The development of word frequency distribution in first language acquisition. An analysis on a spoken language corpus of French children

Andrea Briglia, Massimo Mucciardi, Giovanni Pirrotta

MICHELANGELO MISURACA | GERMANA SCEPI | MARIA SPANO

PROCEEDINGS OF THE 16TH INTERNATIONAL CONFERENCE ON STATISTICAL ANALYSIS OF TEXTUAL DATA

VOLUME 1

VADISTAT
SIMONA BALBI

The development of word frequency-distribution in first language acquisition. An analysis on a spoken language corpus of French children

Andrea Briglia[1], Massimo Mucciardi[2], Giovanni Pirrotta[3]

[1]Sorbonne Université – andrea.briglia@sorbonne-universite.fr

[2]University of Messina – massimo.mucciardi@unime.it

[3]University of Messina – giovanni.pirrotta@unime.it

Abstract

In this paper we present a study on the word-frequency distribution development in French speaking children, which aims to evaluate how their lexical output is related to a standard word-frequency distribution: Zipf's law. We adopted a set of spoken language transcripts of French children named CoLaJE: by using Python tools we turned original transcripts into strings that allowed us to estimate the exponential parameter of the word-frequency distribution (*alpha*) for each child, as well as for parental input. We show how *alpha* values tend to converge to 1 during later development which is coherent with current literature. We also estimate the exponential parameter for parental input and we found that Spearman's rho shows a fairly positive correlation between child's *alpha* and parents' *alpha* in later ages. Finally, we discuss our results in the light of previous studies on the CoLaJE corpus and we compare the obtained values to similar works on children's spoken language transcripts that were sampled in an analogous way, before outlining possible future directions.

Keywords: Zipf's Law; Spearman's rho; Language Acquisition, Natural Language Processing

1. Introduction

Zipf's law has been at the center of a long-lasting scientific debate since its discovery at the beginning of the last century by French stenographer Estoup (Zipf, 1949). This rank-frequency distribution is an inverse relation found in an approximately similar way in a number of human and natural phenomena (Ferrer i Cancho and Solé, 2001). In mathematical terms, this means that the rank of a word is approximately inversely proportional to its frequency and so produces a power law probability distribution. As far as we know, there is not a statistically significant difference in the shape of the plot between written and oral language, as it was suggested by a study on short oral samples similar to those used for our study (Ridley and Gonzales, 1994). Recent advances in NLP have increased understanding of this relation (Piantadosi, 2014) by pointing out how the universal word distribution would be an equilibrium between communication effectiveness and cognitive load that is unconsciously reached by listeners and speakers in a collaborative way (Lestrade, 2017). To our knowledge, research is still lacking regarding the emergence of a Zipfian word frequency distribution in French language during childhood. Our hypothesis is that some sort of pattern, tendency or regularity - which is to say, something that could be interpreted as a preliminary form of a Zipfian distribution - should be found in child language during the acquisition process. Preliminary results on these subjects suggest that a Zipf-like statistical distribution progressively emerges from around 30 months of age. In an analogous way to adults, stop words (especially determiners) are more frequent than

nouns and verbs. However, if we look at the transcriptions, there are a significant number of exceptions that we claim to be determined by context-specific situations (such as "maman" and "papa"), as well as the frequent naming of a toy that was constrained to a specific situation (it is important to bear in mind that our study is based on one-hour samples of spoken language). When we compare these results to previous studies on these children (Morgenstern and Parisse, 2012; Mucciardi et al., 2021), it could be said that children acquiring their native language in a faster way (*e.g* Madeleine) showed an earlier Zipfian-shaped distribution (*i.e.*, closer to *alpha* = 1) compared to the other children.

