The frequency that counts in language acquisition: types or tokens?

Marina Vigário, Sónia Frota & Fernando Martins
Laboratório de Fonética – CLUL
Universidade de Lisboa

Workshop on Frequency i n Phonology
26th October, 2010
Faculdade de Letras da Universidade de Lisboa

Introduction

- **Frequency in the input**
  - Recent research has shown that the frequency of units/patterns in the input plays an important role in language acquisition
  - In the acquisition of phonology, a number of studies point to a correlation between frequency of patterns/units in the input and the order or emergence and/or the frequency of those units/patterns in children's productions (Ingram 1988, Roark & Demuth 2000, Demuth & Johnson 2003, Levelt & Van de Vijver 2004, the papers in Demuth 2006, a.o.)

Introduction

- **What frequency?**
  - Computing the frequency in the input: over the lexicon (list of unique words – *types*) or over the occurence of every instance of words (*tokens*) > to our knowledge, this issue has not been previously addressed in a systematic way in he literature on language acquisition and development
  - Studies vary: data from dictionaries/lexica (types, e.g., Monin, Loevenbruck & Beckman 2007), or based on corpora (tokens, e.g., Prieto 2006); very few compare types and tokens (Marchman & Plunkett 1989)

Introduction

- **Types and tokens crosslinguistically**
  - Type and token frequency values in a single language may not converge, and not all units/patterns vary in the same way (e.g., Leung & Law 2004, Cantonês de Hong Kong; Ota 2006, Japonês)
  - In EP, the frequency data is very scarse in the literature (cf. Viana et al. 1996). The case of subminimal words:

    0,4% (Vigário 2003, **type**)

    7% (Vigário, Frota & Martins 2006, **token**)

    In children's productins, these words appear very early and remain frequent after the initial stage (like in French but not English – diffs. attributed to diff. frequency in the input: Demuth & Johnson 2003, Vigário, Freitas & Frota 2006) > ***token*** frequency

Introduction

- Goals

- Contribute to the understanding of the role of type and token frequency in the acquisition of several aspects of phonology

> Debate on the role of frequency and on the theories of acquisition base on UG vs statistical computation

- Contribute to the knowledge on the frequency of various phonological units/patterns in types and tokens in EP

- Structure

1. Method
2. Results
   - Word shapes
   - Stress patterns
   - Syllable shapes
     Position within the word & stress
   - Segments
     Position within the syllable
   - Point of articulation
     Position within the word & stress
3. Discussion and conclusios

Methodod

- Input

Corpora (part of FrePOP)

ca. 245.000 words

- spontaneous speech (portions of several corpora: *Português Falado* anos 90, CLUL; CRPC, CLUL; CORP-ORAL, ILTEC; C-ORAL-ROM, CLUL)

Tokens: 240.767	Types: 16.702
PWs: 173.355	Clitics: 72.525
Syl: 447.331	Seg: 933.411

FreP (Martins, Vigário & Frota 2009, v.2)
Frequency values in tokens and types; units/patterns from the word-level to features

- Child speech data

Available in the literature:

. Word shapes
(Vigário et al. 2006)

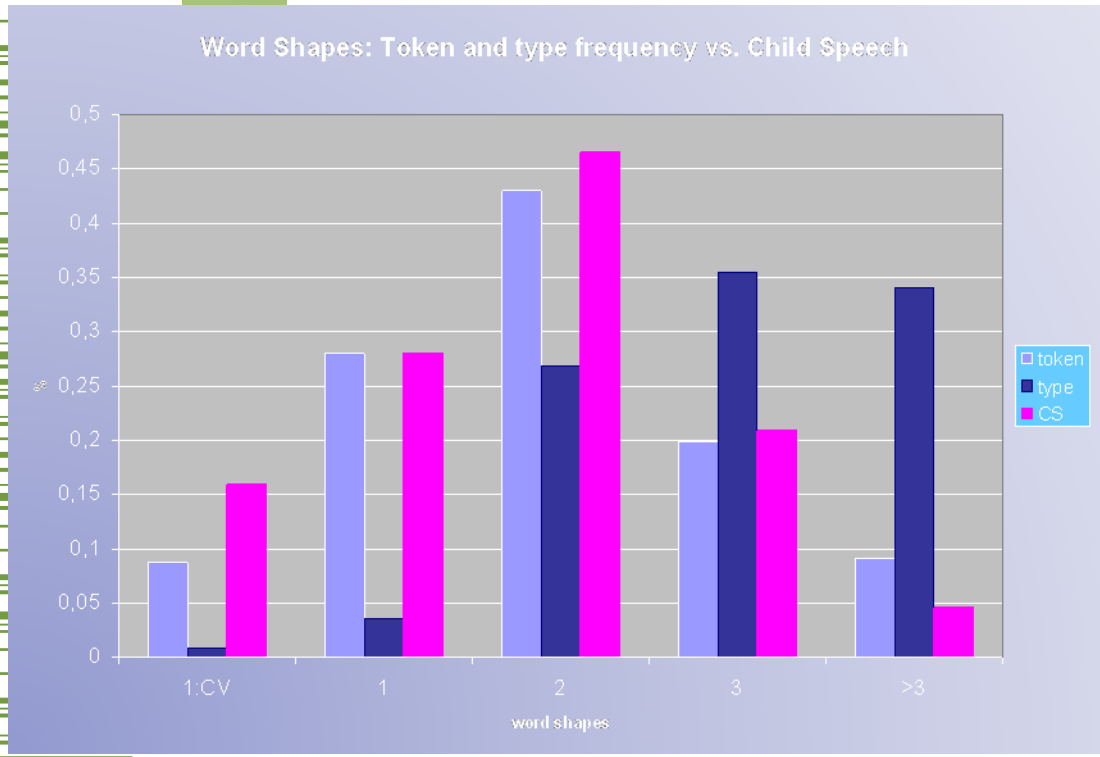
. Sress pattern
(Frota & Vigário 2008, Correia 2008)

