Cadernos de
ESTUDOS LINGUÍSTICOS – (59.3), Campinas, pp. 599-616 - set./dez. 2017

BASIC RESEARCH IN PHONOLOGY, RESOURCES AND APPLICATIONS–THE CASE OF FREQUENCY

MARINA VIGÁRIO[1]
FERNANDO MARTINS[2]
MARISA CRUZ[3]
NUNO PAULINO[4]
SÓNIA FROTA[5]
(UNIVERSITY OF LISBON)

RESUMO: É crescente a valorização da conversão do conhecimento fundamental desenvolvido pelos cientistas em produtos concretos, socialmente relevantes. No Laboratório de Fonética da Universidade de Lisboa tem-se trabalhado nos últimos anos tanto em domínios de investigação fundamental, como de investigação aplicada. Neste artigo é feita uma revisão dos principais recursos recentemente disponibilizados à comunidade por este Laboratório visando o acesso a informação sobre frequência fonológica e lexical. São sumariamente apresentados, em particular, as aplicações informáticas para extração de informação de frequência fonológica e lexical *FreP* (V2016) e *FreLex*, a base de dados alargada *FrePOP* (agora sobre um corpus de cerca de 2 milhões de palavras), e 3 novos léxicos, o *FrePOP Lexicon* (frequência lexical num corpus de 3 milhões de palavras), o *PLEX5 – Um léxico Infantil para o Português Europeu* e o *CDS_EP – Frequência lexical nos dados de fala dirigida à criança da FrePOP (0;11 a 3;04)*. São ainda revistos estudos recentes que fizeram uso destes recursos, ilustrando a sua utilidade em domínios de investigação fundamental e aplicada.
Palavras-chave: Frequência fonológica; Recursos linguísticos; Português

ABSTRACT: Scientific knowledge transfer into society in the form of concrete products is increasingly valued. At the Lisbon Phonetics Lab work has been done for the past years in both basic and applied research. In this paper we survey recent major resources and applications made available by this Lab in areas related to phonological and lexical frequency in Portuguese, in particular, the software applications for extracting phonological and word frequency information *FreP - Frequency in Portuguese* (V2016) and *FreLex – Lexical frequency*, the extended reference database *FrePOP* (ca. 2 million words), and three new lexica, i.e. *The FrePOP Lexicon* (based on an input corpus of 3 million words), *PLEX5 – A production lexicon of child speech for European Portuguese*, *CDS_EP: A lexicon of Child Directed Speech from the FrePOP database (0;11 to 3;04)*. Recent research using these resources is also reviewed, illustrating their usefulness in several areas of applied, experimental and basic linguistics.
Key-words: Phonological frequency; Linguistic resources; Portuguese.

[1] marina.vigario@mail.telepac.pt
[2] fmartins@campus.ul.pt
[3] marisasousacruz@gmail.com
[4] npaulino@campus.ul.pt
[5] sonia.frota@mail.telepac.pt

 doi/10.20396/cel.v59i3.8651000

FOREWORD: *This paper is offered in honor of one of the most prominent phonologists working on Portuguese, Bernadete Abaurre, who has also dedicated a significant part of her professional life to the area of language education, and alphabetization, in particular. Some of the authors of this paper have worked in common projects with Bernadete, and a major outcome of these joint projects was a pioneer and multidisciplinary research exploiting the frequency of phonological objects and patterns in the identification of a phonological change that happened in earlier stages of Portuguese (published in 2012 in Journal of Historical Linguistics, together with Charlotte Galves and Veronica González-López). This research was only possible because at that time we had a (version of a) tool to identify and count relevant phonological units. Such tool in turn could only exist because of prior basic research on phonological features, segments, syllables, prosodic words, word stress. A major goal of this paper is to illustrate how basic and applied investigation in the area of linguistics may feed each other in very productive ways. Some of the applications reviewed are in the area of education and other domains of social interest. With her intelligence, background, experience and topics of interest, we believe Bernadete is certainly one of the scholars best placed to evaluate and hopefully appreciate the work reviewed here.*

1. INTRODUCTION

There are many reasons to invest in making available, in multiple forms, information on the frequency of phonological objects and patterns, as well as on word frequency. Some of these have been reviewed in Frota, Vigário & Martins (2006) and Vigário, Frota, Martins & Cruz (2012), ranging from the role it plays in language acquisition, in language change and in language processing, including reading and writing, to its effects in speech production and in subjects' response in memory tasks related to language (Dell 1990; Saffran, Aslin & Newport 1996; Bybee & Hoper 2001; Coltheart, Rastle, Perry, Langdon & Ziegler 2001; Rastle 2007; and many others).

Because of its relevance, frequency information in Portuguese is more and more taken into account in the production of materials for research in many areas of linguistics, as well as for developing resources for language assessment (see some examples in section 3, below).

Importantly, the frequency of phonological objects and patterns is specific to each language. The tools for extracting this type of information also largely require language specific knowledge. Furthermore, not only databases with frequency information are needed, but tools that allow the extraction of the same type of information from new materials are also necessary. The lack of this type of resources for Portuguese at the time, together with the maturity attained in the studies on Portuguese phonology, as well as a growing perception of the importance of phonological and lexical frequency in language led us to develop a number of resources in this area. In the line of previous reviews (Vigário et al. 2012), in section 2 we present a selection of the newest resources recently made available by our team.

Certainly fundamental for the understanding of human language is to determine the relative importance of grammar and frequency. Even though this is not so often a topic of investigation on its own (but see Diessel & Hilpert 2016 for an updated review), some of the research on Portuguese has begun to highlight not only areas where frequency clearly seems to be relevant, but also domains where results go against predictions based on frequency alone. Some were pinpointed in Vigário et al. (2012), and a few more are identified in section 3.

The paper is organized as follows. In the next section we describe FreP – Frequency in Portuguese (V2016), the first version of FreP available via download, as well as the present state of FrePOP database, and the 3 lexica that have become accessible with the FrePOP package. Section 3 reports on a short selection of recent work where (the information obtained with) some of these resources were exploited. Cases of mismatch between predictions based on frequency and actual results are also highlighted, pointing to the need of further research. Concluding remarks close this paper in section 4.

