Hierarquias de Conceitos para um Ambiente Virtual de Ensino Extraídas de um *Corpus* de Jornais Populares

Maria José Bocorny Finatto[1], Lucelene Lopes[2], Renata Vieira[2], Aline Evers[3]

[1,3]Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul (UFRGS), [1]Pós-Doutoranda ICMC-USP
Av. Bento Gonçalves, 9500 – 91.540-000 – Porto Alegre – RS – Brasil

[2]Programa de Pós-Graduação em Ciência da Computação, Faculdade de Informática – Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)
Av. Ipiranga, 6681 – 90.619-900 – Porto Alegre – RS – Brasil

mfinatto@terra.com.br, lucelene.lopes@pucrs.br, renata.vieira@pucrs.br, aline.evers@gmail.com

Abstract. *In this paper we present conceptual hierarchies automatically obtained from the popular newspaper Diário Gaúcho. The hierarchies were generated by means of ExATOlp tool through the extraction of noun phrases considered concept candidates by applying linguistic and statistic approaches. When accessed in a virtual learning environment, the hierarchies became a differentiated resource for vocabulary teaching and for the journalistic language patterns research in newspapers geared to audiences with lower education.*

Resumo. *Neste artigo são apresentadas hierarquias de conceitos geradas automaticamente a partir do jornal popular Diário Gaúcho. As hierarquias são obtidas pela ferramenta ExATOlp através da extração de sintagmas nominais considerados candidatos a conceitos mediante combinação de técnicas linguísticas e estatísticas. Acessadas em um ambiente virtual de aprendizagem, as hierarquias oferecem recurso diferenciado para o ensino de vocabulário e pesquisa sobre padrões da linguagem jornalística voltada para públicos de menor escolaridade.*

1. Introdução

Apresenta-se aqui um modelo de hierarquia automática de conceitos produzida a partir de uma amostra de textos do jornal popular *Diário Gaúcho* (doravante DG). A hierarquia, em diferentes formatos, é gerada pela ferramenta ExATOlp, acessada gratuitamente em <http://www6.ufrgs.br/textecc/index_porpopular.php>, em que se oferece um ambiente virtual de aprendizagem. Seu acesso é um resultado parcial da pesquisa *Padrões do Português Popular Escrito – Projeto PorPopular*, dedicada a reconhecer o vocabulário e especificidades do texto de jornais populares brasileiros, cujos leitores preferenciais têm poder aquisitivo e graus de escolaridade menores se comparados aos de leitores de jornais tradicionais. A investigação tem apoio do CNPq e conta com a colaboração de pesquisadores do grupo de pesquisa de Processamento de Linguagem Natural (doravante PLN) da PUCRS <http://www.inf.pucrs.br/~linatural/>.

A ferramenta ExATOlp [Lopes *et al*. 2009] retira do *corpus* DG sintagmas nominais candidatos a conceitos (portadores de informação) via combinação de técnicas de extração baseadas em princípios de análise linguística e estatística. Para os itens extraídos, a ferramenta organiza uma hierarquia e uma lista dos contextos verbais nos quais cada conceito foi encontrado. A hierarquia gerada pode ser consultada em diferentes formatos, com maior ou menor detalhamento de informações, dependendo da opção de visualização do usuário, podendo auxiliá-lo na percepção da organização do conteúdo do *corpus* em foco.

Este trabalho está assim organizado: na Seção 2, está a caracterização do quadro geral da pesquisa PorPopular, do *corpus* e do jornal **DG**; na Seção 3, descrevem-se a ferramenta ExATOlp**,** o processo de geração das hierarquias e traz-se uma pequena amostra de hierarquias obtidas; na Seção 4, exemplificam-se aplicações das hierarquias, apresentam-se limitações do trabalho e são indicadas possibilidades para trabalhos futuros.

