

## NURC DIGITAL: UMA PROPOSTA DE PRESERVAÇÃO DOS DADOS DO PROJETO NURC

### **Introdução: identificação do problema**

O Projeto da Norma Urbana Linguística Culta teve seu início em 1969, tendo como proposta inicial documentar e estudar a norma falada culta de cinco capitais brasileiras: Recife, Salvador, Rio de Janeiro, São Paulo e Porto Alegre.

Os dados coletados pelo Projeto NURC desde então têm sido utilizados para a elaboração de um grande número de trabalhos acadêmicos, incluindo dissertações de mestrado, teses de doutorado, artigos publicados em periódicos nacionais e internacionais, e trabalhos apresentados em encontros científicos ao redor do mundo.

Lamentavelmente, os registros magnéticos dos inquéritos do Projeto NURC estão em sério risco de deteriorização. Na verdade, muitos desses registros já se encontram irremediavelmente destruídos pela ação do tempo. É imprescindível, portanto, que este valioso material seja resgatado o quanto antes, mediante a transposição de seus dados analógicos para formatos digitais que garantam a sua preservação e utilização no futuro.

O que se propõe aqui é indicar meios que poderão ser utilizados pelo Projeto NURC em todas as capitais em que está sediado para a preservação e a disponibilização mais efetiva de seus corpora. Para isso, propõe-se a digitalização de um corpus representativo de inquéritos pertencentes ao acervo do NURC, mediante técnicas de digitalização e de arquivagem recomendadas por órgãos internacionais especializados em arquivagem de dados digitais.

### **Estado da arte**

A maior parte dos estudos desenvolvidos a partir dos dados do Projeto NURC deriva de uma série de publicações feitas com transcrições de material selecionado pelos grupos de pesquisadores atuantes em cada uma das capitais em que o Projeto era desenvolvido. Essas coletâneas de transcrições publicadas a partir da década de 80 ficaram conhecidas por Materiais Para o Seu Estudo: Castilho e Preti (1986, 1987), Preti e Urbano (1990), Callou (1992), Callou e Lopes (1993, 1994), Motta e Rollemberg (1994), Hilgert (1997), Sá et al. (1996, 2005). Os estudos feitos a partir dessas publicações desconsideravam, em sua grande maioria, o registro de áudio, baseando-se exclusivamente nas transcrições aí presentes. Essa não era, evidentemente, uma opção dos estudiosos. Tratava-se mesmo de uma questão de dificuldade de acesso aos dados gravados. Todas as gravações feitas pelo Projeto NURC utilizaram, como meio, fitas magnéticas de rolo, que, se por um lado garantia a qualidade das gravações, por outro dificultava o acesso às mesmas, uma vez que reprodutores de fita de rolo eram equipamentos caros e pouco comuns.

Uma outra dificuldade que a utilização do material do Projeto NURC apresentava aos estudiosos era – e continua sendo, em grande parte – a não disponibilização dos dados transcritos em formato digital. Assim, o processo de análise a partir dos textos publicados em formato impresso era – e continua sendo – necessariamente demorado e eventualmente falho, uma vez que não se podia contar com buscas automatizadas de fenômenos linguísticos particulares.

Com o advento da tecnologia, tem-se cada vez mais incentivado a disponibilização de dados linguísticos em formato digital, que possam ser acessados por humanos e máquinas. A simples digitação de dados é apenas um primeiro passo para a criação de um corpus digital. Há, na verdade, uma série de medidas recomendadas por especialistas na área da construção de corpora eletrônicos que precisam ser consideradas, se o objetivo for construir um corpus que seja também legível por máquinas. A observância desses procedimentos metodológicos garantirá a preservação do valioso material do Projeto NURC, de forma que o mesmo possa ser utilizado mais eficazmente não apenas no presente, mas por futuras gerações de pesquisadores.

### **Metodologia e Resultados Esperados**

O que se pretende aqui é propor uma metodologia de organização de um corpus representativo do acervo do Projeto NURC, em formato digital, que servirá como possível modelo a ser adotado para a informatização de todo o material pertencente ao arquivo do Projeto NURC. Para isso, serão levados em conta procedimentos internacionais estabelecidos para a construção de corpus linguístico digital. Esta proposta representa, assim, um importante passo no processo de preservação do valioso acervo do Projeto NURC, que atualmente se encontra em sério risco de deteriorização ocasionada pela ação do tempo. Os resultados provenientes da execução desta proposta beneficiará diretamente a comunidade científica, que passará a ter disponíveis para consulta otimizada dados – anteriormente de difícil acesso – em formato digital de alta qualidade.

