4th International Conference on Statistical Language and Speech Processing
SLSP 2016

Pilsen, Czech Republic
October 12th, 2016

TESTING THE ROBUSTNESS OF LAWS OF POLYSEMY AND BREVITY VERSUS FREQUENCY

Antoni Hernández Fernández , Bernardino Casas, Ramon Ferrer-i-Cancho and Jaume Baixeries

antonio.hernandez@upc.edu

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Complexity and Quantitative Linguistics Lab,
Laboratory for Relational Algorithmics, Complexity and Learning (LARCA),
Departament de Ciències de la Computació /
Institut de Ciències de l'Educació

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

OUR GROUP

Complexity and Quantitative Linguistics Lab,
Laboratory for Relational Algorithmics, Complexity and Learning (LARCA),
Departament de Ciències de la Computació /Institut de Ciències de l'Educació

https://recerca.upc.edu/larca/en
https://www.cs.upc.edu/~cqllab/

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

PROJECT ON THE EVOLUTION OF CHILD LANGUAGE

- Baixeries, J., Elvevåg, B. & Ferrer-i-Cancho, R. (2013). **The evolution of the exponent of Zipf's law in language ontogeny.** *PLoS ONE* 8 (3), e53227. [doi: 10.1371/journal.pone.0053227]
- Casas, B., Català, N., Ferrer-i-Cancho, R. & Baixeries, J. (2014). **The evolution of polysemy in child language**. In: *THE EVOLUTION OF LANGUAGE*, *Proceedings of the 10th International Conference (EVOLANG10),* Cartmill, E. A., Roberts, S., Lyn, H. & Cornish, H. (eds.). Evolution of Language Conference 2014. Vienna (Austria), pp. 409-410. [doi: 10.1142/9789814603638_0068]
- Casas, B., Català, N., Ferrer-i-Cancho, R., Hernández-Fernández, A. & Baixeries, J. (2016). **The polysemy of the words that children learn. (submitted)**
-Ferrer-i-Cancho, R. (2016). **The optimality of attaching unlinked labels to unlinked meanings.** Glottometrics, in press. https://arxiv.org/abs/1310.5884

This conference:

Hernández-Fernández, A., Casas, B., Ferrer-i-Cancho, R. & Baixeries, J. (2016). **Testing the robustness of laws of polysemy and brevity versus frequency.** *4th International Conference on Statistical Language and Speech Processing (SLSP 2016), Pilsen (Czech Republic).* P. Král and C. Martín-Vide (eds.). LNAI 9918, pp. 1–11. [doi: 10.1007/978-3-319-45925-7_2]

SUMMARY

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

- Introduction and objectives
- Materials
- Methods
- Results
- Discussion and future work

ZIPF'S STATISTICAL LAWS OF LANGUAGE

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

- **Introduction**
- Materials
- Methods
- Results
- Discussion

Zipf's law for word frequencies (Zipf 1935, 1949): the frequency of the *i-th most* frequent word in a text follows approximately:

$$f \propto i^{-\alpha}$$

where ***f*** *is the frequency of that word,* ***i*** *their rank or order and* ***α*** *is the exponent* ($\alpha \approx 1$).

Meaning-frequency law (Zipf 1945): the tendency of more frequent words to be more polysemous.

- **Later work shows correlation between frequency and number of synsets in adults L1** (Baayen & Moscoso del Prado Martín 2005; Ilgen & Karaoglan 2007) **and in adults L2 learners** (Crossley et al, 2010).

Zipf's law of abbreviation *or* brevity law (Zipf 1935): the tendency of more frequent words to be shorter or smaller.

Hernández-Fernández et al (2016)

OBJECTIVES

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

- **Introduction**
- Materials
- Methods
- Results
- Discussion

Preliminary Study:

Test statistically the robustness of these linguistic laws in English
Content words (N, V, Adj, Adv)
Different types of speakers (Children vs Adults)
Oral vs Written texts

+

Four different sources of estimation:
CELEX lexical database
WordNet
CHILDES database
SemCor corpus

Hernández-Fernández et al (2016)

MATERIALS:
DATABASES AND STUDY UNITS

- **Word Frequency (tokens)**
  - **CHILDES Database: Four types of speakers (CHI, MOT, FAT, INV).**
  - **SemCor Corpus. http://multisemcor.fbk.eu/semcor.php**
  - **CELEX2 Lexical Database (only English). https://catalog.ldc.upenn.edu/LDC96L14**

