5th Workshop on Phonologica Development

Potsdam, March 31, 2006

FreP – Getting the frequency of phonological units from Portuguese written text. Application to studies on acquisition

M. Vigário[+] & S. Frota*

[+] *Universidade do Minho – ILCH and CEHUM*

**Universidade de Lisboa – FL and ONSET-CEL,*

Structure of the presentation

1. Background – the relevance of frequency information in linguistics (an overview)
2. *FreP* – an electronic tool for extracting frequency information of phonological units
   - → general description of the tool
   - → demo
3. Application to acquisition studies
4. Conclusion

1. Background: The relevance of frequency information in linguistics

Growing body of research devoted to or taking into account frequency issues

(e.g. Bybee, 2001; Bybee & Hopper, 2001a; Pierrehumbert, 2002; Moates, Bond & Stockmal, 2002, among many others…)

Some areas of phonology where frequency has been argued to play a role:

- ✓ in phonetic reduction (Jurafsky, Bell & Girand, 2002; Vigário, 2003)
- ✓ in the regularization of irregular paradigms (Bybee & Hopper, 2001b)
- ✓ in the activation of constraints (Thornton, 1996; Booij, 2005)
- ✓ in the emergence and development of grammatical units such as the syllable and the word (to mention but a few recent studies – e.g., Roark & Demuth, 2000; Beckman & Edwards, 2000; Lleó, 2001a, 2001b, 2004; Demuth & Johnson, 2003; Prieto, to appear)

……

2. *FreP*

Lack of frequency information of various phonological units and their distribution in Portuguese

FreP

an electronic tool designed to extract frequency information of phonological units at the level of the word and below from texts written in Portuguese

2. *FreP* – description of the tool

FreP

- from an ongoing joint project with Fernando Martins (Universidade de Lisboa / ILTEC), started in July 2004
- previous presentations of the tool:
  Vigário, Martins & Frota, 2005, submitted
  Frota, Vigário & Martins, to appear
- web site:
  http://www.fl.ul.pt/LaboratorioFonetica/FreP

Very general features

- extracts frequency information on phonological units:
  prosodic words, clitics, syllables, classes of segments, stress
- runs on written texts, following regular orthographic conventions
- optimized for the European Variety of Portuguese

2. *FreP* – description of the tool

Very general features (cont.)

- freely available to users for scientific, research, or teaching and learning non-commercial purposes
  (via e-mail contact)

- presently in Beta mode

- still in progress:
  - ✓ improving existing facilities
  - ✓ adding new facilities

How it works

a fairly predictable relation between Portuguese orthography and the (lexical) phonology of the language

+

the phonological knowledge of the obligatory segmental phenomena that characterize the language

=

identification and segmentation of phonological units

2. *FreP* – description of the tool

What it does

- automatically extracts (identifies and counts) the following phonological units:
  - classes of segments (consonants, vowels, glides; non-consonantal nasality + empty V positions)
  - syllables
  - phonological clitics
  - prosodic words (PWs)
- provides information on the size of prosodic words & clitics
  - in number of syllables
  - in number of segments
- locates word stress
- provides information on the distribution of stress within words (final, penult & antepenult stress)

2. *FreP* – description of the tool

What it does (cont.)

- provides information on the frequency of different syllable types (CV, V, CVC…)
  - by *position in the word* (initial, internal & final)
  - taking into account the *presence/absence of word stress*
  - taking into account the *position in the word & the presence/absence of stress*

- within the class of phonological clitics, sets *enclitics* and *proclitics* apart

  ⇨ frequency information for both types of units separately

- provides information on orthographic objects

  no. orthographic words
  no. characters

2. *FreP* – description of the tool

Phonological knowledge behind the computational system

Segments: e.g. Andrade & Kihm, 1988; Mateus & Andrade, 2000

Syllables: e.g. Andrade & Viana, 1994; Vigário e Falé, 1994; Mateus & Andrade, 2000

Prosodic Words: Vigário, 2003

Clitics: Vigário, 2003

Some relevant options made

- lexical phonology / obligatory phonological processes (e.g. Mateus, 1975; Mateus & Andrade, 2000; Vigário, 2003)

  Includes: e.g. glide insertion to break a hiatus; nasalization of vowels and glides; nasal diphthongs formation

  Excludes: e.g. optional semivocalization; optional deletions

2. *FreP* – description of the tool

Some relevant options made (cont.)

