# SISTEMA CORPOR: *CORPORA* DO PORTUGUÊS FALADO DE SÃO PAULO

*Zilda Maria Zapparoli*

Universidade de São Paulo, CNPq, FAPESP, Brasil
zmz@usp.br

## Resumo

Utilizando o computador no armazenamento, tratamento e análise de dados autênticos de língua oral, o trabalho contempla a construção de *Bases de Informações Ortográfico-Fonéticas do Português Falado de São Paulo (São Paulo, Campinas e Itu) em Sistema de Banco de Dados Relacional*. As *Bases* incluem informações ortográficas e fonéticas do português falado de São Paulo, organizadas, relacionadas e armazenadas em função de anotações linguísticas e extralinguísticas. Os resultados da utilização de *recursos da Informática na Linguística* podem servir de subsídios às áreas que se servem de *recursos da Linguística na Informática*, a exemplo do processamento automático da língua portuguesa.
**Palavras-chave**: Linguística Informática, Tecnologias Informatizadas nos Estudos Linguísticos, Projeto CorPor, Sistema de Banco de Dados Relacional, Bases de Informações Ortográfico-Fonéticas do Português Falado de São Paulo, *Corpora* Eletrônicos do Português Falado de São Paulo

## 1. Introdução

É numa visão por excelência interdisciplinar entre a Linguística e a Informática que se insere a presente investigação. Estudando fatos da língua em uso e utilizando o computador no armazenamento, no tratamento e análise de dados autênticos de língua oral, o trabalho dedica-se à construção de *Bases (ou Bancos) de Informações Ortográfico-Fonéticas, de Corpora e de Léxicos do Português Falado de São Paulo (São Paulo, Campinas e Itu)*, compatíveis com os sistemas computacionais atuais, a partir dos dados coletados para a tese de doutorado (1980) e das informações contidas nas Bases então geradas em sistemas de computadores de grande porte, conforme em [1].

As Bases estão armazenadas em formato específico de Banco de Dados Relacional, o que oferece aos pesquisadores facilidade, rapidez e confiabilidade na pesquisa (consulta), na recuperação (acesso) e no tratamento (exploração) automáticos de extensos e variados dados do português paulista para o desenvolvimento de estudos de aspectos diversos da língua – fonéticos, fonológicos, lexicais, morfológicos, sintáticos, textuais e discursivos.

Nessa perspectiva, a investigação insere-se no campo da *Linguística Informática*, apoiando-se em áreas que partilham a crença nos resultados positivos advindos da interação entre Linguística e Informática − parte da utilização de *recursos da Informática na Linguística* para a composição de Bases de Informações, que, por sua vez, oferecem subsídios às áreas que se servem de *recursos da Linguística na Informática*, a exemplo do processamento automático da língua portuguesa.

## 2. Procedimentos metodológicos

### 2.1. C*orpus* de língua oral

As amostras das falas dos informantes, recolhidas de 1972 a 1973, totalizam 54 horas de gravações entre documentador e 216 informantes paulistas (São Paulo, Campinas, Itu), de diferentes sexos, escolaridades, faixas etárias e níveis socioeconômicos, num total de 432 diálogos, visto que incluem dois tipos de interação dialógica – entrevistas e conversações.

O *Diagrama de Distribuição dos Informantes* (Figura 1) apresenta a distribuição dos informantes nas variáveis e nos diversos níveis de cada uma delas, demonstrando as várias possibilidades de relações contrastivas.

### 2.2. C*orpus* de fala transcrito para tratamento computacional

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, o tipo de encontro fônico que mantém com a palavra antecedente e com a subsequente) e extralinguísticas (região de origem, sexo, escolaridade, faixa etária, nível socioeconômico, condições de produção do diálogo), do que resulta um código exclusivo para cada item lexical, dentre cerca de 180 mil ocorrências.

A maneira como as informações estão codificadas e estruturadas confere às Bases funcionalidade, com possibilidades de extração de diferentes *corpora* e léxicos por variáveis linguísticas e extralinguísticas.

### 2.3 Sistema gerenciador de banco de dados

As *Bases de Informações* estão armazenadas em um Sistema de Banco de Dados – *Firebird*. A estrutura dos dados segue o modelo de dados relacional, de forma que as Bases contêm informações linguísticas e extralinguísticas com as diferentes relações existentes entre os dados armazenados. As Bases constituem, assim, uma coleção de dados ortográficos e fonéticos do português falado de São Paulo, organizados, relacionados e armazenados em função de anotações linguísticas e extralinguísticas.

