PhonBank and Data Sharing: Recent Developments in European Portuguese

Ana Margarida Ramalho*, M. João Freitas*, Yvan Rose[+]
*University of Lisbon / CLUL; [+]Memorial University
*Lisboa, Portugal; [+]St. John's, NL Canada
mramalho@letras.ulisboa.pt, joaofreitas@letras.ulisboa.pt, yrose@mun.ca

Abstract

This paper presents the recently published RAMALHO-EP and PHONODIS corpora. Both include European Portuguese production data from Portuguese children with typical (RAMALHO-EP) and protracted (PHONODIS) phonological development. The data in the two corpora were collected using the phonological assessment tool CLCP-EP, developed in the context of the *Crosslinguistic Child Phonology Project,* coordinated by Barbara Bernhardt and Joe Stemberger (University of British Columbia (UBC), Canada). Both corpora are part of the PhonBank *Project* (Brian MacWhinney (Carnegie Mellon, USA) and Yvan Rose (Memorial University of Newfoundland, Canada), which is the child phonology component of TalkBank, coordinated by Brian MacWhinney. The data at PhonBank is edited in Phon format, a language tool designed and built by Yvan Rose and Greg Hedlund (Memorial University of Newfoundland) and widely used by researchers working in the field of phonological acquisition. RAMALHO-EP contains production data from 87 typically developing children, aged 2;11 to 6;04, all monolinguals. PHONODIS includes production data from 22 children diagnosed with different types of speech and language disorders, all EP monolinguals, aged 3;2 to 11,05. Both corpora are open access language resources and contribute to enlarge the amount of production data on the acquisition of European Portuguese available in PhonBank.

Keywords: Phonology, Phonetics, Acquisition, Phon, PhonBank, Data sharing, European Portuguese

1. Introduction

Big Data research enables computer-assisted, broad-based generalizations over rich datasets, which cannot be obtained through traditional methods of observation. However, these generalizations often come with partially or completely untested assumptions about the irrelevance of potential noise in the data, the nature of which may influence results in ways which remain elusive to the researcher. In light of this, and also given the fact that some types of Big Data corpora are not easily obtainable, due to the nature of the data, the need for precisely-analyzable corpora remains across many areas of research.

This is the case for the study of phonological development and speech disorders, two disciplines which require access to phonetically-transcribed sets of data documenting different populations of speakers (e.g. typically-developing vs. disordered) across different languages and, in the context of multilingual populations, across different combinations of languages. Current systems for the automatic transcription of speech data are not reliable enough to be used as a basis for corpus building at the moment, especially given that child, disordered, or accented speech are typically transcribed very poorly through these systems. This implies that corpus building in these areas must rely on human transcribers, whose work can be assisted through computer-assisted methods of data measurement (e.g. acoustic analysis) or data annotation (see below). This requirement in turn poses a major roadblock for the development of Big Data corpora for research.

The PhonBank project, which supports researchers and students through software programs and a data sharing platform[1] has been offering compelling solutions to these two challenges, through a combination of technical innovations in the areas of phonological and phonetic data representation and annotation, assembled within a software package called Phon[2] and the building of a public (web-accessible) database documenting phonological development and speech disorders across a number of different languages and speaker populations. A key feature of PhonBank is that it relies on the commitment of the users of Phon toward data sharing in order to construct the database component of this project: As scholars of phonological development and speech disorders benefit from the functionality available through Phon for their research, their publication works and, in particular, the contribution of their corpus data to the PhonBank database enable its construction in ways that make Big Data research possible in the long term.

In this paper, we present an overview of this project and highlight how it offered a framework for corpus building and analysis based on a series of projects recently developed based on European Portuguese (EP) speech data. We begin by situating PhonBank within the larger context of corpus-based research on language development. We then highlight the main components of Phon for corpus building and analysis. This provides the technical basis for our discussion of how these resources have been used for research on European Portuguese, which we highlight through some of the key findings emanating from this research. Finally, we highlight how these results have now been incorporated within PhonBank in ways which contribute to cross-linguistic, Big Data research.

[1] https://phonbank.talkbank.org
[2] https://www.phon.ca

2. CHILDES, TalkBank, PhonBank

PhonBank comes from a rich tradition of corpus-based research in language acquisition, which itself has its roots all the way to Charles Darwin's (1877) description of gestural development by his son. This work provided a model application of Darwin's descriptive techniques for work on natural selection and evolution to the study of human development (MacWhinney, 2000). Limited by the means of research available to individuals until the democratization of computers in the early 1980s, corpus-based research on language acquisition remained central to virtually all studies of language acquisition since Darwin, however in ways that were limiting both the level of annotation obtained and the analysis of these annotations.

The Child Language Exchange System (CHILDES[3], co-founded by Brian MacWhinney and Catherine Snow in 1983, opened the door for the building of the first large-scale database documenting child language. To this day, CHILDES remains the primary source of data and research tools for the study of child language worldwide, through its combination of publicly-accessible datasets documenting child language across different languages and learning situations, and computer programs for the analysis of these datasets, assembled within the CLAN software package.

