A logistic regression model for predicting child language performance

Andrea Briglia, Massimo Mucciardi, Giovanni Pirrotta

A logistic regression model for predicting child language performance

Un modello di regressione logistica per la previsione dell'apprendimento del linguaggio nel bambino

Andrea Briglia, Massimo Mucciardi and Giovanni Pirrotta

Abstract In this paper we propose a logistic regression model to evaluate how different components of language contribute to its acquisition over time. The empirical basis consists of a corpus which can be considered as a series of statistically representative samples taken at regular time intervals. Aim is to show how quantitative methods can contribute to solve linguistic puzzles.

Abstract *In questo articolo si propone un modello di regressione logistica per valutare come differenti componenti del linguaggio contribuiscono alla sua acquisizione nel corso del tempo. La base empirica consiste in un corpus, considerabile come una serie di campioni statisticamente rappresentativi presi ad intervalli di tempo regolari, l'obiettivo è mostrare come fenomeni apparentemente qualitativi possano essere spiegati con metodi quantitativi*

Key words: Natural Language Processing; Logistic Regression; Phonetic Variation, Frequency Effects on Learning

[1] Andrea Briglia, LHUMAIN Université "Paul-Valéry" Montpellier 3; email: andrea.briglia@univ-montp3.fr

Massimo Mucciardi, Dept. of Cognitive Science, Univ. of Messina; email: mucciard@unime.it;

Giovanni Pirrotta, Univ. of Messina; email: giovanni.pirrotta@unime.it

1 Introduction

This paper is to be considered as a continuation of a previous research project [1] [4] in which the phonetic development of children was explored. In the current paper we have extended the level of analysis from a merely phonetic one to a more global view on how phonemes turn into words. The elementary units are Part Of Speech tags (from now POS tags). If phonemes could be metaphorically considered as the atoms of language, in a similar fashion words could be viewed as playing the role of the molecules: though the latter are far bigger than the former, they combine in different ways to form more complex meaningful entities in an analogous manner. Children always need to infer rules and regularities of their native language from a limited amount of input. This is the task they need to follow:

« [..] discover the underlying structure of an immense system that contains tens of thousands of pieces, all generated by combining a small set of elements in various ways. These pieces, in turn, can be combined in an infinite number of ways, although only a subset of those combinations is actually correct. However, the subset that is correct is itself infinite. Somehow you must rapidly figure out the structure of this system so that you can use it appropriately early in your childhood» [6].

Learning a language means learning how to combine in a creative way a set of units that must respect a conventional order (*i.e.* phonotactic constraints and grammar): to reach this cognitive ability, children do not acquire their native language by simply repeating the input received because doing it this way would require much more time than it actually does. Children actively optimize the input received by trying to check whether it fits with adult language: English speaking children often pronounce an irregular verb conjugation by adding the regular "-ed" suffix on it. As "-ed" is the most frequent one, when you do not know how to conjugate a verb that you have never heard, the best option is to not take any risk and treat it as if it was part of the most common category. Another example, Italian speaking children do not need to hear all verbs listed in the "-are", ".ere", "-ire" forms to be sure how to conjugate the corresponding suffixes: once they know which rule applies to the different singular and plural forms of each person, they reach the ability to derive this rule even to verbs that they have never heard (that means the vast majority).

Modeling first language acquisition is a challenging scientific puzzle impossible to tackle in a short paper: what is at stake here is to propose a logistic regression model that has shown good performance in predicting grammatical development.

2 Data optimization strategy and statistical model

CoLaJE [3] is an open access French database part of the broader CHILDES project: seven children have been recorded in a natural setting one hour every

month, from their first year of life approximately until five years of age. Data is available in three different formats: IPA, orthographic norm and CHAT (acronym for Code for the Human Analysis of Transcription), each of them is aligned to the correspondent video recording, allowing researchers to see the original source and to eventually reinterpret every utterance on their own. The main coding structure of the database consists in the fundamental division between "*pho*" (what the infant says) and "*mod*" (what the infant should have said according to the adult standard phonetic/phonological norm): we define every occurrence in which "*pho*" differs from "*mod*" as being a variation. The sampling scheme can influence the range of deductions and generalizations that we could draw from data: for this reason we check if this corpus sampling scheme meets internationally recognized reliability criteria [8] and it does: this means that it is considered as statistically representative in respect to the frequency of the linguistic structures targeted. We transformed all the sentences for the child named *Adrien* from 2 to 5 years old (8,000 sentences, 20,000 words approximately) to machine-readable strings of characters to make them computable by the Python STANZA library [5] software. STANZA library features a language-agnostic fully neural pipeline for text analysis, including tokenization, multiword token expansion, lemmatization, part-of-speech, morphological feature tagging, dependency parsing, and named entity recognition. We tagged all the words of the sentences by assigning a part-of-speech (POS) tag to each. In the second step, we calculated the "Word Phonetic Variation" (WPV) for each word by setting a specific algorithm to compute this difference. At this step, we assume that a correct word is a word that has been correctly pronounced though we are aware that -grammatically speaking- a word needs also to be pronounced in the correct place to be considered fully correct (*i.e* in the correct order). This is the case for the majority of sentences (especially shorter ones). This assumption is to be considered as acceptable because programming a set of grammar-sensitive and context-dependent algorithms is a hard challenge, especially for evolving linguistic structures such as those of children, besides the fact that every language has its own specific grammar. In other words, this analysis has been made *a priori* from the cardinality of the original sentences to which the words belong. We then organize the data in a spreadsheet structure on which we have built the statistical model.

