DE LUCCA, J. L. Identificação de padrões recorrentes no discurso técnico e científico para a extração automática a candidatos definitórios em língua portuguesa. *Revista Intercâmbio*, volume XV. São Paulo: LAEL/PUC-SP, ISSN 1806-275X, 2006.

IDENTIFICAÇÃO DE PADRÕES RECORRENTES NO DISCURSO TÉCNICO E CIENTÍFICO PARA A EXTRAÇÃO AUTOMÁTICA DE CANDIDATOS A CONTEXTOS DEFINITÓRIOS EM LÍNGUA PORTUGUESA

José Luiz DE LUCCA

ABSTRACT: In this article we described the acquisition and classification of an inventory of recurrent patterns, in Portuguese, that allows the construction of a computacional tool destined to the automatic extraction of possible definitory contexts in a corpus of scientific texts. The investigations were made by the professors. Gerardo Sierra and Rodrigo Alarcón (UNAM) were fundamental to our work elaboration.

KEYWORDS: contexts, definitory contexts, extraction, recurrent patterns, corpus

0. Introdução

Esta pesquisa insere-se no âmbito da lingüística computacional, mais precisamente, a extração automática de termos (Automatic Extraction of Terms).

Há muitos trabalhos sobre EAT, mas poucos sobre Extração Automática de Candidatos a Contextos Definitórios (EACDD).

A extração de termos e a conseqüente definição é parte essencial da pesquisa terminológica. Acreditamos, contudo, que além da identificação de termos, é necessário, como parte do trabalho terminológico, a identificação de sua definição correspondente. Assim, para a identificação de padrões recorrentes em textos técnicos e científicos para a extração automática de candidatos a contextos definitórios em língua portuguesa, realizamos uma pesquisa em revistas (de interesse científico) técnicas e científicas a fim de podermos identificar e extrair os contextos definitórios.

DE LUCCA, J. L. Identificação de padrões recorrentes no discurso técnico e científico para a extração automática a candidatos definitórios em língua portuguesa. *Revista Intercâmbio*, volume XV. São Paulo: LAEL/PUC-SP, ISSN 1806-275X, 2006.

As tecnologias lingüísticas e os recursos digitalizados prestam um grande serviço à tradução e à terminografia, tanto para a busca de informação, como quando utilizadas como ferramentas.

Atualmente os maiores recursos são representados pelos bancos de textos. Os bancos textuais apresentam a novidade de apresentar os dados em contexto real – sem fragmentação – proporcionando tantos contextos como ocorrências de uma mesma unidade nos textos. Assim, os bancos textuais converteram-se em laboratórios adequados para a descrição das unidades do discurso.

Diferentemente da lingüística tradicional, aqui o objetivo é a extração apenas de contextos definitórios, abstendo-se da análise e coleta dos contextos explicativos e associativos.

Já foram feitos muitos esforços para identificação de definições em corpus especializados de maneira sistemática (Pearson 1998). No entanto, faltam estudos para identificação dos padrões recorrentes usados pelos autores.

A partir dos padrões, seria possível desenvolver uma ferramenta capaz de extrair automaticamente os contextos definitórios, obtendo-se uma classificação dos possíveis termos e suas definições.

Os contextos definitórios aparecem, geralmente, em três circunstâncias. Quando o autor do texto cita um termo técnico; quando o autor, em uma publicação, eminentemente científica, introduz um termo novo ou pouco conhecido da comunidade; terceiro, quando o termo é conhecido apenas em inglês, por exemplo, o autor informa o equivalente mais próximo em português. Geralmente um neologismo é criado para representar tal equivalência. Nestes casos, com exceção do último, o termo vem acompanhado de sua definição – o que caracteriza o contexto definitório -, ainda que não obedecendo os rigores da definição lexicográfica, é considerada uma definição ou explicação (contexto definitório ou explicativo).