2. FreP (V2016), FrePOP and FrePOP Lexica

In this section, we describe the present form of a number of resources that have started to appear at the Phonetics Lab of the University of Lisbon (CLUL/FLUL) shortly after the turn of the millennium. These new products emerged and were entirely implemented by phonologists, crucially making use of knowledge built in the field of Portuguese phonology in the previous few decades. Electronic resources such as FreP – Frequency in Portuguese and FrePOP were specifically conceived to provide information on the frequency of the major phonological objects and patterns at the level of the word and below. Earlier research on the prosodic word, stress assignment, syllable structure and feature theory provided the linguistic knowledge necessary in the first place to implement FreP. For European Portuguese (EP), this research included, among others, Mateus (1975), Vigário & Falé (1994), d'Andrade & Viana (1994), Mateus & d'Andrade (2000), Vigário (2003). We must also acknowledge the work that has led to a very coherent orthographic system, in particular by Gonçalves Viana, who was decisive in the present-day fixation of the Portuguese orthography (see for instance Viana & Abreu 1885; Viana 1904).

The first version of FreP was presented in Vigário, Martins & Frota (2005) and Frota, Vigário & Martins (2006). Since then the tool has grown and was subject to test, in a systematic way, as well as by individual users interested in different facets of the tool. As a result, improvements were made. Several studies have been conducted using this software, and a reference database, FrePOP (Frota, Vigário, Martins & Cruz 2010), was fed on the basis of frequency information obtained with FreP. In the following subsections we present the latest version of this tool, the extended form of FrePOP database, and the three lexica recently made available at FrePOP web platform.

2.1. FreP V2016 - Main features

In this section we summarize the main features of FreP, V2016, which is now freely available for download for the first time (http://labfon.letras.ulisboa.pt/FreP/tools.html). This version is restricted to scientific and non-commercial purposes (other uses require authors' contact).

Detailed descriptions of earlier versions of the tool have appeared in Frota, Vigário & Martins (2006) and Vigário, Frota, Martins & Cruz (2012). The Manual, which is part of the package for download, contains information on the requirements for use, how to install the application and instructions of use.

The application runs on any recent version of Windows. It also runs on Linux and Macintosh, using the *Wine* emulator. FreP V2016 opens non-formatted, plain text files (.txt files or similar), with a maximum of 250,000 orthographic words. The frequency information provided for a variety of phonetic-phonological and orthographic objects and patterns may be acquired by navigating through the menus accessible on the main screen (see Fig. 1), and may also be exported into an Excel file, by choosing the Export>Report option.

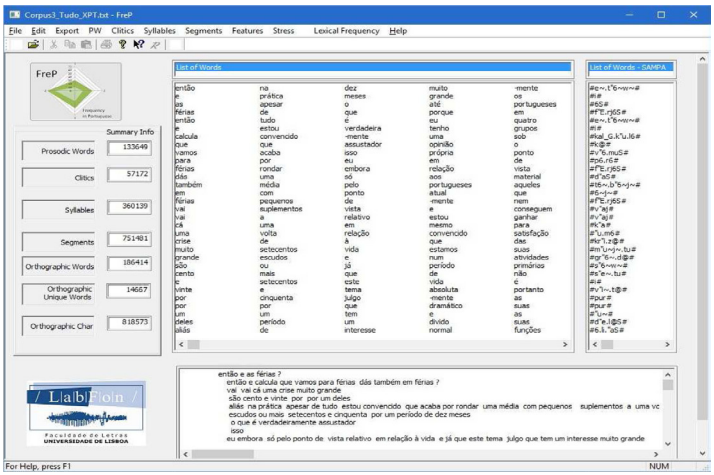

Fig. 1: FreP V2016 main screen.

Taking advantage of a highly predictable relation between Portuguese orthography and (lexical) phonology, FreP allows the automatic extraction (identification, count and listing) of the following phonological units and patterns:

(i) prosodic words

(ii) phonological clitics

(iii) syllables

(iv) segments

(v) articulatory features

(vi) stress patterns

The frequency of these units may be computed taking into consideration specific conditions like the position in the word and/or in the syllable, presence or absence of word stress, size (in number of segments or syllables), direction of phonological attachment (in the case of clitics). Specifically, FreP V2016 provides the frequency of:

(i) prosodic words by word-size (number and list of words with one, two, three, … N syllables or segments)

(ii) phonological clitics by word size and direction of phonological attachment (enclitics and proclitics)

(iii) syllable types (CV, V, CVC…), with the possibility of further taking into account the position in the word or in the syllable, and/or the presence/absence of word stress

(iv) major segment class (consonant, vowel, glide, plus Nasal autosegment and V-slot)

(v) segments by position in the word or in the syllable, and/or in stressed/unstressed syllable

(vi) distinctive features grouped by Point of Articulation, Mode of Articulation, Nasality and Voicing features, taking into consideration the features' position in the word or in the syllable, and/or in stressed/unstressed syllable.

For most of the counts, the exact list of specific items that are extracted from the input text is displayed on the screen.

FreP also gives the frequency of orthographic objects, namely, total number of orthographic words and characters, as well as of individual orthographic words (in alphabetic order or ranked by frequency).

A list with the phonetic transcription of the words extracted and computed with FreP is available on the main screen (in SAMPA, http://www.phon.ucl.ac.uk/home/sampa/).

The list of word types in the text and their respective frequency is also given in the Report. Besides the Report, two additional options in the Export menu of the main screen allow generating two new plain text files, one containing only the prosodic words of the input text file and the other only the phonological clitics.

The criteria for the identification and segmentation of phonological units are detailed in the Manual of FreP V2016, which also includes a section detailing the major limitations of the tool, as well as the ways of avoiding expectable errors. In general, all segmental phonological phenomena that are obligatory are considered, usually following standard assumptions on the phonology of the variety of Portuguese spoken in Lisbon area (see the references at the beginning of this section). Optional less frequent phonological phenomena are ignored by FreP. In the case of the nasal autosegment in syllable final position in word internal position, which obligatorily nasalizes the preceding vowel and deletes, the program displays independent information on its frequency, and consonant clusters that do not conform to the general principles of syllable construction in Portuguese are assumed to display a V-slot position (see the references above, and the Manual of FreP V2016 for further details).