- [k^w] and [g^w] = phonologically labialized consonants (instead of C+G)
- glides between vowels are ambisyllabic (e.g. *areia* 'sand')
- syllables that do not conform to the general principles of syllable construction display a V-slot position (e.g. *obter* 'to obtain' > *o.bV.ter*)

Some limitations/types of errors

- digits are excluded from the analysis procedures
- acronyms and abbreviations are treated as regular words
- morphosyntactic compounds that have more than one word stress but are written as single orthographic words are computed as forming single prosodic words

2. *FreP* – description of the tool

Some implemented solutions

- manual intervention on the input file (e.g. deletion of consonants that are not pronounced; full writting of digits, abbreviations and acronyms)
- introduction of rewriting processes for phonological transparency (e.g. *somente>so mente>só mente; muito>muinto*)
- exclusion of specific sequences from the aplication of a particular rule

pezinho $(pe)_{\omega}(zinho)_{\omega}$
$>(pé)_{\omega}(zinho)_{\omega}$
$\neq$
vizinho $(vizinho)_{\omega}$
$//> (vi)_{\omega}(zinho)_{\omega}$

2. *FreP* – description of the tool

Evaluation of the tool

- A global evaluation of the tool is still in progress
- The automatic calculations generated by *FreP* have been checked by hand with regard to:
  - PW identification
  - phon.clitic identification
  - stress location
  - syllable count

based on two corpora (one from adult speech, with 22994 orthographic Ws, and the other one from child-directed, with 23674 orthographic Ws) – see Vigário, Martins & Frota, 2005; Vigário, Freitas & Frota, to appear

Reliability results:

99,935% for PW & clitic identification

99,930% for stress location

99,709% for syllable count

2. *FreP* – description of the tool

Evaluation of the tool (cont.)

For *syllable structure* and *major classes of segments* the automatic calculations have been checked on subsections of three corpora (one from adult speech, with 2796 ortographic Ws, one from child-directed speech corpus, with 4000 orthographic Ws, and one from child speech, with 1003 orthographic Ws (Frota et al, 2005)

Reliability results:

99,746% for syllable structure

99,746% for major classes of segments

2. *FreP* – description of the tool

Requirements for use

- ✓ takes as input non-formatted, plain text files (.txt files or similar)
- ✓ runs on Windows XP, Windows Millenium, Windows 98, Windows Server 2003

Information display

FreP provides two kinds of outputs:

- ✓ a menu based display (windows design)
- ✓ a formatted text file designed to be read by a spreadsheet or database utility

2. *FreP* – Demo

A note on the data used in this demo

1. Text composed of CHILDES orthographic lines
2. Because orthographic transcription in CHILDES must allow to recover the target, phonological categories (PW and clitics, major classes of segments, i.e. C, V, G, N) and phonological shapes (nº of syllables per PW/Cl; word stress location; syllable types, also by position in the word and status relative to stress) in child speech can be retrieved by *FreP*.

   The actual phonetic form of segments replaced by other segments is not retrieved, though, because it is not transparent from the orthographic form (e.g. "pato" when the child said ['katu]) – at this point, *FreP* does not treat specific segments, and thus this feature has no consequence on the operations performed

2. *FreP* – Demo

3. For *FreP* to run on the orthographic output of *CHILDES*, the text file has to be wiped off of all orthographic characters introduced as special markers (see below; and the *CHAT* manual for a complete list). Most of these characters are automatically removed when *FreP* reads the input text file: namely, punctuation marks and sequences within parentheses and square brackets. The remaining characters have to be removed by the user

   it is possible to implement the addition of these markers to the automatic wiping off function

   e.g. xxx/yyy (unintelligible speech); @ (all @ markers)

