O PAPEL DO LÉXICO NA ELABORAÇÃO DE ONTOLOGIAS COMPUTACIONAIS: DO SEU RESGATE À SUA DISPONIBILIZAÇÃO

Claudia **Zavaglia**
UNESP

O léxico é um dos Recursos Lingüísticos primários no que diz respeito à Engenharia da Linguagem, especialmente para a Engenharia Ontológica. De fato, qualquer sistema aplicativo, para analisar ou processar uma língua natural, não pode prescindir do tesouro vocabular de uma língua, seja ele de domínio especializado ou não. Esse repertório lexical, por sua vez, para que seja utilizado por uma máquina, deve conter informações adequadas e codificadas para que o programa computacional ou o algoritmo possa "compreendê-las" e utilizá-las. As informações contidas em um léxico podem ser de vários níveis lingüísticos, a saber, morfológico, sintático, semântico, pragmático, e, desse modo, existem várias etapas de codificação para esses diferentes módulos, em se tratando de programas computacionais.

Uma base de dados semântica constitui um repositório lexical fundamental como fonte de Reservas e Recursos Lingüísticos para os estudos do Processamento de Línguas Naturais (doravante PLN). De fato, a informação semântica é indispensável para programas que analisam e decodificam textos em língua natural.

A elaboração de bases léxico-ontológicas que possuam informações de natureza morfossintática, semântica, ontológica e *qualia* torna-se essencial para que expedientes lingüísticos das mais variadas espécies possam ser utilizados e recuperados em Sistemas de Processamento de Línguas Naturais (SPLN), tais como a Tradução Automática, a Recuperação da Informação, a Web Semântica, os Motores de Busca.

Por meio do levantamento dos itens lexicais referentes ao subdomínio da Ecologia (*Ecologia de Ecossistemas* – EEc; *Ecologia de Populações* – Ep; *Ecologia de Comunidades* – Ec) do domínio das Ciências Biológicas em língua portuguesa, traçamos a estrutura ontológica desse subdomínio e especificamos as relações semânticas que os itens mantêm com as suas classes, subclasses e unidades lexicais afins. Ademais, esses itens lexicais foram etiquetados manualmente, como explicitado adiante, contendo informações morfossintáticas e informações semânticas concernentes à Estrutura *Qualia* do Léxico

Gerativo de Pustejovsky (1995). Como escopo concreto, elaboramos um modelo de *Ontologia do subdomínio da Ecologia – OntoEco* e sugerimos uma interface computacional de acesso aos seus dados.

A organização conceitual da OntoEco assemelha-se àquela de um *thesaurus*, já que os itens lexicais encontram-se correlacionados e interligados por diferentes tipos de relações semânticas.

Utilizamo-nos da mesma execução metodológica de trabalho empregada por Zavaglia (2002), a começar por diversas análises lingüístico-computacionais, a partir do elenco de itens lexicais extraídos de *corpora* de textos científicos. Para a construção de representações semânticas para os termos selecionados, incluimos informações segundo (i) a estrutura *Qualia* da Teoria do Léxico Gerativo de Pustejovsky (1995) e (ii) a estrutura ontológica de domínios específicos. Em (i), os termos foram descritos a partir dos papéis Constitutivo, Formal, Télico e Agentivo para que fosse possível resgatar e capturar de forma unívoca a dimensão do seu conceito. Nesse sentido, as unidades semânticas foram organizadas em termos de relações semântico-lexicais, tais como: hiperonímia/hiponímia; meronímia/holonímia, sinonímia, antonímia. Em (ii), os termos foram classificados segundo uma organização ontológica do tipo hereditária em que os itens classificados em subclasses herdam as características da classe maior e assim sucessivamente.

Segundo Zavaglia (2002):

> O processamento de uma língua natural bem como a compreensão do fenômeno da linguagem natural são temas de maior interesse, nos dias de hoje, para ciências como a Inteligência Artificial, a Lingüística Computacional, a Tradução Automática, entre outras. A introdução do computador no quotidiano das pessoas afetou a sua maneira de enxergar o mundo, transformando-as em seres mais conscientes e exigentes não somente com o mundo a seu redor mas também com o mundo além-mar, sem fronteiras, atingível e acessível, em segundos, por meio da Internet, a rede mundial de computadores. (ZAVAGLIA, 2002, p. 17)

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 e a 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.

A tese que se defende neste trabalho 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[1]". E ainda: "Modelos de domínio são usados, principalmente, pela comunidade de reuso de software (ARANGO, 1994). Ontologias são aplicadas, principalmente, pela comunidade de inteligência artificial na perspectiva de modelagem de conhecimento".

O delineamento do esquema arbóreo do subdomínio da Ecologia e o agrupamento de seus itens lexicais nas diversas subclasses da ontologia objetiva servir à produção de bases computacionais para SPLN também em outras línguas, além do português, fato esse enriquecedor para um trabalho que pretende sanar uma das centenas de lacunas existentes no mercado lexicográfico brasileiro, i.e., o de obras especializadas, máxime se bilíngües e/ou multilíngües.

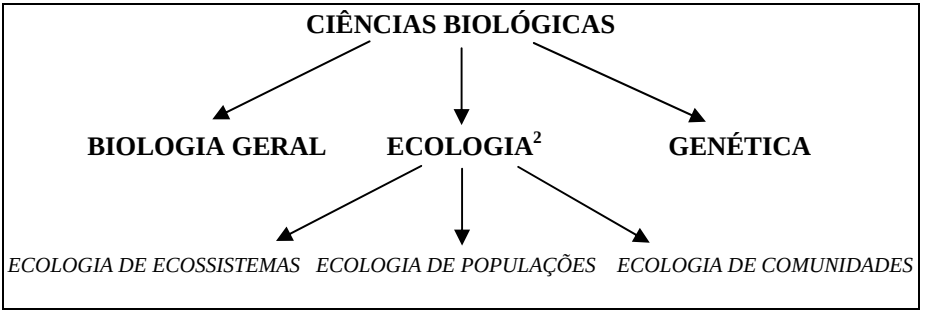
Cumpre ressaltar que a partir do momento que o uso de Ontologias apregoa um entendimento e uma base de informações comuns a uma certa comunidade, uma das tarefas mais densas, e que requer um olhar

[1] "Ambas estão em busca do armazenamento de informações sobre um domínio que possam ser reutilizadas, ou que auxiliem na tarefa de reutilização durante o desenvolvimento de uma nova aplicação daquele domínio. A principal diferença entre eles é que a ontologia não assume a pré-existência de nenhum sistema no domínio a ser modelado. Deste forma, o nível de abstração das ontologias é mais alto que o dos modelos de domínio" (MANGAN *et al*., s.d.)

especial sobre ela, é, justamente, o momento de se traçar e de se conceituar as classes hierárquicas que farão parte dessa estrutura ontológica: como classificar itens lexicais em superclasses, subclasses, instâncias, *slots*? E ainda: de onde resgatar as informações necessárias para o delineamento desse mapa ontológico: como proceder, como reagir, como agir? Partindo-se do pressuposto de que uma ontologia é uma fotografia registrada de um dado instante de uma certa realidade por um certo alguém, o ontológo deverá se conscientizar de que a modelagem das estruturas traçadas terá de se adaptar aos limites impostos pela ferramenta computacional que será utilizada para a modelagem dos dados. Há que se relatar ainda, que existe na tarefa de delinear, demarcar, limitar, caracterizar, definir, conceituar conhecimento, seja ele de qualquer universo de discurso, muito mais arte do que ciência (MANGAN, s.d.), e o "colocar a mão na massa" ainda é o ponto crucial de qualquer pesquisa que requer modelagem de conhecimento, uma vez que é feito com resultados positivos essencialmente por humanos, embora tentativas automáticas estejam sendo levadas a cabo..

1 Estruturação Ontológica

A partir da Ontologia proposta por Zavaglia (2002), para o domínio das Ciências Biológicas, para o qual determinou-se o subdomínio da Ecologia, traçamos o perfil ontológico desse subdomínio, ou seja, individualizamos as classes e subclasses que o compõem. Temos, como ponto de partida, a seguinte estruturação arbórea:



Segundo Almeida (2000):

[2] Em Zavaglia (2002), os subdomínios referentes à Ecologia são Ecologia Teórica, Ecologia de Ecossistemas e Ecologia Aplicada, segundo a classificação do CNPq.

> A metodologia do trabalho terminológico exige o cumprimento de uma seqüência de etapas, e todas elas devem ser desenvolvidas com o assessoramento de especialistas da área-objeto, uma vez que o terminólogo, ao iniciar um projeto com fins terminográficos, não tem o necessário domínio e conhecimento da área escolhida para o trabalho, não podendo, portanto, prescindir da ajuda do(s) especialista(s). Esse conhecimento requerido vai-se (*sic*) construindo à medida que o terminólogo vai-se (*sic*) comprometendo com o projeto, mas, ainda, assim, a presença do especialista é necessária, como podemos perceber na seqüência abaixo descrita. (ALMEIDA, 2000, p. 51-52)

O trabalho do ontolólogo, embora tenha características próprias e limítrofes, mescla-se, e por isso muitas vezes confunde-se, com o trabalho do terminólogo. Tal semelhança se explica pelo fato de os dois pesquisadores trabalharem diretamente com campos conceituais ou nocionais e com listas de unidades lexicais superordenadas em classes. Entretanto, no trabalho do pesquisador em ontologia, perfilam os conceitos de hereditariedade semântica e herança múltipla que unidades lexicais adquirem e herdam dependendo da sua ordenação estrutural em um esquema arbóreo, ao passo que em terminologia fala-se em uma estruturação hierárquica de dados.

Almeida (2000) descreve uma série de etapas a serem seguidas para o trabalho do terminólogo, entre elas: (a) a escolha dos setores envolvidos no projeto, (b) a seleção das fontes (documentos, revistas, obras especializadas, entrevistas) e (c) a seleção dos termos.

Já para a elaboração de ontologias, as etapas a serem seguidas são: (a) o delineamento geral do domínio a ser tratado, (b) a enumeração das unidades lexicais, (c) a definição das classes de forma hierárquica, (d) a definição das propriedades das classes, (e) a definição dos atributos das propriedades e (f) a criação de instâncias.

Vejamos abaixo um exemplo da *Ontologia de Ecossistemas* em campos nocionais estruturados conceitualmente:

1. Ecologia de Ecossistemas
1.1- Ecossistema
1.1.1- Ecossistema natural
1.1.1.1- Ecossistema terrestre
1.1.1.1.1-Florestas
1.1.1.1.1.1- Boreal de Coníferas

1.1.1.1.1.2- Temperadas Deciduas
1.1.1.1.1.3- Tropical
1.1.1.1.1.3.1-Semiperinifólia
1.1.1.1.1.3.2- Úmida Perenifólia
1.1.1.1.2- Tundra
1.1.1.1.3- Campos
1.1.1.1.3.1- Temperados de Gramíneas
1.1.1.1.3.2- Cerrado
1.1.1.1.3.2.1- Campo Limpo
1.1.1.1.3.2.2- Campo Sujo
1.1.1.1.3.2.3- Cerrado Strictu Sensu
1.1.1.1.3.2.4- Cerradão
1.1.1.1.4- Desertos
1.1.1.1.5- Mangue
1.1.1.1.6- Restinga
1.1.1.2- Ecossistema Aquático
1.1.1.2.1- Marinho
1.1.1.2.1.1- Oceano Aberto
1.1.1.2.1.2- Plataforma Continental
1.1.1.2.2- Dulcícola
1.1.1.2.2.1- Lêntico
1.1.1.2.2.2- Lótico
1.1.2- Ecossistema Artificial
1.1.2.1- Reflorestamento
1.1.2.1.1- Nativos
1.1.2.1.2- Introduzidos
1.1.2.2- Represas
1.2- Energia no Ecossistema
1.2.1- Produtividade
1.2.2- Ciclos Biogeoquímicos
1.2.3- Fluxo de Energia

Cada item lexical foi etiquetado a partir dos seguintes preceitos: (i) organização ontológica do tipo hereditária; (ii) presença de informações morfossintáticas e definicionais; (iii) disposição e ordenação de cada um dos itens lexicais a partir do *genus* e suas relações semânticas com o *definiendum* e (iv) classificação dos itens lexicais segundo as relações semânticas da Estrutura *Qualia* e da *Qualia* Ampliada descritas em Lenci (1999) para o português a partir dos papéis Formal, Constitutivo, Télico e Agentivo.