2.2. FrePOP database – Main features

FrePOP is a database that contains information on the frequency of occurrence of a number of phonological objects and patterns in various types of Portuguese corpora. It is freely available at http://frepop.letras.ulisboa.pt (after a simple process of registration). The data may come from originally written corpora or speech that was orthographically transcribed, following in general the current orthographic convention (some specific modifications were made in order to allow better results with FreP). Frequency counts were obtained with a previous version of FreP tool, tested for accuracy (see Vigário et al. 2012). While some of the corpora were collected by the FreP Project's team, a large portion was made available by other researchers or research teams. All the details about the corpora, and their respective sources, can be found clicking the button 'About' in the main screen of the database webpage (option on the left box). The properties of the database are detailed in the Manual (accessible using the option 'How to' in the same box) (see also Vigário et al. 2012 for a previous description of the database, in Portuguese). The actual number of words in FrePOP corpora totals more than 3 million word tokens, but the frequency information presently available covers around two thirds of that amount. It is expected that the frequency information reflecting the full range of corpora will be available soon.

There are a number of search options, some related to the properties of the corpora and some related to the specific linguistic variables for which frequency information is quested. Search options related to the FrePOP corpora include: *Text Type* (written, spoken), subtype (bibliography, CDS=Child Directed Speech, CS=Child Speech, interview, news, spontaneous, technic/scientific); *Transcription* (orthographic or adapted, the latter is found in some orthographic transcriptions which include non-conventional orthography in general aiming at capturing speech properties that are not representable using regular orthography); *Year* (where the date or time frame of the data may be defined); *Age* (where the age range of the subjects that produced the corpus may be selected); *Variety* (presently corpora come from the European variety of Portuguese alone); *Level of Education* (all levels, from primary to university education, plus *illiterate*); and *Occupation* (according to the classification in *CNP – Classificação Nacional de Profissões,* plus the category *student*). If no options are defined, the whole range of corpora is considered in the search.

The linguistic variables that may be selected by the user are the following: Orthographic words (Tokens), Orthographic Characters, Prosodic Words, Clitics, Syllables, Segments and Stress. For the phonological objects, further frequency information is available: (i) Prosodic Words' frequency by size in number of syllables, plus monosyllables formed of open syllables; (ii) Clitics' frequency by size (the same options available for Prosodic Words), as well as by directionality of phonological attachment (proclitic and enclitic); (iii) Syllables' frequency of the ten most frequent syllable types in Portuguese, which in general account for more than 95% of the syllables in the language, by position in the word (initial, final and medial, and in monosyllables), and by stress condition (stress

and unstressed); (v) frequency of the major classes of segments (Consonants, Vowels, Glides, V-Slots) and of individual segments, the latter by position in the word and by stress condition; (vi) Stress distribution frequency patterns (Final, Penult, Antepenult, Stressed Monosyllables), with the possibility of also considering the size of each type of word (starting with the frequency of words with the minimum number of syllables that is required in order to have each specific stress pattern – i.e. for words with final and penult stress, 2 syllable words, and for words with antepenultimate stress, 3 syllable words). On the basis of the information directly available a number of other variables may be computed (like, for example, the frequency of consonants by position in the word or by stress condition).

Three examples of FrePOP screen displays are given in Fig. 2.

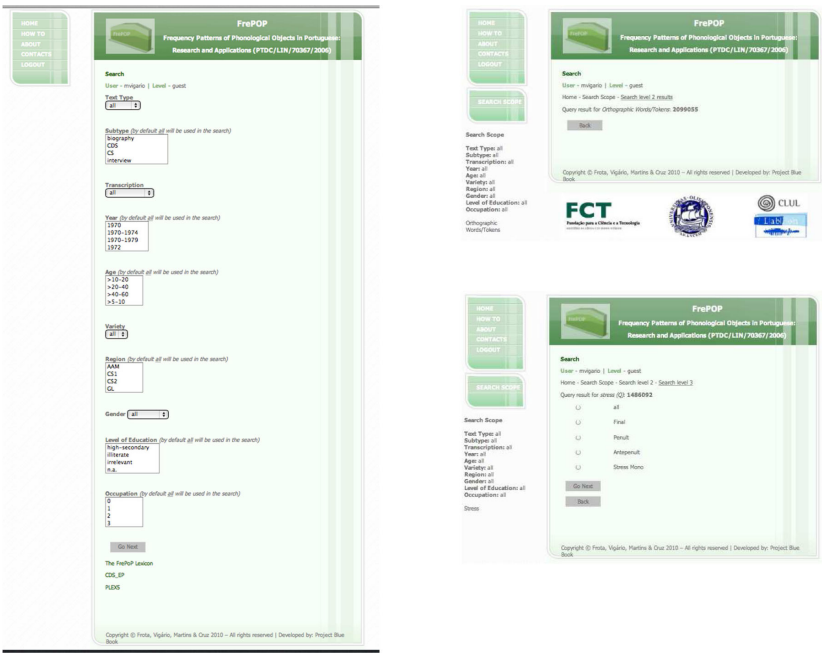

Fig. 2: Screen display of FrePOP menu: the search options related to the properties of the corpora (left panel); the results of the search for number of tokens (right panel, top) and of the search for stress patterns distribution (right panel, bottom).

The lexicon from the corpora that feed FrePOP is made available for the whole dataset, for the dataset coming from Child Directed Speech and from the Child Speech dataset. The three lexica can be accessed in FrePOP's web platform, once logged in. A brief description of these lexica follows in the next sub-sections.

2.3. FrePOP Lexica – A brief description

2.3.1. The FrePOP Lexicon and FreLex

The *FrePOP Lexicon* (Vigário, Cruz, Paulino, Martins & Frota 2015) contains the list of orthographic word forms of all texts that constitute the input for the frequency information in FrePOP, and their respective frequency of occurrence. This list was obtained with an application that generates lexical frequency lists from large input files, *FreLex – Lexical frequency* (Martins, Vigário & Frota 2012).

