Lingüística de corpus no Brasil

Resumos das intervenções

1. Corpora para reconhecimento de terminologias e de linguagens técnico-científicas: o desafio da integração de bases de dados

Maria José Bocorny Finatto - UFRGS

Ao longo da história da Terminologia, vários grupos de pesquisa têm reconhecido terminologias com corpora. Desde os anos 30, na Rússia, já se exploravam grandes acervos de textos; nos anos 80, desenvolveram-se técnicas para identificação de vocabulário científico, das quais apenas há pouco tivemos notícia. A partir de pesquisas mais antigas, de uma série de Conferências sobre Léxico, Corpus e Dicionários do final dos anos 90 e de trabalhos atuais do Grupo TERMISUL e Projeto TEXTQUIM, discutimos aqui alternativas para criação de um núcleo de Terminologia e Lingüística de Corpus na nossa universidade. Esse núcleo visa: a) integrar bases de dados e corpora de diferentes tipos de reconhecimentos de linguagens técnico-científicas; b) oferecer informações sobre o acervo reunido a pesquisadores e estudantes.

2. Métodos estatísticos na avaliação das diferenças entre corpora

Marco Antonio Esteves da Rocha – UFSC

A intervenção pretende focalizar a utilização de métodos estatísticos na construção e avaliação de corpora segundo critérios de equilíbrio e diversidade, com ênfase específica nas questões relacionadas ao tratamento da rede WWW como um corpus. O propósito é aprofundar a discussão de métodos estatísticos para a comparação de corpora, de modo a iniciar um processo de investigação que possa resultar na definição de padrões de semelhança, homogeneidade e diferenciamento de corpora, subcorpora e outras classificações mais específicas de tipos textuais. Espera-se contribuir para o esforço de pesquisa no sentido de criar formas operacionalmente adequadas e lingüisticamente fundamentadas para a construção de corpora a partir da rede WWW segundo os objetivos de cada investigação.

3. Pesquisas com corpora: Tradução, Ensino e Aprendizagem

Stella E. O. Tagnin - USP

Esta apresentação exporá os projetos que envolvem a Lingüística de Corpus e que estão em desenvolvimento no Departamento de Letras Modernas (DLM) da Universidade de São Paulo. O primeiro deles é o CorTec – um corpus de textos técnicos originais em inglês e português em cinco áreas: Direito Comercial (contratos), Informática (segurança na Internet), Meio ambiente (ecoturismo), Cardiologia (hipertensão arterial e insuficiência

renal) e Gastronomia (receitas culinárias). Esse corpus poderá ser acessado na Web, a partir de agosto, para extração de listas de palavras e concordâncias, o que fornecerá subsídios para pesquisas lexicográficas, terminológicas e tradutórias. Há diversas pesquisas em andamento que contemplam essas áreas. Outro projeto é o Corpus de Aprendizes, em fase de implementação, que vai reunir textos produzidos pelos alunos das cinco áreas do DLM (alemão, espanhol, francês, inglês e italiano), tanto dos cursos de graduação quanto dos extracurriculares. Esse corpus permitirá detectar áreas de dificuldade dos aprendizes, seus erros mais comuns, sub-uso ou sobre-uso de certas palavras, expressões ou estruturas, bem como acompanhar seu desenvolvimento, uma vez que será um corpus diacrônico. Permitirá também pesquisas comparativas de aprendizes dos dois tipos de curso e de aprendizes de línguas diversas. Essas pesquisas podem fornecer importantes dados para a confecção de material didático mais adequado a esse público.