Revista Prolíngua – ISSN 1983-9979
Volume 2 - Número 2 - jul/dez de 2008

A RELEVÂNCIA DOS BANCOS DE DADOS PARA O ENSINO DA LÍNGUA PORTUGUESA

Darcilia Simões (AILP-UERJ-PUCSP-UFC-SELEPROT[1])
Eliana Meneses de Melo (UBC/SP-UERJ-SELEPROT)

RESUMO

Considerando o tema desta seção "A Expansão do Português no Mundo" e retomando uma conferência de Ataliba de Castilho intitulada "Reflexões sobre o Português Falado e o Exercício da Cidadania"[2], em que o pesquisador fez uma retrospectiva da construção e consolidação dos estudos linguísticos no Brasil, verificou-se a oportunidade de tratarmos então da Linguística de Corpus, destacando-lhe o objeto: os Bancos de Dados. No I Simpósio Mundial de Estudos de Língua Portuguesa (I SIMELP), realizado em 2008 sob a organização da Universidade de São Paulo (USP), tivemos a oportunidade de participar do Simpósio Ensino do Português e Novas Tecnologias[3] e conhecer experiências inovadoras na difusão sistemática da língua portuguesa. Recursos como o computador e o DVD com gravações de voz e imagem vêm trazendo contribuições espetaculares para o âmbito da pesquisa e do ensino da língua portuguesa. Nesse ambiente de revolução tecnológica, a informática se destaca com suas contribuições de natureza tática para o trabalho com os dados. E é dos aportes trazidos da linguagem digital que pretendemos falar nesta sessão.

PALAVRAS-CHAVE: Português no Mundo, Banco de Dados, Ensino da Língua Portuguesa

ABSTRACT

Considering the theme of this section "The Expansion of Portuguese in the World" and back to a conference by Ataliba Castilho entitled "Reflections on the spoken Portuguese and the Exercise of Citizenship", in which the researcher did a retrospective of the construction and consolidation of the linguistic studies in Brazil, there was an opportunity to deal with the Corpus Linguistics, highlighting the objective: the Database from the I World Symposium for the Study of Portuguese (I SIMELP), held in 2008 under the organization of the University of São Paulo (USP); we had the opportunity to participate in the Symposium of Portuguese education and new technologies and innovative experiences in the known systematic dissemination of the Portuguese language. Resources as the computer and the DVD recordings of voice and images are bringing spectacular contributions to the field of research and teaching of the Portuguese language. In this environment of technological revolution, the computers outstand with their contributions of tactic nature for the work with data. It is of this contribution brought from the digital language that we want to talk in this session.

KEYWORDS: Portuguese World, Database, Teaching of Portuguese Language

[1] Grupo de Pesquisa Nacional SEMIÓTICA, LEITURA E PRODUÇÃO DE TEXTOS.
[2] Proferida no VII Fórum de Estudos Linguísticos da UERJ em 2003 (*apud* HENRIQUES & SIMÕES, 2004).
[3] Coordenado pela Profa. Dra. Liliane Santos - Departamento de Português - Université Charles-de-Gaulle -- Lille 3 - UMR 8136 "Savoirs, Textes, Langage" (CNRS)

Revista Prolíngua – ISSN 1983-9979
Volume 2 - Número 2 - jul/dez de 2008

A fala de Castilho em 2003 procurou enfatizar a importância da constituição da língua falada como objeto de estudo. Para que tal ocorresse, uma ferramenta tecnológica ganhou destaque: o gravador magnetofônico. Veja-se o que disse Castilho:

> Todos os manuais de Linguística sustentam que a língua falada é a manifestação primordial das línguas naturais, sendo a língua escrita uma transposição mais ou menos feliz da primeira.
>
> Apesar dessas convicções, foi preciso aguardar a invenção do gravador portátil para que a língua falada passasse efetivamente a ocupar a atenção dos linguistas. E de fato, os estudos de língua falada percorreram dois momentos bem distintos, separados pela utilização do gravador magnetofônico. (...)
>
> Anteriormente à invenção do gravador portátil, esses estudos se fundamentavam em segmentos conversacionais recolhidos de memória e depois registrados no papel, ou na observação de como os escritores documentavam em seus textos literários a língua falada, muitas vezes erroneamente então denominada "fala popular". (Castilho in HENRIQUES & SIMÕES [orgs.] 2004, p. 17)

E é na esteira dos avanços tecnológicos que procuramos construir essa comunicação, de modo a dar mostras da relevância do progresso cibernético para o desenvolvimento da área dos estudos linguísticos.