   # (pauses); +< (overlaps); < > (all angle brackets markers);

   + (compounds)

3. Application to acquisition studies

Two areas of research on acquisition

1. Vigário, Freitas & Frota (to appear) used *FreP* to extract prosodic word shape frequencies (i.e. frequency of PW with different sizes) both on an adult speech and a child-directed speech corpus

> As seen in this morning's presentation, it was found that the statistical properties of the input do constrain the shape of early words in Portuguese

2. Frota et al (2005) & Freitas et al (submitted) investigate the relevance of syllable type frequency in the input for the emergence of syllable types in child speech

3. Application to acquisition studies

Frota, Freitas, Vigário & Martins (2005)
Freitas, Frota, Vigário & Martins (submitted)

FreP was used to obtain overall frequency of

- ✓ syllable types
- ✓ syllable type distribution by *position within the word* and by *position relative to stress* (e.g. stressed/unstressed)

in different corpora

- ✓ adult speech (AS)
- ✓ child directed speech (CDS)
- ✓ child speech (CS)

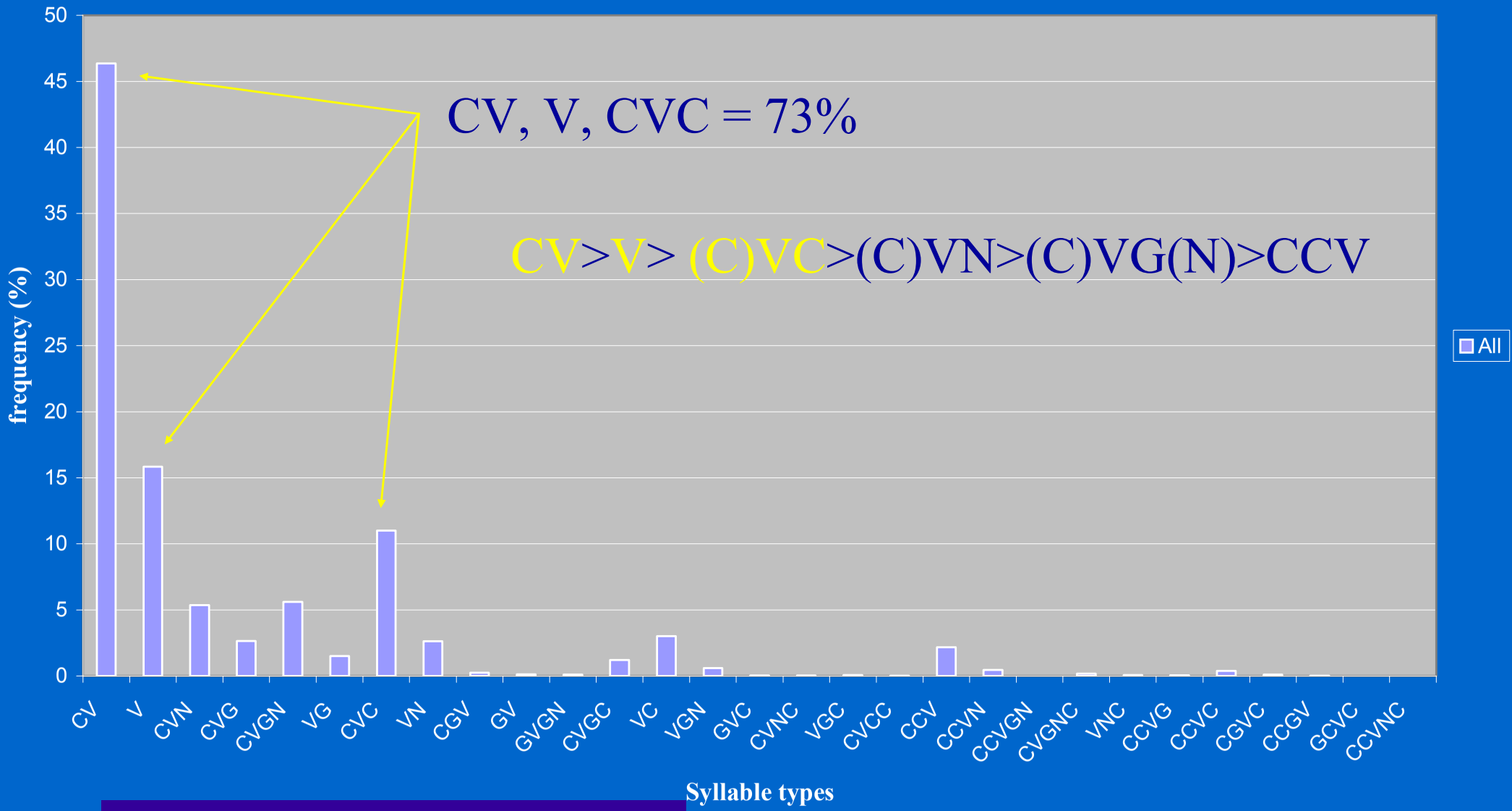
It was found that...

