VOZ E TEXTO ORTOGRÁFICO-FONÉTICO NO SISTEMA CORPOR – *CORPORA* DO PORTUGUÊS FALADO DE SÃO PAULO

Tema: O Estado da Lusofonia

Subtema: Português nos Media e no Ciberespaço

Zilda Maria Zapparoli

Universidade de São Paulo, Brasil

RESUMO

Situada em área interdisciplinar, na interface entre Linguística e Informática, a investigação dedica-se à construção do Sistema CorPor, que inclui Bases de Informações Ortográficas e Fonéticas do Português Falado de São Paulo em Sistema de Banco de Dados Relacional. As informações estão organizadas, relacionadas e armazenadas através de anotações linguísticas e extralinguísticas: a língua oral paulista, observada numa perspectiva sincrônica, é, assim, passível de ser avaliada na sua diversidade − diferenças entre comunidades regionais, sexos, níveis de escolaridade, gerações, classes sociais, condições de produção dialógica. As Bases são o suporte a partir do qual os demais componentes do Sistema – *corpora* e léxicos – são gerados. A recuperação das informações linguísticas através do computador pode ser feita de maneira: a) multissensorial, pelo emprego coordenado de áudio (voz humana) e textos (transcrição ortográfica e fonética da fala); b) integrada, pela utilização simultânea dos meios de comunicação − voz e texto − sob a coordenação do computador; c) interativa, pela maneira com que se faz a recuperação das informações, isto é, ativamente, através de buscas, interligações. Voltada a aspectos pouco explorados nos estudos linguísticos − se são raros, no Brasil, os *corpora* eletrônicos de transcrições ortográficas de fala, mais ainda o são os *corpora* com transcrições fonéticas e com disponibilidade simultânea de voz e texto −, os resultados do trabalho podem oferecer contribuições e benefícios: no âmbito da Linguística, pelo oferecimento de *corpora* digitalizados de voz e 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. O Sistema CorPor está disponível a estudiosos interessados em estudos da língua oral do português do Brasil, para diferentes finalidades, no site www.corpor.fflch.usp.br.

Palavras-chave − Linguagem e Tecnologias, Linguística Informática, Linguística de Corpus, Sistema 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, Voz e Texto Ortográfico-Fonético do Português Falado de São Paulo, Léxicos Eletrônicos do Português Falado de São Paulo, Fonética e Fonologia, Lexicologia

INTRODUÇÃO

Alicerçada em uma interação real entre os estudos da linguagem humana e as tendências atuais de acesso à informação e à comunicação, a investigação que levou à geração do *Sistema CorPor* é por excelência interdisciplinar, situada na interface linguagem / tecnologias.

A utilização do computador como ferramenta auxiliar no decorrer de toda a pesquisa explica-se pela dimensão do *corpus*, para que haja uma interação mais fácil, rápida e segura com os materiais de estudo, e para que os dados possam ser tratados dentro de uma perspectiva quantiqualitativa.

O trabalho justifica-se pela demanda por Bases de Informações, *Corpora* e Léxicos Eletrônicos de Transcrições de Fala em Língua Portuguesa do Brasil, dada a sua restrita disponibilidade no momento em que a tendência internacional de pesquisa caminha no sentido de priorizar o emprego de uma abordagem baseada em *corpus*, pelas suas vantagens de possibilitar investigações com grandes volumes e variedades de textos representativos da língua em uso, com rapidez, exatidão, confiabilidade nos resultados e facilidade de armazenamento, recuperação e tratamento de informações.

Mais particularmente ainda, justifica-se pela carência de Bases de Informações, C*orpora* e Léxicos Eletrônicos que apresentem transcrições ortográficas e fonéticas com acesso simultâneo à voz dos informantes, bem como dados quantiqualitativos sobre o uso da língua portuguesa do Brasil.

Os *corpora*, como também os léxicos, são gerados a partir de *Bases de Informações* Ortográfico-Fonéticas do Português Falado de São Paulo (São Paulo, Campinas, Itu) em arquitetura de banco de dados relacional − *Sistema CorPor*.

Um dos produtos tecnológicos de relevo, recentemente incorporado ao Sistema e ponto central deste trabalho, é o componente áudio − voz humana − com textos − *Corpora* de Língua Oral com *Corpora* de Fala Transcrita Ortográfica e Foneticamente do Português Falado de São Paulo.