1 O SURGIMENTO DA LINGUÍSTICA DE CORPUS

> Está em curso uma verdadeira revolução no pensamento linguístico (...). A mola propulsora dessa revolução é a tecnologia, mais especificamente o computador. Já foi dito que o computador pessoal, com memória poderosa e capacidade de armazenamento, começa a desempenhar, nas ciências humanas, o papel transformador que o telescópio teve na física e nas ciências exatas. Passamos da idealização para a sistematização da observação da evidência. (SARDINHA, 2004, Prefácio)

A epígrafe abre as portas de nossa conversa sobre mudanças significativas no panorama dos estudos linguísticos. A ciência cibernética descortinou mundos novos para a experiência humana. Algo que era típico das ciências exatas e que parecia impossível para as ciências humanas, mostra-se hoje plenamente incorporado no cotidiano da pesquisa das línguas e linguagens. A informática vem possibilitando a manipulação do conteúdo das línguas de forma mais objetiva, trazendo assim amplificação da margem de segurança das conclusões produzidas nos estudos e pesquisas apoiados pelo computador.

Revista Prolíngua – ISSN 1983-9979
Volume 2 - Número 2 - jul/dez de 2008

As dúvidas que emergiam do estudo de dados documentados pela experiência individual e subjetiva do linguista dá lugar, no limiar do terceiro milênio, à discussão subsidiada pela estatística produzida pelas linguagens cibernéticas. Essa mudança se deve à produção de bancos de dados digitais que constituiriam corpora de estudos e pesquisas avançados no âmbito das línguas naturais.

Segundo Tony Berber Sardinha, no ano de 1999 comemorou-se o aniversário de 35 anos da criação do primeiro córpus lingüístico eletrônico, o córpus Brown. Segundo o pesquisador da PUCSP, esse córpus foi lançado em 1964, o *Brown University Standard Corpus of Present-Day American English*, continha uma quantidade invejável de dados para a época: um milhão de palavras (cf. SARDINHA, 2000, p. 323-4).

Apesar da enorme dificuldade na informatização de textos na época – os textos foram transferidos para o computador por meio de cartões, perfurados um a um ─ o córpus Brown significou um desafio à total incredulidade e até hostilidade que então existia, quanto a se gastar tempo e recursos financeiros para a coleta de registros lingüísticos. A prevalência da visão chomskiana de que os dados estavam na mente dos linguistas constituía um óbice a uma possível linguística cibernética.

O aparecimento do Corpus Brown não só demonstrou a aplicabilidade e produtividade da aliança entre a pesquisa linguística e a informática, como também impulsionou a criação da Linguística de Córpus, área recente dos estudos linguísticos que vem crescendo prodigiosamente.

2 O QUE É A LINGUÍSTICA DE CÓRPUS?

> A conseqüência foi uma ampla e intensa valorização, como objeto primário de estudo, das variedades lingüísticas usadas por falantes iletrados, principalmente aqueles que pertencessem a comunidades isoladas, pouco ou supostamente nada sujeitas a influências de outras comunidades. (Mikhail Bakhtin, 1992 p. 301)

Essa afirmação do semioticista russo nos faz relembrarmos dados históricos relevantes para a grande área da Linguística, Letras e Artes. Cunha, em seu artigo “A importância da organização de uma base de dados para pesquisas atuais e futuras”, traz palavras de Nelson Rossi somadas às suas

