Sistema CorPor: uma contribuição para o processamento da fala do português variante brasileira

Z. M. Zapparoli, Professora Associada, USP, Brasil, E. G. Cavalcanti, doutorando, USP, Brasil

Resumo — **Apoiando-se em áreas que partilham a crença nos resultados positivos advindos da interação entre Linguística e Informática, este trabalho insere-se na área da *Linguística Informática* – parte da utilização de *recursos da Informática na Linguística* para a geração do *Sistema CorPor*, que, por sua vez, oferece contribuições às áreas que se servem de *recursos da Linguística na Informática*, a exemplo do processamento automático da língua portuguesa. O *Sistema CorPor* inclui informações ortográficas e fonéticas do português falado no estado de São Paulo (Capital, Campinas, Itu), organizadas, relacionadas e armazenadas em função de anotações linguísticas e extralinguísticas. As informações são de fundamental importância para o desenvolvimento, treinamento e avaliação de sistemas de processamento da fala do português variante brasileira – reconhecimento e síntese –, sobretudo em se tratando de sistemas que utilizam recursos de aprendizagem de máquina através da construção de regras adaptativas.**

Palavras-chave — **Bases de Informações Ortográfico-Fonéticas do Português Falado de São Paulo (*Databanks of Phonetic and Orthographic Information about the Portuguese Language as Spoken in São Paulo*), *Corpora* Eletrônicos do Português Falado de São Paulo (*Electronic Corpora of the Portuguese Language as Spoken in São Paulo*), Linguística Informática (*Linguistic Informatics*), Processamento Automático da Língua Portuguesa (*Automatic Processing of the Portuguese Language*), Sistema CorPor (*CorPor System),* Sistema de Banco de Dados Relacional *(Relational Database System*), Tecnologias Adaptativas nos Estudos Linguísticos (*Adaptive Technologies in Linguistic Studies*).**

I. Introdução

Por envolver o uso de ferramentas informáticas, o trabalho insere-se na interface entre Linguística e Computação e, pois, em área multidisciplinar. Dedica-se à constituição de *Bases de Informações Ortográfico-Fonéticas do Português Falado de São Paulo*, a partir das quais podem ser gerados *corpora* digitalizados de textos orais em português do Brasil, para a sua exploração por recursos computacionais, para diferentes finalidades, como a geração de léxicos, o exame de padrões da língua oral, o processamento de línguas naturais.

Em arquitetura de banco de dados relacional, o Sistema CorPor reúne *Bases de Informações Ortográfico-Fonéticas, Corpora e Léxicos do Português Falado de São Paulo (São Paulo, Campinas, Itu)*, gerados, inicialmente, para a tese de doutorado (1980), em sistemas de computadores de grande porte, conforme em [1].

O armazenamento das bases de textos de língua oral no *Sistema CorPor* facilita a manipulação e o tratamento das informações, contribuindo para suprir a carência de *corpora* eletrônicos com transcrições ortográficas e fonéticas, e de conhecimentos linguísticos necessários ao desenvolvimento de sistemas de processamento da fala.

II. Pressupostos teórico-metodológicos

Numa dimensão mais ampla, o trabalho insere-se na área da *Linguística Informática*. A *Linguística Informática*, como linha de investigação científica, propõe-se, de um lado, à *utilização de recursos da Informática na Linguística* para o armazenamento, processamento e recuperação quantitativa e qualitativa de informações linguísticas; de outro, à *utilização de recursos da Linguística na Informática* para o desenvolvimento de sistemas que exigem equipes multidisciplinares, nas quais se incluem linguistas, como sistemas de tradução automatizada, sistemas de ensino de línguas naturais a distância, sistemas de produção e reconhecimento de línguas naturais.

Ainda, concebendo a *Linguística Informática* como abrangendo as diferentes áreas em que as tecnologias informatizadas estão relacionadas aos estudos da linguagem – *Linguística de Corpus, Linguística Computacional e Processamento de Língua Natural* –, a proposta enquadra-se mais particularmente nos propósitos da *Linguística de Corpus* em uma de suas preocupações, que constitui a condição *sine qua non* para a sua existência – construção de *corpora* eletrônicos a partir de textos e discursos reais. A *Linguística de Corpus* é vista, aqui, mais do que um simples instrumento de trabalho, por acreditarmos que o emprego das tecnologias informatizadas – base da *Linguística de Corpus* – na exploração de grandes quantidades de dados da língua em uso trará informações inéditas sobre as línguas naturais.

