Bases de dados de fala, linguagem e escrita: finalidades e possibilidades para a fonoaudiologia

Speech, language and writing data bases: purposes and possibilities for the speech and language clinics

Base de datos de habla, lenguaje y escritura: propósitos y posibilidades para la fonoaudiología

Regina Maria Freire[*]
Gisele Gouvêa da Silva[*]
Camila Parducci Arruda[*]

Resumo

A pesquisa em Fonoaudiologia, quando o foco é a fala e/ou escrita de sujeitos, pode demandar muito tempo do pesquisador, pois além da coleta de dados, estes precisam ser transcritos e digitados para serem analisados. Esta seria uma das razões para o pouco investimento em pesquisas sobre a fala de sujeitos, quer na instância da aquisição ou da patologia da fala, da linguagem e da escrita. Sugere-se que a Fonoaudiologia conheça as bases de dados em fala, linguagem e escrita disponíveis na rede virtual dado seu potencial para a pesquisa. Objetivo: apresentar as bases de dados no campo das ciências da linguagem para incentivar a pesquisa sobre a linguagem em funcionamento. O método escolhido foi organizá-las por sua filiação, maior ou menor, aos: a) estudos sobre variação linguística; b) estudos sobre aquisição de fala e linguagem, oral e escrita ou, mais particularmente, c) estudos sobre o objeto da Fonoaudiologia e sua clínica. Além disso, foram trazidos o histórico da criação de cada banco, os objetivos, a forma de acesso aos dados, a localização e as características principais. No caso dos estudos específicos sobre o objeto da Fonoaudiologia, o artigo detém-se sobre um banco em particular, detalhando seu conteúdo de forma estatística, as ferramentas para seu acesso e mostrando como a análise de dados interacionais pode instigar o fonoaudiólogo a direcionar suas pesquisas para

*Pontifícia Universidade Católica de São Paulo – PUCSP, São Paulo, Brasil.
Contribuição dos autores: RMF idealizadora e fundadora do Banco, responsável pela ideia do artigo com o qual todos os autores contribuíram igualmente
E-mail para correspondência: Regina Maria Freire - freireregina@uol.com.br

Recebido: 28/07/2016
Aprovado: 26/09/2016

este campo. Ao final, conclui-se que os bancos de dados têm uma importante contribuição aos estudos no campo da Fonoaudiologia.

Palavras-chave: *Base de dados; Linguagem; Escrita; Investigação laboratorial*

Abstract

Researching in Speech and Language clinics can demand a lot of time of the researcher because besides the data collection, these must be transcribed and typed before being analyzed. This would be one of the reasons for low investment in research on the speech of individuals, be it in speech and language acquisition or in speech and writing pathology. It is suggested that the speech and language clinics get acquainted with the databases in speech, language and writing that are available on the virtual network given its potential for research in speech therapy. Objective: to present the databases in the field of language sciences to encourage research on language functioning. The methodology is to present the databases organized by their filiation to: a) studies on language variation; b) studies on the acquisition of speech and language, oral and written, or c) studies on the subject of speech therapy and its clinics. In addition, we brought the history of the creation of each bank, the objectives, and the form of access to data, location and main characteristics. In the case of specific studies on language and speech, the article focuses on a bank in particular, to detail the tools to access and how the analysis of interactional data can instigate the speech and language therapist to direct their research into this field. Finally, it concludes that the databases have an important contribution to the studies in the field of speech therapy.

Keywords: *Databases; Language; Writing; Investigation*

Resumen

La investigación en Fonoaudiología, cuando la atención es el habla y/o la escritura de sujetos, puede tomar mucho tiempo del investigador porque además de la recogida de datos, hay la transcripción y digitación para el análisis. Esta seria una de las razones para la baja inversión en investigaciones sobre el habla de sujetos, sea sobre adquisición o patología del habla, del lenguaje, y de la escritura. Se sugiere que en la Fonoaudiología se conozca las bases de datos sobre el habla, el lenguaje y la escritura, disponibles en la red virtual, dado su potencial para la investigación. Objetivo: presentar las bases de datos en el campo de las ciencias del lenguaje para fomentar la investigación sobre el lenguaje en su funcionamiento. El método escogido fue organizarlas por su pertenencia, mayor o menor, a los: a) estudios sobre la variación lingüística; b) estudios sobre la adquisición del habla y del lenguaje, oral y escrita, y c) estudios sobre el objeto de la Fonoaudiología y su clínica. En la presentación se cuenta la historia de la creación de cada banco, los objetivos, los medios de acceso a los datos, la ubicación y las características principales. En le caso de los estudios sobre el objeto de la Fonoaudiología, el artículo se centra en un banco en particular, para detallar, de forma estadística, su contenido, las herramientas para el acceso y mostrar como el análisis de los datos de interacción puede instigar el fonoaudiólogo a dirigir sus investigaciones a este campo. Por último, se concluye mostrando porque los bancos de datos contribuyen de manera importante a los estudios en el campo de la Fonoaudiología.

Palabras clave: *Bases de datos; Lenguaje; Escrita; Investigación*

Introdução

Uma das razões que nos motivaram a publicizar bases de dados em fala, linguagem e escrita, veio da experiência prática com coleta e transcrição pois, para estudar a linguagem, sua aquisição e seus distúrbios, é necessário que o pesquisador interessado colete seus próprios dados, situação essa que exige uma parcela significativa de tempo. As bases de dados compartilhadas permitem ao pesquisador acessar diretamente o conteúdo nelas existente para pesquisar a linguagem em diferentes tipos de interação, gerando economia de tempo e imprimindo um ritmo ágil às pesquisas. O uso de base de dados em pesquisas no campo fonoaudiológico serve a diferentes métodos de aplicação que produzem leituras e interpretações quantitativas e qualitativas de uma mesma produção de fala, escrita e linguagem. Ainda, as bases de dados são instigadoras de questões e ampliam as discussões interdisciplinares, criando um espaço para as inquietações da área onde há condições de aprofundamento.

