# Estrutura Ontológica e Unidades Lexicais: uma aplicação computacional no domínio da Ecologia

**Claudia Zavaglia[1,2], Leandro Henrique Mendonça de Oliveira[2,3], Maria das Graças Volpe Nunes[2], Sandra Maria Aluísio[2]**

[1]Universidade Estadual Paulista – UNESP/IBILCE – Rua Cristóvão Colombo, 2265 – Bairro: Jardim Nazareth – CEP: 15054-000 – São José do Rio Preto – SP – Brasil

[2]Núcleo Interinstitucional de Lingüística Computacional – Instituto de Ciências Matemáticas e de Computação, USP – CP: 668 – 13560-970 – São Carlos – SP – Brasil

[3] Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) – Embrapa Informática Agropecuária – CNPTIA – Campinas – SP.

zavaglia@ibilce.unesp.br, {leandroh, gracan, sandra}@icmc.usp.br

***Abstract.*** *Ontologies are used for representing information units that contain related semantic understanding of varied real world situations. For systematizing the set of terminological data from a domain, the use of computational tools for term extraction is essential. This work presents the evaluation of statistical, linguistic and hybrid approaches to automatic term extraction for ontology construction. The evaluation was carried out with a reference list of terms from Ecology domain, using precision and recall metrics. The OntoEco ontology predicts three Ecology subdomains: ecosystems, populations and communities. For extracting the ontological lexical units, an Ecology corpus was built – the CórpusEco. After delineating the ontology in classes, subclasses and instances, the data were stored in the tool Protégé-2000.*

***Resumo.*** *Ontologias são usadas para a representação de informações que contêm um entendimento semântico comum de situações variadas do mundo real. Para a sistematização do conjunto de informações terminológicas de um domínio, é fundamental o uso de ferramentas computacionais para a extração de termos. Este trabalho apresenta a avaliação de métodos de extração automática de termos (EAT) das abordagens estatística, lingüística e híbrida para a construção de ontologias. A avaliação é feita com uma lista de referência com termos do domínio da Ecologia, usando as métricas de precisão e revocação. A OntoEco prevê três subdomínios da Ecologia: Ecossistemas, Populações e Comunidades. Para a extração das unidades lexicais ontológicas, confeccionamos um córpus da Ecologia – o CórpusEco. Após a finalização do delineamento da ontologia, em classes, subclasses e instâncias os dados foram armazenados na ferramenta computacional Protégé-2000.*

## 1. Introdução

Ontologias têm sido utilizadas para a representação de informações que veiculem um entendimento semântico comum de situações variadas do mundo real. Na Web, o uso de ontologias pode fornecer uma base de informações comum e padronizada, englobando conceitos-chave que podem ser utilizados por serviços requisitados para cada situação particular. Em comércio eletrônico, por exemplo, o conjunto de informações oferecido pela ontologia pode ser utilizado para unificar e integrar definições de produtos oferecidos pelos mais variados pontos de venda, com um formato padrão e único. Além disso, as ontologias podem ser utilizadas em sistemas de recuperação da informação para melhorar a precisão e revocação (*recall*) das buscas.

Ontologias são consideradas um recurso importante para muitas aplicações terminológicas como tradução de textos especializados, escrita técnica e criação de dicionários. Em particular, para o trabalho terminográfico as ontologias são fundamentais para o desenvolvimento de recursos como os glossários e dicionários, pois permitem, por exemplo, a construção de definições de uma forma sistemática.

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 córpus. Ainda que o critério semântico seja adequado, a extração manual é lenta, sujeita à subjetividade e à omissão de termos importantes. Este cenário tende a mudar com projetos como o *e-Termos* (Almeida et al, 2006).

Este trabalho apresenta a avaliação de métodos de extração automática de termos (EAT) a partir de córpus para a construção de ontologias. A avaliação é feita com uma lista de referência com termos do domínio da Ecologia, usando as métricas de precisão e revocação.

A *OntoEco* prevê três subdomínios da Ecologia: *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. Esses últimos foram etiquetados manualmente, contendo informações morfossintáticas e informações semânticas concernentes à *Estrutura Qualia* do Léxico Gerativo (LG) de Pustejovsky (1995) e implementados computacionalmente na ferramenta computacional Protégé-2000.

