INFORMATIZAÇÃO DE ACERVOS DA LÍNGUA PORTUGUESA[1]

Ataliba Teixeira de Castilho
USP
Giselle Machline de Oliveira e Silva
UFRRJ
Dante Lucchesi
UFBA

Preliminares

Quando Presidente da ABRALIN, o primeiro dos autores confiou a Miriam Lemle a organização de uma Mesa-Redonda voltada para a "questão do estabelecimento de um código de ética para a profissão de lingüista", no contexto da 36a. Reunião Anual da SBPC (São Paulo, 1984).

O enfoque dado à questão foi o da utilização de dados em análises sociolingüísticas, assim desdobrada por Lemle 1984:5-10: *(i) quais os cuidados aos quais estamos obrigados no que diz respeito à publicação de dados colhidos em entrevista concedida para a pesquisa lingüística? (ii) até que ponto é válido aquilo que se faz rotineiramente nas pesquisas psicolingüísticas, nas quais os sujeitos nunca são corretamente informados sobre qual o aspecto do seu desempenho, que o pesquisador quer conhecer? (iii) como formular um princípio justo de reciprocidade entre pesquisador e comunidade pesquisada? (iv) que fazer quando um modelo de produção do saber que entrou em voga é por si só alienante?* Lemle tece algumas considerações sobre esses quesitos, e passa a palavra aos demais componentes da mesa, Sebastião Votre, Claiz Passos e Fernando Tarallo: Votre 1984, Passos 1984, Tarallo 1984.

Votre se ocupou mais diretamente de uma política de banco de dados, principiando por elencar os argumentos que, naquele momento, evidenciavam sérias reticências – para dizer o menos – quanto ao efetivo compartilhamento de acervos duramente constituídos, a respeito dos quais seus responsáveis gostariam de garantir certa prioridade no tratamento de temas neles empiricamente fundados. Ele sugeriu à ABRALIN uma série de iniciativas, tais como a identificação desses acervos (de que enumera quinze, de seu conhecimento) e a listagem dos temas em andamento, inspirados nesses acervos. Passos defende a necessidade de estabelecimento de normas de conduta, e menciona algumas atividades do grupo de linguistas da UFBa, naquela ocasião. Finalmente, Tarallo discute num texto bem humorado a natureza do dado lingüístico, jogando com a equação "dado é dado". Concentrando-se no segmento predicativo dessa expressão, ele argumenta que por *dado* tanto se pode entender o elemento empírico que serve de base à resolução de problemas, quanto sua qualidade de elemento doado, gratuito, o que permite reconhecer que *"na composição de um corpus para análise lingüística, coletam-se determinados elementos / (dados), tratáveis / (dados) e que foram presentados, concedidos (dados)"*. Isso fundamentaria a expectativa do pesquisador de *que "seus dados sejam dados dados, i.e., oferecidos, concedidos, tratáveis e específicos"* (p.30). Ele menciona a seguir a guarda dos dados lingüísticos em bancos, de que enumera vários, para finalmente formular a questão nevrálgica naquela já distante mesa-redonda: *"será justo e lícito supor que não-depositários destes bancos tenham acesso a empréstimos, ou será mais viável instituir um esquema de caderneta de poupança?"*. E logo a seguir, numa atitude lúcida, como sempre foram as suas, Tarallo declara: *"em hipótese alguma, no entanto, o banco e suas atividades devem adormecer em decorrência de uma atitude radical de não-empréstimo a não-depositários" (p.31)*. É necessário, entretanto, continua ele, impedir que *"imagens de pirataria sequer o infiltrem"*, e esse risco pode ser conjurado se se adotar um *"esquema de poupança"*, em que *"cada 5 ou 10 horas dadas ao banco por um dos pesquisadores, estarão sendo acumuladas a outras 5 ou 10 depositadas por outro sócio-pesquisador: todos dados de igual qualidade, portanto, determinados e tratáveis."* Finalizando sua intervenção, Tarallo retoma a tautologia "dado é dado", com que denominou seu texto, e traça em poucas palavras uma diretriz que