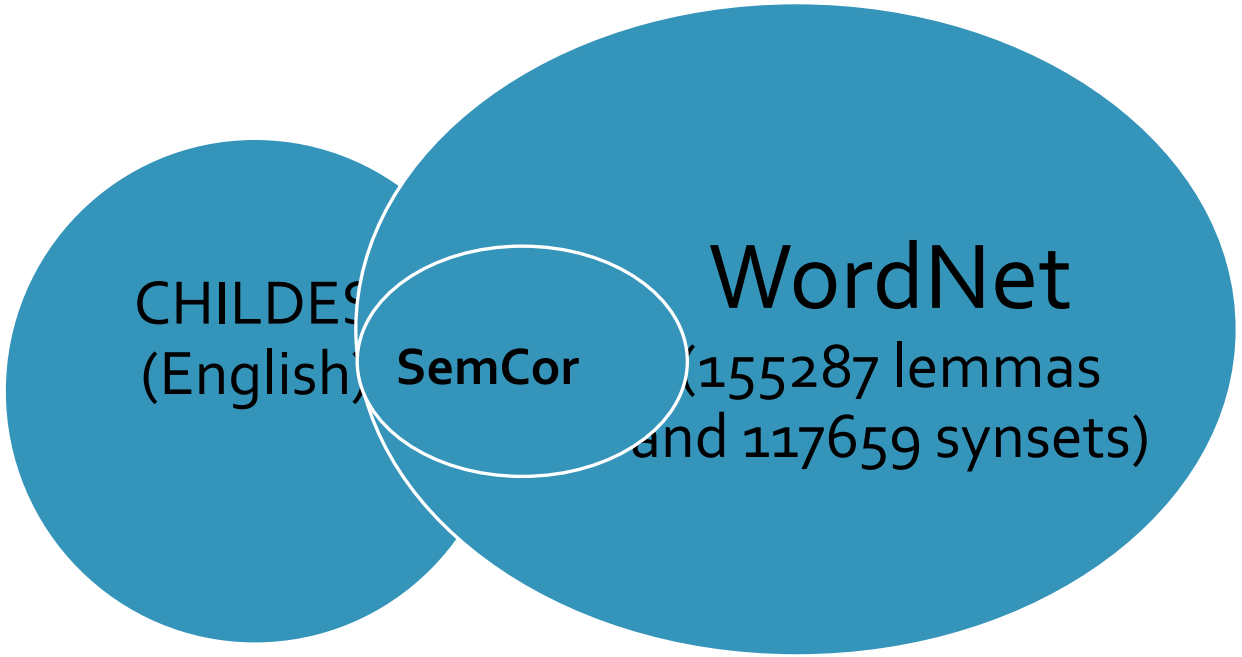
- **Polysemy (number of WordNet synsets):**
  - **WordNet polysemy**: full potential number of synsets of a word.
  - **SemCor polysemy**: measure the number of synsets that are used.

- **Word length** = orthographic length (in characters)

MATERIALS

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

- Introduction
- **Materials**
- Methods
- Results
- Discussion



- **English**
- **Content words** (N,Adj, V, Adv)
- **Separated Roles** *(CHI, MOT, FAT,INV)*
- **Lemmas present in SemCor.**

Role	Tokens	# Lemmas	# Analyzed Lemmas
Child	1,358,219	7,835	4,675
Mother	2,269,801	11,583	6,962
Father	313,593	6,135	4,203
Investigator	182,402	3,659	2,775

Hernández-Fernández et al (2016)

METHODS

We compute the relationship between three variables (for every lemma): length (characters), frequency and polysemy.

Frequency : Frequency of each pair *lemma, syntactic category*

- **SemCor frequency**
- **CELEX frequency.**
- **CHILDES frequency.** By role (CHI, MOT, FAT, INV).

Polysemy : Number of synsets for each pair *lemma, syntactic category*

- **SemCor polysemy.** tagged in the SemCor corpus. Analyzed in SemCor corpus and in CHILDES .
- **WordNet polysemy.** according to the WordNet database. Only analyzed in the CHILDES corpus.

METHODS

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

- Introduction
- Materials
- **Methods**
- Results
- Discussion

Relationship between **frequency, polysemy and lemma length**

1. SemCor frequency and SemCor polysemy.
2. SemCor frequency and lemma length in the SemCor corpus.

1. CELEX frequency and SemCor polysemy.
2. CELEX frequency and WordNet polysemy.
3. CHILDES frequency and SemCor polysemy.
4. CHILDES frequency and WordNet polysemy.
5. CHILDES frequency and lemma length in the CHILDES corpus
6. CELEX frequency and lemma length in the CHILDES corpus.

