Descrição semântica de Definições Terminológicas

Dayse Simon Landim Kamikawachi[1]

[1]Departamento de Letras, Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13.565-905 – São Carlos – SP – Brasil

daysesimon@gmail.com

Abstract. *This paper describes a research of semantic description in Terminological Definitions (DT). We have built a corpus and elaborated a linguistic analysis methodology to map the main semantic relations between the parts of DT text. Thus, quantitative and qualitative results were generated and they will subsidize a computer-aided learning tool for DT writing, that is a resource in Natural Language Processing.*

Resumo. *Este artigo apresenta uma investigação realizada quanto aos aspectos semânticos presentes em Definições Terminológicas (DT). Para tal foram constituídos um corpus e uma metodologia de análise linguística que permitiram mapear as principais relações semânticas entre as partes do texto da definição. Ao final, foram gerados resultados quantitativos e qualitativos que subsidiarão uma aplicação em Processamento de Linguagem Natural de auxílio à redação da DT.*

1. Introdução

Em geral, a criação de dicionários ou vocabulários terminográficos envolve algumas etapas de trabalho, tais como construção do *corpus*, geração de listas de termos, elaboração da ontologia (mapa conceitual do domínio), preenchimento da ficha terminológica e edição do verbete. De fato, todas essas etapas são custosas, pois demandam tempo e abrangem conhecimento linguístico e terminológico. Contudo, graças à Informática e especificamente às áreas de Linguística de *Corpus* e de Processamento de Linguagem Natural (PLN), parte considerável das tarefas da atividade terminográfica tem sido automatizada ou semiautomatizada.

Uma das últimas etapas do trabalho terminográfico é a redação da Definição Terminológica (DT), a qual pode ser explicada como "*una descripción linguística de un concepto. Basada en el listado de un número de características que transmiten el significado del concepto*" [Sager 1993, p. 68]. Este é o momento no qual o terminólogo prevê as necessidades dos consulentes em potencial e, de acordo com as características atribuídas ao termo na área-objeto, adéqua o texto definitório. Portanto, a elaboração da DT se constitui uma tarefa altamente elaborada, ao passo que exige do redator conhecimentos do domínio que está sendo descrito, conhecimentos linguísticos referentes à língua na qual o texto definitório é redigido e conhecimentos de Terminologia, cujos pilares teórico-metodológicos sustentam o trabalho.

Entretanto, ao contrário das demais etapas, a DT ainda carece de descrição a fim de que a sua elaboração se torne em partes automatizada também, propiciando a sua redação mais otimizada aos grupos que desenvolvem produtos terminográficos.

Anais da II Jornada de Descrição do Português, páginas 55–62, Cuiabá, MT, Brasil, Outubro 24–26, 2011.

2. Objetivos

A partir da prática constante da redação de DTs no âmbito do Grupo de Estudos e Pesquisas em Terminologia (GETerm) foi constatado que as DTs: i) referentes a termos formados por substantivo que integram mesmo campo semântico na ontologia de domínio; ii) e que são redigidas por meio do gênero próximo e diferença específica (GPDE)[1] possuem certas regularidades quanto às relações semânticas[2] apresentadas em seus textos [Almeida et al. 2007].

Desta forma, em consonância com a Teoria Comunicativa da Terminologia [Cabré 1999, 1993], que considera os termos como signos linguísticos em funcionamento numa situação de comunicação especializada, nossa proposta de pesquisa foi observar e descrever formalmente essas regularidades desse tipo de texto definitório (GPDE) de termos (substantivos) provenientes de mesmo campo semântico, a fim de servirem como conhecimento linguístico para ser implementado computacionalmente num ambiente de auxílio à redação da DT.

3. Metodologia

O *corpus* utilizado na pesquisa é formado por 500 DTs e seus respectivos termos-entrada, referentes a sete campos semânticos da ontologia do domínio de Revestimento Cerâmico ("matéria-prima", "defeito", "produto acabado", entre outros), e a oito campos da ontologia do domínio da Fisioterapia ("disfunção", "teste e medida", "exame complementar" e etc.). O *corpus* possui no total 28117 *tokens*. Ressalte-se que as DTs utilizadas como *corpus* são textos que já passaram pela revisão linguística/terminológica e pela apreciação minuciosa dos especialistas de ambas as áreas.

