

Elaboração de Dicionário Eletrônico

Aluno: Mauro Ricardo Rebello de Paiva
Orientadora: Prof. Dra. Maria Carmelita Pádua Dias

Introdução

Este projeto insere-se na linha de pesquisa Processamento Automático de Linguagem Natural, na área de Estudos da Linguagem do Programa de Pós Graduação em Letras, e utiliza a metodologia da Lingüística de Corpus. A Lingüística de Corpus é caracterizada pelo estudo e extração de dados lingüísticos textuais presentes em um acervo de textos ou corpus. Nesta pesquisa, esses dados estão sendo utilizados para inclusão de termos do léxico do português em Dicionário Eletrônico, ferramenta básica para o processamento automático de textos.

A construção do corpus foi realizada através da coleta de textos eletrônicos disponíveis na internet sobre o tema “Saúde Pública”. Fez-se posteriormente a conversão da extensão dos textos, sua classificação quanto a gêneros textuais e registro ou grau de formalidade e a etiquetagem gramatical dos itens lexicais simples e complexos neles presentes.

De modo a preencher lacunas existentes em outros dicionários eletrônicos da língua portuguesa, a proposta é a de elaborar dicionários terminológicos, dicionários de compostos e de nomes próprios, além de complementar os conjuntos de etiquetas de cada entrada lexical.

O grupo de pesquisa em Lingüística Computacional da PUC-Rio, utilizando a ferramenta de tratamento de corpus UNITEX disponibilizada pela rede RELEX e pela Universidade de Marne-la-Vallée, França, vem dedicando-se à inserção e formalização de palavras simples e expressões compostas em dicionários.

Objetivos

Esta etapa inicial do projeto teve como objetivo central a coleta de material eletrônico (textos acadêmicos, textos de divulgação científica e relatórios), bem como sua adaptação (conversão de extensão), processamento automático e correção manual de etiquetagem para a formação de um corpus na área de saúde pública.

Metodologia

A pesquisa encontra-se em fase inicial de execução e até o momento foi feita a seleção de parte do corpus e sua preparação para consulta lingüística. A primeira etapa consistiu da conversão de suas extensões *.pdf e *.html para *.txt. A conversão precisou ser corrigida manualmente, pois durante esse processo ocorre a aglutinação de uma série de palavras. Por exemplo, a expressão “geração de corpus” é convertida em “geraçãodecorpus” e deve ser re-segmentada. Também o índice, a bibliografia, as notas de rodapé e marcações de hiperlink devem ser retirados manualmente do corpus.

Feitas estas correções, cada arquivo deve ser catalogado para conter o registro de título, a classificação em gêneros textuais (texto acadêmico, texto de divulgação ou relatório) e grau de formalidade (formal ou informal), para sua posterior etiquetagem, ou seja, a classificação gramatical de itens lexicais simples e complexos que ocorrem nos textos.

O programa UNITEX foi utilizado para a segmentação sentencial e etiquetagem do vocabulário dos textos.

Em seguida, realizou-se a correção manual da etiquetagem de parte do material. Nesta fase, a correção se centrou em nomes próprios e numerais percentuais, uma vez que ambos são itens lexicais complexos reconhecidos de maneira fragmentada como itens lexicais distintos e independentes. O nome próprio “Rio de Janeiro”, por exemplo, aparece classificado como “Rio_N + de_PREP + Janeiro_N”, em que “N”=nome, e “PREP”=preposição, ao invés de “Rio=de=Janeiro_NPROP”, em que “NPROP” corresponde a nomes próprios. Já o numeral percentual “100%”, por exemplo, foi etiquetado como “100_NUM + %_N”, quando deveria ser classificado como “100=%_NUM”, em que “NUM” designa a categoria dos numerais.

Resultados

Até o momento, foi realizada a captação de textos da internet que abordam o tema “Saúde Pública”, bem como a conversão da extensão desses arquivos de *.pdf e *.html para *.txt. Depois de processados os textos, realizou-se a segmentação de unidades lexicais que apareceram aglutinadas por conta da conversão, além da etiquetagem gramatical de parte dos itens lexicais simples e complexos. Por último, foi feito o registro desses textos quanto ao gênero textual e grau de formalidade.

Conclusões

Esta etapa do projeto envolveu a formatação de textos acadêmicos, de divulgação e relatórios na área de saúde pública e a etiquetagem parcial do corpus. Uma etapa posterior da pesquisa envolverá a revisão da etiquetagem de itens lexicais de outras classes gramaticais, a inclusão no dicionário de palavras não reconhecidas pelo etiquetador e a formalização final das entradas lexicais para que passem a fazer parte do dicionário.

Referências

Paumier, Sébastien. UNITEX. [<http://www-igm.univ-mlv.fr/~unitex/>]

Sardinha, T. **Linguística de Corpus: Histórico e Problemática**. D.E.L.T.A, Vol. 16, N 2, 2000 (323-367)

Sinclair, J. **Corpus Concordance Collocation**. Oxford University Press. 1991