SISTEMA CORPOR

BASES DE INFORMAÇÕES, *CORPORA* E LÉXICOS ORTOGRÁFICOS E FONÉTICOS DO PORTUGUÊS FALADO DE SÃO PAULO EM SISTEMA DE BANCO DE DADOS RELACIONAL

As Bases de Informações, a partir das quais os demais componentes do Sistema são gerados – *Corpora* e Léxicos –, são constituídas a partir da transcrição ortográfica e fonética de gravações de 432 diálogos (entrevistas e conversações) entre 216 informantes paulistas (São Paulo, Campinas, Itu) e documentador (26/11/1972 a 02/05/1973). Desde o humilde servente até o intelectual de uma universidade, visto que as Bases incluem informantes analfabetos e dos diferentes níveis de escolaridade, forneceram um material vastíssimo de estudo, com múltiplas possibilidades de análise.

Tempo de gravação de cada informante – quinze minutos:

- cinco minutos iniciais: dados para cadastro dos informantes;
- do sexto ao décimo minutos: diálogo formal – entrevista através de um questionário único para todos os informantes;
- cinco minutos finais: diálogo informal – conversação.

Critérios extralinguísticos considerados na seleção dos informantes:

- região de origem – informantes naturais da região, tendo sempre nela residido, com pais também naturais da região;
- sexo – informantes de ambos os sexos;
- escolaridade – avaliada numa escala de seis pontos, desde o informante com superior completo até o analfabeto;
- faixa etária – em função do nível de escolaridade;
- nível socioeconômico – determinado pela escolaridade e status profissional do chefe de família.

Codificação do informante – de acordo com os critérios considerados na sua seleção:

- Informante – 6 dígitos: cidade / sexo / escolaridade / faixa etária / nível socioeconômico:

- Cidade – 1 digito: 1 – São Paulo / 2 – Campinas / 3 – Itu
- Sexo – 1 dígito: 0 – Feminino / 1 – Masculino
- Escolaridade – 1 dígito: 1 – Curso superior completo com, pelo menos, dois anos de experiência profissional / 2 – Última série de curso superior / 3 – Última série do ensino médio / 4 – Última série do ensino fundamental II / 5 – Última série do ensino fundamental I / 6 – Analfabeto
- Faixa etária – 2 dígitos, uma vez que foi estabelecida uma subdivisão em seis faixas etárias para os informantes de curso superior completo e para os analfabetos: 1.1 – 25 a 29 / 1.2 –30 a 34 / 1.3 – 35 a 39 / 1.4 – 40 a 44 / 1.5 – 45 a 49 / 1.6 – 50 a 54 / 2.0 – 20 a 24 / 3.0 – 17 a 19 / 4.0 – 14 a 16 / 5.0 – 10 a 13 / 6.1 – 25 a 29 / 6.2 – 30 a 34 / 6.3 – 35 a 39 / 6.4 – 40 a 44 / 6.5 – 45 a 49 / 6.6 – 50 a 54
- Nível socioeconômico – 1 dígito: 1 – Classe alta alta / 2 – Classe alta / 3 – Classe média alta / 4 – Classe média baixa / 5 – Classe baixa / 6 – Classe baixa baixa.

Condições extraverbais de produção do diálogo:
1 dígito: 1 – Formal (entrevista) / 0 – Informal (conversação).

Da estruturação das Bases em campos e das anotações que eles contêm resulta:

- um código exclusivo para cada informante, o que permite a sua identificação. Exemplo: o código **101111** refere-se a informante de São Paulo (**1**), do sexo feminino (**0**), com curso superior completo (**1**), de 25 a 29 anos (**1.1**), de classe alta alta (**1**).

Código do Informante				
1	**0**	**1**	**11**	**1**
São Paulo	feminino	superior completo	25 a 29 anos	classe alta alta

- um código exclusivo para cada informante em cada tipo de interação dialógica:

Código do Informante e do Diálogo					
1	**0**	**1**	**11**	**1**	**0**
São Paulo	feminino	superior completo	25 a 29 anos	classe alta alta	diálogo informal

- um código exclusivo para cada item lexical, dentre cerca de 180 mil ocorrências, o qual permite a identificação do informante que o produziu – região de origem, sexo, escolaridade, faixa etária, nível socioeconômico –, das condições de produção do diálogo – entrevista / conversação –, como também da posição do item lexical dentro do enunciado e dentro do discurso. Exemplo:

Código Lexical								
1	**0**	**1**	**11**	**1**	**0**	**01**	**01**	**001**
São Paulo	feminino	superior completo	25 a 29 anos	classe alta alta	diálogo informal	primeiro discurso	primeiro enunciado	primeiro item lexical

Ou seja, esse código refere-se ao primeiro item lexical (001) do primeiro enunciado (01) do primeiro discurso (01), produzido por informante de São Paulo, com curso superior completo, de 25 a 29 anos, de classe alta alta, em situação informal de interação dialógica.

O *Sistema CorPor:*

- inclui uma amostra representativa da variante paulista do português do Brasil;
- tem a finalidade de servir para estudos da língua oral do português paulista em diversas áreas e para diferentes finalidades;
- inclui dados autênticos, provenientes de variedades sociolinguísticas do português falado de São Paulo, coletados em situações reais de uso, em condições formal e informal do contexto interacional;
- reúne conteúdo recolhido criteriosamente, de acordo com diretrizes linguísticas e extralinguísticas, o que torna viável a composição de tantos *corpora* e léxicos quantas são as variáveis controladas e suas possibilidades combinatórias.