Model-based clustering and first language acquisition

Massimo Mucciardi, Giovanni Pirrotta, Andrea Briglia

Book of Short Papers of the 5th international workshop on

Models and Learning for Clustering and Classification

MBC^2 2020, Catania, Italy

Salvatore Ingrassia, Antonio Punzo, Roberto Rocci
(editors)

Ledizioni
The Innovative LEDIpublishing Company

Book of Short Papers of the 5th international workshop on

Models and Learning for Clustering and Classification

MBC^2 2020, Catania, Italy

Salvatore Ingrassia, Antonio Punzo, Roberto Rocci
(editors)

LEDIZIONI

Book of Short Papers of the 5th international workshop on Models and Learning for Clustering and Classification (MBC2 2020, Catania, Italy), Salvatore Ingrassia, Antonio Punzo, Roberto Rocci (editors)

Sito workshop
mbc2.unict.it

Ledizioni: settembre 2021

ISBN: 9788855265393

Ledizioni – LEDIpublishing
Via Antonio Boselli 10 – 20136
Milano, Italia

www.ledizioni.it

Indice

Model-based clustering and first language acquisition

Massimo Mucciardi, Giovanni Pirrotta and Andrea Briglia

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 reasons, we think it is worthwhile to try to explore databases made up of transcripted infant spoken language in order to verify whether and how underlying patterns and recurrent sequences of learning stages work during acquisition. So, we think that model-based clustering method via the Expectation-Maximization (EM) algorithm can be useful to evaluate the development of linguistic structures over time in a reliable way.

Key words: First Language Acquisition, Model-Based Clustering, EM Algorithm, Phonetic Variation Rate, POS Tags

1 General Framework

First language acquisition can be studied and modeled 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 [1]. Computational methods and models 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. The paper is organized as follows: section **2** describes the data structure;

Massimo Mucciardi e-mail: mucciard@unime.it, Andrea Briglia e-mail: abriglia@unime.it
Department of Cognitive Science, Education and Cultural Studies, University of Messina (Italy)

Giovanni Pirrotta e-mail: gpirrotta@unime.it
University of Messina (Italy)

section **3** briefly recalls the **Expectation Maximization** (EM) method, estimation strategy and data analysis. Finally, section **4** provides conclusions and suggestions for future research.

2 Data Structure

CoLaJE [2] 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. In this exploratory research, statistical treatments have been tested only on one child (*Adrien*) because the transcriptions obtained from this corpus are the most complete. The data is transcripted in three forms: **CHI** is what the child says in the orthographic form, **PHO** what the child really says and **MOD** what he should have said according to the adult norm. To make the data uniform in a suitable form for automatic processing, we had to make trade-off like choices: child language is subject to interpretation difficulties by adults trying to decode it: in about 5% of the total number of occurrences, the number of words differs between the three main aforementioned forms in which sounds are coded: we decide to cut off these occurrences because they would have biased the final statistics, since the classification methods need to have an equal number of words related to the same phrase. The resulting data structure is a transformation from the video [3] into a statistically manageable database. In this respect, 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 one *corpus* composed of the overall amount of what the child said through the years. A total of **8214** annotated sentences containing more than 100 variables were collected. Some useful measures have been calculated such as: child age in years (time); Sentence Phonetic Variation Rate (**SPVR**) [8]: the **SPVR** is obtained by comparing *mod* and *pho* in order to measure how the relation between varied and correct form evolves over time. 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 [5] to tag all CHI words by using Universal Dependencies as a standard of reference for part-of-speech classification [11].

3 Data Analysis [1]

The EM algorithm 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

[1] Some results are not shown due to lack of space, they are available upon request.

number of clusters in any empirical situation, by using a two step iterative algorithm: the (**E**) or expectation step and the (**M**) or maximization step. These two steps are repeated until a further increase in the number of clusters would result in a negligible improvement in the log-likelihood, namely a convergence. Accordingly, the program checks how much the overall fit improves in passing from one to two clusters (formed in all possible ways, and selecting the best), then from two to three, etc. If the error function calculated for the solution with K+1 clusters is not marked (e.g at least 5 percent better) more than the simpler solution with K clusters, then the solution with K clusters is considered ideal and retained [9] [10]. Considering the nature of the variables (count data) and assuming their independence, we use finite multivariate Poisson mixtures in the EM procedure. To extend previous research [8], we divide our database in 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=$\leq$33; M=$>$33 and $\leq$66; H$>$66 expressed in percent). In total we get 9 strata (from LL to HH). By framing the analysis in this way, we turn model-based clustering via EM algorithm into a potentially interesting method that could provide a reliable way to observe linguistic structures development over time.