Após o armazenamento de cada uma das 500 DTs num arquivo "txt", o passo seguinte foi a realização da anotação do *corpus*, de forma a otimizar a recuperação das informações contidas em cada arquivo quando fossem realizadas as análises. Nessa etapa, a anotação estrutural (nome do arquivo, domínio, subdomínio, campo semântico, etc.) e a anotação linguística do *corpus* (termo, definição terminológica, gênero próximo, diferença específica, relações semânticas, etc.) foram efetuadas.

O tipo de anotação utilizada seguiu o padrão "xml", que é formado por uma linguagem que permite fazer dos documentos uma estrutura hierárquica, além de permitir a criação de qualquer tipo de etiqueta de anotação.

---

[1] A definição que segue esse modelo GPDE, apresenta-se da seguinte forma: a DT é encabeçada pelo gênero próximo (GP), ou seja, um hiperônimo (ou termo-pai), o qual geralmente é recuperado observando-se o termo superordenado na ontologia. Imediatamente, segue a diferença específica (DE), que estabelece a diferença entre os termos-entradas que possuem o mesmo GP.

[2] Conforme Bodson (2004), os termos "relação semântica" e "relação conceitual" são duplamente empregados nos estudos da área. A escolha de um termo em detrimento do outro reflete, entre outras coisas, o ponto de vista adotado. O autor afirma que "*En effet, la relation sémantique exprime un lien entre deux sens, alors que la relation conceptuelle met plutôt l'accent sur la structure de la connaissance.*"[Bodson 2004, p. 36].

Inicialmente, foi gerado um cabeçalho para cada arquivo "txt". Isso foi feito por meio do Editor de Cabeçalho[3], um programa computacional que "auxilia o linguista a especificar diversas informações sobre os textos" [Aluísio e Almeida, 2006, p. 174].

Em relação à anotação linguística, foi considerada uma tipologia de relações semânticas elaborada nesta pesquisa, com base em Feliu (2004), Seppälä (2004), Pustejovsky (1995) e Jordan (1992). A seguir, são apresentadas as relações semânticas utilizadas na anotação das DTs[4]:

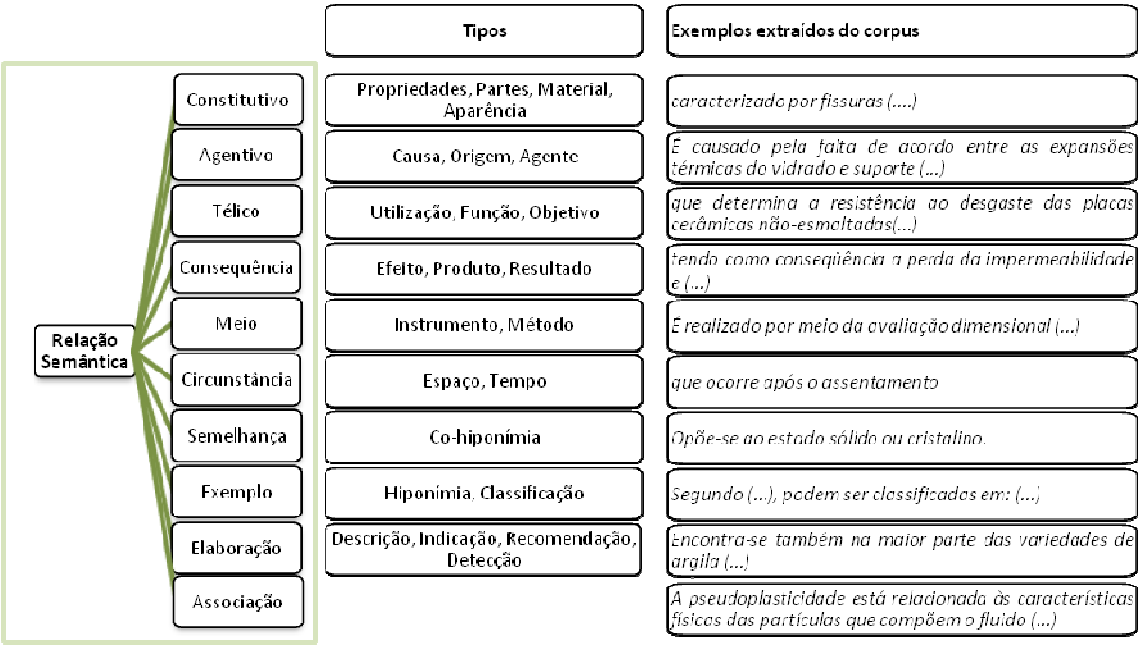

Figura 1. Relações Semânticas utilizadas na anotação das DTs

A título de exemplo, segue um arquivo anotado estrutural e linguisticamente:

[3] Disponível no site: http://www.nilc.icmc.usp.br/lacioweb/

[4] Em Kamikawachi (2009) são apresentados mais detalhes acerca do percurso metodológico da pesquisa.