2. A Pesquisa PorPopular e o *Corpus* Reunido

A pesquisa PorPopular é de natureza linguística, centrada em aspectos lexicológicos e discursivo-gramaticais. Adota a metodologia e ponto de vista da Linguística de *Corpus* (doravante LC) [Berber Sardinha 2004] e enfatiza a descrição com base estatística partindo de acervos textuais em formato digital. Esses acervos são denominados *corpus*/*corpora* e servem tanto para estudos da linguagem quanto para produção de alguma aplicação computacional. A LC concebe a língua como um sistema probabilístico de combinatórias, de modo que não se pode observar as palavras isoladas do vocabulário de um texto ou *corpus*. Isso porque, conforme Stubbs [2001], o conhecimento humano da linguagem e dos textos não se restringe a um conhecimento das palavras isoladas, mas é integrado fundamentalmente pelo conhecimento de combinatórias possíveis e pelo conhecimento cultural que essas combinatórias frequentemente contêm. Assim, os principais focos da pesquisa PorPopular são a descrição e o estudo de padrões associativos do vocabulário para o que são utilizados também métodos, abordagens e produtos do PLN. Visa-se uma caracterização do léxico e da linguagem posta em um texto que é feito, em tese, de um modo mais simplificado, para ser compreendido com facilidade por pessoas com grau de escolaridade relativamente baixo. Na etapa atual da investigação, até o final de 2011, utiliza-se como *corpus* apenas textos coletados do jornal DG, versão impressa, publicado em Porto Alegre-RS, produzido pelo grupo RBS.

O DG impresso não oferece assinatura e é vendido apenas em bancas da cidade de Porto Alegre e região metropolitana. Foi escolhido para estudo em função de sua grande tiragem (168 mil exemplares/dia) e de sua longa existência (11 anos), além de já ter sido objeto de pesquisas na área do Jornalismo [Amaral 2006; Bernardes 2004]. Entretanto, ainda não havia sido explorado no âmbito dos Estudos da Linguagem ou do PLN. Seu número de leitores supera o de jornais da mesma cidade dirigidos a públicos mais tradicionais distribuídos em todo o Estado do Rio Grande do Sul. No *corpus*, estão arquivos de edições completas, coletadas em dias alternados da semana, do jornal impresso em formato somente texto (.txt) do ano de 2008, com pequenas amostras de 2009 e de 2010. O material em formato .txt pode ser compartilhado com pesquisadores e boa parte já se acessa via expressões de busca na seção **Experimente** do *site* do Projeto PorPopular. Materiais e recursos associados a esse *corpus* já estão sendo utilizados para atividades de ensino de língua portuguesa, ensino de vocabulário, como também integram proposta de um dicionário de português como língua estrangeira, dado um caráter *a priori* mais simples dos textos e da linguagem do DG. Entre as aplicações disponíveis, destaca-se neste trabalho o recurso **Hierarquias de Conceitos**, compreendido como uma representação ontológica do conteúdo dos textos reunidos, conforme detalhado na Seção 3 a seguir.

3. A Ferramenta ExATOlp e o Processo de Geração de Hierarquias

O processo de aquisição de hierarquias da ferramenta ExATOlp [Lopes *et al*. 2009] possui base linguística, uma vez que o ponto de partida é a extração de termos (sintagmas nominais), partindo de um *corpus* previamente anotado pelo *parser* PALAVRAS [Bick 2000]. Os sintagmas extraídos passam por uma análise detalhada em que diversas heurísticas são utilizadas para descartar ou aprimorar a qualidade informacional dos sintagmas obtidos.

Em resumo, o processo de extração de termos aplicado consiste em considerar todos os termos (*multi-token*) anotados como sintagmas nominais, ou *tokens* anotados pelo PALAVRAS como sujeito, objeto, complemento de sujeito ou objeto de orações. Em seguida, aplica-se um conjunto de heurísticas propostas na

ferramenta ExATOlp**.** A aplicação dessas heurísticas foi comparada com uma abordagem estatística em [Lopes *et al*. 2010] e o resultado da comparação mostrou a superioridade de cerca de 15% de precisão do método linguístico frente à abordagem puramente estatística. As heurísticas são as seguintes:

- ✓recusar termos que possuem numerais na sua forma textual ou numérica, por exemplo, "sete meses"; "8 horas", *etc*.;
- ✓recusar termos que tenham outros caracteres além de letras acentuadas ou não, por exemplo, "%", "\", "/", "@", *etc*.;
- ✓recusar termos que tenham como núcleo palavras identificadas sintaticamente com outras categorias além de substantivos comuns, próprios, adjetivos ou verbos no particípio passado, por exemplo: "**Eles mesmos** têm de construir um muro". Nesse caso, esse sintagma seria recusado, pois o núcleo é um pronome;
- ✓remover artigos contidos nos termos, por exemplo, "**O Incrível Hulk**" será salvo como "**Incrível Hulk**";
- ✓remover pronomes contidos nos termos, por exemplo, "Ele foi para **sua casa de praia**", o sintagma salvo será **casa de praia**;
- ✓criar termos implícitos detectados pelo uso de conjunções entre adjetivos, por exemplo, "As **pessoas espertas** ou sábias...", nesse caso, dois termos serão salvos: "**pessoas espertas**" e "**pessoas sábias**";
- ✓criar termos genéricos pela remoção sucessiva de adjetivos, por exemplo, do sintagma "**O perigo das doenças virais hemorrágicas**", cria-se mais três termos "**perigo das doenças virais**", "**perigo das doenças**" e "**perigo**"; e
- ✓replicar termos que são sujeitos de mais do que um predicado, por exemplo, "**Pacientes idosos** compram e tomam remédios caros" replica os termos, desdobrando a frase em duas: "**Pacientes idosos** compram **remédios caros**" e "**Pacientes idosos** tomam **remédios caros**".

Os termos considerados segundo a aplicação das heurísticas são salvos em listas de acordo com o número de palavras que os compõem, ou seja, unigramas (uma palavra), bigramas (duas), trigramas (três), *etc*. Além da anotação sintática, o PALAVRAS realiza anotação semântica dos *tokens* de acordo com um conjunto de 16 categorias. Com base nessas etiquetas semânticas, os termos são organizados em uma estrutura de árvore hiperbólica, apresentada na Figura 1.

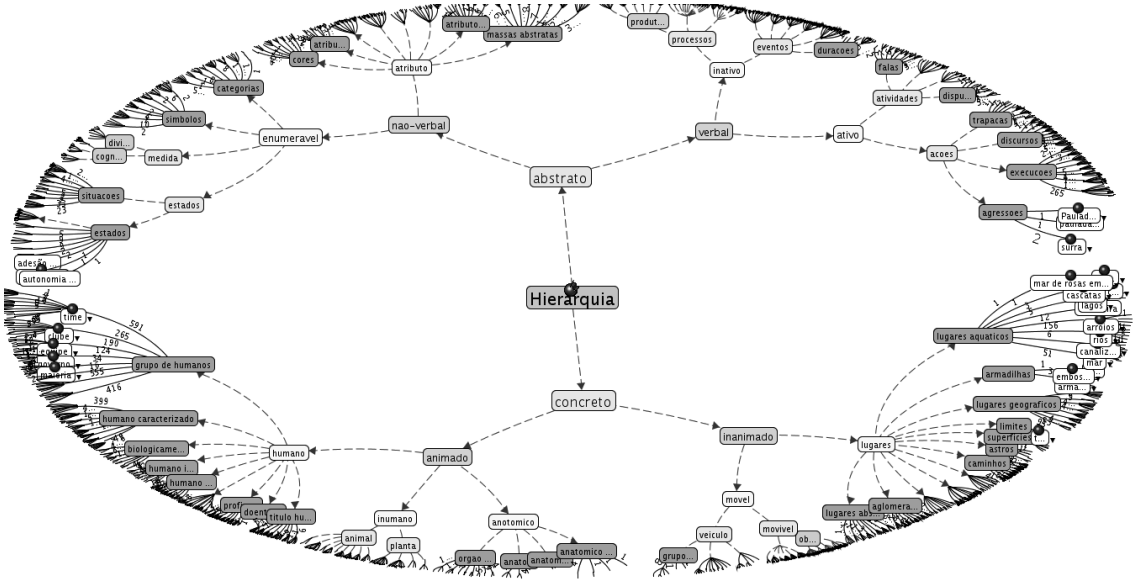

Figura 1. Visão geral do primeiro nível (categorias semânticas) da Hierarquia

Cada uma das categorias semânticas possui diversas etiquetas semânticas, totalizando 174 etiquetas distintas. Dessa forma, cada um dos termos extraídos (que correspondem às folhas da árvore) é relacionado a uma dessas etiquetas. A hierarquia baseada em categorias e etiquetas semânticas produz um primeiro nível qualificado. Em seguida, é feita uma nova estruturação a partir de cada etiqueta semântica, organizando o segundo nível da hierarquia. Nesse segundo nível, os termos (sintagmas nominais) são divididos em subgrupos que possuem o mesmo núcleo. Em seguida, é feita uma organização dos termos de cada subgrupo, considerando hiperônimo de um termo o termo que estiver contido nele.