2 A voz dos teóricos

2.1 O Léxico Gerativo

Pustejovsky (1995) caracteriza o léxico gerativo como sendo um sistema computacional que envolve, no mínimo, quatro níveis de representação: (i) a Estrutura Argumental (*Argument Structure*) em que se tem a especificação do número e do tipo de argumentos lógicos e como eles são realizados sintaticamente; (ii) a Estrutura de Evento (*Event Structure*) que define o tipo de evento de um item lexical em uma frase, tais como estado, processo e transição; (iii) a Estrutura *Qualia* (*Qualia Structure*) que inclui os modos de explicação compostos pelos papéis formal, constitutivo, télico e agentivo; e a (iv) Estrutura de Herança Lexical (*Lexical Inheritance Structure*) em que se tem a identificação de como uma estrutura lexical se relaciona com outras estruturas e a sua contribuição para a organização global do léxico. Dessa maneira, a semântica de um item lexical para o autor é definida como uma estrutura composta por quatro elementos: $\alpha = <$ **A , e , λ , Ý** $>$[3], em que: α = **é o item lexical; A** é a estrutura argumental; **e** a especificação do tipo de evento; **λ** estabelece o vínculo desses dois parâmetros na Estrutura *Qualia*; e **Ý** determina qual informação é hereditária na estrutura lexical global.

Os argumentos na Estrutura Argumental são distintos em quatro tipos para os itens lexicais: (1) Argumentos Verdadeiros (*True Arguments*), ou seja, parâmetros sintaticamente realizados do item lexical; (2) Argumentos *Default* (*Default Arguments*), i.e., parâmetros que participam na expressão lógica na Estrutura *Qualia*, mas que não são necessariamente expressos sintaticamente; (3) Argumentos Sombra (*Shadow Arguments*), i.e., parâmetros que são incorporados semanticamente ao item lexical. Eles podem ser expressos somente por operações de subtipagem (*subtyping*) ou especificação discursiva e (4) Adjuntos Verdadeiros (*True Adjuncts*), ou seja, parâmetros que modificam a expressão lógica, mas que são parte da interpretação situacional e não são vinculados a nenhuma representação semântica de um item lexical particular. Incluem expressões adjuntas de modificação temporal ou espacial.

[3] Adaptação nossa da simbologia da teoria de Pustejovsky.

Por sua vez, na Estrutura de Eventos, Pustejovsky caracteriza o tipo de evento de um item lexical. Os eventos definidos por ele podem conter uma estrutura de subeventos do tipo: Temporalmente ordenados, Completamente simultâneos, Basicamente simultâneos, em que o início de um dos subeventos é anterior ao início do outro.

2.1.2 A Estrutura *Qualia*

Pustejovsky chamou de Estrutura *Qualia* a representação que dá força relacional ao item lexical. O léxico gerativo 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*[4]):

- Constitutivo ou Partes Constituintes (*Constitutive*), i.e., aquele que exprime a relação entre um objeto e suas partes constituintes;
- Formal (*Formal*), ou seja, aquele que identifica o objeto em um domínio mais amplo;
- Télico (*Telic*), aquele que expressa o objetivo/escopo e a função do objeto;
- Agentivo (*Agentive*), i.e., aquele que considera fatores envolvidos na origem do objeto.

Há dois pontos importantes que devem ser considerados com respeito à *Qualia*: (1) toda categoria expressa uma estrutura *Qualia*; (2) nem todos os itens lexicais carregam consigo valores para cada papel *Qualia*. O primeiro item é importante para se entender como um léxico gerativo sustenta uma representação semântica uniforme composicionalmente de todos os elementos de uma frase. Já o segundo é aplicável e específico para classes semânticas particulares.

Para Pustejovsky, *Qualia*, em todos os sentidos, é como um conjunto de propriedades de eventos associado ao item lexical que melhor explica o que aquela palavra significa. Por exemplo, para que se entenda o que itens lexicais como *cookie* (biscoito) e *beer* (cerveja) significam, deve-se reconhecer que eles são um tipo de gênero alimentício e um tipo de bebida, respectivamente. Enquanto *cookie* é um termo que descreve um

[4] Entende-se que, para Pustejovsky, *Qualia* é sinônimo de "significado de uma palavra" (*word's meaning*).

tipo particular de objeto no mundo, a expressão "gênero alimentício" denota uma referência funcional do que se "faz com" alguma coisa, i.e., como se usa essa mesma coisa. Neste caso, o termo é definido, em parte, pelo fato de que alimento é algo que se come. Para *beer*, são feitas observações similares.

O *Telic quale* ou "significado télico" para o nome *food* (comida/alimento) codifica o aspecto funcional do significado, representado como [TELIC[5] = comer]. Da mesma forma, a distinção entre nomes semanticamente relacionados como *novel* (romance) e *dictionary* (dicionário) provém daquilo que "se faz com" esses objetos, que é diferente. Assim, embora esses dois objetos sejam "livros", no sentido geral, o uso de cada um deles difere: enquanto um "romance" serve para "ler", um "dicionário" serve para "consultar". Conseqüentemente, os valores *Qualia* codificam a informação funcional para "romance" e "dicionário" de forma distinta: [TELIC = ler] para "romance" e [TELIC = consultar] para "dicionário". Obviamente, a distinção entre esses dois objetos não se faz somente por meio desses diferentes papéis na estrutura télica de *Qualia*. O tipo de estrutura textual de cada um deles é recuperado pelo papel "constitutivo" da Estrutura *Qualia*. Enquanto que "romance" é caracterizado como uma narrativa ou história, "dicionário" é definido como uma lista de palavras. Assim, temos a representação: [CONST[6] = narrativa] para "romance" e [CONST = lista de palavras] para "dicionário". Esses dois objetos são caracterizados de forma idêntica no papel formal: [FORMAL[7] = livro] para "romance" e [FORMAL = livro] para "dicionário". Ao contrário, diferem, ainda, no papel agentivo da Estrutura *Qualia*, a saber: em como é realizada a "existência" deles, ou seja, enquanto que um "romance" é escrito, um "dicionário" é compilado, ou seja, organizado: [AGENT[8] = escrito] para "romance" e [AGENT = organizado] para "dicionário".

Dada a suposição de que múltiplas dimensões do significado são necessárias para começar a caracterizar unidades lexicais em um nível semântico, a Estrutura *Qualia* tem sido utilizada[9] como um dos princípios

[5] Papel Télico (*Telic*).
[6] Papel Constitutivo (*Constitutive*).
[7] Papel Formal (*Formal*).
[8] Papel Agentivo (*Agentive*).

[9] Um exemplo da utilização da Estrutura *Qualia* como representação do significado pode ser visto em Hathout (1996) onde estão as especificações da elaboração de uma base de conhecimento lexical para o domínio da química, na qual as informações específicas das entidades desse domínio correspondem ao papel Formal da Estrutura *Qualia*.

cruciais de organização para a representação e interpretação do significado lexical de uma frase em sistemas computacionais de complexidade variada. De fato, ela é capaz de suprir o vocabulário básico para expressar aspectos diferentes do significado lexical (*word-meaning*).

Nesse trabalho, a Estrutura *Qualia* foi utilizada para caracterizar, por meio de relações semânticas, os termos ontológicos inseridos na estrutura arbórea da ontologia da Ecologia de Populações utilizada no protótipo computacional que será descrito mais adiante.

2.2 Ontologias

Segundo Ortiz (2000, p. 1-2), a palavra "ontologia" tem gerado diversas controvérsias no campo da Inteligência Artificial (doravante IA), uma vez que possui uma vasta tradição na Filosofia[10] onde é utilizada para referenciar temas relacionados à existência dos seres. 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 IA que há várias décadas buscam uma ontologia que ofereça flexibilidade suficiente para dar conta de representar o conhecimento complexo registrado na mente humana. (cf. ZAVAGLIA, 2002) 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: (i) Ontologia como uma disciplina filosófica; (ii) Ontologia como um sistema conceitual informal; (iii) Ontologia como um cálculo da semântica formal; (iv) Ontologia como uma especificação de uma conceitualização caracterizada por propriedades formais específicas e/ou somente por propósitos específicos; (v) Ontologia como o vocabulário usado por uma teoria lógica; (vi) Ontologia como uma especificação de uma teoria lógica (*meta-level*). Nesse trabalho, a interpretação de Ontologia que nos serve é a de número (4).

[10] Em Ferreira (1999), ontologia significa: "Parte da filosofia que trata do ser enquanto ser, i. e., do ser concebido como tendo uma natureza comum que é inerente a todos e a cada um dos seres: 'Com Kant o universo é uma dúvida: com Locke, é dúvida o nosso espírito: e num destes abismos vêm precipitar-se todas as ontologias' (Alexandre Herculano, Lendas e Narrativas, II, p. 107.)"

Para Gruber (1993), ontologias compartilham e reutilizam o conhecimento de mundo. Com efeito, segundo o autor: “o termo ontologia significa uma especificação de conceitos, isto é, uma ontologia é uma descrição formal dos conceitos e das relações existentes entre estes em um determinado domínio” (*apud* BRAGA *et al*., 2002). Ainda segundo Gruber: “uma ontologia é uma especificação explícita de uma conceitualização” (1993, p. 1). Com efeito, para sistemas computacionais aquilo que existe é somente aquilo que pode ser representado em um formalismo declarativo, para o qual, o conjunto de objetos que pode ser representado é chamado de universo do discurso. Esse conjunto de objetos e as relações existentes entre eles refletem-se no vocabulário de representação com o qual o programa baseado em conhecimento pretende representar o conhecimento em si. Formalmente, segundo o autor, uma ontologia é a declaração de uma teoria lógica (cf. GRUBER, 1993, p. 1-2). Dessa maneira, 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 (no caso de ontologias específicas, por exemplo) e as relações entre elas devem ser conceitualizadas. Tais conceitos nada mais são do que uma visão simplificada e resumida do mundo.

Ressaltamos que, atualmente, no campo do PLN, principalmente em Sistemas de Bases de Conhecimento Lexical, é consensual que a inclusão desse tipo de repositório semântico, i.e., do tipo ontológico para a representação do significado, é essencial. Existe a necessidade de se oferecer de forma estruturada e organizada um léxico comum utilizado em conformidade por um determinado grupo de usuários. 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, onde a sua aplicação pode ser mais eficaz por tratar, justamente, de conjuntos léxicos de número finito. Em uma Base de Conhecimento Lexical, por exemplo, o uso de uma ontologia pode servir como recurso de apoio à informação contida no repositório lexical dessa base para ser possível resgatar o significado de um item léxico de forma unívoca. De fato, os recursos lingüístico-classificatórios que a utilização de uma ontologia pode oferecer para um lingüista e/ou lexicógrafo servem para que ele possa dar conta de individualizar univocamente, dentre os diversos significados ou diversas acepções atribuíveis a um mesmo item lexical, o significado pertinente no interior do feixe de sentidos polissêmicos que a palavra comporta,

neutralizando, dessa maneira, a polissemia própria a esse mesmo item lexical.