Table 1 provides three general indexes describing how child language is developing in quantity, quality and accuracy: these variables are represented respectively in, Child Total Words Tokenized **(CTWT)**, Child Total Distinct Words Tokenized **(CTDWT)** and Normalized Levenshtein Distance **(NLD)**. In particular **NLD** [4] is a string metric for calculating the edit distance between two given words, that means the number of deletion, insertion or substitutions of a single character needed to turn one word into the other. To obtain a realistic picture of the variation rate over a child's ages, we adjust the Levenshtein Distance by normalizing it: this means that the rate will be expressed in relative values, thus obtaining a result capable of comparing shorter and longer sentences We can observe the validity of **NLD** by the fact that it decreases over the three slot of ages as the child improves his language. In a coherent way, **CTWT**, the total number of words pronounced, increases and the **CTDWT**, the total number of different word types (proxy of an index of lexical diversity) increases as well with a similar rate. Table 2 and 3 summarize the main results obtained from clustering through a detailed overview on the most influential POS tags for each strata and its related clusters. In addition, the means of the POS are calculated in each strata **(PSM)**. We recall that the difference between **SPVR** and **NLD** is in the different way of quantifying the variation rate: SPVR counts as a variated form every word that is not pronounced exactly as it should have been pronounced (coarse-grained), while **NLD** gives a percentage of the number of letters by which the pronounced word differs from the target word (fine-grained). These general indexes have been calculated to test the soundness of our dataset: this was necessary because the following analysis and computations applied (parsing and EM) would inevitably be heavily biased by any error occurred in this initial step. Let's move on to comment on the clustering results in detail.

- VERB. We can see that VERB occupies an increasing important role in development: it is almost absent in the earlier age strata (**PSM** = L 0.02; M 0.25; H 0.18), it develops sharply in median age strata (**PSM** = 0.16; 0.62; 0.44) while it is present in almost any sentence in the upper age strata (**PSM** = (0.79; 1.02; 0.67): it is clear also that VERB causes an increase in the error rate, as their values are higher in higher error rate strata (more than 33 percent). We can further explain the fact that VERB is higher in the LM, MM and HM strata by looking at the **CTWT** and **CTDWT** in the corresponding cells in table 1: they both have higher values as compared to the other strata: this because in these strata sentences are longer than the others and - *a fortiori* - they contain more verbs. If we want to know which specific verbs occur in the different clusters of a given strata, it is possible to observe the POS Cluster Mean **(PCM)** (values not shown) and read which kind of sentences have been placed in a specific cluster: from our results, it is possible to see how complex verbs (past and future forms, even in combination with auxiliaries) appear in later age clusters where **PCM** is higher than 0.5 while common verbs such as “to do”, “to be”, “to say”, “to like” occur mainly in their present form in both low and high valued **PCM** in earlier strata clusters without any significant distribution detected. This difference in clustering is probably due to the fact that a two years old child essentially expresses himself through 1-2 words per sentence, so it is hard to divide something that already represents a unit in itself. When the child is four year old the clustering procedure divides in a much clearer way the corpus, helped by the fact that sentences are longer and grammatically richer. **- Morphosyntactic coherence**. If we look at the single sentence [7], we can observe that morphosyntactic coherence is higher in HL, HM clusters compared to those in L layers, which is in line with Parisse’s results, we can also observe that the parts of the speech PRON, VERB, SCONJ - which could be considered as markers of longer sentences - increase their importance (see the PSM in table 2 and 3) along the age progression. Here below a couple of example[2]: *escargot tout chaud* **(CHI)** - ɛskaʁgo tu ʃo **(PHO)** - didago to so **(MOD)** in MH strata; *une souris verte* **(CHI)** - yn suʁi vɛʁtə **(PHO)** - yn ʦoji vatə **(MOD)** in HH strata. In the first, morphosyntactic coherence is expressed in a coherent way in the masculine form, but the pronoun has not been pronounced while in the second sentence the pronoun is correctly there and it is morphosyntactically coherent with the feminine form centered on the noun. We would then say that model-based clustering via EM seems capable to sort syntactically analogous sentences that are part of different error and age classes in a sufficiently precise way. **- NOUN, PROPN and PRON**. We can show how children develop a more abstract and adult-like way to referring to entities by pointing out the evolution of the values of PRON and the sum of the values of NOUN and PROPN: for L 0.02 vs 0.49, 0.20 vs 0.79, 0.09 vs 0.79; for M 0.13 vs 0.25, 0.70 vs 0.55, 0.41 vs 0.39; for H 1.14 vs 0.45, 1.48 vs 0.58, 0.74 vs 0.33. It is clear how children progressively learn to properly use pronouns instead of using nouns: this is reflected and confirmed in the fact that sentences are on average longer and thus children use anaphora in order to avoid the repetition of the noun or proper noun to indicate the main subject of the sentence. These results are in line with current literature on the acquisition of pronouns in French [6].

