". *RITA*, v. 8, n. 2, p. 1-16, 2001. Disponível em: *http://www.inf.ufrgs.br/cg/publications/carla/Freitas- RITA2001.pdf*
24. Furnas, G. "Generalizedfisheye views". *In: Proce edings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 1986, pp 16-23.
25. Hao, M. C.; Hsu, M.; Dayal, U.; Krug, A. *Visual mining large web-based hyperbolic space using hidden links*. Palo Alto: HP Laboratories-Software Technology Laboratory, 1999. 9 p. HPL 1999-20. Disponível em: *http://www.hpl.hp.com/techreports/1999/HPL-1999-20.pdf* .

# TERMINOLOGIA TÉCNICO-CIENTÍFICA: POLÍTICAS LINGÜÍSTICAS E MERCOSUL

Maria da Graça Krieger

*"A constituição de uma terminologia própria marca, em toda ciência, o advento ou o desenvolvimento de uma conceitualização nova, assinalando, assim, um momento decisivo de sua história. Poder-se-ia mesmo dizer que a história particular de uma ciência se resume na de seus termos específicos." (1)*

**TERMOS TÉCNICO-CIENTÍFICOS** Ao salientar o papel da terminologia na constituição do pensamento científico, o célebre lingüista Benveniste também evidencia que seu uso está relacionado a atividades que envolvem um saber especializado. Isso se justifica porque os termos técnico-científicos transmitem conceitos próprios do conhecimento científico, técnico, tecnológico, jurídico, entre outros domínios. Daí a razão pela qual se pode fazer referência à terminologia da química, da biologia, da lingüística, além de muitas outras. Em paralelo aos campos científicos, as terminologias são também designativas de componentes e de produtos resultantes da técnica e da tecnologia, conforme hoje se compreende.

Em razão de sua característica maior, ou seja, delimitar conceitos próprios de uma área, diferenciando-se, nessa medida, da palavra comum, entende-se que:

> "Para os especialistas, a terminologia é o reflexo formal da organização conceptual de uma especialidade e um meio inevitável de expressão e de comunicação profissional". (2)

Mas, além de conjunto de termos de um campo de saber, a referência à terminologia designa um ramo da lingüística, cujo objeto primeiro de investigação é o próprio termo, compreendido como unidade lexical de valor especializado, porquanto integra o conjunto denominativo e conceptual das ciências e das técnicas.

Em linhas gerais, a pesquisa teórica está relacionada ao estudo dos componentes cognitivos e pragmáticos que conferem estatuto terminológico às unidades lexicais de um sistema lingüístico, junto à descrição de suas estruturas morfossintáticas e das formas de comportamento nos cenários comunicativos de que participam. Em sua face aplicada, o trabalho terminológico, entre outros aspectos, busca definir princípios e métodos orientadores da elaboração de glossários, dicionários técnico-científicos, bancos de dados terminológicos, ontologias, além de outros produtos que sistematizam e divulgam termos específicos de uma área.

A terminologia é considerada uma ciência ainda nova; mas, em contrapartida, o uso de termos técnico-científicos vem de tempos remotos :

> A terminologia não é um fenômeno recente. Com efeito, tão longe quanto se remonte na história do homem, desde que se manifesta a linguagem, nos