. Syllable shapes
(Frota et al. 2005, Freitas et al. 2006)

. Segments in the syllable
(Freitas 1997, Jordão 2009)

. Point of articulation
(Costa et al. 2007)

Results: word shapes



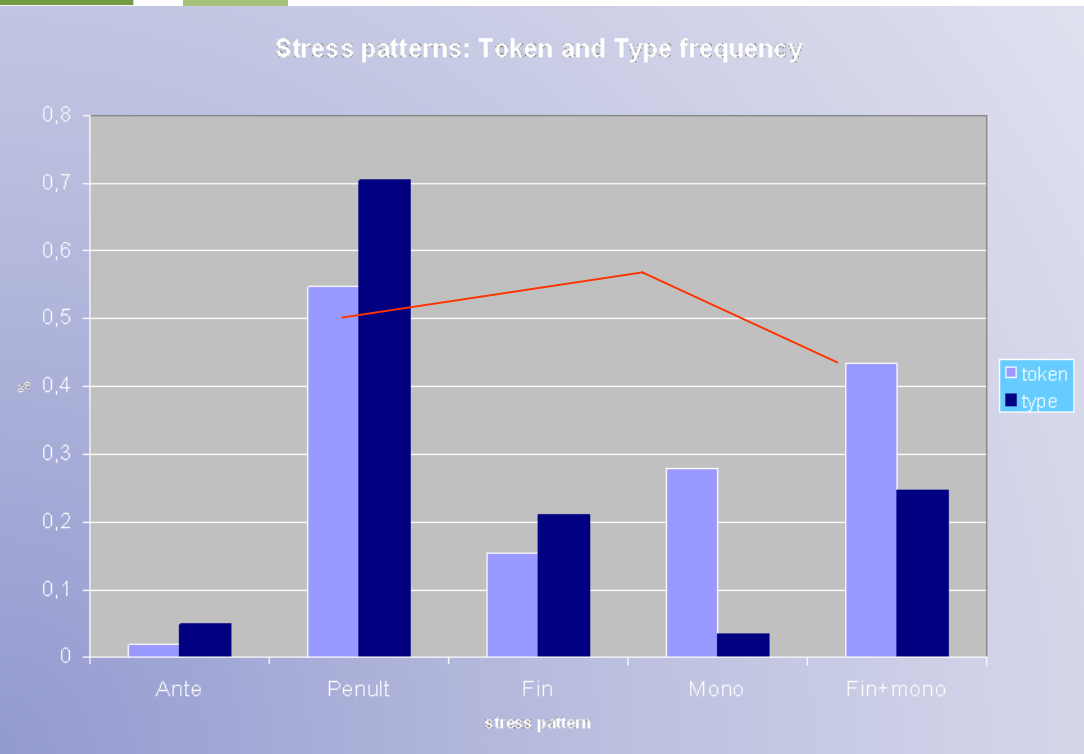
Difference tokens/types:

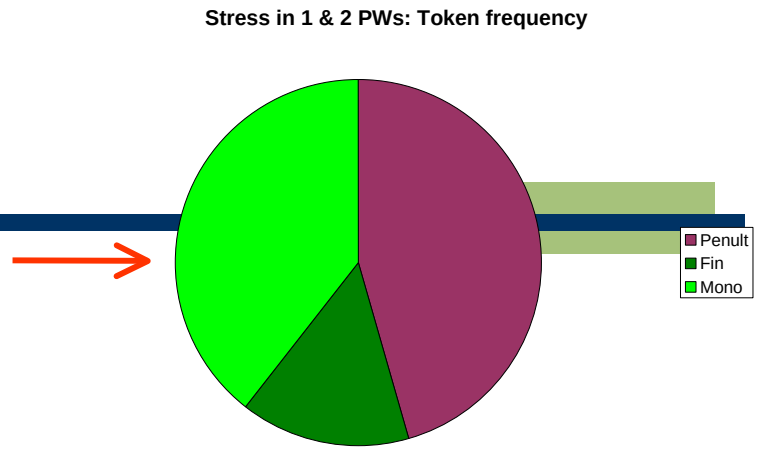
- In all word shapes
- <3, token > type
- =>3, token < type
- 2/3 of types have 3 or more syllables, but the most frequent words have 1 or 2 syllables
- Monosyllables and subminimal words are very rare in types