Encontramos, no campo das ciências da linguagem, uma série de bancos de dados que colecionam discursos, textos e diálogos gravados e transcritos nos mais variados contextos em que a fala ou a escrita é tomada como ponto de partida para a investigação. Estes bancos apresentam particularidades quanto à apresentação do projeto na página da internet e à configuração tecnológica de seus dados. Possuem colaboradores com formações diversas, a saber: linguistas, educadores, fonoaudiólogos, engenheiros da fala, oriundos de instituições, públicas e/ou privadas, voltadas à pesquisa. Vale ressaltar que esses bancos, em sua grande maioria, foram/são assistidos por instituições de amparo à pesquisa de origem nacional (CNPq, CAPES, FAPERJ, FAPESP e FAPEMIG), e por algumas instituições internacionais inseridas em projetos de cooperação internacional, o que possibilita discussões articuladas sobre a formação de pesquisadores, a produção e o desenvolvimento científico em suas áreas e, concomitantemente, em áreas afins. A partir das bases de dados, o fonoaudiólogo é convocado a produzir um dizer sobre as questões que se inscrevem em sua atuação como efeito de posicionamentos que assume. São estes posicionamentos, teóricos/técnicos, que vão configurar o seu trabalho e a continuidade da construção da Fonoaudiologia enquanto disciplina. Estudos desta natureza são importantes para a Fonoaudiologia e para as áreas que se interessam pela linguagem em aquisição e pela fala sintomática, pois estabelecem um funcionamento estrutural linguístico-discursivo do sujeito, desenvolvendo elementos clínicos e científicos para métodos e técnicas a serem usados na clínica.

Objetivos

Apresentar importantes bases de dados no campo das ciências da linguagem visando o desenvolvimento de pesquisas sobre aquisição e patologias de linguagem, a elaboração de conhecimento científico sobre a Fonoaudiologia e a problematização do olhar sobre a linguagem, enquanto objeto da Fonoaudiologia, que direciona seus estudos a partir de seu funcionamento nas dimensões normal e patológica.

Para atingir esses objetivos e orientar o leitor nesse percurso, identificamos as bases de dados que podem servir aos propósitos de pesquisadores da linguagem e que se destacam por trazerem aspectos sobre a língua, a fala e a escrita.

As bases de dados estão organizadas por sua filiação, maior ou menor, aos: a) estudos sobre variação linguística; b) estudos sobre aquisição de fala e linguagem, oral e escrita ou, mais particularmente, c) estudos sobre o objeto da Fonoaudiologia e sua clínica.

Em sua apresentação, traremos o histórico da criação de cada banco, os objetivos, o acesso aos dados, a identificação, a localização e as características principais. No caso dos estudos específicos sobre o objeto da Fonoaudiologia, o artigo detém-se sobre um banco em particular, para detalhar seu acesso e descrever o seu conteúdo, instigando o fonoaudiólogo a direcionar suas pesquisas para este campo. Ao final, conclui-se que os bancos de dados têm uma importante contribuição ao campo da Fonoaudiologia.

Base de Dados de Variação Linguística

O Projeto de Estudo da Norma Linguística Urbana Culta – NURC[1], teve seu início em 1969, no Brasil, quando se vinculou ao *Proyecto de Estudio Conjunto y Coordinado de la Norma Lingüística Culta de las Principales Ciudades de Iberoamérica y de la Península Ibérica*. A proposta do NURC no Brasil consiste em documentar e estudar a norma

falada culta de cinco capitais: Recife, Salvador, Rio de Janeiro, São Paulo e Porto Alegre. Os objetivos desse projeto são descrever os padrões da norma da língua falada pelos brasileiros, em seus aspectos fonológicos, fonético, morfossintático, lexical e estilístico. Em 1985, passou também a investigar a norma falada culta a partir da análise da conversação, análise da narrativa e da análise sócio-pragmática do discurso. Os informantes pertencem aos gêneros feminino e masculino, com ensino superior, divididos por três faixas etárias: a) 25 a 35 anos; b) 36 a 55 anos; c) 56 anos em diante, nascidos e sediados na cidade objeto de estudo por pelo menos três quartas partes de sua vida. O *corpus* constituído em cada cidade compreende três diferentes categorias de texto: elocuções formais, diálogos entre informante e documentador e diálogos entre dois informantes. O *corpus* nacional constitui-se de um total de 1.870 inquéritos gravados, perfazendo, aproximadamente, 1.570 horas de gravação. O acervo do Projeto NURC Rio de Janeiro, pertence à Faculdade de Letras da Universidade Federal do Rio de Janeiro, que disponibiliza o *corpora on-line*. Trata-se de entrevistas gravadas num total de 350 horas, com informantes com nível superior completo, nascidos no Rio de Janeiro e filhos de pais preferencialmente cariocas. As entrevistas entre informante e entrevistador, foram gravadas em fitas de áudio, que atualmente estão digitalizadas e transcritas. O material também foi publicado em 3 volumes. O NURC – São Paulo[2], vinculado à Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, conta com 381 entrevistas, 474 informantes e 316 horas de gravação, disponibilizadas em quatro volumes. O *NURC* Salvador[3] compreende 307 horas e 20 minutos de gravação que documentam o desempenho linguístico de 461 informantes, distribuídos em 360 inquéritos que cobrem as três categorias, distribuídas em: 58 Elocuções formais, 201 Diálogos entre informante e documentador e 101 Diálogos entre dois informantes. As gravações originais, em fitas de rolo e em fitas cassete, bem como, atualmente, em cópias digitalizadas, encontram-se catalogadas no Arquivo Sonoro do Setor de Língua Portuguesa do Instituto de Letras da Universidade Federal da Bahia. O NURC – Recife[4] está vinculado ao Programa de Pós-Graduação em Letras, da Universidade Federal de Pernambuco. O material coletado em Recife foi publicado em dois volumes, os registros sonoros fazem parte do acervo da biblioteca e conta com 346 inquéritos, 417 informantes e aproximadamente 290 horas de gravação. O NURC – Porto Alegre está vinculado ao Instituto de Letras da Universidade Federal do Rio Grande Sul e conta com 375 entrevistas, 472 informantes e 413 horas e 40 minutos de gravação.