III. Procedimentos Metodológicos

A. *Constituição do Corpus de Língua Oral*

Os critérios rigorosos utilizados para a constituição do *corpus* de língua oral para fins do doutorado[1], mediante o controle de variáveis linguísticas – relativas às especificidades da língua falada – e variáveis extralinguísticas – região de origem, sexo, escolaridade, faixa etária, nível socioeconômico, condições extraverbais de interação dialógica – na seleção dos informantes e nos critérios de armazenamento dos dados,

[1] Zapparoli, 1980, v.1, t.1.

permitiram a obtenção de uma amostra representativa do português falado paulista, passível de ser objeto de estudo em diferentes áreas dos estudos da linguagem e de áreas afins. Trata-se de *corpus* compilado, também conhecido como *corpus* de amostragem, porque é fixo, uma vez que foi compilado através de amostras pré-selecionadas.

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.

B. Constituição do Corpus de Fala Transcrito para Tratamento Computacional

O registro[2] dos dados foi planejado para que eles pudessem ser armazenados e recuperados por sistemas computacionais.

O *Diagrama de Registro do Informante* (Fig. 1) mostra a anotação dos dados, a sua estruturação, os seus inter-relacionamentos e as muitas possibilidades de sua recuperação em função do interesse de estudo.

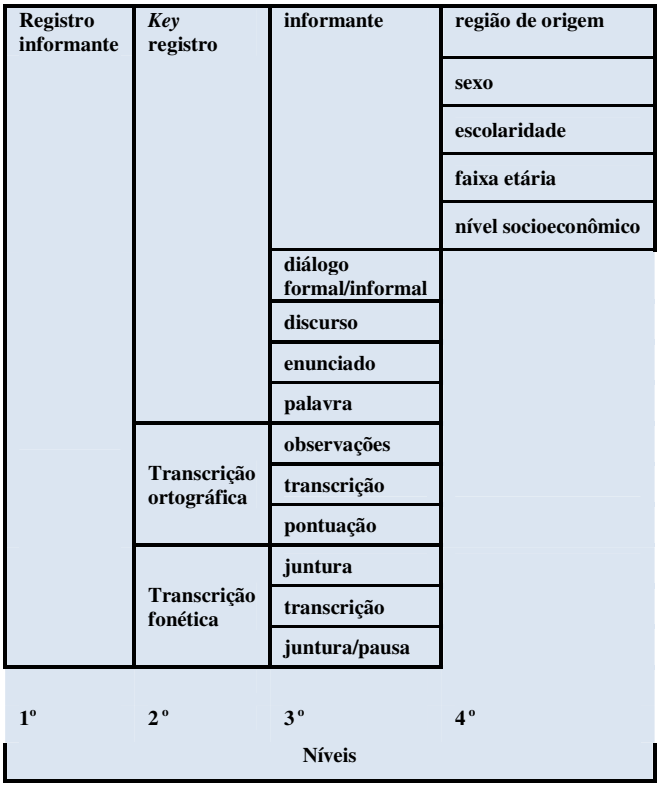

Fig.1. Diagrama de Registro do Informante

Trata-se, então, 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 – juntura – 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.

C. Sistema Gerenciador de Banco de Dados Relacional

As *Bases de Informações* estão armazenadas em Sistema de Banco de Dados e são manipuladas por meio do Sistema Gerenciador de Banco de Dados *Firebird*. A estrutura dos dados segue o modelo relacional, conforme Diagrama de Registro do Informante (Fig. 1), havendo uma correspondência entre os campos da tabela principal do banco de dados e os do diagrama. 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, com as diferentes relações existentes entre os dados armazenados.

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.

IV. Principais Componentes do Sistema

A. Bases de Informações Ortográfico-Fonéticas do Português Falado de São Paulo

As Bases contêm com todas as informações de cada um dos 216 inquéritos pela ordem de registro de gravação e de acordo com os critérios linguísticos e extralinguísticos que foram controlados na seleção dos informantes que forneceram material linguístico para a constituição da amostra.

B. Corpora Eletrônicos do Português Falado Paulista

Também em função das variáveis linguísticas e extralinguísticas, os *corpora* oferecem variadas possibilidades de exploração por programas de análise lingüística, como em [3].

C. Léxico de Frequência Ortográfico-Fonético

O Léxico de Frequência Ortográfico-Fonético traz, para cada palavra em sua transcrição ortográfica, as correspondentes transcrições fonéticas, sem e com separação silábica, com anotação da frequência da unidade fonética e da frequência acumulada da unidade ortográfica.

