

**CONSTRUÇÃO DO ONTOLÉXICO DO DOMÍNIO  
'INDÚSTRIA DO BORDADO DE IBITINGA':  
HERDANDO CONCEITOS E RELAÇÕES DA WORDNET DE PRINCETON**

Erasmio Roberto MARCELLINO\*  
Bento Carlos DIAS-DA-SILVA\*\*

*ABSTRACT: This study focuses on some MS research developments the goal of which is to construct an "ontolexicon" linking both concepts and word forms that are used to conceptualize and to talk about Ibitinga's embroidery industry. Linguistic and Computational Linguistic literature on ontology and lexicon construction underpins the research, which consists of investigating formal ways of developing domain ontologies and aligning such ontologies and wordnets, which, in turn, helps aligning wordnets of different languages. In particular, this paper explores and tests some methodological paths to test ways of inheriting hierarchical relations from Princeton WordNet in the construction of Ibitinga's embroidery industry ontolexicon.*

*KEYWORDS: lexicon; wordnets; ontology; ontolexicon.*

## 1. Introdução

A pesquisa de mestrado, que tem parte de seus desenvolvimentos descrita aqui, visa à sistematização, do ponto de vista linguístico-computacional, do domínio conceitual da Indústria do Bordado Ibitinguense (IBI) em termos de uma ontologia do domínio e nela "ancorar" as parcelas correspondentes dos léxicos correspondentes do português e do inglês. Fundamentando-se na teoria da semântica lexical, pura e computacional, com auxílio de pesquisas em *corpus*, assim como na metodologia de construção de ontologias, a pesquisa deverá culminar com a proposição de um ontoléxico do domínio conceitual da IBI.

Realizamos os objetivos da proposta da pesquisa em termos de duas grandes atividades complementares:

- (I) construção de uma ontologia do domínio conceitual da IBI;
- (II) ancoragem, nessa ontologia, das parcelas dos léxicos correspondentes das duas línguas.

As atividades em I consistem em (Ia) determinarmos os conceitos do domínio e (Ib) descrevê-los em termos de glosas (isto é, definições informais) e dos diferentes tipos de relações que se estabelecem entre eles, para, então, propor a ontologia nos moldes indicados na literatura. Já em II, procede-se à (IIa)<sup>1</sup> seleção, em *corpus* e nas redes WordNet de Princeton e FrameNet de Berkeley, dos itens lexicais que comporão as parcelas dos léxicos da IBI de cada uma das duas línguas e que serão estruturados em termos de sua ancoragem à ontologia, isto é, em termos da (IIc) especificação da relação de significação entre o item lexical e o conceito da ontologia por ele denotado e do (IId) alinhamento semântico entre os itens lexicais das duas línguas decorrente dessa ancoragem, que resulta no que se denomina,

\* Mestrando; UNESP – Universidade Estadual Paulista Júlio de Mesquita Filho, Campus de Araraquara.

\*\* Professor Doutor; UNESP – Universidade Estadual Paulista Júlio de Mesquita Filho, Campus de Araraquara.

<sup>1</sup> Como a atividade Ia' (coleta de itens lexicais) é concomitante à atividade Ia (identificação de conceitos), usamos a' para indicar o paralelismo.

neste estudo, de “ontoléxico” do domínio da IBI, um dos alvos aplicados da pesquisa que, ao ser implementado, tem potencial para gerar automaticamente, a partir dele, um produto como um dicionário bilíngue.

Uma parte da motivação dessa pesquisa advém de estudos e atividades teóricas e práticas de Iniciação Científica, principalmente no que diz respeito à representação do léxico em redes *wordnets*. A outra advém do fato da atividade, a partir da qual se recorta o domínio conceitual, a IBI, ter importância para o artesanato, indústria e cultura da região e oferecer material lexical rico para ser tratado do ponto de vista adotado neste estudo, que se reveste também de originalidade, posto que não há estudos análogos.

Para explicitarmos tanto a motivação quanto a justificativa desta investigação contextualizamos o domínio IBI.

A cidade de Ibitinga, no interior paulista, adquiriu importância graças à atividade do bordado, introduzida na cidade, em meados de 1950, pela imigrante portuguesa Dioguina Sampaio. Desde a década de 60, quando da formação da “Escola de Bordados Singer”, essa atividade vem se desenvolvendo. A cidade, que desde 1974 abriga a tradicional “Feira do Bordado de Ibitinga”, fica reconhecida como a Capital do Bordado entre as décadas de 80 e 90, período de mudanças para a indústria, que se reestrutura, para acompanhar o cenário nacional e mundial, ampliando e diversificando sua produção.