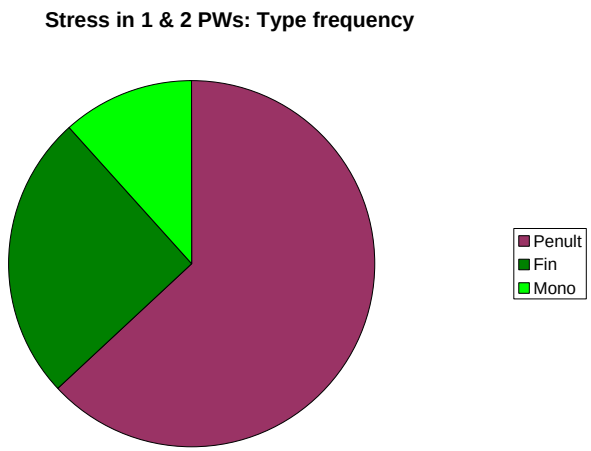
→ It is the frequency computed over **tokens** that correlates best with the data from child speech (CS)

Token results similar to those reported for the input in Vigário et al. (2006), now computed on a corpus 10x larger

Results: stress patterns



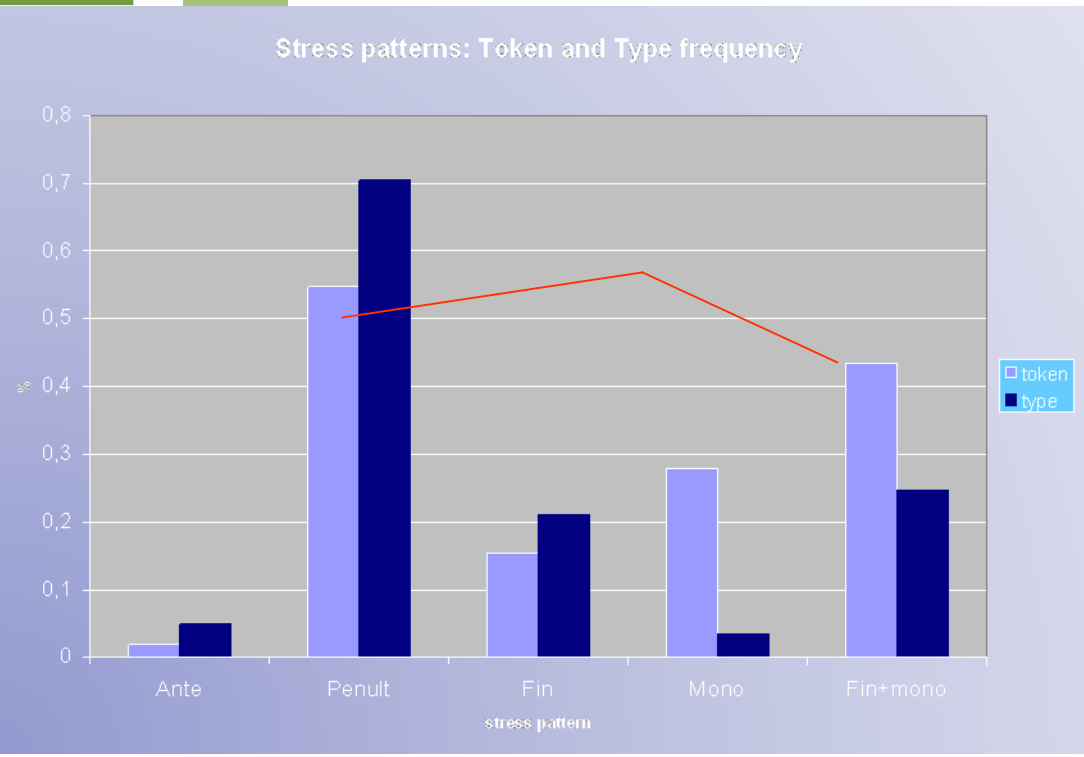


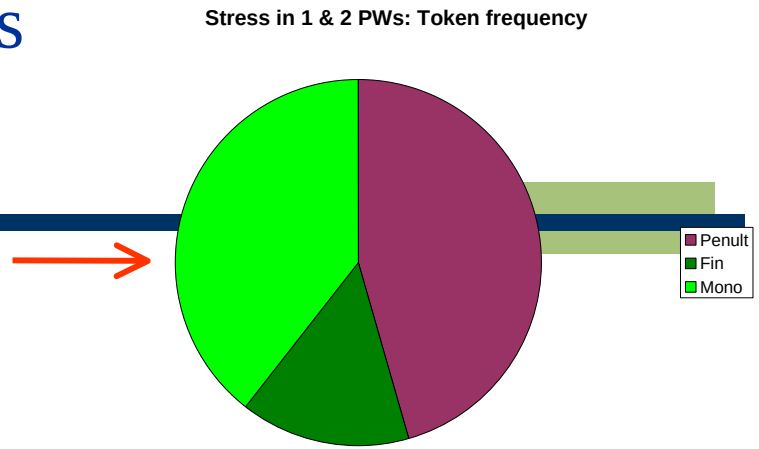


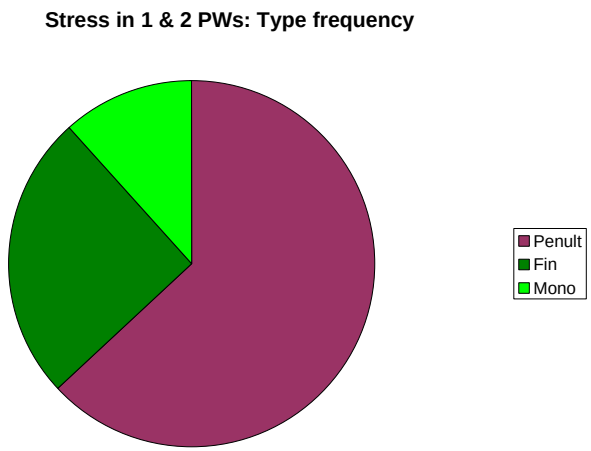
CS: initial difficulty in producing the stress pattern of the target (Correia 2008, Frota & Vigário 2008); stress matching the target: earlier in words with final stress than with penultimate stress (Frota & Vigário 2008)

- Only **token** frequency may explain CS data

Results: stress patterns



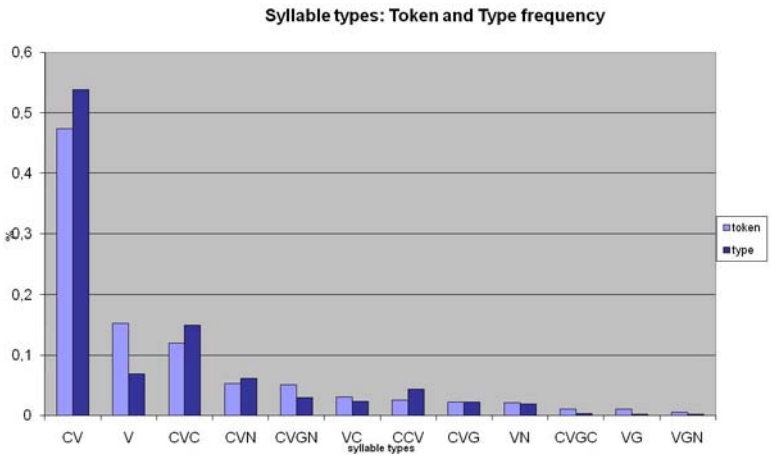


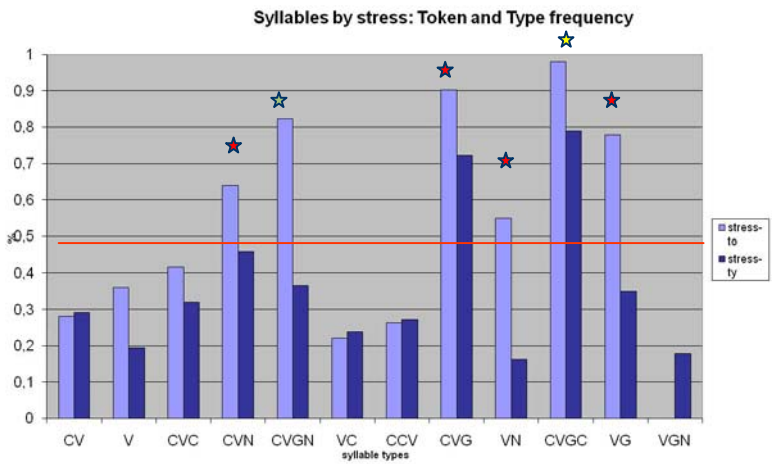


Crucial role of 1σ PWs > pattern like PW final (Frota et al. 2006): complexity of syllable types in monosyllabic PW ~ PW final; proclitics+PWmono yield final stress: final stress>22% (Vigário et al. 2010)