## 2. Extração Automática de Termos (EAT): especificação

Termos são unidades lingüísticas, isto é, palavras ou combinações de palavras designando conceitos ou entidades de um campo altamente especializado da atividade humana. Uma coleção de termos, relacionada com uma área de pesquisa (ou domínio) em particular, usualmente forma um sistema conceitual coerente conhecido como *terminologia* (Bolshakova, 2001). Termos compostos, que correspondem a duas ou mais unidades lexicais, são menos propensos a ambigüidade do que termos simples e aparecem em maior quantidade nos textos especializados, e são mais simples de se extrair. Termos compostos são os preferidos dos métodos de extração automática (Estopà Bagot, 1999).

O crescimento explosivo de dados do tipo texto disponíveis na Web propicia a criação de novos termos e alterações nos seus significados, principalmente, em áreas dinâmicas tais

como a Ciência da Computação. Dado que o desenvolvimento de terminologias é um trabalho difícil quando realizado manualmente, lingüistas computacionais, lingüistas aplicados, tradutores, intérpretes, jornalistas científicos têm se interessado pela extração automática de terminologias de textos. A extração automática de terminologias (EAT) tem sido de grande interesse para todos os tipos de aplicações do Processamento de Línguas Naturais (PLN) que trabalham com domínios especializados e que, conseqüentemente, necessitam de um vocabulário especial.

O gargalo da EAT é a sua avaliação, pois exige a opinião de especialistas, sendo esse processo caro e demorado. Por outro lado, contar com recursos como glossários ou dicionários, isto é, com listas de referências, também traz seus riscos, uma vez que tais recursos são incompletos, dada a constante produção de novos termos. Uma saída pode ser o uso de outras medidas, além das tradicionais medidas de precisão e revocação, como a medida de perplexidade, que mede quão bem um modelo prediz algum dado. Em PLN usa-se perplexidade para comparar a predição de modelos diferentes de língua sobre um córpus (Pantel and Lin, 2001). Outra alterrnativa seria uma avaliação em dois estágios: primeiramente envolvendo várias medidas e, finalmente, os especialistas (Ha, 2004).

Desde o surgimento do TERMINO[1], considerado o primeiro sistema de extração automática de candidatos a termo, diversos projetos têm sido elaborados com a finalidade de projetar extratores (semi-)automáticos de terminologia de naturezas diferentes. No entanto, mesmo com a grande quantidade de estudos realizados nesta direção, o reconhecimento e a delimitação automática das unidades terminológicas a partir de textos ainda não têm apresentado resultados satisfatórios.

A grande maioria dos documentos técnicos e dos artigos científicos contém termos que são explicitamente ou implicitamente definidos pelos autores. É importante que termos recém introduzidos sejam considerados num processamento automático de textos científicos e tecnológicos, pois tais textos apresentam grande quantidade de termos em uso, porém não inseridos nos dicionários por terem sido introduzidos recentemente - denominados *termos de autor*.

Em um aspecto diacrônico, não existe uma fronteira bem definida entre termos de dicionário e de autor. Usualmente, termos nascem como termos de autor. Conforme vão sendo utilizados em vários textos de um dado campo, suas freqüências crescem e eles se convertem em termos de dicionário. As formas usadas para introduzir um termo de autor em um texto variam, resultando em três tipos diferentes de termos de autor (Bolshakova, 2001): a) o termo é explicitamente definido; b) o termo é indefinido (sua definição está ausente), mas ele é visualmente exposto; c) o termo não é nem definido nem exposto, sendo então escondido.

Essas três formas devem ser consideradas pelos métodos de extração automática. A última delas causa grande dificuldade para certos extratores, em razão de que os extratores geralmente utilizam padrões morfológicos e morfossintáticos para reconhecer e delimitar as unidades terminológicas, e o fato de que tais padrões estruturais serem um filtro bastante permissivo para identificar as unidades terminológicas de um determinado domínio impede que tais extratores delimitem todos os termos dos textos especializados. Dessa forma, se forem utilizados padrões referentes somente à forma da unidade, a maioria dos candidatos a termo

[1] TERMINO foi um dos primeiros sistemas de extração automática de terminologia de conhecimento lingüístico. A versão 1.0 deste sistema foi criada em 1989 para o grupo de Recherche et développement en linguistique computationnelle (RDLC) do Centro ATO (Analyse de textes par ordinateur) da Universidade de Quebec Montreal.

apresentará delimitações errôneas. Por esta razão, os extratores também devem possuir conhecimento semântico a fim de detectar e delimitar automaticamente as unidades especializadas de forma mais exaustiva e precisa.

