CORPUS DE FALA TRANSCRITO – CODIFICAÇÃO E ESTRUTURAÇÃO DOS DADOS

Para a *Geração de Bases de Informações Ortográfico-Fonéticas* em plataforma PC, seguimos os mesmos critérios e procedimentos utilizados na constituição das Bases em computadores de grande porte[1], com as devidas adaptações para os sistemas atuais de processamento de dados. Isso porque os critérios e procedimentos adotados na constituição do *corpus* de fala transcrito para tratamento em *mainframes* respondem às exigências apresentadas na literatura atual sobre o assunto, que expressa a tendência internacional de pesquisas linguísticas baseadas em *corpus*. Vejamos o *corpus* da proposta segundo os critérios para definição de *corpus* apresentados por Berber Sardinha (1999c; 2000c):

- a origem: "Os dados devem ser autênticos".

  Os dados são autênticos – são dados provenientes de variedades sociolinguísticas do português falado de São Paulo, coletados em situações reais de uso, em condições de produção formal e informal de diálogos entre o informante e o documentador, colhidos, portanto, de atos reais da fala;
- o propósito: "O corpus deve ter a finalidade de ser um objeto de estudo linguístico".

  O *corpus* foi constituído com a finalidade de servir para estudos da língua oral do português paulista em diversas áreas e para diferentes finalidades;
- a composição: "O conteúdo do corpus deve ser criteriosamente escolhido".

  O conteúdo do *corpus* foi criteriosamente escolhido, em função de diretrizes linguísticas e extralinguísticas que nortearam a sua coleta, expostas na seção anterior e representadas graficamente no *Diagrama de Distribuição dos Informantes*;
- a formatação: "Os dados do corpus devem ser legíveis por computador".

---

[1] ZAPPAROLI CASTRO MELO, Zilda Maria. *Análise do comportamento fonológico da juntura intervocabular no português do Brasil (variante paulista). Uma pesquisa linguística com tratamento computacional*. São Paulo, 1980. Tese (Doutorado em Linguística) – Programa de Pós-Graduação em Linguística do Departamento de Linguística da Universidade de São Paulo. /Disponível na Coleção Didática da Biblioteca Central da FFLCH/USP/. v.1, t.1: 58-135.

A codificação e a estruturação dos dados, conforme *Diagrama de Registro do Informante*, apresentado mais adiante, estão a serviço do armazenamento, processamento e recuperação dos dados por computador;

- a representatividade: "O corpus deve ser representativo de uma língua ou variedade".
  O *corpus* é uma amostra representativa da variante paulista do português do Brasil;

- a extensão: "O corpus deve ser vasto para ser representativo".
  O *corpus* tem a dimensão pequeno-médio, com cerca de 180 mil itens lexicais, dimensão média de *corpora* em uso em pesquisas na área da *Linguística de Corpus* (BERBER SARDINHA, 2002)[2].

Na codificação e estruturação dos dados, conforme exposto a seguir, visamos à adequação da pesquisa a um modelo sistêmico que permitisse ampla análise.

Para que o *corpus* pudesse receber tratamento computacional, na sua constituição, baseamo-nos em critérios predeterminados. O registro[3] dos dados foi planejado para que eles pudessem ser armazenados e recuperados por sistemas computacionais.

Trata-se de *corpus* eletrônico anotado, que traz informações que permitem identificar as variáveis linguísticas (a palavra, a sua posição no enunciado, bem como a do enunciado no discurso, a sua transcrição ortográfica e fonética, a juntura ou o tipo de encontro fônico que mantém com a palavra antecedente e com a subsequente) e extralinguísticas (região de origem, sexo, nível de escolaridade, faixa etária, nível socioeconômico, condições de produção do diálogo) controladas na recolha do *corpus* de língua oral e na sua transcodificação.

Assim, na codificação e estruturação dos dados, levamos em conta não apenas a possibilidade de tratamento estatístico por computador, mas também o oferecimento de materiais de estudo para finalidades diversas. Em função disso, os campos de

[2] Quanto ao tamanho, o *corpus* pode ser pequeno, pequeno-médio, médio, médio-grande e grande, sendo que um *corpus* pequeno tem menos de oitenta mil palavras e um grande, 10 milhões ou mais palavras.

[3] A palavra registro, aqui, é empregada no sentido de conjunto de informações transcritas.

reconhecimento de registro foram estruturados em quatro níveis, para que se pudessem identificar todas as relações lógicas existentes entre eles, para efeito de recuperação.

O *Diagrama de Registro do Informante* e a tabela de *Codificação do Registro do Informante* mostram a anotação dos dados, a sua estruturação, os seus interrelacionamentos e as muitas possibilidades de sua recuperação em função de interesses de estudo.