Em Ibitinga, o bordado – desde aquele confeccionado artesanalmente até o mais tecnológico, com produção em grande escala, fruto de tecnologias desenvolvidas especificamente para o setor – alimenta uma indústria que agrega inúmeros profissionais (bordadeira, costureira, overloquista, dentro outros) e utiliza os mais diversos materiais e instrumentos (linha de bordar, fio de ouro, máquina de bordar e bastidor, por exemplo). Por isso, produzi-lo exige conhecimentos técnicos e artísticos que, em termos linguísticos, traduz-se em um universo lexical rico e específico, e que possibilita a comunicação eficiente entre os profissionais do setor, proporcionando, não só entre esses profissionais, como também entre eles e o público geral, a discursivização de, por exemplo, agentes, técnicas, instrumentos, materiais, suportes, processos e produtos que constituem o universo discursivo dessa importante indústria regional.

Esse universo lexical tem uma contrapartida conceitual, que, conforme dissemos, pode ser sistematizada numa ontologia. Estudar essa forma de representação de conceitos e dos itens lexicais, para esta pesquisa, é estudar a constituição e formalização de léxicos e ontologias e da combinação de ambos em termos de um ontoléxicos, objetos de estudo da seção 3, que é antecedida pela apresentação dos recursos utilizados na pesquisa (seção 2); mais adiante, na seção 4, demonstramos o papel relevante da WordNet de Princeton para a pesquisa, para, então, finalizarmos as discussões na seção 5.

## 2. Recursos para a pesquisa

Os recursos de onde se extraem os conceitos e os itens lexicais do domínio trabalhado na pesquisa constituem-se de: dicionários, enciclopédias, teses, artigos, relatórios, entrevistas, *folders*, catálogos e materiais de divulgação dos produtos comercializados na cidade, livros e glossários que abordam a arte e/ou a indústria do bordado, *corpus*, *framenets* e *wordnets*.

Esses recursos têm uma dupla função: tanto permitem a coleta dos itens lexicais que compõem os léxicos da IBI nas duas línguas alvos da pesquisa (atividade IIa’) quanto auxiliam na identificação dos conceitos e categorias da ontologia do domínio da IBI (atividade Ia). A ontologia, por sua vez, ao mesmo tempo em que ancora conceitualmente os itens lexicais em suas categorias (atividade IIc), também motiva a busca e seleção dos itens lexicais que se associam aos conceitos que nela estão estruturados.

Assim, o trabalho empírico concentra-se na coleta de itens lexicais do português e do inglês que denotam conceitos do domínio conceitual da IBI, sem se descuidar, quando relevante para a descrição do ontolêxico da IBI, da coleta de itens lexicais que denotam conceitos do domínio mais geral da indústria do bordado (IB). Na IB, temos inseridos, então, todos os conceitos relativos à indústria do bordado, que, por sua vez, não lhe são necessariamente exclusivos e podem ser compartilhados com outros domínios (por exemplo, o conceito TESOURA)<sup>2</sup>.

Como este texto foca recursos como as redes *wordnets*, então, uma descrição mais aprofundada é oferecida, aqui. O construto computacional WordNet (FELLBAUM, 1998), doravante WN.Pr, foi desenvolvido por George Miller e sua equipe, entre as décadas de 1980 e 90, na Universidade de Princeton. Trata-se de uma rede que estrutura os conceitos expressos no léxico do inglês norte-americano e organizados em termos de *synsets* (*synonym sets* = conjuntos de sinônimos). Um *synset*, fundamentando-se na sinonímia contextualmente motivada, reúne itens lexicais como *embroidery* e *fancywork*, porque podem ser usados para expressar um mesmo conceito em um dado contexto<sup>3</sup>.

Além da sinonímia, que agrupa os itens lexicais em *synsets*, a WN.Pr abriga outros três tipos de relações *entre os substantivos*: a antonímia (oposição de sentidos), a hiponímia/hiperonímia (subordinação/superordenação) e a meronímia/holonímia (parte-todo), que relacionam os *synsets* (isto é, os conceitos lexicalizados)<sup>4</sup>.

Desse modo, na constituição da rede, cada *synset* é um nó e cada relação que ele estabelece com outros *synsets* é um arco. O Quadro 1 exemplifica as relações que estruturam uma rede como a WN.Pr.

SYNSETS			
(a) { <i>tambour1, embroidery frame, embroidery hoop</i> }			
(b) { <i>framework, frame2, framing</i> }			
(c) { <i>brace, bracing</i> }			
RELAÇÕES SEMÂNTICO-CONCEITUAIS			
hiperonímia / hiponímia		meronímia / holonímia	
(b) é hiperônimo de (a)	(a) é hipônimo de (b)	(b) tem (c) como parte	(c) é parte de (b)

Quadro 1 – Estruturação *léxico-conceitual* da WN.Pr.

Além de *synsets* formados por substantivos, a WN.Pr contém também *synsets* formados por advérbios, verbos, para os quais prevê as relações semânticas como a troponímia e acarretamento<sup>5</sup>, e adjetivos, para os quais prevê a relação de antonímia, também prevista para os *synsets* de substantivos.

<sup>2</sup> A notação em caixa alta nomeia conceitos.

<sup>3</sup> Entendido o fato de que a sinonímia exata é rara em línguas naturais, para a WN.Pr, são considerados sinônimos os itens lexicais que são intercambiáveis em um dado contexto, ou seja, compartilham um mesmo conceito.