CHILDES has since provided a model (and inspiration) for the creation of similar databases, for the study of other topics relevant to human language and communication (e.g. AphasiaBank, BilingBank, FluencyBank) which together form the TalkBank consortium[4]. In addition to the rich documentation it provides to scholars of language, language acquisition, pathologies, etc., one of the most central assets of the TalkBank databases is their adherence to relatively strict data formats, in particular the CHAT format supported by CLAN and its corresponding. This cross-corpus consistency greatly facilitates the large-scale study of the different datasets.

PhonBank, came into existence first as a subset of the CHILDES database; through successive rounds of funding from the National Institutes of Health, PhonBank subsequently grew into becoming an additional sister of CHILDES within the larger TalkBank consortium (Rose and MacWhinney, 2014). There are two key differences between PhonBank and CHILDES. The first is that PhonBank is not confined to the study of child language and developmental disorders; all research areas that relate to phonetics and phonology are welcome to this database (e.g. second language acquisition or disordered data produced by adults). The second difference is the primarily focus of PhonBank, which is on phonetics and phonology (see further below), two areas of investigation which have traditionally not been supported through CHILDES or other TalkBank projects. This lack of support is what has motivated the creation of specialized tools for the study of phonology and phonological development, for example in the area of segmental and syllabic patterns of behaviours (e.g. phones and phonological features; syllable onsets vs. codas, stress, tone, and so on). In addition to this, at the time of PhonBank inception, in 2006, recent progress in phonetic science was making acoustic analyses of child speech increasing compelling for our understanding of phonological development. In order to address this need, Praat libraries for the acoustic analysis speech[5] were subsequently integrated within Phon. We provide more detail on the functions supported by Phon in the next section.

3. Phon

Phon, an open-source (free) software program, was first released in 2006 (Rose et al., 2006). Among other functions, Phon supports the following:

- Media (audio/video) time-alignment to data transcripts
- Orthographic and phonetic transcription
- Segmental, tonal and stress-level annotation of the phonetic transcriptions
- Segment-by-segment alignment between target (model) and actual (speaker-produced) forms
- Data query and reporting in reference to segmental and prosodic aspects of the transcribed forms
- Acoustic analysis

These general functions are supplemented by a series of additional features within Phon which help the creation and maintenance of structured databases, including format checks and the ability to import and export data from and to third-party applications.

After the research has diarized the corpus into speaker-specific utterances (which can range from isolated word forms to full utterances, each associated with a given speaker) and completed the orthographic transcription of the corpus (a task that remains essentially manual in the absence, still today, of reliable and/or easily accessible speech recognition systems), Phon can automatically generate broad phonetic transcriptions (following the standards of the International Phonetic Association[6] using either built-in dictionaries IPA-transcribed forms (current support includes Catalan, Cantonese, Dutch, English, French, German, Icelandic, Italian, Mandarin, Spanish, Portuguese, also with support for major dialects of English and Spanish) or transliteration systems (e.g. Arabic, Ewe, Slovak, Turkish). These broad transcriptions can in turn provide a basis for the transcription of actual speech forms, through modification (by human transcribers) of the standard forms provided by Phon, which greatly speed up this otherwise time-consuming task (McAllister Byun and Rose, 2016).

After each transcription is completed, Phon automatically labels the transcribed forms to identify syllables and syllable positions (e.g. syllable onsets vs.

[3] https://childes.talkbank.org

[4] https://talkbank.org

[5] http://www.fon.hum.uva.nl/praat/

[6] https://www.internationalphoneticassociation.org/

codas). Each time a researcher wants to compare produced forms against standard versions of these forms (e.g. train /tɹaɪn/ produced as [dwaɪn]), Phon automatically performs a phone-by-phone alignment through which we can extract accuracy measures, or describe patterns of segmental substitution, deletion, or epenthesis. In all cases (IPA generation, syllabification labelling, and phone alignment), the forms and annotations generated by Phon are fully modifiable by the user, who retains full control over corpus building and subsequent analysis. Finally, each phone and diacritic present in the IPA transcriptions is assigned descriptive phonological features, which enables research on particular phones and phone classes. Finally, Phon integrates seamlessly with Praat functions for acoustic data analysis; TextGrid files and annotations can be imported from Praat and serve as a basis for the generation of corresponding data records in Phon, or be generated directly from within Phon. After the TextGrids are aligned, the record data and associated TextGrid annotations can be used in tandem to incorporate as many linguistic criteria as needed into the queries (e.g. utterance-, phrase-, word-, syllable-, and phone-level information, including phonological features), as required by the question at hand. In this sense, Phon supplements Praat in the areas of phonological data annotations and query as well as in database management more generally. Data extraction relies on different types of queries (e.g. textual, phonological, or based on regular expressions) and specialized analyses (e.g. the *Percentage of Consonants Correct*; Shriberg et al., 1997), with each query and analysis result visualized as part of structured reports or exported to third-party applications (e.g. spreadsheets; statistical analysis packages) for further processing.

Phon thus offers tremendous database building and analytic support toward virtually all types of research in the areas of phonology and (acoustic) phonetics. The existence of this tool, whose development has been relying primarily on public funding from the NIH, as mentioned already, is ultimately contingent on the building of PhonBank which, like CHILDES and other databases within TalkBank, aims to contribute the broad-based, reliable foundation for large-scale studies of language development and beyond. One direct implication of this is that without data sharing as foundational to the public database building mission it supports, access to the various functions and resources overviewed above would be compromised.