This Lexicon contains a little more than 84,000 unique word forms, from an input corpus of slightly over 3 million word tokens. The original corpus included texts with vocabulary of different types, coming for example from titles of films or books in languages other than Portuguese. It was therefore found useful to provide some additional information for particular classes of words. A manual codification was made of: (i) foreign words not used in the language; (ii) foreign words used in the language with non-native spelling; (iii) acronyms, abbreviations or truncations; (iv) interjections that are spelled in a way that does not comply with the regular grapho-phonotactics of the language. These codes may be useful for various purposes (such as eliminating word forms that clearly do not belong to the language), but they can also be ignored (for example, if users wish to apply different criteria to obtain some of the categories coded or if the categories coded are irrelevant for their purpose). Clear cases of non-words were eliminated and digits in dates and numbers were converted to their full word counterparts.

2.3.2. CDS_EP – A lexicon of Child Directed Speech from the FrePOP database

The *CDS_EP: A lexicon of Child Directed Speech from the FrePOP database (0;11 to 3;04)* (Frota, Cruz, Martins & Vigário 2013) was based on a section of the Child Directed Speech materials included in the FrePOP database (see the details in the FrePOP 'About' option, http://frepop.letras.ulisboa.pt).

This child directed speech resource is part of the supporting documents for the development of the *MacArthur-Bates Communicative Inventories (CDI) for European Portuguese – Short forms* (Frota 2012a, b; Frota, Butler, Correia, Severino, Vicente & Vigário 2016). The main goal behind CDS_EP was to provide information on the child input relevant to the developmental course of lexical acquisition. Lexical frequency lists (word forms, not lemmatized) were obtained with the FreP tool, Version 3.0 (Martins, Vigário & Frota 2011).

CDS_EP comprises data from eight caregiver-child dyads and contains frequency lists for the following age ranges (organized alphabetically and by frequency rank):

CDS_0-18 (types: 2714; tokens: 48746)

CDS_16-30 (types: 5951; tokens: 201871)

CDS_31-36 (types: 2009; tokens: 23078)

CDS_37-40 (types: 1558; tokens: 12923)

CDS_0-40 (types: 7634; tokens: 286618)

2.3.3. PLEX5 – A production lexicon of child speech for European Portuguese

PLEX5 – A production lexicon of child speech for European Portuguese (Frota, Correia, Severino, Cruz, Vigário & Cortês 2012) is a production lexicon of child speech orthographically transcribed. It is publicly available for scientific, educational and clinical purposes and is based on the materials from three public databases of child speech: *EP_Mono. Base de Dados de Aquisição do Português Europeu como Língua Materna (dados monolingues)* (Correia & Costa 2010), *LumaLiDaOn* (Frota, Vigário & Jordão 2008) and *LumaLiDaAudy – Audio Child Speech Database with phonetic transcription and prosodic labeling* (Frota, Vigário, Matos, Cruz & Jordão 2012).

This child speech resource was built as part of the supporting documents for the development of the *MacArthur-Bates Communicative Inventories (CDI) for European Portuguese – Short forms for Level I* (8 to 18 months) and *Level 2* (16 to 30 months) (Frota a,b; Frota et al. 2016). PLEX5 may be a source of information on the developmental course of lexical acquisition. Lexical frequency lists were obtained with the FreP tool Version 3.0 (Martins et al. 2011).

PLEX5 contains the following information, as database fields: (i) orthographic word list, together with frequency of occurrence; (ii) morphosyntactic category; (iii) mean age of emergence; (iv) number of children that produced the word; (v) information on which children produced the word; (vi) lemma list, together with frequency of occurrence; (vii) lemma mean age of emergence; and (viii) number of children that produced the lemma.

For the purposes of PLEX5, the following criteria were established: (i) age of emergence corresponds to the date of the first of two consecutive sessions where the same word is produced; (ii) lemma means the dictionary entry (it collapses all regular flexional forms under a unique entry); (iii) ambiguous word forms were disambiguated according to context, by manual search on the text files, and frequency of occurrence was recalculated.

PLEX5 thus provides both group and individual lexical development data, assembled in a single database. Version 1.1, released in November 2012, includes the lexicon produced between 0;11 and 2;6, comprising 34.517 tokens and 2.166 word types.

The database is currently being extended to include lexical information for 'João' between 2;0 and 2;6, based on new transcriptions by N. Matos, extending Correia & Costa (2010). PLEX5 (version 1.2) will thus comprise 36.426 tokens and 2.816 word types. Work also proceeds in order to extend the lexical data covered until age 3;00 for 4 of the children (version 2.0, including 61.318 tokens and 7.960 word types).

 doi/10.20396/cel.v59i3.8651000

3. RECENT APPLICATIONS

FreP, FrePOP and FrePOP CDS and CS lexica have been used in recent investigation in various domains. Below we briefly report on some of this work.

In Pires, Cavaco & Vigário (2017) FreP was used to obtain data for a number of linguistic variables that are argued in Vigário (2012) to be informative for measuring linguistic complexity in Portuguese. It is proposed that the frequency of phonological variables in a given text (whether originally written or the orthographic transcription of spoken material) may contribute to assess text complexity/difficulty, in particular, word size, words with exceptional features (e.g. with exceptional stress location, rare syllable types, and the very uncommon V-Slot). In addition to FreP, FrePOP was also used in this study for comparing the data under investigation (namely, medicine package leaflets, news and oral texts) with the frequency data obtained with FrePOP. FrePOP data were separated into three groups, using search filters: written/news, spoken/illiterate and primary, and spoken/secondary and above. Within the FrePOP data, the results were clear in systematically showing that spoken data yields frequency values that indicate greater simplicity and/or a clear tendency towards the use of more frequent patterns in the language, in comparison with the written data, and more so in the corpora from the less educated subjects. In the comparison of the results from the study of the medicine package leaflets and a smaller subset of news and spoken corpora of FrePOP, the results showed similar tendencies, revealing that the size of corpus does not seem to have a significant impact in the variability of results. This indicates that the proposed variables may be a contribution for assessing linguistic complexity/difficulty in corpora of varying magnitude.