2.2.1 A utilidade de ontologias

Uma primeira utilização de ontologias seria a representação de informações que possua um entendimento semântico comum de situações variadas do mundo real. Há que se considerar, porém, que a descrição detalhada de informações gerais do mundo ao nosso redor não deve ser considerada uma tarefa banal, ao contrário. Com isso, tem-se preferido representar o conhecimento de domínios específicos, no qual apenas uma parte de mundo é representada e formalizada, e no qual veiculam informações restritas, porém com maior riqueza de detalhes e consenso entre a comunidade que a compõe. Com efeito, almeja-se que os conceitos-chave do domínio restrito sejam definidos senão diretamente por especialistas na área, com a ajuda e a consultoria deles. Em consonância, Vasconcelos (2003, p. 16) diz que "As ontologias se propõem a capturar domínios de conhecimento de forma genérica, para fornecer um entendimento semântico desses domínios que poderá ser utilizado e compartilhado por diversas comunidades e aplicações".

Na Web, o uso de ontologias pode fornecer uma base de informações comum, bem como padronizada, englobando conceitos-chave que possam ser utilizados por serviços requisitados para cada situação particular. Em comércio eletrônico, por exemplo, o conjunto de informações oferecido pela ontologia poderá ser utilizado para unificar e integrar definições de produtos que estejam sendo oferecidos pelos mais variados pontos de venda com um formato padrão e único. Além disso, as ontologias podem ser utilizadas por Motores de Busca existentes na *Web*, uma vez que podem servir como um esquema conceitual de um determinado domínio, já que serve de suporte semântico às buscas ou consultas realizadas. Com efeito, quando as máquinas de busca fazem uso de ontologias para realizarem consultas por palavras-chave, por exemplo, em suas respostas-saída elas podem oferecer além das páginas que contêm a palavra-chave requisitada, outras páginas que contenham informações vinculadas ao conceito das palavras-chave, tais como sinônimos, antônimos, hiperônimos, hipônimos.

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 "populaçao" 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 pós-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.

Em consonância, Melcop *et al* (2002) dizem que existe uma dificuldade visível e preponderante para se encontrar a informação almejada na *Web*. E acrescentam:

> A maioria dos sistemas de indexação e busca na *Web* utiliza técnicas baseadas em palavras-chave. Tais sistemas retornam uma grande quantidade de apontadores (*links*) para páginas de pouco ou nenhum interesse do usuário. A baixa precisão desses sistemas deve-se, entre outros fatores, à sua pequena capacidade de interpretar o conteúdo dos textos, e, em conseqüência, de contextualizar a seleção de documentos de acordo com o interesse do usuário. (MELCOP, 2002, p.1)

Nesse mesmo sentido, Rigo; Vieira (2002) apontam para a deficiência existente na Internet para busca de informações, uma vez que as ferramentas de busca empregadas aliam sua pesquisa apenas nos termos indicados pelo usuário. Essa técnica faz com que o resultado das buscas de um usuário por meio de um único termo lhe reporte, além de *links* que podem possuir informações acerca do conteúdo desejado, uma lista de documentos recuperados cujo termo está presente mas não representa de forma alguma o significado da busca inicial. Isso se dá porque um mesmo termo ou unidade lexical pode possuir diversos significados em diferentes domínios ou num mesmo domínio, dada a polissemia inerente existente na maioria dos itens lexicais de uma língua natural. Além disso, esses autores afirmam que “o formato utilizado na descrição da grande maioria de documentos disponíveis na Internet (HTML) não é adequado à identificação de seu conteúdo” (RIGO; VIEIRA, 2002, p. 1). Acrescentam ainda que para sanar essas deficiências e propiciar ao usuário uma identificação correta do conteúdo de um documento aos seus interesses de busca de informações, são

necessários a utilização de ontologias adequadas e o emprego de formatos de descrição baseados em XML[11]. A partir do momento que um termo encontra-se definido em uma ontologia, a sua associação com um outro termo utilizado para o mesmo conceito é possível, desde que ele também esteja descrito na ontologia. Ademais:

> O uso de ontologias permite também a associação de possíveis regras de inferência relacionadas a termos descritos, de modo a possibilitar também uma outra forma de busca de informações, agora de modo automático, através de agentes de *software*, onde serão levados em conta possíveis relacionamentos entre as informações encontradas nos documentos pesquisados. (RIGO; VIEIRA, 2002, p. 1)

Como suporte à interoperabilidade de informações e de dados, as ontologias podem ser utilizadas como aplicação para banco de dados e recuperação de informação, uma vez que servem de modelos conceituais globais entre entidades e relacionamentos. De fato, em motores de busca, as ontologias servem como grandes esquemas conceituais que suportam o caráter semânticos das consultas, como dissemos anteriormente.

3 O *corpus* da OntoEco

No Brasil, a Lingüística de *Corpus* (LC) vivenciou um crescimento vertiginoso nos últimos anos, propiciando o surgimento de novas propostas de estudos lingüísticos e o interesse crescente em elaborar bases textuais de tipologia variada. Segundo Berber Sardinha (2000), a tecnologia alia-se à LC, na medida em que "permite não somente o armazenamento de *corpora*, mas também a sua exploração" (p. 12). Para a elaboração de um *corpus*, alguns critérios são fundamentais, dentre eles: representatividade, extensão e tipologia. No que diz respeito à tipologia, Berber Sardinha (2000) elenca os seguintes tipos de *corpus* no que diz respeito a: (i) modo: falado ou escrito; (ii) tempo: sincrônico, diacrônico, contemporâneo, histórico; (iii) seleção: de amostragem, monitor, dinâmico ou orgânico, estático, equilibrado; (iv) conteúdo: especializado, regional ou dialetal, multilíngüe; (v) autoria: de aprendiz; de língua nativa; (vi) disposição interna: paralelo ou alinhado; (vii) finalidade: de estudo, de referência, de treinamento ou teste. Em relação à

[11] e**X**tensibile **M**arkup **L**anguage.

representatividade e à extensão, parte-se da premissa de que todo *corpus* possui uma função representativa e para ser representativo o conjunto de textos deve ser o maior possível, ou seja, deve possuir uma dada extensão de um número determinado de palavras e de textos. Segundo Sinclair (1991) "A principal maneira, ou 'salvaguarda', pela qual se pode garantir maior representatividade é através do aumento da extensão do *corpus*. Um *corpus* maior é em geral mais representativo do que um menor devido ao fato de conter mais instâncias de traços lingüísticos raros" (BERBER SARDINHA, 2000).

Nesse sentido, Biderman (2001) revela:

> No desenho do *corpus* é necessário que haja uma proporção equilibrada dos diferentes tipos de textos e/ou de temas nele incluídos. É também importante que o corpus seja representativo dos diferentes gêneros e variedades dos usos lingüísticos, ou seja, impõe-se a representatividade dos diferentes níveis de linguagem para assegurar a inclusão de todos os aspectos do idioma. Só assim o corpus pode representar, em miniatura, o universo multifacetado da língua. Quando se projeta um corpus visa-se extrair de sua observação generalizações sobre a língua. Portanto, não se pode atribuir um peso excessivo a um gênero ou a outro (BIDERMAN, 2001, p. 79).

É importante ressaltar que a produção de *corpus* acentuou-se a partir do momento que os lingüistas "puros" e os lingüistas computacionais comprovaram o mérito de realizarem consultas em um *corpus* como filtro mediador de suas validações, hipóteses e evidências para pesquisas lingüísticas de caráter variado. Com isso, deu-se início tanto à produção maciça de grandes repertórios textuais nas mais variadas línguas quanto de ferramentas e programas para a sua gestão.

Nesse trabalho, elaboramos uma base de textos especiais concernente ao subdomínio da Ecologia para o português do Brasil. Para a sua formação, utilizamo-nos de livros especializados sobre Ecologia com a finalidade de selecionarmos os textos que fariam parte da sua base lingüística. Os textos selecionados, após intensa pesquisa realizada na biblioteca da UNESP/IBILCE e na Internet, foram selecionados porque demonstraram ser os mais adequados ao tipo de pesquisa em questão, além de serem utilizados nos cursos de Biologia da UNESP de São José do Rio Preto e, conseqüentemente, terem sido indicados pelo consultor do projeto, professor de Ecologia do IBILCE. São eles: RICKLEFS,

Robert E. *A Economia da natureza*. Rio de Janeiro Guanabara/Koogan, 1993 – 3. ed. p. 85 a 130; 199 a 258 e 333 a 400 e ODUM, Eugene P. *Ecologia*. Rio de Janeiro: Guanabara, 1983 – p. 9 a 54; 187 a 231 e 233 a 281.

Além desses *subcorpus*, utilizamo-nos de outro formado com textos técnico-científicos do *corpus* do Lácio-Web, disponibilizado pelo NILC, assim caracterizado:
Domínio maior: Ciências Biológicas e Áreas específicas: Recursos Florestais e Engenharia Florestal; Zootecnia; Biologia Geral; Ecologia e Botânica.

Atualmente, o *corpus* da OntoEco, o *CorpusEco*, conta com as seguintes ocorrências em palavras:

Subcorpus: Ecologia de Comunidades -	57.501	
Subcorpus: Ecologia de Ecossistemas -	49.271	
Subcorpus: Ecologia de Populações -	54.645	
Subtotal -		**161.417**
Subcorpus: Recursos Florestais e Engenharia Florestal -	2.579	
Subcorpus: Zootecnia		2.572
Subcorpus: Biologia Geral		27.587
Subcorpus: Ecologia		35.529
Subcorpus: Botânica		16.959
Subcorpus: Zoologia		14.278
Subtotal -		99.504
TOTAL do *CorpusEco*:	**260.921**	

A começar pelos parâmetros descritos acima, almejamos que o *corpus* da Ecologia seja representativo no domínio de *corpora* especiais, a partir do momento que pretendemos alimentá-lo sistematicamente com novos textos. Suas características iniciais apontam para os seguintes critérios de elaboração: foi composto por textos escritos, representando o período de tempo corrente; o seu conteúdo é o especializado, em uma língua natural, no caso, o português; sua autoria é a de aprendizes (língua de chegada), dado que esses textos são traduções do inglês (como a maioria dos textos em Ecologia); e a sua finalidade é a de referência para pesquisas. Esses textos foram escaneados e digitalizados com a finalidade de serem armazenados em uma ferramenta computacional, no caso, o *Folio Views* 3.1.

4 A extração das unidades léxico-ontológicas

Primeiramente, gostaríamos de diferençar o termo que cunhamos, a saber: *Termo Ontológico* (TO) ou *Unidade Léxico-Ontológica* (ULO) em detrimento a *Termo*, este último comumente empregado em Terminologia e Lexicologia. Segundo Alves (1999, p. 70), "as lexias científicas e técnicas constituem os *termos*, que integram conjuntos que obedecem a uma conceitualização rigorosa dos significados e a uma organização normalizada, as terminologias". Estamos, nesse caso, trabalhando com línguas de especialidade (também chamadas de tecnoletos) uma vez que compreendem um domínio de conhecimento especializado que se particulariza dentro da língua geral ou comum.

No domínio da Inteligência Artificial (IA), atuamos em um de seus mais recentes subdomínios, i. e. a Engenharia Ontológica (EO). Uma vez que as linhas limítrofes entre estruturação terminológica e estruturação ontológica ainda se confundem, achamos necessário delimitar também, a exemplo da IA para a EO, o domínio da Terminologia para o subdomínio da Terminologia Ontológica (TEO). De fato, em TEO podemos tratar tanto de domínios gerais quanto de especiais, enquanto os seus preceitos são as especificações de classes (que possuam conceitos), de seus objetos e de suas relações em um dado domínio (genérico ou específico). Dito isso, um TO possui necessariamente um conceito que será herdado por outras unidades léxico-ontológicas se a ele forem vinculados, uma vez que será disposto em uma estruturação ontológica de caráter hereditária.