Revista Prolíngua – ISSN 1983-9979
Volume 2 - Número 2 - jul/dez de 2008

> sob a influência evidente da ideologia romântica do século XIX, criou-se o conceito da "língua pura". Para encontrá-la "impunha-se recorrer às manifestações lingüísticas concretas do povo, principalmente na comunicação oral". (*apud* Cunha 2005,)

A busca pela fidelidade dos dados nas pesquisas linguísticas fez o homem se debruçar sobre a variação das línguas naturais. Com isso, o número de dados a serem considerados atingia proporções inalcançáveis para as pesquisas "manuais", como as que fizeram os precursores da ciência da linguagem. Em função da necessidade de abarcar cada vez maior número de dados a observar, o cientista vê na informática uma possibilidade de trabalho mais eficaz, não só em relação à quantidade de dados com que operaria, mas também com a possibilidade de cruzamento de variáveis, do que poderia nascer resultados mais confiáveis.

Rossi (*apud* Cunha, 2005) nos diz que

> A conseqüência foi uma ampla e intensa valorização, como objeto primário de estudo, das variedades lingüísticas usadas por falantes iletrados, principalmente aqueles que pertencessem a comunidades isoladas, pouco ou supostamente nada sujeitas a influências de outras comunidades. (Enciclopédia Mirador, Verbete **dialectologia**)

Um avanço visível é a Linguística de Corpus e suas ferramentas.

É Sardinha (2004, p.3) quem a define como a Linguística que se ocupa da coleta e da exploração de corpora, ou conjuntos de dados lingüísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística.

Ainda que o termo córpus (e corpora) fosse anterior ao computador, sua produção digital, com volume extraordinário só se torna possível após o advento da informática. No entanto, Randolf Quierk (1959) e sua equipe ao constituírem o *Survey English Usage* – um córpus não computadorizado – criaram uma referência para os córpus contemporâneos, inclusive o Brown.

Os estudiosos dessa nova especialidade já contam com número relevante de corpora eletrônicos de língua inglesa; e a língua portuguesa, que é nosso interesse imediato, já dispõe de corpora significativos, conforme demonstra Berber (2004, p. 9-10) cujo quadro aqui transcrevemos:

Revista Prolíngua – ISSN 1983-9979
Volume 2 - Número 2 - jul/dez de 2008

CORPUS	PALAVRAS	COMPOSIÇÃO	LOCALIZAÇÃO
Banco de Português	233 milhões	Português brasileiro, escrito e falado	PUCSP
Borba-Ramsey Corpus of Brazilian Portuguese**	1,67 milhão	Português brasileiro escrito	Brigham Young University
CETEM (Corpus de Extractos de Textos Electrónicos MCT) Público	229 milhões	Jornal português, "público"	Projeto Linguateca
COMET (Corpus Multilíngue para Ensino e Tradução)*	5 milhões	Parte referente ao português escrito comparável ao inglês	USP
CORDIAL (Corpus de Discurso para a Análise de Língua e Literatura)	Não disponível	Português escrito	UFMG
CORPUS UNESP/Araraquara/Usos do Português	200 milhões	Português brasileiro escrito	UNESP, Araraquara
CR-LW (Corpus de Referência Lácio-Web)*	5 milhões	Português escrito	USP, NILC
CRPC (Corpus de Referência do Português Contemporâneo)	152,6 milhões	Português dos vários países lusófonos, com predominância da variedade européia	CLUL (centro de Linguística da Universidade de Lisboa)
Historical Portuguese Prose**	2,8 milhões	Português escrito (1300 a 1900)	Brigham Young University
Modern Newspapers**	28 milhões	Português escrito, jornalístico e entrevistas publicadas em jornais	Brigham Young University
Modern Portuguese**	315 mil	Português literário (romances)	Brigham Young University
NILC*	35 milhões	Português brasileiro, escrito	NILC (USP, UFSCAR, UNESP Araraquara)
NUPILL (Núcleo de Pesquisas em Informática, Linguística e Letras)	Não disponível	Português escrito	UFSC
NURC (Projeto de Estudo da Norma Linguistica e Letras)	Não disponível (570 mil)**	Português brasileiro, falado	USP, UFRJ, UFBA, UFPE, UFRGS