O Programa de Estudos sobre o Uso da Língua – PEUL[5] – , vinculado à Universidade Federal do Rio de Janeiro, vem coletando diferentes amostras de fala e de escrita, desde 1980. O Banco de dados ficou conhecido como Amostra Censo 1980, quando se propôs a estudar os processos de variação e mudança da língua carioca. Os critérios utilizados na amostra foram estabelecidos segundo as faixas etárias: 15 a 25 anos, 26 a 49 e acima de 50 anos; a escolaridade a partir do 1º e 2º ciclos do ensino fundamental ao ensino médio e os gêneros masculino e feminino. Os falantes foram selecionados aleatoriamente em diferentes bairros da cidade do Rio de Janeiro. Em 1981, foi incluída no acervo uma amostra de 16 falantes na faixa de 7 a 14 anos, alunos de escolas municipais do Rio de Janeiro. As duas amostras somadas perfazem um total de 64 falantes, correspondendo a 64 horas de gravação que foram transcritas em um método semi ortográfico que busca respeitar as características inerentes da fala, disponíveis e explicitadas no site. Entre 1979 e 1981, foi estudada a incorporação do processo de variação e mudança na fala infantil. O acervo da Amostra de Fala Infantil é constituído de 32 entrevistas gravadas e transcritas em ortografia regular, com crianças cariocas, na faixa etária de 4 a 11 anos de idade. A Amostra Recontactados /2000 é caracterizada por novas gravações de 16 falantes que compuseram a Amostra Censo/1980, realizadas no período de 1999 a 2000. A Amostra Interacional consiste em capturar o uso da língua na fala espontânea, não controlada e, em situações de fala. O acervo é composto por 22 gravações realizadas entre 1989 e 1990. A Amostra Censo/2000 seguiu os mesmos procedimentos e método da Amostra Censo/1980 e conta com 32 horas de gravação, transcritas em sistema semi ortográfico. A Amostra Mobral trata do *corpus* da língua falada pelos estudantes do Movimento Brasileiro de Alfabetização de jovens e adultos, na década de 1970. O acervo contava, inicialmente, com 140 entrevistas gravadas eletromagneticamente, com duração de 1 hora em locais e circunstâncias variadas da cidade do Rio de Janeiro, contudo, apenas 59 entrevistas transcritas referentes a 12 informantes, estão dis-

ponibilizadas. A Amostra de escrita é caracterizada pelo discurso jornalístico extraído de jornais no Rio de Janeiro e publicados entre 2000 e 2004, como os populares *Extra* e *O Povo*, e os menos populares, *Jornal do Brasil* e *O Globo*. O acervo é composto por 75 cartas, 75 crônicas, Notas de Coluna Social, Editorial, Horóscopo, Notícias/Reportagens (geral e esportivas) e Artigos de Opinião (100 arquivos de cada gênero, vinte e cinco de cada jornal).

Estes bancos apresentam particularidades quanto à apresentação do projeto na página da internet e à configuração tecnológica de seus dados. Possuem colaboradores com formações diversas, a saber: linguistas, educadores, fonoaudiólogos, engenheiros da fala, oriundos de instituições, públicas e/ou privadas, voltadas à pesquisa. Como dito anteriormente, a maioria dos bancos foram/são assistidos por instituições de amparo à pesquisa de origem nacional (CNPq, CAPES, FAPERJ, FAPESP e FAPEMIG), e por algumas instituições internacionais inseridas em projetos de cooperação internacional, o que possibilita discussões articuladas sobre a formação de pesquisadores, a produção e o desenvolvimento científico em suas áreas e, concomitantemente, em áreas afins.

