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

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

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

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

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*

*Pontifícia Universidade Católica de São Paulo – PUCSP, São Paulo, Brazil.
Author's contributions: RMF - creator and founder of the Bank, responsible for the idea of the article with which all authors contributed equally

Correspondence address: Regina Maria Freire - freireregina@uol.com.br
Received: 28/07/2016
Accepted: 26/09/2016

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 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*

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*

Introduction

One of the reasons that led us to publish speech, language, and writing databases came from practical experience with collecting and transcribing data. To study a language, its acquisition and disorders, it is necessary that the interested researcher gather his own data, which requires a significant amount of time. Shared databases allow the researcher to directly access their existing content to research the language in different types of interaction, saving time and allowing a quicker pace in researching. The use of databases in research in Speech Language Pathology and Audiology fits different application methods that produce quantitative and qualitative readings and interpretations of the same speech, writing, and language production. Databases also instigate questions and broaden interdisciplinary discussions, creating space for concerns in the area where it is possible to do further research.

In the field of language sciences, we find a number of databases that collect speeches, texts, and dialogues recorded and transcribed in various contexts in which speech or writing is taken as a starting point for investigation. These databases have particularities regarding both the presentation of the project on the webpage and the technological configuration of their data. They have contributors with diverse backgrounds: linguists, educators, speech therapists, speech engineers, from public and/or private research-driven institutions. It is noteworthy that the vast majority of these databases were/are assisted by national institutions of support to research (CNPq, CAPES, FAPERJ, FAPESP and FAPEMIG), and by some international institutions that participate in international cooperation projects, which enables articulated discussions on the training of researchers, production and scientific development in their areas, and, concomitantly, in related fields. From the databases, the speech therapist is called upon to give his opinion on matters related to his area and, therefore, assuming a position. These theoretical/technical positions will be the basis for the research and the continued construction of Speech Language Pathology and Audiology as a discipline. Studies of this nature are important for Speech Language Pathology and Audiology and for areas interested in the acquisition of language and in symptomatic speech, for they set the linguistic-discursive structural function of the subject, developing clinical and scientific elements for methods and techniques to be used in speech therapy clinic.

Objectives

To represent important databases in the field of language sciences for the development of research on language acquisition and its pathologies, the development of scientific knowledge on Speech Language Pathology and Audiology, and the positing of language as the object of Speech Language Pathology and Audiology, which directs its studies according to its normal or pathological functions.

To achieve these goals and guide the reader through this course, we identified the databases that can serve language researcher's purposes and are characterized by bringing aspects of language, speech and writing.

The databases are organized by their belonging in a greater or lesser extent to: a) studies of language variation; b) studies on the acquisition of speech and language, oral and written, or more particularly, c) studies on the subject of Speech Language Pathology and Audiology and its clinical practice.

In this presentation we will bring the history of the creation of each database, its objectives, access to data, identification, location and main characteristics. In the case of specific Speech Language Pathology and Audiology studies, the article dwells on a particular database to detail its access and describe its contents, prompting the speech and language therapist to direct his research to this field. In the end, one may conclude that the databases make an important contribution to the field of Speech and Language Pathology and Audiology.

Linguistic Variation Database

The *Projeto de Estudo da Norma Linguística Urbana Culta - NURC1 (*Project for the Study of the Urban Linguistics Norm of the Educated), began in 1969 in Brazil, when it was linked to the *Proyecto de Estudio Conjunto y Coordinado de la Norma Lingüística Culta de las Principales Ciudades de Iberoamérica y de la Península Ibérica (*Project for the Joint and Coordinated Study of the Linguistics Norm of the Educated in the Main Cities of Latin America and the Iberian Peninsula).

