Extração de Contextos Definitórios a partir de Textos em Língua Portuguesa

Igor S. Wendt[1], **Renata Vieira**[1]

[1]Pontifícia Universidade Católica do Rio Grande do Sul (PUC-RS)
Av. Ipiranga, 6681 - 32 - Porto Alegre - RS - 90619-900

igor.wendt@acad.pucrs.br, renata.vieira@pucrs.br

Abstract. *A defining context is a part of a text or a statement that provides information about a concept, based on its use. Defining contexts extraction from texts is an important task in many applications as an aid in the construction of ontologies, the development of material aid to translation, creation of glossaries, dictionaries, among others. Thus, this paper proposes, implements and evaluates a set of grammatical rules to make the automatic extraction of potentially defining contexts in Portuguese texts.*

Resumo. *O contexto definitório é a parte de um texto ou de um enunciado que fornece informação sobre um conceito, com base em seu uso. A extração de contextos definitórios a partir de textos é uma tarefa importante em várias aplicações, como auxílio na construção de ontologias, no desenvolvimento de material de auxílio à tradução, na criação de glossários, dicionários, entre outros. Nesse sentido, este trabalho propõe, implementa e avalia um conjunto de regras gramaticais para fazer a extração automática de contextos potencialmente definitórios em textos de língua portuguesa.*

1. Introdução

Atualmente, diversas tecnologias que utilizam a linguística como base, vêm sendo desenvolvidas com o objetivo de dar suporte a diversas tarefas, entre elas a busca de informações, a elaboração de sumários, como apoio a tradutores, e a elaboração de dicionários e glossários. Uma das etapas da elaboração de dicionários e glossários, foco desse trabalho, consiste, primeiramente, na identificação dos conceitos de um domínio e de sua descrição. A descrição de um conceito, quando extraída a partir de textos, é chamada de contexto definitório ou explicatório e tem como função contribuir para a determinação do seu significado [Finatto 2002].

Para desenvolver esta etapa manualmente, é necessário que especialistas do domínio despendam grande quantidade de tempo na busca desses contextos, o que implica em custos elevados. Devido a isso, ultimamente tem se prestado mais atenção à tarefa de automatizar o processo de extração de contextos definitórios. Essa tarefa também é recorrente no processo de construção de ontologias, pois um dos elementos essenciais das ontologias são os conceitos. Portanto, outra aplicação relevante para as técnicas propostas nesse trabalho é a captura da descrição de conceitos relevantes em ontologias a partir do processamento de corpus de domínio.

Esse trabalho está organizado da seguinte forma, a seção 2 descreve trabalhos relacionados, na seção 3 descrevemos materias e métodos empregados na pesquisa, a

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 115–123,
Cuiabá, MT, Brazil, October 24–26, 2011.

seção 4 descreve os padrões identificados e implementados, na seção 5 descrevemos os experimentos, e concluímos o trabalho na seção 6.

2. Trabalhos Relacionados

O desenvolvimento de um dicionário ou glossário exige que um grande volume de textos especializados em um domínio seja analisado para que, a partir desta análise, possam ser identificados os termos mais representativos do domínio. Esses termos, que podem vir a ser entradas de um glossário ou dicionário, costumam ser explicados nos textos. Os trechos que contêm a explicação dos termos, chamados de contextos definitórios, podem ser identificados por um conjunto de critérios [Picht 2001].

De acordo com [De Lucca 2006], os contextos definitórios aparecem em três circunstâncias: a primeira é quando o autor cita um termo técnico; a segunda é quando o autor, em uma publicação científica, introduz um novo termo ou um termo pouco conhecido pela área; a terceira ocorre quando o termo é conhecido somente em uma língua e o autor informa o equivalente em sua língua nativa. Com exceção do último caso, os termos são seguidos de sua definição (mesmo que essa não siga os rigores da definição lexicográfica encontrada nos dicionários).

Existem trabalhos que visam à extração de definições de termos a partir de textos. Nesse sentido, busca-se o termo seguido de sua definição, como apresentado em [Del Gaudio e Branco 2007]. Esses autores apresentam um sistema baseado em regras gramaticais, em que se procura identificar definições em documentos anotados com informações morfossintáticas. Os textos utilizados são escritos em língua portuguesa e pertencem a três domínios diferentes; E-learning, Tecnologia da Informação e Sociedade da Informação. As regras adotadas têm como objetivo buscar definições nos documentos anotados através dos padrões linguísticos impostos pelas regras. São apresentados três grupos de regras gramaticais: "*Copula definitions*", "*Verbs definitions*" e "*Punctuation definitions*". O "*Copula definition*" é utilizado para encontrar definições onde o substantivo é seguido do verbo "ser". O "*Verbs definition*" procura por definições em que os verbos que seguem o substantivo são diferentes do verbo "ser". Por fim, o "*Punctuation definition*" considera somente as definições introduzidas por dois pontos (:).