poderia ser agora observada, no momento em que retomamos a idéia da criação de um Banco de Dados da Língua Portuguesa: *"mas que nem por isso mesmo sócios-não-depositários se julguem no direto de abusar da força da tautologia 'dado é dado' e dar a ela entoações várias para o lucro próprio. Afinal de contas, nós, como sócios--depositários, devemos fazer valerem as acepções adjetivais de dado: nós também temos o direito e o dever, no cumprimento de nossas regras, de 'permitir', 'oferecer', 'conceder', 'facultar' e 'determinar'"* (p.33).

Para retomar essa discussão, abordaremos cinco questões neste texto: (1) O Projeto Minerva, (2) O Seminário sobre a Informatização de Acervos da Língua Portuguesa (Campinas, 1993), (3) A Oficina de Trabalho sobre Programas de Análise e Tratamento de Dados (São Paulo, 1994), (4) Apresentação dos resultados do levantamento de acervos (Vitória, 1994), e (5) Sugestões para a constituição de um Banco de Dados da Língua Portuguesa.

1. Projeto Minerva

Em 1991, o Prof. Juan Uriagereka, da Universidade de Maryland, propôs a constituição de um Banco de Dados do Português e do Galego.

A primeira versão dessa proposta foi enviada a diversos pesquisadores, entre 1991 e 1992. Manifestaram-se favoravelmente a ela os seguintes pesquisadores: Jairo Morais Nunes, Anton Santamarina, Arthur L. Askins, Marisa Rivero, Ivo de Castro e Ataliba T. de Castilho.

A segunda versão, já agora denominada *Projeto Minerva*, menciona as reuniões havidas em Santiago de Compostela, College Park (Maryland) e Lisboa, para a discussão do documento inicial, além das reações de vários especialistas. Esse texto trata dos seguintes tópicos: a necessidade de um projeto de banco de dados do Português e do Galego, o escopo do projeto e os passos iniciais para sua implantação, tais como a edição de um boletim informativo, prioridade inicial para a informatização de textos contemporâneos, indicação do Centro de Lingüística da Universidade de Lisboa para coordenar as atividades, e nomeação de um Diretor do Projeto, para o que se sugeria o nome da Profa. Rosa Virgínia Mattos e Silva: Uriagereka 1993.

Uma série de documentos, alguns dos quais com sugestões de caráter técnico, foi apensada às duas versões.

2. Seminário sobre a Informatização de Acervos da Língua Portuguesa (Campinas, 1993)

Em reunião havida em dezembro de 1992, em Salvador, Rosa Virgínia de Mattos e Silva e Ataliba T. de Castilho concordaram em que seria de todo interesse que os brasileiros se articulassem, para oferecer uma resposta afirmativa à proposta do Prof. Uriagereka. Ficou acertado que ambos convocariam um Seminário para o exame desse e de outros tópicos, a realizar-se na Universidade Estadual de Campias, no ano subseqüente.

Foi assim que, sob os auspícios da Fundação de Amparo à Pesquisa do Estado de São Paulo, reuniram-se em Campinas, de 4 a 5 de outubro de 1993, diversos responsáveis pro acervos da Língua Portuguesa. A reunião, denominada "Seminário sobre a Informatização de Acervos da Língua Portuguesa", tinha por objetivos: (1) obter informações sobre acervos disponíveis no Brasil, (2) propor um acordo de compartilhamento desses acervos, e (3) tomar as decisões técnicas necessárias para a constituição de um banco de dados, tendo em vista a compatibilização prévia dos programas. O Prof. Uriagereka, especialmente convocado para o Seminário, escusou-se, em carta de 10 de junho do mesmo ano, enviada a Castilho, na qual afirma: "*Eu xa tiréi a primeira pedra, e agora pareceme que ainda que a miña ciscunstancia mo permitira, non podo e non debo entrometerme en cousas que, cando menos, son delicadas. O máis que podo facer e estar aqui*".

