Produção de Ontologias Específicas: a modelagem da *OntoEco*

Claudia Zavaglia[1]

[1]Instituto de Biociências, Letras e Ciências Exatas – Universidade Estadual Paulista (UNESP)
R. Cristóvão Colombo, 2265 – 15.054-000 – São José do Rio Preto – SP – Brasil
zavaglia@lem.ibilce.unesp.br

Abstract. *This paper describes the specific ontologies of restricted domains, in this case, the domain of Ecology. We attempt to demonstrate how the construction of an ontology and its formalization work in practice. We relate the steps that we followed and the choices that we made with the intention of serving an effective computational treatment of the proposed tree-like structure.*

Keywords. *Ontology**;** Computational Linguistics; Qualia Structure.*

Resumo. *Este artigo trata de ontologias de domínios restritos, no caso o da Ecologia. Procuramos demonstrar como se verifica, na prática o delineamento de uma ontologia e a sua formalização. Relatamos os passos que seguimos e as decisões tomadas com o intuito de servir a um tratamento computacional eficiente da estruturação arbórea proposta.*

Palavras-chave. *Ontologia**,** Lingüística Computacional; Estrutura Qualia.*

1. Introdução

Nossa pesquisa faz parte de um projeto maior que busca investigar a problemática da elaboração de base de dados lexicais e de ontologias específicas, sendo estas de domínios especiais, desenvolvido na Universidade Estadual Paulista – Câmpus de São José do Rio Preto – Brasil. A partir da ontologia proposta por Zavaglia (2002) para o subdomínio da Ecologia, reestruturamos o conhecimento especializado desse subdomínio em classes e subclasses, segundo os preceitos de Gruber (1993), e descrevemos formalmente seus conceitos e as relações existentes entre eles. O uso de ontologias tem sido amplamente empregado em representações do conhecimento de domínios restritos, máxime para sistemas de busca de informação e indexação de documentos, em Processamento de Línguas Naturais - PLN, no qual a sua aplicação pode ser mais eficaz por tratar, justamente, de conjuntos léxicos de número finito. Por sua vez, em Bases de Conhecimento Lexical – BCL, por exemplo, o uso de uma ontologia pode servir como recurso de apoio à informação contida no repositório lexical dessa base para que seja possível resgatar o significado de um item léxico de forma unívoca.

A *OntoEco* prevê três sub-domínios da Ecologia, a saber: *Ecologia de Ecossistemas* – EEc; *Ecologia de Populações* – Ep; *Ecologia de Comunidades* – Ec, que se revelaram altamente produtivos tanto no processo de categorização, quanto no

processo de nomeação de termos.

Em se tratando de armazenamento de dados, de registro de informações, de organização, estruturação e busca de conhecimento, o computador é visto atualmente como a principal ferramenta para auxiliar a todas essas tarefas. Em se tratando de estocagem de dados com informação semântica, como é o caso da armazenagem de conhecimento ontológico, torna-se necessária a "existência de representações de conhecimento explícitas que possam armazenar informações de forma acessível aos programas", como pontuam, Mangam et al. (s.d.). Esses mesmos autores apontam para a não disponibilização do conhecimento, fato esse que se deve a dois fatores, primordialmente: (i) a não existência de uma representação computacional que esteja disponível para esse tipo de conhecimento formalizado, além da possibilidade de o conhecimento ser intratável computacionalmente, ao mesmo tempo em que o conhecimento pode ser tratável mas deve ser resgatado de forma adequada e (ii) quando a representação semântica está disponível, ela é inadequada aos padrões de processamento que se almeja, ou então, o resgate do conhecimento já ocorreu, mas a representação selecionada para a estocagem de dados é imprópria para o processamento específico para o qual se destina.