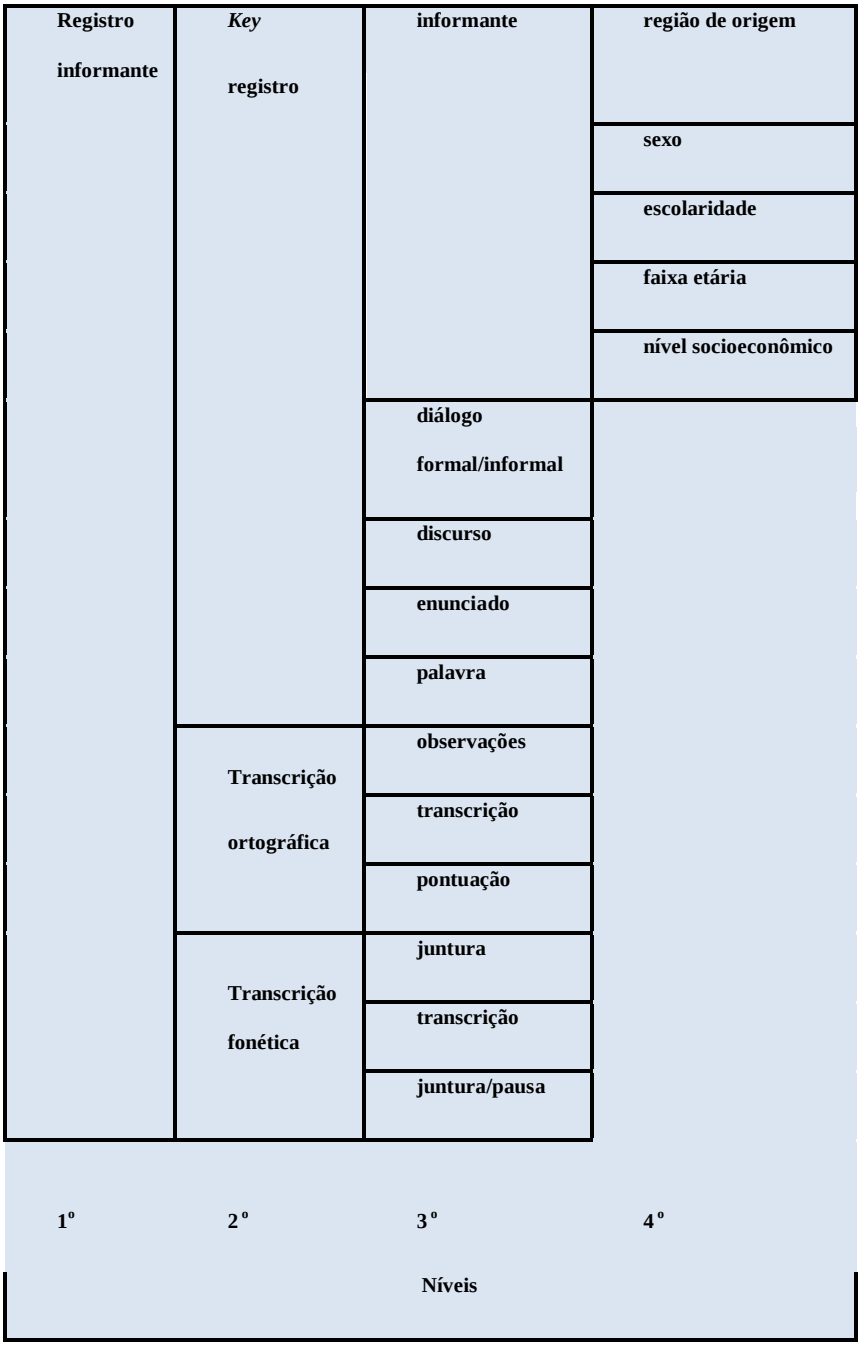

Figura – Diagrama de Registro do Informante

Conforme observado no *Diagrama de Registro do Informante*, os registros são compostos de campos – *Key* Registro ou Código Lexical (Identificação do Informante, Tipo de Diálogo, Discurso, Enunciado, Palavra), Transcrição Ortográfica (Observações − relativas a crases, desvios léxico-morfossintáticos, estrangeirismos, silabadas − Transcrição, Pontuação), Transcrição Fonética (Juntura Sílaba Inicial, Transcrição, Juntura Sílaba Final / Pausa, Divisão Silábica / Acento).