O projeto VARSUL[6] - Variação Linguística na Região Sul do Brasil foi criado em 1982, com o objetivo de descrever o português falado e escrito de 12 cidades sócio culturalmente representativas da região Sul do Brasil. O projeto conta com a participação de quatro universidades brasileiras: UFRGS, PUC-RS, UFSC e UFPR, que implantaram três acervos:

1. Banco de Dados VARSUL: caracterizado por 288 entrevistas, gravadas em cassete, com duração de 45 minutos cada, que resgatam os aspectos da vida pessoal do entrevistado e da história da cidade. O perfil social dos participantes resultou de 12 células sociais, nas quais levaram em consideração o gênero, a escolaridade (1 a 4 anos de escolaridade, 5 a 8 anos de escolaridade e de 9 a 11 anos de escolaridade), falar apenas português, ter morado ao menos 2/3 da vida na cidade, não ter morado fora da região mais que um ano durante a aquisição de linguagem. Não foram incluídos naquele momento os iletrados e universitários, nem a faixa etária abaixo de 25 anos.
2. Amostra Digital VARSUL: contém amostra digital em áudio, entre 5 e 15 minutos, de 24 entrevistas com falantes das três capitais do sul do país, de 8 entrevistas do Banco Monguilhott e 8 do banco Brescancini, representativos de zonas não urbanas de Florianópolis, a fim de servir como fonte de informação sobre as variedades sociolinguísticas e como fonte para a pesquisa sociocultural dos informantes e pode ser acessada diretamente pelo site do projeto.
3. Banco de Dados Diacrônico: está vinculado ao projeto nacional Para a História do Português Brasileiro. Trata-se de um acervo formado por documentos de cartório, arquivos públicos e privados e de textos de jornais da região sul do país, representativos da escrita sulista dos séculos XIX e XX, com acesso pela internet de alguns documentos, dado estar em processo de desenvolvimento.
4. O BDSer[7]: Banco de Dados de Fala da Serra Gaúcha - foi criado em 2000 e está vinculado ao Departamento de Letras da Universidade Caxias do Sul e conta com pesquisas sobre a língua falada na região serrana do Rio Grande do Sul. Os critérios para seleção dos informantes são residir na zona rural ou urbana, ser do gênero masculino e feminino, e das seguintes faixas etárias: 15 a 25 anos, 30 a 45 anos, 50 a 65 anos, 70 ou mais anos, escolaridade e nível superior. Os dados foram coletados por meio de Entrevistas Sociolinguísticas, realizadas em ambiente familiar e gravadas. O acervo conta ainda com a Ficha Social e a Ficha de Redes de comunicação. A equipe elaborou um roteiro de questões para a entrevista, para poder analisar a língua falada pelo informante de forma descritiva, narrativa e argumentativa. Obteve-se 64 informantes para cada município da área de abrangência.

O Projeto ASPA[8] - Avaliação Sonora do Português Atual é vinculado ao LABFON – Laboratório de Fonética da Faculdade de Letras da UFMG - e tem como objetivo construir conhecimento probabilístico da estrutura do Português-Brasileiro, contribuindo, especificamente, para a análise e mapeamento dos tipos fonológicos, silábicos e segmentais do Português. Esse projeto se apoia teoricamente nos modelos de Fonologia de Uso e Teoria dos Exemplares, em que a multirepresentacionalidade da linguagem está em questão. O Projeto ASPA ainda não está totalmente disponível online, porque atualmente passa pelo procedimento de verificação e correção do *corpus*. Os interes-

sados devem realizar um cadastro e enviar por e-mail uma solicitação dos critérios de busca para consultar o *corpus*, que será retornado por e-mail no formato PDF. O usuário pode restringir sua busca da maneira que quiser, criando uma combinação de regras em quaisquer campos cadastrados :

1. Ortografia: exemplo: palavras que comecem com "in", palavras que possuem "nha", palavras que terminam em "ado".
2. Categoria: substantivo, verbo, adjetivo, advérbio, artigo, conjunção, interjeição, preposição, pronome, numeral.
3. Morfologia: flexionado verbal, flexionado plural, derivado, original.
4. Origem: africana, indígena, nenhuma (outra que não seja as duas anteriores).
5. Transcrição: quaisquer segmentos, sequências de segmentos, sílabas específicas, índice de tonicidade, etc. Ex: palavras oxítonas que terminem com [S], palavras que possuem a sequência [di] seguida de consoante fricativa, palavras que possuem a sílaba [bu].

O Projeto Norte Vogais[9] segue os objetivos do PROBRAVO[10-11], grupo de pesquisa "Descrição Sócio-Histórica das Vogais do Português (do Brasil)", criado em 2005 e, atualmente, envolve 21 universidades brasileiras. Tem como objetivo investigar e descrever as realizações fonéticas das vogais nos dialetos do Sul ao Norte do Brasil. Este grupo de pesquisa busca compreender:

1. Como são realizadas foneticamente as vogais no Português (do Brasil)?
2. Como se explica ou o que motiva, a diversidade de realizações fonéticas?
3. Como os falantes do Português (do Brasil) se entendem apesar das diversidades da qualidade vocálica?
4. É possível explicar essa diversidade gramaticalmente?

O projeto Norte Vogais busca caracterizar o sistema vocálico átono e suas variantes, com base em amostra estratificada e em termos variacionistas, assim como analisar qualitativamente e explicar o processo de variação das vogais médias pretônicas e postônicas não finais no português falado no Norte do Brasil condicionado por fatores internos. As pesquisas são realizadas a partir da investigação de *corpora* com amostras de fala das variedades linguísticas do português da Amazônia paraense situadas na zona do português regional paraense. Os *corpora* do projeto Norte Vogais possuem um número total de 318 (trezentos e dezoito) informantes nativos da Amazônia Paraense, originários de cinco variedades locais: Belém, Cametá, Breves, Breu Branco e Mocajuba, em suas zonas rural e urbana, cuja idade varia de 24 a 72 anos. Os informantes que compõem o *corpus* foram estratificados socialmente em sexo, escolaridade, faixa etária e procedência e a situação de fala predominante é a de narrativas de experiência pessoal. Todo o *corpus* encontra-se transcrito grafematicamente e com os dados que atestam ocorrência do fenômeno-alvo – vogais médias pretônicas – transcritos foneticamente. Além das transcrições, o *corpus* contém o áudio das gravações realizadas em trabalho de campo.