Some of the frequency data resources described in the previous section have also been used in the construction of materials for language assessment and for experiments in psycholinguistics. For instance, Jesus, Valente & Hall (2015) used FreP and the FrePOP database to evaluate if the Portuguese version of "The North Wind and the Sun" was phonetically balanced for the language. This passage has been translated and phonetically transcribed in many languages, 29 of which published in *The Handbook of the International Phonetic Association* (1999), European Portuguese included (Cruz-Ferreira 1999). Jesus and colleagues conclude that the EP version of the text is a phonetically balanced text for EP. These results are relevant because the fact that translations of this particular text are available for a great number of languages makes it an evident choice for work in many areas where the comparison across languages is useful. The same text was chosen, for example, in an ongoing project that involves, among other things, comparing the performance of two large groups of patients in different stages of Parkinson's disease, speaking French and Portuguese, respectively, in a reading-aloud task (Pinto et al. 2016).

Another text using FrePOP data is evaluated in Mendes et al. (2012, 2014). This text was built with the purpose of being phonetically balanced and function as an instrument for speech production assessment (i.e. to be used in a reading-aloud task as a screening instrument for Motor Speech Disorders). After determining the sensitivity of the text to the European-Portuguese (EP) dialectal variations, the authors conclude that it can be applied to any EP speaking dialects.

In Coutinho (2014) a screening instrument was built for testing subjects with dyslexia, where phonological frequency obtained with FreP family tools was intensively used. The instrument, which included tasks of repetition and reading aloud of words and pseudowords, and auditory discrimination of pseudowords minimal pairs, was applied in a pilot-experiment. The results showed in general that the higher the phonological probability index associated to a pseudoword, the higher the success rate of the subjects, with subjects diagnosed with dyslexia performing significantly poorer than the group of subjects not diagnosed as dyslexic.

Ribeiro (2011) also develops a new a tool very much grounded on phonological frequency. In this case, phonological frequency information was used to create pseudowords to be employed in a pseudoword repetition task with Portuguese population. This type of instrument is useful because pseudoword repetition has been found to be a sensitive marker for SLI (Specific Language Impairment) (see Bishop, North & Donlan 1996, Ribeiro 2011, Girbau 2016 and references therein).

FrePOP CDS lexicon and PLEX5 were also intensively used in the recent adaptation of the *MacArthur-Bates Communicative Inventories (CDI) for European Portuguese – Short Form I* and *Short Form II* in the process of selecting the words to appear in the forms (Frota et al. 2016). Word frequency patterns in PLEX5 and CDS were instrumental to ensure that the lexical items included show varying ages of acquisition (early as well as late-appearing words) and varying frequency profiles (high, mid and low frequency words), and that individual, regional and other kinds of biases for different age ranges were avoided. CDS and adult speech frequency data were also used as reference to guaranty that the words chosen reflect to a certain degree the language's frequency patterns of word shapes (size), stress distribution and syllable types. This adaptation of the short forms of the CDI is most valuable not only because these instruments were found to be very informative of infants' and toddlers vocabulary growth and language development in numerous languages, also providing indication on babies' subsequent language development, but also because *The MacArthur-Bates Communicative Inventories (CDI) for European Portuguese – Short Form I* and *Short Form II* are the first published instruments for the assessment of language development in EP-learning infants and toddlers (see the literature review and the normative data from the EP population in Frota et al. 2016).

FreP family of resources has also been involved in the construction of the materials for experimental research. In a recent investigation on the perception of word stress placement, Correia, Butler, Vigário & Frota (2015) take into account FrePOP frequency information both in the selection of the form of the pseudowords that were used in the experiments and in the discussion of the results. In this study it was found that, at odds with the frequency of stress distribution patterns (predominance of penultimate stress), European Portuguese subjects show great difficulty in perceiving word stress, patterning like speakers of languages with fixed stress, described in the literature as *stress deaf*, such as French or Finnish. Considering that very often frequency goes on pair with subjects' behavior in language processing tasks, these results call for a deeper understanding of the factors behind *stress deafness* effects, which are discussed in the paper.[6]

Phonological frequency also guided the construction of the experimental materials of Ferreira et al. (2014). In this study, the intelligibility of 4 synthetic voices available for EP were tested. For the intelligibility tests, a corpus of pseudowords was elaborated which mirrored the frequency patterns in the language in terms of the syllable formats, word size and stress patterns. FrePOP and FreP were used for obtaining the frequency information. In the task, after listening to each pseudoword subjects had to write it as accurately as possible. The effect in error rate of word size, stress pattern and syllable structure was evaluated. The results did not reveal effects of the particular synthetic voice under test in terms of particular categories or patterns. However, interestingly enough across voices, including the control (=natural voice), the most marked stress pattern (i.e. antepenult stress, which is clearly the least frequent), was associated with much higher error rate. Similar tendencies were also found across synthetic voices for more errors as word length increased, though not in natural voice, where word length was only a (positive) factor in monosyllables.

Some investigation was also conducted focusing on phonological frequency effects in various domains. For example, the frequency of word size and stress patterns in EP coming from published work using FreP was analysed in Afonso, Freitas & Alves (2009), who experimentally tested word segmentation abilities by children between 4 and 7 years-old. The hypotheses put forth included the expectation of better results in the segmentation of syllables in shorter words, which are more frequent in the language, and in words with the most frequent stress pattern. Response accuracy and reaction times were measured. Globally, results confirmed the hypotheses entertained in that the less frequent patterns were associated with lower accuracy and longer reaction times in syllable segmentation. In the same work the effect of syllable structure was also tested. CV and V syllables showed better results than syllables with complex onset. While here the authors interpret the results as springing from the differences in syllable complexity, we may observe that

[6] The relevant factors seem to be related to the phonological grammar and distribution of reduced vowels and pitch accents in EP. As pointed out in Vigário et al. (2010, 2012), another area where frequency clearly seems to be overridden by grammar is pronominal clitic distribution, since while phonological clitics are overwhelmingly proclitic in the language, an increasing tendency for enclisis is found in contexts of proclisis in present-day EP.