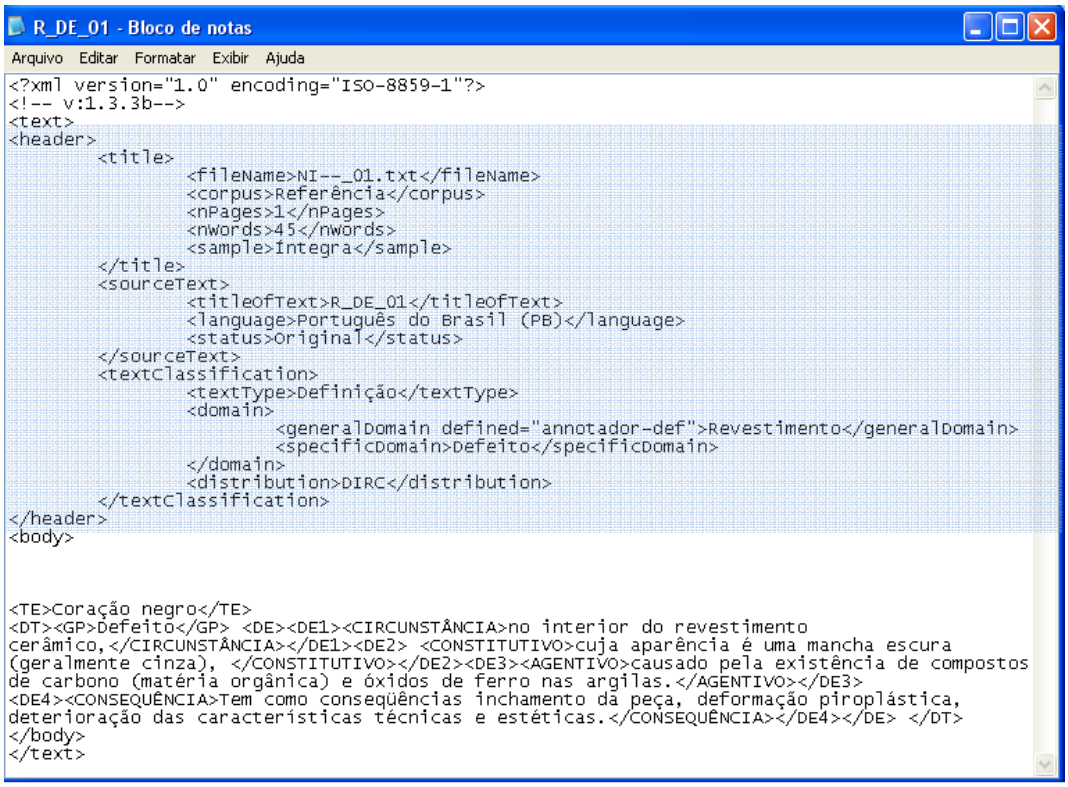

Figura 2. Cabeçalho com etiquetas "xml" gerado pelo Editor de Cabeçalho

4. Análise das DTs

Na análise do *corpus*, foi utilizado o programa computacional *WordSmith Tools,* versão 3, da autoria de Mike Scott, da Universidade de Liverpool. A escolha por esse programa se deu pelo fato de que ele apresenta: i) um bom desempenho estatístico; ii) manipulação de vários arquivos simultaneamente e iii) leitura de etiqueta *xml*. Foram utilizadas, especificamente, as ferramentas *Wordlist* (sobretudo a janela dos dados estatísticos referentes aos arquivos selecionados) e *Concord* (para realização de buscas específicas das relações semânticas).

A descrição das DTs foi efetuada em vários níveis, os quais são descritos a seguir.

4.1. Informações gerais

Considerando os dois domínios eleitos, foram analisadas 500 DTs agrupadas em 15 campos semânticos diferentes. Todas as DTs do *corpus* apresentam no mínimo **1 DE** e no máximo **5 DEs** (totalizando a frequência de 1241 DE_ns), o que comprova a afirmação de Bessé [1996, *apud* Seppälä, 2004, p. 149] quanto ao aspecto estrutural de que "*une définition (terminographique) ne comporte pas plus de cinq spécifiques*".

Quanto à quantidade de DE_ns que compõe as DTs, os números mais expressivos correspondem a DTs formadas por **2 DEs** e **3 DEs** (46,8% e 33,0% das DTs, respectivamente).