Thankfully, researchers and students from around the world have been generous in their support for PhonBank; the database has been growing steadily since its inception, both in the number of case and cross-sectional studies it offers and in the number of language learning contexts documented by these studies. Each of these studies offers additional data which, through their standardized annotations and formats, contribute as many elements toward our ability to perform reliable Big Data research in phonological and phonetic development, speech disorders, and beyond.

In the next section, we present recent examples of such contributions, coming from the study of phonological development and speech disorders in European Portuguese. As we will see, this work illustrates the tandem expressed above between Phon-assisted research and PhonBank building: as the scholars involved in this research have benefited from specialized software tools to conduct their research, the databases constructed for this research are now available to all students and researchers worldwide for additional research on European Portuguese and beyond.

4. Phonological Development and Speech Disorders in European Portuguese: new corpora in PhonBank

Over the last four decades, a considerable amount of research outputs on phonological acquisition has been published (see Kager et al., 2004; Fikkert, 2007; Demuth, 2009 for state of the art). More recently, the interest on protracted phonological development has increased substantially (Bishop and Leonard, 2000; Ball et al., 2008; Dinnsen and Gierut, 2008). Research on phonological acquisition has been mostly performed on the basis of production and perception data sets often not publicly-accessible. The PhonBank project was crucial to change this scenario by promoting data sharing worldwide, thus contributing to develop research in the field of phonological acquisition from a crosslinguistic perspective.

Another project that provided tools to enhance this crosslinguistic approach was the *CrossLinguistic Child Phonology Project* (CLCP Project)[7]. It gathers researchers (linguists and speech and language therapists (SLTs)) from several countries whose main interest is phonological acquisition. Under this project, researchers develop tools to assess the children's phonological knowledge in different languages, all theoretically framed in the nonlinear phonology approach (Bernhardt and Stemberger, 1998, 2000). These tools enable researchers to perform comparative studies based on results from different languages since data collection and analysis share the same methodological procedures.

The two new corpora presented in this paper profit from the two projects: PhonBank and CLCP. RAMALHO-EP (Ramalho, 2017, 2019[8]) and PHONODIS (Freitas, Ramalho, Lousada, Oliveira and Pereira, 2019[9]) are both available at PhonBank and provide production data from monolingual Portuguese children, either with a typical phonological profile, in the former, or with a protracted phonological development, in the latter. The

[7] http://phonodevelopment.sites.olt.ubc.ca; coordinators: Barbara Bernhard and Joe Stemberger, University of British Columbia, Canada.

[8] https://phonbank.talkbank.org/access/Romance/Portuguese/Ramalho.html

[9] https://phonbank.talkbank.org/access/Clinical/PhonoDis.html

data in both corpora were collected using the assessment tool CLCP-EP, developed under the CLCP Project (Ramalho, 2017; Ramalho et al., 2014[10]). Phon was used to edit the data. The edition of these corpora in an open access format follows the goal underlying the edition of other corpora on EP already available in PhonBank: to make web-acessible as much data as possible on the acquisition of EP phonology.

In the sections below, we will briefly present the CLCP-EP test and provide data from RAMALHO-EP and PHONODIS on the acquisition of EP branching onsets, a syllable structure commonly reported as problematic both in typically developing children (Fikkert, 1994, 2007; Gnanadesikan, 1995; Freitas, 1997; Pater and Barlow, 2003; Bernhardt and Stemberger, 2018; among many others; see Rose, 2000 and Almeida, 2011 for different results in French) and in children with a protracted phonological profile (Marshall et al., 2002; Gallon et al., 2007; Tamburelli and Jones, 2013; Bernhardt and Stemberger, 2018; among others; see Ferré et al., 2015 for different results in French). The main goal is to show that the CLCP-EP is able to discriminate age groups and different language type groups (typical *versus* atypical development); no discussion framed on the phonological aspects of our results will be performed.

4.1. The CLCP-EP Assessment Test

In the Portuguese community of SLTs, as in many other countries, most tools designed to assess phonological development use exclusively the *segment*, which may be synonymous of a phoneme or a speech sound, depending on the characteristics of the test. These tools do not enable SLTs to explore possible correlations between segments, their internal structure (features) and their prosodic distribution. Despite the long use of other constituents to perform descriptions on the phonology of many linguistic systems worldwide, language assessment tools in the clinical practice tend not to make use of these constituents. The CLCP project was created to reverse this tendency, by adopting the nonlinear phonology approach (Bernhardt and Stemberger, 1998, 2000, 2008). The *Crosslinguistic Child Phonology – European Portuguese* test (CLCP-EP), thus follows this theoretical approach, as shown below.

Under the nonlinear phonology framework (see Nespor and Vogel 1986), the knowledge of the sound structure of languages is represented in terms of different tiers of phonological information, corresponding to different hierarchically organized segmental and prosodic constituents (such as the *feature*, the *segment*, the *syllable*, the *foot* and the *prosodic word* - the most studied in language acquisition -, along with the *clitic group*, the *phonological phrase* and the *intonational phrase*). Based on multiple descriptions of the children's phonological development in different languages over the last decades, the CLCP project assumes that such a model accurately accounts for child phonology. These constituents were shown to be productive in the description of the children's phonological systems and in the establishment of correlations between their phonological knowledge and their language profiles.