From a statistical point of view, we developed a logistic regression model [2] to examine which factors can predict child's performance as part of our methodology. The binary variable was set as follows: WPV=1 if there is a phonetic variation in the spoken word and WPV=0 if there is no phonetic variation.

We choose 3 predictors to explain WPV: AGE, COMPLEX and CLASS.

-AGE is the main driver of development: the more a child is exposed to his environment, the more he will learn from it.

-COMPLEX relates to the difficulty a child has to get over long and semantically rich sentences where more cues need to be spotted.

-CLASS provides a way to evaluate whether a child can or cannot use a given grammatical element, giving an indirect measure of his grammatical development.

In particular:

1) AGE - It represents Adrien's age from 1 to 5 years[1];
2) COMPLEX - It represents the type/token ratio, meaning the percentage of different or distinct words per sentence (*i.e* a proxy of lexical richness)
3) CLASS - Specifies the class to which the word belongs to: Open, Closed and Other. This framework of three classes has been taken from the Universal Dependencies project[2].

The following "logit" equation (1) will give the probability of WPV based on the three regressors:

$$P(Y \mid \mathrm{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 AGE + \beta_2 COMPLEX + \beta_3 CLASS)}} \tag{1}$$

Based on the equation analysis results, we can see (Table 1) which variable among AGE[3], COMPLEX and CLASS variables is statistically significant. The likelihood ratio (LR) test is significant indicating that the logistic model provides a better fit to the data than the intercept-only model. Furthermore, with a cut-off = 0.5, overall correctly classified cases are equal to 72.2%. Odds ratios (OR) in Table 1 suggest that AGE is the main regressor: as it increases, the likelihood of reporting a higher WPV decreases consistently. COMPLEX works differently: an increase in lexical richness causes an increase in WPV too, but this relation becomes weaker over time, as is confirmed by the graph in Figure 1.
CLASS is composed by three categories. "Open" contains lexical words such as verbs, common and proper nouns, adjectives. These classes contain large numbers of elements and are subject to change (a new entry can be added, another can be deleted). "Closed" contains functional words such as auxiliaries, pronouns and determiners. These classes contain few but highly occurring elements that are not subject to change (no new entries at all). "Other" contains everything that cannot be classified in the previous categories (punctuation, acronyms, etc). Figure 1 plots the WPV probability profiles with respect to the variable CLASS based on nine pre-

[1] We transform the variable from months to years for a better representation of the data.
[2] Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features and syntactic dependencies) across different human languages [9].
[3] Before modelling AGE as linear, we tried to create three successive yearly time slots to see how the two other regressors behave if taken apart, but the resulting correctly classified number of cases was lower than the proposed model. We then choose to model it as linear because first language acquisition is a highly non-linear phenomenon and the only certainty linguists have is that – roughly speaking – it develops in a cumulative way over time. We tried to model the interaction effects between COMPLEX and CLASS too, but it turned out to be less precise than the model proposed: in fact, COMPLEX showed a counterintuitive result in which its increase in value causes a decrease in WPV (models are available on request).

defined scenarios. We consider a range for the AGE variable from 1 to 5 years while for the COMPLEX variable a (hypothetical) range from 0.1 to 0.9. To give an example "A1-C0.1" means one year of age, and a percentage of distinct words of 10% per sentence. "A" stands for age and "C" stands for complexity. It can be observed how Open class words are easier to learn compared to Closed class words: the difference between the two profiles shows an (almost) constant value of 0.2 up to 4 years old. When the child has almost completed his growth (after 4 years) the two profiles tend to be similar. This is because children are more at ease in naming things and persons with their names instead by using more abstract pronouns which impersonally refers to them, and because verbs are easier to put in a sentence rather than auxiliaries, whose place must respect precise grammar rules that requires time to be learned.