CV and V syllables are also noticeably more frequent than CCV syllable types (see Vigário, Frota & Martins 2010 and references therein). The fact that CV are in fact more complex than V syllables, but do not yield differences in response accuracy and reaction times may indicate that frequency is more relevant than complexity for the syllable segmentation task. Interestingly, within CCV syllable types, Afonso and colleagues also found a difference between onsets formed of a consonant plus a liquid versus a tap. Here, the two syllable formats are similar in terms of complexity, but not in terms of frequency. Results show that children were also clearly better at segmenting CrV syllables (47,3% correct segmentation) than ClV syllables (16,8% correct segmentation), in line with the differences found in earlier work in the frequency of both syllable formats in the language (Vigário & Falé 1994). A final result reported in this work is worth noticing. In the same task, reaction times did not follow the trend of correct response rate, i.e. segmentation of CrV formats showed slower reaction times than ClV formats. Thus, for reaction times in this type of task other factors besides frequency must be playing a (more relevant) role.

In her experimental work on phonological awareness in EP children in primary school (first four years of obligatory formal education), Alves (2012) also systematically considers the frequency of occurrence of segments in the adult speech. On the one hand, segmental frequency is used in the construction of the stimuli for the phonological awareness assessment tools newly developed for EP. On the other hand, frequency is also investigated as a possible factor in the results observed. Alves concludes that, in contrast with what is observed in the order of emergence in the acquisition of segments, which often mirrors the order of frequency of occurrence in the adult language, higher frequency does not seem to promote accuracy in phonological (segmental) awareness tasks of the tested Portuguese children.

The effects of phonotactic frequency in the processing of words and non-words in EP were also very recently investigated in Leone-Fernandez, Vigário, Jerónimo, Alter & Frota (2017). Previous studies had shown that phonotactic knowledge is relevant in multiple dimensions for word processing: high frequency sound sequences are learned easily, faster and earlier than low frequency ones; incongruous words and pseudowords, unlike phonotactically illegal sequences elicit a specific brain response (an N400); pseudowords and illegal sequences display a similar early processing (see references in Leone-Fernandez et al. 2017). In this ERP study, the frequency of segments was controlled in the construction of the stimuli. Words, pseudowords with high and low frequency phonotactic sequences, and non-words, with illegal phonotactics were tested. The results showed similarities and differences in several brain wave regions related to the processing of the different types of words. Among other things, an effect was found in the brain wave at an early temporal stage for stimuli with low or null phonotactic frequency, indicating that this type of object was acoustically detected as 'strange' in the language. A second effect was found in the wave shortly after, distinguishing the items with impossible sound sequences from those with low phonotactic frequency. This was interpreted as an early effect of phonological grammar in word processing. This was the first time such effects were observed because the frequency of pseudoword sound patterns had not been previously taken into account in this type of study.

To conclude this brief review, we would like to report on theoretical work on Portuguese phonology where frequency has been given a critical role. An early example was pointed out in Vigário et al. (2012), involving the status of Word Minimality constraints in Portuguese. While a Word Minimality constraint is evoked in many languages in order to account for the phonological patterns and processes in those languages, the high frequency of monosyllables with open syllables has been taken by phonologists working Portuguese to indicate that this constraint does not play a role in Portuguese phonology. In a recent work, Garcia (2017) also examines the role of frequency patterns in Portuguese, in this case in the distribution of stress in non-verbs. Using FrePOP data in the analysis implemented, Garcia proposes a probabilistic approach to stress assignment, which contrasts with the categorical view classically assumed by specialists in Portuguese phonology.

4. CONCLUSION

In this paper we have presented the main properties of three major classes of resources that have recently become freely available: FreP V2016 and FreLex, the extended version of FrePOP database (presently with over 2 million words) and three lexica (FrePOP Lexicon, CDS_EP and PLEX5). These resources give access to information previously unknown, or insufficient on phonological and lexical frequency in Portuguese. Importantly, not only reference information on phonological and lexical frequency in different types of databases has been made available, but tools for extracting frequency data from new texts, as well as for creating new frequency lexica are also made available to the scientific community.

Recent studies that use some of these resources were also surveyed, illustrating the impact that this work may have in clinical and educational areas, as well as in experimental and basic research on language.

An autonomous line of basic research, only barely touched in this revision paper, concerns the division of labor between grammar and frequency in multiple domains, such as language development, language processing, and language variation and change. For some scholars, *frequency is not just a performance phenomenon, distinct from mental grammar. Rather, the frequency with which linguistic forms are experienced is at the heart of our grammatical knowledge* (Diessel & Hilpert 2016). While we already know that frequency plays a role in many areas of linguistics, as research proceeds domains also stand out where grammar (in a classical sense) seems to override frequency. Under which conditions this happens is certainly a most relevant question in the field of language research, and one we very much wish to contribute to answer with further work.

REFERENCES

AFONSO, C., FREITAS, M. J., ALVES, D.,<HTML><METAHTTP-EQUIV=”content-type” CONTENT=”text/html;charset=utf-8”>05230 (2009). Complexidade prosódica e segmentação depalavras em crianças dos 4 aos 6 anos de idade. *Cadernos de Saúde* 2(2): pp. 31-41.

ALVES, D. (2012). *Efeito das Propriedades Segmentais em Tarefas de Consciência Segmental, de Leitura e de Escrita*. PhD Dissertation, Universidade de Lisboa.

D`ANDRADE, E., VIANA, M.C. (1994). Sinérese, diérese e estrutura silábica. *Actas do IX Encontro Nacional da Associação Portuguesa de Linguística,* pp. 31-42.

BISHOP, D. V. M., NORTH, T., DONLAN, C. (1996). Nonword repetition as a behavioural marker for inherited language impairment: Evidence from a twin study. *Journal of Child Psychology and Psychiatry and Allied Disciplines* 37(4): pp. 391-403.

