SISTEMA DE PESQUISA EM MÚLTIPLOS NÍVEIS PARA A FALA ESPONTÂNEA

Diego Furtado de Souza, Sandro Renato Dias,
Maryualê Malvessi Mittmann, Heliana Mello, Tommaso Raso

Universidade Federal de Minas Gerais - Estudos Linguísticos Baseados em Corpora
Início: Setembro/2011. Previsão de Conclusão: Dezembro/2012

Este trabalho apresenta uma ferramenta de manipulação e análise de corpus anotado em diversos níveis. Relatamos o desenvolvimento de um sistema de pesquisa em múltiplos níveis (lexical, morfossintático, informacional, ilocucionário, sociolinguístico) para o corpus de fala espontânea C-ORAL-BRASIL (RASO; MELLO, 2012). O sistema envolve um banco de dados dinâmico, em MongoDB (10GEN, 2011), e uma interface de pesquisas online, em JSONQuery. O MongoDB é um SGBD orientado a objetos, dinâmico, baseado em documentos. Conta com documentos embutidos, mas permite a referência entre os diversos documentos, possibilitando o cruzamento de informações contidas nos documentos de metadados (informação sociolinguística e dados das gravações) com os documentos de transcrição e de áudio. Baseia-se em documentos BSON (Binary JSON), um tipo simplificado de objeto web, suportado pela maioria dos navegadores. JSON (Java Script Object Notation) (CROCKFORD, 2006) é uma linguagem de intercâmbio de dados alternativa ao XML, porém mais simples e leve. O desenvolvimento do sistema envolve a migração dos dados, o desenvolvimento dos módulos e a implantação no servidor. A migração consiste na transformação dos dados contidos na estrutura atual (XML) gerados pelos softwares WinPitch (MARTIN, 2004), que contém as informações do alinhamento do texto com o som, e PALAVRAS (BICK, 2000), que tem a etiquetagem morfossintática. O sistema será construído nos seguintes módulos: (1) Autenticação de usuários. Controla o acesso dos usuários ao sistema e a segurança relacionada às permissões de usuários, possibilitará estatísticas sobre acesso, manutenção de usuários, bloqueio de acessos indevidos, etc. (2) Módulo Principal. Centraliza todas as configurações, páginas e acessos do sistema criadas tfem menus de acesso visíveis em todas as telas, trazendo uma maior usabilidade e segurança na utilização. (3) Módulo de Pesquisa. Parte central do sistema, consiste na pesquisa e filtragem de todos os dados disponíveis na base, permitindo o cruzamento dos dados dos documentos no BD. (4) Módulo de Detalhes dos Arquivos. Destina-se a mostrar os detalhes de cada arquivo. (5) Módulo Estatísticas e Gráficos. Destina-se à consolidação de dados, identificação de vícios de linguagem, escala de tempo, análise do corpus em geral e a pesquisas previamente definidas através de gráficos e estatísticas. (5) Módulo de Relatórios. Exporta as informações de uma pesquisa ou estatísticas do sistema para análise futura, com geração de arquivo pdf e/ou impressão em papel. (6) Módulo Help. Destinado a melhorar a usabilidade, com dicas de campos e um manual de utilização. (7) Módulo JSONQuery Console. Interface para pesquisas rápidas e otimizadas no banco de dados. O sistema acima descrito permitirá buscas complexas e cruzamento de dados em diferentes níveis de etiquetagem do corpus (morfossintática, informacional, ilocucionária) com dados sociolinguísticos. O usuário terá acesso a estatísticas e ao áudio dos enunciados resultantes da pesquisa.

10GEN. **MongoDB**. 2011. Disponível em: <www.mongodb.org>. Acesso em: 12 Maio 2012.

BICK, Eckhard. **The Parsing System Palavras**: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework. Aarhus: Aarhus University Press, 2000.

CROCKFORD, Douglas. **JSON**: Java Script Object Notation. 2006. Disponível em: <json.org>. Acesso em: 12 Maio 2012.

MARTIN, Philippe. **WinPitch Corpus**: A text to Speech Alignment Tool for Multimodal Corpora. Lisbon: LREC. Maio, 2004. Disponível em: <http://lablita.dit.unifi.it/coralrom/papers/index.html>. Acesso em: 12 Maio 2012.

RASO, Tommaso; MELLO, Heliana (Org). **C-ORAL-BRASIL I**: Corpus de referência do Português Brasileiro falado informal. Belo Horizonte: UFMG, 2012.