O tempo se revelou escasso para um encaminhamento dos temas agendados em Campinas. Durante o Seminário, foram lidos e debatidos os seguintes relatórios, previamente encomendados:

1. Rosa Virgínia Mattos e Silva e Dante Lucchesi Ramaciotti – "O Banco de Dados do Programa para a História da Língua Portuguesa – PROHPOR".

2. Dino Preti e Zilda Maria Zapparoli Castro Melo – "Corpus Informatizado do Projeto NURC/SP".

3. Dinah I. Callou – "Projeto NURC/RJ: situação atual do corpus".

4. Jacyra Andrade Mota – "Relatório sobre o Projeto NURC em Salvador".

5. Suzana Alice Marcelino Cardoso – "Relatório sobre Acervo de Língua Portuguesa de Base Dialetal".

6. Alzira Tavares de Macedo – "Constituição do Corpus do PEUL".

7. Leila Bárbara – "Banco de Texto da PUC-SP".

8. Francisco da Silva Borba – "O corpus do Dicionário de Usos do Português (DUP)".

9. Angela Cecília de Souza Rodrigues – "A linguagem popular de São Paulo".

10. Perpétua Gonçalves *et alii* – "Panorama do Português Urbano de Maputo [Moçambique]".

11. Dante Lucchesi Ramaciotti e Alan Baxter – "Falas da Comunidade Afro-Brasileira".

12. Maria Antonieta Cohen – "Corpus de Textos Escritos para Pesquisa em Lingüística Histórica".

13. Maria del Rosário Suárez de Albán – "Dados para pesquisa em Literatura Popular".

Num segundo momento, os pesquisadores intercambiaram pontos de vista sobre uma política de compartilhamento de materiais e a constituição de um banco de dados. Os argumentos expedidos foram os seguintes:

a) Reconhecem-se como válidos os esforços pela institucionalização de uma política acadêmica voltada para a informatização de *corpora* disponíveis sobre a Língua Portuguesa, e sua abertura à consulta, sob certas condições.

b) Muitos dos levantamentos de dados relatados neste Seminário foram feitos com dinheiro público, e por isso deveriam ser abertos aos pesquisadores, discriminando-se usuários contribuintes de usuários não-contribuintes. Estes últimos deveriam ter acesso aos dados com algum tipo de ônus: pagamento em dinheiro, prestação de serviços, ou intercâmbio com material de interesse.

c) O acesso aos dados deve fazer-se acompanhar de uma troca de informações sobre as pesquisa elaboradas e a elaborar a partir deles, para que se estabeleça um relacionamento frutífero entre pesquisadores que estejam operando em uma mesma área de estudos.

d) A ABRALIN deveria ser envolvida na identificação de novos acervos, e na tomada de decisões sobre o compartilhamento dos dados e os aspectos técnicos envolvidos. A Profa. Suzana A. M. Cardoso, Presidente dessa Associação e participante do Seminário, dispôs-se a adotar as medidas necessárias para esse efeito.

e) A execução de um projeto de constituição de um banco de dados pode ser feita de maneira centralizada, encarregando-se de uma ou mais de uma universidade para seu gerenciamento, ou de maneira descentralizada, encarregando-se cada grupo participante de informatizar seus materiais, a partir de parâmetros previamente acordados.

Na sessão de encerramento, foram aprovadas as seguintes recomendações: (1) Que as Profas. Leila Bárbara e Zilda M.Z.Castro Melo organizem em São Paulo um seminário para a demonstração de equipamentos e de *softwares*, tendo em vista a constituição de um banco de dados. (2) O Centro de Documentação Lingüística e Literária Alexandre Eulálio, do Instituto de Estudos da Linguagem da UNICAMP, preparará um modelo de ficha, para o recenseamento dos acervos. Após aprovada, a ficha será encaminhada pela ABRALIN aos pesquisadores. (3) Uma Comissão integrada por Ataliba T.de Castilho, Giselle Machline de Oliveira e Silva e Dante Lucchesi Ramaciotti examinará os materiais assim coletados, apresentando à reunião da ABRALIN em Vitória, julho de 1994, um texto com sugestões para a implantação de um *Banco de Dados da Língua Portuguesa*.