A problemática da delimitação lexical nas línguas de especialidade (incluindo a Terminologia e a Terminologia Ontológica) subsiste, segundo Alves (1999) porque:

> A delimitação dos termos-sintagmas de uma área de especialidade é problemática por causa da dificuldade no estabelecimento de fronteiras entre um segmento frásico, sintagma livre, e um segmento frásico lexicalizado, que se tornou (ou está se tornando) um novo termo. Na verdade, como ocorre a passsagem de um segmento frásico a uma estrutura lexicalizada? (ALVES, 1999, p. 73)

Essa mesma autora cita diversos outros teóricos da terminologia que abordam essa problemática, tais como Kocourek, 1991; Cabré, 1993; Rey, 1995; Arntz; Picht, 1995.

A principal utilidade da elaboração do *Corpus*Eco foi, justamente, servir de base textual para a extração dos TOs vinculados a OntoEco.

Para o protótipo de nossa ontologia, detivemo-nos na caracterização da subontologia do Ecossistema de Populações e assim, a extração dos termos que será relatada a seguir trata somente da extração de unidades desse subdomínio.

Essa extração foi feita, primeiramente, de forma manual, utilizando-se o critério semântico no processo de extração. De fato, utilizamos a metodologia da onomasiologia[12], 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.

Como resultados dessa extração manual temos:

Adaptação; Agregação; Amostra; Área; Área amostral; Área de Distribuição; Atributo; Ciclo Vital; Classe etária; Colônia; Competição; Densidade; Densidade Populacional; Dinâmica da População; Dispersão; Distribuição; Distribuição Agregada; Distribuição etária; Distribuição Geográfica; Distribuição Homogênea; Ecologia Animal; Ecologia Vegetal; Emigração; Espécie; Espécime; Estação de Reprodução; Estrutura de População; Estrutura de População; Estrutura etária; Expectativa Adicional de Vida; Fator denso-dependente; Fator Endógeno; Flutuação Populacional; Fronteira Geográfica; Fronteira Natural; Genética de Populações; Genótipo; Habitat; Hiperdispersão; Imigração; Índice de Lincoln; Indivíduo; Interação; Maninhos de Serpentina; Método de marcação e recaptura; Método do Quadrado; Migração; Movimento Aleatório; Movimento Centrífugo; Mutação; Oscilação Populacional; Pirâmide de Idades; População; Probabilidade de Sobrevivência; Seleção Natural; Sistema de Acasalamento; Sobrevivência; Subárea; Subpopulação; Supervivência; Tabela de Vida; Tabela de Vida da Coorte; Tabela de Vida de Tempo Específico; Tabela de Vida Estática; Tabela de Vida Horizontal; Tabela de Vida Vertical; Tamanho de Vizinhança; Taxa de Fecundidade; Taxa de Fertilidade; Taxa de Mortalidade; Taxa de Morte; Taxa de Nascimento; Taxa de Predação; Variação Populacional.

Quadro 1: Extração Manual de Termos Ontológicos em Ecologia de Populações. Resultados: 69

Existem algumas características morfológicas (que chamaremos aqui de padrões morfológicos) presentes nas línguas de especialidade. Segundo Alves (1999), a mais freqüente delas é a composta por (i) <substantivo genérico acompanhado por um adjetivo determinante - [S+A]>. Com efeito, temos no domíno da Ecologia de Populações TOs do tipo: <área amostral>, <ciclo vital>, <seleção natural>, entre outros, com bastante freqüência. Outro padrão bastante comum é aquele representado por (ii) um <substantivo determinado acompanhado de uma

[12] A onomasiologia opõe-se à semasiologia, i.e. metolodologia de caráter semântico que parte do significante para o significado.

preposição e de outro substantivo [S+Prep+S]>: <tamanho de vizinhança>; <taxa de nascimento>; <taxa de morte>. Além desses, um padrão também identificado foi aquele composto por (iii) <substantivo determinado acompanhado de uma preposição e de um outro substantivo determinado acompanhado de um adjetivo determinante [S+Prep+S+A]>: <tabela de vida vertical>; <tabela de vida estática>; <tabela de vida horizontal>. Outros padrões morfológicos nos chamam a atenção neste domínio e são eles: (iv) <substantivo determinante acompanhado de preposição seguido de um outro substantivo acompanhado de preposição e substantivo [S+Prep+S+Prep+S]>: <tabela de vida da Coorte>; (v) <substantivo determinante acompanhado de preposição seguido de um outro substantivo acompanhado de preposição, substantivo e adjetivo determinante [S+Prep+S+Prep+S+A]>: <tabela de vida de tempo específico>; (vi) <substantivo determinante acompanhado de preposição seguido de um outro substantivo mais uma conjunção aditiva acompanhado de outro substantivo [S+Prep+S+Conj+S]>: <método de marcação e recaptura>; (vii) <substantivo acompanhado de adjetivo determinante seguido de preposição mais um outro substantivo [S+A+Prep+S]>: <expectativa adicional de vida>; (viii) <substantivo único [S]>: <supervivência>; <subpopulação>; <área>.

Sobre essas formações, Alves (1999) pontua:

> Essas formações, que recebem variadas denominações (unidades sintagmáticas segundo Guilbert (1975a, p. 249), termos-sintagmas segundo Kocourek (1991, p. 135), entre outras) resultam, na verdade, da lexicalização de segmentos frásicos. Constituem uma conseqüência do caráter onomasiológico da disciplina terminológica, em que o conceito usualmente precede a criação do termo correspondente. Desse modo, a explanação de um conceito, expressa por segmentos frásicos, muitas vezes condiciona a lexicalização desses segmentos, que se tornam termos. Esses termos-sintagmas são em geral transparentes, facilmente interpretáveis por causa da junção de seus elementos integrantes. (ALVES, 1999, p. 72)

Como dito anteriormente, em Engenharia Ontológia (EO), adotamos a terminologia de *Termos Ontológicos* (TOs) ou *Unidades Léxico-Ontológicas* (ULOs), uma vez que esses segmentos sintagmáticos ontológicos, além de expressarem um conceito, têm o caráter de hereditariedade conceitual que é trasmitida a todos os TOs ou ULOs que

a eles forem vinculados. Convém ressaltar que isso ocorre a partir do momento que em EO, TOs ou ULOs se trasformam em CLASSES e SUBCLASSES em uma cadeia hierárquica por meio da relação ISA (“is_a”, ou seja, “é_um”). Por sua vez, esses TOs possuem instâncias que são vinculadas a eles por possuírem, exatamente, todas as características semânticas herdadas pelo seu “pai”. Nesse sentido, instâncias são os “filhos” que cada “pai” possui, ou “folhas” e “galhos” em uma perspectiva arbórea, respectivamente. Vejamos um exemplo concreto: o TO “Área” é uma SUBCLASSE da CLASSE “CLASSES” que possui instâncias do tipo: “Estreito de Bering”; “Minas Gerais”; “São Paulo”; “Fortaleza”; “Roma”. Numa perspectiva terminológica, todas essas instâncias seriam consideradas termos, cada uma com suas características próprias. Já em EO, essas mesmas possuem o conceito de “Superfície ocupada por uma comunidade ou táxon” em que novas particularidades, por meio de propriedades, são inseridas na medida da necessidade da instância.

Cumpre ressaltar, neste ponto, a diferença de enfoque que perfila entre *Terminologia* e *Engenharia Ontológica*. A primeira busca (a) a sistematização, estudo e descrição de conjuntos vocabulares da especialidade; (b) a seleção de fontes (documentos, obras, revistas especializadas) para a extração de termos; (c) a seleção de termos; (d) o mapa ou estrutura conceitual ou árvore de domínio como auxiliar na distribuição e “encaixe” de termos para uma uniformidade conceitual da área-objeto e as relações existentes entre eles no processo de elaboração do produto final de conjuntos vocabulares, como objeto de estudo e finalidade e (e) a produção de glossários e vocabulários especializados. A segunda, por sua vez, visa (f) ao delineamento do domínio a ser focalizado (genérico ou específico); (g) à seleção de fontes (documentos, obras, revistas especializadas ou não) para a extração de classes, subclasses e termos ontológicos; (h) à definição de conceitos de classes como taxonomia hierárquica; (i) à definição das propriedades das classes, dos atributos das propriedades (chamados de *slots*) e de instâncias; (j) à estruturação arbórea de domínio para permitir a reutilização de conhecimento, como objeto de estudo; (l) ao compartilhamento de entendimento comum entre pessoas e agentes de *software*; (m) à interoperabilidade entre sistemas e (n) à utilização de bases de conhecimento comum para diversas aplicações, tais como: comércio eletrônico, Buscas e Recuperação de Informação, Tradução Automática, Web Semântica, entre outros.

Em seguida, partimos para a extração automática dos TOs, com o auxílio de uma ferramenta computacional que extrai de forma automática candidatos a termos, cujo critério adotado foi o da freqüência (cf. TELINE *et al.*, 2003). 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 mesmos.

A extração automática de itens lexicais é uma interessante etapa na identificação de itens terminológico-ontológicos, uma vez que possibilita a identificação de lexias complexas que podem passar desapercebidas numa extração manual, dada a precariedade de tal metodologia. Com efeito, Teline *et al*. (2003) acentuam que:

> Extrair manualmente do corpus (...) os candidatos a termo faz com que o terminólogo enfrente uma das maiores dificuldades na pesquisa terminológica, qual seja, o terreno movediço que há entre palavra (unidade da língua geral) e termo (unidade das comunicações especializadas). Uma das etapas fundamentais de qualquer pesquisa desse tipo é a coleta de termos nos textos especializados. Ora, que critérios deveremos utilizar para efetuar essa tarefa a contento? Dito de outro modo: como saber, ao certo, se aquela unidade selecionada é termo ou palavra, já que o terminólogo, na maioria das vezes, não é um especialista da área que está sendo objeto de investigação? (TELINE *et al*., 2003, p. 2)

Vejamos a relação de unidades ontológicas (bigramas ou lexias complexas) resgatadas de forma automática:

Abundância relativa; Ciclos populacionais (ciclo populacional); Ciclo vital; Classe etária; Competição intra-específica; Competição interespecífica; Controle biológico; Crescimento exponencial; Crescimento geométrico; Crescimento logístico; Crescimento populacional; Curva logística; Densidade absoluta; Densidade bruta; Densidade populacional; Densidade relativa; Dinâmica populacional; Distribuição agregada; Distribuição aleatória; Distribuição espacial; Distribuição etária; Distribuição uniforme; Diversidade genética; Ecologia vegetal; Energia líqüida; Equação logística; Esforço reprodutivo; Estrutura etária; Fatores ambientais (fator ambiental); Fatores bióticos (fator biótico); Fator-chave; Fatores climáticos (fator climático); Fatores ecológicos (fator ecológico); Fatores limitantes (fator limitante); Flutuações populacionais (flutuação populacional); Fluxo energético; Freqüências

relativas (freqüência relativa); Genética quantitativa; Idade reprodutiva; Longevidade ecológica; Meio ambiente; Níveis tróficos (nível trófico); Pirâmides etárias (pirâmide estária); Populações estáveis (população estável); Potencial biótico; Potencial reprodutivo; Regulação populacional; Resistência ambiental; Seleção artificial; Seleção natural; Sistemas abertos (sistema aberto); Taxa instantânea; Taxa intrínseca; Taxa reprodutiva.