2. Data and NLP tools

CoLaJE (Morgenstern and Parisse, 2012; ver. 2.4) is a database made up of eight children[1] recorded *in vivo* one hour per month from their first year until they were approximately five. For each child there are nearly 8000 sentences and 20000 words with a mean length of utterance of three words on average. Child-directed speech has also been recorded and it is transcribed by using FAT and MOT lines. As adult input is of primary importance (Goodman, 2008) we decide to calculate the *alpha* for the parents too. This means that we calculate the sum of every occurrence of FAT (father) and MOT (mother). We have not, however, considered occurrences of OBS (the researcher filming the conversation), because this person spends just one hour per month with the observed child, so her language does not influence child development in a significant way. Adults in CoLaJE corpus are quite often speaking in the so-called "motherese": it is difficult to estimate how this influences the emerging word-frequency distribution in the examined children. This set of spoken language transcripts is considered statistically representative for studies in language acquisition (Yamaguchi, 2018). We analyze the corpus by using the following libraries: PyLangAcq (Jackson et al., 2016) for the manipulation of CHAT format (Mac Whinney, 2000) and Stanza (Qi et al., 2020) for the Part Of Speech (POS) tagging of children spoken utterances. Data are available in three different types[2]: CHI represents what the child should have said according to the adult norm in standard French orthography; MOD represents what the child should have said according to the adult norm in International Phonetic Alphabet and PHO represents what the child actually said in IPA. To obtain our results, we decided to choose CHI for a number of reasons. First, it is the only transcription available for all children, whilst MOD and PHO are available only for Adrien and Madeleine. Secondly, PHO poses more substantial problems for the purposes of the study, despite its advantage as the only transcription that exactly corresponds to real language. Indeed, a given word can be pronounced in many different ways with different degrees of variations, which makes computations complex: it is difficult to establish when a word means what it was meant to mean for a child, and to what extent different variated forms refer to the same entity, especially in earlier ages (Vihman, 1994). We thought that this choice would have biased our word-frequency counts, resulting in too many different word types that in most cases do not really represent different named entities. On the other hand, CHI represents what the child says in its correct form, which gives us an adult-biased interpretation of child vocabulary. It has nonetheless the advantage of gathering words that differed on what was imagined to be a phonetic level in a stable and fairly reliable way that in the majority of cases would not affect

[1] We decided to include the child named Philippine, even though her first recording is at 4 years old.

[2] Here is an example: *https://ct3.ortolang.fr/tools/trjsbrowser/trjs.html?f=/data4/colaje/adrien/ADRIEN-21-3_00_15/ADRIEN-21-3_00_15.tei_corpo.xml*

the lexical level. By proceeding in this way, we accepted a heavy reliance on the transcribers' choices.

3. Results and discussion

For every child, we estimated the *alpha* parameter according to the well-known Zipf's law (Zipf, 1949; Mandelbrot, 1961) showed in formula (1):

$$f(r) \sim \frac{1}{r^{\alpha}} \quad (1)$$

where *r* is the ranking of the word (from common to rare), f(r) is its frequency distribution for words in the child corpus and α (*alpha*) is the exponent of the law (a parameter defining the shape). We analyze a total of 237 corpora covering a developmental temporal range from 9.23 to 83.90 months. Figure 1 shows one of the 237 estimated models[3]. For each corpus examined we estimate the *alpha* parameter, obtaining the results shown in Table 1 and Figure 2 where alpha values are fitted according to a smoothing spline (Fan and Gijbels, 1996)[4]. If we look at figure 2, we can observe how the variability around alpha=1 is higher in earlier ages. It is possible to observe how the *alpha* parameter converges towards the value of 1 over time. In particular, after 48 months (data not shown) the *alpha* average is almost 1 (*i.e.* 0.99). We remind the reader that when *alpha* goes above 1, this means that there is a decrease in lexical richness, and vice versa: while *alpha* goes below 1 there is a growing lexical richness. Despite a limited vocabulary and an undeveloped syntax, we observe that the cognitive underpinnings behind the universality of Zipf are at play from the very beginning. The progressive emergence of a Zipf-like probability distribution in children can be influenced by an increased exposure to their parents' input. In our results, this appears by estimating Spearman's non-parametric correlations (Table 2). In fact, during the initial age of the children (<= 24 months), the correlation between the *alpha* parameter of the children and the *alpha* parameter of the parents is almost zero (0.005). With increasing age, the correlation appears to be significant and increases between 24 and 48 months (0.221). From the age of 48 months, it grows to a significant value of 0.330. We find these results coherent to previous findings on the same corpus (Mucciardi, 2021), in which we observe[5] two graphs representing the evolution of POS tags. We observe that for both children (despite the fact that the girl learns faster than the boy) there is a clearly significant increase in the quantity and quality of speech production from two-and-a-half-year-old (30th month): this is coherent with what we observe in Figure 2, as well as with the difference in the Spearman's values for the same age range over time. It is coherent in the sense that three different ways of measuring child language development (*i.e.* alpha evolution over time, Spearman correlation and the evolution of POS tags) show a similar convergence in which we can see how children's utterances are becoming closer to adult language both in quantitative and qualitative terms. Furthermore, these results are in line with the graphs produced by the

