Sistematização linguístico-computacional do léxico do domínio conceitual "Indústria do Bordado de Ibitinga"

Erasmo Roberto Marcellino[1], Bento Carlos Dias da Silva[1,2]

[1]Centro de Estudos Linguísticos e Computacionais – Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)
Caixa Postal 174 – 14.800-901 – Araraquara – SP – Brasil

[2]Departamento de Letras Modernas – Universidade Estadual Paulista "Júlio de Mesquita Filho"– Araraquara – SP – Brasil

erasmorm@hotmail.com, bento@fclar.unesp.br

Abstract. *This work discusses a proposition for organizing the lexical items from the conceptual domain labeled THE EMBROIDERY INDUSTRY OF IBITINGA in terms of a natural ontology. It also aims to establish the alignment between this ontology and the bases WordNet.Pr and WordNet.Br.*

Resumo. *Este trabalho discute uma proposta para a organização dos itens lexicais do domínio conceitual rotulado INDÚSTRIA DO BORDADO DE IBITINGA em termos de uma ontologia natural. Também visa estabelecer o alinhamento entre essa ontologia e as bases WordNet.Pr e WordNet.Br.*

1. Introdução

A produção do bordado exige conhecimentos técnicos e artísticos que, em termos linguísticos, traduz-se em um universo vocabular rico que possibilita a comunicação eficiente entre os profissionais, proporcionando, não só entre esses profissionais, como também entre eles e o público geral, a discursivização dos agentes, das técnicas, dos instrumentos, dos materiais, dos suportes, dos processos e dos produtos que constituem o universo discursivo dessa importante indústria. Este trabalho, que se insere nos domínios da Linguística, Linguística Computacional e Estudos do Léxico, foca uma parte da investigação desse universo vocabular, aqui denominado INDÚSTRIA DO BORDADO DE IBITINGA, doravante IBI, necessário para a comunicação dos profissionais envolvidos nesse setor industrial.

2. Objetivos

O estudo de natureza teórico-aplicada aqui apresentado visa (i) coletar itens e expressões lexicais do domínio conceitual IBI e (ii) selecionar, descrever e organizar esses itens em termos de uma base relacional de dados de conhecimento lexical. Essa base estrutura o estoque vocabular desse domínio em termos de uma rede semântica particular, como se demonstrará a seguir, capaz de dar forma a esse estoque não só em termos taxonômicos como também em termos de dois tipos de conexão: conexões conceituais, que se estabelecem entre uma unidade lexical e o conceito que ela

lexicaliza, isto é, entre itens lexicais e os conceitos lexicalizados que organizam-se na ontologia[1] do domínio conceitual alvo do estudo, e conexões linguísticas, que se estabelecem entre os itens lexicais do português e do inglês que lexicalizam os mesmos conceitos. Desse modo, dentro do âmbito das ciências do léxico, em que diversas áreas do conhecimento se inter-relacionam [Krieger e Finatto 2004], adverte-se que este trabalho não se insere nos domínios da lexicografia ou da terminologia, embora possa contribuir para trabalho dessa natureza, uma vez que o seu objetivo não é a produção de, por exemplo, um dicionário geral ou técnico-científico. Seu domínio é o do PLN (processamento automático de língua natural), ao propor um tipo específico de base de conhecimento lexical (uma BDL) passível de ser acoplado a sistemas de PLN [Dias-da-Silva 2006].

3. Metodologia

A estratégia para investigar o "revestimento lexical" do domínio conceitual da IBI tem motivação nas *wordnets*[2], cujo principal exemplar é a de Princeton, a WordNet.Pr (WN.Pr) [Fellbaum 1998], uma rede relacional de conceitos expressos no inglês norte-americano em termos de *synsets*[3]. Essa metodologia de estruturação de léxicos e de sua ancoragem conceitual é extraída, além da WN.Pr, também da EuroWordNet [Vossen 1998] – uma *multiwordnet* em desenvolvimento para línguas da União Europeia – e da WordNet.Br (WN.Br) [Dias-da-Silva 2008], base análoga à da WN.Pr em construção para o português brasileiro.