3. Oficina de Trabalho sobre Programas de Análise e Tratamento de Textos (São Paulo, 1994)

No dia 25 de março de 1994, realizou-se no Centro de Informática da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, por iniciativa de Leila Bárbara e Zilda Maria Zapparoli Castro Melo a Oficina de Trabalho sobre Programas de Análise e Tratamento de Textos.

Foram apresentados e demonstrados os seguintes *softwares*:

1. Leland Emerson McCleary – *Notebuilder e Wordcruncher.*
2. Heloísa Collins – *Instrumentos para uma análise lexical em uma perspectiva do discurso: o MicroConcord e o Wordlist Suite*

3. Francisco da Silva Borba – *Folio Views.*
4. Ruth E.Lopes Moino – *Varbrul.*
5. Maurício Pereira Nunes e Zilda Maria Zapparoli Castro Melo – *Palestra-demonstração sobre o Stablex.*
6. Ivone Isidoro Pinto – *O Tact.*
7. Roxane R.H.Rojo – *The Ethnograph.*

As conclusões da Oficina de Trabalho apontaram para a constituição de uma Comissão Técnica, integrada por especialistas em informática e por lingüistas, para a tomada de decisão na órbita do Banco de Dados da Língua Portuguesa.

4. Resultados do levantamento de acervos: a reunião de Vitória ES

O Seminário de Campinas tinha recomendado a identificação de acervos e o contacto com seus titulares, para se avaliar o interesse em uma eventual integração num Banco de Dados da Língua Portuguesa.

A 2 de maio de 1994, a ABRALIN enviou ao corpo associativo uma ficha preparada por Rodolfo Ilari e aprovada pelo CEDAE, solicitando que os resultados fossem encaminhados diretamente ao primeiro autor deste relatório.

Até 12 de julho do mesmo ano, tinham sido recebidas 25 respostas, totalizando 51 acervos identificados. Alguns acervos mencionados nos trabalhos de Votre 1984 e Tarallo 1984 ainda não responderam, sendo necessário contactar os respectivos titulares para esse fim.

Uma primeira análise das respostas foi apresentada durante as atividades da ABRALIN, desenvolvidas no contexto da 45a. Reunião Anual da SBPC (Vitória, 17 a 22 de julho de 1994). Além do presente texto, Dante Lucchesi leu seu trabalho "Perspectivas para a Lingüística do *corpus*".

Uma rápida análise das respostas obtidas revelou que predominaram os acervos institucionais, 84% do total, seguindo-se 16% de acervos em mãos dos particulares que os constituíram. Considerando-se apenas os acervos institucionais, constata-se que a universidades federais cobrem 67% de seu número, seguindo-se 19% de universidades estaduais e 14% de universidades privadas.

Quanto à modalidade de Língua Porguesa recolhida nesses acervos, constata-se uma forte predominância de dados da língua falada, 74,5%, seguindo-se 20% de língua escrita, e 5,5% de acervos mistos.

No caso da língua falada, predomina a documentação da fala de crianças, totalizando 44%. Segue-se a fala dos adultos analfabetos ou de baixa escolaridade, 29%, a dos jovens que cursaram o primeiro e/ou o segundo graus, 15%, e a dos adultos de formação universitária 12% – embora se concentrem aqui os acervos mais extensos. Não foi possível quantificar esta variedade, dada a incompletude dos dados.

No caso da língua escrita, os acervos provilegiaram a língua moderna e contemporânea, 46%, vindo logo depois a arcaica e a clássica com 27% cada.

É surpreendente o número de acervos informatizados, entendendo-se por isto aqueles que foram digitados e são disponíveis em disquetes: 54%. Dos restantes 46%, muitos estão em processo de digitação.