No trabalho apresentado em [Przepiórkowski et al. 2007], a busca de definições é realizada de forma semelhante ao trabalho de Del Gaudio e Branco (2007), porém essas definições estão presentes em textos do domínio de *E-learning* e escritos em língua búlgara, polonesa e tcheca. Os textos são anotados com informações linguísticas e então para cada língua é utilizado um conjunto de regras adaptadas à linguagem em que são aplicadas. Para a língua búlgara são aplicadas 8 regras, para a língua polonesa são utilizadas 34 regras e para a língua tcheca 147 regras. O trabalho foi desenvolvido com o propósito de apoiar o especialista na construção de um glossário. Portanto, a abrangência torna-se mais importante que a precisão, pois é mais rápido e fácil para o especialista avaliar o que já foi extraído do que buscar as definições nos documentos.

Por fim, o trabalho proposto em [Iftene et al. 2007], apresenta um grupo de regras gramaticais utilizadas para extrair definições de textos em língua romena. Nesse trabalho foi utilizado um *corpus* de 56 documentos que tem como tema o uso de computadores na educação. Esse *corpus* contém aproximadamente 700.000 palavras e cada documento foi anotado com informações linguísticas. Com o *corpus* anotado, foi aplicado um conjunto

de regras que definem um padrão específico de definições. As definições foram divididas em cinco categorias: *is_def* (definições contendo o verbo "ser", em romeno "esta"), *verb_def* (definições contendo verbos romenos específicos, diferentes do verbo "esta"), *punct_def* (definições que usam pontuação como travessão, parênteses e vírgula), *layout_def* (definições que podem ser deduzidas pelo layout), *pron_def* (definições anafóricas) e *other_def* (definições que não podem ser incluídas nas outras categorias). As regras utilizadas para cada uma das categorias fazem uma busca por esses padrões nos documentos anotados e então extraem as frases que correspondem a essas regras.

Verifica-se que o uso de anotação linguística torna-se indispensável para os trabalhos que fazem extração de definições a partir de textos. Isso ocorre porque a anotação fornece o suporte necessário para o desenvolvimento de regras gramaticais que façam a identificação e recuperação de contextos potencialmente definitórios.

Nesse sentido, esse trabalho propõe o uso de regras gramaticais para realizar a extração de contextos definitórios a partir de textos anotados com informações linguísticas. Algumas das regras gramaticais utilizadas foram apresentadas por diferentes autores e outras foram expandidas de acordo com uma análise realizada sobre o *corpora*.

3. Materiais e Métodos

A extração de contextos potencialmente definitórios é dividida em diversas etapas. Primeiro, o *corpus* precisa ser anotado com informações linguísticas para então serem extraídos os termos mais relevantes. Em seguida, são recuperados os contextos definitórios desses termos através do uso de um conjunto de regras gramaticais.

3.1. *Parser*

O parser utilizado para fazer a anotação linguística do *corpora* utilizado foi o PALAVRAS [Eckhard 2000]. Esse *parser* recebe um texto em formato `ASCII` e, então, faz uma análise sintática, produzindo um arquivo em formato `XML` que contém todas as palavras do documento junto com suas características morfológicas.

3.2. Extrator de termos

A extração da lista de termos mais relevantes do *corpora* foi realizada através da ferramenta EχATO$_{LP}$. Essa ferramenta recebe o *corpus* anotado e extrai automaticamente todos os sintagmas nominais (SN) classificando-os segundo o número de palavras (*tokens*) que o compõem. Baseado em informações linguísticas e estatísticas, o EχATO$_{LP}$ utiliza algumas heurísticas para refinar o processo de extração de termos. Detalhes sobre esta ferramenta, são apresentados em [Lopes et al. 2010].