O ambiente de programação utilizado é o *Delphi*, produzido pela *Borland Software Corporation*, que utiliza a Linguagem Pascal com extensões orientadas a objetos (*Object Pascal*), associada a recursos da Linguagem Estruturada de Pesquisa (*Structured Query Language* – SQL) [2].

Além de recursos de pesquisa – para o acesso às informações das Bases –, o Sistema abrange recursos de um editor de textos – para os trabalhos de edição dos resultados das pesquisas às Bases de Informações.

Para acesso dos usuários e para pesquisas por meio dos comandos da linguagem SQL, não somente as Bases de Informações Ortográfico-Fonéticas, como também *Corpora* e Léxicos (Dicionários) gerados a partir delas, integram o *Sistema CorPor*, cada um deles compondo um módulo com seus registros e campos.

## 3. Principais produtos do sistema

### 3.1. Bases de informações ortográfico-fonéticas do português falado de São Paulo

As *Bases de Informações Ortográfico-Fonéticas do Português Falado de São Paulo* contêm todas as informações de cada um dos 216 informantes. As informações estão organizadas pela ordem de registro de gravação e de acordo com os procedimentos de anotação e de estruturação adotados. Ou seja, as Bases trazem a informação lexical organizada tendo em vista relações com dados linguísticos e extralinguísticos. A Tabela 1 traz um extrato das *Bases*.

### 3.2. Corpora eletrônicos do português falado de São Paulo (bases de dados textuais)

*Corpora Eletrônicos do Português Falado de São Paulo* (*Bases de Dados Textuais*) podem ser extraídos das *Bases de Informações Ortográfico-Fonéticas*, com variadas possibilidades de exploração por programas de análise linguística, como em [3], e suscetíveis de aplicação em diferentes áreas dos estudos da linguagem e de áreas afins. Podem ser gerados tantos *corpora* quantas são as variáveis linguísticas e extralinguísticas anotadas e suas diferentes possibilidades combinatórias. Abaixo, extrato do *corpus* do português falado culto de São Paulo – *corpus* de fala transcrito de informantes paulistanos com curso superior completo. Nas *Bases de Dados Textuais*, os códigos de pontuação foram substituídos pelos sinais correspondentes.

Código Lexical: 1011111 – Informante de São Paulo (1), do sexo feminino (0), com curso superior completo (1), 25 a 29 anos (11), classe alta alta (1), registro formal de interação dialógica (1)

*De profissional ou...*

*Nossa mãe! depende do dia —isso que é o problema, entende?— Eu optei um curso de complementação pedagógica e, agora, tem uns trabalhos, para apre/ apresentar, então, eu estou fazendo esses trabalhos: tem o de sociologia —para entregar— e um sobre o INCRA; tem uma tese que eu estou corrigindo a parte de português, toda parte de ortografia e construção —é de minha prima que tra/ trabalha no Butantã, sabe?; ela está fazendo uma tese sobre educação e saúde; também estou dando uma olhada na tese dela de manhã—. Tsu que mais que eu faço de manhã?... tempo de aulas, corrige-se provas; agora vai mudar —engano— vou mudar também; agora, de manhã, vou dar aula no Mackenzie; à tarde, venho para cá —varia—.*

### 3.3. Léxico de frequência ortográfico-fonético do português falado de São Paulo

O *Léxico de frequência*, constituído a partir do *corpus* integral, traz, para cada palavra em sua transcrição ortográfica (coluna 3), as correspondentes transcrições fonéticas, com e sem separação silábica (colunas 5 e 4 respectivamente), com anotação da frequência da unidade fonética (coluna 2) e da frequência acumulada da unidade ortográfica (coluna 1), conforme amostra apresentada na Tabela 2.

### 3.4. Léxico de junturas intervocabulares do português falado de São Paulo

O *Léxico de Junturas Intervocabulares*, também construído a partir das Bases *de Informações Ortográfico-Fonéticas*, inclui a categoria de juntura (coluna 1), a combinatória acentual das sílabas intervocabulares (coluna 2), a transcrição fonética silábico-lexical das ocorrências de juntura intervocabular – manifestações de encontros fônicos lexicais que se dão nos limites de duas ou mais fronteiras de palavras – (colunas 3, 4, 5 e 6), com a correspondente transcrição ortográfica (colunas 7, 8, 9 e 10), de acordo com a amostra exposta na Tabela 3.