Revista Prolíngua – ISSN 1983-9979
Volume 2 - Número 2 - jul/dez de 2008

PHPB (Projeto para a História do Português Brasileiro)*	Não disponível	Português escrito	UFPE, UFBA, UFMG, UFRJ, EFSC, UFPB, USP
PORTEXT	30 milhões	Português escrito de vários países	Universidade de Nice
Português falado do Ceará	Não disponível	Português brasileiro, falado	UFC, URCA
Tycho Brahe Parsed Corpus of Historical Portugueses	1,9 milhão	Português antigo (1550 a 1850)	Unicamp
VARPORT (Análise Contrastiva de Variantes do Português)*	Não disponível	Português escrito e falado, brasileiro e europeu	UFRJ, CLUL
VARSUL (Variação Linguistica Urbana da Região Sula)*	Não disponível	Português falado	UFSC, UFRGS, UFPR
*Pinheiro, Oliveira, Tagnin, Aluísio: http://www.nilc.icmc.usp.br/iiiencontro/programacao			
** Davies, Mark: http://davies.linguistic.byu.edu/personal/texts.asp			

A relevância da manipulação de dados da fala ordinária vem-se amplificando. A Universidade Federal do Ceará, por exemplo, vem desenvolvendo o projeto **Variação e Processamento da Fala e do Discurso: Análises e Aplicações – PROFALA**[4] que tem como objetivo geral a implantação de um sistema baseado em tecnologia da informação para análises e aplicações à língua falada e ao discurso. O projeto se desenvolve em parceria com o Programa de Pós-Graduação em Teleinformática, da mesma UFC e destina-se à realização de Análises Lingüísticas (fonético-fonológicas, léxicas, morfossintáticas, pragmáticas, discursivas); Dialetais (variações diatópicas do falar do Ceará e de outros estados nordestinos); Sociolingüísticas (variações diastráticas do falar do Ceará e de outros estados nordestinos) e Psicolingüísticas (processamento da fala e do discurso). Para essas análises, o projeto conta com um banco de dados com *corpora* já existente na UFC: O Português Não-Padrão do Ceará, o Português Oral Culto de Fortaleza, Projeto AliB-CE, *Corpus* de Língua Inglesa Falada.

[4] O *córpus* do projeto foi coletado em cidades da região do Cariri, especialmente Barbalha, Nova Olinda, Juazeiro, Várzea Alegre, Altaneira, Mauriti, Caririaçu e Brejo Santo.

Revista Prolíngua – ISSN 1983-9979
Volume 2 - Número 2 - jul/dez de 2008

Os resultados advindos desse tipo de pesquisa podem ser muito positivos não só para a descrição e conhecimento da língua falada atualmente no Ceará, mas poderão, principalmente, ser utilizados para estudos comparativos com outras fases da língua portuguesa, além da aplicação para o ensino da Língua Portuguesa, especialmente na alfabetização e no ensino fundamental, mas, também, no ensino médio e superior. Não se deve esquecer a significativa contribuição que pode ser dada para a tomada de decisões no âmbito das políticas lingüísticas no Brasil[5].

Buscando relatos sobre os avanços da Linguística para fortalecer o que ora falamos sobre a contribuição da Linguística de Corpus, encontramos o artigo "A Lingüística e os Estudos de Linguagem Rumo ao Século XXI" de Sônia Bastos Borba Costa[6], do qual retiramos um trecho em que a autora resume decisões que se refletem na diferenciação dos rumos dados à ciência linguística nos últimos tempos. Segundo Bastos (idem),

> os interessados no fenômeno das línguas começaram o século XX com a utopia do recorte objetivo, da documentação empírica, do isolamento do objeto para observações sistemáticas, etc.. O avançar do século evidenciou as suas limitações. Estamos adentrando o século XXI com a utopia da multi/interdisciplinaridade.