Outro ponto importante nessas respostas é a quase categórica decisão de compartilhar os dados, invertendo-se a posição dos participantes da reunião de 1984. A opção de compartilhar dados foi manifestada por 90% dos responsáveis pelos acervos. Quanto às condições para isso, 49,3% aceitariam fazê-lo à base de reciprocidade, 46,5% à base de cobertura de custos, e 4,2% optaram por outras formas, nem sempre especificadas. Estas proporções não são muito confiáveis, pois a grande maioria dos que responderam optaram ao mesmo tempo pela reciprocidade e pela cobertura dos custos.

A segunda autora organizou em agosto de 1994 um Banco de Dados com as respostas obtidas, ordenando-as segundo os seguintes parâmetros: (1) título do acervo, região e período coberto pelos dados; (2) nome do responsável, endereço e cidade; (3) Estado, agência de fomento, idade e escolaridade dos informantes, suporte do documento; (4) tipo de entrevista, gênero, horas-fita, tipo de transcrição; e (5) linguagem de computador adotada, outras informações.

Desta análise, resultou a necessidade de complementar as informações, de acordo com uma lista de quesitos transcritos no item 5 deste documento, para o qual chamamos a atenção dos detentores de acervo que responderam ao primeiro questionário.

Para uma divulgação informal das respostas, segue-se uma relação dos acervos identificados, arranjados por períodos históricos, indicando-se o responsável, o título do acervo, e a instituição a que está ligado; a omissão deste dado implica em que se trata de acervo privado.

Obtidas as respostas ao questionário suplementar, novas decisões deverão ser tomadas pela ABRALIN, em sua próxima reunião (São Luís, julho de 1995), ou em outra reunião a combinar.

1. Português Arcaico e Clássico

1.1 – Dante Lucchesi e Rosa Virgínia Mattos e Silva – *Banco de Textos para a História da Língua Portuguesa,* UFBa.

1.2 – Célia Maria Moraes de Castilho – *Português Arcaico: séculos XIII-XVI.*

1.3 – Maria Antonieta A. de Mendonça Cohen – *Banco de Textos para a Lingüística Histórica,* UFMG.

1.4 – Rosane de Andrade Berlinck – *O Português dos séculos XVIII a XX: correspondências, peças de teatro, relatos de viagem e autos de inquéritos.*

1.5 – Vanda de Oliveira Bittencourt – *Textos do Português do Brasil do século XVI a XX,* UFMG.

2. Português Contemporâneo: Modalidade Falada.

2.1 – Aquisição da Linguagem

2.1.1 – Leonor Scliar-Cabral *et alii – Arquivo brazil.tar – lista Childes-noneng,* UFSC.

2.1.2 – Maria Denilda Moura *et alii – Língua Utilizada em Alagoas,* UFAL.

2.1.3 – Regina Ritter Lamprecht – *Banco de Dados: a Linguagem da Criança com Desvios Fonológicos Evolutivos,* PUC-RS.

2.1.4 – Ana Maria de Mattos Guimarães *et alii – Linguagem da Criança na Fase de Letramento,* PUC-RS.

2.1.5 – Vânia Maria B. Arruda Fernandes *et alii – Variantes Lingüísticas empregadas pela Escola e pelos Alunos,* UFUberlândia.

2.1.6 – Regina Maria Freire – *Retardo de Linguagem,* PUC-SP.

2.1.7 – Idem – *Corpora de Crianças de 1.6 a 7 anos,* PUC-SP.

2.1.8 – Idem – *Desenvolvimento da Linguagem,* PUC-SP.

2.1.9 – Idem – *Discurso da Saúde,* PUC-SP.

2.1.10 – Idem – *Mãe-Bebê,* PUC-SP.

2.1.11 – Roxane H. R. Rojo – *Banco de Dados sobre Letramento Emergente e Aquisição de Narrativas.*

2.2 – Linguagem Culta

2.2.1 – Dino Preti e Ataliba T.de Castilho – *Projeto NURC-SP*, USP e CEDAE-UNICAMP.

