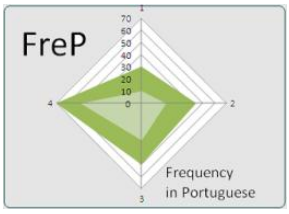


FreP – Frequency in Portuguese

Manual

Fernando Martins
Universidade de Lisboa
Marina Vigário
Universidade de Lisboa
Sónia Frota
Universidade de Lisboa

Last Update – June 2011

Contents

Introducing *FreP*

FreP is an electronic tool that allows the extraction of frequency information of Portuguese phonological units at the word-level and below. It runs on written texts, following the current orthographic conventions. The acronym comes from the expression ***Fre****quency in* ***P****ortuguese*.

Taking advantage of a highly predictable relation between Portuguese orthography and (lexical) phonology, this tool allows the automatic extraction (identification, count and listing) of the following phonological units: articulatory features, segments, classes of segments (consonants, vowels, glides), syllables, phonological clitics and prosodic words. In addition, (i) it locates word stress, and provides information on the distribution of stress within words (i.e. number and list of words with final, penult and antepenult stress), (ii) it counts the number of different features, segments, classes of segments and syllable types (CV, V, CVC…), by position in the word (initial, internal and final), or taking into account the presence/absence of word stress, or both (position in the word and presence/absence of word stress), (iii) it provides information on the size of words (number and list of words with one, two, three, N syllables or segments), and does so for prosodic words as well as for clitics, and (iv) within the class of phonological clitics, it sets enclitics and proclitics apart, providing the number of both types of units separately, together with their respective size. The tool also gives information on orthographic objects, namely, number of orthographic words and characters.

FreP belongs to the public domain, and it is user-friendly, as the information is structured in a transparent way and a system of windows and commands based on the *Windows* format is used.

This manual was developed for version 3.0, June 2011.

Historical Note

FreP emerged from a joint project involving Marina Vigário, Fernando Martins and Sónia Frota, which started in July, 2004.

All three authors have worked on every aspect of the tool, and thus all three are ultimately responsible for each part of the program. There are nevertheless some areas that have been worked more in depth by each one of the authors. Most work on the programming and visual design of the tool was performed by Fernando Martins, with close collaboration of Marina Vigário. The original idea and most decisions regarding the functionalities, as well as the phonological information behind the implemented ideas, come from Marina Vigário, who is also largely responsible for this manual. The precise organization of the functionalities of *FreP*, is largely attributed to Sónia Frota, who also tested the program several times and helped improving the tool.

The tool is under test and still in progress. All comments and suggestions are most welcome and may be sent to the following e-mail addresses:

fmartins@fl.ul.pt
marina.vigario@mail.telepac.pt
sonia.frota@mail.telepac.pt

How to Get/Update *FreP*

FreP was conceived as a public domain tool, with the restriction of being used for scientific, non-commercial, purposes.

The tool may be obtained by request to the following e-mail address:

fmartins@fl.ul.pt

The program will be sent in a zipped folder together with a password, required for the program setup.

The users will sign a user agreement, whereby they agree to properly cite the tool (name of the tool, authors, date and version of the tool) when used for the extraction of information to be made public, and to send the authors of the tool the Excel files created by FreP when running in new corpora. These filed will be used to enlarge the frequency database that is being developed at Laboratório de Fonética (FLUL) – FrePOP – Frequency of Phonological Objects in Portuguese. The provenience of these files will be acknowledged in the database.

As the tool is still in progress, updated versions are planned to appear in the near future. Please keep in touch with us!

Requirements for Use

FreP runs on Windows XP, Windows Vista and 7, Windows Server 2003 and 2008 R2. It also runs on Linux, using the emulator *Wine*.

The tool opens non-formatted, plain text files (.txt files or similar). The files should not exceed 250.000 orthographic words.