Nas corpora de conteúdo especializado, podemos observar que a presença de contextos definitórios tem uma freqüência muito significativa. No entanto, nas aplicações teminográficas ou na lexicografia especializada por tema, baseadas em corpora, e que, portanto, ilustram cada sentido com uma frase extraída dos corpora, tanto quando se trata de dicionários em papel, como bancos de dados, observamos uma ausência muito significativa dos

contextos definitórios. Esta constatação não difere muito, se observarmos outros tipos de aplicações terminológicas.

O estudo permitirá uma melhor compreensão do complexo fenômeno dos contextos definitórios assim como adentrar no universo do discurso, que se mostrou fundamental para o reconhecimento dos Contextos definitórios em textos técnicos e científicos.

1. Motivação

Quatro problemas fundamentais motivaram a presente pesquisa. O primeiro relativo à ausência quase total em língua portuguesa de pesquisa sobre contexto definitório em discurso técnico e científico. O segundo, relativo ao tipo de critérios que deveriam ser estabelecidos para o reconhecimento de contextos definitórios em unidades pertencentes a um domínio técnico ou científico.

O terceiro refere-se à discrimininação desses contextos pelo tipo de contexto definitório.

Por fim, a construção de ferramentas computacionais para distinguir esses contextos.

2. Metodologia

A metodologia empregada nesta pesquisa consistiu na detecção e análise manual de contextos definitórios em textos científicos.

Esta investigação compreendeu 49 documentos, todos em formato impresso, na área de ciências. Estes documentos são publicações periódicas brasileiras – não traduzidas - .

Esta classificação de padrões permitiu sintetizar os contextos definitórios e criar paradigmas de cada elemento constituinte.

Escolha das fontes:

As fontes foram escolhidas com base em indicações dos especialistas da Escola Politécnica da Universidade de São Paulo. Segundo esses profissionais, são fontes que têm o respaldo dos docentes e/ou pesquisadores que atuam nesta área.

Coleta dos contextos:

DE LUCCA, J. L. Identificação de padrões recorrentes no discurso técnico e científico para a extração automática a candidatos definitórios em língua portuguesa. *Revista Intercâmbio*, volume XV. São Paulo: LAEL/PUC-SP, ISSN 1806-275X, 2006.

A coleta dos contextos será feita apenas em textos em Língua Portuguesa, o que não descarta o aproveitamento de contextos que se refiram a termos em outras línguas.

Critério de seleção dos contextos:

Partiremos primeiro da digitalização de um corpus de treinamento, para, a seguir, definir os critérios de seleção e extração dos contextos definitórios dos corpora.

3. Objetivos

Nosso principal objetivo é cobrir satisfatoriamente uma carência na área da lingüística textual, da terminologia e da análise do discurso e

1. Desenvolver um método para identificar e extrair contextos definitórios do discurso técnico e científico;
2. Considerar revistas e jornais escritos em português, no Brasil, como relevantes para essa pesquisa;
3. Aprimorar uma metodologia de pesquisa terminológica que seja eficaz em todas as etapas do trabalho: escolha das fontes, coleta de termos, critério de seleção dos termos que funcionarão como *entradas* e redação dos verbetes;
4. Disseminar os resultados obtidos a partir dessa sistematização metodológica através da Internet e para as instituições envolvidas na pesquisa terminológica baseada no texto, de forma a facilitar a confecção de glossários e pesquisas em segmentos da terminologia.

Este trabalho tem o objetivo de apresentar os critérios para a identificação e a descrição de Contextos Definitórios típicos dos discursos técnico e científico, presentes no corpus escrito e eletrônico em língua portuguesa. Tratar do Contexto Definitório, em textos técnicos e científicos, significa entrar no terreno da delimitação de unidades complexas, tema que vem merecendo a atenção por parte de muitos estudiosos nos últimos anos. Implica, sobretudo, o estabelecimento de critérios para reconhecer e

estabelecer os limites entre contextos de diferentes conteúdos semânticos. Aqueles que trabalham com terminologia sabem que é complexa a tarefa de estabelecer esses limites.