2. Objetivos

O trabalho tem como objetivo geral realizar um levantamento dos itens lexicais referentes ao subdomínio da Ecologia em língua portuguesa. A partir das unidades ontológicas detectadas, temos como escopo específico traçar a Ontologia desse subdomínio, para a qual especificaremos as relações semânticas que os itens mantêm com as suas classes, subclasses e itens lexicais afins. Ademais, esses itens lexicais serão etiquetados manualmente, contendo informações morfossintáticas e informações semânticas concernentes à Estrutura *Qualia* do Léxico Gerativo (LG) de Pustejovsky (1995). Como escopo concreto, pretendemos implementar computacionalmente todos os dados que serão elaborados para a *OntoEco*, na ferramenta computacional Protégé-2000.

Neste artigo, a tese que se defende é a de que o conhecimento pode ser disponibilizado para sistemas computacionais, desde que seja utilizada uma técnica de representação para o domínio tecnológico que seja aceita pela sua comunidade. Uma dessas técnicas é, justamente, a modelagem do conhecimento por meio de Ontologias. Com efeito, Mangan et al (s.d.) dizem que "Ontologias e Modelagem de Domínios são duas técnicas de grande aceitação no domínio tecnológico da gestão do conhecimento". E ainda: "Modelos de domínio são usados, principalmente, pela comunidade de reuso de software". Ontologias são aplicadas, principalmente, pela comunidade de inteligência artificial na perspectiva de modelagem de conhecimento".

3. Ontologias: caracterização

Em Lingüística Computacional, ou seja, no campo de ação da representação formal do conhecimento, a ontologia pressupõe um enlace entre os símbolos da linguagem natural e as entidades do mundo real que ela representa. Nesse sentido, pode-se considerá-la como "uma especificação de uma conceptualização" (GRUBER, 1993). Essa especificação tem sido objeto de grande esforço por parte dos investigadores em Inteligência Artificial que há várias décadas buscam uma ontologia que ofereça

flexibilidade suficiente para dar conta de representar o conhecimento complexo registrado na mente humana. Em consonância, Santos et. al (2001) acentuam:

> Para a filosofia, ontologia é o estudo da existência do ser. Em Inteligência Artificial, ontologia pode ser definida como "uma especificação formal e explícita de uma conceituação compartilhada". A palavra conceituação refere-se a uma abstração, visão simplificada do mundo que desejamos representar para algum propósito, construído através da identificação dos conceitos e relações relevantes. O termo explícita indica que os tipos de conceitos e as restrições ao seu uso são explicitamente definidos. Formal significa que a ontologia deve ser compreensível por um computador (não pode ser somente escrita em linguagem natural). (SANTOS, 2001, p.2)

Guarino, N. & Giaretta, P. (1995, p.1) elucidaram diversas interpretações que vêm sendo utilizadas para a palavra "ontologia" com o escopo de esclarecer terminologicamente a escolha técnica do uso desse item lexical. Vejamos as possibilidades de interpretação elencadas por eles: (1) *ontologia* como uma disciplina filosófica; (2) *ontologia* como um sistema conceitual informal; (3) *ontologia* como um cálculo da semântica formal; (4) *ontologia* como uma especificação de uma conceitualização caracterizada por: (i) propriedades formais específicas e (ii) somente por propósitos específicos; (5) *ontologia* como o vocabulário usado por uma teoria lógica; (6) *ontologia* como uma especificação de uma teoria lógica (*meta-level*). Nesse trabalho, a interpretação de *Ontologia* que nos serve é a de número (4). Nesse sentido, inferimos que "conceitualização" é a palavra chave para a representação do conhecimento de maneira formal. Objetos, conceitos e outras entidades existentes em determinada área do conhecimento e as relações entre elas devem ser conceitualizadas. Tais conceitos nada mais são do que uma visão simplificada e resumida do mundo.