Além dos *corpora*, o *Sistema CorPor* inclui: a) Léxico de Frequência Ortográfico-Fonético do Português Falado de São Paulo; b) Léxico Ortográfico-Fonético de Junturas Intervocabulares do Português Falado de São Paulo; c) Léxico de Frequência Ortográfico-Fonético de Junturas Intervocabulares do Português Falado de São Paulo.

O *Léxico de Junturas Intervocabulares*, construído a partir do exame de diferentes manifestações de encontros fônicos que se dão no contexto intervocabular, representa estudo inédito. Dessa forma, no âmbito dos estudos fonológicos da língua contemplada pela pesquisa, a investigação não se limita à descrição dos segmentos fônicos – alofones – segundo o estruturalismo europeu, mas se estende, a partir dos pressupostos teóricos da Fonologia Gerativa, ao estudo da fonologia sintática − ou fonologia combinatória lexical −, resvalando, assim, o terreno da Morfofonologia − descrição dos processos fonológicos e/ou fonéticos que os segmentos fônicos sofrem quando se combinam na constituição de unidades maiores, as sequências vocabulares.

O Sistema ainda contém o menu *Ajuda*, com artigos, apresentações em eslaides e textos explicativos dos procedimentos metodológicos adotados na constituição do *corpus* de língua oral, na constituição do *corpus* de fala transcrito para tratamento computacional e na geração do Sistema CorPor.

Pautado em trabalhos que vimos realizando há cerca de quarenta anos para a descrição e análise de aspectos fonológicos, lexicais, textuais e discursivos do português falado de São Paulo, através de pesquisas baseadas em Bancos de Dados e em C*orpora* Eletrônicos, o *Sistema CorPor*: a) contempla investigação de natureza interdisciplinar, que envolve o recurso a contribuições de métodos e técnicas diversos e atuais, extrapolando, assim, as abordagens tradicionais; b) responde à preocupação atual dos estudos linguísticos, no que diz respeito à composição de acervos textuais eletrônicos para o exame da língua em situações reais de uso; c) disponibiliza para outras pesquisas Sistemas de Banco de Dados, C*orpora* e Léxicos Eletrônicos da Língua Oral Paulista com informações sonoras, ortográficas e fonéticas; d) utiliza tecnologias informatizadas e de comunicação na pesquisa e no conhecimento da língua portuguesa do Brasil, com contribuições para a implementação de sistemas com vistas à obtenção, representação e uso desse conhecimento através do computador.

1 PRESSUPOSTOS TEÓRICO-METODOLÓGICOS

O trabalho insere-se na área da *Linguística Informática* − parte da utilização de recursos da Informática na Linguística para a composição de Bases de Informações, *Corpora* e Léxicos do Português em Sistema de Banco de Dados, que, por sua vez, servirão de subsídios às áreas que se servem de *recursos da* Linguística na Informática, a exemplo do *Processamento Automático da Língua Portuguesa*.

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 pesquisa 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.

O trabalho fundamenta-se, também, nos quadros teóricos: a) da *Linguística Descritiva*, em sua preocupação com o que é dito ou escrito, por quem, onde e quando; b) da *Linguística Aplicada*, em sua concepção atual – que vai além de sua aplicação ao ensino/aprendizagem de línguas –, enquanto área multidisciplinar, dedicada às situações de uso da língua e, pois, ao desenvolvimento de pesquisas a partir da análise de *corpora*; c) da *Linguística Conversacional*, no cuidado em respeitar, na transcrição dos dados, as características específicas do discurso oral, evitando-se, na medida do possível, as normas tradicionais da linguagem escrita; d) da *Fonética* e da *Fonologia*, nos critérios que nortearam a transcrição e o exame do comportamento fonológico dos encontros fônicos que se dão na juntura lexical, ou seja, nos limites de duas ou mais fronteiras de palavras; e) da *Sociolinguística*, nas variáveis extralinguísticas que foram controladas na seleção dos informantes que forneceram material linguístico para a constituição dos *corpora*, com consequente oferecimento dos perfis dos dialetos e situações de uso contemplados através da técnica baseada em *corpus*; f) da *Lexicologia,* nos critérios que orientaram a questionada definição e delimitação da palavra para a segmentação do enunciado nos seus constituintes léxicos, bem como a constituição de léxicos; g) *da Linguística Textual e da Análise do Discurso Oral*, na descrição lexical quantiqualitativa, que oferece subsídios para estudos do texto e do discurso.