Quadro 2: Extração Automática de candidatos a Termos Ontológicos – Bigramas – em Ecologia de Populações pelo método da Freqüência (legitimadas pela LR). Resultados: 54

Além desses TOs, a extração automática nos resgatou outros 132 candidatos a termos que, embora não tenham sido legitimados pela lista de referência[13] (LR) na sua totalidade, pelo menos 50 deles foram reconhecidos como termos legítimos pelo especialista na área. Tal fato é decorrente, certamente, da não representatividade da LR no momento da extração dessas unidades. São eles:

seres humanos; população humana; taxa específica; variação genética; populações naturais; faixa etária; condições ambientais; tamanho populacional; taxa exponencial; tamanho inicial; peso corporal; processos populacionais; densidade ecológica; indivíduos marcados; equação diferencial; modelo logístico; flutuações aleatórias; crescimento rápido; variância fenotípica; equação exponencial; densidade máxima; ciclos limitados; dinâmica espacial; taxa finita; taxa máxima; valores fenotípicos; taxa líqüida; padrão sigmoidal; escala aritmética; proporção direta; forma integrada; aumento populacional; equações diferenciais; forma carbonária; oscilações amortecidas; logaritmos neperianos; natalidade específica; desenvolvimento larval; forma sigmoidal; energia assimilada; sobrevivência específica; material genético; modelo exponencial; escalas aritméticas; escala logarítmica; estrutura etária; idade precoce; crescimento sigmoidal; taxa média; controle natural; variações aleatórias; estrutura espacial; estrutura populacional; composição etária; curva sigmoidal; taxa anual; intervalos sucessivos; crescimento natural; escala semilogarítmica; natalidade bruta; efeitos deletérios; população inicial; população laboratorial; variância V; crescimento instantâneo; estágios juvenis; planta herbácea; desenvolvimento ontogenético; oscilações regulares; correlações genéticas; taxas relativas; curva exponencial; áreas naturais; formas disseminantes; genética populacional; reino animal; espécies migratórias; Equação diferencial; oscilação amortecida; eventos aleatórios; processo logístico; mortalidade adulta; taxas intrínsecas; dinâmica metapopulacional; assíntota superior; flutuações ambientais; variação populacional; modelo geométrico; distribuições não-aleatórias; abundâncias máximas; área sombreada; ambiente ilimitado; base genética; fatores meteorológicos; fatores extrínsecos; mortalidade populacional; dispersão aleatória;

[13] *Glossário de Ecologia* (1997*), Dicionário de Ecologia e Ciências Ambientais* (2001), além do consultor Prof. Dr. Luiz Zanini Branco da UNESP de São José do Rio Preto, especialista da Área de Ecologia.

ciclos regulares; relações ecológicas; alelo deletério; taxa geométrica; forma exponencial; organismos patogênicos; repartição energética; abundância máxima; superfície aquática; taxas instantâneas; período reprodutivo; seleção diferencial; análise estatística; nível populacional; fatores dependentes; fatores independentes; estrutura populacional; aumento exponencial; tamanho corporal; ajustamento diferencial; mudança ambiental; população máxima; ciclos vitais; oscilações cíclicas; condições naturais; natalidade máxima; sistemas genéticos; alelo dominante; taxa finita; vida reprodutiva; crescimento positivo; função exponencial; limites geográficos; ação combinada; animais exotérmicos.

Quadro 3: Extração Automática de candidatos a Termos Ontológicos – Bigramas – em Ecologia de Populações pelo método da Freqüência (não legitimados pela LR). Resultados: 132

A metodologia da extração automática, no que diz respeito aos unigramas, foi de grande valia, e nos revelou um número expressivo de candidatos a termos validados pela RF. Vejamos:

abundância; acasalamento; agregação; aleatório; alimentação; alimento; ambiente; amostra; área; árvores (árvore); atributos (atributo); autofertilização; base; biomassa; biótico; campo; ciclos (ciclo); classe; colonização; comportamento; comunidades; condições (condição); controle; corpos-lúteos (corpo-lúteo); crescimento; densidade; desvio; dispersão; distribuição; diversidade; ecologia; ecossistema; emigração; energia; endocruzamento; escala; espécie; exocruzamento; extinção; fator-chave; fecundidade; fenótipo; fertilidade; florestas (floresta); forma; freqüência; gene; genótipo; geração; grupo; habitat; hereditariedade; hibernação; indivíduos; imigração; irrupção; isolamento; melanismo; metapopulação; microrganismos; migração; modelo; modo; mutação; natalidade; ninhada; nutrientes (nutriente); organismo; oscilações (oscilação); padrão; parasitismo; período; peso; plantas (planta); população; predação; predador; probabilidade; produção; recursos; região; regulação; reprodução; retardamento; retroalimentação; seleção; sobrevivência; taxa; territorialidade; território; tipo; variação; variância; vegetação; vida.

Quadro 4: Extração Automática de candidatos a Termos Ontológicos – Unigramas – em Ecologia de Populações pelo método da Freqüência (legitimados pela LR). Resultados: 95

5 O delineamento da OntoEco

Na elaboração de ontologias, alguns passos têm necessariamente de serem seguidos e algumas etapas precisam ser cumpridas. Vejamos, a seguir, quais são eles, concernentes à ferramenta computacional Protégé:

5.1. Projeção

Algumas questões são abordadas, tais como: (a) *Qual o domínio especificado?* A Ontologia que propomos neste trabalho abrange o subdomínio da "Ecologia" que pertence ao domínio maior "Ciências Biológicas". O subdomínio "Ecologia" foi subdividido em 3 subdomínios, como especificamos no início deste: "Ecologia de Ecossistemas" – EEc; "Ecologia de Populações" – Ep; "Ecologia de Comunidades" – Ec; (b) *Qual a utilidade?* A Ontologia deverá servir para aplicações diversas, tais como a Tradução Automática na Internet, para a Recuperação de Informação em sites na Internet e para a Web Semântica na Busca de Informações mais precisas e refinamento de *Queries*. Além disso, a ontologia deverá servir para o reuso e a interoperabilidade por e com sistemas computacionais; (c) *Qual tipo de informação veicula?* Numa Ontologia objetiva-se o provimento de informações para perguntas/buscas de tipo diversificado. Na *OntoEco*, deveremos prover sua base para servir a informações do tipo: (i) Representação geral dos sub-domínios de Ecologia de Populações, de Comunidades e de Ecossistemas. Possíveis perguntas: como são formados, qual seu mapa conceitual, seus sub-domínios são constituídos do que, formados por, etc; (ii) Informações específicas sobre Ecossistemas terrestres, por exemçlo: "pastagens" e "florestas" são a mesma coisa? E "desertos" e "tundras"? (iii) Tipos de relacionamentos: "um lago artificial pode ser considerado um Ecossistema Aquático, como rios e mares?" O que eles têm em comum? E de diferente? (iv) O São Francisco é um rio ou um lago? (aplicação imediata: TA); (v) Em Recuperação da Informação, se um usuário pretender realizar uma busca sobre "População/Populações", o sistema de pergunta, estando vinculado a OntoEco, poderá lhe retornar informações sobre "Populações" em Ecologia e lhe sugerir sites para isso.

5.2. Utilização

A OntoEco servirá para a reutilização de/ou por sistemas computacionais para aplicações variadas, tais como: Buscas e Recuperação de Informação e Tradução automática.

5.3. Enumeração dos termos na ontologia

Nessa etapa, a questão levantada é (a) *Quais são as unidades lexicais ontológicas?* A resposta encontra respaldo conforme descrito na seção 4 deste.

5.4. Definição das Classes

No processo de construção da OntoEco, utilizamo-nos tanto da abordagem (i) *top-down* que define primeiramente os conceitos do senso comum e, em uma segunda etapa, abarca o conhecimento especializado: **Ecologia → Ecologia Comunidades, Ecologia de Ecossistemas, Ecologia de Populações** como a *bottom-up* que parte de um número pré-definido de bases especializadas para, em seguida, integrar os conceitos gerais que fazem parte do senso comum: **Florestas e Desertos → Ecossistema Terrestre → Ecologia de Ecossistemas**. Optamos pela combinação dessas duas abordagens dado que à medida que encontramos termos específicos, criamos classes para eles e conseqüentemente superclasses. Ao contrário, detectados conceitos muito gerais, criamos superclasses para que a elas fossem agrupadas subclasses e termos específicos. Esse é o chamado *Processo de Desenvolvimento Híbrido*. Com isso, cria-se a taxionomia hierárquica da ontologia que, em termos gerais, define-se a partir do seguinte axioma: "Se a categoria A é a superclasse da categoria B, então toda instância de B é também uma instância de A. Em outras palavras: a categoria B representa um conceito que é 'tipo_de'".

5.5. Definição dos *slots* (propriedades das classes)

Decidimos por manter alguns *slots* já disponíveis na ferramenta, além de definir e criar outros deles pontualmente, a partir das relações da *Qualia* Extendida de Lenci (1999), a começar das relações da Estrutura *Qualia* de Pustejovsky (1995) (cf. ZAVAGLIA, 2002). Existem dois tipos de propriedades que podem se tornar *slots* na ontologia: "propriedades intrínsicas" – como a cor do mar, o sabor do mar, etc. e "propriedades extrínsicas" – como o nome de um mar, rio, como Mar Atlântico.

5.6. Definição das facetas dos *slots*

No Protégé, existem os seguintes tipos de facetas que podem ser acionados na definição de um *slot*: **Cardinalidade** que define quantos valores um *slot* pode ter; **Tipo de valor para o slot** que expõe quais os tipos de valores que podem ser preenchidos num *slot* e **Domínio e Extensão (Range) de um slot** em que as classes consentidas ou

reconhecidas para *slot*s do tipo Instância são freqüentemente chamadas de extensão (*range*) de um *slot*. Na *OntoEco* "florestas" é a extensão do *slot* VEGETAL, por exemplo.

5.7. Criação de instâncias

Para se definir uma instância individual de uma classe é preciso: (i) escolher uma classe, (ii) criar uma instância individual daquela classe e (iii) preencher o campo de valores para os *slot*s.

5.8. Construção das Estruturas Arbóreas

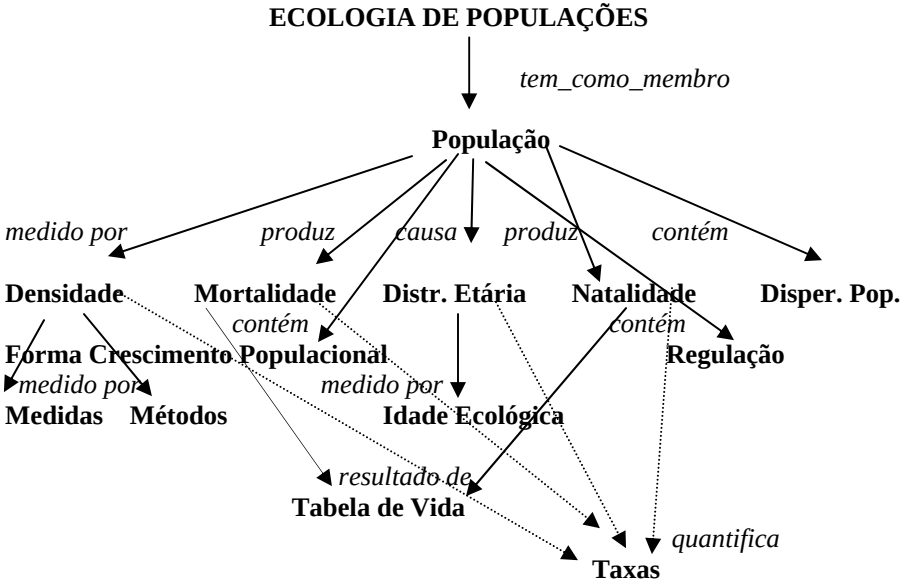


Figura (1) - Ecologia de Populações – relacionamentos.