---

[2] **PHO** and **MOD** are the equivalent of the line in standard orthographic form **CHI** but have been translitterated in IPA (International Phonetic Alphabet). See for more details https://www.internationalphoneticalphabet.org/.

4 Conclusion

There are of course exceptions to these grouping tendencies but, besides that, 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 pronounciation). Until now, we can cautiously say that in this first stage of research the model-based clustering via EM algorithm can provide us some mild descriptions in the classification of POS tags. In other words, the unsupervised automatic procedure seems to be able to confirm a general grammatical development over time. This because cluster memberships are made up of grammatical categories that are differently learnt at different ages. Next step will be to focus on particular POS tags development over time by scanning every cluster and looking to confirm more specific learning tendencies.

Table 1: Corpus index by strata

Corpus index	**LL**	**LM**	**LH**	**ML**	**MM**	**MH**	**HL**	**HM**	**HH**
NLD	0.01	1.04	2.27	0.04	0.84	1.88	0.11	0.69	1.47
CTWT	1.52	2.52	1.54	1.88	3.67	2.34	4.54	5.43	3.01
CTDWT	1.19	2.09	1.26	1.53	3.10	1.98	3.69	4.48	2.49
# of sentences	611	184	914	851	626	1136	1762	1242	888

Table 2: Clustering results by strata (# - clusters number in brackets - POS sorted for ANOVA post-hoc F-test (in bold) p<0.05)

Ordered POS	**LL** (3)	**PSM**	**LM** (2)	**PSM**	**LH** (4)	**PSM**	**ML** (5)	**PSM**	**MM** (3)	**PSM**
POS1	**INTJ**	0.13	**VERB**	0.25	**PRON**	0.09	**CCONJ**	0.05	**ADP**	0.18
POS2	**DET**	0.09	**PROPN**	0.04	**ADV**	0.36	**PRON**	0.13	**ADV**	0.65
POS3	**ADP**	0.01	ADV	0.59	**DET**	0.08	**NOUN**	0.22	**DET**	0.28
POS4	**NOUN**	0.47	NOUN	0.75	**VERB**	0.18	**AUX**	0.05	**SCONJ**	0.04
POS5	**SYM**	0.02	INTJ	0.18	**NOUN**	0.62	**VERB**	0.16	**CCONJ**	0.04
POS6	**ADV**	0.56	PRON	0.20	**INTJ**	0.06	**NUM**	0.04	**INTJ**	0.17
POS7	**PROPN**	0.02	DET	0.17	**PROPN**	0.05	**SYM**	0.02	**NOUN**	0.52
POS8	**PRON**	0.02	AUX	0.10	**AUX**	0.04	**ADV**	0.83	**ADJ**	0.09
POS9	**VERB**	0.02	NUM	0.07	**ADJ**	0.02	**DET**	0.09	**NUM**	0.04
POS10	**X**	0.02	CCONJ	0.05	**SCONJ**	0.00	**PROPN**	0.03	**PROPN**	0.04
POS11	**CCONJ**	0.02	ADP	0.03	**CCONJ**	0.01	**ADP**	0.03	**AUX**	0.28
POS12	**SCONJ**	0.01	X	0.03	**ADP**	0.01	**X**	0.03	**VERB**	0.62
POS13	**AUX**	0.01	ADJ	0.02	**NUM**	0.02	**INTJ**	0.18	**PRON**	0.70
POS14	**NUM**	0.10	SCONJ	0.02	**SYM**	0.00	**ADJ**	0.01	**SYM**	0.01
POS15	ADJ	0.00	SYM	0.00	X	0.00	SCONJ	0.01	X	0.00