2.2.2 – Zilda M. Zapparoli Castro Melo – *Corpus Informatizado do Projeto NURC-SP*, USP.

2.2.3 – Idem – *Corpus Informatizado do Português Falado do Brasil, Variante Paulista*, USP.

2.2.4 – Carlota Silveira *et alii* – *Projeto NURC-SSA*, UFBa.

2.2.5 – Maria da Piedade Moreria de Sá *et alii* – *Projeto NURC-Recife*, UFPe.

2.2.6 – José Lemos Monteiro – *Projeto de Descrição do Português Oral de Fortaleza*, Universidade Estadual do Ceará.

2.2.7 – Jorge de Vasconcelos *et alii* – *Programa "Certas Palavras"*, CEDAE-UNICAMP.

2.2.8 – Rosane de A. Berlinck – *A Fala dos Universitários de Curitiba.*

2.3 – Linguagem não-padrão

2.3.1 – Suzana A.M. Cardoso – *Acervo de Língua Portuguesa de Base Dialetal: o Atlas Prévio dos Falares Baianos e o Atlas Lingüístico de Sergipe*, UFBa.

2.3.2 – Maria do Socorro Silva de Aragão – *Atlas Lingüístico da Paraíba*, UFPB.

2.3.3 – Ângela Cecília de Souza Rodrigues – *Português Popular de São Paulo.*

2.3.4 – Anthony J.Naro *et alii* – *Amostra Censo da Variação Lingüística no Rio de Janeiro*, UFRJ.

2.3.5 – Vanderci de A.Aguilera – *Atlas Lingüístico do Paraná*, Universidade Estadual de Londrina (UEL).

2.3.6 – Idem – *A Arcaicidade na Fala Popular de Ortigueira PR*, UEL.

2.3.7 – Idem – *Aspectos Lingüísticos da Fala Londrinense*, UEL.

2.3.8 – Idem – *O Léxico da Costura em Londrina*, UEL.

2.3.9 – Idem – *Aspectos Lingüísticos da Fala Popular de Porecatu*, UEL.

2.3.10 – Oswaldo A.Furlan e Hilda Gomes Vieira – *Atlas Lingüístico-Etnográfico da Região Sul do Brasil, Secção SC*, UFSC.

2.3.11 – Luiza Galvão Lessa e Lindinalva M.Chaves – *Atlas Lingüístico do Estado do Acre*, UFAc.

2.3.12 – Carlos Vogt e Maurizio Gnerre – *Cafundó*, CEDAE-UNICAMP.

2.3.13 – Eni P.Orlandi – *Estudos e Linguagem Rural*, CEDAE-UNICAMP.

2.3.14 – Maria del Rosário Suarez de Albán – *Programa de Estudo e Pesquisa da Literatura Popular*, UFBa.

2.3.15 – Regina Célia F.C.Trindade – *O Português Falado pelas Comunidades Remanescentes de Quilombos do Nordeste Paraense*, UFP.

2.3.16 – Sílvia F.Brandão, Maria Emília B.da Silva e Edila Vianna da Silva – *Atlas Etnolingüístico dos Pescadores do Estado do Rio de Janeiro*, UFRJ.

2.3.17 – Mary Francisca do Careno – *Projeto Vale do Ribeiro*, UNESP-Assis.

2.4 – Linguagem Culta e Linguagem não-Padrão

2.4.1 – Leda Bisol, Paulino Vandresen, Iara Bemquerer Costa *et alii* – *Banco de Dados VARSUL: Paraná, Santa Catarina e Rio Grande do Sul*, UFPR, UFSC, UFRGS.

2.4.2 – Dermeval da Hora Oliveira *et alii* – *Projeto Variação Lingüística no Estado da Paraíba*, UFPB.

2.4.3 – Maria Irene Francisco Canovas – *Variação Fônica em Falantes de Salvador.*

2.5 – Patologias da Linguagem

2.5.1 – Regina Maria Freire – *Discurso Fonoaudiólogo – Sujeito Afásico*, PUC-SP.