Do *corpus* desta pesquisa, que possui 1.137.847 palavras, foram extraídos inicialmente 176.287 sintagmas nominais. Após aplicação das heurísticas de refinamento, 83.107 foram rejeitados. Os sintagmas aproveitados foram organizados em uma hierarquia em dois níveis: no primeiro, o nível das categorias semânticas; e, no segundo, nível hierarquia estrutural do termo. Por exemplo, o nodo **mulher** liga-se ao nodo **humano** num primeiro nível. Dessa forma, coloca-se hierarquicamente em relação a outros nodos semanticamente próximos, como **homem**, por exemplo. Partindo da estrutura lexical, são considerados filhos do nodo **mulher** os nodos que correspondem aos sintagmas **mulher de minha vida – mulheres nascidas este dia – mulheres que vivem em Rio de Janeiro – mulheres escravas**, etc.

A saída de todo o processo de geração de hierarquias realizado pela ferramenta ExATOlp são arquivos no formato XML. Arquivos desse tipo podem ser utilizados como formato de entrada de vários sistemas com diferentes aplicações. Uma delas é a visualização hiperbólica da hierarquia para analistas do discurso/texto, como nesta pesquisa com o DG. Nessa modalidade, é possível a manipulação dos termos por árvores interativas que facilitam tanto uma visão ampla dos nodos da hierarquia (Figura 2) quanto uma visão mais restritiva somente do segundo nível (Figura 3). Adicionalmente, as hierarquias possuem um terceiro nível, que corresponde aos verbos associados aos nodos, indicando-se a ocorrência do termo e seu respectivo contexto verbal. A Figura 4 traz a hierarquização do termo **arroios** e seus respectivos verbos. A hierarquia estabelece a seguinte ordem: hierarquia > concreto > inanimado > lugares > lugares aquáticos > **arroios** (que está relacionado com os verbos: **limpar**, **alagar**, **ter**, **transbordar**, **encontrar**, **apresentar**, **ser**). Note-se que o termo **arroios** possui hipônimos (**arroio local**, **arroio que corta a restinga**, *etc*.) que só aparecem quando não se observa seus contextos verbais (Figura 3).