Table 3: Clustering results by strata (# - clusters number in brackets - POS sorted for ANOVA post-hoc F-test (in bold) p<0.05)

Ordered POS	**MH** (3)	PSM	**HL** (4)	PSM	**HM** (5)	PSM	**HH** (5)	PSM
POS1	**PRON**	0.41	**PRON**	1.16	**NOUN**	0.55	**AUX**	0.26
POS2	**AUX**	0.20	**DET**	0.32	**DET**	0.47	**NOUN**	0.31
POS3	**NOUN**	0.31	**VERB**	0.79	**PRON**	1.48	**VERB**	0.67
POS4	**DET**	0.16	**NOUN**	0.42	**ADJ**	0.13	**DET**	0.20
POS5	**ADP**	0.11	**SCONJ**	0.15	**AUX**	0.37	**PRON**	0.74
POS6	**ADV**	0.38	**ADP**	0.23	**VERB**	1.02	**NUM**	0.09
POS7	**PROPN**	0.08	**AUX**	0.21	**ADP**	0.26	**ADJ**	0.09
POS8	**SCONJ**	0.02	**ADV**	0.73	**ADV**	0.67	**ADP**	0.12
POS9	**VERB**	0.44	**ADJ**	0.09	**SCONJ**	0.10	**ADV**	0.31
POS10	**INTJ**	0.06	**CCONJ**	0.12	**X**	0.02	**X**	0.03
POS11	**NUM**	0.03	**SYM**	0.02	**CCONJ**	0.11	**PROPN**	0.02
POS12	X	0.01	**NUM**	0.08	**NUM**	0.04	**SCONJ**	0.04
POS13	SYM	0.00	**X**	0.02	SYM	0.01	**CCONJ**	0.04
POS14	ADJ	0.10	**PROPN**	0.03	INTJ	0.15	**INTJ**	0.08
POS15	CCONJ	0.01	**INTJ**	0.16	PROPN	0.03	SYM	0.00

References

1. Saffran J. R., Aslin R. N., Newport E. L.: *Statistical learning by 8-Month-Old infants*. Science, vol. 274,. 1926-1928, (1996)
2. Morgenstern A., Parisse C.: *The Paris Corpus*. French language studies 22. 7-12. Cambridge University Press. Special Issue, (2012)
3. CoLaJE Corpus, retrieved from http://colaje.scicog.fr/index.php/corpus, (2020)
4. Damerau J. *A technique for computer detection and correction of spelling errors*. Communications of ACM, 7 (3):171-176, (1964)
5. Zhang Y.; Zhang Y.; Bolton J.; Manning C. D. *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. In Association for Computational Linguistics (ACL) System Demonstrations, (2020)
6. Morgenstern A., Sekali M. *What can child language tell us about prepositions?*. Jordan Zlatev, Marlene Johansson Falck, Carita Lundmark and Mats Andrén. Studies in Language and Cognition, Cambridge Scholars Publishing, pp.261-275,ffhalshs-00376186, (2009)
7. Parisse C., Le Normand M. T. *How children build their morphosyntax: The case of French*. Journal of Child Language, Cambridge University Press (CUP), 27, pp.267-292., (2000)
8. Briglia A., Mucciardi M., Sauvage J.: *Identify the speech code through statistics: a data-driven approach*. Proceedings SIS 2020 (Pearson Editions), (2020)
9. Dempster A.P., Laird N.M., Rubin D.B.: *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society. Series B: Methodological 39: 1–38. (1977)
10. Witten, I.H., Frank, E.: *Data Mining*. Carl Hanser, München and Wien (2011)
11. Universal Dependencies, retrieved from https://universaldependencies.org/fr/pos/index.html