2 PROCEDIMENTOS METODOLÓGICOS

2.1 Constituição do *Corpus* de Língua Oral

O *corpus* de língua oral foi constituído a partir da gravação de diálogos – em situação de entrevistas e de conversações – entre o entrevistador e 216 informantes de três regiões do Estado de São Paulo – a Capital e duas regiões do interior, Campinas e Itu −, selecionados por critérios sociolinguísticos − região de origem, sexo, escolaridade, faixa etária, nível socioeconômico −, num total de 54 horas de gravação, de 432 diálogos e de cerca de 180 mil ocorrências de itens lexicais.

2.2 Constituição do *Corpus* de Fala Transcrito para Tratamento Computacional

Para a geração das *Bases de Informações Ortográfico*-Fonéticas, procurou-se responder às exigências apresentadas na literatura atual sobre o assunto, que expressa a tendência internacional de pesquisas linguísticas baseadas em *corpus:* a) os dados são autênticos – 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; b) o *corpus* foi constituído com a finalidade de servir para estudos da língua oral do português paulista em diversas áreas e para diferentes finalidades; c) o *corpus* tem o propósito de ser um objeto de estudo linguístico; d) o conteúdo do *corpus* foi criteriosamente escolhido, em função de diretrizes linguísticas e extralinguísticas que nortearam a sua coleta; e) a codificação e a estruturação dos dados estão a serviço do armazenamento, processamento e recuperação dos dados por computador; f) o *corpus* é uma amostra representativa da variante paulista do português do Brasil; g) 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*.

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.

Para a transcrição ortográfica, adotou-se o sistema de sinais escritos do alfabeto latino utilizado pela língua portuguesa, com convenções para a distinção de palavras homógrafas. Através de códigos, representaram-se as pausas, entonações e outras informações contextuais características do código falado.

A transcrição fonética é alofonêmica, por especificar alofones da língua. Utilizam-se os caracteres do *Alfabeto Fonético Internacional* e anotam-se, por códigos, a pausa efetivamente realizada na fala e o comportamento de encontros fônicos na juntura intervocabular.

3 Sistema CorPor − Sistema Gerenciador de Banco de Dados Relacional

Estudando fatos da língua em uso e utilizando o computador no armazenamento, na recuperação e no tratamento e análise de dados autênticos de língua oral, o *Sistema CorPor* reúne Bases de Informações Ortográfico-Fonéticas, Corpora e Léxicos do Português Falado de São Paulo em arquitetura de banco de dados relacional.

3.1 Armazenamento das informações

As *Bases de Informações* estão armazenadas no *Sistema CorPor*, Sistema de Banco de Dados Relacional, e são manipuladas por meio de Sistema Gerenciador de Banco de Dados (SGBD) − um conjunto de programas computadorizados − *software* ou ferramenta −, desenvolvidos numa determinada linguagem, que possibilitam o gerenciamento das funções de edição, consulta, controle e remoção de registros, campos ou tabelas de um Banco de Dados. Esse procedimento oferece a possibilidade de se estabelecerem relacionamentos entre os dados do Banco para a extração e análise de novas informações. 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 armazenamento das Bases em formato específico de Banco de Dados Relacional tem o propósito de oferecer a estudiosos do português facilidade, rapidez e confiabilidade na pesquisa (consulta), na recuperação (acesso) e no tratamento (exploração) automáticos de extensos e variados dados autênticos do português paulista para o desenvolvimento de estudos de aspectos diversos da língua – fonéticos, fonológicos, lexicais, morfológicos, sintáticos, textuais e discursivos − e para o desenvolvimento de sistemas de processamento da fala.

Não somente as Bases de Informações Ortográfico-Fonéticas, como também *Corpora* e Léxicos gerados a partir delas, integram o Sistema CorPor, cada um deles compondo um módulo − ou componente − com seus registros e campos.

3.2 Recuperação das informações

A maneira como as informações estão codificadas e estruturadas confere às Bases funcionalidade, com possibilidades de recuperação automática de diferentes *corpora* e léxicos por variáveis linguísticas e extralinguísticas. É possível extrair desde o *corpus* integral e conjunto, constituído pelo total das informações das 432 interações dialógicas realizadas com os 216 informantes, até diferentes *subcorpora* quantas são as variáveis linguísticas e extralinguísticas anotadas e suas diferentes possibilidades combinatórias, para posterior tratamento por programas de análise linguística.

Assim, a língua portuguesa, observada numa perspectiva sincrônica, é passível de ser avaliada na sua diversidade: diferenças entre comunidades regionais, diferenças entre sexos, diferenças entre níveis de escolaridade, diferenças entre gerações, diferenças entre meios sociais, diferenças ligadas às condições de produção do diálogo.