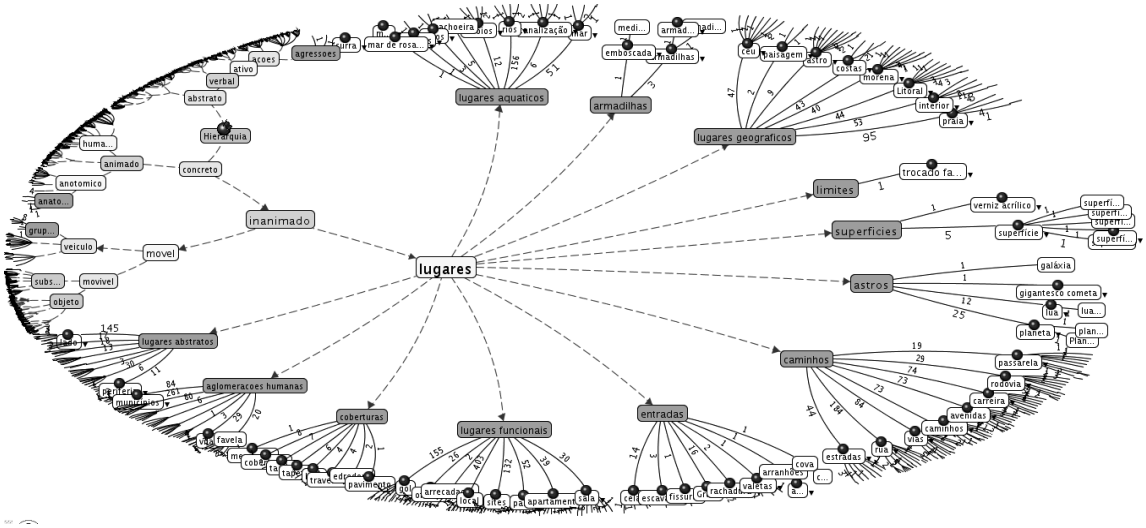

Figura 2. Visão focada na etiqueta semântica <lugares> para o *corpus* DG

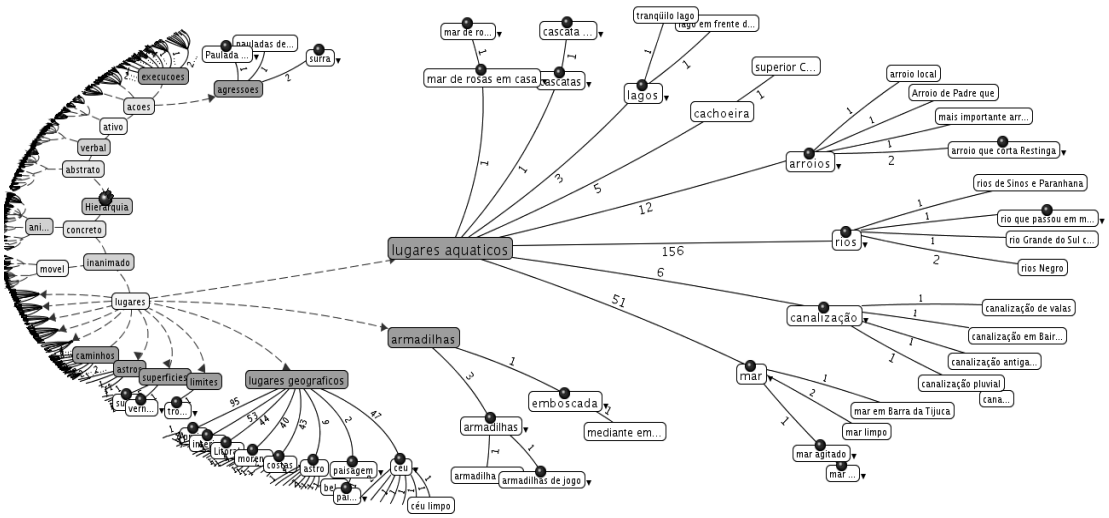

Figura 3. Visão do segundo nível da Hierarquia para <lugares> no *corpus* DG

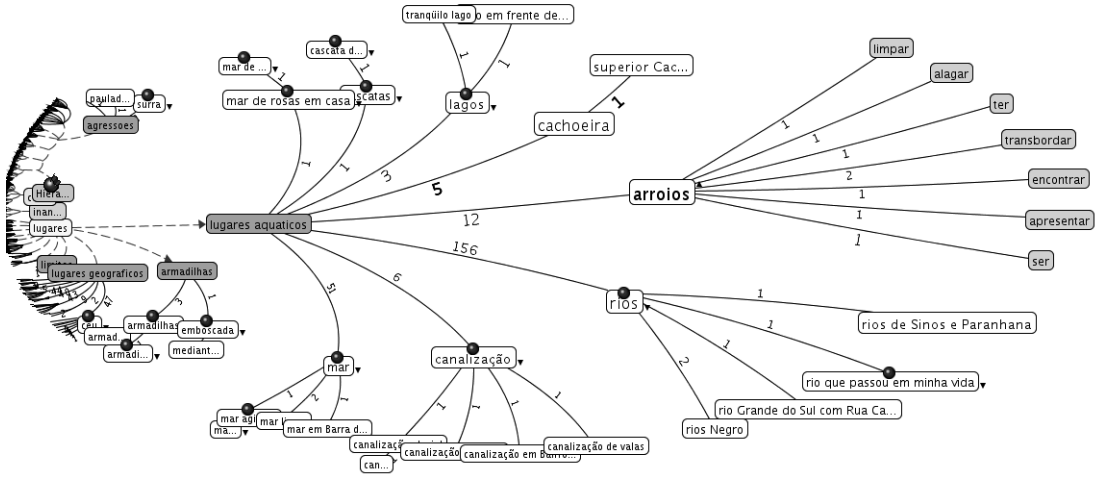

Figura 4. Visualização dos verbos na Hierarquia de conceitos do *corpus* DG

4. Utilizações das Hierarquias Geradas, Limitações e Trabalhos Futuros

A aplicação mais direta das hierarquias produzidas pela ferramenta ExATOlp é a geração de dados para auxiliar a reconhecer padrões de organização de conteúdo de um *corpus*, o que permite também um aproveitamento para o reconhecimento de padrões vocabulares e textuais em diferentes cenários comunicativos e em distintas categorias de textos (redações escolares, textos jornalísticos, textos científicos, *etc*.) que conformem um dado *corpus*. No caso do jornalismo popular, um segmento ainda pouco estudado entre os pesquisadores de Comunicação Social/Jornalismo e de Letras/Linguística, o auxílio é bastante importante, sobretudo pelo tipo e desenho de informação que o recurso oferece ao investigador da linguagem em foco.