The **overall** input frequency is **not** a good predictor of the order of emergence of syllable types

- ➢ overall frequency predicts the order CV > V > CVC > CVG/N; ≥
- ➢ the development pattern in child speech displays both CV and V at the beginning, and CVG/N before CVC ≥≥

3. Application to acquisition studies

(from Vigário et al, submitted) **ADS**



3. Application to acquisition studies

The two mismatches (AS, CDS *vs.* CS) can be explained if the distribution of syllable types by position within the word & by position relative to stress is taken into account

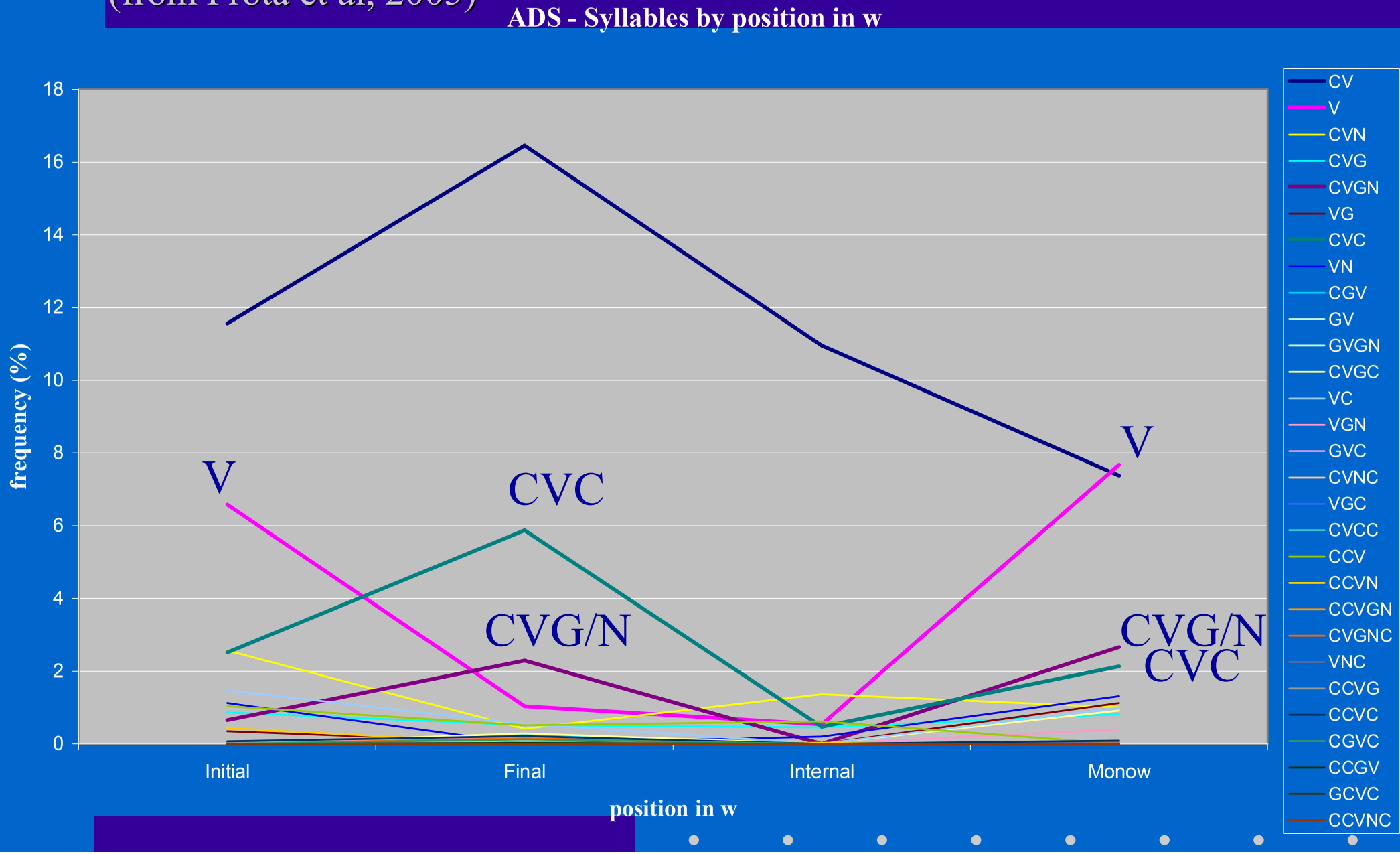
V and CVG/N are syllable types that *appear mainly in prominent* positions in the input (i.e. at word-edges and/or in stressed position)