Todas as unidades léxicas têm uma freqüência associada que corresponde ao número de vezes que elas aparecem em um córpus. A partir desta informação, é possível saber se uma palavra pode ou não ser um termo. Ou seja, substantivos que aparecem mais de um certo número de vezes podem ser considerados termos candidatos; palavras de outras categorias devem ser mantidas a fim de completar o processamento de termos compostos. Existem, porém, estatísticas mais elaboradas para a seleção de candidatos a termos, por exemplo, Informação Mútua, Coeficiente *Log-Likelihood* (Daille, 1996) e Coeficiente *Dice*[2], que serão descritas neste trabalho. Uma das abordagens para a realização da tarefa de extração usa medidas estatísticas – são os *Sistemas Baseados em Estatística*. Outra abordagem encontrada na literatura é a *lingüística,* na qual os sistemas detectam padrões recorrentes de unidades terminológicas complexas, tais como "substantivo–adjetivo" e "substantivo–preposição–substantivo", por exemplo; e a *híbrida*, na qual os sistemas começam a detectar algumas estruturas lingüísticas básicas, tal como expressões nominais, e depois de os termos candidatos terem sido identificados, uma estatística relevante é usada para decidir se eles correspondem a um termo. O inverso também é possível, começando-se com uma lista de candidatos levantados estatisticamente, sendo que a informação lingüística, neste caso, é usada para filtrar termos válidos desta lista.

## 3. Metodologia e Desenvolvimento

### 3.1. O *Córpus-Eco*

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 260.921 ocorrências e está armazenado em uma ferramenta computacional que administra grandes quantidades de dados, o *Folio Views 4.1*. Os textos foram extraídos de partes dos livros "A Economia da natureza" e "Ecologia", da editora Guanabara Koogan, e de revistas, presentes no Córpus de Referência do Projeto Lácio-Web[3]. A lista de referência utilizada para a avaliação dos métodos foi confeccionada com 694 termos das partes dos livros acima extraídos com o critério semântico, além de dois glossários especializados, e mais 1105 termos do Dicionário On-line do Jornal do Meio Ambiente[4]. Após eliminação de termos duplicados e intersecção com o *CórpusEco*, a lista totalizou 520 termos.

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.

[2] http://www.d.umn.edu/~tpederse/Group01/bsp.txt
[3] http://www.nilc.icmc.usp.br/lacioweb/
[4] http://www.jornaldomeioambiente.com.br/

### 3.2. As abordagens da EAT

Em seguida, partimos para a extração automática dos candidatos a termos e avaliamos métodos simples de EAT para uni, bi e trigramas das três abordagens existentes, a saber: a *estatística*, a *lingüística* e a *híbrida* para a elaboração da *OntoEco*.

Esses métodos, num total de quinze, foram desenvolvidos no projeto *ExPorTer*[5] e empregam recursos simples como:

(a) uma stoplist para eliminar palavras como advérbios, pronomes e artigos;

(b) padrões sintáticos para os termos do domínio, por exemplo, <*substantivo adjetivo*>, <*substantivo preposição adjetivo*>, levantados após a aplicação do etiquetador *Part-Of-Speech* MXPOST treinado no NILC[6] (precisão de 97%) ;

(c) uma *lista de expressões* e *palavras* características de definições, descrições, classificações como "definido(a)(s) como", "caracterizado", "chamado(a)(s)", "significa", entre outras que são concentradoras de termos.

### 3.3. Métodos da Abordagem Estatística

As medidas estatísticas utilizadas nesse trabalho são quatro: *Freqüência*, *Log-likelihood*, *Informação Mútua* e *Coeficiente Dice*, implementadas no pacote para a extração de n-gramas NSP[7] (*N-gram Statistics Package*), com objetivo de eleger a melhor medida estatística para a extração automática de unigramas, bigramas e trigramas (termos que apresentam, respectivamente, o número de *tokens* igual a 1, 2 e 3). Entende-se como a melhor medida estatística aquela que apresentar a maior precisão, embora tenham sido calculadas a revocação e a medida F.

