Sequential Patterns of POS Labels Help to Characterize Language Acquisition

Isabelle Tellier, Zineb Makhlouf, Yoann Dupont

Sequential Patterns of POS Labels Help to Characterize Language Acquisition

Isabelle Tellier[1,2], Zineb Makhlouf[1], Yoann Dupont[1]

(1) Lattice, CNRS - UMR 8094, (2) University Paris 3 - Sorbonne Nouvelle

Abstract. In this paper, we try to characterize various steps of the syntax acquisition of their native language by children with emerging sequential patterns of Part Of Speech (POS) labels. To achieve this goal, we first build a set of corpora from the French part of the CHILDES database. Then, we study the linguistic utterances of the children of various ages with tools coming from Natural Language Processing (morpho-syntactic labels obtained by supervised machine learning) and sequential Data Mining (emerging patterns among the sequences of morpho-syntactc labels). This work thus illustrates the interest of combining both approaches. We show that the distinct ages can be characterized by variations of proportions of morpho-syntactic labels, which are also clearly visible inside the emerging patterns.

Keywords. language acquisition, POS labeling, CRF, Sequential Data Mining, emerging patterns

1 Introduction

The acquisition of their native language by children, especially how grammatical constructions are gradually mastered, is a process which largely remains mysterious. Some researches address this issue within a Natural Language Processing framework, for example by implementing programs trying to mimic the learning process [CM06,Ali10]. Our approach in this paper is different: we do not target to reproduce, but to *mine* children productions, from a morphosyntactic point of view. More precisely, we study the linguistic utterances of children of various ages, seen as sequences of part-of-speech (POS) labels, with sequential data mining tools.

Sequential data mining can be applied to any kind of data following an order relation. This relation is often related to time; for texts, it is only the linear order of words in sentences. Sequential data mining allows to extract sequential patterns, that is sequences or sub-sequences of itemsets that repeatedly occur in the data. This domain has given rise to many works [AS95,SA96,Zak01,NR07]. If the extracted sequences are contiguous portions of texts, patterns coincides with the older notion of repeated segments [Sal86].

When data are composed of natural language texts, the itemsets are not necessarily reduced to words: lemmas and POS labels can also be taken into account. The use of sequential data mining technics in such a linguistic context has recently been tested for the extraction of Named Entities [NAFS13], the discovery

of relations between entities in the biological field [CPRC09,CCP10,BCCC12] or the study of stylistic differences between textual genres [QCCL12]. As we look at the emergence of grammatical constructions in children, we are mainly interested here in patterns of morpho-syntactic labels. As a matter of fact, they are more general than words or lemmas and provide more abstract characterizations of a given age. We seek in particular to exhibit specific *emerging patterns* for different age groups.

The remaining of the article is as follows. First, we present the way our corpora of children's productions of different age groups have been collected. Then, we explain how we processed their morpho-syntactic analysis. Observing that usual POS taggers available for French made many mistakes on our data, we have built a new one, by training a machine learning device (a CRF model) on a reduced set of manually corrected data. We show that, despite this reduced set of manual corrections, the new tagger obtained behaves far better than the previous one on our data. Finally, the last part of the paper describes the technique used for the extraction of n-grams of morpho-syntactic labels of each specific age group and provides quantitative and qualitative analyses of the corresponding emerging patterns.

2 Corpora

2.1 The CHILDES Corpus

Several resources collecting children's productions exist online, as those available in the CNRTL[1]. But the best known and most widely used database is CHILDES[2] [Elm01], a multilingual corpus of transcriptions of recorded interactions between adults and children. In this article, we are only interested in the French part of these data. The recordings of a child cover several months or years, the age of the children may therefore vary from one record to another. Relying on the transcription manual[3] which explicits the meta-data associated with the corpus, we created six different sub-corpora corresponding to six age groups: from the "1-2 years" to the "6-7 years".

2.2 Pretreatments

In this corpus, children and parents communicate by speech turns. Each speech turn is transcribed and delimited by a period. In the following, we consider that each line corresponds to a "sentence". The transcriptions are annotated and are often followed by additional information in a (semi-)standard format allowing to describe elements of the situation (e.g. objects which are in the scene). We performed a preprocessing step to focus only on linguistic productions. We have

[1] Centre National des Ressources Textuelles et Linguistiques (http://www.cnrtl.fr for children's production): see Traitement de Corpus Oraux en Français (TCOF) corpus
[2] http://childes.psy.cmu.edu/
[3] http://childes.psy.cmu.edu/manuals/CHAT.pdf

removed all special characters related to standards of transcription, as well as all information of phonetic nature, which are not relevant for the analysis of syntactic constructions and prevent the use of a tagger. We have also eliminated from our data all adult utterances.