Installing *FreP*

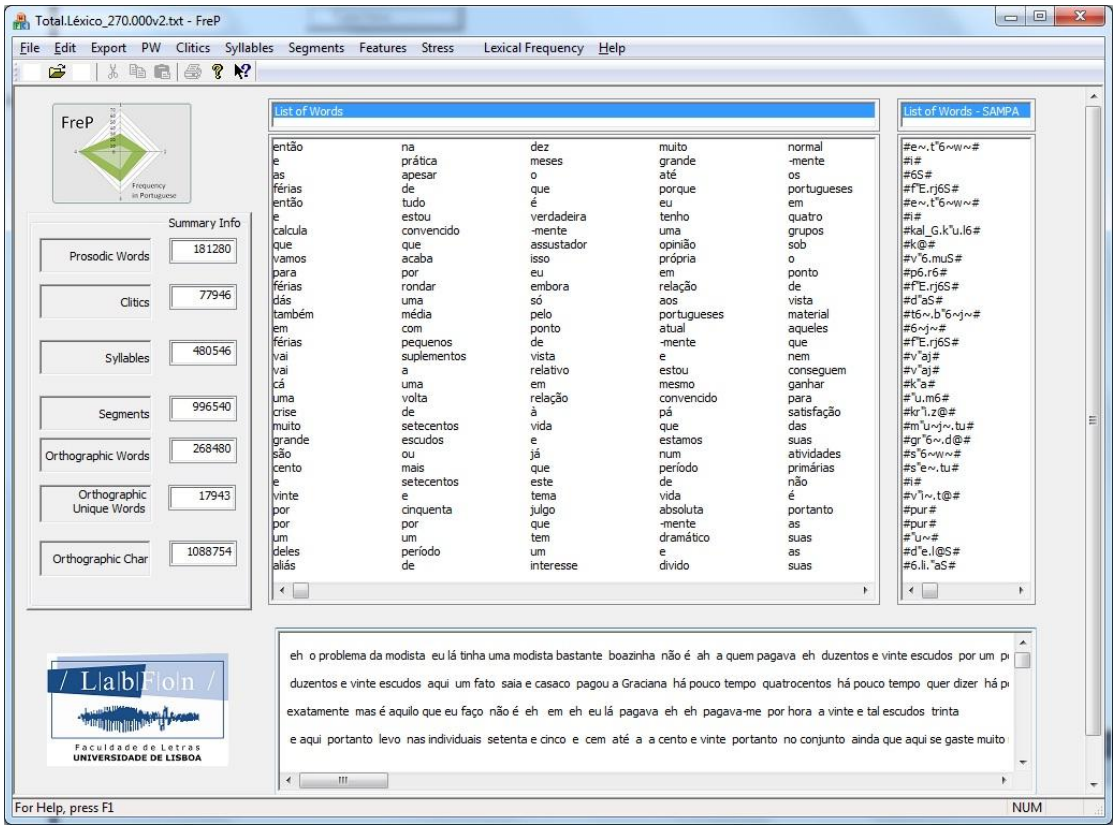
Unzipp the *FreP_setup* folder. Execute the program by clicking on the *FreP_setup* file, and follow the instructions on the screen. Use the password sent together with the program.

Loading a file

In the File menu open the text file you want *FreP* to run on. Navigate throughout the menus (see the details in the following sections).

Map of the Tool

The following picture shows *FreP*'s Opening Screen.



The diagram below presents the options available from the Opening Screen Menus.

Available Menus

Export Menu
- *Report*
- *PW File*
- *Clitics File*

PW Menu
- *Size*
  - *Syllable Count*
  - *Syllable Count with Clitics*
  - *Segment Count*

Clitics Menu
- *Size*
  - *Syllable Count*
  - *Segment Count*
- *Directionality*
  - *Enclitic*
  - *Proclitic*

Syllable Menu
- *Types*
  - *All*
  - *by Position*
  - *by Stress Condition*
  - *by Position and Stress*
- *Actual Syllables*
  - *All*
  - *by Position*
  - *by Stress Condition*
  - *by Position and Stress*

Segments Menu
- *All*
- *by Position*
  - in the Word
    - as a function of the Word
    - as a function of the Syllable
  - in the Syllable
- *by Stress Condition*
- *by Position and Stress*
  - in the Word
    - as a function of the Word
    - as a function of the Syllable
  - in the Syllable

(cont.)