Essas idas e vindas decorrem de uma infinda busca de construção de uma ciência que possa abarcar seu objeto de uma forma o mais universalizante possível. No entanto, nesses momentos de extrema paixão científica, o estudioso se esquece de que a ciência, como produção humana, sempre será particularizante, independentemente de seu traçado e que as tendências contemporâneas à multi-, inter- ou transdisciplinaridade não trarão resultados menos específicos do que os obtidos nas molduras traçadas anteriormente.

Assim sendo, a contribuição dos bancos de dados e da Linguística de Corpus, evidentemente, trarão significativas novidades para as ciências da linguagem, sem, contudo, qualquer hipótese de exaustividade cogitada por alguns apaixonados.

[5] cf. http://www.profala.ufc.br/historico.htm
[6] *In* http://www.prohpor.ufba.br/alinguis.html

Revista Prolíngua – ISSN 1983-9979
Volume 2 - Número 2 - jul/dez de 2008

3 A IMPORTÂNCIA DOS BANCOS DE DADOS PARA OS ESTUDOS SÓCIO-VARIACIONISTAS.

> A "interface" entre a Linguística e a Informática pode ser analisada sob prismas diversos. Uma área das ciências da computação que tem dedicado uma atenção particular ao estudo das línguas naturais é, como se sabe, a Inteligência Artificial, cujas relações com a Linguística e a Psicologia Cognitiva são de tal forma estreitas que se pode falar de um metadomínio científico em que elas se aglutinam, a Ciência Cognitiva. (Matos[7], 1988)

Trouxemos essa epígrafe a esta seção do artigo, para documentar diálogos anteriores sobre as interseções entre a Linguística e outras ciências, em especial, as da computação. Estas, por sua vez, se avizinham das ciências do cérebro e da cognição, a partir do que seus laços com a metacognição se concretizam, e sua importância para as ciências da linguagem tomam vulto. Soma-se a isso interpenetração da informática em todas as áreas do conhecimento humano e que faz com que a Linguística se entrelace com a inteligência artificial na busca de resultados mais eficientes para suas investigações.

As interseções entre Lingüística e Psicanálise, as incursões pela Filosofia da linguagem e os estudos do cérebro e da inteligência artificial propulsionaram investigações que resultaram na criação de ferramentas digitais para processamento das línguas naturais nos computadores.

Sardinha (1999), em "**Usando WordSmith Tools na investigação da linguagem",** a adoção do computador como ferramenta de trabalho na análise da linguagem tem acontecido de modo tardio. Há mais ou menos 40 anos, o Córpus Brown acenou com uma nova estratégia de pesquisa linguistica e, acrescida a popularização dos computadores nas universidades, verifica-se um aumento significativo nas práticas de análise e armazenamento de dados em arquivos digitais. Mas segundo Sardinha, "mesmo assim, a parcela da pesquisa linguística assistida por computador ainda é minoritária", mantendo-se um modelo de pesquisa sobre pequenas quantidades de dados, coletados, mantidos, e analisados à mão, o que faz com que admitamos como ainda válido certo comentário de 20 anos atrás que assim circunscrevia a análise linguística à manipulação de dados que cabem no quadro-negro.

Felizmente esse cenário está sendo modificado nos anos dois mil. Hoje a Linguística já se mostra articulada com a inteligência artificial, e sua entrada no mundo digital veio a propiciar não só a manipulação de um número infinito de dados, como também levantamentos mais sofisticados de traços que distingam as variedades de uma língua natural.

[7] Third European Science Foundation summer school, (Pisa, Julho/Agosto de 1988), sobre Computational Linguistics and Lexicography.

Revista Prolíngua – ISSN 1983-9979
Volume 2 - Número 2 - jul/dez de 2008

Além dos pontos mencionados, tenhamos em mente as pesquisas em PLN (Processamento de Língua Natural) envolvendo componentes lexicais que expressem informações semânticas no desenvolvimento de sistemas que tenham como agente motivar o tratamento da especificidade dos sentidos das palavras em traduções simultâneas.