The EP lexical stimuli included in the CLCP-EP were selected on the basis of a subset of constituents assumed within the CLCP project: the *feature*, the *segment*, the *syllable*, the *foot*, and the *prosodic word.* The phonological variables selected in accordance with the nonlinear phonological framework were: i) the EP *consonantal inventory*, viewed in terms of an exhaustive representation of all place of articulation, manner of articulation and voicing contrasts (all segments and their prosodic distribution); ii) the *syllable constituency* (all syllable constituents); iii) the *word stress* (target structures tested in stressed and unstressed positions); iv) the *position within the word* (structures tested in initial, medial and final positions); v) the *word length* (measured in number of syllables).

CLCP-EP, as all assessment tools within the CLCP Project, is a naming test based on the story of a rabbit living in a family of humans. Different scenarios were built to create semantic diversity (the school, a visit to the dentist, a party, a visit to the zoo, among others). This type of elicitation has the advantage of triggering either a word naming or a story telling task. The test includes 150 words elicited on the basis of 42 pictures (grouped by semantic network contexts). The tool was tested with a sample of 87 typically developing monolingual Portuguese children (organized by age groups: 3;0-4;0, 4;0-5;0, 5;0-6;06) and a sample of 3 monolingual Portuguese children with phonological disorders. The data were transcribed within the PHON software (Ramalho, 2017).

To validate the CLCP-EP, procedures and norms required in the literature on language assessment tests were considered. Both linguistic (phonological structures, lexical knowledge, semantic content of the items) and non-linguistic criteria (characteristics of the drawings) were taken into account. For this purpose, the *Classic Theory of Tests* (Brown, 1976; Almeida and Freire, 2003) was used as the support to the CLCP-EP construction and validation; *item difficulty index*, *validity* and *reliability* measures were considered. A statistical analysis was performed to check validity (*content* and *criterion*) and reliability (*internal consistency*, *inter-reliability*, *intra-reliability*). Table 1 characterizes the psychometric measures of the CLCP-EP (Ramalho, 2017):

[10] available at www.clul.ulisboa.pt/pt/24-recursos/851-clcp-pe-crosslinguistic-child-phonology-project-portugues-europeu, and http://phonodevelopment.sites.olt.ubc.ca/practice-units/portuguese-european/

Reliability	**Internal Consistency**	α= 0.98
	Inter-reliability	98% agreement (2nd transcriber analysed 10% of the sample)
	Intra-reliability	97.9% agreement
Validity	**Construct Validity**	Age differentiation (statistically differences showed in at p<0.001)
	Content Validity	**Nonlinear phonology content** validated by 2 field expert judges; **Content Validity Index** applied to an expert panel (IVC=0.99) assured to: phonological variables, lexical items, visual stimuli, and instruction protocol. **Phonological Content Validity u**sed for EP: **Phonological Index of Difficulty** (*word match measure)* with statistical differences in all age groups (p<0.001).

Table 1: Psychometric measures for the CLCP-EP (Ramalho, 2017).

4.2. New EP corpora in PhonBank

We will now present the two new corpora on EP recently edited in the PhonBank: RAMALHO-EP and PHONODIS. The corpora contain production data from typically developing monolingual Portuguese children, in the case of RAMALHO-EP, and from monolingual Portuguese children with protracted development, in the case of PHONODIS.

The data collection procedures for RAMALHO-EP and PHONODIS were identical: i) informed consents were gathered according to the CLCP project guidelines; ii) the test used for data collection was the CLCP-EP: a sequence of 42 digital images were presented to the child on an IPAD screen, in a picture naming format; iv) audio recording was performed during the session; v) the recordings were made at each child's school, in individual sessions ; vi) data collection and child anamnesis was performed by an SLT. Phon was used to perform orthographic and phonetic transcriptions, and to implement the procedures to validate the test in terms of its phonological characteristics (Ramalho, 2019).

4.2.1. RAMALHO-EP corpus

The RAMALHO-EP corpus consists of experimental cross-sectional data produced by 87 typically developing children from the Lisbon area, all EP monolinguals, aged 2;11 to 6;04. In the validation procedure carried out by Ramalho (2017), the children were organized by age groups: Group 1 (G1) gathers children aged 2;11 to 4;0; Group 2 (G2) gathers children aged 4;0 to 5;0; Group 3 (G3) gathers children aged 5;0 to 6;04. All children were assessed by an SLT (the first author in this paper) and the following assessment tests and procedures were used to confirm no history of hearing or language deficits or disorders: each child's clinical history was checked ; standardized EP comprehension and production language tests were used (TALC: *Teste de Avaliação da Linguagem na Criança* (Sua-Kay and Tavares, 2007) - within 1.5 SD) and oral-motor structure and function were assessed.

Initially, the audio files were all transcribed by the first author in this paper and revised by the second author (see inter-reliability and intra-reliability at table 1). The data in the corpus were recently (in 2019) revised by a MA student in Linguistics, highly trained in phonetic transcription; after this revision, detailed phonetic transcription for target liquids was displayed.

The RAMALHO-EP corpus is now available at PhonBank[11]. Although, the phonetic transcriptions and audio files are available only for children whose parents had consent it, all audio files are available for investigation purposes and can be accessed by contacting the author[12].