- Only **token** frequency may explain CS data

Results: syllable shapes





Token: CV>**V**>CVC>CVN>CVGN>CCV

Type: CV>CVC>CVN>**V**>CCV>CVGN

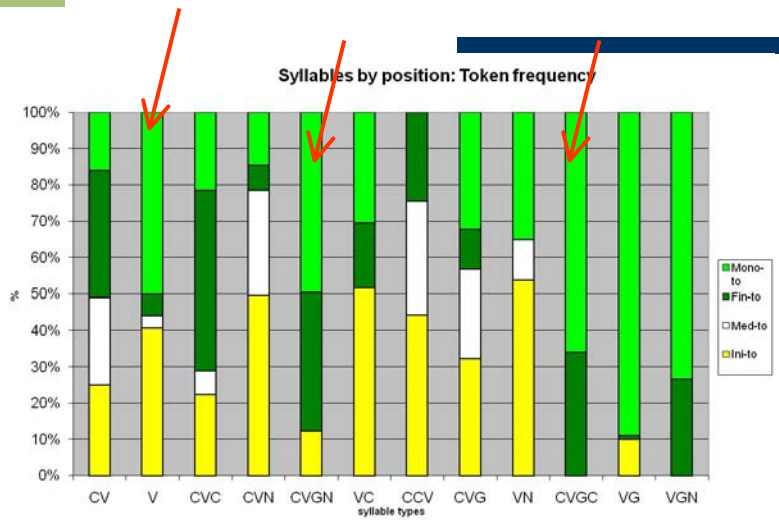
CS: CV, V > (C)VN > (C)VG > (C)VC

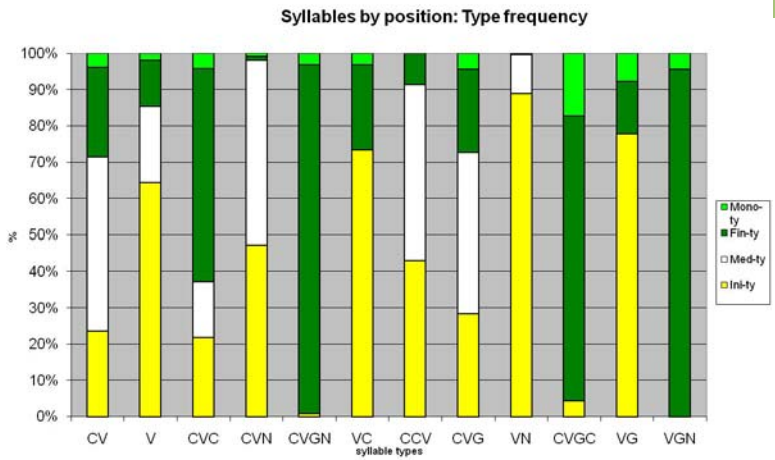
Token closer to CS
but CV,V and CVN/G before CVC?

Frota et al. (2005) and Freitas et al. (2006) explain the emergence of V and CVN/G by the interaction of frequency (only in **tokens**) and prominent positions

Stress factor promotes (C)VG and (C)VN if tokens are considered

Results: syllable shapes

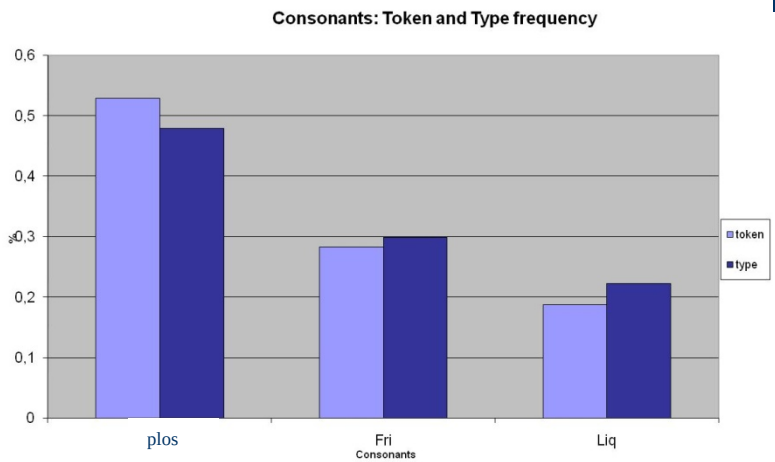


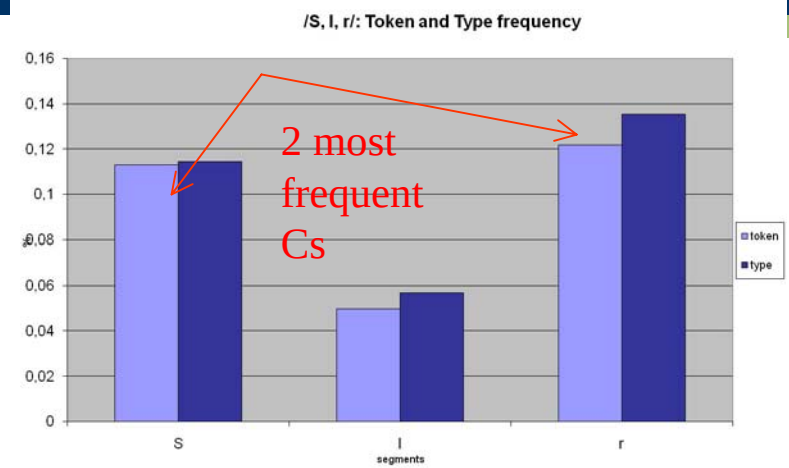


PW edges as prominent positions (Vigário 2003):
90% of V occur in initial position (including proclitics) > tokens
50%-60% of CVG(N/C) syllables occur in stressed monosyllables > tokens
[against only 20% of CVC in both cases]

Tokens frequency in prosodically strong positions predicts the emergence pattern in CS (this confirms and extends previous results)

Results: segments





Token and type frequence show the same distribution

CS: plosives > fricatives, liquids (non-branching onsets – Freitas 1997)

