320 **MARCAÇÃO DE CORREFERÊNCIA EM CORPUS LINGUÍSTICOS UTILIZANDO O MMAX.**
*Genessa Robinson, Renata Vieira***.**(Programa Interdisciplinar de Pós-Graduação em Computação Aplicada - UNISINOS).

O projeto COMMOn-REFs (Um Modelo Computacional Unificado para o Tratamento de Referências) tem por objetivo a marcação lingüística de *corpus*. Marcação é o ato de destacar elementos lingüísticos do *corpus*, através de etiquetas. O presente projeto visa a marcação de expressões referenciais. Existe uma tendência de padronização para as marcações lingüísticas. Um padrão de linguagem de marcação que tem sido utilizado é a Linguagem XML (*Extensible Markup Language*). O XML é uma linguagem que permite a criação de etiquetas de acordo com as informações a serem destacadas. Um aplicativo que trabalha com marcação lingüística de *corpora*, utilizando XML é o MMAX, adotado por este projeto. Através do MMAX, é possível tratar expressões referenciais e manter as informações salvas em um arquivo de marcação. Nosso projeto adotou esta ferramenta, pois ela permite destacar as expressões referenciais, suas correferências e classificá-las de acordo com o esquema de marcação pretendido. O processo de transformação do *corpus* para o formato XML, compatível com a ferramenta MMAX, consiste primeiramente na utilização de scripts na Linguagem PERL para criação dos arquivos descritos a seguir. O primeiro deles é o word.xml, que atribui um identificador para cada palavra ou elemento de pontuação do *corpus*. Uma vez tendo identificado estes itens de um *corpus*, o próximo passo é a criação do arquivo text.xml, que contém a marcação da estrutura das sentenças e parágrafos do *corpus*, identificando seus intervalos de acordo com o arquivo anterior gerado. Estes dois arquivos são utilizados no MMAX pelo lingüista que fará a marcação manual das expressões referenciais do *corpus*. Toda a marcação feita pelo lingüista no *corpus* é salva em um arquivo markable.xml. Um *script* foi criado para adaptar a marcação para o formato XML de outros *corpora* anotados, que foram utilizados em projetos anteriores. Ao termos um *corpus* marcado, podemos utilizá-lo como um *corpus* de treinamento para várias atividades do processamento da linguagem natural, por exemplo, para uma ferramenta de marcação automática de expressões referenciais em outros *corpora*. (Fapergs).