The characteristics of each of our initial sub-corpora are presented in the table of Figure 1. There are differences between them: the corpus for the age of "6-7 years" is the smallest one. To balance the corpora of the different age groups, we have sampled them according to the *number of words*: this feature is more reliable than the number of sentences, because the length of the sentences is a key factor which significantly varies from one age to another (see the following). To have comparable sub-corpora, the number of words is thus more reliable than the number of sentences.

corpus	number of sentences	number of words	nb of distinct words	average length of the sentences
1-2 years	41786	63810	3019	1.23
2-3 years	115114	324341	8414	2.15
3-4 years	60317	243244	8479	4.62
4-5 years	16747	74719	4465	4.71
5-6 years	4542	29422	938	6.96
6-7 years	3383	21477	841	6.88

Fig. 1. Characteristics of the initial sub-corpora

2.3 Sampling

The smallest corpus in terms of words (the one of "6-7 years") is the reference sample for the other age groups. So, we chose to take 20,000 words per corpus, with a rate of 0.01% tolerance. To build our new corpora from the initial ones, we sampled sentences randomly until the sum of all words in all sentences reaches this size. After the sampling, we have six new corpora, whose properties are given in the table of Figure 2.

The corpora now have comparable size in terms of words. The number of sentences in each corpus have of course decreased, but we note that the average lengths of the sentences follow the same evolution than in the initial corpora. This is crucial because, as long as the children grow up, they tend to produce longer sentences. This is a well-known key feature of language acquisition [Bro73,MC81]. To go further in our exploration, we will now label the productions of the children with morpho-syntactic labels.

corpus	number of sentences	number of words	nb of distinct words	average length of the sentences
1-2 years	14284	20348	1086	1.42
2-3 years	9075	20504	1427	2.26
3-4 years	5043	21051	1575	4.17
4-5 years	4433	20949	1806	4.73
5-6 years	3047	20514	805	6.73
6-7 years	3147	20525	819	6.52

Fig. 2. Characteristics of the sampled sub-corpora

3 POS labeling

3.1 Use of an existing tagger

As we want to characterize the acquisition of syntactic constructions, we need more information than simple transcriptions of words. Our experiments in this article rely on a morpho-syntactic tagging of children's productions: we must thus assign to each word in the sub-corpora a label corresponding to its grammatical category. Several tools are available to annotate plain text in French with "Part of Speech" (POS) labels, such as TreeTagger [Sch94]. In our work, we have used SEM[4] [TDE+12], which was obtained by training a linear CRF (Conditional Random Fields) model on the French Treebank [ACT03]. The set of labels adopted in SEM, similar to the one of [CC08], includes 30 different categories among which the main important ones for the following are: NC (for common nouns), V (for verbs), DET (for determiners), P (for prepositions), I (for interjections) and CLS (for subject clitic). SEM also integrates the external lexical resource Lefff [CSL04] to help achieve a better labeling.

SEM has been learned with labeled sentences extracted from the French newspaper "Le Monde". Our texts of children productions have very different properties, and we therefore expect many annotation errors. Indeed, the corpus CHILDES is composed of oral transcription, whose conventions differ from those of writing (especially concerning punctuations). Furthermore, children utterances are often far from standard French. It has already been observed that, even if SEM is supposed to reach 97% accuracy on texts similar to those on which it has been learned, it reaches 95.6% accuracy on more casual written texts from blogs, and only 81.6% on oral productions of adults.

To assess the quality of SEM on our data, we have randomly selected 200 sentences from each of our six corpora, tagged them with SEM and manually corrected the labeling errors, following the annotation conventions of the French Treebank. The accuracy of SEM on these samples (see table of Figure 4) ranges from 70% (2-3 years) to 87% (6-7 years). The detailed F-measures of the main categories for each age group can also be seen in the table of Figure 3: the label interjection (I), very rare in the French Treebank but very frequent in our

[4] http://www.lattice.cnrs.fr/sites/itellier/SEM.html

corpora, are particularly not well recognized by SEM (the F-measures goes from 33.33 for the "1-2 years" age group to 0 for the the "6-7 years" one).

3.2 Learning a New tagger

As we want to perform statistical measures on the morpho-syntactic labels, labeling errors must be reduced as much as possible. In [TDEW13], it has been shown that to learn a good tagger by supervised machine learning, it is more efficient to have a small annotated corpus similar to the target data than to have a large too different training set. So, we decided to use the labelled sentences which have been manually corrected for the evaluation of SEM as training data to learn a new tagger adapted to our corpora.