The proposal of *NURC* in Brazil is to document and study the spoken norm of the educated of five capitals: Recife, Salvador, Rio de Janeiro, São Paulo, and Porto Alegre. The objectives of this project are to describe the patterns of the norm of the language spoken by Brazilians in their phonological, phonetic, morph-syntactic, lexical, and stylistic aspects. In 1985, it also began to investigate the spoken norm of the educated from the analysis of conversation, analysis of narrative, and the socio-pragmatic analysis of discourse. The informants are males and females with higher education, divided into three age groups: a) 25-35, b) 36-55, and c) 56+ years, born and residents of the city of study for at least three-quarters of their lives. The *corpus* in each city comprises three different categories of text: formal speeches, dialogues between informant and documenter, and dialogues between two informants. The national *corpus* consists of a total of 1,870 recorded inquiries, totaling approximately 1,570 recorded hours. The collection of *Projeto NURC Rio de Janeiro* belongs to *Faculdade de Letras da Universidade Federal do Rio de Janeiro*, which provides the *corpora online*. It is comprised of 350 hours of recorded interviews with respondents who have a college degree, born in Rio de Janeiro, and preferable children of *carioca* parents. Interviews between informant and interviewer were recorded on audiotapes, which have since been digitalized and transcribed. The material was also published in three volumes. *NURC - São Paulo*[2], linked to *Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo*, is comprised of 381 interviews, 474 respondents and 316 hours of recording, available in four volumes. *NURC Salvador*[3] consists of 307 hours and 20 minutes of recording documenting the linguistic performance of 461 respondents, with over 360 surveys covering the three categories, divided into: 58 formal speeches, 201 dialogues between informant and documenter, and 101 dialogues between two informants. The original recordings, tape reels and cassette tapes, and currently also in digital format, are catalogued in the *Arquivo Sonoro do Setor de Língua Portuguesa do Instituto de Letras da Universidade Federal da Bahia*. The *NURC - Recife*[4] is linked to the *Programa de Pós-Graduação em Letras, da Universidade Federal de Pernambuco*. The material collected in Recife was published in two volumes. The sound recordings are part of the library collection and have 346 inquiries, 417 respondents and approximately 290 hours of recording. The *NURC - Porto Alegre* is linked to the *Instituto de Letras da Universidade Federal do Rio Grande do Sul* and is comprised of 375 interviews, 472 respondents, and 413 hours and 40 minutes of recording.

The *Programa de Estudos sobre o Uso da Língua - PEUL*[5] (Studies on Language Usage Program) - linked to the *Universidade Federal do Rio de Janeiro*, has been collecting different samples of speech and writing since 1980. The database became known as *Amostra Censo 1980* when it proposed to study the processes of variation and change in the *carioca* language. The criteria used in the sample were established according to age groups: 15-25 years, 26-49, and, 50+ years; from 1st and 2nd cycles of elementary school to high school, and male and female genders. The speakers were selected at random from different parts of the city of Rio de Janeiro. In 1981 a sample of 16 speakers ages 7-14 years, students of municipal schools of Rio de Janeiro, was included in the collection. Both samples added together make up a total of 64 speakers, corresponding to 64 hours of recording that were transcribed in a semi-orthographic method that seeks to respect the inherent characteristics of speech. These are available and explained on the site. Between 1979 and 1981, the incorporation of process variation and change in children's speech was considered. The *Amostra de Fala Infantil* collection consists of 32 interviews recorded and transcribed using correct spelling, with *carioca* children, ages 4-11 years old. New recordings of 16 speakers who composed the Amostra Census/1980 held from 1999 to 2000 characterize the Amostra Recontactados/2000. The interactional sample captures the use of language in non-controlled, spontaneous speech, and in conversational situations. The collection consists of 22 recordings made between 1989 and 1990. The *Amostra Censo/2000* followed the same procedures and method of *Amostra Censo/1980* and has 32 hours of recording, transcribed in a semi-orthographic system. The *Amostra Mobral* is the *corpus* of spoken language of students of the *Movimento Brasileiro de Alfabetização (*Brazilian Movement of Literacy*)* for youth and adults in the 1970s. The collection had initially 140 electromagnetically recorded interviews, lasting 1 hour each, in varied circumstances and places of the city of Rio de Janeiro. However, only 59 transcribed interviews