Zavaglia (2003) afirma que "A semântica é capaz de resolver muitos casos de homografia na linguagem falada e escrita". Sua pesquisa é direcionada para elaboração de uma Base de Conhecimento Lexical (BCL) cuja finalidade reside em, através de informações ontológicas, qualia, morfossintática, definicional e pragmática, construir bases de dados direcionadas aos pesquisadores sobre universos de discursos específicos.

Trazer para o cenário de nossa discussão pesquisas como as de Zavaglia, tem por finalidade ressaltarmos as dimensões de aplicação de conhecimento em Língua Natural envolvendo áreas que se afinam à ciência da computação. Ao mesmo tempo, somos levados a refletirmos sobre a formação do docente em Língua Portuguesa .De maneira geral, profissionais e pesquisadores lançam seus olhares para os processos hipermidiáticos da Internet, tendo como muito distante outras investigações e ferramentas.

Oportuno se torna lembrarmos que a ampliação do conhecimento em áreas da computação e pesquisas interdisciplinares, gerou novas ferramentas e aperfeiçoamento dos instrumentais de armazenamento de dados. Neste sentido, associando à nossa temática, trabalhos como de Regine Robin[8], no início da década de setenta, se fossem processados em bases atuais, os resultados ultrapassariam os dados numéricos, tão criticados por muitos pesquisadores.

Pela importância dos Bancos de Dados para o Ensino da Língua Portuguesa perpassam a relevância de estudos vindos etnolinguísticas, geolinguística e sociolinguística. São pesquisas que nos conduzem ao universo da diversidade cultural explícita na Língua Portuguesa.

Em rápidas pinceladas, destacamos variantes próprias da identidade social, os termos e meta-termos nos quais residem dados comportamentais. Seja no nível diatópico ou diastrático, ou ainda nos recortes sincrônicos e/ou diacrônicos, a riqueza dos diferentes falares

[8] Regine Robin (1977) e Michel Pêcheux (1990) em contato com o conceito de Formação Discursiva, proposto por Michel Foucault na Arqueologia do Saber, reconfiguram-no à luz do materialismo histórico e produzem uma mudança substancial em relação à concepção de discurso e de *córpus*. O discurso não pode mais ser visto fora das condições históricas de produção (...) e os *corpora* devem, então, ser analisados considerando que se inscrevem no interior de determinadas condições de produção, definidas em relação à história das formações sociais.

Revista Prolíngua – ISSN 1983-9979
Volume 2 - Número 2 - jul/dez de 2008

de nossa cultura lingüística devem ser compartilhados através dos recursos e todos as ferramentas que a tecnologia hoje nos oferece.

É justamente reconhecendo a importância de nossa diversidade cultural e lingüística e os múltiplos olhares investigativos que ressaltamos a importância do Banco de Dados com fonte de pesquisa e memória disponibilizada. Recurso imprescindível quando direcionamos nossas inquietações para a expansão da Língua Portuguesa e para a inclusão das diferentes comunidades.

Como sabemos, temos hoje mais de 240 milhões de falantes da Língua Portuguesa. Ela é a quinta língua no Planeta e, se considerarmos o mundo ocidental, é a terceira. Como oficialidade linguística, temos: "Angola, Brasil, Cabo Verde, Guiné Bissau, Guiné Equatorial, Macau, Moçambique, Portugal, São Tomè e Príncipe e Timor- Leste, sendo também falada nos antigos territórios da Índia Portuguesa (Goa, Damão, Diu e Dadrá e Nagar- Avali), além de ter também estatuto oficial na União Européia, no Mercosul, União Africana."[9]

Consideradas as preocupações hodiernas, em especial no Brasil, com um ensino de língua voltado para a inserção social, impõe-se falar de trabalho de pesquisa relacionado à variação linguística esboçado na Antiguidade Clássica. Veja-se o que diz Rossi (*apud* Cunha, 2005):

