FreP

Fernando Martins
ILTEC / Universidade de Lisboa
Marina Vigário
Universidade do Minho
Sónia Frota
Universidade de Lisboa

Manual

Last Update – January, 2006

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: classes of segments (consonants, vowels, glides, as well as subclasses), 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 syllable types (CV, V, CVC…), and does so 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, and their respective size. The tool also gives information on orthographic objects, namely, number of orthographic words and characters.

Some more general properties of *FreP* are the following: it belongs to the public domain (via download and password assignment), 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.

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 of programming and visual design of the tool was performed by Fernando Martins. 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 in Beta mode 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.

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 Widows XP, Windows Millenium, Windows 98, Windows Server 2003.

The tool opens non-formated, plain text files (.txt files or similar).

Installation

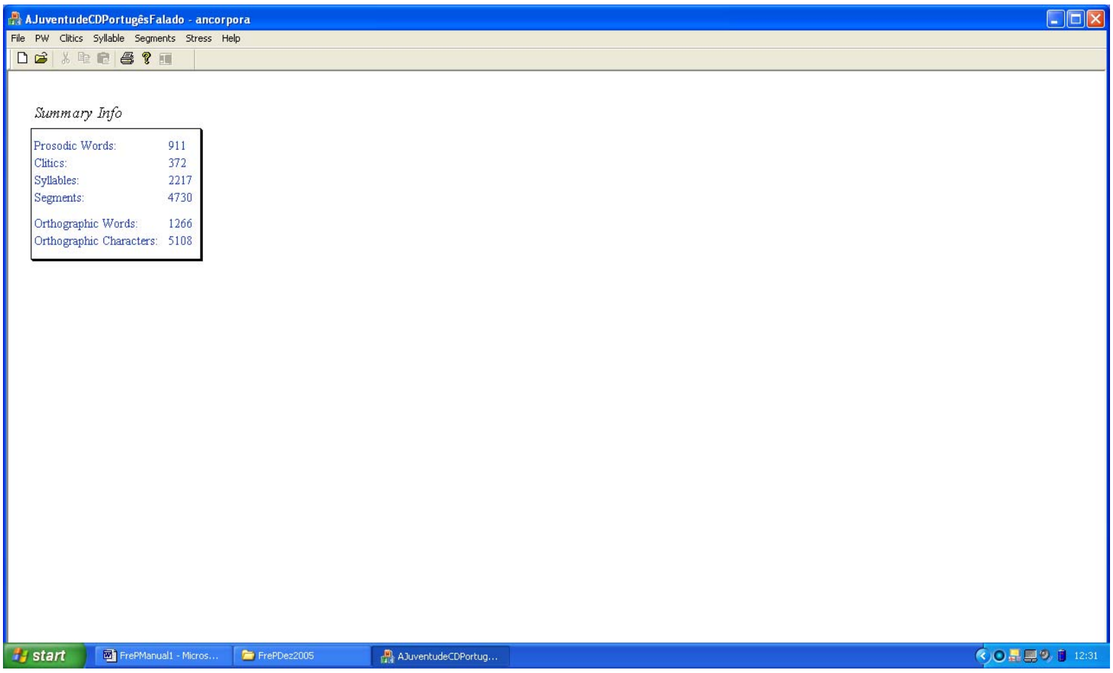
Unzipp the *FreP_setup* folder. Execute the program by clicking on the *FreP_setup* file, and follow the instructions on the screen. Use the password sent together with the program.

Loading a file

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

Map of the Tool

The following picture shows *FreP*’s Main Screen, after opening a file. The *Summay Info* is displayed at all time.



Summary Info

Prosodic Words:	911
Clitics:	372
Syllables:	2217
Segments:	4730
Orthographic Words:	1266
Orthographic Characters:	5108

The diagram below shows the options available from the Main Screen Menus, and their organization.

Main Screen Menus

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

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

Syllable Menu
- *Type*
- *by Position*
- *by Stress Condition*
- *by Position & Stress*

Segments
- *Type*

Stress
- *Location*

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* and *Orthographic Characters*. This box will be permanently available.

Main Screen Menus

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* ▸ *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 of Syllables by Syllable Type (e.g. V, CV, CVC,…). The list of all syllable types is available

- ▸ *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

- ▸ *Type*
  Provides the number of Segments by 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.