Esta investigação compreende, pois, o estudo em dois domínios complementares: (a) no domínio linguístico, em que serão investigados e sistematizados informações sobre o domínio da IBI e dados lexicais do português e do inglês coletados do *corpus* de referência do projeto[4], estudando-se certos padrões de lexicalização dos substantivos e a estruturação do léxico [Handke 1995, Hirst 2004], e (b) no domínio linguístico-computacional, em que será representada, de modo formal, toda a descrição construída no domínio anterior de uma maneira que o computador possa processar [Dias-da-Silva 2006].

3.1. Amostragem do procedimento analítico

Os dados lexicais coletados em *corpus* englobam itens e expressões lexicais que serão sistematizados em termos de uma ontologia de conceitos que fornecerá a estrutura

---

[1] Dentre as diferentes perspectivas sobre ontologias [Vossen 2003], a noção de relação ontológica aqui adotada baseia-se em Ding e Foo (2002), que a define como uma representação formal do conhecimento conceitual compartilhado em alguns domínios de interesse que se estrutura por meio de relações e funções.

[2] Nessas redes, os itens lexicais organizam-se principalmente em termos das relações de sinonímia, antonímia, hiponímia e meronímia.

[3] Cada *synset* é um conjunto de itens lexicais que compartilham o mesmo conceito.

[4] O *corpus* de referência do projeto (em fase de montagem) constitui-se, principalmente, de: materiais impressos referentes à indústria do bordado (manuais, catálogos, teses, dissertações, entre outros); inquéritos de informantes; informações, nas duas línguas, localizáveis em textos pelo motor de busca *Google*™; definições e abonações de dicionários; exemplos das gramáticas e informações lexicais contidas das bases da WN.Br e da WN.Pr.

semântico-conceitual para as diferentes categorias e subcategorias conceituais do domínio; parte dessas categorias ontológicas já foi estabelecida em um estudo inicial que contou com o auxílio de informantes: PROFISSIONAIS DA ÁREA, TIPOS DE MATERIAIS EMPREGADOS, MAQUINÁRIOS E INSTRUMENTOS ENVOLVIDOS, TIPOS DE SUPORTES PARA OS BORDADOS, PROCEDIMENTOS E PROCESSOS ENVOLVIDOS NA PRODUÇÃO DAS CONFECÇÕES e TIPOS DE PONTOS. O quadro 1 exemplifica parte da sistematização e possíveis alinhamentos entre *synsets* criados para o português, a partir de alguns itens lexicais coletados no *corpus*, e os já existentes para o inglês.

Quadro 1. Categorização e estruturação exploratórias

CATEGORIAS ONTOLÓGICAS	ITENS LEXICAIS	SYNSETS	
		PORTUGUÊS	INGLÊS
PROFISSIONAIS DA ÁREA	bordadeira; overloquista; arrematador	{bordadeira}	{embroideress}
TIPOS DE MATERIAIS EMPREGADOS	linha; codornê; lantejoula	{linha, linha de bordar}	{thread, yarn}
MAQUINÁRIOS E INSTRUMENTOS ENVOLVIDOS	bastidor; máquina de bordar; agulha	{bastidor}	{tambour1, embroidery frame, embroidery hoop}
TIPOS DE SUPORTES PARA OS BORDADOS	guardanapo; toalha; colcha	{semaninha}	{dish towel, tea towel}*