related to 12 informants are available. The *Amostra de Escrita* is characterized by journalistic discourse extracted from newspapers in Rio de Janeiro and published between 2000-2004, such as the tabloids *Extra* and *O Povo*, and the regular newspapers, *Jornal do Brazil* and *O Globo*. The collection consists of 75 letters, 75 chronicles, social column entries, editorials, horoscopes, news/reports (general and sports) and opinion articles (100 files each, 25 from each newspaper).

These databases have unique characteristics regarding the presentation of the project on the webpage and technological configuration of their data. They have collaborators with diverse backgrounds: linguists, educators, speech therapists, and speech engineers, who come from research-driven public and/or private institutions. As stated earlier, most databases were/are assisted by institutions of national origin that offer support to research (*CNPq, CAPES, FAPERJ, FAPESP* and *FAPEMIG*), and by some international institutions involved in projects of international cooperation. This enables discussions on training of researchers, scientific production and development in their areas and, at the same time, in related fields.

The *Projeto VARSUL*[6] *- Variação Linguística na Região Sul do Brasil* (Linguistic Variation in Southern Brazil) was created in 1982 with the objective of describing the spoken and written Portuguese of 12 social and culturally representative cities in Southern Brazil. The project includes the participation of four Brazilian universities: *UFRGS, PUC-RS, UFSC* and *UFPR*, having set up three collections:

1. *VARSUL Database*: characterized by 288 taped interviews of 45 minutes each, recording aspects of the interviewee's personal life and of the city's history. The participants social profile resulted from 12 social groups in which was taken into account -- gender, educational level (1-4, 5-8, and 9-11 years of schooling), speak Portuguese only, have lived at least 2/3 of their life in the city, and not having lived outside the region more than one year during language acquisition. At that time, illiterates, university students, and those under 25 years old, were not included.
2. *Amostra Digital VARSUL*: contains digital audio samples between 5-15 minutes of 24 interviews with speakers of the three most Southern capitals of the country, 8 interviews of each Monguilhott and Brescancini databases, representative of non-urban areas of Florianópolis. These serve as a source of information on sociolinguistic varieties and as a source for the informants' sociocultural research and they can be accessed directly from the project site.
3. *Banco de Dados Diacrônico*: it is linked to the *Para a História do Português Brasileiro* National Project (For the History of the Brazilian Portuguese). It is a collection made up of notary public documents, public and private archives, and newspaper articles from the South of the country, representative of the nineteenth and twentieth centuries' Southern writing, with Internet access to some of these documents, since it is in process of development.
4. The *BDSer*[7]*: Banco de Dados de Fala da Serra Gaúcha* - was established in 2000 and is linked to the Language Department of Universidade Caxias do Sul and is made up of research on the spoken language in the mountainous region of Rio Grande do Sul state. The criteria for selection of the informants are: residing in rural or urban area, male or female, pertaining to one of the following age groups: 15-25, 30-45, 50-65, 70+ years, educational level, and higher education. Data was collected by recorded sociolinguistic interviews held in a family environment. The collection also includes their Social Profile and the Record of Networking. The team prepared a list of questions for the interview in order to analyze the language spoken by the informant in a descriptive, narrative, and argumentative manner. They obtained 64 informants for each municipality in the coverage area.