3.3. *Corpus* de Geologia Geral

Esse *corpus* é composto por 137 textos em português da área de Geologia Geral, sendo 119 artigos, 9 dissertações e 9 teses. Para construí-lo foram adotados alguns critérios para que fossem coletados apenas textos científicos (artigos, teses e dissertações). Esses arquivos, geralmente em formato `.pdf`, são livres e estão disponíveis na internet. Após coletados, esses textos foram encaminhados para especialistas do domínio, que avaliaram e selecionaram os textos de melhor qualidade para o estudo da área. Após a análise e seleção dos especialistas, os textos aprovados foram convertidos para arquivos no formato `.txt` com o auxílio da ferramenta Entrelinhas [Silveira 2008].

3.3.1. Extração de Termos do *Corpus* de Geologia Geral

Para verificar quais termos extraídos pela ferramenta EχATO$_{LP}$ podem ser considerados entradas de um glossário ou dicionário, foram utilizados dois glossários de referência e uma enciclopédia *online*: o glossário da MINEROPAR[1] com 3.078 termos, o glossário da UnB[2] com 1.447 termos e a Wikipédia[3] que, mesmo não sendo uma fonte específica de Geologia, contém uma grande quantidade de informações úteis para o domínio em questão.

Como resultado desse trabalho, foi obtida uma lista contendo 1.367 unigramas, 488 bigramas e 268 trigramas, onde 10 de cada grupo são posteriormente utilizados para a extração de contextos definitórios.

3.4. *Corpus* de Química Geral

Esse *corpus* é composto por 8 textos da área de Química Geral, sendo 4 da obra de [Atkins et al. 2001] e 4 da obra de [Russel 1994]. Esses textos são compostos por uma seleção dos capítulos mais relevantes para o conhecimento da área de Química Geral. Esse *corpus* foi desenvolvido pela equipe TEXTQUIM do Instituto de Letras e pela equipe da Área de Educação Química (AEQ) da Universidade Federal do Rio Grande do Sul.

3.5. Lista de Termos

O projeto TEXTQUIM oferece um banco de expressões e de termos técnicos para auxiliar as tarefas de tradução, redação, revisão e ensino de tradução. O banco disponibilizado no *site* TextQuim[4] possui atualmente 513 termos técnicos, sendo que 295 desses apresentam uma definição. Esses 295 termos estão divididos em 215 bigramas, 78 trigramas e 2 quadrigramas.

4. Heurísticas

Para a identificar os padrões de contextos definitórios, foi desenvolvido um concordanciador para apresentar os contextos que continham os termos extraídos. O concordanciador é uma ferramenta utilizada para listar as ocorrências de palavras ou frases de um texto. Em seguida, foi realizada uma análise manual, através da uma leitura sistemática dos contextos recuperados. Os padrões identificados através desta análise foram divididos em quatro grupos, apresentados a seguir. As heurísticas descritas nesses padrões foram implementadas em linguagem Java, utilizando o *parser* JAXP.

- Padrões sintáticos:
  Os padrões sintáticos apresentam apenas uma forma sintática, sendo o predicado verbal utilizado o verbo "ser" ou suas flexões. Este padrão recupera somente contextos em que o termo seja diretamente seguido do verbo "ser" ou suas flexões, ou seja, é verificado no documento anotado se os atributos desse verbo são: lemma="ser" e pos="v-fin".
  Exemplo do verbo "Ser": "A sismo estratigrafia **é** um método estratigráfico de análise e interpretação de dados sísmicos (...)"

[1] http://www.mineropar.pr.gov.br
[2] http://www.unb.br/ig/glossario/
[3] http://pt.wikipedia.org
[4] http://www6.ufrgs.br/textquim/

- Padrões tipográficos:
  Neste padrão é verificado no documento anotado se o atributo da palavra que segue o termo da lista é: word=":" ou word="(".
  Exemplo de dois pontos ":": "Granulometria**:** Medição do tamanho dos grãos que compõem uma rocha sedimentar."
- Padrões verbais:
  Este padrão tem por finalidade utilizar verbos que indiquem a presença de um possível contexto definitório. Não é necessário que o termo identificado seja diretamente seguido desses verbos, basta que o contexto apresente o termo e um dos seguintes verbos, apresentados abaixo, anotados no atributo "lemma": Chamar, Formar, Compor, Constituir, Denotar, Mostrar, Representar, Definir, Consistir, Indicar, Significar, Simbolizar, Caracterizar, Conter, Apresentar e todas as suas flexões.
  Exemplo do verbo Conter: "Cristais de plagioclásio **contêm** freqüentes inclusões de zircão, apatita e minerais opacos (...)"
- Padrões indicativos:
  Neste padrão são utilizadas expressões que indicam uma explicação prévia de determinado termo como "Conhecido como", "Reconhecido como" e "Isto é", que indica uma explicação sobre determinado contexto ou introduz o termo sobre o qual estava sendo discutido. Para tal, é verificado no documento anotado se o contexto recuperado contém o termo buscado e as expressões anotadas como lemma="conhecer" seguido de lemma="como" ou lemma="reconhecer" seguido de lemma="como" ou word="isto" seguido de word="é".
  Exemplo de "Reconhecido como": "Na análise de testemunhos efetuada em dois poços na região do Campo de Merluza, foram **reconhecidos como** principal litofácies reservatório, os arenitos maciços de granulometria fina a grossa e seleção pobre a moderada do Membro Ilhabela."