Como uma outra aplicabilidade futura associada também ao *corpus* DG, embora indireta, cita-se a construção de um dicionário *on-line* de português para estrangeiros, projeto em fase inicial ao Projeto TEXTECC <http://www6.ufrgs.br/letras/dicionarioportuguesle/>. Com novas hierarquias geradas desse

corpus DG, segmentado por temáticas ou assuntos do jornal, esse dicionário *on-line* poderia dispor seus verbetes por nodos em vez de privilegiar uma ordem estritamente alfabética, ou até mesmo oferecer definições e verbetes dinâmicos, permitindo visualização da rede de relações estabelecida através da saída da ferramenta. Por meio da apresentação da hierarquia de conceitos, o usuário teria uma visão amplificada, por exemplo, de um determinado domínio semântico e que conseguiria estabelecer ou reconstruir as relações de forma mais dinâmica e proveitosa, construindo seu conhecimento de língua de forma mais concreta. Esse recurso também teria bom aproveitamento em dicionários para falantes nativos do português.

O trabalho realizado também poderia ser expandido para que as hierarquias incluam outras relações além da relação taxonômica "é um" representada na árvore hiperbólica. Assim, seria possível localizar termos que são sujeitos e objetos de verbos específicos e criar relações não-taxonômicas. Uma opção, por exemplo, seria extrair da ocorrência de uma frase tal como "Os alunos compram doces" e criar uma relação "comprar" entre o termo "aluno" e o termo "doce". Este tipo de relação não-taxonômica seria fácil de extrair a partir da versão atual da ferramenta, mas seria necessário encontrar uma forma de escolher os verbos para criar as relações e o modo de visualização desse tipo de informação, pois uma árvore hiperbólica não se prestaria para isso, já que o conjunto de relações na maioria dos casos poderia gerar ciclos. Cabe salientar que o processo de aquisição de hierarquia de conceitos executado automaticamente pela ferramenta integra um trabalho de doutorado em andamento junto ao Grupo de Processamento da Linguagem Natural da Faculdade de Informática da PUCRS, de modo que novas configurações para a seleção dos sintagmas podem ser testadas.

A ferramenta ExATOlp gera representações de conteúdo dos textos que se prestam a um sem-número de aplicações. Integrada a um ambiente virtual de aprendizagem que aproveita textos de um jornal popular, oferece uma visualização panorâmica sobre seu conteúdo em diferentes opções. Da integração entre o PLN, a geração de ontologias e os Estudos da Linguagem, especialmente com a LC, tem-se uma nova opção de leitura e de descobertas para os conteúdos e palavras postos nesse tipo de texto, sem contar que a amplitude e a dinamicidade das representações de conteúdo de textos e dos *corpora* tende a entusiasmar estudantes e pesquisadores.

Referências

Amaral, M. F. (2006), Jornalismo Popular, São Paulo, Contexto.

Berber Sardinha, T. (2004), Linguística de Corpus, Barueri, São Paulo, Manole.

Bernardes, C. B. (2004), As Condições de produção do jornalismo popular massivo: o caso do Diário Gaúcho, Universidade Federal do Rio Grande do Sul, Faculdade de Biblioteconomia e Comunicação, Programa de Pós-Graduação em Comunicação e Informação, Diss. Mestrado.

Bick, E. (2000), The parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework, PhD thesis, Arhus University.

Lopes, L.; Fernandes, P.; Vieira, R.; Fedrizzi, G. (2009) ExATOlp – "An automatic tool for term extraction from Portuguese language corpora" In *Proceedings of the 4th Language and Technology Conference: Human Language Technologies as a challenge for computer science and linguistics (LTC'09)*. Adam Mickiewicz University.

Lopes, L.; Oliveira, L. H.; Vieira, R. (2010) "Portuguese term extraction methods: comparing linguistic and statistical approaches" In *PROPOR 2010 – International Conference on Computational Processing of Portuguese Language*.

Stubbs, M. (2001), Words and phrases: Corpus studies of lexical semantics, Oxford, Blackwell.