4.2.2. PHONODIS corpus

To our knowledge, PHONODIS is the first corpus on atypical phonological development in EP to become available in a publicly-accessible format. The main purpose of this corpus is to provide production data on protracted phonological development in EP, to be used in academic and in clinical contexts. The sample does not follow any predetermined criteria in terms of type of diagnosis or sociolinguistic variables; the data are to be used according to the goals of each researcher or SLT. New data will be edited in the corpus in a near future.

PHONODIS contains experimental cross-sectional data collected by each child's SLT (Marisa Lousada, Patrícia Oliveira, Margarida Ramalho). Currently, the corpus contains production data from 22 Portuguese children diagnosed with different types of speech and language disorders. They are all EP monolinguals, aged 3;2 to 11,05. The children are from 3 different districts in the country (Aveiro, Évora, Leiria); for further detail, see table 2 below:

Code (gender: M/F)	Age	Diagnosis
A (M)	11;02	Secondary language disorder with mostly morphosyntactic and phonological compromises associated to an intellectual developmental disorder.
B (F)	11;05	Secondary language disorder; phonetic and phonological disorders associated to Down's Syndrome
C (F)	11;00	Secondary language disorder with mostly morphosyntactic and phonological compromises associated to an intellectual developmental disorder.

[11] https://phonbank.talkbank.org/access/Romance/Portuguese/Ramalho.html

[12] Email: mramalho@letras.ulisboa.pt

D (M)	4;11	Developmental Language Disorder (DLD) with a notorious disruption of the morphosyntactic and phonological domains
E (F)	5;04	Secondary language disorder and a phonological disorder associated to an intellectual developmental disorder
F (M)	8;03	DLD with a notorious disruption of the morphosyntactic and phonological domains.
G (M)	7;09	DLD with a notorious disruption of the pragmatic and phonological domains; phonological impairment
H (F)	7;04	Secondary language disorder and a phonetic-phonological impairment associated to an intellectual developmental disorder
I (F)	6;10	Developmental Verbal Dyspraxia (DVD)
J (F)	6;00	Secondary language disorder and a phonological impairment associated to an intellectual developmental disorder
K (F)	5;09	DVD
L (F)	7;06	DLD
M (M)	4;00	DVD
N (M)	4;00	DLD
O (M)	4;09	DLD with phonological and morphosyntactic impairments
P (M)	14	Secondary language disorder; phonetic and phonological disorders associated to Down's Syndrome
Q (M)	3;02	DVD
R (M)	5;11	DLD
S (M)	05;02	Speech Sounds Disorder (SSD)
T (M)	5;03	SSD
U (M)	05;05	SSD
V (M)	6;00	Central auditory processing disorder; under study, possible DVD.

Table 2 - PHONODIS corpus: Age and information on diagnosis

The data collection followed the procedures described for the RAMALHO-EP corpus: i) Informed consents were gathered according to the CLCP project guidelines; ii) the test used to elicitate production was the CLCP-EP: a sequence of 42 digital images were presented to each child individually on an IPAD or a computer screen, in a picture naming format; iv) audio recording was performed; v) the recordings were made in the therapy session by the child's SLT.

The data were transcribed by a linguist highly trained in phonetic transcription and revised by another linguist and two SLTs involved in the data collection. Phon was used to perform the data transcription and edition. The PHONODIS corpus is now available at PhonBank (https://phonbank.talkbank.org/access/Clinical/PhonoDis.html)

4.3. Some results on branching onsets in RAMALHO-EP and in PHONODIS

In order to briefly illustrate the use of the two corpora in the description of phonological acquisition, we will focus on branching onsets. This syllable structure has been reported in the literature as problematic for many children acquiring different phonological systems, both with typical and atypical profiles (see references in section 4 above). This was also attested in EP (see Freitas, 2003; Almeida et al., 2015; Amorim, 2014; Ramalho, 2017 for typically developing children).

We will first focus on the data in RAMALHO-EP, and test the CLCP-EP ability to discriminate age groups. We will then compare the data in both corpora to test the CLCP-EP ability to discriminate language profiles: typically developing children (RAMALHO-EP) and children with phonological disorders (PHONODIS). For data extraction, the search tools available in Phon were used.

Branching onsets in EP (Mateus and Andrade, 2000) are syllable-initial consonant clusters consisting of an *obstruent+liquid*; table 3 displays EP examples for all possible segmental combinations in onset domain; the items were extracted from the CLCP-EP list of lexical stimuli (except for the examples illustrating *tl* and *gl*, not tested in the tool for lack of EP lexical items that match the children's lexicon knowledge):

plosive+ rothic	*plosive+ lateral*	*fricative+ rhotic*	*fricative+ lateral*
[pɾ]etas *"black"*	[pl]anta *"plant"*	[fɾ]ango *"chicken"*	[fl]or *"flower"*
[bɾ]aços *"arms"*	bi[bl]ioteca *"library"*	li[vɾ]os *"books"*	
[tɾ]ês *"three"*	a[tl]eta "athlete"		
[dɾ]agão *"dragon"*			
[kɾ]eme *"cream"*	[kl]ube *"club"*		
[gɾ]andes *"big - pl"*	[gl]obo *"globe"*		

Table 3: Types of Branching onsets in EP.