Syllable diversity & complexity:

Final>Monow>Initial>Internal

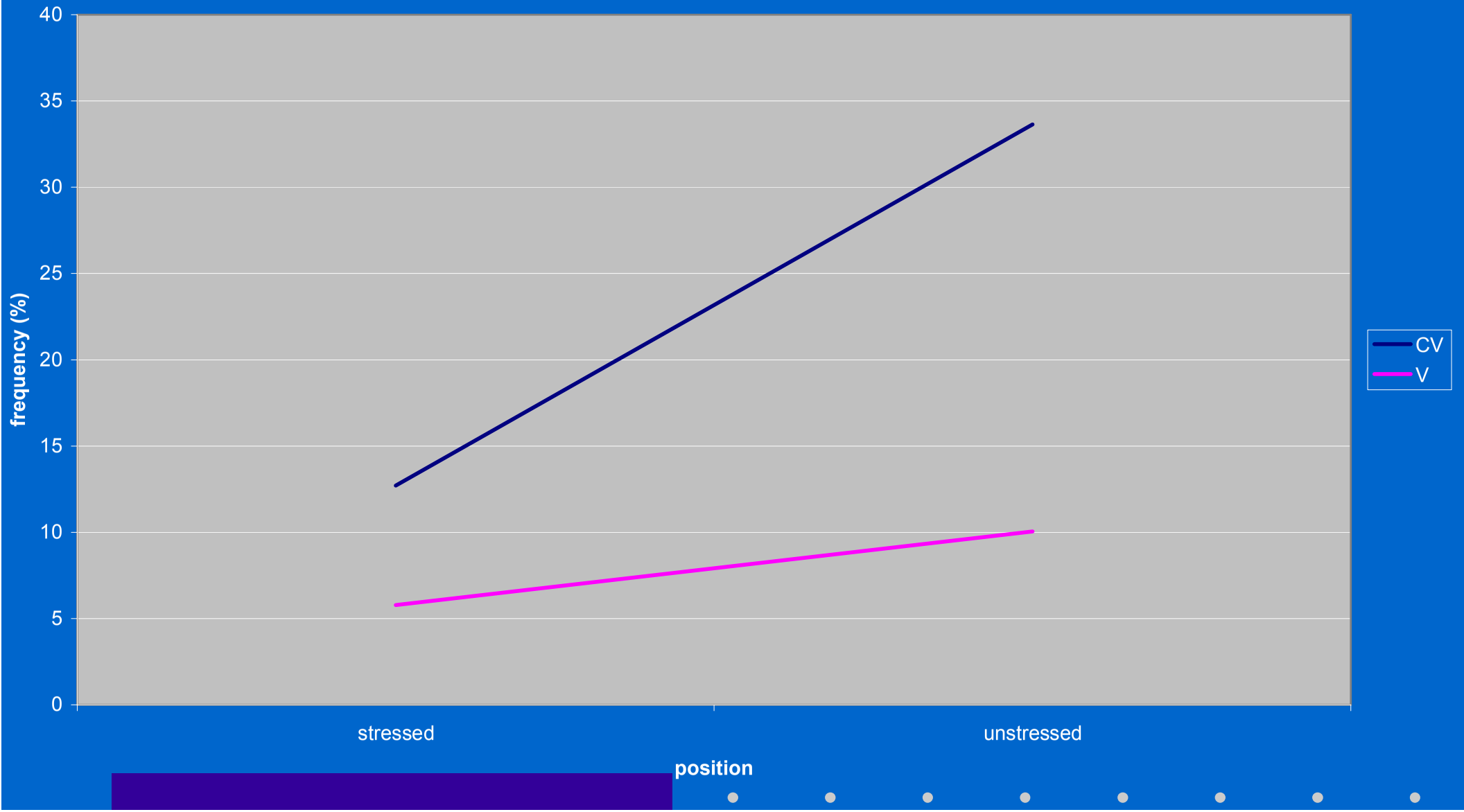
3. Input: word-edges vs. internal position