3. Português Contemporâneo Escrito

3.1 – Maria Bernadete M.Abaurre, Maria Laura T.M.Sabinson e Raquel S.Fiad – *Aquisição da Representação Escrita da Linguagem*, UNICAMP.

3.2 – Eni P.Orlandi – *Análisedo Discurso Indígena*, CEDAE-UNICAMP.

3.3 – Vanderci de A. Aguilera – *Banco de Dados para Confronto Oralidade-Escrita*, UEL.

3.4 – Maria Angélica F.da Cunha – *Discurso e Gramática*, UFRN.

3.5 – Ataliba T. de Castilho – *Corpus Diacrônico do Português da Cidade de São Paulo, USP.*

4. Português Contemporâneo Falado e Escrito

4.1 – Cláudia Lemos – *Projeto de Aquisição da Linguagem Oral e Escrita*, CEDAE-UNICAMP.

4.2 – Sebastião Votre – *Corpus Discurso e Gramática*, UFRJ.

4.3 – Virgínia Colares S.F.Alves – *Interação Verbal na Justiça. Tomada de Depoimentos*, Universidade Católica de Pernambuco.

4.4 – Alice Maria Teixeira de Saboia e Deusa Fonseca Raposo de Medeiros – *Lexicologia e Lexicografia*, UFMT.

4.5 – Nadja da Costa R.Moreira – *Base de Dados da Escrita Infantil*, UFC.

4.6 – Maria Vicentina de Paula do A. Dick – *Atlas Toponímico do Estado de São Paulo*, USP.

4.7 – Alice M.T.de Saboia – *Vocalização do Português Oficial*, UFMT.

4.8 – Elisabeth Silveira e Sebastião Votre – *NUPLEn- D e G*, UFRJ.

4.9 – Maria Antonieta Alba Celani *et alii* – *Direct*, PUC-SP.

4.10 – Leila Bárbara – *Português Acadêmico*, PUC-SP.

Os debates havidos em Vitória concluíram pela indispensabilidade da constituição de um Banco de Dados da Língua Portuguesa, tendo-se apresentado as sugestões que figuram no item 5 deste documento.

5. Sugestões para a constituição de um Banco de Dados da Língua Portuguesa.

Tendo em vista os interesses manifestados messas respostas, a Comissão autora deste documento gostaria de apresentar à discussão algumas formas de operacionalização de um Banco de Dados da Língua Portuguesa.

Alguns passos de caráter prático deveriam ser considerados:

1. Uma primeira observação deve destacar a dificuldade da empresa, visto que muitos acervos de língua falada ainda não estão devidamente transcritos, e visto que acervos da língua escrita ainda não foram especificados. Uma solução interessante para este problema é que seus detentores providenciassem localmente tais transcrições, encaminhando-as à Comissão adiante mencionada somente após essa providência. No caso dos textos escritos, os interessados escolheriam a edição considerada mais autorizada..

2. Completamento das informações contidas nas fichas de consulta, preenchidas pelos detentores de acervos, em 1994. Giselle Machline de Oliveira e Silva, que está inserindo os dados em DBASE, solicita os seguintes dados, em complementação áqueles já enviados:

(1) Período de tempo coberto pelo acervo.

(2) Indicar o tipo de transcrições adotado: fonológico, grafemático.

(3) Especificar a entidade que concedeu recursos para a constituição do acervo, e indicar até que ponto os responsáveis pelo acervo podem dispor dele.

(4) No quesito "descrição do acervo", tentou-se obter dados a respeito de uma série de informações que passamos a listar:

a) faixa etária dos informantes: no presente estágio da tabulação, dividimos aproximadamente essas faixas em B = bebês, C = crianças, A = adolescentes, J = jovens, D = adultos, V = velhos.

b) grau de instrução dos falantes: 0 = nenhuma, ou inferior a 4 anos; P = 4 anos, G = de 4 a 8 anos, C = segundo grau, S = superior.

c) quanto à extensão dos *corpora*, é necessário informar o número de horas gravadas, o número de fitas e de entrevistados, ou então o número de páginas de documentos, número de perguntas e questionários.