In order to describe the acquisition of EP branching onsets in the RAMALHO-EP corpus, the children were organized into 3 age groups: G1 = 2;11 – 4;0; G2 = 4;0 – 5;0; G3 = 5;0 – 6;04. Statistical data for success rates on the production of branching onsets per age group are presented below (percentages in Ramalho 2017: 217):

Age Group	***N***	***Success Rate (mean ± standard deviation)***	***Error in standard deviation***
G1	28	17,2 ± 23,2	4,39
G2	30	36,0 ± 25	4,57
G3	29	48,5 ± 18,6	3,45

Table 4: Statistical data for EP branching onsets (Ramalho 2017: 217)

We may observe in table 4 that success rates are low for all age groups, showing that this structure is highly problematic for Portuguese children. The Kruskall-Wallis test was used to identify statistical significance; the result (H=19,820) corresponds to p≤0,001, which reveals statistical significance for all age groups, when branching onsets are taken into account. These results show the CLCP-EP ability to discriminate age groups on the basis of this specific target syllable structure (for results on other structures, see Ramalho, 2017).

If we look into de data considering exclusively the success rate for the second member in the cluster (/l/ or /ɾ/), traditionally considered as the most problematic consonant in the focused structure (Fikkert, 1994; Goad and Rose, 2004), an asymmetry emerges; see table 5 for success rates (percentages) per age group:

Age Group	***/l/***	***/ɾ/***
G1	8,3	21,6
G2	24,7	50,1
G3	35,9	68,7

Table 5: /l, ɾ/ in branching onsets (Ramalho 2017: 230)

The data gathered in table 5 show higher rates for branching onsets with /ɾ/, when compared to /l/. Clusters with /l/ showed up as extremely problematic for the children assessed. Different behaviours are displayed in different age groups. The low success rates, when compared with other studies in EP, may be due to the type of stimuli in the test: several polysyllabic words with branching onsets were included in the test, which is not the case in other studies for EP (Mendes et al., 2009 does not include polysyllabic words with branching onset; Amorim, 2014 only includes one polysyllabic word with branching onset; for the discussion of these results, see Ramalho, 2017).

In order to test the CLCP-EP assessment test in the discrimination of typically *versus* atypically developing Portuguese children, we used RAMALHO-EP and PHONODIS. Due to the age of children with protracted phonological development in this study, we compared only the typically developing children from G3 in Ramalho (2017), aged 5;0 to 6;04, with the children with an atypical profile aged 5;0 on. Since we are focused on phonological processing, only children with phonological disorders were considered (children B and P in PHONODIS (see table 2), diagnosed with Down's Syndrome, show also a phonetic disorder; for this reason, they were excluded from this analysis). The graphic 1 represents the success rates (percentages) for branching onsets produced by children in the oldest age group in RAMALHO-EP (G3 = 5,0 – 6;04) and by children with phonological disorders in PHONODIS aged 5;0 on:

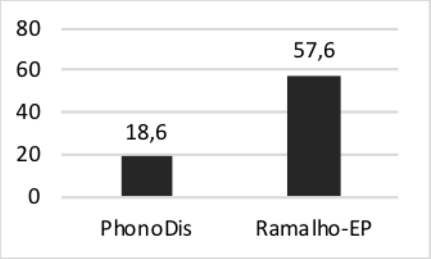

Figure 1: success rates for branching onsets in RAMALHO-EP and in PHONODIS.

The data reveals a clear contrast: the typically developing children showed a mean success rate of 57,6%, while children with a protracted phonological profile showed an extremely low rate of target-like structures (18,6%). The results may be used as empirical evidence to propose branching onsets as a potential clinical marker in Portuguese children, to be tested in samples organized by types of diagnosis.

The high complexity of branching onsets follows the tendency attested for other languages (Fikkert, 1994 for Dutch; Gnanadesikan, 1995 and Barlow, 2007 for English; Freitas, 2003 for Portuguese; Marshall et al., 2002; Marshall and van der Lely, 2009) but contrasts with reports on French (Rose, 2000; Almeida, 2011; Ferré et al., 2015), where codas, and not branching onsets, are the most problematic syllable structures.

The results disseminated in the literature are crucial to test and set different clinical markers for different language pathologies according to the children's language profiles. Publicly-accessible corpora on typical and atypical phonological development are therefore, directly or indirectly, of extreme relevance to help SLTs in the task of performing accurate diagnoses in their clinical practice.

5. Conclusion and further work

We have presented RAMALHO-EP and PHONODIS, two corpora recently edited in PhonBank. RAMALHO-EP contains production data from typically developing Portuguese children. PHONODIS gathers data produced by Portuguese children with atypical phonological development. Both corpora display experimental cross-sectional data elicited by using the CLCP-EP, a phonological assessment test developed under the *CrossLinguistic Child Phonology Project.*

In order to illustrate the use of both corpora in the study of phonological development in EP, we briefly focused on branching onsets. First, we described the data in RAMALHO-EP, organized by age groups, and showed significant contrasts by age group, which argues for the validity of the CLCP-EP as an assessment tool able to discriminate samples by age intervals. Then we focused on the same target structure and compared the data in RAMALHO-EP with those in PHONODIS. Again, results showed that the CLCP-EP is able to discriminate groups of children based on their language profile (typical *versus* atypical development).

As future work, we intend to expand both RAMALHO-EP and PHONODIS corpora by adding new production data elicited on the basis of the CLCP-EP assessment tool. This will provide empirical data to continue the validation procedure of this tool, which may be used in clinical settings for assessing the children's language systems in a phonological detailed mode.