6. Estruturação e organização do conhecimento semântico

A partir do levantamento dos TOs, descrevemos formalmente o conhecimento semântico inerente a cada um deles, explicitados ora em classes ora em itens lexicais.

A abordagem teórica utilizada, qual seja a Teoria do Léxico Gerativo, particularmente os postulados da Estrutura *Qualia*, nos

permitiu capturar dimensões do significado de um item lexical, por meio das relações semânticas que perfazem os papéis Télico, Agentivo, Constitutivo e Agentivo.

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 TOs 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 essa relação 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, 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 classes e subclasses. Dessa forma, temos como subclasses da superclasse **CLASSES**: **INTERAÇÃO; POPULAÇÃO; COMUNIDADE; ECOSSISTEMA; ENERGIA; BIOTA; ÁREA; DENSIDADE; NATALIDADE; MORTALIDADE; DISTRIBUIÇÃO_ETÁRIA; PRODUTIVIDADE; CICLOS_BIOGEOQUÍMICOS; FORMA_CRESCIMENTO_POPULACIONAL; REGULAÇÃO; DISPERSÃO_POPULACIONAL; FLUXO_DE_ENERGIA; FAUNA; VEGETAÇÃO; CLIMA; TEMPERATURA; UMIDADE; PRECIPITAÇÃO; IDADE_ECOLÓGICA; TABELA_DE_VIDA; MEDIDA; MÉTODO; ECOLOGIA; TAXAS**.

Vejamos a sua implementação no Protégé:

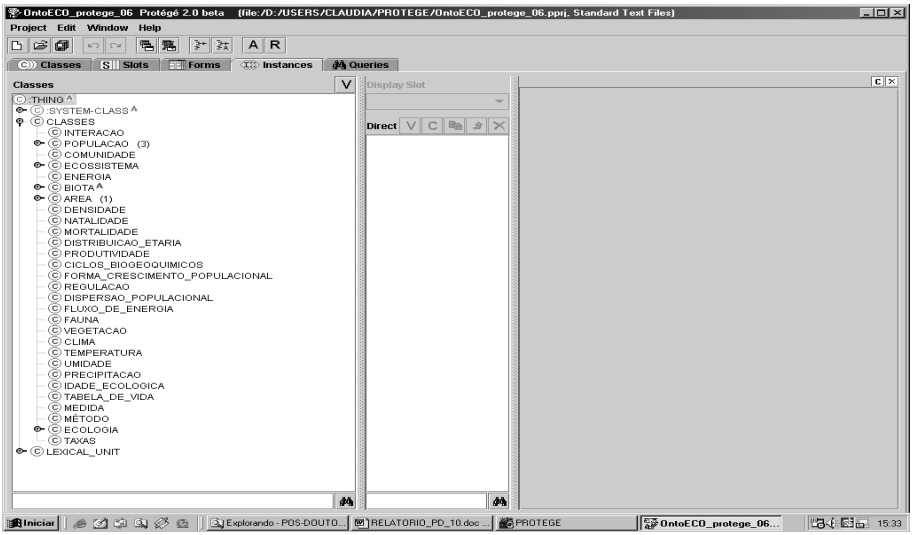


Figura (2) Implementação da classe CLASSES no Protégé.

Vejamos as classes que possuem a relação *é-um* cujas subclasses herdam todas as características da superclasse:

- Classe **ECOSSISTEMA** → Subclasses vinculadas por *é-um*: **ECOSSISTEMA_ARTIFICIAL** e **ECOSSISTEMA_NATURAL**: *é-um* (<ecossistema>,<eco_artificial>), em que temos o axioma "todo eco_artificial é um ecossistema" e *é-um* (<ecossistema>,<eco_natural>), em que "todo eco_natural é um ecossistema".

Há casos em que as subclasses são também superclasses a partir do momento que a elas sejam vinculadas subclasses por meio da relação *é-um*, como é o caso de:

- Classe **ECOSSISTEMA_ARTIFICIAL** → Subclasses vinculadas por *é-um*: **REFLORESTAMENTO** e **REPRESAS**: *é-um* (<ecossistema_artificial>,<reflorestamento>), em que temos o axioma "todo reflorestamento é um ecossistema-artifical" e *é-um*(<ecossistema_artificial>,<represas>), em que "toda represa é um ecossistema_artificial".

- Classe **REFLORESTAMENTO** → Subclasses vinculadas por *é-um*: **REFLORESTAMENTO_NATIVOS** e **REFLORESTAMENTO_INTRODUZIDOS**: *é-um* (<reflorestamento>,<reflorestamento_nativos>), em que temos o axioma “todo reflorestamento_nativos é um reflorestamento” e *é-um* (<reflorestamento_reflroestamento_introduzidos>), em que “todo reflorestamentointroduzido é um reflorestamento”.

A seguir, listaremos as outras classes e subclasses que seguem o mesmo padrão de estruturação semântica demonstrado acima:

- Classe **ECOSSISTEMA_NATURAL** → Subclasses vinculadas por *é-um*: **ECOSSISTEMA_TERRESTRE** e **ECOSSISTEMA_AQUÁTICO;**
- Classe **ECOSSISTEMA_TERRESTRE** → Subclasses vinculadas por *é-um*: **FLORESTAS, TUNDRAS, CAMPOS, DESERTOS, MANGUES, RESTINGAS** e **CAATINGA;**
- Classe **FLORESTAS** → Subclasses vinculadas por *é-um*: **BOREAL, TEMPERADA** e **TROPICAL;**
- Classe **CAMPOS** → Subclasse vinculada por *é-um*: **CERRADO;**
- Classe **ECOSSISTEMA_AQUÁTICO** → Subclasses vinculadas por *é-um*: **MARINHO** e **DULCÍCOLA;**
- Classe **MARINHO**→ Subclasses vinculadas por *é-um*: **OCEANO_ABERTO** e **PLATAFORMA_CONTINENTAL;**
- Classe **DULCÍCOLA**→ Subclasses vinculadas por *é-um*: **LÊNTICO** e **LÓTICO;**
- Classe **BIOTA**→ Subclasses vinculadas por *é-um*: **VEGETAL** e **ANIMAL;**
- Classe **ANIMAL**→ Subclasses vinculadas por *é-um*: **HERBÍVORO** e **CARNÍVORO;**
- Classe **ÁREA**→ Subclasses vinculadas por *é-um*: **REGIÃO** e **HABITAT;**
- Classe **HABITAT**→ Subclasses vinculadas por *é-um*: **HABITAT_TERRESTRE** e **HABITAT_AQUÁTICO;**
- Classe **ECOLOGIA**→ Subclasses vinculadas por *é-um*: **ECOLOGIA_DE_COMUNIDADES;** **ECOLOGIA_DE_ECOSSISTEMAS**e **ECOLOGIA_DE_POPULAÇÕES.**

Utilizamo-nos também da relação ***é_um*** para estabelecer que todo item lexical implementado como subclasse da classe **LEXICAL_UNIT** é um substantivo a partir do momento que as unidades léxico-ontológicas (ULOs) existentes na OntoEco estão vinculadas à subclasse **SUBSTANTIVO** que por sua vez está vinculada à **LEXICAL_UNIT**. Temos então que "agregação" é um substantivo que por sua vez é uma unidade lexical ontológica. Toda subclasse **SUBSTANTIVO** possui como característica os *slot*s implementados como *frames*: **Antônimo** (que traz a ULO cujo conceito é contrário à entrada), **Sinônimo** (que traz a ULO cujo conceito é sinônimo à entrada), **Contexto** (que traz a contextualização da ULO extraída do *Corpus*Eco), **Morfologia** (que traz informações sobre o número e o gênero da entrada) e **EquivIt** (que traz o equivalente tradutório em língua italiana da unidade em questão).

A definição da ULO é reportada em **Documentation**, *frame* já previsto pela própria ferramenta. Além disso, para que pudéssemos relacionar a ULO como sendo uma unidade lexical ativa em uma determinada classe ou subclasse prevista na ontologia, vinculamos a ULO à classe ou subclasse por meio do *frame* **Superclasses**, já previsto pela ferramenta também. Desse modo, toda ULO terá pelo menos, duas Superclasses, fato esse que a caracterizará como sendo uma subclasse que possui múltiplos parentes ou herança múltipla, caracterizado pelo "M" em sobrescrito à ULO.

Vejamos, a seguir, a entrada "agregação", apresentada em forma de tabela ontológica e a sua implementação no Protégé:

Agregação

SemU:	<agregação>
Tipo:	[População]
Supertipo:	[Ecologia de Populações]
Domínio:	*Ecologia*
Formal:	*é_um*(<agregação>,<conjunto>)
Agentivo:	*<Nil>*
Constitutivo:	*conjunto_de*(<agregação>,<indivíduo>) *tem_como_membro*(<agregação>,<indivíduo>) *está_em*(<agregação>,<habitat>)
Télico:	<Nil>
Glossário:	Conjunto de indivíduos de uma mesma espécie agrupados em conseqüência de diferentes estímulos ambientais, como a atração sexual ou as diferenças entre habitat.
Exemplo:	*O agrupamento, ou **agregação**, resulta das tendências sociais dos indivíduos a formar grupos.*

PDD:	NOME
MORFOL:	FEM SING
SemU_syn	<Agrupamento>
SemU_ant	<Nil>
Equiv_It	*Aggregazione*

Tabela (1) - Termo Ontológico "agregação"

Vejamos, a seguir, "agregação" implementada no Protégé:

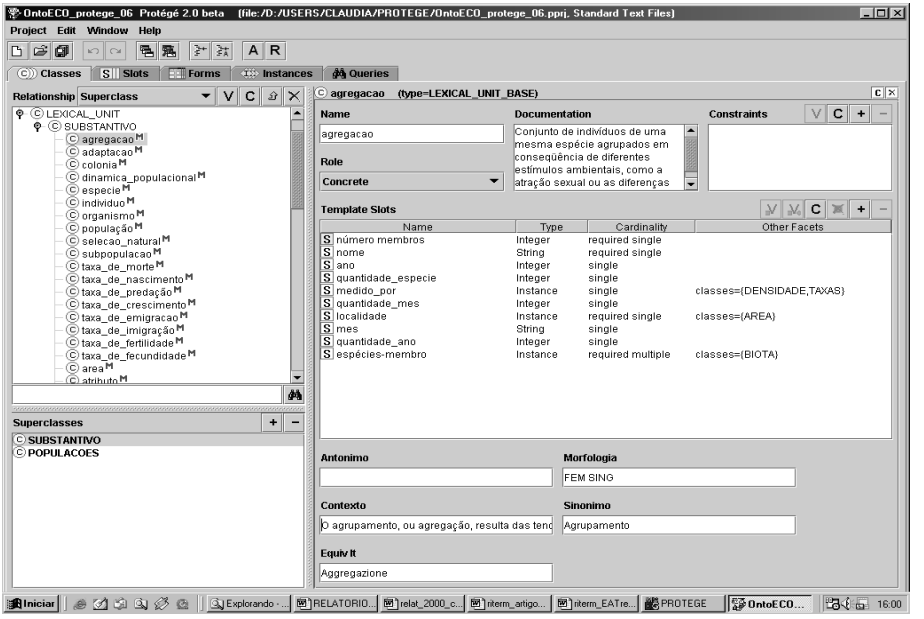


Figura (3) Implementação de "agregação" no Protégé.