Sintetizando, a proposta das bases de dados apresentadas é documentar e estudar a norma falada culta de algumas capitais e regiões do Brasil, em seus aspectos fonológico, fonético, morfossintático, lexical e estilístico, contribuindo, especificamente, para a análise e mapeamento dos tipos fonológicos, silábicos e segmentais do Português.

Neste momento, caberia perguntar: o que o fonoaudiólogo pesquisador tem a ver com a norma falada culta ou com os aspectos linguísticos que foram levantados? Ou melhor, o que a norma falada culta tem a ver com o profissional interessado na fala sintomática? De que forma ele poderia aproveitar ou extrair saberes dos dados dos bancos para a sua clínica? A resposta pode transitar por várias searas: primeiro, os bancos em questão nos dizem que é possível armazenar dados de fala e trazem, de forma clara, a metodologia usada no levantamento e armazenamento de dados de fala e linguagem e os requisitos para que um dado seja considerado confiável; segundo, com base nessas informações pode-se elaborar um banco próprio; terceiro, o normal pode nos fornecer parâmetros para pensar-se o patológico e, ainda, encontrar na normalidade o funcionamento dialógico que, tomado como manejo, pode afetar os dizeres dos que vem a ter na clinica fonoaudiológica.

Base de dados de fala e linguagem

A Plataforma de Documentos Sonoros do IEL-UNICAMP[12-13] é uma importante base de registros em áudio de pesquisas sobre a linguagem realizadas no interior do instituto, disponibilizados no formato MP3. Conta, por exemplo, com a coleção do

Projeto de Aquisição de Linguagem, coordenado pela Dra. Claudia De Lemos, em que a fala de oito crianças brasileiras com idades entre 11 meses e cinco anos foram gravadas em rolo, digitalizadas em 423 horas e transcritas em 14 mil páginas de diálogos, situações cotidianas das crianças[14], como almoço ou jantar, momentos de brincadeira ou contação de histórias, em interação com pais, irmãos e amigos. Os dados foram coletados entre 1976 e 1981. O *corpora* já serviu de base empírica para artigos científicos, dissertações de mestrado e teses de doutorado sobre a aquisição da linguagem pela criança e suas relações com a língua e com a fala do outro, não somente na área da Linguística como na Fonoaudiologia.

O Banco de Dados do Centro de Estudos de Linguagem e Fala – CELF – está vinculado ao Programa de Pós-Graduação em Distúrbios da Comunicação Humana, da Universidade Federal de Santa Maria - RS. O CELF realiza pesquisas na área da Fonoaudiologia[15-16], especificamente, sobre fonologia clínica e aquisição da linguagem normal e com desvios, desde 1997. O acervo do Banco CELF é constituído por material extraído de entrevistas, avaliações e sessões de terapia fonoaudiológica realizadas na clínica-escola da universidade, pelos alunos de graduação e pós-graduação. O acesso pela internet nos permite encontrar uma série de artigos, dissertações e teses que fizeram uso dos dados do banco. Há por exemplo, pesquisas que realizam o estudo comparativo[17] de diferentes modelos de terapia fonológica e o estudo sobre a generalização[18] em três modelos de terapia fonológica em crianças com diferentes graus de severidade do desvio fonológico.

No interior do LAFAPE – Laboratório de Fonética e Psicolinguística vinculado à linha de pesquisa COGITES – cognição, interação e significação do Instituto de Estudos da Linguagem – IEL – UNICAMP, encontra-se o APHASIACERVUS, que trata de dados linguísticos-interacionais de afasia, coletados no Centro de Convivência dos Afásicos de 2003 até 2012. O CCA é um espaço de interação entre pessoas afásicas e não afásicas cujo objetivo é desenvolver estudos linguísticos e neurolinguísticos, bem como garantir às pessoas afásicas efeitos terapêuticos e sociais possibilitados por um conjunto variado de experiências interacionais cotidianas. Mais uma vez, o problema encontrado para implantar o banco de dados foi a transcrição e a interpretação dos dados de afasia coletados e estudados, uma vez que conta com interações entre duas ou mais pessoas afásicas ou não afásicas. As particularidades dos aspectos verbais e não verbais dos falantes, como as manifestações sintomáticas da afasia e da dimensão multimodal da linguagem e da interação, contribuem para a complexidade do banco de dados. O material que compõe o *AphasiAcervus* ainda não está totalmente digitalizado e transcrito, mas conta com dispositivos computacionais para captura e manipulação do material audiovisual. Outro ponto importante a esclarecer é que o processo de regras e usos do APHASIACERVUS ainda não foi concluído, o que restringe o acesso aos integrantes do LAFAPE. A composição do acervo é caracterizada por:

1. acondicionamento de dados físicos e digitais dos encontros semanais do CCA, registrados em formato audiovisual;
2. digitalização do registro audiovisual dos encontros semanais do CCA;
3. sistemas heterogêneos de transcrição do material registrado em vídeo;
4. registros de ações tomado *in situ*, realizados por pesquisador presente aos encontros do CCA;
5. descrição do perfil linguístico/neurolinguístico dos participantes do CCA;
6. em construção: acervo de pesquisas já concluídas (individuais e coletivas) baseadas no *corpus AphasiAcervus*; ações documentais.