Stress
Provides frequency information on Stress location

- ▸ *Location*
  Provides the number of (prosodic) words as a function of Stress location (words with final, penult, antepenult Stress). Examples are also available by clicking on the *List* box that immediately follows each number

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’

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

The identification of syllable boundaries essentially follows the description and analyses proposed in Vigário & Falé (1994) and Mateus & Andrade (2000). In line with such work, glides between two vowels, as in *areia* ‘sand’, are ambisyllabic, and all 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 total number of V-slots in a given file is provided under the menu Segments. 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).

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

Available Options

[in progress]

FreP is being prepared to allow some optional settings. For example, for all operations involving Syllable Count it will be possible to choose to include or exclude the syllables obtained via V-slot insertion.

Limitations and Tips for Overcoming Them

Due to cases where phonology may not be predicted from the orthographic conventions, *FreP* inevitably yields some errors. These 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* (mono)$_{\text{PW}}$ (gamia)$_{\text{PW}}$ [ˈmɔnɔ gɐˈmiɐ] ‘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* (pe)$_{\text{PW}}$ (zinho)$_{\text{PW}}$ [ˈpɛ ˈziɲu] ‘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* (agil)$_{\text{PW}}$ (mente)$_{\text{PW}}$ [ˈaʒiɫ ˈmẽtɨ] ‘skillfully’). Rather frequent words such as *difícilmente* ‘hardly’ and *facilmente* ‘easily’ have been successfully excluded from this set of errors;

- Spelt consonants that are not pronounced are incorrectly parsed as existing consonants by *FreP* (e.g. ó**p**timo [ˈɔtimu] ‘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.

Extensions in Progress

Some of the extensions of *FreP* that are still in progress include the following

Segments:

Subclasses of segments (obstruents, sonorants, liquids, laterals, vibrants, some subclasses of vowels)

Combinations of segments by demand

Stress:

Number and list of PWs with Initial Stress

Features:

(At least) some segmental features (as many as possible): presence and frequency

The possibility of running *FreP* on files using SAMPA is a major goal for the near future. Besides the intrinsic value of such a possibility, it will also allow, for example, the extraction of all the information regarding phonological features, as well as to avoid the errors that *FreP* yields induced by orthography.

At this point, printing the information is only available through the Print Screen option on the keyboard. The possibility of creating and printing files with selected information is still in progress.

Work Already Available Using *FreP*

Papers

VIGÁRIO, Marina, Maria João Freitas & Sónia Frota (to appear) Grammar and frequency effects in the acquisition of the Prosodic Word in European Portuguese. *Language and Speech* (*Special Issue on the Acquisition of the Prosodic Word)*, guest-edited by Katherine Demuth).

VIGÁRIO, Marina, Fernando Martins & Sónia Frota (2005) Frequências no Português: a ferramenta *FreP*. In Inês Duarte & Isabel Leiria (eds.) *Actas do XX Encontro Nacional da Associação Portuguesa de Linguística*, 897-908.

Talks

VIGÁRIO, M., F. Martins & S. Frota (2005) A ferramenta *FreP* e a frequência de tipos silábicos e classes de segmentos no Português. Paper given at the *XXI Encontro Nacional da Associação Portuguesa de Linguística*, Porto, September 2005.

FREITAS, M. J., S. Frota, M. Vigário & F. Martins (2005) Efeitos prosódicos e efeitos de frequência no desenvolvimento silábico em Português Europeu. Paper given at the *XXI Encontro Nacional da Associação Portuguesa de Linguística*, Porto, September 2005.

FROTA, Sónia, Maria João Freitas, Marina Vigário & Fernando Martins (2005) Prosody and frequency effects on the development of syllable structure in European Portuguese. Paper given at the *Symposium "Exploring the Effects of Prosody, Morphology, Frequency and Representation on the Development of Syllable Structure in Romance Languages", Xth International Congress for the Study of Child Language*, Berlim, July 2005.

FREITAS, M.J., M. Vigário, & S. Frota (2004) The acquisition of the Prosodic Word in European Portuguese. Paper given at the *Second Lisbon Meeting on Language Acquisition*, Lisboa, June 2004.

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.