PROTOCOL

1. **Correlation test. Pearson, Spearman and Kendall correlation tests:** *cor.test* (R)
2. **Plot data: show the density of points.**
3. **Nonparametric regression:** *locpoly* **(R) -** showed in the previous plot.

Hernández-Fernández et al (2016)

RESULTS: *COR.TEST*

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

MAIN RESULTS

•**For all the speaker roles (including CHILDREN):**

• ***Positive*** **correlation** ***for the frequency-polysemy relationship.***

• ***Negative*** ***correlation between frequency and lemma length.***

•**p-value** ***near zero in all cases (<10^{-16})*** **=** ***correlations are*** **significant.**

Hernández-Fernández et al (2016)

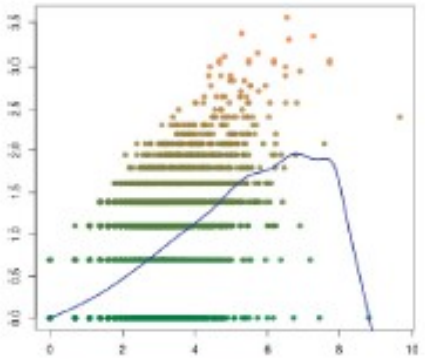
RESULTS:
NON PARAMETRIC REGRESSION

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

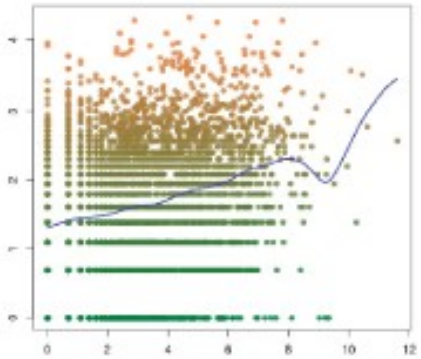
- Introduction
- Materials
- Methods
- **Results**
- Discussion

The **nonparametric regression** function ***draws a line*** with a ***positive slope*** *for* the frequency-polysemy relationship

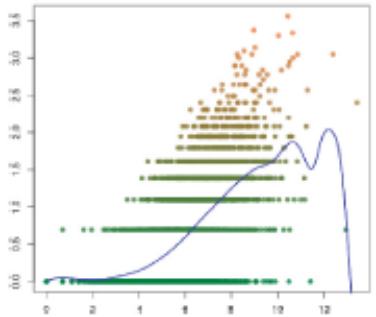
SemCor freq. vs SemCor pol.



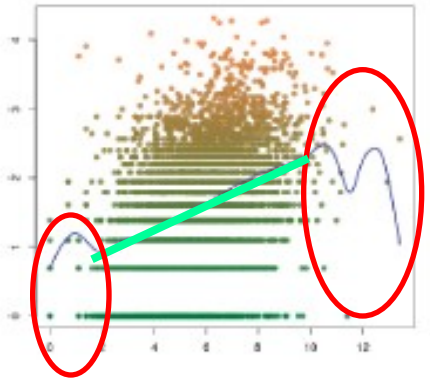
CHILDES freq. vs CHILDES pol.



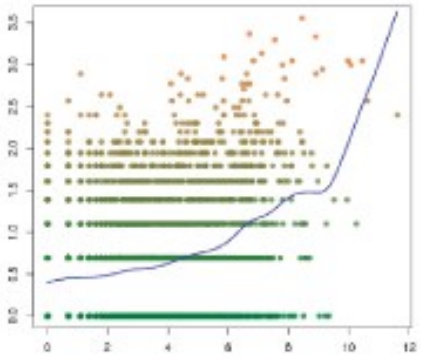
Celex freq. vs SemCor pol.



Celex freq. vs CHILDES pol.



CHILDES freq. vs SemCor pol.