4.1. Ranqueamento

Considerando que os contextos definitórios mais relevantes geralmente apresentam o *definiendum* como sujeito da frase, desenvolvemos uma fórmula para pontuar a posição do termo na frase. A fórmula leva em consideração a posição do termo na frase, sendo que, quanto mais no começo da frase o termo estiver, maior será seu peso, visto que a estrutura frasal mais comum é SVO (Sujeito, Verbo, Objeto) ou SOV (Sujeito, Objeto, Verbo) [Botelho 2010].

A pontuação dada pela fórmula varia de 0,0000 a 1,0000 de acordo com a sua posição na frase. Abaixo é apresentamos a fórmula, onde "*A*" é o total de termos da frase e "*B*" a quantidade de termos presentes antes do termo em questão.

$$\chi = \frac{A - B}{A}$$

Através do uso dessa fórmula, são ranqueadas as definições através da pontuação que *definiendum* recebe, facilitando a visualização e seleção de contextos mais relevantes.

5. Experimentos

Nesta seção são apresentados os experimentos realizados utilizando o *corpus* de Geologia Geral e o *corpus* de Química Geral.

5.1. *Corpus* de Geologia Geral

Este experimento foi realizado utilizando as 9 dissertações e 9 teses do *corpus* de Geologia Geral visto que esses documentos possuem uma característica mais explicativa do que artigos, que são documentos científicos escritos de especialistas para especialistas, tendo como característica a omissão de explicações sobre a terminologia utilizada [Pearson 1999].

A partir desses documentos foram extraídos, através da ferramenta EχATO$_{LP}$, os unigramas, bigramas e trigramas candidatos a possuírem definições. Desta lista, foram escolhidos os 10 termos mais frequentes de cada categoria (unigramas, bigramas e trigramas) e presentes na lista previamente validada (seção 3.3.1). Fizemos este corte devido a grande quantidade de contextos extraídos para o conjunto total de termos. Utilizando esses 30 termos, foram extraídos todos contextos em que os mesmos apareciam, gerando assim um total de 1498 contextos. Esses contextos foram analisados por um terminólogo, que os classificou em Bom, Potencial e Ruim.

5.1.1. Resultados

Os contextos classificados como "Bom" e "Potencial" somam no total 152 contextos, desses, 45 foram marcados com Bom e 107 com Potencial. Analisando todos os contextos extraídos do *corpus* (1498), verifica-se que somente 10,1% (152) desses são considerados úteis como contextos definitórios. Através do uso das heurísticas, a quantidade de contextos extraídos é reduzida de 1498 para 552 contextos. Desses 552 contextos, 37 foram classificados como "Bom" e 48 classificados como "Potencial", somando 85 contextos relevantes. Os valores de Precisão, Abrangência a F-Measure são apresentados na Tabela 1, onde adotamos como *baseline* a recuperação de todos os contextos, ou seja, sem o uso de heurísticas.

Tabela 1. Resultado da extração de contextos a partir do *corpus* de Geologia Geral.

–	# Bom	# Potencial	B & P/Total	P	A	F
Sem Heurísticas	45	107	152/1498	10,1%	100%	18,3%
Com Heurísticas	37	48	85/552	15,4%	55,9%	24,1%

Analisando a Tabela 1, constata-se que 552 contextos foram extraídos através das heurísticas, sendo desses, 85 classificados como Potencial ou Bom, resultando em uma precisão de 15,4% e que destes 552 contextos, 152 são válidos (Bom / Potencial), o que resulta em 55,9% de abrangência. Cabe ressaltar que 6 contextos assinalados como "Ruim" apresentaram o termo como constituinte da definição e não como *Definiendum*. Para amenizar esse tipo de ocorrência, utilizamos a fórmula de ranqueamento apresentada na seção 4.1. Através dos valores obtidos pela fórmula, calculamos a média para o conjunto de contextos assinalados com "Bom", "Potencial" e "Ruim", resultando que grande parte dos contextos com valor abaixo de 0,7000 foram considerados ruins pelo avaliador.