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

D. Léxico de Junturas Intervocabulares

O Léxico de Junturas Intervocabulares inclui as categorias de juntura intervocabular – encontros fônicos lexicais que se dão nos limites de duas ou mais fronteiras de palavras –, a transcrição ortográfica das ocorrências de juntura intervocabular com a correspondente transcrição fonética silábico-lexical e a combinatória acentual das sílabas intervocabulares.

Pelas limitações de espaço de um artigo, estendemo-nos, aqui, na apresentação das Bases, visto que elas são o suporte para a geração dos demais componentes.

As *Bases de Informações Ortográfico-Fonéticas do Português Falado de São Paulo* contêm todas as informações de cada um dos 216 informantes, num total de 432 inquéritos, visto que incluem, para cada informante, dois tipos de interação dialógica – entrevistas e conversações. As informações estão organizadas pela ordem de registro de gravação e de acordo com os procedimentos de anotação e de estruturação adotados.

Além da transcrição ortográfica e da transcrição fonética de cerca de 180 mil registros de itens lexicais, as *Bases* incluem anotações relativas a variáveis linguísticas (especificidades da língua oral, categorias de encontros fônicos intervocabulares) e a variáveis extralinguísticas que foram controladas na seleção dos 216 informantes que forneceram material linguístico para a constituição da amostra (região de origem, sexo, escolaridade, faixa etária e nível socioeconômico) e na produção dos diálogos (formal e informal). Ou seja, as Bases trazem a informação lexical organizada em função de relações com dados linguísticos e extralinguísticos, o que permite diferentes possibilidades combinatórias.

A Tabela I traz uma amostra das Bases.

TABELA I. 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]
126	10111100302001		chegamos			$& 'G9 MU	101
127	10111100302002		na		101	NA	101
128	10111100302003	6	França	4	101	'F>@ S	5
129	10111100302004		aquele		5	A 'K& LI	101
130	10111100302005		problema		101	P>O 'BL7 M	5
131	10111100302006		assim	2	5	A 'S1	1
132	10111100302007		a			A	101
133	10111100302008		guerra		101	'GE X	5
134	10111100302009		ainda		5	A '1 DA	101
135	10111100302010		está		101	'T	5
136	10111100302011		ali		5	A 'LI	101
137	10111100302012		presente	4	101	P>& 'Z3) TI	101
138	10111100302013		sabe	9	101	'SA BI	1
139	10111100302014		então	4		1 'T@%	101
140	10111100302015		você		101	'S&	37
141	10111100302016		entra		37	'3) T>A	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

Segue, a título de exemplificação, recorte discursivo extraído das Bases.

Código Lexical: 1011211 – Informante de São Paulo (1), do sexo feminino (0), com curso superior completo (1), 30 a 34 anos (12), classe alta alta (1), registro formal de interação dialógica (1):

Assim, eu... eu acho... eu... o indivíduo, quando escolhe a profissão por... por escolha, independente de influência de qualquer indivi/ qualquer pessoa, tem, ahn..., muito mais possibilidade de realizar se dentro do campo que escolheu. Por exem/ eu acho que dé/ dentro do sta/ do status atual, biblioteconomia é um campo altamente explorável, com boa remuneração econômica e com grande, ah, possibilidade de atividades e especializações; é um campo novo, com poucos especialistas, ih, dentro da o... dentro de São Paulo; quase todos eles são englobados pela Universidade de São Paulo. Acho assim, por exemplo, no momento, nós contamos, aqui na faculdade, com oito bibliotecários, todos de curso superior, dos quais quatro têm especialização em ciências biomédicas — inclusive eu tenho especialização—. Ah o ambiente de trabalho é ideal; não sei, porque não conheço, uhn..., éh..., nenhum; trabalhei um... durante dois anos como bibliotecária da... tsi... do Conselho Regional de Contabilidade do Estado, mas era uma biblioteca independente, com um único profissional; então, você não pode avaliar bem a... o relacionamento; isso eu vim sentir mais aqui na universidade; eu acho um campo... por ser um campo muito novo, todo profissional é muito unido; acho ideal, um trabalho muito bom, e nós trabalhamos aqui, em equipe; embora nós tenhamos todas nós setores bem definidos, pela carga de trabalho ser muito grande, nós todas trabalhamos em comum acôrdo, a ponto de podermos qualquer uma substituir a outra, a qualquer momento. É, eu acho que deu, assim, uma amplitude de trabalho muito grande, o que, muitas vezes, não se verifica em outros campos, né?; nós, graças a Deus, não tivemos esse problema.