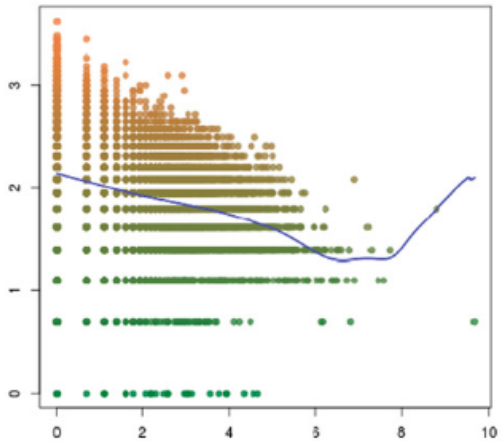
Hernández-Fernández et al (2016)

RESULTS:
NON PARAMETRIC REGRESSION

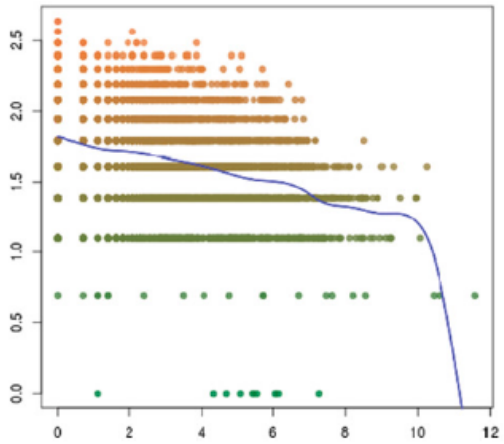
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

- Introduction
- Materials
- Methods
- **Results**
- Discussion

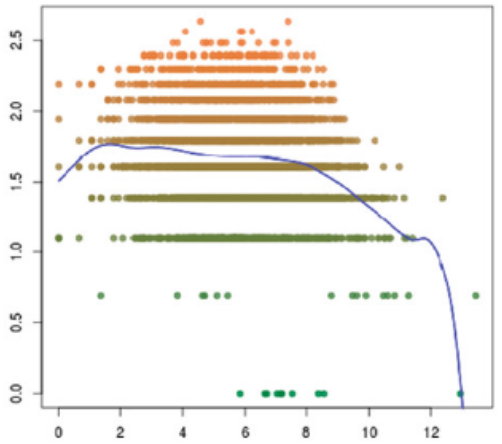
The nonparametric regression function ***draws a line*** with a ***negative slope*** *for the* frequency-length relationship.



SemCor freq.
vs
lemma length



CHILDES freq.
vs
lemma length



Celex freq.
vs
lemma length.

draws a line = this function is a **quasi-line** in the central area of the graph.
This tendency is not maintained at the extreme parts of graph, where the density of points is lower.

Hernández-Fernández et al (2016)

RESULTS: *COR.TEST*

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

- Introduction
- Materials
- Methods
- **Results**
- Discussion

Corpus	ρ	ρ_S	τ_K	Corpus length
SemCor frequency versus SemCor polysemy				
SemCor	0.209	0.627	0.555	23341
CHILDES frequency versus CELEX polysemy				
CHILDES (children)	0.084	0.249	0.177	4675
CHILDES (mothers)	0.081	0.281	0.202	6962
CHILDES (fathers)	0.084	0.279	0.202	4203
CHILDES (investigators)	0.062	0.211	0.153	2775
CELEX frequency versus WordNet polysemy				
CHILDES (children)	0.073	0.353	0.249	4406
CHILDES (mothers)	0.085	0.366	0.261	6577
CHILDES (fathers)	0.089	0.373	0.264	3989
CHILDES (investigators)	0.075	0.341	0.24	2654
CHILDES frequency versus SemCor polysemy				
CHILDES (children)	0.211	0.230	0.178	4675
CHILDES (mothers)	0.186	0.252	0.197	6962
CHILDES (fathers)	0.201	0.256	0.200	4203
CHILDES (investigators)	0.189	0.219	0.171	2775
CELEX frequency versus SemCor polysemy				
CHILDES (children)	0.201	0.607	0.477	4406
CHILDES (mothers)	0.197	0.602	0.474	6577
CHILDES (fathers)	0.226	0.595	0.463	3989
CHILDES (investigators)	0.228	0.585	0.451	2654

- **Positive** correlation *for the frequency-polysemy relationship.*

- ***Negative*** *correlation between frequency and lemma length.*

Corpus	ρ	ρ_S	τ_K	Corpus length
SemCor frequency versus lemma length				
SemCor	−0.062	−0.301	−0.229	23341
CHILDES frequency versus lemma length				
CHILDES (children)	−0.099	−0.324	−0.24	4675
CHILDES (mothers)	−0.076	−0.373	−0.278	6962
CHILDES (fathers)	−0.092	−0.366	−0.277	4203
CHILDES (investigators)	−0.096	−0.318	−0.242	2775
CELEX frequency versus lemma length				
CHILDES (children)	−0.091	−0.132	−0.095	4406
CHILDES (mothers)	−0.084	−0.124	−0.089	6577
CHILDES (fathers)	−0.087	−0.142	−0.102	3989
CHILDES (investigators)	−0.099	−0.172	−0.126	2654