Available Menus (contd.)

Features Menu
- *PoA*
  - *All*
  - *by Position*
    - in the Word
      - as a function of the Word
      - as a function of the Syllable
    - in the Syllable
  - *by Stress Condition*
  - *by Position and Stress*
    - in the Word
      - as a function of the Word
      - as a function of the Syllable
    - in the Syllable

Stress Menu
- *Location*
- *Size*

Lexical Frequency Menu
- *Types and Tokens*

Commands

The Opening Screen displays a box with a *Summary Info*, showing the total number of units (in the opened file) of the following type: *Prosodic Words, Clitics, Syllables, Segments, Orthographic Words*, *Orthographic Unique Words* and *Orthographic Characters*. This box is permanently available.

The Opening Screen also includes:
- The list of extracted words, given in a box at the centre of the page. Note that these do not coincide with orthographic words since enclitics are listed as a separate word, keeping the dash to their left, signalling the enclitic rather than proclitic status (e.g. 1 orthographic word: *deu-me* > two extracted words: *deu* and *–me*), and the prosodic words that integrate a single orthographic word are listed as separate words (e.g. 1 orthographic word: *alegremente*> two extracted words: *alegre* and *–mente*).
- The phonetic transcription (in SAMPA), available for each word extracted on a box at the right-hand of the page.
- The original text, appearing in a box at the bottom of the page.

A scrollbar is available to view the whole rage of data in the three boxes.

Menus

Export	Several types of files may be created, saved by default in the same directory as the non-formatted opened file
▸ *Report*	Creates an Excel file containing part of the information provided by the FreP display (e.g. the list of *words* extracted from the text analysed; the syllable boundaries and templates, point of articulation labels, stress mark, number of syllables and PW/Clitic status for each word; number of PW and Clitics and size; total number of syllables, and syllable types, by position, stress condition and position and stress condition);
▸ PW File	Generates a file containing only the Prosodic Words of the text file, excluding all the Clitics
▸ Clitics File	Generates a file containing only the Clitics of the text file, excluding all the Prosodic Words
PW Menu	Provides frequency information on Prosodic Words
▸ *Size*	Provides frequency information on the size of Prosodic Words
▸ *Size* ▸ *Syllable Count*	Provides the number of Prosodic Words with *one, two, three, N Syllables.* Examples are also available by clicking on the *List* box that immediately follows each number
▸ *Size* ▸ *Syllable Count with Clitics*	Similar to Syllable Count, but including Clitics
▸*Size* ▸ Segment Count	*Provides the number of Prosodic Words with* one, two, three, N Segments. Examples are also available by clicking on the List box that immediately follows each number
Clitics Menu	Provides frequency information on Clitics
▸ *Size*	Provides frequency information on the size of Clitics
▸ *Size* ▸ *Syllable Count*	Provides the number of Clitics with one or two *Syllables.* Examples are also available by clicking on the *List* box that immediately follows each number
▸ *Size* ▸ *Segment Count*	Provides the number of Clitics with one, two,

	three, N *Segments*. Examples are also available by clicking on the *List* box that immediately follows each number
▸ *Directionality*	Provides frequency information on Enclitics and Proclitics
▸ *Directionality* ▸ Enclitics	Provides the number of Enclitics. Examples are also available by clicking on the *List* box that immediately follows each number
▸ *Directionality* ▸ Proclitics	Provides the number of Proclitics. Examples are also available by clicking on the *List* box that immediately follows each number
Syllables Menu	Provides frequency information on Syllables
▸ *Type*	Provides the number and list of Syllables by Syllable Types (e.g. V, CV, CVC,…). The Actual Syllables and their frequency is also given
▸ *by Position*	Provides the number of Syllables by Syllable Type as a function of the position within the word (i.e. initial, internal and final). The list of all syllable types in each position is also available by clicking on the *List* box that immediately follows each number
▸ *by Stress Condition*	Provides the number of Syllables by Syllable Type as a function of the presence / absence of Stress. The list of all syllable types in each Stress condition is also available by clicking on the *List* box that immediately follows each number
▸ *by Position & Stress*	Provides the number of Syllables by Syllable Type as a function of the position within the word (i.e. initial, internal and final) and the presence / absence of Stress. The list of all syllable types by position and stress is also available by clicking on the *List* box that immediately follows each number
Segments	Provides frequency information on (Classes of) Segments
▸ *General*	Provides the number of Segments and Major Classes (Consonants, Vowels, Glides). It also gives information on the number of occurrences of the Nasal autosegment and of V-Slots inserted between consonants that would otherwise violate principles of syllable