Por sua vez, a relação *parte-todo*, ou seja, a meronímia/holonímia pode ser implementada de três maneiras no Protégé-2000:

(1) por meio de *slots* que ativam as propriedades de cada classe e subclasse da ontologia;

(2) como subclasses da classe **DIRECTED-BINARY-RELATION** que é subclasse da classe **RELATION** que tem como superclasse **SYSTEM-CLASS** e

(3) em *frames*. No nosso caso, foram configurados os frames **:CONSTITUTIVO**; **:TÉLICO**; **:AGENTIVO**; **:FORMAL** para todas as classes que forem vinculadas a **META_CLASS**, no caso a **SEM_CLASS_BASE**.

A relação de meronímia pode ser equiparada às informações previstas pelo papel Constitutivo da Estrutura *Qualia*. A tipologia das

relações que vincula as classes e subclasses da *OntoEco* (Ecologia de Populações) é: ***tem_como_membro; medido_por; produz; causa; contém; resultado_de; quantifica*** que foram especificadas pela *Qualia* Extendida (LENCI, 1999). Essas relações parte-todo foram estabelecidas como *slots*, e servem para vincular classes e subclasses. Além desses, temos outros tipos de meronímia, tais como: ***é_parte_de; tem_como_parte; instrumento; relaciona; estado_resultante; é_um_seguidor_de; feito_de; está_em; vive_em; tem_como_cor; atividade_constitutiva; transfere; produzido_por; propriedade_de; tem_como_propriedade; concerne; inclui; relacionado_a/com; sucessor_de; tem_como_efeito; típico_de.*** Por sua vez, a maioria dessas relações foi implementada no *frame* **:CONSTITUTIVO** que implementamos para a caracterização de classes e subclasses, i.e., como especificação do significado que o conceitua.

Uma vez que as dimensões Télica e Agentiva também podem permitir a codificação de informações multifacetadas, implementamos os *frames* **:TÉLICO** e **:AGENTIVO** para toda classe e subclasse. Em nossa ontologia, a faceta Télica, cuja proposta é a de apontar a finalidade, o escopo ou o objetivo associado a um item lexical foi mais ativa do que a faceta Agentiva, que prevê a origem do objeto, cuja distinção nem sempre é possível. O papel Télico prevê as seguintes relações: ***télico_indireto; propósito; é_a_atividade_de; é_a_habilidade_de; é_o_hábito_de; usado_para; usado_por; usado_contra; usado_como***. Já o papel Agentivo prevê: ***resultado_de; causa_agentiva; experiência_agentiva; causado_por; origem; criado_por; derivado_de***.

Para a classe **POPULAÇÃO,** estabelecemos os *slot*s *quantidade_mês*, *quantidade_ano*, *quantidade_espécie*, *medido_por*, com o propósito de quantificarmos uma dada população. Já os *slot*s *nome*, *ano*, *localidade*, *mês* servem para caracterizá-lo. Vejamos:

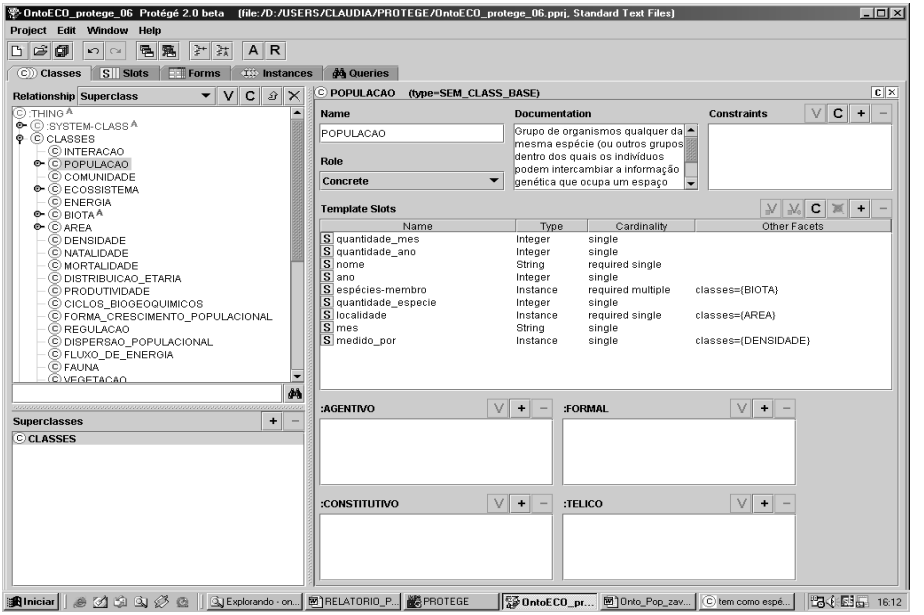


Figura (4) Slots de População no Protégé-2000.

Considerações finais

A mediação realizada por meio de ontologias em busca de informações na Internet consente uma maior precisão e importância nos documentos que são encontrados e investigados, uma vez que podem servir como um esquema conceitual de um determinado domínio, servindo de suporte semântico às buscas ou consultas realizadas. Com efeito, quando as máquinas de busca fazem uso de ontologias para realizarem consultas por palavras-chave, por exemplo, em suas respostas-saída elas podem oferecer além das páginas que contêm a palavra-chave requisitada, outras páginas que contenham informações vinculadas ao conceito das palavras-chave, tais como sinônimos, antônimos, hiperônimos, hipônimos e termos relacionados ou dependentes.

Modelamos e estruturamos conceitualmente três subdomínios do domínio da Ecologia, a saber: o subdomínio Ecologia de Populações, de Ecossistemas e de Comunidades, todos de forma manual. Todas as classes desses subdomínios foram conceituadas e implementadas na ferramenta Protégé, respeitando os limites impostos pela máquina. Por sua vez, os termos ontológicos estudados, estruturados e conceitualizados amiúde, respeitando todos os campos de informação previamente propostos, foram os pertencentes à Ecologia de Populações, uma vez que

se tratava apenas de um protótipo de reuso e interoperabilidade da representação ontológica registrada.

Dito isso, implementamos 29 SUBCLASSES da classe CLASSES, com suas respectivas subclasses que somam 36. Convém ressaltar que definimos letras maiúsculas para CLASSES e SUBCLASSES e letras minúsculas para os termos ontológicos relacionados a cada CLASSE e/ou SUBCLASSE, implementados como SUBCLASSES da SUBCLASSE SUBSTANTIVO da CLASSE LEXICAL_UNIT. Além disso, usamos o plural para definirmos apenas CLASSES e/ou SUBCLASSES, sendo que para os TOs optamos pelo singular. Até o presente momento implementamos 64 TOs todos com os campos "name" (correspondente ao nome do termo ontológico); "documentation" (que possui a definição do TO); "role" (apresentando todos a característica de "concrete"); "Template *Slot*s" (que atribuem as propriedades, por meio de *slot*s, a cada TO); "antônimo" (quando existente, atribui o TO antônimo à entrada); "sinônimo" (quando existente, atribui o TO sinônimo à entrada); "morfologia" (traz informações sobre gênero e número do TO); "contexto" (reporta a contextualização autêntica do TO retirada do *Corpus*Eco); Equiv_it ou Equivalente em Italiano (que traz o equivalente do TO em língua italiana) preenchidos quando existem as informações para casa um deles.

No momento de transportar as subontologias geradas manualmente para a ferramenta Protégé, deparamo-nos com alguns e variados limites impostos pelo formalismo da ferramenta:

- O primeiro deles, e o que nos limitou sobremaneira, foi ter de criar um *slot* que fosse o mais genérico possível, uma vez que, a partir do momento que criamos um *slot* com determinado nome a ferramenta não permite uma duplicação nominal, não existindo, portanto homônimos nesse formalismo. Lembremos que um *slot* pode ser utilizado para descrever as propriedades de diversas e diferentes classes, mas, se o *slot* contiver uma particularização de valor, por exemplo, em "Template Values" ou em "Default", quando estiver caracterizado como "Instance", as Classes asssociadas a esse *slot* serão herdadas para todas as classes as quais o *slot* será vinculado. Para a classe TEMPERATURA, por exemplo, precisávamos usar o *slot* "medido-por" que foi implementado como uma instância, já que serviu a muitas classes com essas características. Entretanto as classes a ele associadas não serviam para a classe

TEMPERATURA, que precisava ser medida por “grau” e não por “densidadades” ou “taxas” . A solução encontrada para isso foi criarmos uma relação binária na classe RELATION “medido-por (grau)” e a implementamos no valor CONSTITUTIVO da classe TEMPERATURA. Por sua vez, se o *slot* for marcado como “String”, aí então poderemos particularizar informações para os valores “Template Values” e “Default” que poderão ser alterados de acordo com a classe a ele associada (e que serão herdadas por todas as sub-classes dessa classe).

- O segundo grande impedimento da ferramenta foi o fato de a sua estruturação ser toda centrada na relação ISA, enquanto que nem todas as subclasses das classes das sub-ontologias possuíam essa relação. Desse modo, as subclasses que não possuiam a ISA com a sua classe foram implementadas como classes independentes vinculadas a sua superclasse inicialmente. É o caso de super classe REGIÃO que possui a relação de “tem-como-propriedade” com as subclasses FAUNA, CLIMA e VEGETAÇÃO. As subclasses foram implementadas como classes independentes, vinculadas no valor “superclasses” à classe REGIÃO, tendo adquirido o *slot* “tem-como-propriedade” com os valores “fauna, vegetação e clima”. Isso ocorreu com diversas classes.

O trabalho que realizamos foi de encontro com os objetivos propostos no início deste, ou seja, demonstrar o quão expressivo é o uso do léxico de uma língua em aplicações e reusos computacionais, uma vez que não expusemos apenas uma lista de palavras com informações morfossintáticas extraídas de um *corpus*, antes. Apresentamos a proposta de elaboração de uma Base Lexical que possui um conhecimento semântico de forma refinada e detalhada, por meio de frames e *slot*s, que são uma forma de representação comum e bastante utilizada em Lingüística Computacional, cujo aplicação se verifica por meio da ferramenta Protégé.

Para a sistematização do conjunto de informações terminológicas de um domínio é fundamental o uso de ferramentas computacionais para a extração de termos. Para o português do Brasil, muitos projetos de construção de repertórios terminológicos ainda utilizam o critério semântico para a extração de termos, em uma abordagem manual a partir de *corpus*. Ainda que o critério semântico seja adequado, a extração manual é lenta, sujeita à subjetividade e à omissão de termos importantes.

Nesse sentido, a extração automática de termos torna-se uma etapa essencial no processo de delimitação semântica de um domínio.

Este trabalho pretendeu revelar a importância de se utilizarem ferramentas computacionais no trabalho prático em Lingüística e suas subáreas Lexicologia, Lexicografia, Terminologia e Terminografia.

Tivemos como objetivo relatar o quão cuidadoso deve ser o trabalho do lingüista-ontólogo no processo de elaboração de estruturas ontológicas, cujo olhar deve ser extremamente atento e direcionado para o campo minado da conceitualização de categorias de mundo e de seus domínios específicos.

Nesse sentido, o trabalho desenvolvido justifica-se e torna-se relevante na medida em que revelou facetas até então inexploradas por lingüistas em domínios específicos.

Referências Bibliográficas

ALMEIDA, G. M. de B. **Teoria Comunicativa da Terminologia (TCT**): uma aplicação. 2000. Tese (Doutorado) - UNESP, Araraquara, 2000. p. 26-36.

ALVES, I. M. A Delimitação da Unidade Lexical nas Línguas de Especialidade. **PaLavra**, Rio de Janeiro, Grypho, v. 5, 1999.