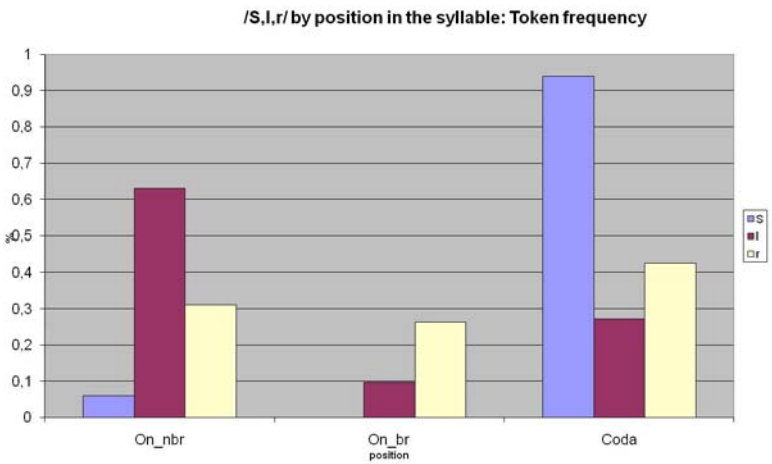
This is the order of emergence predicted by frequency (token and type)

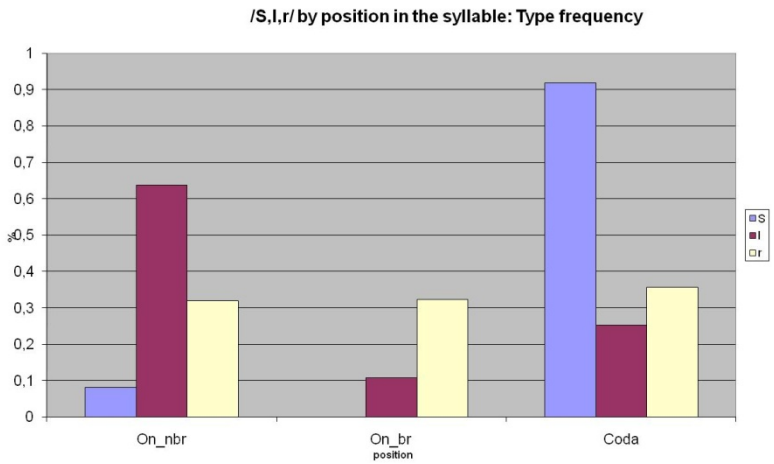
CS: /ʃ, l, ɾ/ in onset: S > l, r (Jordão 2009)
in coda: S > l > r (Freitas 1997)
S > l,r (Jordão 2009)

Not predicted by frequency!

Results: distribution of segments in the syllable

Distribution of /ʃ, l, ɾ/ in the syllable: token and type show the same pattern





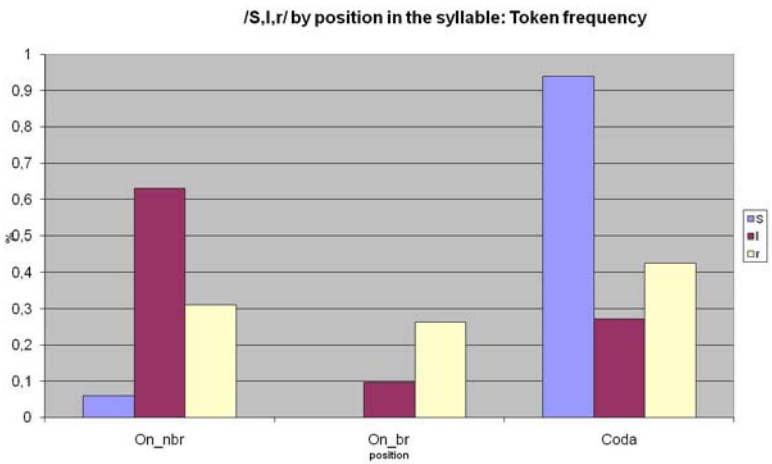
Frequency data do not correlate with order of emergence in onset (l>r>S vs. S>l,r); prediction based on frequency for order of emergence in coda S>r>l ou S>l,r , parcially borneout (S>l>r or S>l,r)

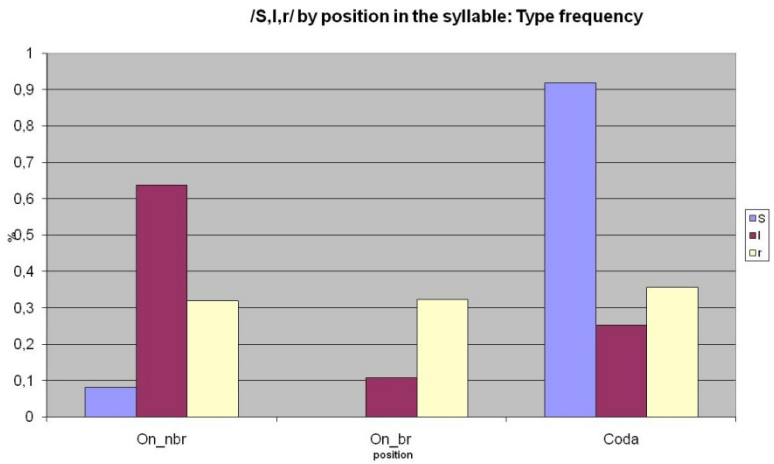
The case of /ʃ/: the 2nd most frequent C in the language occurs in coda >90%; but emerges first in onset (Freitas 1997, Jordão 2009)