O Projeto E-LABORE[19] - Laboratório Eletrônico de Oralidade e Escrita - da Faculdade de Letras da Universidade Federal de Minas Gerais (UFMG) contém um acervo de redações de crianças de 6 a 12 anos de idade, que frequentam escolas da rede pública e particular em Belo Horizonte, Minas Gerais. Esse laboratório busca investigar os desvios ortográficos que ocorrem na escrita da criança durante o processo de alfabetização. O material consta da escrita de crianças às quais foi oferecida uma folha pautada para uma produção (redação ou do desenho) sobre um tema escolhido pelos professores. As escolas participantes recebem uma lista de 8 instruções e um questionário e a identificação de cada produção contém uma sequência de 8 dígitos, como um código de barras. Dois tipos de informações são coletadas: a primeira com o nome, data de nascimento, gênero e série da criança e, a segunda, com dados da escola, como endereço, telefone, tipo de escola e pessoal responsável. Todas as produções

escritas das crianças são digitalizadas e digitadas pela equipe do E-LABORE, ainda indisponíveis ao público interessado. Segundo os criadores do laboratório, esse processo de digitação encontra muitas dificuldades, tanto pela exigência de um número elevado de redações para ser estatisticamente representativo, como da interpretação da letra de crianças que ainda atravessam a aquisição da escrita. A padronização das digitações seguiu um conjunto de 7 regras:

1. Digitação das redações: os textos devem ser copiados de forma que a versão digitada seja a mais parecida possível com o original, como por exemplo, as quebras de linhas realizadas pelos autores são apontadas na digitação.
2. Marcação de parágrafo: os textos devem ser digitados respeitando a quebra de parágrafo, por exemplo, o digitador coloca uma sequência de dois "enter", para iniciar um novo parágrafo.
3. Marcação de erros: os erros devem ser copiados entre chaves e o digitador deve incluir o correto em colchetes.
4. Dificuldade de leitura de uma ou mais palavras: a palavra ilegível será substituída por um * (asterisco).
5. Ausência de uma ou mais palavras: ao notar que falta uma palavra que dê sentido à produção da criança será inserido um sinal de +, e, entre colchetes incluir a palavra que o digitador considere que esteja faltando.
6. Começo e fim de texto contínuo: o digitador deverá colocar um $ (cifrão) para marcar o início e o fim de um texto.
7. Hifenização: o digitador deverá copiar as marcas de hifenização do texto da criança.

Estas bases de dados registram pesquisas sobre a aquisição da linguagem e sobre a linguagem com desvios, além de caracterizar a escrita de crianças em processo de escolarização. Neste caso são bancos que estreitam as relações entre a linguística – tanto na aquisição como nos casos de afasia – e a Fonoaudiologia. As relações entre os estudos em aquisição de linguagem e a Fonoaudiologia foram amplamente divulgadas em publicações da área. Por outro lado, os estudos sobre desvios estão diretamente relacionados com o campo, pois entre os pesquisadores há fonoaudiólogos que, pela montagem do banco, identificam suas vantagens. Já o banco de escrita pode trazer dados para o mapeamento das andanças da criança em processo de aquisição da escrita. Permite, ainda, que se discuta a relação entre letramento e alfabetização. Estes são apenas alguns dos usos e alternativas para a Fonoaudiologia em seu diálogo com os bancos de dados de fala, linguagem e escrita.

O Banco de Dados de Fala e Escrita

O Banco de Dados de Fala e Escrita é um exemplo de base de dados específico da Fonoaudiologia, que se utiliza de uma coleção organizada de registros audiovisiográficos de falantes e escreventes. Trata-se de um banco digital, disponibilizado pela internet por meio de site específico, com acesso livre aos usuários interessados em pesquisar a aquisição e as patologias de linguagem oral e escrita, sob a perspectiva de quaisquer modelos teóricos e metodológicos. O Banco está vinculado à linha de pesquisa Linguagem e Subjetividade[20] do PEPG em Fonoaudiologia da PUC/SP.

Apesar de sua fundação ter ocorrido apenas em 2001, o seu início se deu muitos anos antes, no final da década de 1970, quando uma fonoaudióloga e pesquisadora pretendia investigar a aquisição de linguagem sob a perspectiva da interação mãe e criança, utilizando como material de pesquisa a gravação e transcrição de diálogos em situações lúdicas de uma criança em processo de aquisição de linguagem e sua mãe. Essa pesquisa teve o caráter de investigar a aquisição de linguagem a partir de um desenho de estudo longitudinal, tendo como foco os processos dialógicos que aconteciam nas interações entre mãe e criança.

Na década de 1980, quando a pesquisadora ministrava a disciplina aquisição de linguagem e realizava supervisão clínica em uma instituição educacional na cidade de São Paulo, que atendia crianças de 0 a 5 anos e 11 meses, resolveu, juntamente com seus alunos de graduação em Fonoaudiologia, investigar a aquisição de linguagem de crianças institucionalizadas e não institucionalizadas, utilizando como estratégia prática e investigativa a gravação de situações de interações dialógicas de crianças em diferentes faixas etárias, tendo como tipificação do desenho de estudo o recorte transversal. Na década de 1990, a pesquisadora investigou, de forma paralela, o processo de aquisição de linguagem de duas das crianças institucionalizadas e outra que atendia em seu consultório com queixa de atraso de linguagem e diagnóstico médico de fissura palatina. Esse trabalho serviu de base para

a sua pesquisa de doutorado, posteriormente publicada em livro. Nas décadas seguintes manteve-se esse tipo de investigação e coleta com a parceria dos alunos de Fonoaudiologia e passou-se a incluir alguns casos clínicos de sujeitos com sintomas de linguagem que eram atendidos pela pesquisadora e fonoaudióloga, por apresentarem, por exemplo, atraso de linguagem por surdez, distúrbio específico da linguagem, dislalias, gagueiras, disfonias, atraso de aquisição da escrita, distúrbios de aprendizagem e afasias. Esse tipo de coleta de dados para investigação do funcionamento da linguagem se deu até o ano de 2004, momento em que foi preciso interromper as coletas para organizar o material, digitalizá-lo e formatá-lo de maneira padronizada, para disponibilizá-lo aos usuários da rede. Salienta-se que o banco foi aprovado pelo comitê de ética da Pontifícia Universidade Católica de São Paulo sob o nº 202/2009 .