Após a geração das listas de freqüência para unigramas, bigramas e trigramas, foram realizados os cálculos da informação mútua, do *log-likelihood* e do *coeficiente dice* para bigramas, que utilizam como entrada a lista de freqüência gerada para os bigramas do córpus. Em seguida, foram realizados os cálculos da informação mútua e do *log-likelihood* para trigramas, que utilizam como entrada a lista de freqüência gerada para os trigramas encontrados no córpus. Para unigramas somente foi realizado o cálculo da freqüência, pois é a única medida para unigramas disponível no pacote NSP.

O método estatístico usando a medida de *Freqüência* para unigramas teve seu corte estabelecido em 20, sendo a sua Precisão de 9,48% e Revocação de 34,27%. Para bigramas usando a medida de *Freqüência* o corte foi de 18, para a medida de *Informação Mútua* de 0,0097, para a *Log-likelihood* de 53,0782 e para o *Coeficiente Dice* de 0,1689, sendo que os quatro valores de Precisão foram 20,31% e os quatro de Revocação, 14,44%. Para os trigramas, a medida de *Freqüência* utilizou corte de 18, para a *Informação Mútua* de 0,0066 e de *Log-likelihood* de 113, 2980, os três com Precisão de 2,41% e Revocação de 10,23%. Deve-se ressaltar que mantivemos os mesmos cortes resultantes de análises realizadas no projeto ExPorTer, embora o córpus lá utilizado tenha o dobro do tamanho do *CórpusEco*. Também mantivemos a mesma *stoplist*, embora os gêneros tratados fossem diferentes. Acreditamos que essas decisões possam ter afetado as precisões dos métodos estatísticos e híbridos.

[5] http://www.nilc.icmc.usp.br/nilc/projects/termextract.htm
[6] http://www.nilc.icmc.usp.br/nilc/projects/mestradorachel.html
[7] http://www.d.umn.edu/~tpederse/nsp.html

### 3.4. Métodos da Abordagem Lingüística

O método lingüístico implementado neste artigo baseia-se em expressões lingüísticas e indicadores estruturais, bem como nos padrões morfossintáticos dos termos do domínio de Ecologia. Dessa maneira o método lingüístico baseou-se tanto no trabalho de Heid et al (1996), no sentido de realizar um pré-processamento lingüístico no córpus utilizado e posteriormente a realização de consultas sobre o mesmo, quanto no trabalho de Klavans e Muresan (2000; 2001a; 2001b), no sentido de realizar uma busca por expressões lingüísticas e indicadores estruturais que introduzem definições e os termos definidos. Entretanto, este trabalho não se assemelha totalmente ao método proposto por Heid et al. em razão do córpus não ter sofrido o processo de lematização. Por outro lado, o método aqui implementado, chamado de *ExPorTer_lingüístico*, fugiu um pouco da proposta feita por Klavans e Muresan, em razão de não terem sido realizadas buscas somente de expressões de definições, mas também de classificações, descrições e outras que concentram termos, além de não ter sido utilizado um módulo de análise gramatical, responsável por identificar definições introduzidas por fenômenos lingüísticos mais complexos, tais como anáforas e apostos. As precisões para uni, bi e trigramas foram 2,74%, 1,31% e 0,89% e as Revocações, 89,18%, 62,22% e 82,95%.

Baseando-se, a princípio, no trabalho de Klavans e Muresan (2000; 2001a; 2001b), realizamos um levantamento de expressões lingüísticas e indicadores estruturais que geralmente vêm acompanhados de definições, descrições, classificações e de outros tipos de orações que concentram termos, para identificar o termo ou termos que aparecem nelas (Tabela 1). Essas expressões e indicadores estruturais foram levantados a partir do *CórpusEco*, com sua lista de referência, desenvolvido no Projeto Bloc-Eco[8].

**Tabela 1. Padrões morfossintáticos**

| Padrões morfossintáticos utilizados[9] | | |
|---|---|---|
| **Para unigramas** | **Para bigramas** | **Para trigramas** |
| n / np / adj / verb | n_adj / n_n / adj_n / adj_adj<br>n_adv | n_prep_n / n_prep_np / n_n_adj / n_adj_adj<br>n_prep_adj |

### 3.5. Métodos da Abordagem Híbrida

Para a abordagem híbrida, foi gerado um conjunto de orações do córpus, aqui chamado de subcórpus, que apresentassem as expressões dos padrões lingüísticos definidas no método lingüístico, de maneira que cada oração é impressa no subcórpus somente uma vez, independentemente do número de expressões que pode apresentar. Este procedimento representou a "parte lingüística" do método híbrido. O subcórpus de saída, constituído pelas orações que apresentaram alguma expressão lingüística, é tomado como entrada para o pacote NSP, representando assim a parte estatística deste método; chamamos o método híbrido de *ExPorTer_híbrido,* subcategorizado pela estatística utilizada.