The case of /ɾ/: the most frequent C in the languade occurs ~evenly in the 3 positions; but order of emergence: onset>coda>branching onset (Freitas 1997, Jordão 2009)

Results: distribution of segments in the syllable

Distribution of /ʃ, l, ɾ/ in the syllable: token and type show the same pattern





Non-branching Onset: l>r>S vs. S>l, r

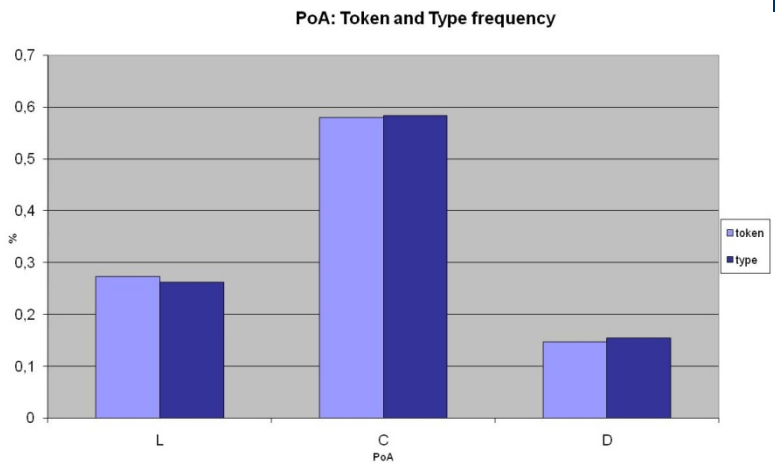
Coda: S>r>l ou S>l,r vs. S

The case of /ʃ/: >90% coda v

The case of /ɾ/: 3 positions v nset

Interaction between structure and frequency, to be explored

Results: point of articulation



Token and type show the same distribution:

C > L > D (L and D very close)

- **CS**

Stage where initial position becomes autonomous: Labial and Dorsal (vs. *Dorsal in Dutch, attributed to the distribution in the input)

- Frequency in production:

Initial position: L>C>D

Medial and final: C>L>D or C>L,D

- Frequency in targets:

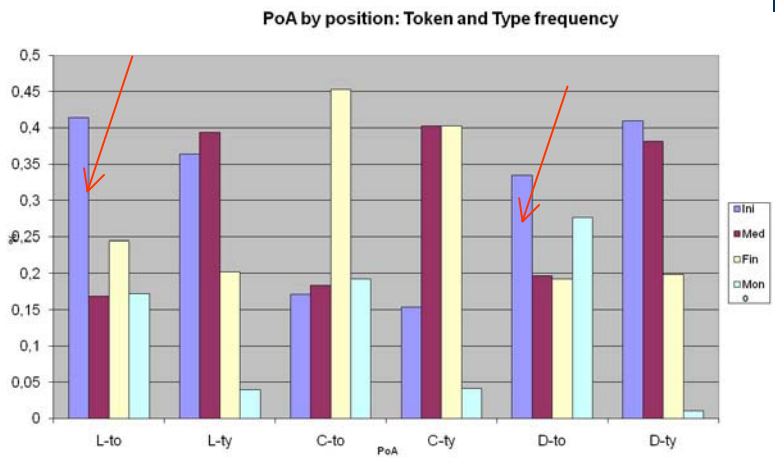
Initial position: L>C,D

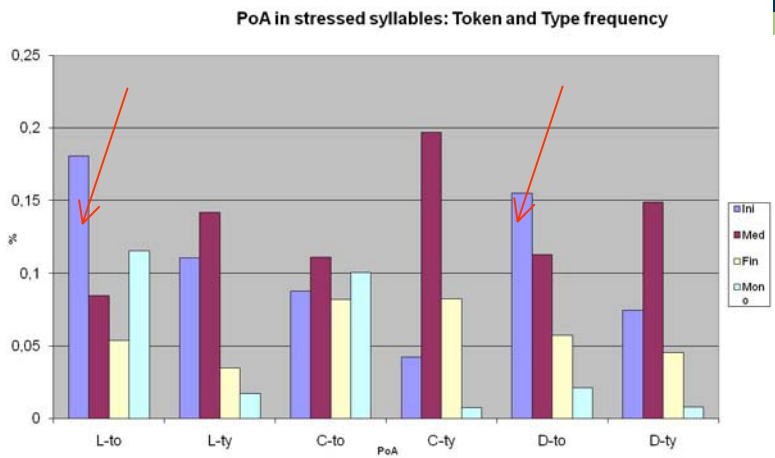
Medial position: C>L>D or C>L,D

Final position: C>L>D

(Costa et al. 2007)

Results: point of articulation





Different distribution of L and D vs. C according to position; differences betweem tokens and types