O Banco de Dados de Fala e Escrita pretende fornecer: (i) subsídios para a descrição do processo de aquisição da fala e da escrita; (ii) condições para análise e desenvolvimento de teorias fonoaudiológicas interessadas no falante; (iii) condições para formação de novos pesquisadores por meio do ensino de técnicas de observação, análise e compreensão do funcionamento normal e patológico da linguagem; (iv) subsídios para programas educacionais, promovendo o conhecimento do processo de aquisição da linguagem, da fala e da escrita, em diferentes faixas etárias e em diferentes meios linguísticos; (v) subsídios para a pesquisa do funcionamento da linguagem, pelas perspectivas clínica e não clínica; (vi) subsídios para a pesquisa sobre a terapêutica, em particular no atendimento de sujeitos com problemas de linguagem oral ou escrita; (viii) meios de divulgação do conhecimento gerado a partir de pesquisas com dados de fala, linguagem e escrita, armazenados no banco de dados.

Possui, atualmente, dados referentes à 321 sujeitos, totalizando cerca de 641 *corpora*. Há dados de interação mãe-criança, criança-criança, terapeuta-paciente e terapeuta-familiares e de interação entre adultos. O acesso ao *corpus* escolhido para análise é intermediado por ferramentas de busca que permitem localizar: coletas transversais e/ou longitudinais, *corpora* de crianças ou de adultos, de ambos os sexos, em situações dialógicas de tipo diádico, triádico e polidiádico, lúdico e terapêutico. Os *corpora* estão digitalizados em arquivos de texto, áudio e vídeo e, ainda, referem-se a coletas de linguagem em aquisição e/ ou sintomática, sendo que, neste caso, as coletas se organizam por procedimentos clínicos: entrevista, avaliação, atendimento, orientações. Foram estruturados com base nas regras da Linguística de *Corpus*, ou seja, o material discursivo coletado obedece a critérios que asseguram homogeneidade e legitimidade ao material como: origem, propósito, composição, formatação, representatividade e extensão dos dados. Os corpora foram transcritos de acordo com as regras do Projeto de Estudo da Norma Lingüística Urbana Culta de São Paulo [21] .

Para ter acesso à coleção de corpus é necessário apenas que o pesquisador preencha o cadastro na plataforma: www.bancofalaescrita.org , e aguarde a liberação para o uso. Após a análise do preenchimento, o acesso pode ser liberado pelos administradores do banco.

Várias pesquisas foram e têm sido realizadas a partir dos dados disponibilizados pelo Banco. Podemos citar autores (2007)[22], (2011)[23] e (2016)[24].

Uma das vantagens deste Banco em relação aos outros é a sua neutralidade, ou seja, o fato de que os dados são coletados, e deixam ao pesquisador a escolha da vertente que deseja alçar para analisar seus dados. Cabe lembrar que além desta, temos o fato de que os dados, por serem de natureza clínica e não clínica e longitudinais e transversais, permitem o acompanhamento de um evento de fala ao longo do tempo ou seu reaparecimento em diversas crianças no mesmo momento do processo de aquisição. Outra vantagem bastante relevante é o fato de que o acesso aos dados é direto e aberto àqueles que se cadastrarem. Os dados em vídeo, por questões de segurança podem ser acessados a partir da presença do interessado na instituição, ou seja, a partir de um software específico e local.

Conclusão

Os bancos, de uma forma geral, e o Banco de Dados de Fala e Escrita, de forma particular, se inscrevem como iniciativa para que as pesquisas científicas extrapolem os meios impressos e adquiram um alcance mais extenso. Estudos já realizados com os dados dos Bancos apresentados corroboram para explicar a estrutura e o funcionamento da fala, da linguagem e da escrita. O almejo pela expansão da transmissibilidade de dados advém da importância em destinar, a um número cada vez maior de pessoas, novas fontes e resultados de pesquisas.

Para cumprir tal direcionamento de informações, a internet é um dispositivo imprescindível, uma vez que possibilita a divulgação de dados e permite compartilhar informações. Conclui-se que o diálogo com a comunidade fonoaudiológica e a disseminação do uso de *corpora* dos Bancos apresentados contribuiria para o desenvolvimento de estratégias clínicas e científicas para a elaboração e o manejo de métodos e técnicas a serem usados tanto na clínica quanto na pesquisa, apontando sua importância para a Fonoaudiologia, bem como para as áreas que se interessam pela Linguagem em aquisição e por seus sintomas na fala e escrita.

Referências bibliográficas

1. Silva LA. Projeto NURC: Histórico. Linha d'Água Rev. Julho, 1996; (10) 83-9. Disponível em http://www.letras.ufrj.br/nurc-rj/

2. Urbano, H. "Apresentação". In: Preti, D. e Urbano, H. (org.). A linguagem falada culta na cidade de São Paulo". São Paulo: T.A. Queiroz/FAPESP, 1988, v. III, p. 1-12.