> O conhecimento empírico, às vezes parcial ou intuitivamente sistematizado, da "diversidade na unidade lingüística", documenta-se na cultura ocidental pelo menos desde a Grécia antiga, onde se reconhecia a existência de quatro dialetos, o eólico, o dórico, o jônico e o ático. (Enciclopédia Mirador, Verbete **dialectologia**)

A Relevância dos Bancos de Dados para o Ensino da Língua Portuguesa, no percurso trilhado por nossas reflexões, é de fundamental importância por ser instrumento através do qual se ampliam conhecimento, aceitação, memória social e inclusão das diferentes falas da Língua Portuguesa em suas dimensões planetárias.

Necessário se faz, também, que a formação do docente contemple estes aspectos, através de disciplinas que ampliem o conhecimento em tecnologia computacional, habilitando-os ao uso eficaz das ferramentas existentes, tanto para a pesquisa, como para o enriquecimento das práticas pedagógicas.

É indiscutível que um assunto como esse "daria panos pra manga", contudo, os limites de um artigo para comunicação em congresso impõem que passemos à conclusão.

[9] http://pt.wikipedia.org/wiki/L%C3%ADngua_portuguesa

Revista Prolíngua – ISSN 1983-9979
Volume 2 - Número 2 - jul/dez de 2008

4 ATANDO OS PONTOS DESSE TECIDO

A popularização dos computadores pessoais na década de 80 e do acesso à Internet na década de 90 promoveu a inserção da informática nas mais variadas áreas do saber. A partir de então, a informática passa a desempenhar um papel cada vez mais fundamental no mercado profissional e no meio acadêmico.

Desde a década de 60 os recursos tecnológicos vinham sendo empregados na pesquisa lingüística. Benefício maior da tecnologia recaiu sobre a Lingüística de Corpus, que vem se desenvolvendo e sendo aplicada a diferentes tópicos relacionados à linguagem. As ferramentas de processamento de dados textuais vêm prestando relevante serviço à linguística no que tange à análise lexical, sintática e discursiva, bem como para pesquisa e ensino de línguas estrangeiras, tradução, estudos culturais, descrição linguística e várias outras práticas, em uma dada língua ou comparativamente.

Dentre os ganhos da aliança entre Linguística em Informática destaca-se a constituição digital de córpus ─ coletânea de textos em formato eletrônico, compilada segundo critérios específicos, considerada representativa de uma língua (ou da parte que se pretende estudar), destinada à pesquisa, em especial, de dados empíricos (cf. Stella Tagnin, 2004). Ainda segundo Tagnin, o foco concentra-se no uso, medido pela freqüência de ocorrências, protegendo de alguma forma as conclusões nascidas da intuição do pesquisador.

Na perspectiva sociolinguística de tratamento igualitário dos dados da língua, independentemente da camada social que represente, a Linguística de Corpus traz uma contribuição substancial. Pois a busca num córpus demonstrará empregos, estruturações, que estão em uso; não fornecerá apenas a forma correta, mas principalmente as formas mais usuais na língua sob investigação.

Retomando a perspectiva de Castilho, ao fazer uma retrospectiva da construção e consolidação dos estudos linguísticos, cumpre então destacar a importância dos bancos de dados para os estudos contemporâneos, uma vez que, trazendo em abundância dados empíricos da língua em ação, permitirá aos estudiosos um mapeamento bastante próximo da realidade da distribuição da língua portuguesa no mundo.

Revista Prolíngua – ISSN 1983-9979
Volume 2 - Número 2 - jul/dez de 2008

À GUISA DE CONCLUSÃO

Para encerrar as considerações sobre a importância do ensino da língua portuguesa no mundo, vem ao texto um fragmento de autoria de José Manuel Matias[3110], extraído do sítio do Instituto Internacional da Língua Portuguesa[11] ─ IILP ─ que nos deixa uma mensagem de alta relevância no que tange à difusão da língua portuguesa.