<sup>4</sup> A WN.Pr estrutura-se, então, em termos de relações lexicais (entre os itens lexicais sinônimos que compõem os *synsets*) e relações conceituais (entre os conceitos da rede, lexicalizados nos *synsets*).

<sup>5</sup> Troponímia é um termo cunhado pelos desenvolvedores da WN.Pr para denotar a relação de hiponímia entre *synsets* de verbos. Por exemplo: {*embroider, broider*} tem como tropônimo {*purl*} (bordar com linha de ouro ou prata), ou seja, este codifica um modo particular de executar a ação codificada naquele; já a relação (unilateral) de acarretamento entre *synsets* de verbos pode ser exemplificada pelos *synsets* {*dream*} e {*sleep, kip, slumber, log Z's, catch some Z's*}, em que o primeiro acarreta o segundo.

Seguindo a metodologia de montagem da WN.Pr para a descrição do léxico do inglês norte-americano, outros projetos foram propostos para o desenvolvimento de outras wordnets e de redes *wordnets* multilíngues, como a EuroWordNet (VOSSSEN, 1998), uma rede *multiwordnet* que alinha semanticamente as redes *wordnets* em construção para as línguas da União Europeia. A WordNet.Br (DIAS-DA-SILVA, 2006, 2004), doravante WN.Br, motivadora de estudos e produções de nossa Iniciação Científica (MARCELLINO, 2008; MARCELLINO; DIAS-DA-SILVA, 2008; RODRIGUES; MARCELLINO; DIAS-DA-SILVA, 2008) e da pesquisa, aqui, descrita, é uma iniciativa, em andamento, de construção da rede *wordnet* para o português brasileiro.

A importância desse tipo de rede se reflete nos diversos trabalhos de PLN (processamento automático de línguas naturais) que o utilizam de várias maneiras, inclusive aproveitando a ontologia que lhe subjaz: “A ontologia implícita nas hierarquias dos substantivos têm recebido especial atenção dos linguistas computacionais” (FELLBAUM, 1998, p. 44 – tradução livre).

### 3. Três construtos-chave: ontologia, léxico e ontoléxico

#### *Ontologias*

Vossen (2003) discute que, no processamento de informações, valemo-nos de informações de naturezas distintas, armazenadas em léxicos e ontologias. Para ele, não há consenso na identificação de exatamente quais são as semelhanças e as diferenças entre léxicos e ontologias, pois as informações que ambos os construtos podem conter podem se sobrepor umas às outras, além de ambos poderem também ser abordados de diferentes maneiras. Por exemplo, a estruturação do conhecimento em ontologias depende de como uma dada teoria aborda os itens lexicais e os conceitos e do propósito a ser atingido com a estruturação. Uma vez que tradições teóricas diferentes propõem diferentes concepções de ontologia para atingir os seus objetivos, não é tarefa fácil estabelecer um consenso sobre o que seja uma ontologia. Diante dessa indefinição, precisamos contextualizar e definir o que entendemos por ontologia: adotamos a noção vigente no âmbito da Representação do Conhecimento (GELLER, PERL e LEE, 2004), que é a que se utiliza no estudo do PLN.

Como mostram Geller, Perl e Lee (*op. cit.*), em levantamento histórico, quando Ross Quillian publicou o artigo *Semantic Memory*, em 1968, descrevendo um programa de computador que gerava expressões simples de língua natural, ele alcançou um feito que inspirou, dentre outras coisas, o desenvolvimento do campo de estudos que seria denominado Representação do Conhecimento. Um dos grandes marcos nesse campo deu-se no início da década de 1990, com Thomas Gruber, que lhe oferece uma abordagem diferenciada, a da construção de ontologias: “Uma especificação de um vocabulário representacional para um domínio de discurso compartilhado – definições de classes, relações, funções e outros objetos, é chamada ontologia” (GRUBER, 1993, p. 199 – tradução livre).

Consideradas um tipo de especificação explícita de uma conceitualização<sup>6</sup>, as ontologias a que se refere Gruber têm como objetivo:

[...] prover conhecimento sobre domínios específicos que seja inteligível tanto por computadores quanto pelos seus desenvolvedores. Especificamente, as ontologias enumeram os conceitos de um domínio e as relações entre eles. Elas podem também

---

<sup>6</sup> De acordo com Gruber (1993, p.199), toda base ou sistema de conhecimento está, explícita ou implicitamente, envolvido com alguma conceitualização, ou seja, com uma visão do mundo abstrata e simplificada que por algum motivo se deseja representar.

definir explicitamente propriedades, funções, restrições e axiomas. (ZHOU, 2007, p. 242 – tradução livre)

Esse objetivo é detalhado em Chishman (2009, p. 113):

[...] (i) compartilhar conhecimento estruturado de informações comuns entre pessoas e máquinas (sistemas computacionais); (ii) possibilitar o reuso do conhecimento de determinado domínio; (iii) tornar explícito o conhecimento sobre determinado domínio; (iv) separar o conhecimento de um domínio do conhecimento operacional de construção de um sistema; (v) analisar o conhecimento de um domínio.