O âmbito desta pesquisa restringir-se-á a identificar padrões recorrentes utilizados pelos autores para introduzir conceitos. Para tanto será necessário um inventário representativo da variedade de formas e estilos que os autores utilizam para intodução de novos termos.

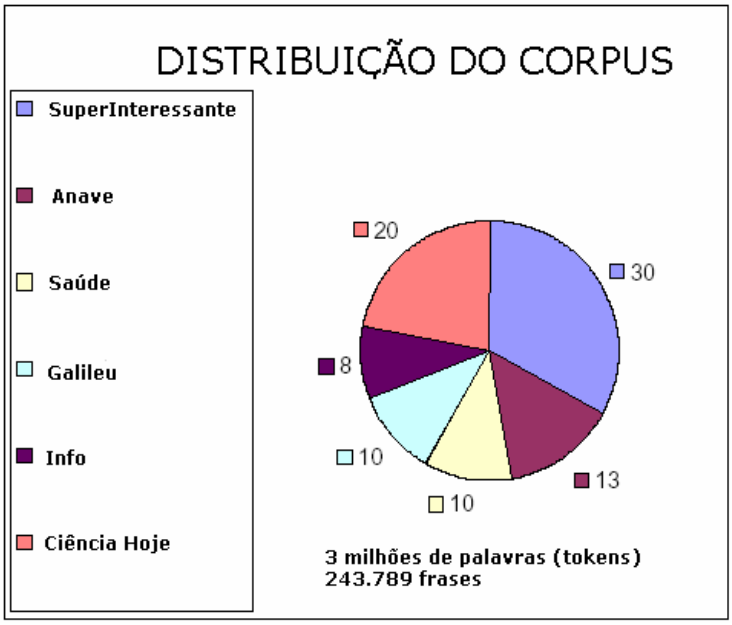
A partir de tal inventário, a próxima etapa seria o desenvolvimento de uma ferramenta computacional capaz de extrair automaticamente contextos definitórios.

A terceira e última etapa será a incorporação destes contextos definitórios como exemplos de uso (paradigma pragmático) na construção de dicionários especializados.

4. Corpus

Corpus feito exclusivamente de artigos extraídos de revistas (de interesse científico) técnicas e científicas.

A maior parte do corpus é constituída pelas Revista SuperInteressante e Ciência Hoje, respectivamente, 30% e 20%. As revistas "Saúde" e "Galileu" o terceiro e quarto lugares respectivamente, com 13% e 10%. As revistas com menor representatividade foram: Info(8%) e Anave (13%).



5. Classificação dos padrões

Utilizando uma metodologia simples de detecção e análise manual dos textos existentes em nosso corpus, buscamos, através da leitura sistemática, os contextos definitórios - contextos que claramente apresentavam uma definição de um termo. Com a finalidade de simplificar a identificação de padrões, delimitamos os contextos definitórios como aquelas estruturas compostas de um termo (T) e de sua definição (D). Estes padrões recorrentes foram agrupados em quatro grupos distintos: tipográficos, sintáticos, mistos e compostos.

Padrões tipográficos

DE LUCCA, J. L. Identificação de padrões recorrentes no discurso técnico e científico para a extração automática a candidatos definitórios em língua portuguesa. *Revista Intercâmbio*, volume XV. São Paulo: LAEL/PUC-SP, ISSN 1806-275X, 2006.

Os padrões tipográficos caracterizam-se por fatores de formato de texto: vírgula, hífen ou parêntesis, especialmente e ausência de predicado verbal.

Exemplo: "smart cards" – cartões que têm um "chip" acoplado para armazenar dados –. (T+'-'+D+'-')
Exemplo 2: Precatórios (decisões judiciais que consolidam dívidas públicas). (T+'('+D+')')

Padrões sintáticos

Os padrões sintáticos apresentam apenas uma característica sintática.