6. Acknowledgements

We are deeply grateful to Brian MacWhinney (Carnegie Mellon, USA) for hosting PhonBank within the Talkbank Project, following his tradition to promote worldwide open access to data on child language. We are also deeply grateful to Barbara Bernhardt and Joe Stemberger for having invited us to join the *Crosslinguistic Child Phonology Project,* hosted at the University of British Columbia (Canada). Last but not least, we express our gratitude to Rodrigo Pereira, who transcribed all the data in PHONODIS, and to Marisa Lousada and Patrícia Oliveira, who collected a substantial part of the data included in PHONODIS. Finally, we thank all the parents and children who accepted to collaborate in our data collection.

The work reported in this paper was partially supported by national funds through FCT – Fundação para a Ciência e a Tecnologia (CLUL funding: PEst-OE/LIN/UI0214/2013; UID/LIN/00214/2013; individual PhD research grant SFRH/BD/88966/2012). The research was indirectly supported by grant RO1-HD051698 from NIH-NICHHD to Brian MacWhinney and Yvan Rose, and by the Social Sciences and Humanities Research Council of Canada to Barbara Bernhardt and Joe Stemberger, coordinators of the *CLCP Project*, University of British Columbia.

7. Bibliographical References

Almeida, L. (2011). Acquisition de la Structure Syllabique en contexte de Bilinguisme Simultané Portugais-Français. Tese de Doutoramento, Universidade de Lisboa.

Almeida, L., Freire, T. (2003). Metodologia da investigação científica em Psicologia e Educação. Braga: Psiquilíbrios.

Almeida, L., Y. Rose & M. J. Freitas (2015) Acquisition des attaques branchantes par des enfants bilingues simultanés portugais-français. *LIDIL – Revue de Linguistique et de Didactique des Langues*, 14 pp. 407-436.

Amorim, C. (2014). Padrão de aquisição de contrastes do PE: a interação entre traços, segmentos e sílabas. Tese de Doutoramento, Universidade do Porto.

Ball, M., Perkins, M., Muller, N. and Howard, S. (org.s) (2008). The Handbook of Clinical Linguistics. Cambridge: Blackwell.

Barlow, J. (2007). Constraint conflict in the acquisition of clusters in Spanish. In F. Martínez-Gil and S. Colina (org.s). Optimality-Theoretic Studies in Spanish Phonology. Amsterdam: John Benjamins.

Bernhardt, B., and Stemberger, J. (1998). Handbook of phonological development from the perspective of constraint-based nonlinear phonology. San Diego: Academic Press.

Bernhardt, B. and Stemberger, J. (2000). Workbook in Nonlinear Phonology for Clinical Applications Austin, Texas: Pro-Ed.

Bernhardt, B. and Stemberger, J. (2008). Constraint-based nonlinear phonological theories : application and implications. In M. Ball, M. Perkins, N. Muller and S. Howard (org.s) *The Handbook of Clinical Linguistics.* Cambridge: Blackwell, pp. 423-438.

Bernhardt, B. and Stemberger, J. (2018), Acquisition of Word-initial Rhotic Clusters in Languages with Taps or Trills". Special issue of *Clinical Linguistics and Phonetics,* 32.

Bishop, D. and Leonard, L. (org.s) (2000). Speech and Language Impairment in Children. Causes, Characteristics, Intervention and Outcome. Hove/NY: Psychology Press.

Brown, F. (1976). Principles of educational and psychological testing. (2nd ed.). New York: Holt, Rinehart and Winston.Drost, E. A. (2011). Validity and Reliability in Social Science Research. Education Research and Perspectives, 38(1), 105–123.

Darwin, C. (1877). A Biographical Sketch of an Infant. *Mind* 2. 285–294.

Demuth, K. (2009). The prosody of syllables, words and morphemes. In E. Bavin (org.) *Cambridge Handbook on Child Language*. Cambridge: Cambridge University Press, pp. 183-198.

Dinnsen, D. and Gierut, J. (orgs) (2008). Optimality Theory, Phonological Acquisition and Disorders. London, UK: Equinox.

Ferré, S., Dos Santos, C. and Almeida L. (2015) Potential clinical markers for SLI in bilingual children. In E. Grillo and K. Jepson (eds) Proceedings of the 39th Boston University Conference on Language Development. Sommerville: Cascadila Press, 152-164.

Fikkert, P. (1994) On the Acquisition of Prosodic Structure. Leiden: HIL.

Fikkert, P. (2007). Acquiring phonology. In P. de Lacy (org.), *Handbook of Phonological Theory*. Cambridge, MA: Cambridge University Press.

Freitas, M. J. (1997). Aquisição da Estrutura Silábica do Português Europeu. Tese de Doutoramento. Faculdade de Letras da Universidade de Lisboa.

Freitas, M. J. (2003). The acquisition of Onset clusters in European Portuguese. In J. Meisel (org.) *Probus. International Journal of Latin and Romance Linguistics.* Vol. 15 (1), pp. 27-46.