Table 1: Logistic regression estimates[1]

Variables in the equation	***B***	***SE***	***WALD***	***df***	***OR-Exp(B)***
AGE	-1.854*	0.033	3189.8	1	0.157
COMPLEX	1.007*	0.087	132.6	1	2.739
CLASS#			602.3	2	
Class (Open)	2.555*	0.374	46.6	1	12.875
Class (Closed)	3.478*	0.375	85.9	1	32.410
Constant	1.777*	0.389	20.8	1	5.914

baseline CLASS = "Other" *p<0.01- Overall percentage correct = 72.2%
Nagelkerke R Square = 0.29 - Initial -2 Log Likelihood =23846.685 - Final -2 Log Likelihood = 19560 - (LR test p<0.01) - Sample size=19093 words

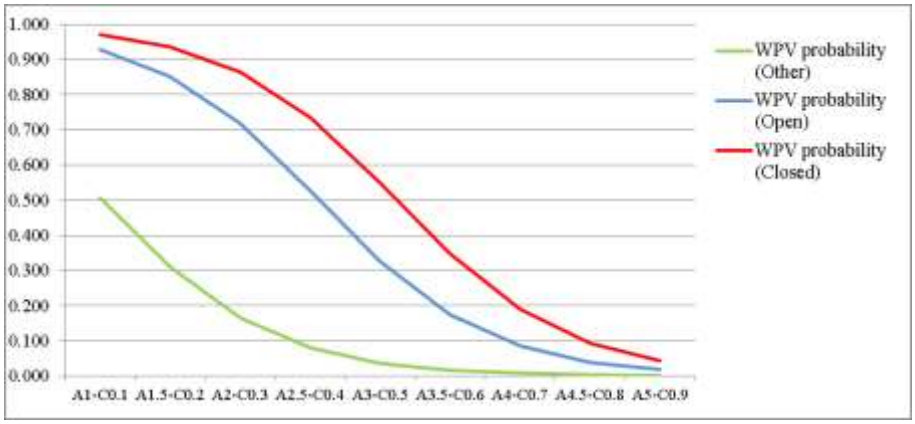

Figure 1: Predicted probability according to 9 scenarios by CLASS category – In abscissa A = Age (1 – 5 years); C = Complex Index (range 0.1 – 0.9)

[1] All calculations are performed with STATA ver. 15

3 Conclusion and future directions

The logistic regression model could be a fair way to represent child language acquisition through quantitative and graphical tools. The predicted outcome is fairly good but needs to be improved by taking into account how the place a word occupies in the sentence structure influences the WPV. Attempts to create a model closer to child development, in which age is modelled in a non-linear way and in which the complexity of a sentence influences (and is influenced by) the grammatical elements contained in it turned out to be too difficult and unpredictable. These difficulties could be interpreted in the following way: as we do not know exactly how AGE influences learning (WPV) and how COMPLEX and CLASS interact with each other, it seems to be better to model these regressors in the simplest possible way (AGE as linear, COMPLEX and CLASS as independent from each other). This being said, this can be true only at an initial stage of research: this statistical model should be applied to other similarly sampled children. By doing so, it would become possible to test the generalizability of the claims made in this paper and improve current knowledge on first language acquisition by comparing children between them and children learning similarly grammatical structured language between them [7]. A new research project on these themes is currently in progress, new tests on accounting for the non-linear effects of age and the interaction between regressors will be made.

References

1. Briglia, A., Mucciardi, M., Sauvage, J.: Identify the speech code through statistics: a data-driven approach, Book of Short Papers SIS (2020)
2. Hosmer, D., Lemeshow, S.: Applied logistic regression. New York: Wiley (1989)
3. Morgenstern, A., Parisse. C.: The Paris Corpus. French language studies 22. 7-12. Cambridge
4. Mucciardi, M., Pirrotta, G., Briglia, A.: EM Clustering method and first language acquisition. Workshop in Models and Learning for Clustering and Classification, Poster Session, Catania (2020)
5. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.J.: Stanza: a Python Natural Language Processing toolkit for many human languages. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (2020)
6. Saffran, J.: Statistical language learning: mechanisms and costraints. Current directions in Psychological Science. Vol.12 No 4. P 110-114. (2003)
7. Sekali, M.: First language acquisition of French grammar (from 10 months to 4 years old). French Language Studies 22, 1-6 . (2012)
8. Tomasello, M., Stahl, D.: Sampling children's spontaneous speech: How much is enough?. Journal of Child Language, 31:101–121. (2004)
9. UD (Universal Dependencies).: Retrieved from https://universaldependencies.org (2021)