L and D occur mostly in initial position in tokens (vs. C): L>D>C

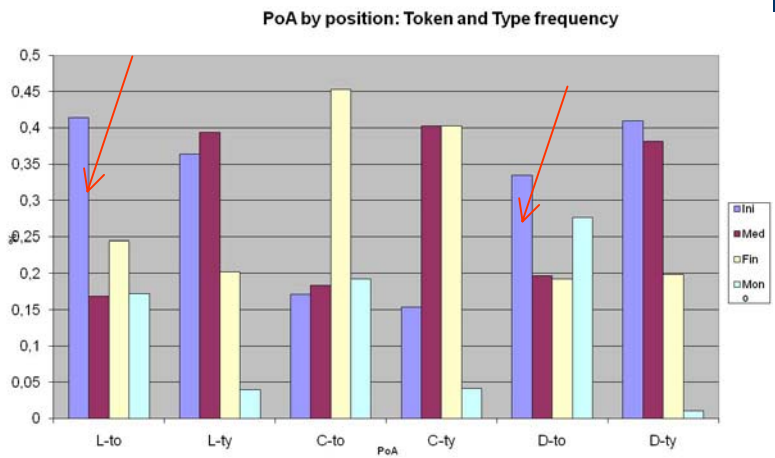
D dominates in monosyllables

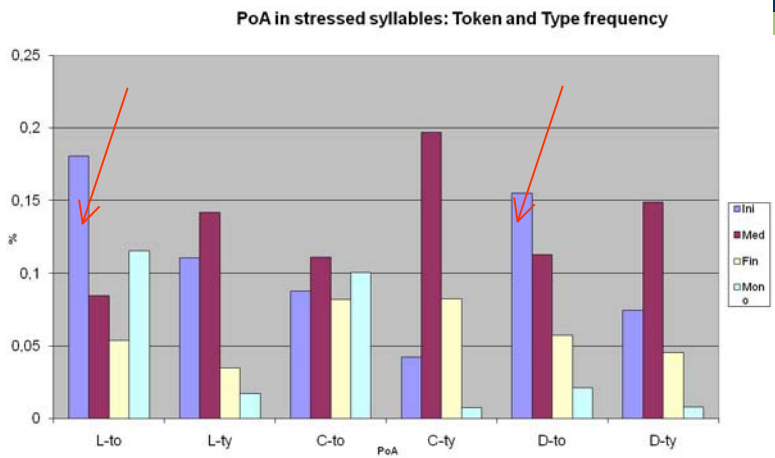
Different distribution of L and D vs. C according to stress+position; Large differences token / type

L and D stand out in initial stressed position (token)

L and D occur in prominent positions: $_{PW}$[_ and $_{PW}$[+<u>stress</u>

Results: point of articulation





L and D stand out in initial position in token frequency (occuring mostly in this position vs. C):
L>D>C
D dominates in monosyllables

L and D stand out in initial stressed position (token)

L and D occur in prominent positions:
$_{PW}$[_ and $_{PW}$[+<u>stress</u>

Distribution in the input (**token** only) predicts the autonomization of L and D in initial position, L being more frequent than D (this confirms and extends the results of the input in Costa et al. 2007)

Discusion and conclusions

Frequency: types or **tokens**?

Frequecy in the input is a good predictor of CS data for

- word shapes (emergence and frequency in production)
- stress pattern (emergence and evolution)
- syllable formats (emergence),
- major classes of segments (emergence)
- segments in coda (emergence)
- the acquisition of PoA

- Whenever types and tokens do not converge, **tokens** are the preditors **:**
  - word shapes
  - stress
  - syllable formats
  - point of articulation

Type and token frequency coincide in the classes of segments and segments in coda

Token frequency is the one relevant in acquisition

Discusion and conclusions

- Frequency and structure

✓ For various aspects of phonology, the frequency data that are relevant are those computed over tokens and taking the structure of the language into consideration, i.e., prosodically proeminent positions: syllable formats and PoA

Proeminent positions stand out in perception and favour the retrieval of the elements that occur there frequently

- Language specificities

✓ The role of **token** frequency highlights the importance of the distribution of the units and patterns that are *indeed presente* in input, which may vary crosslinguistically

✓ It highlights the relevance of the ***use*** of the language, and of studying the possible *different inputs* within a single language to which different children may be exposed

(>different paths in acquisition?)

→ methodological implications

Discusion and conclusions

- Other implications

The relevance of **token** frequency points in the direction of recent proposals on the representation of phonological knowledge (*exemplar-based models*) and the process of language acquisition (*usage-based account of language learning)* (e.g., Pierrehumbert 2001, Bybee 2003, Bybee & McClelland 2005), in its moderate version: interaction with structure

- To be explored

- Other interactions between frequency and structure (e.g., syllable structure and the emergence of segments)
- Systematic analysis of type and token frequency for units and patterns of EP phonology in different corpora (FrePOP: 3.500.000 words + FreP functionalities)
- Predictions of patterns in the acquisition of the language

Acknowledgments

- Research assistents of FreP project:

  Marisa Cruz e Nuno Matos

- Mary Beckman, who has first questioned us on which frequency data, token or type, may be relevant for the computation of word size distributions

  →freq. monomoraic PW~Minimal Word Constraint

- All those who have made corpora available for the frequency analysis, with special reference to Fernanda Bacelar e Tiago Freitas

Projecto FreP: Padrões de Frequência na Fonologia do Português - Investigação e Aplicações PTDC/LIN/70367/2006

www.fl.ul.pt/LaboratorioFonetica/frep