Freitas, M.J., Ramalho, A.M., Lousada, M., Oliveira, P., Pereira, P. (2019). PHONODIS - Corpus on Atypical European Portuguese Phonological Development. https://phonbank.talkbank.org/access/Clinical/PhonoDis.html

Gallon, N., Harris, J. and van der Lely, H. K. J. (2007). Non-word repetition: an investigation of phonological complexity in children with Grammatical SLI. *Clinical Linguistics & Phonetics*, 21(6), 435–55. 2007

Gnanadesikan, A. (1995/2004). Markedness and faithfulness constraints in child phonology. In R. Kager, J. Pater and W. Zonneveld (org.s). *Constraints in phonological acquisition*. Cambridge: CUP.

Goad, H. and Rose, Y. (2004). Input Elaboration, Head Faithfulness and Evidence for Representation in the Acquisition of Left-edge Clusters in West Germanic. In R. Kager, J. Pater & W. Zonneveld (org.s). *Constraints in Phonological Aacquisition.* Cambridge: CUP, pp. 109-157.

MacWhinney, B. (2000). The CHILDES project: tools for analyzing talk. 3rd ed. Mahwah, NJ: Lawrence Erlbaum.

McAllister Byun, T. and Rose, Y. (2016). Analyzing Clinical Phonological Data Using Phon. *Seminars in Speech and Language* 37(2):85–105.

Marshall, C., Ebbels, S., Harris, J., and van der Lely, H. (2002). Investigating the impact of prosodic complexity on the speech of children with Specific Language Impairment. In R. Vermeulen and A. Neeleman (Ed.), UCL Working Papers in Linguistics (Vol. 14, pp. 43-68).

Marshall, C. R. and van der Lely, H. (2009). Effects of word position and stress on onset cluster production: evidence from typical development, specific language impairment, and dyslexia. *Language*, 85(1), 39–57.

Mateus, M.H., and d'Andrade, E. (2000). *The Phonology of Portuguese*. Oxford: University Press.

Mendes, A., Afonso, E., Lousada, M. and Andrade, F. (2009/2013). *Teste Fonético Fonológico – Avaliação da Linguagem Pré-Escolar (TFF-ALPE)*. Universidade de Aveiro, Aveiro.

Nespor, M. and Vogel, I. (1986/2007). *Prosodic Phonology*. Dordrecht: Foris.

Pater, J. and Barlow, J. (2003). Constraint conflict in cluster reduction. *Journal of Child Language* 30, pp. 487-526.

Ramalho, A.M., Almeida, L. and Freitas, M.J. (2014). CLCP-PE (Avaliação Fonológica da Criança: Crosslinguistic Child Phonology Project – Português Europeu). Portuguese registration number: IGAC: 67/2014.

Ramalho, A.M. (2017). Aquisição fonológica na criança: tradução e adaptação de um instrumento de avaliação interlinguístico para o PE. Tese de Doutoramento, Universidade de Évora, Évora (http://rdpc.uevora.pt/handle/10174/23564).

Ramalho, A.M. (2019). (2019). RAMALHO-EP - Corpus on Typical European Portuguese Phonological Development (https://phonbank.talkbank.org/; final link to be added as soon as the Phon team finishes the corpus editing procedures)

Rose, Y. (2000). Headedness and Prosodic Licensing in the L1 Acquisition of Phonology. PhD These, McGill University, Montréal.

Rose, Y. and MacWhinney, B. (2014). The PhonBank Project: Data and Software-Assisted Methods for the Study of Phonology and Phonological Development. In J. Durand, U. Gut and G. Kristoffersen (eds.), *The Oxford Handbook of Corpus Phonology*, 380–401. Oxford: Oxford University Press.

Rose, Y., B. MacWhinney, R. Byrne, G. Hedlund, K. Maddocks, P. O'Brien and Wareham, T. (2006). Introducing Phon: A Software Solution for the Study of Phonological Acquisition. In D. Bamman, T. Magnitskaia and C. Zaller (eds.), Proceedings of the 30th Annual Boston University Conference on Language Development, 489–500. Somerville, MA: Cascadilla Press.

Shriberg, L.D., D. Austin, B.A. Lewis, J.L. McSweeny and Wilson, D.L. (1997). The Percentage of Consonants Correct (PCC) Metric: Extensions and Reliability Data. *Journal of Speech, Language, and Hearing Research* 40(4):708–722.

Sua-Kay, E., Tavares, D. (2007). Teste de Avaliação da Linguagem na Criança (TALC). Oficina Didática, Lisboa, 4th edition.

Tamburelli, M., & Jones, G. (2013). Investigating the relationship between nonword repetition performance and syllabic structure in typical and atypical language development. *Journal of Speech, Language, and Hearing Research : JSLHR*, 56(2), 708–720.

8 Language Resources References

CHILDES: https://childes.talkbank.org

PhonBank: https://phonbank.talkbank.org

TalkBank: https://talkbank.org

Cross-Linguistic Child Phonology (CLCP) Project: http://phonodevelopment.sites.olt.ubc.ca

CLCP-EP: www.clul.ulisboa.pt/pt/24-recursos/851-clcp-pe-crosslinguistic-child-phonology-project-portugues-europeu; http://phonodevelopment.sites.olt.ubc.ca/practice-units/portuguese-european/

RAMALHO-EP: https://phonbank.talkbank.org/access/Romance Portuguese/Ramalho.html

PHONODIS: https://phonbank.talkbank.org/access/Clinical/PhonoDi .html