4. Metodologia e Desenvolvimento

Para o desenvolvimento desta pesquisa, elaboramos uma base de textos especiais, o *CórpusEco*, concernente ao subdomínio da Ecologia para o português do Brasil. Esse repertório textual conta hoje com cerca de 300.000 ocorrências e está armazenado em uma ferramenta computacional que administra grandes quantidades de dados, o *Folio Views 4.1*. A principal utilidade da elaboração desse córpus foi, justamente, servir de base lingüística para a extração dos termos ontológicos vinculados a *OntoEco*. Essa extração foi feita, primeiramente, de forma manual, utilizando-se o critério semântico no processo de extração. De fato, utilizamos a metodologia da onomasiologia, a partir do momento que partimos do significado ou conceito de um item lexical para o seu significante, ou seja, a identificação da sua forma. Em seguida, partimos para a extração automática dos itens lexicais, com o auxílio de uma ferramenta computacional que extrai de forma automática candidatos a termos, cujo critério adotado foi o da freqüência. Adotamos tal critério uma vez que a alta incidência de ocorrência de um candidato a termo nos indica que ele tem uma grande probabilidade de, com efeito, ser um termo da área especializada em questão, podendo servir como um critério identificador para a seleção dos termos. Em mãos desses resultados, delineamos a ontologia para cada um dos subdomínios especificados anteriormente.

Na etapa seguinte, essas unidades lexicais foram etiquetadas com informações semânticas referentes à Estrutura *Qualia* do LG de Pustejovsky (1995). Esse autor chamou de Estrutura *Qualia* a representação que dá força relacional ao item lexical. O LG analisa todos os itens lexicais como relacionais; o modo em que a sua propriedade é expressa difere de categoria para categoria, bem como entre classes semânticas. A Estrutura *Qualia* especifica quatro papéis essenciais do significado de uma palavra (ou *Qualia*): (i) **Constitutivo ou Partes Constituintes** (*Constitutive*), i.e., aquele que exprime a relação entre um objeto e suas partes constituintes; (ii) **Formal** (*Formal*), ou seja, aquele que identifica o objeto em um domínio mais amplo; (iii) **Télico** (*Telic*), aquele que expressa o objetivo/escopo e a função do objeto e (iv) **Agentivo** (*Agentive*), i.e., aquele que considera fatores envolvidos na origem do objeto.

Após a distribuição dos itens lexicais na estrutura ontológica delineada, definimos e mapeamos as relaçãoes semânticas existentes entre eles presentes nos papéis da *Qualia*. Cada unidade lexical ativa no âmbito de domínio da Ecologia foi delineada a partir de vários campos de valor, como explicitado a seguir:

Tabela 1. Campos de valor dos termos ontológicos

SemU:	Unidade Semântica – unidade lexical ou ontológica.
Tipo:	Subclasse a que o termo pertence.
Supertipo:	Classe a que o termo pertence.
Domínio:	O domínio com o qual trabalhamos, no caso, Ecologia.
Formal: **Agentivo:** **Constitutivo:** **Télico:**	Relações semânticas existentes na Estrutura *Qualia*
Glossário:	Definição do termo.
Exemplo:	Contextualização autêntica da SemU extraída do*CórpusEco*
PDD:	Parte do Discurso – informamos a classe gramatical a que o termo pertence.
MORFOL:	Morfologia do termo.
SemU_syn	Termo sinônimo, caso exista.
SemU_ant	Termo antônimo, caso exista.

4.1. A ferramenta computacional Protégé-2000

Protégé-2000 originou-se a partir de um Projeto desenvolvido no Departamento de Informática Médica (SMI - *Stanford Medical Informatics*), pelo Desenvolvido pelo KMG (*Knowledge Modeling Group*4) da Faculdade de Medicina da Universidade de Stanford. É uma ferramenta computacional integrada, mais especificamente um editor de ontologias, usada para o desenvolvimento de sistemas baseados em conhecimento. Essa feramenta tem por objetivos: (i) consentir a interoperabilidade com outros sistemas de representação do conhecimento; (ii) ser uma ferramenta de aquisição de conhecimento que seja fácil de se configurar e manejar; (iii) ser extensível. Seu modelo de conhecimento é representado por meio de *classes* (conceitos no domínio de discurso – constituem uma hierarquia taxonômica), *instâncias* dessas classes, *slots* (que descrevem as propriedades e atributos das classes e instâncias), *facetas* (que são restrições de informações, especificando informações adicionais sobre propriedades) e *axiomas* que especificam contrastes adicionais, em que: (a) é baseado em *frames*, ou seja, construções em blocos de uma base de conhecimento; (b) usa a arquitetura de *metaclasse*, ou seja, um *template* que é usado para definir novas classes em uma

ontologia; (metaclasse é uma classe cujas instâncias também são classes); (c) possibilita a especificação de herança múltipla e de classes abstratas.