Do total de 1241 DE_ns analisadas, as 10 relações semânticas utilizadas para anotar as DE_ns apresentam a seguinte frequência:

Tabela 1. Relações semânticas no *corpus* de acordo com a frequência

RELAÇÃO	FREQ. REL.
CONSTITUTIVO	**31,3%**
TÉLICO	**23,2%**
ELABORAÇÃO	**20,0%**
AGENTIVO	**9,8%**
EXEMPLO	3,8%
MEIO	3,7%
CIRCUNSTÂNCIA	3,3%
CONSEQUENCIA	3,0%
ASSOCIAÇÃO	1,6%
SEMELHANÇA	0,4%
TOTAL	**100,0%**

4.2. Informações semânticas

Quanto à ordem na qual as relações semânticas se apresentam nas DTs, observamos que, considerando o critério de 2 DE_{ns} em sequência (desprezando a posição se a relação integrava a DE1, DE2, DE_n), temos que as três sequências mais recorrentes no *corpus* são de **TÉLICO + ELABORAÇÃO** (18,0%), **CONSTITUTIVO + TÉLICO** (16,1%) e **CONSTITUTIVO + AGENTIVO** (14,4%).

Também foram analisadas as relações semânticas segundo o lugar (DE_n) que as mesmas ocupam no texto da DT. Chegamos aos seguintes resultados:

Tabela 2. Relações semânticas nas posições DE1 à DE5

DE1	FREQ. REL.	DE2	FREQ. REL.	DE3	FREQ. REL.	DE4	FREQ. REL.	DE5	FREQ. REL.
CONSTITUTIVO	**58%**	**ELABORAÇÃO**	**23%**	**ELABORAÇÃO**	**47%**	**ELABORAÇÃO**	**50%**	**ELABORAÇÃO**	**56%**
TÉLICO	**30%**	**TÉLICO**	**20%**	TÉLICO	18%	TÉLICO	14%	CONSTITUTIVO	22%
CIRCUNSTÂNCIA	7%	**AGENTIVO**	**17%**	AGENTIVO	11%	CONSEQUÊNCIA	11%	AGENTIVO	11%
AGENTIVO	2%	**CONSTITUTIVO**	**17%**	CONSTITUTIVO	6%	AGENTIVO	7%	TÉLICO	11%
ELABORAÇÃO	1%	EXEMPLO	7%	EXEMPLO	5%	MEIO	7%		
MEIO	1%	MEIO	7%	CONSEQUÊNCIA	5%	CONSTITUTIVO	5%		
EXEMPLO	0%	CONSEQUÊNCIA	4%	MEIO	4%	EXEMPLO	4%		
		ASSOCIAÇÃO	4%	ASSOCIAÇÃO	2%	CIRCUNSTÂNCIA	2%		
		SEMELHANÇA	1%	CIRCUNSTÂNCIA	1%				
		CIRCUNSTÂNCIA	0%	SEMELHANÇA	1%				

No que se refere ao campo semântico no qual o termo a ser definido se insere, agrupamos os campos dos dois domínios que apresentam características semelhantes, a fim de avaliar: i) em que medida o campo semântico interfere na escolha das relações semânticas das DTs; ii) o quanto as relações são produtivas, mesmo em se tratando de domínios de conhecimento distintos.

Segue um exemplo da análise dos campos semânticos "Instrumento ou Equipamento (IN)" do domínio de Revestimento Cerâmico e "Instrumento de medida (IM)" do domínio da Fisioterapia:

Tabela 3. Relações semânticas dos campos IN e IM

REVESTIMENTO CERÂMICO	RELAÇÃO	FREQ. REL.
IN	TÉLICO	43%
	CONSTITUTIVO	25%
	ELABORAÇÃO	25%
	EXEMPLO	5%
	CIRCUNSTÂNCIA	1%
	TOTAL	100%
FISIOTERAPIA	**RELAÇÃO**	**FREQ. REL.**
IM	TÉLICO	55%
	CONSTITUTIVO	30%
	ELABORAÇÃO	10%
	SEMELHANÇA	5%
	TOTAL	100%

4.3. Informações quanto às expressões linguísticas

Por meio da análise do *corpus*, foi possível mapear as expressões linguísticas que se referem às relações semânticas adotadas na pesquisa. Segue amostra das expressões encontradas:

Tabela 4. Expressões linguísticas

RELAÇÃO SEMÂNTICA	EXPRESSÃO LINGUÍSTICA	RELAÇÃO SEMÂNTICA	EXPRESSÃO LINGUÍSTICA
AGENTIVO	causado por	CONSTITUTIVO	caracterizado pelo
	como proveniente		constituído de
	decorrente de		é composto por
	desenvolvida por		feito de
ASSOCIAÇÃO	depende de	EXEMPLO	os principais tipos são
	está relacionada ao		pode ser classificado em
	está associado ao		um exemplo comum é
	está relacionado com		um exemplo desse tipo
CIRCUNSTÂNCIA	empregada na	SEMELHANÇA	conhecido como
	quando		contrário a
	que atua no		espécie de
	que ocorre após		opõe-se ao
CONSEQUÊNCIA	causando	TÉLICO	a fim de
	compromete		é empregado como
	leva ao		promove
	provoca		tem por objetivo

5. Resultados

Como resultados da pesquisa, foram gerados:

i) uma proposta geral de sistematização quanto à redação da DT do tipo GPDE, a qual contempla uma base de exemplos de DTs, base de relações semânticas e uma base de suas respectivas expressões linguísticas;

ii) um *corpus* anotado em arquivo *txt*, composto de 500 DTs, e;

iii) dados estatísticos das DTs em diferentes níveis (combinação de relações mais produtivas e sua ordem no texto; quantidade de itens léxico por DE e por relação semântica; etc.) e também dados quanto à produtividade de determinadas relações por campo semântico analisado de ambos os domínios.

A partir dos dados obtidos nas análises será possível sugerir ao usuário, por exemplo, quais as combinações e a ordem de relações mais produtivas na DT, quais expressões linguísticas no português se referem à determinada relação semântica, ou seja, a orientação dar-se-á do planejamento da estrutura para a redação da DT (*top down*).

6. Considerações Finais

A presente pesquisa pôde contribuir com a investigação de tipo de texto definitório considerado como ideal tanto pela comunidade lexicográfica, como pela terminológica. Por meio do uso de *corpus* e de ferramentas de análise linguística foi possível o desenvolvimento de uma sequência de passos metodológicos válidos e ainda a descrição de uma grande quantidade de dados que, se feita manualmente, não seria satisfatoriamente fiável e nem tampouco realizada em curto tempo.

E por fim, o conhecimento linguístico gerado como consequência da aproximação entre a Terminologia e a área de Processamento de Linguagem Natural será útil para ambas as comunidades.

Referências

Almeida, G.M.B.; Souza, D.S.L.; Pino, D.H.P. (2007), A definição nos dicionários especializados: proposta metodológica. *Debate Terminológico*, v. 3, p. 1-20.

Aluísio, S.M.; Almeida, G.M.B. (2006) O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários corpora para pesquisa lingüística. *Calidoscópio,* Porto Alegre, v. 4, n. 3, p. 155-177.

Bodson, C. (2005), *Termes et relations sémantiques en corpus spécialisés : rapport entre patrons de relations sémantiques (PRS) et types sémantiques (TS).* 2005. 298f. Tese (Doutorado em Linguística) - Faculté des études supérieures, Université de Montréal, Montréal.

Cabré, M. T. (1993), *La terminología: teoría, metodología, aplicaciones.* Tradução de Carles Tebé. Barcelona: Editorial Antártida/Empúries.

Cabré, M. T. (1999), *La terminología: representación y comunicación. elementos para uma teoría de base comunicativa y otros artículos.* Barcelona: IULA/Universitat Pompeu Fabra.

Feliu, J. (2000), *Relacions conceptuals i variació funcional: elements per a un sistema de detecció automática.* (Trabalho de pesquisa) - UPF/IULA, Barcelona.

Jordan, M.P. (1992), An Integrated Three-Pronged Analysis of a Fund-Raising Letter. In: W.C. Mann and S.A. Thompson (eds.), *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text,* p. 171-226.

Kamikawachi, D. S. L. (2009), *Aspectos semânticos da Definição Terminológica (DT): descrição linguística e proposta de Sistematização.* (Dissertação de Mestrado), Departamento de Letras, Universidade Federal de São Carlos, São Carlos.

Pustejovsky, J. (1998), *The generative lexicon.* Londres: Cambridge/MIT Press.

Sager, J. C. (1993), *Curso práctico sobre el procesamiento de la terminología.* Tradução de Laura C. Moya. Madrid: Fundación Germán Sánchez Ruipérez/Pirámide.

Seppälä, S. (2004), *Composition et formalisation conceptuelles de la définition terminographique***.** Tese (Doutorado em Tratamento Informático Multilingue), École de traduction et d'interprétation, Université de Genève, Genebra.