BATEMAN, J. A. Ontology construction and natural language. In: N. GUARINO, N.; POLI, R. THE INTERNATIONAL WORKSHOP ON FORMAL ONTOLOGY IN CONCEPTUAL ANALYSIS AND KNOWLEDGE REPRESENTATION, 1993, Padova. **Proceedings…..**Padova, 1993.

BIDERMAN, M. T. C. Conceito Lingüístico de Palavra. **PaLavra**, Rio de Janeiro, Grypho, v. 5, 1999.

BIDERMAN, M. T. C. **Teoria Lingüística**: teoria lexical e lingüística computacional. 2. ed. São Paulo: Martins Fontes, 2001.

BLECUA, J. M. *et al*. **Filología e informática**: Nuevas tecnologias em los estudios filológicos. Barcelona: Editorial Milenio i Universitat Autònoma de Barcelona, 1999.

CAMPOS, F. C. A.; SANTOS, N.; BRAGA, R. M. M. Ontologias para o Domínio da Educação Mediada pela *Web*. In: WORKSHOP DE ONTOLOGIAS PARA A CONSTRUÇÃO DE METODOLOGIAS DE BUSCA NA WEB POR CONTEÚDOS EDUCACIONAIS – SBIE, 13., 2002, São Leopoldo. **Anais...** Unisinos. São Leopoldo, RS, 2002.

CASTELLVÍ, M. T. C. Informática y terminología. In: BLECUA, J. M. *et al.* (Ed.) **Filología e Informática**. Nuevas tecnologías en los estudios filológicos. Seminario de Filología e Informática, Departamento de Filología Española: Unversidad Auntónoma de Barcelona, 1999.

CHISHMAN, R. L de O. Ontologias e Relações Semânticas: uma aplicação. In: WORKSHOP DE ONTOLOGIAS PARA A CONSTRUÇÃO DE METODOLOGIAS DE BUSCA NA WEB POR CONTEÚDOS EDUCACIONAIS – SBIE, 13., 2002. **Anais...** Unisinos. São Leopoldo, RS, 2002.

DAMASCENO, F. O.; OLIVEIRA, A. de P. Uso de Ontologias para Suporte à Classificação Automática de Documentos. In: WORKSHOP DE ONTOLOGIAS PARA A CONSTRUÇÃO DE METODOLOGIAS DE BUSCA NA WEB POR CONTEÚDOS EDUCACIONAIS – SBIE, 13., 2002, São Leopoldo. **Anais...** Unisinos. São Leopoldo, RS, 2002.

DA SILVA, G. C., LIMA, T. de S. **RDF e RDFS na Infra-estrutura de Suporte à Web Semântica**. 2002. Disponível em: http://www.sbc.org.br/reic/edicoes/2002e1/cientificos/RDFeRDFSnaInfraEstruturadeSuporteaWebSemantica.pdf Acesso em 20/01/2004.

FERREIRA, A. B. H. **Novo Dicionário Aurélio Eletrônico século XXI.** Rio de Janeiro: Nova Fronteira. Versão 3.0, 1999.

FREIRE, R. C.; OLIVEIRA, A. P.; JHAM, A. M. C. Uso de Ontologia Léxica para Captura de Informações contidas em Manchetes Jornalísticas. In: WORKSHOP DE ONTOLOGIAS PARA A CONSTRUÇÃO DE METODOLOGIAS DE BUSCA NA WEB POR CONTEÚDOS EDUCACIONAIS – SBIE, 13., 2002, São Leopoldo. **Anais...** Unisinos. São Leopoldo, RS, 2002.

GOÑI, J. L.; FERNANDES, M. C. P.; LUCENA, C. J. P. Geração de Ontologias usando Protégé-2000 para reuso de conteúdos educacionais

numa arquitetura multiagente. In: WORKSHOP DE ONTOLOGIAS PARA A CONSTRUÇÃO DE METODOLOGIAS DE BUSCA NA WEB POR CONTEÚDOS EDUCACIONAIS – SBIE, 13., 2002, São Leopoldo. **Anais...** Unisinos. São Leopoldo, RS, 2002.

GUARINO, N. The Ontological Level. In: CASATI, R. *et al* (Ed.) **Philosophy and the Cognitive Sciences**. Vienna: Hölder-Pichler-Tempsky, 1994.

GUARINO, N. **Formal Ontology, Conceptual Analysis and Knowledge Representation**. Special issue on Formal Ontology, Conceptual Analysis and Knowledge Representation edited by N. Guarino and R. Poli. 1995. Disponível em: http://www.loa-cnr.it/Papers/FormOntKR.pdf Acesso em 20/01/2004.

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.

GUARINO, N. **Semantic Matching**: Formal Ontological Distinctions for Information Organization, Extraction and Integration. Berlin: Springer Verlag, s/d. p.139-170.

GRUBER, T. R. **Toward principles for the design of ontologies used for knowledge sharing**. Presented at the Padua workshop on Formal Ontology, March 1993, to appear in an edited collection by Nicola Guarino.

MAHESH, K.; NIRENBURG, S. A situated Ontology for Practical NLP. In: WORKSHOP ON BASIC ONTOLOGICAL ISSUES IN KNOWLEDGE SHARING. 1995, Montreal. **Proceedings…..** Montreal, 1995.

MANGAN, M. A. S.; MURTA, L. G. P.; SOUZA, J. M. WERNER, C. M. L. **Modelos de Domínio e Ontologias:** uma comparação através de um estudo de caso prático em hidrologia. s.d. (in mimmeo)

MELCOP, T. *et al*. Uma Ferramenta para Recuperação e Categorização de Páginas Web para Domínios Específicos. In: WORKSHOP DE ONTOLOGIAS PARA A CONSTRUÇÃO DE METODOLOGIAS DE

BUSCA NA WEB POR CONTEÚDOS EDUCACIONAIS – XIII SBIE'2002, São Leopoldo. **Anais...** Unisinos. São Leopoldo, RS, 2002.

NOY, N. F. *et al.* Creating Semantic Web Contents with Protege-2000. **IEEE Intelligent Systems,** v. 16, n. 2, p. 60-71, 2001.

NOY, N. F.; FERGERSON, R. W.; MUSEN, M. A. The knowledge model of Protege-2000: Combining interoperability and flexibility. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE ENGINEERING AND KNOWLEDGE MANAGEMENT (EKAW'2000), 12., 2000, Juan-les-Pins. **Anais...** Juan-les-Pins, France, 2000.

OLTRAMARI, A. *et al*. Il ruolo dell'ontologia nella disambiguazione del significato. **Networks**, v. 2, n. 14, 2003. Disponível em: http://lgxserve.ciseca.uniba.it/lei/ai/networks/03-2/guarino.pdf Acesso em 21/01/2004.

ORTIZ, A. M. Diseño e implementación de un Lexicón Computacional para lexicografía y Traducción Automática. **Estudios de Lingüística Española,** v. 9, 2000. Disponível em: http://elies.rediris.es/elies9/index.htm Acesso em 14/06/2002.

PIZZATO, L. A. S. ; LIMA, V. L. Estrutura Multitesauro para Recuperação de Informações. In: WORKSHOP DE ONTOLOGIAS PARA A CONSTRUÇÃO DE METODOLOGIAS DE BUSCA NA WEB POR CONTEÚDOS EDUCACIONAIS – SBIE, 13., 2002, São Leopoldo. **Anais...** Unisinos. São Leopoldo, RS, 2002.

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

RIGO, S.; VIEIRA, R. Busca de informações auxiliada por ontologia. In: WORKSHOP DE ONTOLOGIAS PARA A CONSTRUÇÃO DE METODOLOGIAS DE BUSCA NA WEB POR CONTEÚDOS EDUCACIONAIS – SBIE, 13., 2002, São Leopoldo. **Anais...** Unisinos. São Leopoldo, RS, 2002.

SAIAS, J.; QUARESMA, P. Construção automática de ontologias e sua utilização em sistemas de recuperação de informações em texto. In:

WORKSHOP DE ONTOLOGIAS PARA A CONSTRUÇÃO DE METODOLOGIAS DE BUSCA NA WEB POR CONTEÚDOS EDUCACIONAIS – SBIE, 13., 2002, São Leopoldo. **Anais...** Unisinos. São Leopoldo, RS, 2002.

SANTOS, E. T.; BARROS, L. N.; VALENTE, V. C. P. Projetando uma Ontologia de Geometria Descritiva. In: SIMPÓSIO NACIONAL DE GEOMETRIA DESCRITIVA E DESENHO TÉCNICO, 15.; INTERNATIONAL CONFERENCE ON GRAPHICS ENGINEERING FOR ARTS AND DESIGN, 4., São Paulo. **Anais....** São Paulo, 2001. Disponível em: http://www.cin.ufpe.br/~sas/chat/ontologiafinal.pdf Acesso em: 03/02/2003

SILVA, T. M. S. ; FREITAS, F. L. G.; BITTENCOURT, G. Extração de Informação no Mater-Web baseada em ontologias. In: WORKSHOP DE ONTOLOGIAS PARA A CONSTRUÇÃO DE METODOLOGIAS DE BUSCA NA WEB POR CONTEÚDOS EDUCACIONAIS – SBIE, 13., 2002, São Leopoldo. **Anais...** Unisinos. São Leopoldo, RS, 2002.

TELINE, M. F.; ALMEIDA, G. M. de B.; ALUÍSIO, S. M. Extração manual e automática de terminologia: comparando abordagens e critérios. In: TIL 2003; BRAZILIAN SYMPOSIUM ON COMPUTER GRAPHICS AND IMAGE PROCESSING –SIBGRAPI, 16., 2003, São Carlos. **Proceedings…** São Carlos, UFScar, 2003. v. 1, p. 1-12.

TISCORNIA, Daniela. Una metodologia per la rappresentazione della conoscenza giuridica; l'ontologia formale applicata al diritto. Articolo per conferenza di filosofia del diritto. Bologna, 1995. (in mimmeo)

USCHOLD, M.; GRUNINGER, M. Ontologies: principles, methods and applications. **The Knowledge Engineering Review**. v. 11, n. 2, p. 93-136, 1996. User's Guide for Protégé. Disponível em: http://protege.stanford.edu/doc/users_guide/index.html

VASCONCELOS, K. F. **OntoEditor**: Um editor para manipular ontologias na Web. 2003. Dissertação (Mestrado) - Campina Grande/PB, 2003. Disponível em: http://www.dsc.ufcg.edu.br/~copin/pesquisa/bancodissertacoes/2003/KarineFreitas.pdf. Acesso em: 10/11/2003

ZAVAGLIA, C.; KASAMA, D. Y. O delineamento de uma Ontologia com vistas ao tratamento computacional: uma proposta para o subdomínio de Ecologia de Comunidades. **PaLavra**, v. 12, n. 1, p. 1-18, 2004. Volume Temático: Processamento Computacional do Português organizado por M. C. Dias e V. Quental.

ZAVAGLIA, C.; GREGHI, G. Homonymy in Natural Language Processes: a representation using Pustejovsky's Qualia Structure and Ontological Information. In: MAMEDE *et al.*, N. J. (Ed.). PROPOR 2003, INAI 2721. ENCONTRO PARA O PROCESSAMENTO COMPUTACIONAL DA LÍNGUA PORTUGUESA FALADA E ESCRITA, 6., 2003, Berlin. **Anais...** Springer-Verlag: Berlin Heidelberg, 2003, p. 86-93.

ZAVAGLIA, C. **Análise da homonímia no português**: tratamento semântico com vistas a procedimentos computacionais. 2002. Tese (Doutorado) – UNESP, Araraquara, 2002. v. 1, p. 320; v. 2, p. 199.