V. CONTRIBUIÇÕES

Voltada a aspectos pouco explorados nos estudos linguísticos – se são raros, no Brasil, os *corpora* eletrônicos de transcrições de fala, mais ainda o são os *corpora com* transcrições fonéticas –, os resultados da investigação podem oferecer contribuições e benefícios: no âmbito da Linguística, pelo oferecimento de *corpora* digitalizados de textos autênticos da língua oral paulista para o desenvolvimento de estudos diversos; 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.

Estamos certos de que o êxito do processamento de línguas naturais depende tanto do avanço tecnológico como de novos conhecimentos linguísticos. A tarefa que nos cabe, como linguistas e falantes da língua portuguesa como língua

materna, consiste em oferecer contribuições para a aquisição de novos conhecimentos do português. Nesse sentido, o *Sistema CorPor*, que armazena as *Bases* em formato específico de Banco de Dados Relacional, oferece a estudiosos do português facilidade, rapidez e confiabilidade na pesquisa (consulta), na recuperação (acesso) e no tratamento (exploração) automáticos de extensos e variados dados do português falado paulista para o desenvolvimento de estudos de aspectos diversos da língua – fonéticos, fonológicos, lexicais, morfológicos, sintáticos, textuais e discursivos.

No que diz respeito a avanços na área da computação, destacamos que, em se tratando de sistemas com base em tecnologias adaptativas, os ganhos são significativos pela possibilidade de reconhecimento automático de padrões da língua oral paulista e, pois, pelo oferecimento de uma Base de Conhecimentos, indispensável na arquitetura de um sistema de processamento de língua natural.

VI. Considerações Finais

Para concluir, retomamos a referência inicial que fizemos à área da Linguística Informática. Neste trabalho de movimento duplo entre a Linguística e a Informática, de um lado, ressaltamos que as vantagens da utilização das Novas Tecnologias Digitais nas pesquisas linguísticas que desenvolvemos são indiscutíveis; de outro, vislumbramos resultados positivos de uma convergência do *Sistema CorPor* com a área da Inteligência Computacional, através de uma conexão do formalismo já desenvolvido a mecanismos adaptativos, para a geração de uma Base de Conhecimentos da língua oral paulista.

Expressamos o convite a estudiosos interessados no desenvolvimento dessa empreitada.

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.

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 orientada por Francis Henrik Aubert, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 1980.

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

[3] Z. M. Zapparoli e A. Camlong, *Do léxico ao discurso pela informática*. São Paulo: EDUSP/FAPESP, 2002.

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

Zilda Maria Zapparoli nasceu em Itu, São Paulo, Brasil, em 2 de agosto de 1945. É professora associada aposentada junto ao Departamento de Linguística da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo (FFLCH-USP), instituição em que obteve os títulos de Mestre, Doutor e Livre-Docente, e onde continua desenvolvendo atividades de ensino, pesquisa e orientação no Curso de Pós-Graduação em Linguística, área de Semiótica e Linguística Geral. É Bolsista de Produtividade em Pesquisa do CNPq e líder do Grupo Interdisciplinar de Pesquisas em Linguística Informática. Tem mais de trinta anos de atuação em Linguística Informática, com tese de doutorado, tese de livre-docência e trabalhos publicados na área. Integrou comissões e colegiados na USP, destacando-se os trabalhos relativos ao processo de informatização da FFLCH-USP, enquanto Membro da Comissão Central de Informática da USP e Presidente da Comissão de Informática da FFLCH-USP por cerca de treze anos.

Edenis Gois Cavalcanti nasceu em São Paulo, Brasil, em 2 de fevereiro de 1962. Possui graduação em Filosofia pela Faculdade de Filosofia Nossa Senhora Medianeira (1986) e mestrado em Semiótica e Linguística Geral pela Universidade de São Paulo (2005). Atualmente, desenvolve projeto de doutorado – Construção de Software para o Estudo da Língua Grega Clássica: o Sistema Nominal do Dialeto Ático na Visão Temática – pelo Departamento de Linguística, USP, área de Linguística Informática. É desenvolvedor de *software* nas linguagens Delphi e C# .NET (Visual Studio), plataformas Desktop e Web, integradas com bando de dados FireBird, MySql e Sql Server. Tem implantado – como metodologia de trabalho em sala de aula, rede municipal de ensino, São Paulo, Capital – projeto de avaliação *on-line*, SIGA WEB .NET (em www.fcavalcanti.com), no qual os alunos, com acesso totalmente gratuito, consultam conteúdos e realizam suas avaliações via *Internet*.