O componente *Corpora Eletrônicos do Português Falado Paulista − Bases de Dados Textuais* − passou a disponibilizar, recentemente, recursos multimídia, com a opção de recuperação simultânea de áudio e texto, do que resultam *corpora* de língua oral que incorporam o componente acústico – as gravações das vozes dos informantes – mais a transcrição ortográfica e fonética da fala. Dessa forma, é possível a recuperação das informações linguísticas, através do computador, de maneira multissensorial, integrada e interativa: a) multissensorial, pelo emprego coordenado de áudio (voz humana) e textos (transcrição ortográfica e fonética da fala); b) integrada, pela utilização simultânea dos meios de comunicação − voz e texto − sob a coordenação do computador; c) interativa, pela maneira com que se faz a recuperação das informações, isto é, ativamente, através de buscas, interligações, construção de informações novas.

Seguem, a título de exemplificação, transcrição ortográfica e fonética de recortes discursivos extraídos das Bases. Trata-se de extratos de informante de São Paulo, do sexo feminino, com curso superior completo, 25 a 29 anos, classe alta alta, registro informal de interação dialógica.[1]

Já viajei um bocadinho.

['ʒa via'ʒej ʒũ boka'dʲɲu ||]

Eu fui pela Associação dos Professores de Francês, sabe?

['ew 'fuj pel_asosja'sãw̃ dus pɾofe'soɾiz di fɾã'sej_'sabi ||]

Olha, o curso em si não... não sei se é o curso em si se vale a pena, entende?,

['ɔʎa u 'kuɾsw̃_ĩ 'si || 'nũ || 'nũ 'sej sj_'ɛ_w 'kuɾsw̃_ĩ 'si si 'valj_a 'pẹ̝n_ĩ'tẽʒ̃dʲi ||]

mas eu acho que, para criar uma maior maturidade, principalmente, no pessoal

[majz_'ew 'waʃu ki || pɾa kɾi'a_'wma maj'jɔɾ maturi'dadi põsipał'mẽʒ̃tʲi nu pe'swał]

que eu fui, eu achei uma... eu achei uma... um pessoal tão imaturo, um pessoal

[kj_'ew 'fuj 'jew wa'ʃej 'jʊ̟ma || ũ pe'swał 'tãw̃ ima'tuɾu || ũ pe'swał]

[1] Na transcrição fonética, representa-se por || a pausa efetivamente realizada na fala e por _ os casos de juntura lexical em que a fronteira vocabular é desrespeitada foneticamente, deixando de haver coincidência entre limite silábico e limite vocabular.

que chorava, porque estava vinte graus abaixo de zero, estava doendo o dedo, umas

[ki ʃo'ɾava puki 'tava 'vĩt̪i 'gɾawz_a'bajʃu ɖi 'zɛɾo || 'tava do'ẽj̃d_u 'ded_'ṵmas]

coisas assim; então, eu notei que o brasileiro, mesmo depois de uma faculdade,

['kojzaz_a'sĩ ʃĩ'tãw̃ 'ew no'tej kj_u bɾazi'leɾu 'mezmu ɖi'pojz ɖj̃_'ṵma fakuɬ'daɖi ||]

ele é imaturo; não se fala no pessoal... eu pensei: bom, só eu de Mackenzie...

['elj_'ɛ_jma'tuɾu || 'nũ si 'fala nu pe'swaɬ 'ew pẽj̃'sej 'bõ 'sɔ 'ew ɖi ma'kẽj̃zi]

— dizem que o pessoal de Mackenzie é filhinho de mamãe, de papai, né?; não é nada disso —

['dizĩ kj_u pe'swaɬ ɖi ma'kẽj̃zj_'ɛ || fi'l̪ɲu ɖi mą̇'mãj̃ ɖi pa'paj 'nɛ || 'nũ 'ɛ 'nada 'ɖisu ||]

pessoal formado por USP, etc., não sabia viver sozinho, entende?

[pi'swaɬ foɾ'madu puɾ_'uspj_eti'sɛteɾa || 'nũ sa'bija vi've sɔ'zi̪ɲw̃_ĩ'tẽj̃di ||]

Chegamos na França, aquele problema assim: a guerra ainda está ali presente, sabe?;

[ʃe'gą̇mu na 'fɾãs_a'keli pɾo'blę̇m_a'sĩ || a 'gɛx_a'ĩda 't_a'li pɾe'zẽj̃ti 'sabi ||]

então, você entra no metrô, reservam, ahn, lugar para mutilados de guerra, coisas assim;

[ĩ'tãw̃ 'se 'ẽj̃tɾa nṵ me'tɾo xe'zɛɾvũ ã || lu'gax pɾa muti'laduz ɖi 'gɛxa 'kojzaz_a'sĩ ||]

um pessoal super conscientizado, super amadurecido — pelo menos, o pessoal que eu conheci —.