	construction.
▸ *by Position*	Provides the number of Segments as a function of the position within the word (i.e. initial, internal and final) and the position within the syllable.
▸ *by Stress Condition*	Provides the number of Segments as a function of the presence / absence of Stress.
▸ *by Position & Stress*	Provides the number of Segments as a function of the position within the word (i.e. initial, internal and final) and the position within the syllable, and the presence / absence of Stress.
Features	Provides frequency information of articulatory features of C, V and G
▸ *PoA* ▸*General*	Provides frequency of Point of Articulation features
▸ *PoA* ▸ *by Position*	Provides frequency of Point of Articulation features as a function of the position within the word (i.e. initial, internal and final) and the position within the syllable.
▸ *PoA* ▸ *by Stress Condition*	Provides frequency of Point of Articulation features as a function of the presence / absence of Stress.
▸ *PoA* ▸*by Position & Stress*	Provides frequency of Point of Articulation features as a function of the position within the word (i.e. initial, internal and final) and the position within the syllable, and the presence / absence of Stress.
▸ *MoA* ▸*General*	Provides frequency of Mode of Articulation features
▸ *MoA* ▸ *by Position*	Provides frequency of Mode of Articulation features as a function of the position within the word (i.e. initial, internal and final) and the position within the syllable.
▸ *MoA* ▸ *by Stress Condition*	Provides frequency of Mode of Articulation features as a function of the presence / absence of Stress.
▸ *MoA* ▸*by Position & Stress*	Provides frequency of Mode of Articulation features as a function of the position within the word (i.e. initial, internal and final) and the position within the syllable, and the presence / absence of Stress.
▸ *Nasality* ▸*General*	Provides frequency of *nasal/non-nasal* features
▸ *Nasality* ▸ *by Position*	Provides frequency of nasal/non-nasal features as a function of the position within the

	word (i.e. initial, internal and final) and the position within the syllable.
▸ *Nasality* ▸ *by Stress Condition*	
	Provides frequency of nasal/non-nasal features as a function of the presence / absence of Stress.
▸ *Nasality* ▸*by Position & Stress*	
	Provides frequency of nasal/non-nasal features as a function of the position within the word (i.e. initial, internal and final) and the position within the syllable, and the presence / absence of Stress.
▸ *Voicing* ▸*General*	Provides frequency of voiced/non-voiced features
▸ *Voicing* ▸ *by Position*	Provides frequency of voiced/non-voiced features as a function of the position within the word (i.e. initial, internal and final) and the position within the syllable.
▸ *Voicing* ▸ *by Stress Condition*	
	Provides frequency of voiced/non-voiced features as a function of the presence / absence of Stress.
▸ *Voicing* ▸*by Position & Stress*	
	Provides frequency of voiced/non-voiced features as a function of the position within the word (i.e. initial, internal and final) and the position within the syllable, and the presence / absence of Stress.
Stress	Provides frequency information on Stress location
▸ *Location*	Provides the number of prosodic words with the three stress patterns (final, penult, antepenult Stress). The complete list of words is also given by clicking on the *List* box that immediately follows each number
▸ Size	Provides the number of prosodic words with the three stress patterns by PW size
Lexical Frequency	Provides type and token frequency, and the frequency ranking of words

Note on SAMPA symbols:

The symbols used for the phonetic transcription are those listed in http://www.phon.ucl.ac.uk/home/sampa/portug.htm, and http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm for the representation of velarization and labialization (used to transcribe dark laterals and labialized plosives, respectively); stress is also signalled in monosyllabic PWs.