Exemplo: Bronquite é a inflamação dos brônquios. (T+P+D)

Padrões mistos

Padrões mistos são aqueles que combinam os padrões tipográficos e sintáticos.

Exemplo: A dihydroziacetona (DHA) reage com a queratina (proteína presente na camada superficial da pele), garantindo o tom bronzeado. (T+P+'('+D+')'+P)

Padrões compostos

Os padrões compostos podem ser de dois tipos:
1. Um contexto definitório serve para introduzir dois ou mais termos distintos.
2. A definição de um termo serve como contexto definitório para a introdução de um novo conceito.

Exemplo: A artrite é a inflamação das juntas, enquanto a artrose é o desgaste da articulação. (T+P+D+P+T+P+D)

De acordo com Sierra y Alarcón, a cada uma das formas corresponde um T (termo), uma D (definição) e uma P (predicado verbal, nominal ou informação pragmática).

6. Tipos de Contextos

Segundo o conteúdo semântico, os contextos podem ser divididos em três categorias: contextos definitórios, contextos explicativos e contextos associativos.

contextos definitórios – Os contextos com caráter de definição são os que contêm os descriptores essenciais do conceito.

exemplo: Isso ocorre porque a capsaicina, substância ativa das pimentas, não é solúvel em água.

contextos explicativos – Os contextos explicativos nos informam apenas alguns traços dos aspectos do conceito do termo.

exemplo: Em uma economia de mercado, o sistema de regulação oficial de preços na atividade econômica ameaça distorcer os preços relativos e a livre concorrência.

contextos associativos – Os contextos associativos, ao contrário dos dois acima citados, servem apenas para demonstrar o uso do termo no campo temático em estudo.

exemplo: O México não pode resistir muito mais sem desvalorizar *(Ámbito Financiero, Argentina, 15/01/1996)*

Conclusão

Nossa pesquisa, baseada em um corpus composto por 49 documentos, ou aproximadamente 600 mil palavras, revelou que, segundo nossa amostragem, 42% dos contextos caracterizam-se por padrões mistos; 33% por padrões sintáticos; 17% por padrões compostos; 8% dos contextos definitórios

A pesquisa representa um ganho e um valor agregado aos sistemas informacionais existentes. O resultado mostrou a existência de diferentes

padrões de contextos definitórios em textos de divulgação científica. A partir deste resultado, será possível desenvolver algoritmos para a extração automática de contextos definitórios em textos de divulgação científica, em língua portuguesa, com a finalidade de construir vocabulários especializados.

REFERÊNCIAS BIBLIOGRÁFICAS

CABRÉ, T., Estopà, R., Vivaldi, J.: Automatic term detection. A review of current systems. In Natural Language Processing. Amsterdam: John Benjamin's. 2001.

CARDERO, A.: El procesamiento de una terminología. Referencia especial a la terminología de control de satélites en el área de las telecomunicaciones en México. PhD Thesis, México: UNAM, 2001.

DE LUCCA, J.L. (2005). Identificação de padrões recorrentes em textos técnicos e científicos para a extração automática de candidatos a contextos definitórios em língua portuguesa. 15º InPLA. São Paulo:CD

PEARSON, J.: Terms in context. Amsterdam: John Benjamin's, 1998.

RODRÍGUEZ, C.: Extraction of knowledge about terms from metalinguistic activity in texts In: A. Gelbukh (ed.) Proceedings of the Conference on Intelligent Text Processing and Computational Linguistics, CICLING-2000. Mexico: Instituto Politécnico Nacional, 2000.

SIERRA, G., Alarcón, R. Identificación de Patrones recurrentes para la extracción automática de contextos definitorios. http://iling.torreingenieria.unam.mx/Clase/clase5.htm,2004