[ũ pesu'aɬ 'supeɾ kõsiẽj̃ti'zadu || 'supeɾ_amaduɾe'sidu pelu 'mę̇n_u pe'swaɬ kj_'ew koŋe'si ||]

Então, para mim, como questão de amaturecimento, de viajar sozinha e conhecer países

[ĩ'tãw̃ pɾa 'mĩj̃ || 'kọ̇mu kes'tãw̃ dj_amatuɾesi'mẽtu ɖi vija'ʒa sɔ'zi̪ɲ_i koŋe'se pa'izis]

diferente, sozinha, foi excelente; agora, como curso mesmo, não dá para muito, né?;

[ḍife'ɾẽj̃ti sɔ'zi̪ŋa 'foj jese'lẽj̃t̪i || a'gɔɾa 'kọ̪mu 'kuxsu 'mezmu 'nũ 'da pɾa 'mũj̃tu 'nɛ ||]

o pessoal que eu conheci lá no curso também foi bom.

[u pe'swał kj_'ew koɲe'si 'la nu 'kuxsu t̪ą'mẽj̃ 'foj 'bõw̃ ||]

4 − Tratamento das informações

A possibilidade de extração, a partir das Bases, de diferentes *Corpora* por variáveis linguísticas e extralinguísticas torna viável a sua exploração por programas de análise linguística para estudos de aspectos diversos do português.

Há programas disponíveis que são indexadores e servem para a busca textual – permitem a indexação das palavras de um texto, ou seja, a identificação de sua localização no texto, a recuperação por listagens em forma de concordâncias (o conjunto de ocorrências de cada palavra, em ordem alfabética, com seu contexto imediato e sua localização). Possibilitam, também, a busca de colocados (de combinações de palavras - listas de palavras que ocorrem à esquerda e à direita da palavra de busca selecionada, em ordem de frequência) e de padrões de colocados (frases comuns - palavras que coocorrem com outras com certa frequência), bem como a pesquisa de grupos de palavras (com o uso de coringas e expressões lógicas, é possível a busca de palavras que guardam alguma relação). Os programas ainda permitem um tratamento quantitativo dos dados e alguns, quantiqualitativo.

Pesquisas linguísticas baseadas em *corpora* eletrônicos vêm tendo interesse crescente em diversas áreas dos estudos da linguagem. Daí o fortalecimento dos estudos na área da *Linguística de Corpus* e a intensificação dos trabalhos que envolvem pesquisas em grandes *corpora*, bem como do número de pesquisadores interessados nas investigações de dados linguísticos autênticos. Nesse sentido, disponibilizamos alguns estudos descritivos do português no *Sistema CorPor*, esperando oferecer uma contribuição para os estudos na área, em especial no que diz respeito à construção de léxicos e aos exames dos padrões da linguagem − e, pois, ao processamento de línguas naturais, área lacunar no Brasil.

CONCLUSÃO

Destacam-se os seguintes pontos: a) a investigação refere-se a aspecto pouco explorado nos estudos da língua portuguesa – construção de Sistemas de Banco de Dados Relacional com Bases de Informações, C*orpora* e Léxicos Eletrônicos do Português que contemplem transcrições ortográficas e fonéticas − se são raros, no Brasil, os *corpora* eletrônicos de transcrições de fala, mais ainda o são, se não inexistentes, os *corpora* com transcrições fonéticas e com recursos multimídia; b) a metodologia utilizada para a constituição do *corpus* de língua oral e do *corpus* de fala transcrito para tratamento computaconal e, pois, para a geração das *Bases de Informações Ortográfico-Fonéticas do Português Falado de São Paulo*, é interessante por possibilitar a extração de diferentes *Corpora* e Léxicos por variáveis linguísticas e

extralinguísticas, bem como a sua exploração por programas de análise linguística para estudos do português.