A freqüência, única medida estatística para unigramas encontrada no pacote NSP, foi calculada para os unigramas do subcórpus, utilizando-se o mesmo corte determinado na abordagem estatística, ou seja, 20, sendo a Precisão de 12,76% e a Revocação de 23,25% . O cálculo da *Freqüência* também foi efetuado para os bigramas e os trigramas do subcórpus, realizando o corte de 18 na *Freqüência,* tanto para bigramas quanto para trigramas, como estipulado na abordagem estatística, com Precisão de 41,18% e de 18,75% e Revocação de

[8] http://www.nilc.icmc.usp.br/nilc/projects/bloc-eco.htm
[9] n = nome; np = nome próprio; adj = adjetivo; verb = verbo; adv = advérbio

7,78% e 3,41%, respectivamente. A medida de *Informação Mútua* para o brigramas foi calculada sem corte, com Precisão de 1,68% e Revocação de 65%.

### 3.6. A *OntoEco*

A *OntoEco* foi implementada no Protégé-2000. Essa ferramenta foi desenvolvida para diferentes linguagens para a Web Semântica, entre as quais RDF e RDF Schema, que permitem a estruturação de informações de um domínio específico e possibilitam a comunicação, por meio de um vocabulário comum, entre agentes de software e páginas da Web. 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. Esse modelo é baseado em *frames*, usa a arquitetura de *metaclasse*, ou seja, um *template* que é usado para definir novas classes em uma ontologia, e possibilita a especificação de herança múltipla e de classes abstratas. Em sua implementação, 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. 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 classes e subclasses Dessa forma, temos como subclasses da superclasse CLASSES: INTERAÇÃO; POPULAÇÃO; COMUNIDADE; ECOSSISTEMA; ENERGIA; entre outras. Após a distribuição dos itens lexicais na estrutura ontológica delineada, definimos e mapeamos as relações semânticas existentes entre eles presentes nos papéis da *Estrutura Qualia* (Pustejovsky, 1995). Cada unidade lexical terminológica ativa no campo da Ecologia foi delineada a partir de vários campos de valor, distribuídas em tabelas.

Após a distribuição dos itens lexicais na estrutura ontológica delineada, definimos e mapeamos as relações semânticas existentes entre eles, presentes nos papéis da *Estrutura Qualia* de Pustejovsky. Cada unidade lexical terminológica ativa no campo da Ecologia foi delineada a partir de vários campos de valor, conforme a Tabela 2.

**Tabela 2: Campos de valor**

| | |
|---|---|
| **SemU:** | Unidade Semântica – unidade lexical ou ontológica. |
| **Tipo:** | Subcategoria a que o termo pertence. |
| **Supertipo:** | Categoria a que o termo pertence. |
| **Domínio:** | O domínio com o qual trabalhamos, no caso, ecologia. |

| | |
|---|---|
| **Formal:** | Relações semânticas existentes a Estrutura *Qualia* |
| **Agentivo:** | |
| **Constitutivo:** | |
| **Télico:** | |
| **Glossário:** | Definição do termo. |
| **Exemplo:** | Exemplo retirado do nosso. córpus |
| **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. Resultados e Análise

Como os valores de precisão foram baixos para os quinze métodos calculados (veja resumo na Tabela 3), analisamos o total de candidatos a termos extraídos para cada abordagem que são efetivamente unidades terminológicas nos 150 primeiros termos dos métodos com melhor precisão (Tabela 4). A idéia era verificar se pelo menos os métodos conseguem distinguir os termos com melhores escores, isto é, os primeiros das listas de candidatos. No caso de empate para a precisão, a lista da freqüência foi escolhida.