3. Projeto Atlas Linguístico do Brasil. Salvador: UFBA. 2013 (acesso em 2016 abr 10). Disponível em: twiki.ufba.br/twiki/bin/view/Alib/AlibNurc

4. Programa de Pós - Graduação em Letras Universidade Federal de Pernambuco. Pernambuco: UFPE 2016. (acesso em 2016 abr 04). Disponível em: http://www.pgletras.com.br/programa-nucleos-nurc.htm

5. PEUL - Programa de Estudos Sobre o Uso da Linguagem. Rio de Janeiro: UFRJ (acesso em 2016 abr 24). Disponível em: www.letras.ufrj.br/peul

6. Projeto Varsul – Variação Linguística na Região Sul do Brasil. UFRGS, PUC-RS, UFSC, UFPR. (acesso em 2016 abr 10). Disponível em: www.varsul.org.br

7. Battisti E. BDSer: Um Banco de Dados de Fala da Serra Gaúcha. 5º Encontro do Círculo de Estudos Linguísticos do Sul; 2002; Curitiba. Curitiba, Mídia Curitibana, 2003.

8. Cristófaro-Silva T, Almeida LS, ASPA: a formulação de um banco de dados de referência da estrutura sonora do português contemporâneo. In Anais do XXV Congresso da Sociedade Brasileira de Computação, 2005, São Leopoldo.

9. Cruz R. Vogais na Amazônia Paraense. Alfa Rev. Linguíst. 2012; 56 (3). Disponível em: seer.fclar.unesp.br/alfa/article/view/4948

10. Probravo - Grupo de Pesquisa sobre a Descrição Sócio-Histórica das Vogais do Português (do Brasil). Belo Horizonte: UFMG. (acesso em 2016 jul 30). Disponível em: relin.letras.ufmg.br/probravo

11. Lee, SH. Vogais Além de Belo Horizonte, FALE/UFMG, 2012. (acesso em 2016 jul 28) Disponível em: www.letras.ufmg.br/site/e-livros/VogaisAlemdeBH2012.pdf

12. Sarmento SR, Siqueira BS. IEL - Instituto de Estudos da Linguagem. Campinas: UNICAMP. 2007 (acesso em 2016 abr 24). Disponível em: www.iel.unicamp.br/projetos/cogites/pdf/cca_descricao.pdf

13. CEDAE - Centro de Documentação Cultural Alexandre Eulálio. Campinas: UNICAMP. 2014 (acesso em 2016 jul 28). Disponível em: eulalio.iel.unicamp.br/sys/audio/

14. De Lemos CTG. Das vicissitudes da fala da criança e de sua investigação. Cad. de Estudos Linguísticos. Janeiro, 2002; 42(1) 41-69. Disponível em: revistas.iel.unicamp.br/index.php/cel/article/view/1599/1178

15. Backes FT, Pegoraro SP, Costa VP, Mota HB. Caracterização das estratégias de reparo incomuns utilizadas por um grupo de crianças com desvio fonológico. Rev. CEFAC. Junho, 2013. Disponível em:www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-18462013005000031&lng=en.

16. Pereira AS, Keske-Soares M. Patologia de linguagem e escuta fonoaudiológica permeada pela psicanálise. Psico. Outubro, 2010; 41(4):517-24.

17. Keske-Soares M, Brancalioni AR, Marini C, Pagliarin KC, Ceron MI. Eficácia da terapia para desvios fonológicos com diferentes modelos terapêuticos. Pró-Fono R. Atual. Cient. Setembro, 2008; 20 (3).

18. Ceron MI, Attoni TM, Quintas VG, Keske-Soares M. A generalização estrutural silábica no tratamento do desvio fonológico. Rev. CEFAC. 2011;13 (1) 35-40.

19. Cristófaro-Silva T, Almeida LS, Oliveira-Guimarães DML; Martins RMF; e-Labore – Laboratório Eletrônico de Oralidade e Escrita. Belo Horizonte: Faculdade de Letras Universidade Federal de Minas Gerais. 2009 (acesso em 2016 agosto 08). Disponível em: www.projetoaspa.org/elabore

20. Linguagem e Subjetividade. São Paulo: Estudos Pós-Graduados em Fonoaudiologia da PUC-SP (acesso em 2016 jun 30). Disponível em: www.pucsp.br/linguagemesubjetividade

21. Castilho, A. & Preti, D. (orgs.) (1986). A Linguagem Falada Culta na Cidade de São Paulo. Materiais para seu estudo. São Paulo: TAQ/Fapesp, vol. I, Elocuções Formais.

22. Gouvêa GS. Por uma multiestratificação estrutural dos sintomas de linguagem.(Dissertação). São Paulo (SP): Pontifícia Universidade Católica de São Paulo. Mestrado em Programa de Estudos Pós Graduados em Fonoaudiologia; 2007.

23. Gouvêa G, Freire RMAC, Dunker CIL. Sanção em Fonoaudiologia: um modelo de organização dos sintomas de linguagem. Cad. de Estudos Linguísticos. 2011; 53 (1) 07-25.

24. Lieber, SN Aspectos da constituição de uma criança surda pela fala do ouvinte: de traços a significantes. Dissertação). São Paulo (SP): Pontifícia Universidade Católica de São Paulo. Mestrado em Programa de Estudos Pós Graduados em Fonoaudiologia; 2015.