(5) Para facilitar a vida dos consulentes, é necessário detalhar o seguinte:

a) se se trata de entrevistas, questionários ou textos.

b) no caso de entrevistas, se são do tipo laboviano (ou DID do NURC), dialógicas (D2 do NURC), ou ainda de outro tipo, descrevendo-o.

c) se são escritos, dizer detalhadamente em que gênero se enquadram (por exemplo, manuscritos oficiais, cartas familiares ou de autores célebres, documentos literários, etc.).

(6) No caso de informantes que integram alguma etnia ou profissão, especificar (por exemplo, índios, pescadores, etc.).

(7) Finalmente, as colunas relativas a linguagem e *software* ficaram extremamente confusas. Construímos agora uma resposta com múltipla escolha, para disciplinar o quesito. Pedimos que os responsáveis pelos acervos nos mandem esses dados:

a) Editores de texto
( ) Qedit
( ) Word
( ) Carta Certa
( ) Wordstar
( ) ASCII
( ) Outro – qual*

b) Linguagem de Programação:
( ) Pascal
( ) Clipper
( ) Dbase III
( ) Outro – qual*

c) Equipamento
( ) MacIntosh-Apple
( ) 86-486
( ) PS1-PS2-IBM
( ) Grande porte – VAX~
( ) Grande porte – IBM
( ) Grande porte – UNISYS
( ) Outro – qual*

d) Aplicativos para procura e/ou concordância
( ) TACT
( ) Outro – qual*

3. Designação pela ABRALIN de uma Comissão Nacional encarregada de gerenciar a criação e a operacionalização do Banco. Dessa Comissão deveriam participar especialistas em Informática.

Suas atribuições seriam: (i) Completar a identificação dos acervos, para o que os titulares de acervo devem responder aos quesitos acima. (ii) Recolher materiais prontos para informatização, na forma do item anterior, submetendo-os previamente a um tratamento arquivístico, para assegurar sua recuperação. Preparação dos inventários descritivos. (iii) Tomar as decisões de caráter técnico: escolha dos *softwares* redatores, idem dos de busca eletrônica dos dados (programas de concordância), tipo de banco de dados a selecionar. (iv) Escolha do processo de implementação, conforme atrás indicado, ouvidos os detentores dos acervos: centralização? descentralização? Na primeira hipótese, que universidade ou universidades assumiriam o compromisso de gerenciar o Banco de Dados? (v) Especificação das formas de acesso aos dados coletados. (vi) Busca de financiamento.

4. Todo o processo de negociação entre os detentores de acervos seria conduzido por essa Comissão, que relataria seus passos à Diretoria da ABRALIN, para conhecimento e manifestação do respectivo Conselho.

Em 1995, solicitou-se aos detentores dos acervos o envio das informações suplementares aqui indicadas. Em sua reunião de julho do mesmo ano, a ABRALIN encarregou Gisselle M. de Oliveira e Silva da dar continuidade aos trabalhos. Perdemos, infelizmente, essa Colega em abril de 1996, e no momento se aguardam novas decisões para a retomada das atividades.

Notas

1 A primeira versão deste texto foi publicada no Boletim da Associação Brasileira de Lingüística (ABRALIN) 17 (julho de 1995):143-154.

Referências

LEMLE, M. 1984. Texto gerador. *Boletim da ABRALIN* 6:5-11.

PASSOS, C. 1984. Reflexões sobre a profissão do lingüista. *Boletim da ABRALIN* 6:17-26.

TARALLO, F. 1994. Dado é dado. *Boletim da ABRALIN* 6:27-33.

URIAGEREKA, J.1993. *Project "Minerva". Final Report.* College Park, University of Maryland, mimeo.

VOTRE, S. 1994. Para uma política de banco de dados. *Boletim da ABRALIN* 6:12-16.