## 4. Conclusões

A investigação, reflexo de uma sintonia real entre as tendências atuais dos estudos da linguagem e as tecnologias de ponta, pode oferecer contribuições e benefícios: (1) para responder à demanda, no Brasil, de *corpora* eletrônicos de transcrições de fala que

contêm transcrições fonéticas; (2) para a ampliação do intercâmbio científico e tecnológico e para o enriquecimento da interação entre as ciências exatas e a ciência da linguagem; (3) no âmbito da Linguística, para a disseminação do uso de pesquisas baseadas em *corpora* e de tecnologias informatizadas nos estudos da língua em uso; (4) na interface entre a Linguística e a Informática, pelo oferecimento de conhecimentos linguísticos para o desenvolvimento, treinamento e avaliação de sistemas de processamento da fala do português variante brasileira − reconhecimento e síntese −, uma das áreas de maior complexidade do Processamento de Línguas Naturais.

## 5. Agradecimentos

Agradeço a Manoel Vidal Castro Melo a assessoria em análise e programação para o desenvolvimento do Sistema em *Mainframe* e a Edenis Gois Cavalcanti, para a criação do Sistema em PC.

## 6. Referências

[1] Z. M. Zapparoli Castro Melo, "Análise do comportamento fonológico da juntura intervocabular no português do Brasil (variante paulista). Uma pesquisa linguística com tratamento computacional", Tese de Doutorado, Universidade de São Paulo, São Paulo, SP, Brasil, 1980.

[2] C. Szyperski, *Component Software: Beyond Object-Oriented Programming*. Boston: Addison-Wesley, 1998.

[3] Z. M. Zapparoli, A. Camlong, *Do Léxico ao Discurso pela Informática*. São Paulo: EDUSP/FAPESP, 2002, 256 p. + CD-ROM.

[4] International Phonetic Association, *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press, 1999.