Conforme ensina Zhou (2007), o desenvolvimento de uma ontologia envolve:

- (a) a representação formal, que, além de tornar a ontologia compreensível por computadores e humanos, deve também possibilitar inferências eficientes;
- (b) a aquisição ou a criação dos conteúdos, como conceitos e relações, que, na maioria das vezes, depende de engenheiros do conhecimento ou de especialistas do domínio;
- (c) a avaliação, para aprimorar a qualidade da ontologia e a interoperabilidade entre sistemas; e
- (d) a manutenção, que envolve a organização, a pesquisa e a atualização das ontologias existentes.

Ou ainda, como discutem Ding e Foo (2002), uma ontologia pode ser criada do zero, a partir de ontologias já existentes, de fontes de informação provenientes de *corpus* ou de uma combinação dessas duas últimas, variando, no que diz respeito aos graus de automação, desde o totalmente manual, passando pelo semi-automatizado, até o totalmente automatizado. Quanto ao método que gera uma ontologia, ele pode ser *bottom-up*, parte dos conceitos mais específicos em direção aos mais gerais, *top-down*, parte dos conceitos mais gerais em direção aos mais específicos, ou *middle-out*, parte dos conceitos mais importantes em direção aos mais gerais e aos mais específicos.

Além de se beneficiar das ontologias, esta pesquisa, como os estudos do PLN, busca também agregar às suas investigações os léxicos computacionais.

### **Léxicos**

Handke (1995) lembra-nos de que os itens lexicais podem ser armazenados na mente, em livros de referência e em dispositivos de armazenamento conectados a computadores, conforme mostra a Figura 1.

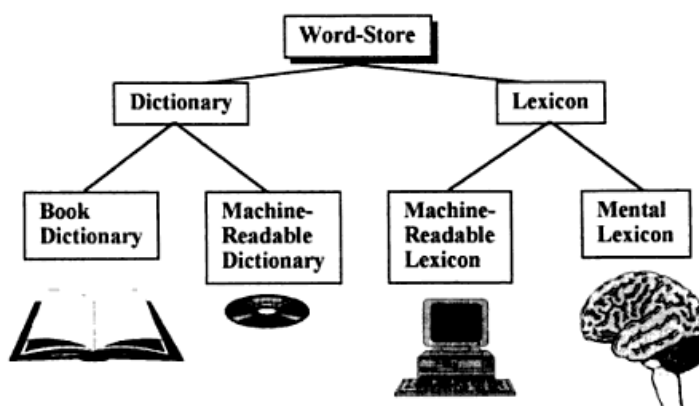


Figura 1 – Uma tipologia de acervos de itens lexicais (HANDKE, 1995, p. 49).

Os recursos para esta pesquisa descritos na seção 2 podem ser, então, conformados com a representação dos tipos de acervo de palavras descrito na Figura 1: de um lado, temos, para *Dictionary*, tanto obras impressas quanto obras em CD-ROM, de outro, temos, para *Lexicon*, constructos como as redes *wordnets* e *framenet*.

Neste ponto, é importante ressaltar que um léxico computacional é a representação formal, com vistas a aplicações em PLN, de parte de um léxico, e que sua capacidade representacional depende do refinamento das relações hierárquicas que contém e da sua ancoragem conceitual.

Constituindo parte significativa do acervo de itens lexicais de uma língua natural, o léxico é, pois:

[...] o módulo central de um sistema de processamento de língua natural, seja do homem ou da máquina. Ele interage intimamente com os outros componentes do processador da língua e fornece informações detalhadas sobre as palavras a serem produzidas ou compreendidas. (HANDKE, 1995, p. 50 – tradução livre)

As informações associadas aos itens lexicais são bem complexas e, por motivos de eficiência na estocagem dessas informações, mas não só por esse motivo, a organização dos itens lexicais no léxico requer o estabelecimento de diversas relações dentro dele (HANDKE, 1995, p. 108). Organizar itens lexicais por meio de relações conceituais (hiponímia, meronímia, etc.) é o que se tem feito na construção de léxicos computacionais como as redes *wordnets*. No entanto, é possível também estabelecer relações entre conceitos sem a ancoragem em línguas naturais, e é o que se tem feito na construção de ontologias. Da proposta de metodologia que se ampare nesses dois construtos nasce a ideia de construção dos ontoléxicos.

### ***Ontoléxicos***

Sobre o conhecimento que as ontologias e os léxicos abarcam, problematiza Vossen (2003): “[...] a diferença entre ontologias e léxicos não é bem definida e geralmente há uma grande sobreposição na informação que eles contêm.” (p. 465 – tradução livre). Chishman (2009, p. 105) explicita o cruzamento que pode haver entre as descrições de ambos:

Os léxicos computacionais, do ponto de vista linguístico, possuem uma relação estreita com as ontologias. As ontologias como estrutura conceitual, que apresentam relações de significados entre os diferentes conceitos que estruturam um determinado conhecimento de mundo, podem incluir ou não o conhecimento linguístico. De uma forma geral, as ontologias que descrevem conceitos mais gerais são conhecidas como ontologias de *nível superior*, ou *top-level*. As ontologias de domínio descrevem o vocabulário relacionado a uma área em especial.

