VISUALIZING CLUSTER OF WORDS: A GRAPHICAL APPROACH TO GRAMMAR ACQUISITION

Massimo Mucciardi[1], Giovanni Pirrotta[2], Andrea Briglia[3] and Arnaud Sallaberry[4]

[1] Department of Cognitive Science, Education and Cultural Studies, University of Messina, (e-mail: massimo.mucciardi@unime.it)

[2] University of Messina, (e-mail: gpirrotta@unime.it)

[3] Université "Paul Valéry » Montpellier3 (e-mail: andrea.briglia@univ-montp3.fr)

[4] LIRMM, University of Montpellier, CNRS, & AMIS, Université "Paul Valéry » Montpellier3 (e-mail: arnaud.sallaberry@lirmm.fr)

ABSTRACT: Language has been traditionally considered as a qualitative phenomenon that mainly requires hermeneutical methodologies in order to be studied, yet in recent decades thanks to advances in data storage, processing and visualization - there has been a growing and fertile interest in analysing language by relying on statistics and quantitative methods. In light of these motivations, we think it is worthwhile to try to explore databases made up of transcripted infant children spoken language in order to verify whether and how underlying patterns and recurrent sequences of learning stages work during acquisition. So, we think that the Expectation Maximization clustering method combined with an innovative graphical visualization can be useful to evaluate the development of linguistic structures over time in a reliable way.

KEYWORDS: first language acquisition, EM clustering, graphical visualization, phonetic variation rate, POS Tags.

1 General Framework

First language acquisition can be studied and modelled by using statistical tools: experiments have shown how specific *innately biased statistical learning mechanisms* are activated during in vitro settings where children easily learn how to keep memory of the transitional probability between syllables to spot word' boundaries [6]. Statistical and computational methods have contributed to important advances in the understanding of language acquisition: corpus analysis is one of the most rigorous ways to account for pattern, regularities and learning stages in a sound and replicable procedure [2]. In a very abstract form, first language acquisition could be viewed as a mixture of deterministic and random processes. It is deterministic because rules and constraints applied to human cognition are partly known. It is partly a random process because the amount of variability between children and within a single child is largely acknowledged and represents at the same time what is interesting and what is di cult in modelling child language studies. Romberg and

Saffran [4] assert that in language acquisition, the term `statistical learning' is most closely associated with tracking sequential statistics in word segmentation or grammar learning tasks. Knowing these rules and constraints does not allow us to predict the outcome of a child beginning to be immersed in his/her native language. All we know is that around the age of 5/6, she/he will master his/her own language/s. We know approximately the learning stages, the date of his/her first word, and the rough order of consonant acquisition. Interesting theories have been developed about the patterns of errors (e.g. phonetic variation) that the child will most likely make, but it is to date vary hard to model language acquisition. The types of patterns tracked by a statistical learning mechanism could be quite simple, such as a frequency count, or more complex, such as conditional probability [4]. In other words, learning a language (here conceived as a statistical structure of the environment) is in some ways a process that bring a child to minimize long-term prediction error. Clustering text is an important phase in data analysis. The common task in text clustering is to handle text in a multidimensional space, and to partition corpora into groups, where each group contains sentences that are similar to each other according to some grammatical indicators. Considering the above, in this paper we propose a new statistical strategy to evaluate the development of child linguistic structures over time in a reliable way based on clustering and visualization of words. The clusters are sufficiently explanatory for understanding first language acquisition as well as seem efficient for clustering performance. The paper is organized as follows: section 2 describes the data structure and the model applied; section 3 briefly provides the analysis strategy, the principal elaborations and visual interface for clustering.

2 Data Structure and Model

CoLaJE is a database composed of seven children that have been videorecorded in vivo approximately one hour every month from their first year of life until they were five (see https://www.ortolang.fr/). In this exploratory research, statistical treatments have been tested only on two children (Adrien and Madeleine) because the transcriptions obtained from these corpora are the most complete. Code for the Human Analysis of Transcripts (CHAT) provides a standardized format for producing computerized transcripts of conversational interactions. By analyzing, cleaning, filtering and normalizing all the available original CHAT transcripts we aimed at producing two corpora composed of the overall amount of what the children said through the years. A total of 8214 and 7168 database annotated sentences containing more than 100 variables were collected[1]. Some useful measures have been calculated such as: child age in years (Time) and Sentence Phonetic Variation Rate (SPVR) [1]: the SPVR is obtained by comparing mod and pho in order to measure how the relation between varied and correct form evolves over time. In the single sentence i (with $i = 1....N$),

[1] Due to lack of space in this paper we present the results for the Adrein dataset only. All other calculations are available on request.

$$SPVR_i = \left(TNPV_i / CTWT_i\right) \cdot 100 \quad (1)$$

where TNPV is the Total Number of Phonetic Variations of the words - total number of the difference between what the child really says (called "pho") and what he should have said according to the adult norm called ("mod") - and CTWT is the Child Total Words Tokenized. Hence, SPVR can assume the value 0% when the child does not make any error and 100% when the child does not pronounce all the words contained in the sentence correctly. Then, we applied Part-Of-Speech Tagger (POS Tags), a software that reads text in a given language and assigns parts of speech to each word such as noun, verb, adjective. We used Stanza Core NLP engine [22] to tag all CHI words by using Universal Dependencies as a standard of reference for part-of-speech classification [7]. Considering the nature of the variables (count data), we use finite multivariate Poisson mixtures in the EM procedure. We recall that EM clustering is an iterative method relying on the assumption that the data is generated by a mixture of underlying probability distributions, where each component represents a separate group, or cluster. The method provides the optimal number of clusters in any empirical situation, by using a two step iterative algorithm [3]. According to this approach to estimate mixture parameters we computing the maximum likelihood estimate (MLE) with the EM algorithm. In the next paragraph we will see the results of the principal estimates