5. Conhecimento Semântico: estruturação e organização

A *OntoEco* encontra-se dividida em duas grandes classes: **CLASSES** e **LEXICAL_UNIT**. A classe **CLASSES** possui uma **META-CLASS** por meio da subclasse **STANDARD-CLASS** implementada como a subclasse **SEM_CLASS_BASE**, ou seja, a Classe semântica-base que definirá o padrão de configuração de todas as classes e subclasses que estiverem vinculadas a elas. O mesmo ocorre para a classe **LEXICAL_UNIT**, que possui uma **META-CLASS** por meio da subclasse **STANDARD-CLASS** implementada como a subclasse **LEXICAL_UNIT_BASE**, ou seja, a Unidade lexical-base que definirá o padrão de configuração de todas as classes e subclasses (itens ontológicos) que estiverem vinculadas a elas. A relação de hiponímia/hiperonímia ou ***é_um*** **(*is-a*)** serviu para organizar diversos termos-conceito. De fato, todos os termos que fazem parte da ontologia possuem a relação ***é_um***, como identificadora do *genus terminus* que a conceitua. Ademais, a relação ***é_um*** é considerada a base de qualquer taxonomia e, de conseqüência, a sua aplicação foi bastante incisiva, como se esperava, a partir do momento que a relação ***é_um*** determina todas as subclasses das duas classes principais **CLASSES** e **LEXICAL_UNIT**. Como tipologia, temos, além de ***é_um*****,** a relação ***é_um_tipo_de***, cuja demarcação limítrofe nem sempre é clara e evidente. À luz da Teoria do Léxico Gerativo, a relação de hiperonímia corresponde às informações veiculadas pelo papel Formal da Estrutura *Qualia*. No Protégé-2000, essa relação está representada por classes e subclasses. Além disso, previmos um *frame* **:FORMAL** para cada classe e subclasse, quando for necessária a sua especificação para a recuperação do conceito veiculado pelas categorias e subcategorias. Vejamos a implementação desssas informações no Protégé-2000:

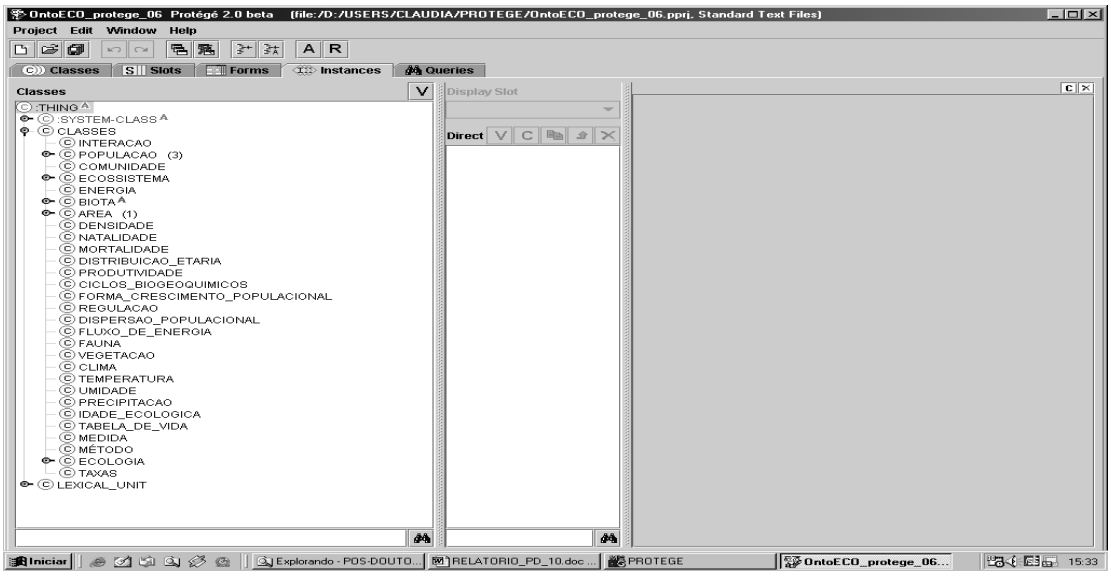

Figura 1. Interface de CLASSES no Protégé-2000

6. Considerações finais

Modelamos e estruturamos conceitualmente três subdomínios do domínio da Ecologia. Todas as suas classes foram conceituadas e implementadas na ferramenta Protégé-2000, respeitando os limites impostos pela máquina. Por sua vez, para esse artigo, os termos ontológicos estudados, estruturados e conceitualizados amiúde foram os pertencentes à Ecologia de Populações, uma vez que tratou-se apenas de um protótipo de reuso e interoperabilidade da representação ontológica proposta. Dito isso, implementamos 29 **SUBCLASSES** da classe **CLASSES**, com suas respectivas subclasses que somam 36. No que diz respeito às unidades lexicais, todos os 64 termos foram implementados com os campos *name* (significante do termo); *documentation* (que possui a definição do termo); *role* (apresentando todos a característica de "concrete"); *template slots* (que atribuem as propriedades, por meio de slots, a cada termo); *antônimo*; *sinônimo*; *morfologia*; *contexto* (reporta a contextualização autêntica do termo retirada do *CórpusEco*). Como exemplificação de uma utilização concreta da *OntoEco* em uma máquina de busca, ao entrarmos com as palavras-chave "População e Ecologia" poderíamos recuperar o conceito expresso por "ecologia de populações", "ecologia" e "população" que o usuário busca conhecer, caso seja a sua intenção de pesquisa, ao invés de termos como respostas páginas que somente nos informam sobre: "Áreas de atuação (em CVs)/atividade/conhecimento"; "grade curricular"; "nomes de disciplinas/programas e linhas de pesquisa de departamentos de universidades/instituto/núcleos (graduação e pos-graduação)"; "sites de professores de ensino fundamental e médio"; "ementas de editais de concurso"; "listagem de projetos", que é o que acontece hoje, em português, ao fazermos uma pesquisa desse tipo.

7. Referências bibliográficas

GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing. *Workshop on Formal Ontology*, Padua, mar. 1993, Edited collection by Nicola Guarino.

GUARINO, N. & GIARETTA, P. Ontologies and Knowledge Bases. *Towards a Terminological Clarification*. Padova, Italy, 1995. Disponível em: http://www.loa-cnr.it/Papers/KBKS95.pdf. Acesso em 20/01/2004.

MANGAN, M. et al. Modelos de Domínio e Ontologias: uma comparação através de um estudo de caso prático em hidrologia. s.d. (in mimmeo)

PUSTEJOVSKY, J. *The Generative Lexicon.* Cambridge: The MIT Press, 1995.

SANTOS, E. T. et al. *Projetando uma Ontologia de Geometria Descritiva*. In: 15 SIMPÓSIO NACIONAL DE GEOMETRIA DESCRITIVA E DESENHO TÉCNICO – IV INTERNATIONAL CONFERENCE ON GRAPHICS ENGINEERING FOR ARTS AND DESIGN. 2001, São Paulo, Brasil. Anais eletrônicos...São Paulo. Disponível em: http://www.cin.ufpe.br/~sas/chat/ontologiafinal.pdf. Acesso em: 03/02/2003

ZAVAGLIA, C. Análise da homonímia no português: tratamento semântico com vistas a procedimentos computacionais. 2002. p.360, v.I; p.199, v.II. Tese (Doutorado em Letras. Área de Concentração: Lingüística e Língua Portuguesa) – Faculdade de Letras, Universidade Estadual Paulista, São Paulo.