(p-value *near zero in all cases = correlations are* significant)

Hernández-Fernández et al (2016)

DISCUSSION

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

- Introduction
- Materials
- Methods
- Results
- **Discussion**

Zipf's laws of language

- **Our analysis confirm a positive correlation between the frequency of the lemmas and the number of synsets (Zipf's meaning-frequency law) and a negative correlation between the length of the lemmas and their frequency (Zipf's law of abbreviation)**

No differences between roles

- **There are no big qualitative differences in the analysis of correlations for the different speakers (roles) in Childes, independently from the corpora analyzed and independently from the source used.**

Conclusion

- **It suggest that adults and children share the same general statistical (linguistic) biases related with frequency, polysemy and word length.**

Hernández-Fernández et al (2016)

FUTURE WORK

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

- Analysis of more **extensive databases** (Wikipedia,…).
- Extension to **other languages** (CHILDES, CELEX2,…).
- **Longitudinal study** of polysemy in children (case-study).
- More detailed **mathematical study** to understand variations displayed in the plots (*non-parametric regression*).
- Brevity law in **speech technologies** (grapheme-phoneme conversion): homophony, coarticulation…

Casas, B., Català, N., Ferrer-i-Cancho, R., Hernández-Fernández, A. & Baixeries, J. (2016). **The polysemy of the words that children learn.**

Gonzalez Torre, I., Luque, B., Lacasa, L., Luque, J. & Hernández-Fernández, A.: **Emergence of linguistic laws in human voice** (2016, submitted).

POLYSEMY... IN OUR BRAINS

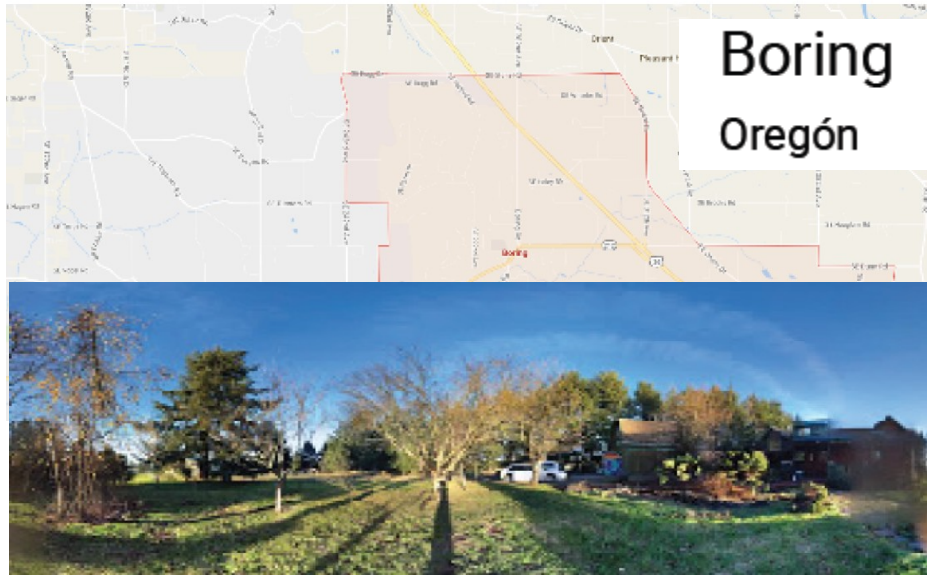
UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

- Introduction
- Materials
- Methods
- Results
- **Discussion**

ADJECTIVE

Boring Not interesting; tedious.

'I've got a boring job in an office'



Hernández-Fernández et al (2016)

4th International Conference on Statistical Language and Speech Processing
SLSP 2016

Pilsen, Czech Republic
October 12th, 2016

THANK YOU FOR YOUR ATTENTION!

MNOHOKRÁT DĚKUJI !

(SPECIALLY AFTER A BEER-NIGHT...)

Antoni Hernández Fernández , Bernardino Casas, Ramon Ferrer-i-Cancho and Jaume Baixeries

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

antonio.hernandez@upc.edu

MINECO project APCOM, Computational Learning and Communication (TIN2014-57226-P, 2014-2016).