Tabela – Codificação do Registro do Informante

Nível	**Campo**	**Conteúdo**
1.0 Registro – Informante	alfanumérico	*key* do registro / transcrição ortográfica / transcrição fonética
1.1 *key* do registro	14 dígitos	informante / registro / discurso / enunciado / palavra
1.1.1 Informante	6 dígitos	cidade / sexo / escolaridade / idade / nível socioeconômico
1.1.1.1 Cidade	1 digito	1 – São Paulo 2 – Campinas 3 – Itu
1.1.1.2 Sexo	1 dígito	0 – Feminino 1 – Masculino
1.1.1.3 Escolaridade	1 dígito	1 – curso superior completo com, pelo menos, dois anos de experiência profissional 2 – Último ano de curso superior 3 – Último ano de ensino médio 4 – Último ano de ensino fundamental II 5 – Último ano de ensino fundamental I 6 – Analfabeto
1.1.1.4 Idade[4]	2 dígitos	11 – 25 a 29 12 –30 a 34 13 – 35 a 39 14 – 40 a 44 15 – 45 a 49 16 – 50 a 54 20 – 20 a 24

[4] Utilizamos dois algarismos para a indicação das faixas etárias dos informantes de curso superior completo e dos analfabetos, uma vez que foi estabelecida uma subdivisão em seis faixas etárias para os mesmos. Dada a necessidade de padronização para o número de registro de todos os informantes, na codificação dos entrevistados dos outros níveis de escolaridade, foi acrescentado um algarismo – 0 (zero) – após o número correspondente à faixa etária.

Nível	Campo	Conteúdo
		30 – 17 a 19 40 – 14 a 16 50 – 10 a 13 61 – 25 a 29 62 – 30 a 34 63 – 35 a 39 64 – 40 a 44 65 – 45 a 49 66 – 50 a 54
1.1.1.5 Nível sócioeconômico	1 dígito	1 – Classe alta alta 2 – Classe alta 3 – Classe média alta 4 – Classe média baixa 5 – Classe baixa 6 – Classe baixa baixa
1.1.2 Diálogo	1 dígito	condições extraverbais de produção do diálogo: 0 – Informal 1 – Formal
1.1.3 Discurso	2 dígitos	unidade no diálogo – corresponde à elocução do informante compreendida entre duas intervenções do entrevistador
1.1.4 Enunciado (ou frase)	2 dígitos	unidade no discurso – segmentação do discurso nos enunciados que o compõem.
1.1.5 Palavra (ou unidade léxica)	3 dígitos	unidade no enunciado – segmentação do enunciado nos seus constituintes léxicos
1.2 Transcrição ortográfica	caracteres	observações / transcrição / pontuação
1.2.1 Observações	1 dígito	1 – alguns desvios lexicais, morfológicos e sintáticos 2 – crase 3 – crase e desvio 4 – palavra estrangeira 5 – silabada 6 – nome próprio 7 – palavra estrangeira e nome próprio 8 – sigla 9 – desvio e nome próprio
1.2.2 Transcrição	alfa	conforme sistema ortográfico
1.2.3 Pontuação	2 dígitos	conforme codificação utilizada
1.3 Transcrição fonética	caracteres	juntura sílaba inicial / transcrição / juntura sílaba final e pausa

Nível	Campo	Conteúdo
1.3.1 Juntura sílaba inicial	2 dígitos	de 2 a 101 – categorias de juntura
1.3.2 Transcrição	alfa	conforme *Alfabeto Fonético Internaiconal*
Juntura sílaba final / Pausa	2 dígitos	1 – pausa real 2 a 101 – categorias de juntura

O Sistema CorPor − com o Banco Informatizado do Português Falado de São Paulo, numa estrutura de anotação de variáveis lingüísticas e extralingüísticas por níveis e subníveis, e com recursos de pesquisas da linguagem SQL e de um editor de textos – oferece a vantagem de facilidade e flexibilidade para a recuperação automática de um bom número de *corpora*[5] de estudo. Além disso, pela sua organização lexical, permite a exploração de diferentes relações estabelecidas entre os campos – para cada item lexical – e entre os registros – entre os vários itens lexicais.

É possível, então, extrair desde o *corpus* integral e conjunto, constituído pelo total das informações dos 432 inquéritos realizados (216 informantes em dois tipos de inquérito – entrevistas e bate-papo), até diferentes *subcorpora* quantas são as variáveis que foram controladas e indexadas – *corpus* por regiões, por sexos, por faixas etárias, por níveis de escolaridade, por níveis socioeconômicos, por condições extraverbais de produção do diálogo –, *corpora* menores – constituídos pelo conjunto das informações de seis inquéritos – e muitos outros, de dimensões várias, pelas possibilidades de cruzamentos das variáveis anotadas.

Ou seja, como já observado e se pôde visualizar no *Diagrama de Registro do Informante* e na tabela de *Codificação do Registro do Informante*, a estrutura do Sistema permite a recuperação dos dados por quaisquer campos ou pelo cruzamento deles – todos os campos podem ser cruzados –, do que resultam tantos *corpora* quantas são as variáveis lingüísticas e extralingüísticas anotadas e suas diferentes possibilidades combinatórias.

Seguem algumas possibilidades de composição de *corpora* para posterior tratamento por programas de análise lingüística:

[5] Arquivos de texto.