The *ASPA Project*[8] *- Avaliação Sonora do Português Atual* (Current Portuguese Sound Rating) is linked to *LABFON - Laboratório de Fonética da Faculdade de Letras da UFMG* (Phonetics Laboratory of UFMG) - and its objective is to build probabilistic knowledge of the Brazilian-Portuguese structure, contributing specifically to the analysis and mapping of phonological, syllabic, and segmental types of Portuguese. This project finds theoretical support in models of the Phonology of Use and Theory of the Samples, where the language multi-representativeness is concerned. The *ASPA Project* is not yet fully available online because it is currently undergoing a verification procedure and correction of the *corpus*. Interested persons should fill out a form and send by email

their search criteria request to query the *corpus*, which will be returned by e-mail in PDF format. The user can narrow the search as desired, creating a combination of criteria in any registered field:

1. Spelling -- example: words beginning with "*in*", words having "*nha*", words ending in "*ado*".
2. Category -- noun, verb, adjective, adverb, article, conjunction, interjection, preposition, pronoun, numeral.
3. Morphology: verbal inflected, plural inflected, derivative, original.
4. Source: African, Amerindian, neither of the above.
5. Transcript: any segment, sequences of segment, specific syllables, tone index, etc. For example: *oxítona* (stress on the last syllable) words that end with [S], words that have the [di] sequence followed by a fricative consonant, words that have the [bu] syllable.

The *Projeto Norte Vogais*[9] (Northern Vowels Project) follows the objectives of *PROBRAVO*[10-11], the research group *"Descrição Sócio-Histórica das Vogais do Português"* (Socio-Historical Description of the Portuguese Vowels (in Brazil)), created in 2005 and currently involving 21 Brazilian universities. It aims to investigate and describe the phonetic realizations of vowels in dialects from the South to the North of Brazil. This research group seeks to understand:

1. How are the vowels in Brazilian Portuguese phonetically realized?
2. How to explain or what motivates the diversity of phonetic realizations?
3. How do the speakers of Brazilian Portuguese understand each other in spite of differences in vowel attributes?
4. Is it possible to grammatically explain this diversity?

The *Projeto Norte Vogais* seeks to characterize the unstressed vowel system and its variants, based on a stratified sample and in variationism terms, as well as qualitatively analyze and explain the process of variation in pre-stressed mid-vowels and non-final post-stressed ones in the Portuguese spoken in Northern Brazil, conditioned by internal factors. The surveys are conducted using the *corpora* research with speech samples of linguistic varieties of the Portuguese spoken in the *Paraense* Amazon located in the regional Portuguese language zone in the state of Pará. The *Projeto Norte Vogais* corpora have a total of 318 (three hundred and eighteen) native informants of the *Paraense* Amazon from five different localities: Belém, Cametá, Breves, Breu Branco, and Mocajuba, in its rural and urban areas, whose ages vary between 24-72 years. The informants making up the *corpus* were socially stratified by gender, education, age, and origin. The predominant speech type is personal experience narratives. The entire corpus is graphically transcribed and the data that proves the occurrence of the target phenomenon – pre-unstressed middle vowels – phonetically transcribed. In addition to the transcripts, the corpus contains the audio recordings made in fieldwork.

In summary, the proposal of the databases presented is to document and study the spoken norm of the educated in some of the capital cities and regions of Brazil, in their phonological, phonetic, morph-syntactic, lexical, and stylistic aspects, contributing specifically to the analysis and mapping of phonological, syllabic and segmental types of the Portuguese language.

At this point it could be asked: what does the speech language researcher have to do with the spoken norm of the educated or the linguistic aspects that have been gathered? Or rather, what does the spoken norm of the educated have to do with the professional interested in symptomatic speech? How could he/she utilize or extract knowledge from databases for clinical practice? The answer may come from different directions: First, the databases in consideration show that it is possible to store speech data and clearly present the methodology used in the survey and storage of speech and language data, and the requirements for data to be deemed reliable; Second, based on this information it is possible to setup one's own database; Third, the norm can provide the parameters to reflect on what is pathological as well as to find dialogic function in normalcy which, taken as management, can affect the speech of those who come to the Speech Language and Audiology clinic.