Com base nessas considerações, os resultados da investigação podem oferecer contribuições e benefícios: a) para responder à demanda, no Brasil, de *corpora* eletrônicos e de léxicos com transcrições de fala e com informações estatísticas de usos do português do Brasil, como fonte para diversos estudos; b) para a ampliação do intercâmbio científico e tecnológico e para o enriquecimento da interação entre as ciências exatas e as humanidades em geral, e, em especial, entre as ciências exatas e a ciência da linguagem; c) no âmbito da Linguística, pela disseminação do uso de pesquisas baseadas em *corpora* e de tecnologias informatizadas nos estudos da língua em uso, sobretudo nas áreas da Lexicologia, pelas possibilidades de aplicações imediatas na produção de dicionários e de glossários, e da Fonologia, pelo conhecimento dos padrões reais de uso do português falado; d) no ensino de línguas, pelas possibilidades de estudos da padronização linguística; e) a estudiosos do português, pelo oferecimento de Bases de Informações como fontes de usos reais, vivos e atestados, para uma descrição do emprego efetivo dos recursos da língua por variáveis linguísticas e extralinguísticas, com a possibilidade, ainda, de estudos comparativos entre esses usos e normas de emprego da gramática normativa; f) para a Fonoaudiologia, pelo estabelecimento de parâmetros da população sadia com vistas à confecção de instrumental para avaliações em áreas correlatas, com especial contribuição para a área de neuropsicolinguística; g) na interface entre a Linguística e a Informática, pelo oferecimento de conhecimentos linguísticos para a construção de sistemas de transcrição fonética automática e de sistemas computacionais de representação do conhecimento linguístico e, portanto, para o processamento da língua portuguesa, principalmente 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.

Para acesso público, as Bases de Informações, *Corpora* e Léxicos delas derivados e resultados de seus estudos estão publicados em meios eletrônicos, que carecem de textos transcritos − há textos escritos e não transcrições de fala −, bem como de recursos multimídia, através do *site* <www.corpor.fflch.usp.br>, para que o seu *download* possa ser feito para a máquina do pesquisador através de transferência de dados em redes de computadores.

O Sistema está disponível para a comunidade acadêmica, para, de um lado, com ela compartilhar parte dos muitos anos de utilização de tecnologias informatizadas nos estudos linguísticos; de outro, para que os usuários possam reportar dificuldades e problemas encontrados, e apresentar sugestões para a sua melhoria.

Para tornar o *Sistema CorPor* acessível aos interessados de maneira mais fácil, rápida, segura e amigável, tem-se a intenção de disponibilizá-lo em plataforma Web, seguindo as tendências atuais de produção, armazenamento e distribuição de conteúdos, o que significa converter o sistema atual em outro sistema com ferramentas web, de forma a viabilizar a sua utilização e pesquisa *on-line*, em tempo real.

Para concluir, retoma-se a referência feita ao trabalho de movimento duplo entre Linguagem e Tecnologias, ressaltando, de um lado, que as vantagens da utilização das Novas Tecnologias Digitais nas pesquisas linguísticas são indiscutíveis; de outro, vislumbrando resultados positivos de uma convergência

do *Sistema CorPor* com a área da Inteligência Computacional para a geração de uma Base de Conhecimentos da língua oral paulista, indispensável na arquitetura de um sistema de processamento de língua natural.

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 materiais para observações de aspectos diversos da língua.

AGRADECIMENTOS

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

BIBLIOGRAFIA

INTERNATIONAL PHONETIC ASSOCIATION (1999). *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press.

MCENERY, Tony; WILSON, Andrew (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

SCHANE, Sanford A. (1975). *Fonologia gerativa*. Rio de Janeiro: Zahar.

SCOTT, Mike (2004). *WordSmith Tools*. 4. vers. Oxford: Oxford University Press.

STUBBS, Michael (1996). *Text and Corpus Analysis - Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell.

SZYPERSKI, C. (1998). *Component Software: Beyond Object-Oriented Programming*. Boston: Addison-Wesley.

ZAPPAROLI CASTRO MELO, Zilda Maria (1980). *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.

ZAPPAROLI, Zilda Maria; CAMLONG, André (2002). *Do Léxico ao Discurso pela Informática*. São Paulo: EDUSP/FAPESP, 256 p. + CD-ROM.

Nota − Este artigo foi publicado em ZAPPAROLI, Zilda Maria. Voz e Texto Ortográfico-Fonético no Sistema Corpor – *Corpora* do Português Falado de São Paulo. In: 16º Colóquio da Lusofonia, 2011, Santa

Maria, Açores, Portugal. *Actas*… São Miguel, Açores: Colóquios da Lusofonia, 2011. p. 233-242, ISBN 9789899589186. Em formato digital – CD-ROM.