Há um interesse cada vez maior na união desses dois tipos de conhecimento (linguístico e ontológico), sendo esse, de acordo com Prévot, Borgo e Oltramari (2005), “[...] um ponto central para as ferramentas da próxima geração enviadas pela Web Semântica, onde compartilhamento de conhecimento, integração de informação, interoperabilidade e adequação semântica são os principais requisitos.” (p. 91 – tradução livre). Nesse cenário, o que se busca com os sistemas e aplicações do PLN é:

[...] acessar o conteúdo informacional de textos através da interpretação de suas estruturas linguísticas. Para realizar as tarefas, os sistemas de PLN precisam conhecer as partes relevantes do conhecimento a ser identificado nos textos bem

como esse conhecimento é codificado nas expressões linguísticas. O papel dos recursos ontolégicos é suprir os sistemas de PLN com esses dois tipos cruciais de informação. (LENCI, 2010, p. 242 – tradução livre)

A interface ontolégico é, então, “[...] uma tentativa de resposta à crescente necessidade de se modelar as complexas inter-relações entre léxicos e ontologias, que estão cada vez mais assumindo a forma de ricos recursos ontolégicos.” (LENCI, 2010, p. 242 – tradução livre).

Hirst (2004, p. 222) contrasta as ‘lexically based ontologies’ e os ‘ontologically based lexicons’, ou seja, tanto a possibilidade de uma ontologia poder servir de base para a construção de léxicos quanto a possibilidade de léxicos estruturados semanticamente poderem servir de base para a construção de uma ontologia, sobretudo, em se tratando da construção da ontologia de um domínio técnico, no qual a correspondência entre os itens lexicais e os conceitos da ontologia do domínio é mais próxima do que na construção de ontologias para domínios gerais, como a dos conceitos que são denotados por itens lexicais da língua geral:

[...] em domínios técnicos onde existem vocabulários explícitos (incluindo glossários, léxicos, dicionários de termos técnicos, etc., apoiados ou não por uma autoridade), uma ontologia existe pelo menos implicitamente [...] E onde uma ontologia explícita existe, um vocabulário explícito certamente também existe; na verdade, frequentemente se diz que a construção de qualquer ontologia de domínio específico implica a construção paralela de um vocabulário para ela [...] (HIRST, 2004, p. 223 – tradução livre)

Quando Hirst (2004) explica que é comum haver a construção paralela de uma ontologia e de parcelas de léxico (vocabulário), ele está descrevendo procedimentos aproximados aos que adotamos na pesquisa aqui descrita: a determinação dos conceitos da ontologia, prevista na atividade (Ia), é acompanhada pela determinação do relacionamento de significado entre itens lexicais e conceitos da ontologia (atividade IIc); ou seja, parte da ontologia é construída juntamente com parte do revestimento lexical ancorado a ela no processo de edificação do ontolégico.

Os desenvolvimentos em direção à concretização dessa proposta de ontolégico são descritos na Seção 4.

#### **4. Herdando informações da WN.Pr na construção do ontolégico**

É importante ressaltar que as atividades apresentadas na Introdução e detalhadas nesta seção permeiam os três níveis de investigação da metodologia para o PLN proposta por Dias-da-Silva (1996, 2006): o linguístico, o linguístico-computacional e o computacional, que correspondem, respectivamente, a:

[...] a “extração do solo” (isto é, a explicitação dos conhecimentos e habilidades linguísticas), a “lapidação” (isto é, a representação formal desses conhecimentos e habilidades) e a “incrustação” (isto é, a construção do programa de computador que codifica essa representação). (DIAS-DA-SILVA, 2006, p. 122)

Em cada nível do PLN, se entrecruzam variadas disciplinas, como a Inteligência Artificial, as Ciências da Computação, a Filosofia da Linguagem, a Linguística, etc., cada uma oferecendo os recursos teóricos e metodológicos de sua especialidade. No que diz respeito à construção de ontologias, a interação com os recursos linguísticos vem acrescentando novas possibilidades aos produtos desenvolvidos, pois ontologias

linguisticamente motivadas, ou ontológicos, segundo já nos atestaram Chishman (2009) e Prévot, Borgo e Oltramari (2005), são o futuro da Web Semântica.

A WN.Pr, conforme adiantamos na seção 2, vem sendo utilizada à exaustão em diversas empreitadas do PLN, sendo igualmente importante para a nossa pesquisa, pois, como veremos, sua metodologia de construção fornece técnicas auxiliares para algumas de nossas mais importantes atividades.

Para representar uma parte do léxico mental da língua para a qual é produzida, uma rede *wordnet* formaliza partes desse léxico nos *synsets*. Para o português brasileiro, por exemplo, construímos o *synset* da Figura 2 para representar o conceito BORDADOR.