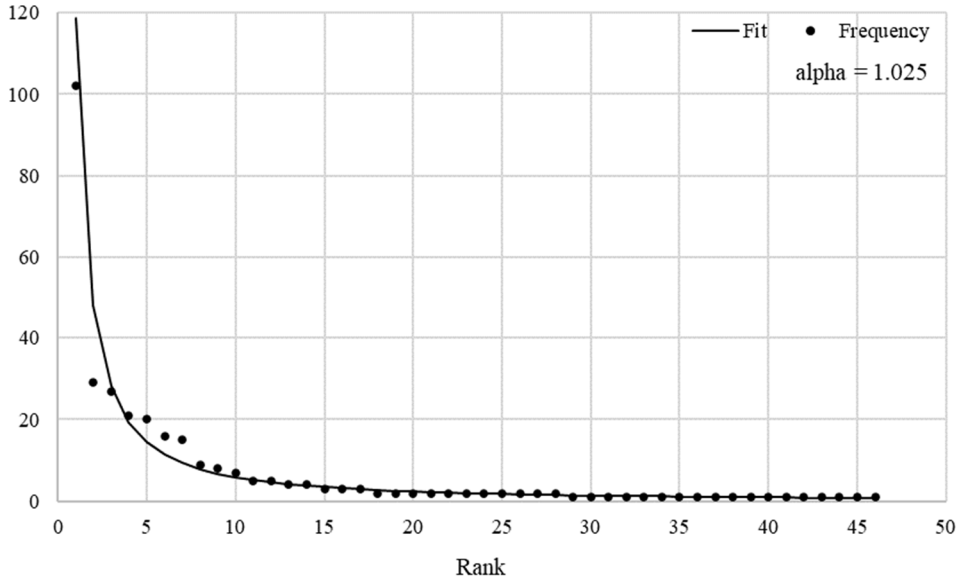
[3]All the 237 models estimated are highly performing in terms of goodness of fit and are available upon request. Data analysis was performed using STATA ver. 15.0 software.

[4] We use a spline function with a kernel function= Epanechnikov and bandwitdh=5 (see for more details Fan and Gijbels, 1996).

[5] *http://advanse.lirmm.fr/EMClustering* Adrien (dataset 1) and Madeleine (dataset 2). Click on the bottom-left part to choose to visualize the "evolution of POS tags" in absolute values (default is in relative values).

authors who led the constitution of the CoLaJE corpus: as we can see in their seminal paper (Morgenstern and Parisse, 2012), the mean length of utterance, type/token ratio and the number of utterances per hour evolve in a similar fashion. The progressive emergence of different grammatical categories (Mucciardi, 2021) is coherent to the graph proposed in the review article cited in the introduction (Piantadosi, 2014, fig.7). Additionally, analogous gaps between children with different paces of development can be observed in both graphs, as is the case for Madeleine and Adrien, while at around five-years old, children usually display less variability between their outputs (*alpha* values, mean length of utterance or type/token ratio).

