Projeto NURC Digital: Resultados Preliminares

O Projeto da Norma Urbana Linguística Culta (NURC), inaugurado em 1969, teve como objetivo inicial documentar e estudar a norma falada culta de cinco capitais brasileiras. Os dados que fazem parte do acervo do Projeto NURC têm sido, ao longo desses anos, utilizados para a elaboração de um grande número de trabalhos acadêmicos de grande importância, como, por exemplo, a *Gramática do Português Falado* (Castilho 1990), grande e ambicioso projeto nacional que envolveu dezenas de pesquisadores na área da linguística e resultou em uma série de volumes, todos contendo análises de materiais extraídos dos dados do Projeto NURC. As gravações desses dados foram todas feitas em fita magnéticas de rolo, uma tecnologia hoje caduca, o que dificulta o seu acesso a pesquisadores interessados não apenas na transcrição, mas também nos registros feitos em áudio. O Projeto NURC Digital, financiado pelo CNPq (Chamada Universal, Processo: 472918/2012-5), tem por objetivo geral propor uma metodologia de processamento, organização e disponibilização de um corpus representativo do acervo do Projeto NURC, em formato digital, que servirá como possível modelo a ser adotado para todo o material pertencente ao arquivo do Projeto NURC. O projeto encontra-se em fase de pleno desenvolvimento. O objetivo aqui é apresentar para a comunidade científica resultados preliminares das ações relativas ao seu desenvolvimento. Em particular, serão apresentados: (i) os procedimentos técnicos que foram considerados e testados no processo de digitalização das gravações analógicas dos inquéritos pertencentes ao Projeto NURC seção Recife (NURC/RE), justificando o método e as características relativas ao formato digital final adotados, (ii) o método utilizado para minimizar ruídos de *pitch* fixo – *hum* e *whistles* –, em geral associados a gravações analógicas em fitas magnéticas, (iii) o procedimento utilizado para a segmentação e a anotação alinhada do corpus compartilhado do Projeto NURC/RE, mediante uso dos aplicativos Praat (Boersma & Weenink, 2015) e ELAN (Wittenburg et. al., 2006), (iv) a rotina para a anotação gramatical automática do corpus, a partir do parser PALAVRAS (Bick, 2000), (v) o método adotado para a composição dos metadados, (vi) o portal na internet onde estão sendo depositados todos os arquivos relativos ao Projeto NURC Digital, incluindo o corpus compartilhado do NURC/RE, totalmente anotado; (vi)i a ferramenta de busca do portal, baseada no sistema *Spock* (Janssen & Freitas 2010), que permitirá consulta avançada do corpus anotado, dando acesso não apenas ao material de texto, mas também ao fragmento de áudio correspondente; (viii) o projeto de reedição, em formato digital, dos volumes organizados pelo pesquisadores do Projeto NURC/Recife, os *Materiais Para o Seu Estudo*, desta vez acompanhados de link para download de todos os inquéritos em áudio, totalmente anotados em formato que possibilite buscas avançadas, (ix) a estratégia de preservação a longo termo dos arquivos digitais, mediante depósito em bancos de dados internacionais, e (x) possíveis desdobramentos das ações do projeto. A apresentação de resultados preliminares do Projeto NURC Digital à comunidade científica busca não apenas divulgar o projeto em si e o corpus por ele disponibilizado, mas sobretudo suscitar comentários e sugestões num estágio de desenvolvimento em que serão bastante importantes e potencialmente úteis.

Referências:

Bick, E. (2000) The parsing system “PALAVRAS”: automatic grammatical analysis of Portuguese in a constraint grammar framework. Aarhus: Aarhus University Press. 412p. Disponível em: <http://visl.sdu.dk/~eckhard/pdf/PLP20-amilo.ps.pdf>. Acesso em 30 abr. 2015.

Boersma, P. & Weenink, D. (2015). Praat: doing phonetics by computer. Disponível em: <http://www.praat.org/>. Acesso em 30 abr. 2015.

Castilho, A. (org.) (1990). Gramática do português falado. Campinas: Editora da Unicamp; São Paulo: Fapesp.

Janssen, M. & T. Freitas (2010) Spock – a spoken corpus client. In: Oliveira Jr., M. Estudos de Corpora: da Teoria à Prática. Lisboa: Edições Colibri, p. 111-126.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In: Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.