## Figura e tabelas

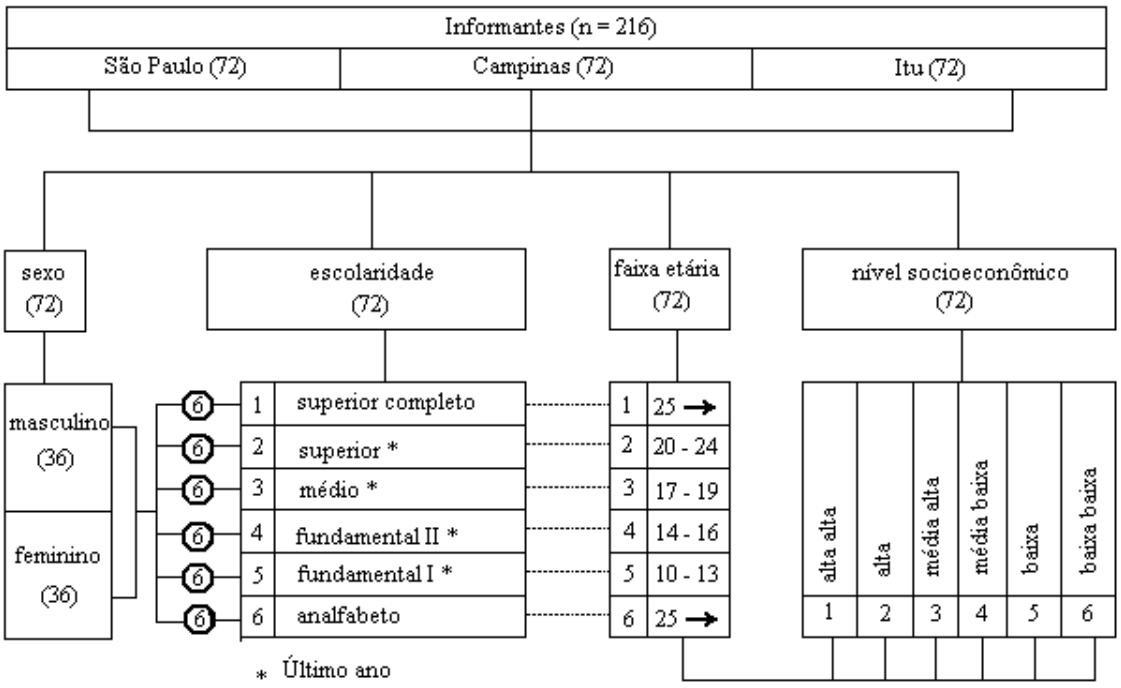


Figura 1. *Diagrama de Distribuição dos Informantes*

Tabela 1. *Bases de informações ortográfico-fonéticas do português falado de São Paulo*

| ***Chave[1]*** | ***Código Lexical[2]*** | ***Obs.[3]*** | ***Transcrição Ortográfica[4]*** | ***Pont.[5]*** | ***J / SI[6]*** | ***Transcrição Fonética[7]*** | ***J SF/ P[8]*** |
|---|---|---|---|---|---|---|---|
| 1 | 10111100101001 | | já | | | 'ʒa | 101 |
| 2 | 10111100101002 | | viajei | | 101 | vi a 'ʒej | 38 |
| 3 | 10111100101003 | | um | | 38 | ʃũ | 101 |
| 4 | 10111100101004 | | bocadinho | 1 | 101 | bo ka 'di ɲu | 1 |
| 5 | 10111100201001 | | eu | | | 'ew | 101 |
| 6 | 10111100201002 | | fui | | 101 | 'fuj | 101 |
| 7 | 10111100201003 | | pela | | 101 | pe l | 5 |
| 8 | 10111100201004 | 6 | Associação | | 5 | a so sja 'sãw̃ | 101 |
| 9 | 10111100201005 | | dos | | 101 | dus | 100 |
| 10 | 10111100201006 | | Professores | | 100 | pɾo fe 'so ɾiz | 101 |
| 11 | 10111100201007 | | de | | 101 | di | 101 |
| 12 | 10111100201008 | 6 | Francês | 4 | 101 | fɾã 'sej | 32 |
| 13 | 10111100201009 | | sabe | 7 | 32 | 'sa bi | 1 |
| 14 | 10111100301001 | | olha | 4 | | 'ɔ ʎa | 37 |
| 15 | 10111100301002 | | o | | 37 | u | 101 |
| 16 | 10111100301003 | | curso | | 101 | 'kuɾ sw̃ | 15 |
| 17 | 10111100301004 | | em | | 15 | ĩ | 101 |
| 18 | 10111100301005 | | si | | 101 | 'si | 1 |
| 19 | 10111100301006 | | não | 3 | | 'nũ | 1 |
| 20 | 10111100301007 | | não | | | 'nũ | 101 |

[1] Ordem
[2] Codificação para identificação do item lexical − informante, tipo de diálogo, discurso, enunciado e palavra
[3] Codificação para desvios léxico-morfossintáticos, siglas, nomes próprios, palavras estrangeiras
[4] Transcrição ortográfica
[5] Codificação para pontuação
[6] Codificação para juntura sílaba inicial
[7] Transcrição fonética [4]
[8] Codificação para juntura sílaba final / pausa real

Tabela 2. *Léxico de frequência ortográfico-fonético do português falado de São Paulo*

| ***Freq Ort Acum*[1]** | ***Freq Fon*[2]** | ***Transcrição Ortográfica*[3]** | ***Transcrição Fonética*[4]** | ***Transcrição Fonética / Sílaba*[5]** |
|---|---|---|---|---|
| 2 | 2 | abacate | aba'kaʦi | a ba 'ka ʦi |
| 1 | 1 | abacaxi | abaka'ʃi | a ba ka 'ʃi |
| 1 | 1 | abacaxis | jabaka'ʃiz | ja ba ka 'ʃiz |
| 1 | 1 | abaixo | a'baʃu | a 'ba ʃu |
| 2 | 1 | abaixo | a'bajʃu | a 'baj ʃu |
| 1 | 1 | abalado | aba'ladu | a ba 'la du |
| 1 | 1 | abandonar | abãdo'na | a bã do 'na |
| 2 | 2 | abandonei | abãdo'nej | a bã do 'nej |
| 1 | 1 | abandonou | abãdo'no | a bã do 'no |
| 1 | 1 | abatida | aba'ʦida | a ba 'ʦi da |
| 3 | 3 | aberta | a'bɛɾta | a 'bɛɾ ta |
| 4 | 1 | aberta | a'bɛrta | a 'bɛr ta |
| 1 | 1 | abertas | a'bɛɾtas | a 'bɛɾ tas |
| 1 | 1 | aberto | a'bɛɽtu | a 'bɛɽ tu |
| 2 | 1 | aberto | a'bɛʀtw | a 'bɛʀ tw |
| 4 | 2 | aberto | a'bɛɾtu | a 'bɛɾ tu |
| 6 | 2 | aberto | a'bɛrtu | a 'bɛr tu |
| 7 | 1 | aberto | ja'bɛɾt | ja 'bɛɾ t |
| 8 | 1 | aberto | ja'bɛɾtu | ja 'bɛɾ tu |