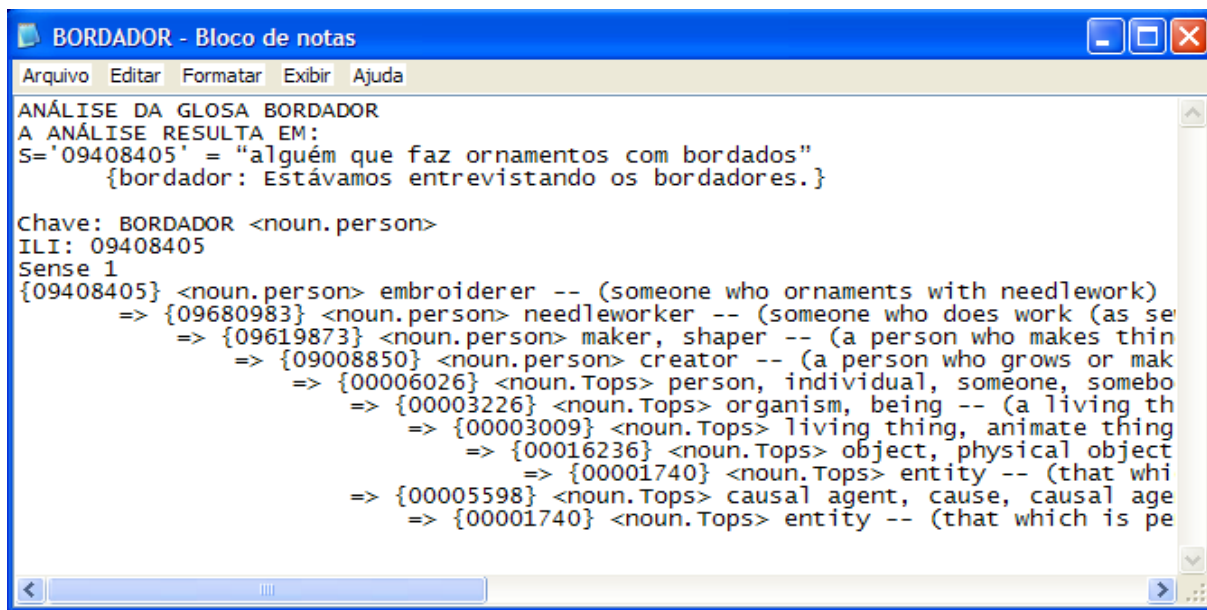


Figura 2 – O *synset* {bordador} proposto para a WN.Br.

Dessa maneira, se ao conceito C da ontologia da IBI for associado o *synset* P do português e o E do inglês, P e E serão alinhados, porque denotam o mesmo conceito, isto é, são *synsets* alinhados por meio da relação de *EQ\_SYNONYM*, conforme proposta de alinhamento de *synsets* descritas em Ide, Greenstein e Vossen (1998), em que são também descritos estes tipos complexos de alinhamento indireto:

- P é mais específico que E: P *EQ\_HAS\_HYPERONYM* E;
- P é mais genérico que E: P *EQ\_HAS\_HYPONYM* E;
- P associa-se a E e a E' por *EQ\_SYNONYM*: P *EQ\_NEAR\_SYNONYM* E;
- P e P' associam-se a E por *EQ\_SYNONYM*: P *EQ\_NEAR\_SYNONYM* E.

Assim, além de permitir a construção do *synset* P da Figura 2, com sua glosa e frase-exemplo extraída de *corpus*, o alinhamento desse *synset* ao seu correspondente da WN.Pr, {*embroiderer*}, permite também **herdar a estrutura hierárquica dessa rede**. Ponto crucial no desenvolvimento da pesquisa que decidimos expor neste trabalho, essa herança contribui para a construção da ontologia da IBI em pelo menos três aspectos muito importantes, pois ela:

- permite a “navegação” pelos conceitos; por exemplo, verificamos, através do *synset* hiperônimo {*needleworker*}, presente na Figura 2, a quais outros profissionais que trabalham com agulha, a rede WN.Pr dá acesso, ou seja, é

possível a identificação de novos conceitos (*synsets*) e de novas relações léxico-conceituais estabelecidas entre eles;

- permite o “reuso” de conhecimento; por exemplo, ao estipularmos a organização da ontologia da IBI com base na WN.Pr, temos uma hierarquização com potencial de “interagir” com outras ontologias;
- permite a visualização, mesmo que parcial, de como pode ser a ontologia da IB, na qual se encaixa a da IBI.