Speech and language database

The *Plataforma de Documentos Sonoros do IEL-UNICAMP*[12-13] is an important audio database of language research done within the institute, available in MP3 format. It offers, for example, the *Projeto de Aquisição de Linguagem* (Language

Acquisition Project) collection, coordinated by Dr. Claudia De Lemos, in which the speech of eight Brazilian children ranging in age from 11 months to 5 years were recorded on reel tapes, put into 423 hours of digital format, and transcribed unto 14,000 pages of dialogues, from everyday situations of the children[14] such as lunch or dinner, play moments or storytelling, and in interaction with parents, siblings and friends. Data was collected between 1976 and 1981. The *corpora* have already served as empirical basis for scientific articles, masters dissertations and doctoral theses on child language acquisition and the relationships with the language and the speech of the other, not only in the area of Linguistics but also in Speech Language Pathology and Audiology.

The *Centro de Estudos de Linguagem e Fala - CELF* database is linked to the *Programa de Pós-Graduação em Distúrbios da Comunicação Humana* (Post-graduate Program in Human Communication Disorders), at the *Universidade Federal de Santa Maria - RS*. Since 1997 *CELF* carries out research in the area of Speech Language Pathology and Audiology[15-16], specifically in clinical phonology and acquisition of normal language and deviations. *CELF* database collection consists of material extracted from interviews, evaluations, and speech therapy sessions in the teaching clinic of the university by undergraduate and graduate students. Internet access allows us to find a number of articles, dissertations, and theses that made use of the database. There is, for example, research that carries out a comparative[17] study of different models of phonological therapy and the study on generalization[18] in three models of phonological therapy in children with different degrees of severity of phonological disorder.

Inside *LAFAPE – Laboratório de Fonética e Psicolinguística* (Laboratory of Phonetics and Psycholinguistics), linked to the research branch *COGITES – Cognição, Interação e Significação* (Cognition, Interaction, and Signification) of the *Instituto de Estudos da Linguagem - IEL- Unicamp*, we find *APHASIACERVUS*, which deals with linguistic-interactional data of aphasia, collected at the *Centro de Convivência dos Afásicos* (Community Center for Aphasic People) from 2003 until 2012. CCA is a place for interaction between aphasic and non-aphasic people and its goal is to develop linguistic and neurolinguistic studies and ensures aphasic people therapeutic and social benefits made possible by a wide range of everyday interactional experiences. Once again, the problem in setting up the database was the transcription and interpretation of collected and studied aphasia data since there are two or more interactions between aphasic or non-aphasic individuals. The peculiarities of the verbal and nonverbal aspects of the speakers, such as the symptomatic manifestations of aphasia and multimodal dimension of language and interaction, contribute to the complexity of the database. The material of which *AphasiAcervus* is composed is not yet fully digitized and transcribed, but it has computing devices that capture and handle audiovisual material. Another important point to clarify is that the *APHASIACERVUS* usage and rules criteria have not yet been completed, which restricts access to members of *LAFAPE*. The collection is characterized by:

1. preparation of physical and digital data of weekly *CCA* meetings, recorded in audiovisual format;
2. digitalization of the audiovisual record of weekly CCA meetings;
3. heterogeneous systems of transcription of the video recorded material;
4. records of actions taken *in situ,* recorded by researchers present at *CCA* meetings;
5. description of linguistic/neurolinguistic profile of *CCA* participants;
6. under construction: collection of research already completed (individual and groups) based on *AphasiAcervus corpus*; documentary actions.