**Tabela 3: Precisão para cada um dos 15 métodos de extração de termos utilizados**

| Abordagem | Métodos | Precisão | Revocação |
|---|---|---|---|
| **Estatística** | **Freqüência – unigramas** | **9,48** | **34,27** |
| | **Freqüência – bigramas** | **20,31** | **14,44** |
| | **Log-Likelihood – bigramas** | **20,31** | **14,44** |
| | **Informação mútua – bigramas** | **20,31** | **14,44** |
| | **Dice – bigramas** | **20,31** | **14,44** |
| | **Freqüência – trigramas** | **2,41** | **10,23** |
| | **Informação mútua – trigramas** | **2,41** | **10,23** |
| | **Log-Likelihood – trigramas** | **2,41** | **10,23** |
| **Lingüística** | **ExPorTer_lingüístico – unigramas** | **2,74** | **89,18** |
| | **ExPorTer_lingüístico – bigramas** | **1,31** | **62,22** |
| | **ExPorTer_lingüístico – trigramas** | **0,89** | **82,95** |
| **Híbrida** | **ExPorTer_híbrido c/ Freqüência – unigramas** | **12,76** | **23,25** |
| | **ExPorTer_híbrido c/ Freqüência – bigramas** | **41,18** | **7,78** |
| | **ExPorTer_híbrido c/ Freqüência – trigramas** | **18,75** | **3,41** |
| | **ExPorTer_híbrido c/ Informação mútua – bigramas** | **1,68**[10] | **65,0** |

**Tabela 4: Resultados das análises para os 150 primeiros termos dos métodos com melhor precisão**

| 150 candidatos a termos analisados | Abordagem estatística | Abordagem Lingüística | Abordagem híbrida |
|---|---|---|---|
| **unigramas** | 42 | 21 | 45 |
| **bigramas** | 51 | 4 | 19 |
| **trigramas** | 10 | 1 | 3 |
| **TOTAL** | **103** | **26** | **67** |

Foi possível observar, na abordagem estatística, que, em sua maioria, os *unigramas* que não são termos são substantivos flexionados ou não, verbos conjugados ou no infinitivo, adjetivos, advérbios, abreviações ou letras avulsas. Dos 150 candidatos a termos analisados em ordem decrescente, 42 (28%) candidatos verificaram-se como termos efetivos, cuja categoria

[10] Calculada sem corte.

gramatical freqüente foi a do substantivo. Já para os *bigramas*, obtivemos o melhor resultado de toda a extração, ou seja, 51 (34%) termos de combinatória gramatical "Substantivo + Adjetivo". Por sua vez, dos 150 *trigramas* levantados, 10 (6,6%) são efetivamente termos do tipo "Substantivo + Preposição + Substantivo".

Na abordagem lingüística, todos os *unigramas* efetivamente termos são de categoria substantiva, isto é, 21 deles (14%). Os outros candidatos a termos extraídos são do tipo adjetivo, como "dinâmico", "mortos", "caídos" ou do tipo verbo flexionado e no infinitivo, como "mudando", "determinam", "dividem", "preferem" e "viver", "mascarar" respectivamente, ou ainda palavras formadas por hífen, tais como "focas-elefante", "besouros-de-farinha". Já para os *bigramas* que se caracterizaram de fato como termos, a combinatória gramatical foi a do tipo "Substantivo + adjetivo", no singular ou no plural. A extração aponta para outras combinatórias que não resultaram na efetivação de termos, tais como: "Pronome + substantivo" (quaisquer registros), "Adjetivo + Substantivo" (tremendos aumentos), "Advérbios + Substantivos" (muitas populações). Por sua vez, a extração de *trigramas* foi extremamente baixa: somente um candidato a termo caracterizou-se como unidade terminológica, cuja combinatória gramatical foi "Substantivo + Preposição + Substantivo". As outras combinatórias foram do tipo: "Substantivo + Adjetivo + Adjetivo" (troncos mortos caídos) e "Substantivo + Preposição + Substantivo" (indivíduos à densidade, noite por satélite, cidades de países, árvores das florestas), essa última com uma alta freqüência de ocorrência.

Na abordagem híbrida, obtivemos o melhor desempenho do extrator para a captura de *unigramas*, a saber: 45 termos (30%), cuja categoria gramatical mais freqüente foi a de substantivo, flexionado ou não em número. As classes gramaticais dos outros candidatos a termos que não se efetivaram foram do tipo "abreviação" (fig), "verbal" (pode, podem), "adjetiva" (grande, maior), "pronominal" (tais). Para os *bigramas*, a combinatória de todos as unidades terminológicas foi "Substantivo + Adjetivo". Os candidatos a temos que não se efetivaram apontam para as combinatórias "Adjetivo + Substantivo" (novas espécies) e "Substantivo + Adjetivo"( populações naturais). Já os trigramas efetivamente termos são do tipo "Substantivo + Preposição + Substantivo" e os que não se caracterizaram como termos também, tais como *número de espécies, número de indivíduos, ponto de vista*.