Se, de maneira análoga ao *synset* da Figura 2 – que foi construído no arquivo padrão que segue as especificações necessárias para que seja possível sua implementação no editor<sup>7</sup> da WN.Br – fosse elaborado um *synset* para cada conceito e categoria identificados no domínio da IBI, seria possível promover um exercício de implementação, por meio do alinhamento entre redes *wordnets*, do ontoléxico alvo desta pesquisa, cumprindo todas as atividades propostas: uma vez determinado o conceito da ontologia (atividade Ia); procede-se à coleta do item lexical (ou itens) no português que o denota, isso caso o conceito já não tenha sido determinado a partir do item lexical, mas em qualquer caso (atividades IIa’ e IIc); com os dados conceituais e lexicais já levantados, o *synset* pode ser construído, frases-exemplos extraídas dos recursos e uma glosa proposta (parte da atividade Ib); o *synset* pode, então, ser alinhado ao ILI, que corresponde ao seu conceito e ao qual também foi alinhado um *synset* da WN.Pr, o que promove a sua conexão direta, via relação *EQ\_SYNONYM* com {*embroiderer*} (parte da atividade Id), e indireta, com os *synsets* da rede, como **herança de relações semânticas** (parte da atividade Ib).

Além do alinhamento por meio da relação de sinonímia entre *synsets* – como demonstramos em {*bordador*} *EQ\_SYNONYM* {*embroiderer*} – podemos, conforme adiantaram Ide, Greenstein e Vossen (1998), nos deparar com a impossibilidade de estabelecer o alinhamento direto, o que reflete lacunas nas estruturas das redes *wordnets* ou lacunas nas línguas, o que é previsto por pesquisadores como Vossen et al. (1998), que identificam os seguintes fenômenos:

- “lacuna cultural” (*cultural gap*), que ocorre quando não há, no léxico da língua, um item lexical específico, porque, para a comunidade de falantes dessa língua, não há o conceito para ser lexicalizado. Por exemplo: o *synset* {*almofadron*}<sup>8</sup>, do português brasileiro, lexicaliza um conceito não partilhado pela comunidade de falantes do inglês;
- “lacuna pragmática” (*pragmatic gap*), que ocorre quando a equivalência se estabelece por meio de construções. Por exemplo: o conceito lexicalizado no *synset* “{06855518} <noun.event> grassfire -- (an uncontrolled fire in a grassy área)”, não se lexicaliza no português, mas é expresso por meio de construções sintagmáticas como “incêndio incontrolável numa área de muita grama”;
- “divergência morfológica” (*morphologic mismatch*), que ocorre quando a equivalência se estabelece entre estruturas gramaticais distintas nas duas

<sup>7</sup> Esse editor possibilita o alinhamento entre a base da WN.Br e a rede WN.Pr através do ILI (Inter-Lingual-Index), uma listagem de todos os *synsets* da WN.Pr e seus respectivos conceitos glosados.

<sup>8</sup> O conceito ALMOFADRON, glosado por “uma almofada que, quando desempacotada, se transforma num edredon”, certamente não é compartilhado por falantes de outras línguas. No entanto, faz parte da ontologia da IBI, do mesmo modo que outros conceitos da ontologia da IB que porventura sejam desconhecidos pelos falantes do português ou que não se lexicalizem nessa língua integrarão a ontologia que estamos desenvolvendo. A título de curiosidade, informamos que o *almofadron*, que é suporte para vários tipos de bordados, foi criado em 2008 na cidade de Ibitinga.

línguas. Por exemplo, o conceito SAUDADE lexicaliza-se, no português, pelo substantivo *saudade*; no inglês, esse conceito é expresso por um significado particular do verbo *to miss*;

- “lacuna por incompletude das bases”, que ocorre quando, em pelo menos uma das redes *wordnets*, não há registro de *synsets* potenciais. Por exemplo: na base da WN.Pr, não há *synset* contendo o substantivo *overlock machine*. Esse fato impossibilita o estabelecimento do alinhamento direto do *synset* {*máquina de overloque*};
- “lacuna semântica”, que ocorre quando o conceito alvo de uma equivalência não está lexicalizado em uma das redes *wordnets*. Por exemplo: da base da WN.Pr, constam *synsets* que contêm o verbo *lump*; mas não consta o *synset* que lexicaliza o conceito representado no *synset* {*embolotar, encaroçar*} da Wn.Br, embora esse conceito seja partilhado com a comunidade de língua inglesa, conforme atesta o exemplo: “Stir the gravy so that it doesn't lump”.

A existência de lacunas como essas levam Hirst (2004) a nos alertar sobre a impossibilidade de concebermos uma ontologia partindo apenas do léxico, pois este “[...] omitirá qualquer referência a categorias ontológicas que não são lexicalizadas na língua – categorias que requereriam uma descrição multi-palavra (possivelmente longa) para serem referidas na língua.” (p. 218 – tradução livre).

## 5. Conclusão

Neste estudo, procuramos apresentar uma proposta de construção de um ontolético para o domínio conceitual da IBI, até então abordado em outros trabalhos apenas do ponto de vista econômico ou do sócio-cultural. Concluímos que representar os conceitos da ontologia desse domínio em termos de *synsets* é uma possibilidade, bem como aproveitar os tipos de relações fundamentados pela metodologia de alinhamento entre *synsets* de redes *wordnets* diferentes, o que possibilita o compartilhamento de conceitos e a herança de relações entre eles. Essas investigações apontam também que se faz necessário procurar outros tipos de estruturação dos itens do ontolético que não os previstos na metodologia das redes *wordnets*, para que seja possível, por exemplo, relacionar os conceitos BORDADOR, MÁQUINA DE BORDAR e BORDADO. Estudos futuros pretendem investigar, na rede FrameNet, que implementa uma semântica de frames, complementações para essa questão do relacionamento de conceitos.