Partindo dos resultados obtidos, aplicamos um ponto de corte em 0,7000, onde somente os contextos que possuíam o termo identificado com valor acima de 0,7000 foram mantidos. Desse modo foram extraídos 264 contextos, sendo 37 avaliados como "Bom",

36 avaliados como “Potencial” e 191 avaliados como “Ruim”. Esses números resultam em uma precisão geral de 27,7% e abrangência de 48%, conforme apresentado pela Tabela 2.

Tabela 2. Resultado da extração de contextos a partir do *corpus* de Geologia Geral com uso da fórmula de ranqueamento.

–	# Bom	# Potencial	B & P/Total	P	A	F
Sem Heurísticas	45	107	152/1498	10,1%	100%	18,3%
Com Heurísticas	37	48	85/552	15,4%	55,9%	24,1%
Com Heurísticas & Ranqueamento	37	36	73/264	27,7%	48%	35,1%

Com o uso da fórmula de ranqueamento nota-se que são removidos mais de 50% dos contextos extraídos e classificados com “Ruim”, aumentando a precisão para 27,7% enquanto a abrangência diminui em menor proporção, para 48%. Comparando os resultados obtidos através do uso das heurísticas e da fórmula de ranqueamento com o resultado obtido a partir da extração de todos os contextos dos termos da lista, nota-se que a precisão aumenta de 10,1% para 27,7%, e que também houve um aumento considerável da F-measure de 18,3% para 35,1%. Cabe ressaltar que a quantidade de contextos é reduzida em 82% (de 1498 para 264) diminuindo a quantidade de contextos a serem analisados de 5 para 1.

5.2. *Corpus* de Química Geral

Para realizar a extração de contextos potencialmente definitórios do *corpus* de Química Geral, foram selecionados aleatoriamente 10 bigramas e 10 trigramas da lista de termos discutida na seção 3.5. Através desses termos, foram localizados 246 contextos os quais foram avaliados por um terminólogo do projeto TextQuim, que os analisou e classificou conforme o expe-rimento anterior. Nessa análise, foram marcados 122 dos 246 contextos como bons ou potenciais, ou seja, 49,6% destes contextos são úteis para constituir uma definição. Esse resultado demonstra a melhor adequação desse *corpus* para a tarefa. Com o uso das heurísticas, foram recuperados 102 contextos, destes, 58 são válidos (Bom ou Potencial) o que gera a Precisão de 56,9%, a Abrangencia de 47,5% e 51,7% de F-measure, conforme apresentado na Tabela 3

Tabela 3. Resultado da extração de contextos a partir do *corpus* de Química Geral.

–	# Cont. Extraídos	# Cont. Válidos	P	A	F
Sem Heurísticas	246	122	49,6%	100%	66,3%
Com Heurísticas	102	58	56,9%	47,5%	51,7%

Utilizando a fórmula de ranqueamento com ponto de corte em 0,7000, a queda de abrangência e F-measure foi muito acentuada. Portanto, averiguamos diferentes pontos de corte, variando de 0,7000 até 0,4000.

Analisando a Tabela 4, nota-se que, quanto maior o ponto de corte, maior a precisão, porém como esse *corpus* é rico em contextos definitórios, com o aumento do ponto de corte são removidos contextos bons. Um ponto a ser observado é que o uso da fórmula de ranqueamento gera um ganho satisfatório de precisão, porém também diminui

Tabela 4. Resultado da extração de contextos a partir do *corpus* de Química Geral com o uso da fórmula de ranqueamento.

–	# Contextos Extraídos	# Contextos Válidos	P	A	F
Sem Corte	102	58	56,9%	47,5%	51,7%
Corte 0,7	37	27	72,9%	22,1%	33,9%
Corte 0,6	55	37	67,2%	30,3%	41,7%
Corte 0,5	69	43	62,3%	35,2%	44,9%
Corte 0,4	73	45	61,6%	36,9%	46,2%

a abrangência. Esse ponto demonstra que a fórmula de ranqueamento contribui melhor quando aplicada em *corpus* que é pobre em definições, visto que reduz significativamente a quantidade de contextos a serem analisados pelo especialista, retornando resultados mais precisos.