The *Project E-LABORE*[19] *– Laboratório Eletrônico de Oralidade e Escrita* (Electronic Laboratory of Orality and Writing) – of the *Faculdade de Letras da Universidade Federal de Minas Gerais (UFMG)* has a collection of essays of children ages 6-12 years who attend private and public schools in Belo Horizonte, Minas Gerais. This laboratory investigates spelling deviations that occur in the child's writing during the literacy process. The material consists of writings of children who received a ruled sheet of paper for a production (writing or drawing) on a topic chosen by the teachers. Participating schools receive a list of 8 instructions and a questionnaire. The identification of each production contains an 8-digit sequence, as though it were a bar code. Two types of information are collected: the first with the name, date of birth, gender, and the child's school grade, and the second with school

data such as address, phone number, school type, and personnel responsible. All written productions of the children are scanned and typed by the *E-LABORE* team and are not yet available to the public. According to the laboratory's creators, this process of typing has many difficulties: the required high number of essays to be statistically representative and the interpretation of the handwriting of children who are still undergoing literacy acquisition. The standardization of data entry followed a set of seven rules:

1. Typing of essays: the texts should be copied so that the digitized version is as similar as possible to the original. For example, line breaks made by the authors are indicated in the data entry.
2. Paragraph marking: texts must be typed respecting paragraph breaks. For example, the typist clicks a sequence of two "enters" to start a new paragraph.
3. Error marking: errors should be copied between braces and the typist must type the correct one in brackets.
4. Difficulty in reading one or more words: the illegible word will be replaced by an asterisk (*).
5. Absence of one or more words: when noting that a word that gives meaning to the child's production is missing, a + sign will be inserted and, in brackets, the word considered to be missing.
6. Start and end of continuous text: the typist should put a $ (dollar sign) to mark the beginning and end of a text.
7. Hyphenation: the typist must copy the hyphenation marks from the child's text.

These databases contain research on language acquisition and on language deviations, in addition to characterizing the writings of children in school. In this case they are databases that narrow the relationship between Linguistics – both in acquisition and in cases of aphasia – and Speech Language Pathology and Audiology. Relationships between studies in Language Acquisition and Speech Language Pathology and Audiology were widely publicized in publications related to the subject. On the other hand, studies on language deviations are directly related to the field because among researchers there are audiologists that, by assembling the database, identify its advantages. The writing database can bring data to the mapping of the child's pathway in the writing acquisition process. It also allows us to discuss the relationship between literacy and alphabetizing. These are just some of the uses and alternatives to Speech Language Pathology and Audiology in its dialogue with speech, language, and writing databases.

The Speech and Writing Database

The Speech and Writing database is an example of a specific Speech Language Pathology and Audiology database that uses an organized collection of audio-visual-graphic records from speakers and writers. This is a digital database, available on the Internet through a specific site, with free access to users interested in researching oral and written language acquisition and pathologies from the perspective of any theoretical and methodological model. The database is linked to the research of *Linguagem e Subjetividade*[20] *of PEPG em Fonoaudiologia da PUC/SP* (Language and Subjectivity of the Speech Language Pathology and Audiology Department of PUC/SP University).

Although its foundation occurred in 2001, it had begun many years earlier, in the late 1970s, when a speech therapist and researcher wanted to investigate language acquisition from the perspective of mother-child interaction, using as research material recordings and transcripts of dialogues in playing situations of a child in the process of language acquisition and his/her mother. This research investigated language acquisition from a longitudinal study design, focusing on the dialogic processes taking place in the interactions between mother and child.

While administering the Language Acquisition course and performing clinical supervision in the 1980s in an educational institution in the city of São Paulo (which served children aged 0-5 years and 11 months), this researcher, along with her Speech Language Pathology and Audiology undergraduate students, decided to use a practical and investigative strategy of researching the acquisition of language by institutionalized and non-institutionalized children by recording situations of dialogic interactions of children from different age groups. The study design used a crosscut as representative. In a similar way, in the 1990s the researcher investigated the language acquisition process of two of the institutionalized children and another one who she was seeing in her office, who presented with complaints of language delay and a medical diagnosis of cleft palate. This work formed the

basis for her doctoral research, later published as a book. In the following decades this type of research and data collection in partnership with Speech Language Pathology and Audiology students was continued. They started to include a few clinical cases of subjects with language symptoms that were seen by the researcher and therapist, because they had, for example, language delay due to deafness, specific language disorder, dyslalia, stuttering, voice disorders, writing acquisition delay, learning disorders, and aphasia. This type of data collection for investigation of language functionality lasted until 2004, when it was necessary to stop collecting data in order to organize the material, digitize it, and format it in a standardized manner to make it available to web users. It should be noted that the database was approved by the Ethics Committee of the *Pontifícia Universidade Católica de São Paulo* with authorization number 202/2009.