BYBEE, J., HOPPER, P. (2001). *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins..

COLTHEART, M., RASTLE, K., PERRY, C., LANGDON, R., ZIEGLER, J. (2001). DRC: A dual cascade model of visual word recognition and reading aloud. *Psychological Review*, 108(1), pp. 204-356.

CORREIA, S., COSTA, T. (2010). *EP_Mono. Base de Dados de Aquisição do Português Europeu como Língua Materna (dados monolingues)*. Laboratório de Psicolinguística, CLUL/Projecto PhonBank. http://www.clul.ul.pt/pt/investigacao/159-acquisition-in-european-portuguese-resources-and-results

CORREIA, S., BUTLER, J., VIGÁRIO, M., FROTA, S. (2015). A Stress “Deafness” Effect in European Portuguese. *Language and Speech* 58(1), 24–47. doi:10.1177/0023830914565809.

COUTINHO, D. (2014). *Processamento Fonológico de Pseudopalavras Linguisticamente Motivadas em Crianças com Dislexia*. MA Dissertation, Universidade do Algarve.

CRUZ-FERREIRA, M. (1999). European Portuguese. In: International Phonetic Association (ed.) *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, pp. 126–130. Cambridge University Press, Cambridge.

DELL, G. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes* 5, 313–349.

DIESSEL, H., HILPERT, M. (2016). Frequency Effects in Grammar. In: Mark Aronoff (ed.) *Oxford Research Encyclopedia of Linguistics*. New York: Oxford University Press.

FERREIRA, J., VIGÁRIO, M., FERNANDES, F., BELCHIOR, F., AZEVEDO, S., NECA, A. R. (2014). Inteligibilidade em Voz Sintetizada. Talk given at *ISAAC Conference 2014*, Julho. Lisboa. doi: 10.13140/2.1.3124.2249.

FROTA, S. (coordinator) (2012a): *MacArthur-Bates Communicative Inventories (CDI) for European Portuguese - Short Form I / CDI para o Português Europeu - Forma reduzida: Nível I*. Laboratório de Fonética (CLUL/FLUL), Lisboa. http://labfon.letras.ulisboa.pt/babylab/pt/CDI_Portugues_Europeu.html, ISBN 978-989-95713-5-8.

FROTA, S. (coordinator) (2012b*): MacArthur-Bates Communicative Inventories (CDI) for European Portuguese - Short Form II / CDI para o Português Europeu – Forma reduzida: Nível II*. Laboratório de Fonética (CLUL/FLUL), Lisboa. http://labfon.letras.ulisboa.pt/babylab/pt/CDI_Portugues_Europeu.html, ISBN 978-989-95713-4-1.

FROTA, S., GALVES, C., VIGÁRIO, M., GONZÁLEZ-LÓPEZ, V., ABAURRE, B. (2012). The phonology of rhythm from Classical to Modern Portuguese. *Journal of Historical Linguistics* 2(2): pp. 173–207.

FROTA, S., CORREIA, S., SEVERINO, C., CRUZ, M., VIGÁRIO, M., CORTÊS, S. (2012). *PLEX5 – A production lexicon of child speech for European Portuguese*. Laboratório de Fonética (CLUL/FLUL), Lisboa. http://labfon.letras.ulisboa.pt/babylab/english/PLEX5.html, ISBN: 978-989-95713-6-5, ISLRN: 124-998-098-334-1.

FROTA, S., CRUZ, M., MARTINS, F., VIGÁRIO, M. (2013). *CDS_EP: A lexicon of Child Directed Speech from the FrePOP database (0;11 to 3;04)*. Laboratório de Fonética (CLUL/FLUL), Lisboa. http://labfon.letras.ulisboa.pt/babylab/english/CDS_EP.html, ISBN 978-989-95713-8-9.

FROTA, S., BUTLER, J., CORREIA, S., SEVERINO, C., VICENTE, S., VIGÁRIO, M. (2016). Infant communicative development assessed with the European Portuguese MacArthur-Bates Communicative Development Inventories Short forms. *First Language*, 36(5): 525-454. doi: 10.1177/0142723716648867.

FROTA, S., VIGÁRIO, M., JORDÃO, R. (2008). *LumaLiDaOn*. Version 1. Laboratório de Fonética (CLUL/FLUL), Lisboa. http://labfon.letras.ulisboa.pt/LumaLiDa.htm, ISBN 978-989-95713-0-3, ISLRN 710-484-042-477-2.

FROTA, S., VIGÁRIO, M., MARTINS, F., CRUZ, M. (2010). *FrePOP* (version 1.0). Laboratório de Fonética (CLUL/FLUL), Lisboa. http://frepop.letras.ulisboa.pt, ISBN: 978-989-95713-2-7, ISLRN: 064-984-771-090-2 (extended in 2012 to ca. 2, 000 000 words)

FROTA, S., M. VIGÁRIO, F. MARTINS (2006). FreP – An Electronic Tool for Extracting Frequency Information of Phonological Units from Portuguese Written Text. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 2224–2229. Genoa.

FROTA, S., VIGÁRIO, M., MATOS, N., CRUZ, M., JORDÃO, R. (2012). *LumaLiDaAudy - Audio Child Speech Database with phonetic transcription and prosodic labeling*. Laboratório de Fonética (CLUL/FLUL), Lisboa. http://labfon.letras.ulisboa.pt/lumalidaaudy.htm, ISLRN 433-882-165-666-8.

GARCIA, G. D. (2017). Weight Gradience and Stress in Portuguese. *Phonology*, 34(1): 41-79.

GIRBAU, D. (2016) The Non-word Repetition Task as a clinical marker of Specific Language Impairment in Spanish-speaking children. *First Language*, 36(1): 30-49.

INTERNATIONAL PHONETIC ASSOCIATION (ed.) (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge.

JESUS, L., VALENTE, A., HALL, A. (2015). Is the Portuguese Version of the Passage "The North Wind and the Sun" Phonetically Balanced? *Journal of the International Phonetic Association* 45(1), 1-11. doi: 10.1017/S0025100314000255.