Figure 1 - Estimated model for Theophile (age = 27 months): observed (frequency) and estimated (fit) values – R^2_{adj}=0.97 Anova test F=1416 (p<0.001)



Child	Mean (*alpha*)	N= corpus	Std. Deviation	Age (range in months)
ADRIEN	1.194	27	0.199	15.30 – 51.87
ANAE	1.021	32	0.082	16.67 – 71.00
ANTOINE	0.965	43	0.168	9.23 – 75.27
JULIE	1.010	27	0.179	11.60 – 70.23
LEONARD	1.024	14	0.080	20.30 – 38.83
MADELEINE	1.062	34	0.117	12.17 – 83.90
PHILIPPINE	0.907	20	0.190	56.90 – 67.53
THEOPHILE	1.035	40	0.181	14.33 – 75.93
Total	1.028	237	0.172	9.23 – 83.90

Table 1 - Descriptive statistics of alpha values on 8 children analyzed

Spearman's rho correlation		
Age <=24 (months)	Age >24 and <=48 (months)	Age >48 (months)
0.005	0.221*	0.330*
*p<0.05		

Table 2 - Spearman correlations between children alpha and parental alpha by age group

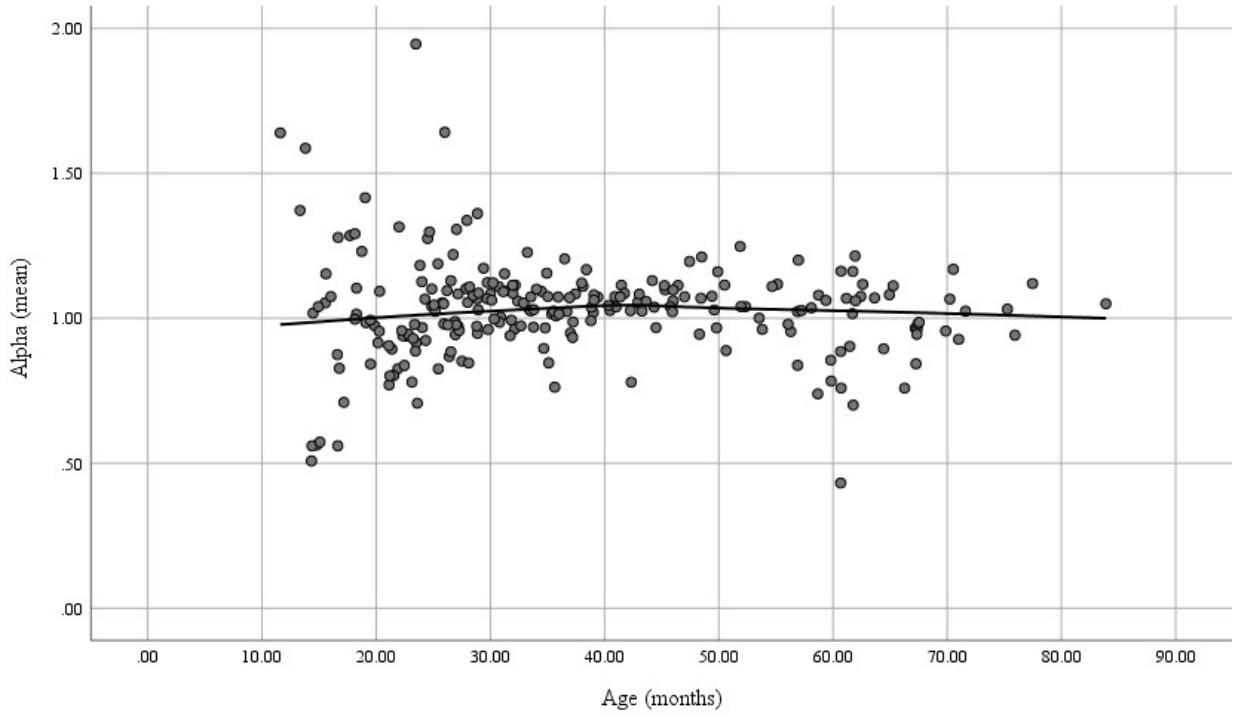


Figure 2 - Temporal trend of the alpha parameter for the 8 children
(solid line = fitted spline curve; point = alpha parameter)