(from Frota et al, 2005)



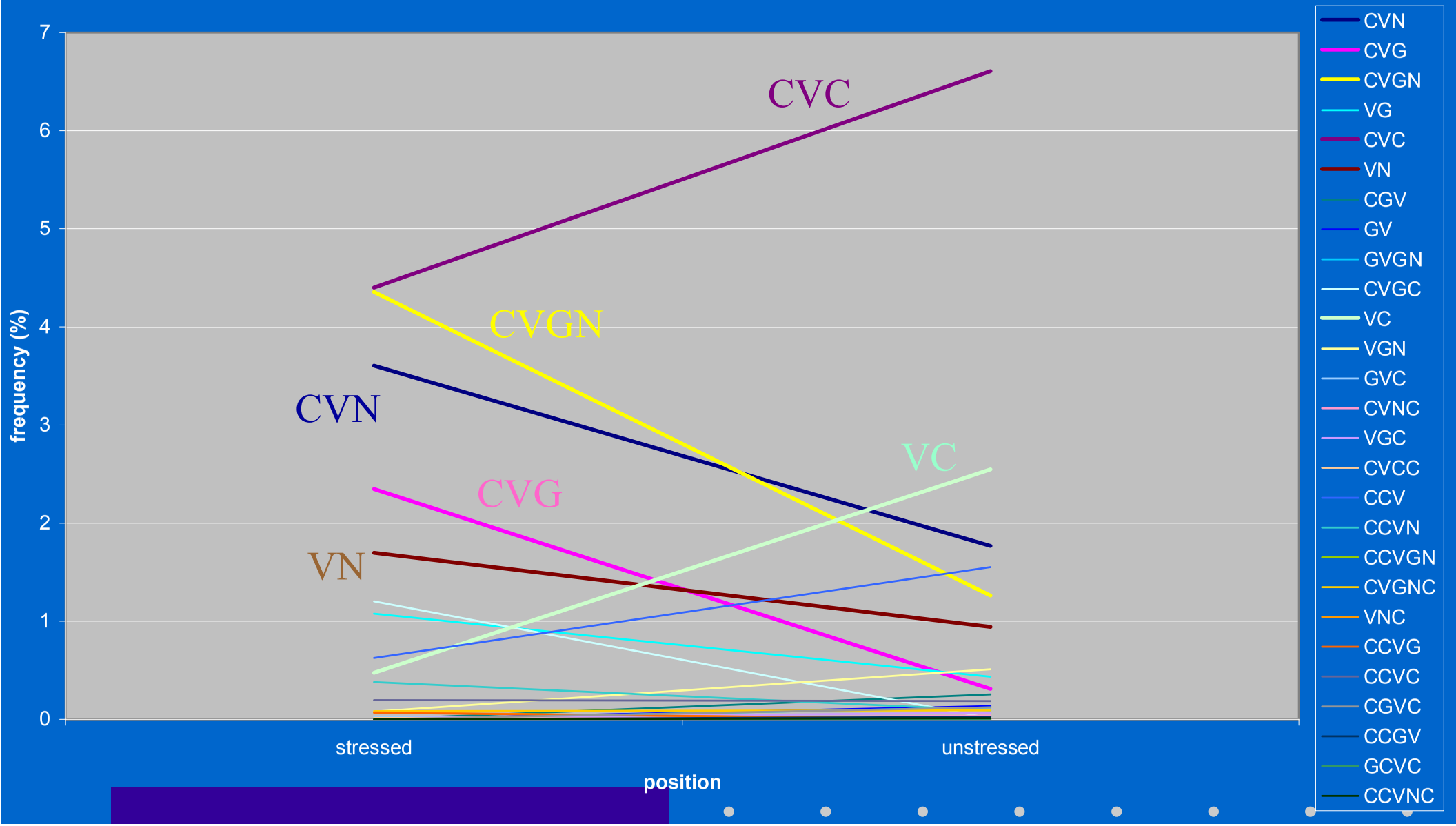
3. Word stress: CV, V

(from Frota et al, 2005)



3. Word stress: CVN, CVG

(from Frota et al, 2005) **ADS - The effect of stress**



3. Application to acquisition studies

Hypothesis entertained:

frequency

&

structural information

(i.e. prosodic prominence)

↓

order of emergence of syllable types in European Portuguese

4. Conclusion

Areas of application of FreP

- ✓ studies on phonological development (... + Costa, in progress - the impact of frequency in the development of phonological features)
- ✓ studies on adult grammar (Vigário et al, 2005; submitted; Frota et al, to appear)
- ✓ studies on speech evaluation and therapy (Guerreiro, in progress; Silva, in progress; Correia, in progress)
- ✓ for teaching/learning purposes

…

4. Conclusion

Planned extensions include:

- ✓ the ability to extract frequency information for different varieties of Portuguese, namely Brazilian Portuguese
- ✓ the ability to provide a SAMPA output from the written text
- ✓ the frequency count of segmental features, like manner, place of articulation, or laryngeal features

…

Final remark: Languages with alphabetic writing systems, and thus a grapheme – phoneme relation, can in principle take advantage of the *FreP* concept

Work done resorting to FreP – selected references

Freitas, M. J., Frota, S., Freitas, M. J., Vigário, M. & Martins, F. (submitted) *Efeitos prosódicos e efeitos de frequência no desenvolvimento silábico em Português Europeu*

Frota, S., Freitas, M. J., Vigário, M. & Martins, F. (2005) Prosody and frequency effects on the development of syllable structure in European Portuguese. Paper given at the *Xth International Congress for the Study of Child Language*, Berlin.

Vigário, M., Freitas, M. J. & Frota, S. (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 K. Demuth).

Vigário, M., Martins, F. & Frota, S. (2005) Frequências no Português: a ferramenta *FreP*. In *Actas do XX Encontro Nacional da Associação Portuguesa de Linguística* (pp. 897—908). Lisboa: Colibri/APL.

Vigário, M., Martins, F. & Frota, S. (submitted) *A ferramenta FreP e a frequência de tipos silábicos e classes de segmentos no Português*.

Thank you!

Acknowledgements:

Our collaborators

Fernando Martins (> *FreP* construction)

João Freitas (> application to acquisition studies)

Ana Lúcia Santos (> child speech data)

Psycholinguistics Laboratory – University of Lisbon