LEONE-FERNANDEZ, B., VIGÁRIO, M., JERÓNIMO, R., ALTER, K., FROTA, S. (2017). Processing words and non-words: An ERP study on the impact of phonotactic frequency and phonological grammar. Talk presented at the *International Symposium of Psycholinguistics*, Braga, pp. 5-8 April.

MARTINS, F., VIGÁRIO, M., FROTA, S. (2011). *FreP - Frequency in Portuguese*. Version 3.0. Software in CD-ROM (IGAC, nº 6722/2011). Laboratório de Fonética (CLUL/FLUL), Lisboa.

MARTINS, F., VIGÁRIO, M., FROTA, S. (2012). *FreLex – Lexical frequency*. Laboratório de Fonética (CLUL/FLUL), Lisboa. http://labfon.letras.ulisboa.pt/FreP/tools.html

MARTINS, F., VIGÁRIO, M., FROTA, S. (2016). *FreP - Frequency in Portuguese*. Version V 2016. Software in CD-ROM.

MATEUS, M. H. (1975). Aspectos da Fonologia Portuguesa [2nd ed.–revised, 1983]. INIC, Lisboa.

MATEUS, M. H., D'ANDRADE, E. (2000). *The Phonology of Portuguese*. Oxford University Press, Oxford.

MENDES, A. P. B. G., COSTA, A., MARTINS, A., FERNANDES, A., VICENTE, S., FREITAS, T. (2012). Contributo para a construção de um Texto Foneticamente Equilibrado para o Português-Europeu. *Revista CEFAC – Atualização Científica em Fonoaudiologia e Educação* 5: 910-917.

MENDES, A. P. B. G., MOREIRA, M., COSTA, A., MURTINHEIRA, A., JORGE , A. (2014). Validade e sensibilidade do texto foneticamente equilibrado para o Português-Europeu "O Sol". *Revista de Distúrbios da Comunicação* 26, 2: 277-286.

PINTO, S., CARDOSO, R., SADAT, J., GUIMARÃES, I., MERCIER, C., SANTOS, H., ATKINSON-CLEMENT, C., CARVALHO, J., WELBY, P., OLIVEIRA, P., D'IMPERIO, M., FROTA, S., LETANNEUX, A., VIGÁRIO, M., CRUZ, M., PAVÃO MARTINS, I., VIALLET, F., FERREIRA, J. (2016). Dysarthria in individuals with Parkinson's disease: a protocol for a binational, cross-sectional, case-controlled study in French and European Portuguese (FraLusoPark). *BMJ Open 2016* (Neurology), 6:e012885. doi:10.1136/bmjopen-2016-012885.

PIRES, C., CAVACO, A., VIGÁRIO, M. (2017). Towards the definition of linguistic metrics for evaluating text readability in Portuguese. To appear in *Journal of Quantitative* Linguistics. doi: 10.1080/09296174.2017.1311448.

RASTLE, K. (2007). Visual word recongnition. In: Gareth, M. (ed.) *The Oxford Handbook of Psycholinguistics*, pp. 71-87. Oxford University Press, Oxford.

RIBEIRO, V. I. C. (2001). *Instrumento de Avaliação de Repetição de Pseudo-palavras. Estudo Piloto*. MA Dissertation, Instituto Politécnico de Setúbal/Universidade Nova de Lisboa.

SAFFRAN., J. R., ASLIN, R. N., NEWPORT, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science* 274(5294), pp. 1926-1928.

VIANA, A. R. GONÇALVES (1904). *Ortografia Nacional. Simplificação e uniformização sistemática das ortografias portuguesas*. Lisboa: Viuva Tavares Cardoso.

VIANA, A. R. GONÇALVES & G. DE VASCONCELOS ABREU (1885). *Bases da ortografia portuguesa*. Lisboa: Imprensa Nacional.

VIGÁRIO, M. (2003). *The Prosodic Word in European Portuguese*. Mouton de Gruyter, Berlin/New York.

VIGÁRIO, M. (2012). *Measures of phonological complexity*. University of Lisbon, Ms.

VIGÁRIO, M., CRUZ, M., PAULINO, N., MARTINS, F., FROTA, S. (2015). *The FrePOP Lexicon* (version 1.0, based on an input corpus of 3 million words). Laboratório de Fonética (CLUL/FLUL), Lisboa. ISLRN 661-393-864-944-9. Accessible from the FrePOP web platform, http://frepop.letras.ulisboa.pt

VIGÁRIO, M., FALÉ, I. (1994). A Sílaba no Português Fundamental: uma descrição e algumas considerações de ordem teórica. In: *Actas do IX Encontro da Associação Portuguesa de Linguística*, pp. 465–477. Colibri/APL, Lisboa.

VIGÁRIO, M., FROTA, S., MARTINS, F. (2010). A frequência que conta na aquisição da fonologia: *types* ou *tokens*. In: Ana Maria Brito, Fátima Silva, João Veloso & Alexandra Fiéis (eds.) *XXV Encontro Nacional da Associação Portuguesa de Linguística. Textos seleccionados*. Porto: Associação Portuguesa de Linguística, pp. 749- 767.

VIGÁRIO, M., FROTA, S., MARTINS, F., CRUZ, M. (2012). Frequência na Fonologia do Português: recursos e aplicações. In: Costa, A., Duarte, I. (eds.) *Nada na linguagem lhe é estranho. Estudos em homenagem a Isabel Hub Faria*, pp. 613–631. Edições Afrontamento, Porto.

VIGÁRIO, M., MARTINS, F., FROTA, S. (2005). Frequências no Português: a ferramenta FreP. In: Duarte, I., Leiria, I. (eds.) *Actas do XX Encontro Nacional da Associação Portuguesa de Linguística*, pp. 897–908. Colibri/APL, Lisboa.

VIGÁRIO, M., MARTINS, F., FROTA, S. (2006). A ferramenta FreP e a frequência de tipos silábicos e classes de segmentos no Português. In: Fátima Oliveira & Joaquim Barbosa (eds.) *XXI Encontro da Associação Portuguesa de Linguística. Textos Seleccionados*. Lisboa: APL, pp. 675-687.