6. Conclusão e trabalhos futuros

O conjunto de heurísticas desenvolvidas, apresentam resultados satisfatórios, principalmente quando aplicadas sobre um *corpus* pobre em contextos definitórios. Esse ponto pode ser observado no experimento realizado sobre o *corpus* de Geologia Geral, no qual se obteve um aumento de precisão de 10,1% para 27,7% e em F-measure de 18,3% para 35,1%, quando comparado a um concordanciador.

Utilizando um corpus rico em contexto definitorios como o *corpus* de Química Geral, a precisão aumenta de 49,6% para 56,9% porém, a F-measure diminui, visto que a quantidade de contextos recuperados passa de 246 (sem heurísticas) para 102 (com heurísticas). Com o uso da fórmula de ranqueamento é possível obter até 72,9% de precisão, porém a abrangencia é reduzida, prejudicando a F-measure.

Ainda existe a possibilidade aplicar novas heurísticas a fim de recuperar contextos com outras características. Uma opção seria utilizar outros verbos indicativos frequentes, porém essa opção teria que ser aperfeiçoada através de experimentos em outro *corpora*, para então determinar as opções que melhor se adequam ao conjunto. Ainda poderia ser utilizada a informação presente nos documento anotados de quais palavras são Sintagmas Nominais, reduzindo assim, a quantidade de contextos em que o termo identificado não é o termo que está sendo definido.

Em comparação com trabalhos que utilizam métodos de extração de contextos definitórios semelhantes, o presente trabalho apresentou resultados satisfatórios, visto que, Del Gaudio e Branco apresentaram o resultado de 51% de F-measure e Przepiórkowski *et al.* apresentaram com melhor resultado 33,9% de F-measure. Os valores discutidos não devem ser considerados de forma competitiva pois são relativos a experimentos que levam em conta diferentes conjuntos de dados e formas de avaliação.

Referências

P.W. Atkins, L. Jones e I. Caracelli (2001). *Princípios de Química: questionando a vida moderna e o meio ambiente*. Bookman.

J.M. Botelho (2010). A ordem dos termos em português e a topicalização. *Revista Philologus*, páginas 45–61.

J. L De Lucca (2006). Identificação de padrões recorrentes no discurso técnico e científico para a extração automática de candidatos a contextos definitórios em língua portuguesa. *Revista Intercâmbio*, 15.

R. Del Gaudio e A. Branco (2007). Supporting e-learning with automatic glossary extraction: Experiments with Portuguese. Em *RANLP Workshop: Natural Language Processing and Knowledge Representation for eLearning Environments*.

B. Eckhard (2000). *The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Aarhus University Press.

M.J.B. Finatto (2002). O papel da definição de termos técnico-científicos. *Revista da ABRALIN*, 1(1):73–97.

A. Iftene, D. Trandabă e I. Pistol (2007). Grammar-based automatic extraction of definitions and applications for Romanian. Em *Proceedings of RANLP workshop"Natural Language Processing and Knowledge Representation for eLearning environments*, páginas 978–954. Citeseer.

L. Lopes, L. Oliveira e R. Vieira (2010). Portuguese term extraction methods: Comparing linguistic and statistical approaches. *International Conference on Computational Processing of Portuguese Language, PROPOR*.

J. Pearson (1999). Comment accéder aux éléments définitoires dans les textes spécialisés. *Terminologie et intelligence artificielle (TIA'1999)*, páginas 21–38.

H. Picht (2001). *Korpora als Ausgangspunkt für die Extraktion von terminologischen Daten*, volume 8. Synaps.

A. Przepiórkowski, Ł. Degórski, B. Wojtowicz, M. Spousta, V. Kuboň, K. Simov, P. Osenova e L. Lemnitzer (2007). Towards the automatic extraction of definitions in slavic. Em *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing: Information Extraction and Enabling Technologies*, páginas 43–50. Association for Computational Linguistics.

J.B. Russel (1994). Química Geral, vol. 2. *São Paulo: Makron*.

F. P. Silveira (2008). Entrelinhas - uma ferramenta para processamento e análise de corpus. *Dissertação de Mestrado, PUCRS*.