The *Banco de Dados de Fala e Escrita* (Speech and Writing Database) aims to provide: (i) assistance for the description of speech and writing acquisition process; (ii) conditions for analysis and development of speech language pathology and audiology theories related to the speaker; (iii) conditions for training of new researchers through teaching techniques of observation, analysis and understanding of normal and pathological functioning of language; (iv) assistance to educational programs promoting knowledge of the language, speech, and writing acquisition process in different age groups and different linguistic environments; (v) assistance to research of the function of language through clinical and non-clinical perspectives; (vi) assistance to research on therapy, in particular in the treatment of individuals with oral or written language problems; (viii) means of publicizing knowledge generated from speech, language, and writing research stored in the database.

Currently this database has data on 321 subjects, totaling about 641 *corpora*. There is data of mother-child, child-child, therapist-patient, and therapist-family interactions and interactions between adults. Access to the corpus chosen for analysis is mediated by search engines that allow you to find: cross-section and/or longitudinal collections, corpora of children or adults, from both sexes, in dialogic situations of the following types: dyadic, triadic and polyadic, playful and therapeutic. The *corpora* are digitalized into text, audio, and video files, and also refer to the collections of language acquisition and/or symptomatic language. In this last case the collections are organized by clinical procedures: interview, assessment, care, and orientations. The *corpora* were structured based on the rules of Corpus Linguistics, that is, the collected discursive material meets the criteria to ensure consistency and legitimacy to the material, such as: origin, purpose, composition, formatting, representativeness and extent of data. The corpora were transcribed according to the rules of the *Projeto de Estudo da Norma Lingüística Urbana Culta de São Paulo* [21] (Project for the Study of the Linguistic Norm of the Educated of São Paulo).

To access the corpus collection it is only necessary that the researcher complete the registration form on the platform: www.bancofalaescrita.org, and wait for the release to use it. After analyzing the registration, access can be released by the database administrators.

Several researches were and are being carried out using the data provided by the database. Here are some examples of authors (2007) [22], (2011) [23], and (2016) [24].

Compared to others, an advantage of this database is its neutrality, or the fact that the data is collected and the choice of the path desired to analyze the data is left up to the researcher. It is noteworthy that in addition to this advantage, there is the fact that the data, being of clinical and non-clinical nature, and longitudinal as well transverse, allows for the monitoring of a speech event over a period of time or its reappearance in several children at the same time of the process of language acquisition. Another very important advantage is the fact that access to the data is direct and open to those who register. For security reasons, the video data can be accessed at the institution, from a specific local software.

Conclusion

Databases in general and the *Banco de Dados de Fala e Escrita* in particular, are an initiative for scientific researches to go beyond the printed media and extend their reach. Previous studies with the databases presented collaborate to explain the structure and operation of speech, language, and writing. The desire for expanding the transmissibility of data comes from the importance of allocating to an increasing number of people new sources and research results. To accomplish such a transmission of

information, the Internet is an indispensable means as it enables the dissemination of data and permits the sharing of information. We may conclude that the dialogue with the speech language pathology and audiology community and the dissemination of the use of *corpora* in the databases presented will contribute to the development of clinical and scientific strategies for the elaboration and management of methods and techniques to be used both in the clinical practice as well as in research. This points to the importance of the databases to Speech Language Pathology and Audiology, as well as to other areas interested in language acquisition and in its symptoms in speech and writing.

References

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.