Criteria for the Identification and Segmentation of Phonological Units

In order to identify and segment the phonologic units some decisions have been made.

In general, all segmental phonological phenomena that are obligatory are taken into account whenever relevant. Optional (less frequent) phonological phenomena are ignored by *FreP*. In the case of the nasal autosegment in syllable final position, which obligatorily nasalizes the preceding vowel and deletes, the program displays independent information on its frequency.

Some of the rules that have implications to the computation of *FreP* and that have been considered obligatory are the following:

- Glide insertion to break hiatus, as in *passeio* (cf. Mateus 1975; Vigário 2003: Ch. 3)
- Semivocalization yielding rising diphthongs (only) in posttonic position, as in *família* 'family'; and in the sequence -ion-, -eon-

Other decisions are listed below:

- [k_w] and [g_w], in words like *quando* 'when', *guardanapo* 'napkin', are assumed to be labialized underlying consonants (cf. Andrade & Viana 1994; Vigário & Falé 1994)
- Deletion of schwas is not taken into account, not even in word final position, where the process usually applies in intonational phrase internal position (see Vigário 2003). This option is taken to be the most convenient for syllabification purposes, and, since FreP provides the number of schwas in word final position, it is always possible to exclude these instances

The identification of syllable boundaries essentially follows the description and analyses proposed in Vigário & Falé (1994) and Mateus & Andrade (2000). In line with that work, Glides between two vowels, as in *areia* 'sand', are ambisyllabic, and Syllables that do not conform to the general principles of syllable construction in Portuguese have been treated as displaying a V-slot position (e.g. *obter* 'to obtain' is syllabified as *o.bV.ter*); the sequences of plosive plus coronal fricative have been excluded from this. The total number of V-slots in a given file is provided under the menu Segments, where the list of words where V-slots appear are also given. All counts involving number of Syllables include the syllables obtained via V-slot insertion (in the near future, it will be possible to choose not to include such cases – see *Available Options*, below). V-slots are not computed for the number of Segments counts (in the Summary, PW size, Clitics size). For stress location, there are two options: one considering V-slot (e.g. in this count a word like 'facto' has antepenultimate stress) and one without V-slot (e.g. in this count the same word 'facto' has penultimate stress).

The each of the components of an ambisyllabic glide are considered in the count of syllable types (e.g. *areia*> V.CVG.GV), segments by position in the word as a function of the syllable (e.g. in *areia*> #6.r"6j.j6#, the G counts as appearing both in a word internal syllable and in a word final syllable), segments by stress condition (e.g. in *areia*> #6.r"6j.j6#, the G counts as appearing both in a stressed syllable and in a stressless syllable), and segments by position in the word as a function of the syllable and stress condition. In all other contexts, ambisyllabic Glides are counted only once.

The identification of Prosodic Words and Clitics and the directionality of cliticization follows the proposals in Vigário (2003).

Limitations and Tips for Avoiding Errors

Due to cases where phonology may not be predicted from the orthographic conventions, *FreP* inevitably yields some errors. It is possible to avoid them by changing the original orthography, so that FreP interprets the forms correctly. These errors are very limited in number, and are of the following types:

- Abbreviations and acronyms are treated as regular words: thus, APL (which contains three Prosodic Words) is treated as *apl* (a single Prosodic Word)

  Tips for avoiding this type of error: use the *Find* facility of text editor programs to find the words written in caps (go to the *Format* option, and select *All caps*) and convert the letters into their full text names (e.g. *APL* → <á pê ele>)

- Digits, like other non-orthographic symbols, are ignored by *FreP*

  Tips for avoiding this type of error: use the *Find* facility of editor text programs to find digits (go to the *Special* option, and select *Any Digit*) and convert them into full text (e.g. *110* → <cento e dez>)

