016 **FORMATO DE ANOTAÇÃO DE CORPUS DA LÍNGUA PORTUGUESA.** *José Guilherme Camargo de Souza, Renata Vieira (orient.)* (UNISINOS).

Para que a web semântica estabeleça-se em larga escala, faz-se necessário que um grande número de documentos seja anotado. Para esse propósito, ferramentas que possibilitam a anotação semi-automática de documentos têm sido desenvolvidas. Para isso, é preciso fazer a anotação lingüística dos textos. Existem diversas ferramentas para anotação manual e automática de corpus com informações lingüísticas de vários níveis. Essas informações devem ser codificadas de uma forma eficiente. Por eficiente, entendemos que os repositórios de dados com anotações lingüísticas devem permitir a expansão e a facilidade de uso e reuso dessas informações. Por tratar-se de uma área recente, modelos de anotação que atendam às exigências acima ainda estão sendo estudados e propostos. Neste trabalho, é proposto um formato de codificação para anotação lingüística para a Língua Portuguesa, baseado na linguagem de marcação XML. O formato codifica informações estruturais sintagmáticas, morfossintáticas e referenciais e procura atender às características acima. Além disso, procura-se estar em conformidade com padrões internacionais e formatos que vêm sendo adotados pela comunidade de Processamento de Linguagem Natural, entre eles, os desenvolvidos pelo grupo ISO TC37 SC 4 (International Standards Organization - Language Resources Standards) e os formatos utilizados por anotadores lingüísticos como o PALAVRAS, além de projetos como o MuchMore, TIGER e outros. Este trabalho faz parte do projeto PLN-BR, um esforço conjunto de diversas Universidades do Brasil, que tem como objetivo estudar e desenvolver ferramentas de Processamento de Linguagem Natural para a Língua Portuguesa. A proposta será discutida e avaliada dentro do contexto deste projeto, com o fim de tornar-se um formato padrão de anotação lingüística para a Língua Portuguesa.