3 Principal Results

To extend previous research [1], we divide our database in nine strata considering 3 different age classes of the child (L=1.97-2.64; M= 2.71-3.39 H=3.46-4.33 expressed in years and months) and 3 classes of SPVR (L= 33; M=>33 and 66; H>66 expressed in percent). In total we get 9 strata (from LL to HH). By framing the analysis in this way, we turn EM clustering algorithm into a potentially interesting method that could provide a reliable way to observe linguistic structures development over time. In tables 1 we summarize the main results obtained from clustering through a overview on the most influential POS tags for each strata and its related clusters for the dataset examined. In addition, the means of the POS are calculated in each strata (data not shown). We can observe that VERB occupies an increasing important role in development: it is almost absent in Adrien (dataset 1) during the earlier ages strata, it develops sharply in median age strata while it is present in almost any sentence in the upper age strata. It is clear that VERB causes an increase in the SPVR, as their values are higher in higher error rate strata (more than 33 percent). We can also observe that the parts of speech such as PRON (pronoun), VERB, SCONJ (subordinating conjunction) - which could be considered as markers of longer sentences - increase their importance. For visualization of clusters, we propose an interactive and visual interface to better this analysis. It has been designed considering a list of requirements defined in regards of the data structures and variables extracted by the clustering technique and the tasks one should be able to perform on such data. These are the main features: 1) visualize the clusters by age and SPVR; 2) visualize the distribution of POS tags in the clusters;

3) visualize the different values characterizing the clusters (age, SPVR, number of POS tags, number of sentences) and the POS tags (number of occurrences in a cluster, percentage, mean, Fisher coefficient, p-value); 4) visualize the list of sentences of a cluster; 5) visualize the relative and absolute evolution of the number of POS tags when child grows up (see the following link for all the details http://advanse.lirmm.fr/EMClustering/). In conclusion, we would suggest that these preliminary results represent a fair attempt to visualize child language development through clusters of words grouped by several criteria (age, grammatical properties, correct pronunciation). We can cautiously say that in this first stage of research the EM algorithm can provide us some mild descriptions in the classification of POS tags.

Table 1. *EM clustering results by strata - Dataset 1 (Adrien) (# - clusters number in brackets - POS sorted for ANOVA post-hoc F-test (in bold) p <0.05) (First 10 POS)*

Ordered POS	LL (3)	LM (2)	LH (4)	ML (5)	MM (3)	MH (3)	HL (4)	HM (5)	HH (5)
POS1	**INTJ**	**VERB**	**PRON**	**CCONJ**	**ADP**	**PRON**	**PRON**	**NOUN**	**AUX**
POS2	**DET**	**PROPN**	**ADV**	**PRON**	**ADV**	**AUX**	**DET**	**DET**	**NOUN**
POS3	**ADP**	ADV	**DET**	**NOUN**	**DET**	**NOUN**	**VERB**	**PRON**	**VERB**
POS4	**NOUN**	NOUN	**VERB**	**AUX**	**SCONJ**	**DET**	**NOUN**	**ADJ**	**DET**
POS5	**SYM**	INTJ	**NOUN**	**VERB**	**CCONJ**	**ADP**	**SCONJ**	**AUX**	**PRON**
POS6	**ADV**	PRON	**INTJ**	**NUM**	**INTJ**	**ADV**	**ADP**	**VERB**	**NUM**
POS7	**PROPN**	DET	**PROPN**	**SYM**	**NOUN**	**PROPN**	**AUX**	**ADP**	**ADJ**
POS8	**PRON**	AUX	**AUX**	**ADV**	**ADJ**	**SCONJ**	**ADV**	**ADV**	**ADP**
POS9	**VERB**	NUM	**ADJ**	**DET**	**NUM**	**VERB**	**ADJ**	**SCONJ**	**ADV**
POS10	**X**	CCONJ	**SCONJ**	**PROPN**	**PROPN**	**INTJ**	**CCONJ**	**X**	**X**

References

[1] BRIGLIA A., MUCCIARDI M., SAUVAGE J. 2020. Identify the speech code through statistics: a data driven approach. Proceedings SIS 2020 (Pearson Editions).

[2] CHATER, N. MANNING, C. D. 2006. Probabilistic models of language processing and acquisition Trends in Cognitive Sciences 10(7), 335-44.

[3] DEMPSTER A.P., LAIRD N.M., RUBIN D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B: Methodological 39: 1-38.

[4] ROMBERG, A.R, SAFFRAN, J.R. 2020. Statistical learning and language acquisition. Wiley Inter discip Rev Cogn Sci. 1(6): 906-914.

[5] QI, P., ZHANG Y., ZHANG Y., BOLTON J., MANNING C. J. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations.

[6] SAFFRAN J. R., ASLIN R. N., NEWPORT E. L. 1996 Statistical learning by 8-Month-Old infants. Science, vol. 274,. 1926-1928.

[7]UNIVERSAL DEPENDENCIES. 2021. Retrieved from https://universaldependencies.org/fr/ pos/index.html