[1] Frequência acumulada da transcrição ortográfica
[2] Frequência da transcrição fonética
[3] Transcrição ortográfica do item lexical
[4]Transcrição fonética do item lexical sem divisão silábica [4]
[5] Transcrição fonética silábico-lexical [4]

Tabela 3. *Léxico de Junturas Intervocabulares do português falado de São Paulo*

| ***Junt*[1]** | ***Ac*[2]** | ***Fon1*[3]** | ***Fon2*[4]** | ***Fon3*[5]** | ***Fon4*[6]** | ***Ort1*[7]** | ***Ort2*[8]** | ***Ort3*[9]** | ***Ort4*[10]** |
|---|---|---|---|---|---|---|---|---|---|
| 101 | TA | 'ʒa | vi a 'ʒej | | | já | viajei | | |
| 101 | AA | j̃ũ | bo ka 'di ɲu | | | um | bocadinho | | |
| 101 | TT | 'ew | 'fuj | | | eu | fui | | |
| 5 | AA | pe l | a so sja 'sãw̃ | | | pela | Associação | | |
| 100 | AA | dus | pɾo fe 'so ɾiz | | | dos | Professores | | |
| 101 | AA | di | fɾã 'sej | | | de | Francês | | |
| 37 | AA | 'ɔ ʎa | u | | | olha | o | | |
| 15 | AA | 'kuɾ sw̃ | ĩ | | | curso | em | | |
| 101 | TT | 'nũ | 'sej | | | não | sei | | |
| 33 | ATA | sj | 'ɛ | w | | se | é | o | |
| 15 | AA | 'kuɾ sw̃ | ĩ | | | curso | em | | |
| 101 | TA | 'si | si | | | si | se | | |
| 2 | AA | 'va lj | a | | | vale | a | | |
| 17 | AA | 'pẹ n | ĩ 'tẽj̃ ʤi | | | pena | entende | | |
| 27 | AT | maj z | 'ew | | | mas | eu | | |
| 101 | AA | 'wa ʃu | ki | | | acho | que | | |
| 101 | AA | pɾa | kɾi 'a | | | para | criar | | |
| 101 | AA | 'w ma | maj 'jɔɾ | | | uma | maior | | |

[1] Categoria de juntura intervocabular
[2] Tonicidade das sílabas intervocabulares – combinatória acentual do contexto intervocabular (T= tônica; A= átona)
[3] Transcrição fonética do vocábulo 1 da sequência vocabular [4]
[4] Transcrição fonética do vocábulo 2 da sequência vocabular [4]
[5] Transcrição fonética do vocábulo 3 da sequência vocabular [4]
[6] Transcrição fonética do vocábulo 4 da sequência vocabular [4]
[7] Transcrição ortográfica do vocábulo 1 da sequência vocabular
[8] Transcrição ortográfica do vocábulo 2 da sequência vocabular
[9] Transcrição ortográfica do vocábulo 3 da sequência vocabular
[10] Transcrição ortográfica do vocábulo 4 da sequência vocabular

Nota − Este texto é a versão em português do artigo publicado em ZAPPAROLI, Zilda Maria. CorPor System: Corpora of the Portuguese Language as Spoken in São Paulo. In: Iberian SLTech 2009 - I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages, 2009, Porto Salvo, Portugal. *Proceedings*… Porto Salvo: Designeed, 2009. p. 35-38, ISBN 9789899687819 e em formato digital – CD-ROM.