## Referências

- CHISHMAN, R. L. O. Integrando léxicos semânticos e ontologias: uma aproximação a favor da Web Semântica. *Informação & Informação*, Londrina, v. 14, n. esp., p. 103-124, 2009.
- DIAS-DA-SILVA, B. C. O estudo linguístico-computacional da linguagem. *Letras de Hoje*, Porto Alegre, v. 41. p. 103-138. 2006. ISSN 0101-3335.
- \_\_\_\_\_. Wordnet.Br: an exercise of human language technology reserch. *Palavra*, Rio de Janeiro, v. 12, p. 15-24. 2004. ISSN 1413-7763.
- \_\_\_\_\_. *A face tecnológica dos estudos da linguagem: o processamento automático das línguas naturais*. Araraquara, 1996. 272 p. Tese (Doutorado em Letras) – Faculdade de Ciências e Letras, Universidade Estadual Paulista, Araraquara. 1996.

- DING, Y.; FOO, S. Ontology Research and Development part 1 – A review of ontology generation. *Journal of Information Science*, [S.l.] v. 28, n. 2, p. 123-136, abr. 2002.
- FELLBAUM, C. (Ed.) *WordNet: an electronic lexical database*. Cambridge, Massachusetts: Cambridge University Press, 1998.
- GELLER, J.; PERL, Y.; LEE, J. Editorial: Ontology Challenges: A Thumbnail Historical Perspective. *Knowledge and Information Systems*, London, v. 6, n. 4, p. 375-379, 2004.
- GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition*, Stanford, v. 5, n. 2, p. 199-220, jun. 1993.
- HANDKE, J. *The structure of the Lexicon: human versus machine*. Berlin: Mouton de Gruyter, 1995.
- HIRST, G. Ontology and the Lexicon. In: STAAB, S.; STUDER, S. (Ed.). *Handbook on Ontologies*. Berlin: Springer-Verlag, 2004, p. 209-229.
- IDE, N.; GREENSTEIN, D.; VOSSSEN, P. Special Issue on EuroWordNet. *Computers and the Humanities*, Netherlands, v. 32, n. 2-3, 1998.
- LENCI, A. The life cycle of knowledge. In: HUANG, C.; CALZOLARI, N.; GANGEMI, A.; LENCI, A.; OLTRAMARI, A.; PREVOT, L. (Eds.) *Ontology and the Lexicon: A Natural Language Processing Perspective*. Cambridge: Cambridge University Press, 2010, p. 241-257.
- MARCELLINO, E. R. A representação das lacunas pragmáticas no alinhamento de wordnets. In: SEMINÁRIO DO GEL, n. 56, 2008, São José do Rio Preto. *Programação...* São José do Rio Preto: [s.n.], 2008.
- MARCELLINO, E. R.; DIAS-DA-SILVA, B. C. A aplicação de relações de equivalência complexas na co-indexação entre wordnets. In: CONGRESSO DE INICIAÇÃO CIENTÍFICA DA UFSCAR, n. 16, 2008, São Carlos. *Anais...* São Carlos: [s.n.], 2008, p. 729, v. 4. 1 CD-ROM
- PRÉVOT, L.; BORGIO, S.; OLTRAMARI, A. Interfacing Ontologies and Lexical Resources. In: ONTOLEX, 2005, Jeju Island. *Proceedings...* Jeju Island: [s.n.], 2005. p. 91-102.
- RODRIGUES, J. O.; MARCELLINO, E. R.; DIAS-DA-SILVA, B. C. Co-Indexação léxico-semântica de synsets de substantivos entre wordnets. In: SIMPÓSIO INTERNACIONAL DE INICIAÇÃO CIENTÍFICA DA USP, n. 16, 2008, São Paulo. *Anais...* São Paulo: [s.n.], 2008.
- VOSSSEN, P. Ontologies. In: MITKOV, R. (Ed.). *The Handbook of Computational Linguistics*. Oxford: Oxford University Press, 2003. p. 464-482.
- VOSSSEN, P. *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Dordrecht: Kluwer, 1998.
- VOSSSEN, P.; BLOKSMA, L.; ALONGE, A.; MARINAI, E.; PETERS, C.; CASTELLON, I.; MARTI, A.; RIGAU, G. Compatibility and interpretation of relations in EuroWordNet. *Computers and the Humanities*, Netherlands, v. 32, n. 2-3, p. 153-184, 1998.
- ZHOU, L. Ontology learning: state of the art and open issues. *Information Technology and Management*, [S.l.], v. 8, n. 3, p. 241-252, 2007.