Os *synsets* da terceira e quarta colunas alinham-se pela relação de EQ-SYNONYM [Peters et.al. 1998], isto é, por "equivalência sinonímica" entre os *synsets*, com exceção do assinalado com um asterisco na última linha, que é um caso especial: o alinhamento, parcial, se dá via a relação de EQ-HYPERNYM ("equivalência hiperonímica"), pois o conceito que, no português é lexicalizado por *semaninha* ("jogo de sete panos de prato bordados ou pintados, cada um deles representando um dia da semana")[5], não se realiza no inglês. Esse exemplo ilustra que diferenças de naturezas diversas podem ocorrer entre as línguas como, por exemplo, quando há lacunas no léxico do português ou do inglês ou quando os significados dos itens ou expressões lexicais nas duas línguas não se correspondem integralmente. Essas características do léxico das línguas tornam complexa a investigação de possíveis relações semânticas entre léxicos distintos, o que exige uma especificação de tipos específicos de alinhamento que se verificam entre os *synsets* das duas bases, como os exemplificados neste parágrafo e outros apontados na literatura, como EQ-NEAR-SYNONYM ("equivalência por sinonímia aproximada") e EQ-HYPONYM ("equivalência por hiponímia").

[5] Não há registro dessa forma nos dicionários. A definição sugerida baseou-se em consulta ao *corpus* de referência do projeto e em entrevistas informais com profissionais. Esta frase-exemplo atesta o contexto de uso: "Entrei na mania de bordar semaninhas: um pano de prato para cada dia da semana". Disponível em: <http://www.vesoloski.eti.br/blogdagabi/2008/04/semaninha-em-ponto-cruz.php>. Acesso em: 23 jul. 2009.

4. Conclusão

O estabelecimento de uma estruturação entre os itens lexicais tanto dentro do domínio escolhido como entre os itens lexicais das duas línguas sob análise, conforme ilustra a sistematização exploratória resumida no quadro 1, já é suficiente para demonstrar a viabilidade de sistematização dos conceitos desse domínio em termos do que Hirst (2004) denomina "ontologia natural", isto é, uma ontologia que se constrói com a identificação de categorias conceituais que são implicitamente codificadas em itens lexicais de uma língua natural e, em particular, nos *synsets* do português e do inglês.

Estudos aprofundados no âmbito da Semântica Lexical, Pura e Computacional, e da Linguística Cognitiva, como já foi citado, são os alicerces teórico-metodológicos para que essa sistematização se concretize de modo robusto, coerente e linguisticamente sólido, podendo, além de se constituir em uma BDL, gerar produtos eletrônicos de referência, como dicionários e glossários bilíngues para auxiliar a comunicação de estudantes e profissionais desse importante setor industrial.

5. Referências Bibliográficas

Bento Carlos Dias-da-Silva (2008). The automatic mapping of Princeton WordNet lexical-conceptual relations onto the Brazilian Portuguese WordNet database. *Proceedings of The Sixth International Conference on Language Resources and Evaluation* (LREC 2008), páginas 335–342. ELRA/ELDA, Paris.

_____. (2006). O estudo lingüístico-computacional da linguagem. *Letras de hoje: periódico do Curso de pós-graduação em Letras*, 41:103–138, PUCRS, Porto Alegre.

Ying Ding e Schubert Foo (2002). *Ontology research and development part 1 – a review of ontology generation*. Journal of Information Science, 28(2):123–136.

Christiane Fellbaum, editora (1998). *WordNet*: an electronic lexical database. The MIT Press, Cambridge.

Jürgen Handke (1995). *The structure of the lexicon*: human versus machine. Mouton de Gruyter, Berlim.

Graeme Hirst (2004). Ontology and the lexicon. Em Steffen STAAB e Rudi STUDER, editores, *Handbook on ontologies. International handbooks on information systems*, páginas 209–229. Springer, Berlin.

Maria da Graça Krieger e Maria José Bocorny Finatto (2004). *Introdução à terminologia*: teoria e prática. Contexto, São Paulo.

Wim Peters, Piek Vossen, Pedro Díez-Orzas e Geert Adriaens (1998). Cross-linguistic alignment of wordnets with an Inter-Lingual-Index. *Computers and the Humanities*, 32(2,3):221–251.

Piek Vossen (2003). Ontologies. Em Ruslan Mitkov, editor, *The handbook of computational linguistics*, páginas 464–482. Oxford University Press, Oxford.