> Poderá a língua portuguesa assumir um espaço relevante no contexto do mundo globalizado do século XXI?
>
> A expansão de determinada língua é condicionada por factores extralinguísticos, é consequência não da vontade dos seus falantes, não de políticas de língua isoladas, mas sim do discurso científico que produz, da expressão cultural e artística que cria, e acima de tudo das relações económicas que veicula.
>
> As línguas desempenham uma função crucial na génese das culturas e civilizações e o Português só desempenhará esse papel neste século, na medida, em que se impuser como língua de ciência, de expressão cultural e que seja um "meio de afirmação e uma poderosa vertente da economia de um país", como recentemente escreveu a linguista portuguesa Helena Mira Mateus num artigo no semanário "Expresso".
>
> Sendo inegável que muito se tem feito pela defesa da língua em Portugal e pela sua expansão, consolidação e divulgação no mundo, empreendimentos alguns de grande mérito, constatamos, porém, que muito ainda há por fazer. Investir no ensino do Português em países pertencentes a blocos políticos regionais emergentes, como a SADC e o MERCOSUL, com perspectivas de desenvolvimento económico futuro, será medida acertada. Investir no ensino do Português em países que necessitam impreterivelmente do Português para estruturação do seu próprio Estado e da sua coesão nacional, sob pena de se transformarem em entidades ingovernáveis, será outra medida acertada. A formação de professores de Português nestes países deveria ser considerada prioritária. A construção de materiais didácticos informatizados e de dicionários electrónicos seria outra medida fundamental.

[10] Vice-presidente da Sociedade da Língua Portuguesa e co-coordenador editorial do Ciberdúvidas
[11] http://www.iilp-cplp.cv/index.php?option=com_content&task =blogcategory&id=16&Itemid=69

Revista Prolíngua – ISSN 1983-9979
Volume 2 - Número 2 - jul/dez de 2008

REFERÊNCIA

BAKHTIN, Mikhail. **Estética da criação verbal**. Tradução de Maria Emarentina G. Gomes Pereira. São Paulo: Martins Fontes, 1992.

CASTILHO, Ataliba de. "Reflexões sobre o Português Falado e o Exercício da Cidadania". In HENRIQUES, C. C. & SIMÕES, Darcilia (orgs.). Língua **e Cidadania: novas perspectivas para o ensino**. Rio de Janeiro: Europa, 2004. [p. 15 -33]

PÊCHEUX, M. **L'inquietude du discours.** Textes choisis par D.Maldidier. Paris: Cendres. 1990.

ROBIN, Regine. **História e Lingüística**. São Paulo: Cultrix, 1977.

SARDINHA, Tony Berber. "Lingüística de córpus: histórico e problemática". In

Documentação de Estudos em Lingüística Teórica e Aplicada – D.E.L.T.A., Vol. 16, N.º 2, 2000 (323-367)

SARDINHA, Tony Berber. **Linguística de Corpus**. São Paulo: Manole, 2004.

ZAVAGLIA, C. "Bases de Conhecimento Léxico- Ontológico para o Português do Brasil: uma proposta de modelo. São Carlos: **Anais do 1º Workshop em Tecnologia da Informação e da Linguagem Humana**, NILC, 2003.

Fontes digitais

COSTA, Sônia Bastos Borba. "A linguística e os estudos de linguagem rumo ao Século XXI". 2003. Disponível em http://www.prohpor.ufba.br/alinguis.html

CUNHA, Cláudia de Souza. "A importância da organização de uma base de dados para pesquisas atuais e futuras". 2005.

Disponível em http://www.gel.org.br/estudoslinguisticos/edicoesanteriores/4publica-estudos-2006/sistema06/784.pdf

MATOS, Sergio. "Linguistica e informática. Perspectivas recentes do uso do computador em Linguistica aplicada e descritiva". 1988. Disponível em <http://ler.letras.up.pt/uploads/ficheiros/2594.pdf>

TAGNIN, Stella. "Corpora: o que são e para que servem". 2004. Disponível em: http://www.fflch.usp.brldlmlcometl