Dados de áudio do *corpus compartilhado* do Projeto NURC/Recife serão digitalizados observando-se as recomendações propostas pelo Open Archival Information System (OAIS) , que é um modelo de referência, com padrão ISO (14721:2003), adotado pelos bancos digitais de dados linguísticos mais recentes. As informações referentes aos arquivos de áudio e às transcrições (metadados), serão registradas seguindo os preceitos do Metadata Encoding and Transmission Standard (METS) , também adotados por bancos de dados internacionais. As transcrições dos dados serão registradas no aplicativo ELAN , que possibilita o seu alinhamento com os arquivos de áudio a que se referem, além de permitir que áudio, transcrições e metadados sejam pesquisáveis local e virtualmente.

O corpus informatizado será arquivado localmente, nos servidores da Universidade Federal de Alagoas e da Universidade Federal de Pernambuco, para livre consulta pela comunidade científica, e depositado em bancos internacionais, tais como o do Dokumentation Bedroter Sprachen (DoBes) , com o intuito de garantir a sua preservação.

Uma vez constituído e devidamente arquivado, o corpus digitalizado será apresentado aos atuais coordenadores do Projeto NURC, em todas as capitais, como modelo a ser discutido e, eventualmente, adotado, para a informatização e preservação de todo o material coletado por este importante projeto na área da linguística.

### **Referências Bibliográficas**

Callou, D. M. I. (org.) 1992). *A Linguagem Falada Culta na Cidade do Rio de Janeiro. Materiais para seu estudo*. Rio de Janeiro: UFRJ/FJB, vol. I, Elocuções Formais.

- Callou, D. M. I. e Lopes, C. R. (orgs.) (1993). *A Linguagem Falada Culta na Cidade do Rio de Janeiro. Materiais para seu estudo*. Rio de Janeiro: UFRJ/CAPES, vol. II, Diálogo entre Informante e Documentador.
- Callou, D. M. I. e Lopes, C. R. (orgs.) (1994). *A Linguagem Falada Culta na Cidade do Rio de Janeiro. Materiais para seu estudo*. Rio de Janeiro: UFRJ/CAPES, vol. III, Diálogos entre dois informantes.
- Castilho, A. (org.) (1990). *Gramática do português falado*. Campinas: Editora da Unicamp; São Paulo: Fapesp.
- Castilho, A. (org.) (1993). *Gramática do português falado*. Campinas: Editora da Unicamp; São Paulo: Fapesp.
- Castilho, A. (2007). Fundamentos teóricos da Gramática do português culto falado no Brasil. *Alfa*, São Paulo, 51 (1): 99-135.
- Castilho, A. & Basílio, M. (orgs.) (1996). *Gramática do português falado*. Campinas: Editora da Unicamp; São Paulo: Fapesp.
- Castilho, A. et al. (1995). Informatização de acervos da Língua Portuguesa. *Boletim da Associação Brasileira de Lingüística* 17: 143-154.
- Castilho, A. & Preti, D. (orgs.) (1986). *A Linguagem Falada Culta na Cidade de São Paulo. Materiais para seu estudo*. São Paulo: TAQ/Fapesp, vol. I, Elocuções Formais.
- Castilho, A. & Preti, D. (orgs.) (1987). *A Linguagem Falada Culta na Cidade de São Paulo. Materais para seu estudo*. São Paulo: TAQ/Fapesp, vol. II, Diálogos entre dois informantes.
- Hilgert, J. G. (org.) (1997). *A Linguagem Falada Culta na Cidade de Porto Alegre*. Passo Fundo: Ediupe / Porto Alegre: Ed. Universidade/Ufrgs, vol. I, Diálogos entre informante e documentador.
- Motta, J. e Rollemberg, V. (orgs.) (1994). *A Linguagem Falada Culta na Cidade de Salvador. Materiais para seu estudo*. Salvador: Instituto de Letras da UFBA, vol. I, Diálogos entre Informante e Documentador.
- Preti, D. e Urbano, H. (orgs.) (1990). *A Linguagem Falada Culta na Cidade de São Paulo. Materiais para seu estudo*. São Paulo: TAQ/Fapesp, vol. III, Diálogos entre o Informante e o Documentador.
- Sá, M. P. M. et al. (orgs.) (1996). *A Linguagem Falada Culta na Cidade do Recife*. Recife: Universidade Federal de Pernambuco, Programa de Pós-Graduação em Letras e Lingüística, vol. I: Diálogos entre informante e documentador.
- Sá, M. P. M. et al. (orgs.) (2005). *A linguagem falada culta na cidade do Recife: elocuições formais*. Recife: Universidade Federal de Pernambuco/Programa de Pós-Graduação em Letras e Lingüística, 2005.