DESCRIÇÃO DO SISTEMA CORPOR

Pautada por uma interação real entre os estudos da linguagem humana e as tendências atuais de acesso à informação, a investigação que levou à geração do Sistema CorPor é por excelência interdisciplinar, inserida entre a Linguística e a Informática, na interface linguagem/tecnologias, portanto, em área lacunar dos estudos linguísticos da língua portuguesa, pois se refere a aspecto pouco explorado nos estudos da língua portuguesa - inexistência de Sistemas de Banco de Dados Relacional, de C*orpora* e de Léxicos Eletrônicos do Português que contemplem transcrições ortográficas e fonéticas (se são raros os *corpora de língua portuguesa*, mais ainda o são, se não inexistentes, os *corpora* com transcrições fonéticas).

O *Sistema CorPor* inclui os seguintes componentes: (a) Bases de Informações Ortográfico-Fonéticas do Português Falado de São Paulo; (b) *Corpora* Eletrônicos do Português Falado de São Paulo (Bases de Dados Textuais Ortográficas e Fonéticas, com emprego coordenado de áudio − voz humana − e textos; (c) Léxico de Frequência Ortográfico-Fonético do Português Falado de São Paulo; (d) Léxico de Freqüência Ortográfico-Fonético de Junturas Intervocabulares do Português Falado de São Paulo.

O Sistema ainda contém exposição da abordagem teórico-metodológica, dos procedimentos metodológicos (constituição do c*orpus* de língua oral, do c*orpus* de fala transcrito para tratamento computacional e do sistema gerenciador de banco de dados relacional), envolvendo parâmetros e critérios de constituição de *corpus* linguístico por procedimento informático; bibliografia; estudos e publicações.

A possibilidade de extração de diferentes Corpora Eletrônicos do Português Falado de *São Paulo* por variáveis linguísticas e extralinguísticas torna viável a sua exploração por programas de análise linguística para estudos de aspectos diversos do português. Investigações 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 e dada a restrita disponibilidade de Bancos de Dados, *Corpora* e Léxicos Eletrônicos de Transcrições de Fala em Língua Portuguesa do Brasil no momento em que a tendência internacional de pesquisa caminha no sentido de priorizar o emprego de uma abordagem baseada em *corpus*, disponibilizamos alguns estudos descritivos do português no Sistema, 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).

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 (este léxico despertou interesse especial em pesquisadores do *Signal Processing Lab* do Instituto de Telecomunicações, *Department of Electrical and Computer Engineering*, da Universidade de Coimbra): (a) os resultados podem ser úteis para a construção de sistemas de transcrição fonética automática; (b) os resultados podem oferecer contribuições para a construção de sistemas computacionais de representação do conhecimento linguístico e, pois, 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.

Um dos produtos tecnológicos de relevo é o Áudio com Texto − *Corpora* de Língua Oral com *Corpora* de Fala Transcrita Ortográfica e Foneticamente do Português Falado de São Paulo.

Ressaltamos o caráter inovador de se disponibilizar recuperação simultânea de voz e texto, ou seja, *corpora* do português falado de São Paulo que incorporem o componente acústico − as gravações das vozes dos informantes (*corpora* de língua oral) – mais a transcrição ortográfica e fonética da fala − Bases de Dados Textuais Ortográficos e Fonéticos −, para possibilitar a recuperação das informações linguísticas, através do computador, de maneira multissensorial, integrada e interativa: a) multissensorial, pelo emprego coordenado de som (voz humana) e textos; b) integrada, pela utilização simultânea dos meios de comunicação (som 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.

Destacamos que se trata de Bases de Informações, Corpora e Léxicos Eletrônicos do Português que contemplam transcrições ortográficas e fonéticas (se são raros os *corpora do português falado*, mais ainda o são, se não inexistentes, os *corpora* com transcrições fonéticas).

Voltada, pois, 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 e de léxicos 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.

Num primeiro momento, desenvolveu-se a versão para *desktop* do *Sistema CorPor* em plataforma compatível com *Windows XP* da *Microsoft*.

Para acesso aos pesquisadores, a versão *desktop* do *Sistema CorPor* está hospedada no *site* < http://tamisa.uspnet.usp.br/corpor>, para que o *Sistema* possa ser acessado via

FTP e, pois, para que o seu *download* possa ser feito para a máquina do pesquisador através de transferência de dados em redes de computadores. Trata-se de um protocolo genérico independente de *hardware* e de sistema operacional, que transfere arquivos por livre arbítrio, tendo em conta as suas propriedades e restrições de acesso. A transferência de dados em redes de computadores envolve normalmente transferência de arquivos e acesso a sistemas de arquivos remotos com a mesma interface usada nos arquivos locais.

Nessa versão, o Sistema Corpor é, pois, um sistema monousuário, de computador de mesa, sem disponibilidade na Web − o usuário faz o *download* do Sistema, instala-o em sua máquina, onde é feito o processamento.

Apesar de a disponibilidade do Sistema CorPor através de FTP representar avanços significativos em relação aos produtos das etapas anteriores, destacam-se algumas desvantagens, como a ocorrência de problemas na instalação no micro do usuário devido à incompatibilidade de sistemas operacionais, de interfaces e de requerimentos de memória.

Para tornar o Sistema CorPor acessível aos *interessados* de maneira mais fácil, rápida, segura e amigável, seguindo as tendências atuais de produção, armazenamento e distribuição de conteúdos, gerou-se o *Sistema CorPor* em ambiente da *World Wide Web* (WWW ou Web), por meio do emprego de tecnologias de computação de ponta, no contexto da plataforma .NET, o que significou converter o sistema *anterior* em outro sistema com ferramentas Web, de forma a viabilizar a sua utilização e pesquisa *on-line*, em tempo real.

Assim sendo, uma das grandes vantagens da versão Web consiste na possibilidade de acesso universal às informações do *Sistema CorPor* com maior rapidez, segurança e confiabilidade, por não depender de recursos do computador do usuário – o usuário opera o sistema *on-line*. O Sistema, nesse ambiente, poderá ser acessado através de qualquer dispositivo (microcomputador, *notebook*, *netbook*, celular) que tenha comunicação com a *Internet*, portanto, de forma democrática, com controle centralizado e independência geográfica.

A estrutura das Bases de Informações em Sistema Gerenciador de Banco de Dados Relacional (do inglês *Relational Database Management System* – RDBMS) permaneceu, mas com migração para plataforma Web, em função do que se fez necessária a seleção de outro conjunto de *software* (Banco de Dados e Linguagem de Programação) mais recomendado para esse ambiente. Alteraram-se, pois, os procedimentos metodológicos no que diz respeito à constituição do *Sistema CorPor* – armazenamento das informações com vistas à sua recuperação −, acrescentaram-se módulos e recursos de pesquisa ao Sistema. Para fins de estudos descritivos do português, aplicaram-se ferramentas do programa *Stablex*, desenvolvidas em função de propostas de análise de textos do método matemático-estatístico-computacional de André Camlong (Universidade de Toulouse II).

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. Destaca-se, ainda, que 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.

Para acesso público, as Bases de Informações, Corpora e Léxicos delas derivados, bem como resultados de seus estudos, estão publicados nos *sites* <http://www.corpor.ilexis.net.br> e <http://www.corpor.fflch.usp.br>, na plataforma .NET.

O Sistema está disponível para a comunidade acadêmica em geral, 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.