4. Conclusion and future directions

To explore exponent evolution in child language development, we adopt as a reference a previous work (Baixeries, 2013) in which the authors challenged the assumption according to which the exponent of Zipf's law remains constant independently from the complexity of the communication system. They do so by pointing out how in children speaking Germanic languages the exponent "evolve from a high value of alpha to the value of alpha of adults at least from about 20 months onwards" (Baixeries, 2013). As for our study, these results have been tested on CHILDES data (Mac Whinney, 2000), which are similarly sampled spoken language transcripts. The results we obtained on French are on the one hand similar to the results of this study: alpha values for children show an analogous tendency from a value higher than 1 to a value closer to 1. On the other hand, our results show a significant number of alpha values below 1. We do not think that this difference is due to the fact that the languages examined are part of two different families (Romance and Germanic). We would rather think that this difference could derive from a different methodological choice: differently from Baixeries (2013), we do not use a length normalization of samples. To improve our work, we

could repeat the alpha estimation by using a text normalization procedure: this could lead to an improvement in comparability (Baixeries, 2013). Another additional strategy could be to explore the relationship of the inter-related evolution between *alpha* and the complexity of linguistic output: differently from what has been previously done we claim that it would be better to replace MLU (mean length of utterance) with a more fine-grained index of linguistic complexity as the ISC (Szmrecsanyi, 2004). We are currently developing a Python script to obtain the ISC score for every child utterance.

References

Baixeries J., Elvevag B. and Ferrer-i-Cancho R. (2013). *The Evolution of the Exponent of Zipf's Law in Language Ontogeny*. PLoS ONE 8(3): e53227.

Fan, J., and Gijbels I. (1996). *Local polynomial modelling and its applications*. London: Chapman & Hall.

Ferrer i Cancho R. and Solé R. V. (2001). *The small world of human language*. Proceedings of Royal Society of London B 2260-2265.

Goodman J., Dale P. and Li P. (2008). *Does frequency count? Parental input and the acquisition of vocabulary*. Journal of Child Language, 35(03), 515–531.

Jackson L., Burkholder R., Flinn G. and Coppess E. (2016). *Working with CHAT transcripts in Python*. Technical report TR-2016-02, Department of Computer Science, University of Chicago.

Lestrade S. (2017). *Unzipping Zipf's law*. PlosOne.

Mac Whinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Mandelbrot B. (1961). *On the theory of word frequencies and on related markovian models of discourse.* In: Jacobson R, editor, Structure of Language and its Mathematical Aspects, Providence, R. I.: American Mathematical Society. pp.190-219.

Morgenstern A. and Parisse C. (2012). *The Paris Corpus*. French language studies 22. 7- 12. Cambridge University press - https://hdl.handle.net/11403/colaje/v2.4.

Mucciardi M., Pirrotta G., Briglia A. and Sallaberry A. (2021). *Visualizing cluster of words: a graphical approach to grammar acquisition*. Proceedings of 13th Scientific Meeting of the Classification and Data Analysis Group, Firenze University Press, pp.392-395.

Qi P., Zhang Y., Zhang Y., Bolton J. and Manning C. (2020). *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. In Association for Computational Linguistics (ACL) System Demonstrations.

Piantadosi S. (2014). *Zipf's word frequency law in natural language: A critical review and future directions*. Psychon Bull Rev.; 21(5): 1112–1130.

Ridley D.R and Gonzales E. A. (1994). *Zipf's law extended to small samples of adult speech*. Percept Mot Skills. 79(1 Pt 1):153-4.

Szmrecsanyi B. (2004). *On operationalizing syntactic complexity*. JADT 2004.

Vihman, M. M. and McCune L. (1994). *When is a word a word?* Journal of Child Language, 21(3), 517–542.

Yamaguchi N. (2018). *What is a representative language sample for word and sound acquisition?* Canadian Journal of Linguistics / Revue canadienne de linguistique, University of Toronto Press 63 (04), pp.667-685.

Zipf G.K. (1949). *Human behaviour and the principle of least effort*. Cambridge (MA), USA: Addison-Wesley.