- *Corpus* integral: inquéritos formais e informais de 216 informantes.
- *Corpus* por condições extraverbais de produção dos diálogos, com 216 inquéritos cada um: (a) formal; (b) informal.
- *Corpus* por região, com 72 inquéritos formais e 72 inquéritos informais cada um: (a) São Paulo; (b) Campinas; (c) Itu.
- *Corpus* por sexo, com 108 inquéritos formais e 108 inquéritos informais cada um : (a) masculino; (b) feminino.
- *Corpus* por nível de escolaridade, com 36 inquéritos formais e 36 inquéritos informais cada um: (a) superior completo; (b) superior incompleto – último ano; (c) segundo grau (atual ensino médio) – último ano; (d) primeiro grau (atual ensino fundamental) – último ano; (e) primeiro grau – quarto ano; (6) analfabeto.
- *Corpus* por faixa etária, com 36 inquéritos formais e 36 inquéritos informais cada um: (a) a partir de 25 anos – superior completo; (b) 20 a 24 anos; (c) 17 a 19 anos; (d) 14 a 16 anos; (e) 10 a 13 anos; (f) a partir de 25 anos –.analfabeto.
- *Corpus* por faixa etária para os informantes de curso superior completo e para os analfabetos – subdivisão em seis faixas etárias, com uma dupla (homem / mulher) em cada cidade, num total de 12 informantes de curso superior completo e de 12 analfabetos em cada região, com os dois tipos de inquéritos: (a) 25 a 29 anos; (b) 30 a 34 anos; (c) 35 a 39 anos; (d) 40 a 44 anos; (e) 45 a 49 anos; (f) 50 a 54 anos.
- *Corpus* por nível socioeconômico, com 36 inquéritos formais e 36 inquéritos informais cada um: (a) classe alta alta; (b) classe alta; (c) classe média alta; (d) classe média baixa; (e) classe baixa; (f) classe baixa baixa.

Alguns cruzamentos possíveis:

- *Corpus* por região / tipo de diálogo, com 72 inquéritos cada um: (a) São Paulo / formal; (b) São Paulo / informal; (c) Campinas / formal; (d) Campinas / informal; (e) Itu / formal / (f) Itu / informal.

- *Corpus* por região / sexo, com 36 inquéritos cada um: (a) São Paulo / masculino; (b) São Paulo / feminino; (c) Campinas / masculino; (d) Campinas / feminino; (e) Itu / masculino; (f) Itu / feminino.
- *Corpus* por região / nível de escolaridade, com 12 inquéritos cada um: (a) São Paulo / superior completo; (b) São Paulo / superior incompleto – último ano; (c) São Paulo / médio – último ano; (d) São Paulo / fundamental – último ano; (e) São Paulo / fundamental – quarto ano; (f) São Paulo / analfabeto; (g) Campinas / superior completo; (h) Campinas / superior incompleto – último ano; (i) Campinas / médio – último ano; (1j) Campinas / fundamental – último ano; (k) Campinas / fundamental – quarto ano; (l) Campinas / analfabeto; (m) Itu / superior completo; (n) Itu / superior incompleto – último ano; (o) Itu / médio – último ano; (p) Itu / fundamental – último ano; (q) Itu / fundamental – quarto ano; (r) Itu / analfabeto;
- *Corpus* por região / sexo / nível de escolaridade, com 6 inquéritos cada um: (a) São Paulo / masculino / superior completo; (b) São Paulo / feminino / superior completo; (c) São Paulo / masculino / superior incompleto – último ano; (d) São Paulo / feminino / superior incompleto – último ano; assim por diante.
- *Corpus* por região / tipo de diálogo / sexo / nível de escolaridade, com 6 inquéritos cada um: (a) São Paulo / formal / masculino / superior completo; (b) São Paulo / informal / masculino / superior completo; (c) São Paulo / formal / feminino / superior completo; (d) São Paulo / informal / feminino / superior completo; (e) São Paulo / formal / masculino / superior incompleto – último ano; (f) São Paulo / informal / masculino / superior incompleto – último ano; (g) São Paulo / formal / feminino / superior incompleto – último ano; (h) São Paulo / informal / feminino / superior incompleto – último ano; assim por diante.
- *Corpus* por região / nível socioeconômico, com 12 inquéritos cada um: (a) São Paulo / classe alta alta; (b) São Paulo / classe alta; assim por diante.
- *Corpus* por região / sexo / nível socioeconômico, com 6 inquéritos cada um: (a) São Paulo / masculino / classe alta alta; (b) São Paulo / feminino / classe alta; assim por diante.

A exploração de cada *corpus* por programas gerenciadores do léxico, além de apresentar o léxico por variável ou conjunto de variáveis selecionadas na sua extração, possibilita estudos contrastivos com os tratamentos efetuados para outras composições.