- Morphosyntactic compounds with more than one word stress that are not separated by a blanc space and form more than one Prosodic Word (PW) – e.g. *monogamia* $(\text{mono})_{PW}$ $(\text{gamia})_{PW}$ ["mOnO g6"mi6] 'monogamy'), are treated as a single PW;

- Given that internal PWs of derived words with *z-avaliative* suffixes (see Villalva 2001) are computed as separate words, there may be some non-derived words ending in a sequence of segments coinciding with the form of *z-avaliative* suffixes that are incorrectly parsed into two PWs. Rather frequent words such as *vizinho/a(s)* 'neighbour' and *cozinha(s)* 'kitchen' have been successfully excluded from this set of errors;

- The morphological base of words with *z*-evaluative suffixes and *–mente* are treated as separated Prosodic Words; however, in monosyllables ending with an oral vowel, such bases are not assigned word stress and thus fail to be assigned PW status (e.g. *pezinho* $(\text{pe})_{PW}$ $(\text{zinho})_{PW}$ ["pE "ziJu] 'foot-dim.'). Rather frequent words such as *sozinho* 'alone', *somente* 'only' have been successfully excluded from this set of errors;

  Additionally, in longer bases with exceptional stress location, stress will be misplaced, as if the base was regular (e.g. *agilmente* $(\text{agil})_{PW}$ $(\text{mente})_{PW}$ ["aZil_G "me~t@] 'skillfully'). Rather frequent words such as *difícilmente* 'hardly' and *facilmente* 'easily' have been successfully excluded from this set of errors;

- Written consonants that do not correspond to a sound are incorrectly parsed as existing consonants (e.g. ó**p**timo ["Otimu] 'great'). Tips for avoiding this type of error: as such cases are limited to the sequences <pt>, <ct>, and <cç>, the graphemes that do not correspond to an existing sound may be manually deleted in the text file by using the *Find* facility of text editor programs.

- <qu> and <gu> followed by <e> of <i> are computed by FreP as [k,g], following the general rule; however, in some words, these consonants are labialized (e.g. *frequente, linguista*). There is no automatic way of avoiding this error; only integrating these words in the list of exceptions inside the tool will prevent this kind of error. A manual count is therefore suggested, using the *Find* facility of the text editor programs.

- Other problems related to the phonetic transcription include: (i) the vowel quality of non-high vowels that exceptionally do not undergo vowel reduction – the use of grave stress mark in the text file allows FreP to treat these vowels as open (e.g. *règública* 'republic; this must be done manually; (ii) the vowel quality of stressed non-high vowels in many cases (e.g. <e> in *colher* 'pick' is a mid vowel and in *colher* 'spoon' is a low vowel) – FreP treats these vowels as mid by default; using acute stress mark in the text file allows to treat the relevant vowels as low; (iii) the phonetic transcription of <x>, which may in many contexts is not predictable – FreP treats these cases as [S] by default; it is possible to manually change <x> to <z>, <ss>, etc. in the text file.

New tools based on FreP

- FreP_B

FreP_B has all the features of FreP, but is optimized for the Portuguese Variety of Portuguese

- FreP_LUDO

FreP_LUDO is a simplified version of FreP, which includes most, but not all of the functionalities of FreP (e.g., it excludes PW size counts including Clitics, as well as all the counts in the menu Features).

This tool is written in Portuguese, includes short definitions of basic concepts used, and was created for non-specialized users, such as undergraduate students and basic and secondary school teachers.

More information

More information about FreP, including a Demo and a list of work done using FreP may be found at http://www.fl.ul.pt/LaboratorioFonetica/frep. For further details, please contact the authors.

References

ANDRADE, E. & M.C. Viana. 1994. Sinérese, diérese e estrutura silábica. In *Actas do IX Encontro da Associação Portuguesa de Linguística*. Lisboa: APL/Colibri, 31-42.

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

MATEUS, M.H. & E. Andrade. 2000. *The Phonology of Portuguese*. Oxford: Oxford University Press.

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

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