## 5. Considerações Finais

Em um primeiro momento, a extração automática de candidatos a termos alimentou o delineamento, propriamente dito, da estrutura arbórea da ontologia na medida em que nos forneceu unidades terminológicas que se caracterizaram como classes ou subclasses, tais como: *população, comunidade, energia, área, ecologia, tabela de vida*, entre outros. Entretanto, a grande utilidade da extração automática foi, justamente, para a seleção das unidades lexicais para a ontologia, que podem ser implementadas tanto como subclasses quanto como instâncias.

Dos resultados obtidos e das análises feitas, embora os métodos tenham alcançado baixa precisão, constatamos que a abordagem híbrida foi a que nos trouxe melhores resultados no que concerne à qualidade da extração, embora a quantidade de candidatos a termos extraídos tenha sido mediana, como as outras duas abordagens. De fato, para unigramas, a abordagem híbrida teve o melhor desempenho, ao passo que, numa análise geral, a abordagem estatística teve seus resultados superiores para bigramas e trigramas, sendo inferior somente para os unigramas em relação à abordagem híbrida. Já a abordagem lingüística carece de maior detalhamento ainda para que possamos obter resultados melhores em uma próxima tentativa de extração. Embora os valores dessa nossa avaliação não tenham sido altos, eles revelam que a

aplicação de métodos automáticos para a captura de unidades lexicais, após serem refinados e especializados para os gêneros tratados, assim como para o tamanho de córpus em uso, poderá auxiliar de maneira eficaz o trabalho de extração do terminólogo, dado que são capazes de identificar diversos itens lexicais considerados efetivamente termos, de forma bastante rápida. Em uma extração de termos com base no critério semântico, essa captura de unidades lexicais demandaria tempo e seria certamente muito mais lenta e subjetiva, se comparada à máquina.

## Agradecimentos

Os autores agradecem o suporte do CNPq e da CAPES para o desenvolvimento desta pesquisa.

## Referências

Almeida, G. M. B., Oliveira, L. H. M., Aluisio, S. M. "A Terminologia na era da Informática". Ciência e Cultura (SBPC), v.58, p.42 - 45, 2006.

Bolshakova, E. "Recognition of Author's Scientific and Technical Terms". LNCS 2004, 2001 p. 281-90.

Bourigault, D. "Surface grammatical analysis for the extraction of terminological noun phrases". In Proceedings of the 14th International Conference on Computational Linguistics, COLING 1992, 1992. p. 977-981.

Daille, B. "Combined approach for terminology extraction: lexical statistics and linguistic filtering". PhD thesis, University of Paris 7, 1994.

Estopà Bagot, R. "Extracció de terminologia: elements per a la construcció d'un SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Significació Especialitzada)". Tese de Doutorado. Universidade Pompeu Fabra, 1999.

Ha, L.A. "Co-training applied in automatic term extraction: an experiment". In: 7th Annual CLUK Research Colloquium, University of Birmingham, Jan 2004. Disponível em http://www.cs.bham.ac.uk/~mgl/cluk/titles.html.

Heid, U.; Jau, S.; Krüger, K.; Hohmann, A. "Term extraction with standard tools for córpus exploration". In: 4th International Congress on Terminology and Knowledge Engineering, Wien. August, 1996.

Klavans, J. L.; Muresan, S. "DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from Online Text". In: Proceedings of AMIA, 2000.

Klavans, J. L.; Muresan, S. "Evaluation of DEFINDER: A System to Mine Definitions from Consumer-oriented Medical Text". In: Proceedings of JCDL, 2001a.

Klavans, J. L.; Muresan, S. "Evaluation of the DEFINDER System for Fully Automatic Glossary Construction". In: Proceedings of AMIA, 2001 b.

Pantel, P.; Lin, D. A statistical corpus-based term extractor. In: E. Stroulia e S. Matwin (Ed.), AI 2001, Lecture Notes in Artificial Intelligence, Springer-Verlag, 2001. p. 36–46.

Pustejovsky, J. The Generative Lexicon. Cambridge: The MIT Press, 1995.