For this, we have used the same tools as those used to learn SEM, that is CRFs (Conditional Random Fields), introduced by [LMP01] and implemented in the software Wapiti [LCY10]. CRFs are graphical models that have proven their effectiveness in the field of automatic annotation by supervised machine learning [TTA09,TDE+12]. They allow to assign the best sequence of annotations y to an observable sequence x. For us, the elements of x are words enriched with endogenous attributes (presence of caps, digits, etc.) or exogenous ones (e.g. associated properties in Lefff), while y is the corresponding sequence of morpho-syntactic labels.

We trained our new tagger thanks to $200 * 6 = 1200$ annotated and manually corrected sentences (which is a very small number to learn a POS tagger), and we tested it on $50 * 6 = 300$ other independent sentences, equally sampled from the 6 distinct sub-corpora. The table of Figure 3 gives the F-measures of the main labels obtained by SEM and by the re-learned tagger for each age group, while the accuracy of both taggers are provided in the table of Figure 4.

corpus	CLS	DET	I	NC	P	V
1-2 years	100/100	80/100	33.33/57.14	76.92/84.21	0/0	80/100
2-3 years	71.43/93.33	66.67/54.55	12.5/90.91	71.43/80	40/33.33	71.43/63.64
3-4 years	77.42/100	80/78.26	13.33/88.89	88.89/94.74	71.43/71.43	83.87/94.74
4-5 years	89.8/94.55	80.95/89.36	8.7/97.78	75.76/93.15	90.91/80	88.89/95.89
5-6 years	81.08/97.56	91.18/93.15	0/94.74	86.32/96.08	78.05/88.89	92.96/90.14
6-7 years	96.55/100	87.88/97.14	0/80	90/92.13	89.47/87.8	93.88/89.36

Fig. 3. F-measures of the main distinct labels before (with SEM) /after the re-learning

We observe that the relearning leads to a significant improvement of the accuracy of about 10% in average. SEM is better for only 4 cells out of 36 in the table of Figure 3, probably thanks to its better vocabulary exposure: the French Treebank on which SEM was learned was about ten times larger than our training corpus. The improvement brought by relearning is larger for oral-specific labels such as I. It is therefore very beneficial, despite a very small training corpus. This

corpus	SEM	re-trained tagger
1-2 years	82%	85%
2-3 years	70%	80%
3-4 years	73%	88%
4-5 years	75%	90%
5-6 years	80%	92%
6-7 years	87%	90%
average	**77.83%**	**87.5%**

Fig. 4. Impact of the re-learning on the accuracy of the distinct age groups

can be explained by the fact that the vocabulary used in our texts is relatively limited and redundant: few data are therefore sufficient to obtain a tagger which is effective on our corpus, even if it is not uniformly better than SEM on every label (it would obviously be much less effective on other types of data). In the following, we systematically use the new version of the tagger.

3.3 Analysis of POS labels

Figure 5 shows the distribution of the main morpho-syntactic categories in the different age groups. For example, we see that the curve of the label I (interjection) is decreasing (except for the 4-5 years age group): it seems that children use fewer and fewer interjections in their productions as long as grow up. In contrast, the label P (preposition) is strictly increasing, which is consistent with an acquisition of increasingly sophisticated syntactic constructions. Curves for the labels CLS (subject clitic) and V (verb) follow very similar variations, probably because they are often used together: they increase till the age of 4, then decrease from 4 to 6, and finally stabilize at the age of 6. Observing labels DET (determiner) and NC (common nouns), we notice that until the age of 4 years, NC is the most common label, but not yet being systematically associated with a DET. It is only at the age of 4 that both curves become parallel (most probably when most NC is preceded by a DET). We finally note that from the age of 5 years, the proportions of different labels stabilize.

The residual errors of the tagger (there is more than 10% remaining labeling errors) lead us to be prudent with these observations. But it is clear that some of the phenomena observed here would not have been possible without re-learning: interjections, for example, were the words most poorly recognized by the original SEM, because they are very rare in newspaper articles. However, their production appears to be an important indicator of the child's age group. Example sentences like "ah maman" ("ah mom") or "heu voilà " ("uh there") were respectively labeled as "ADJ NC" and "ADV V" with the original SEM tagger. After the re-learning, the labels became "I NC " and "I V", which is at least more correct.

Although we can already draw some interesting conclusions from these curves, we cannot characterize the syntactic acquisition of children from single isolated

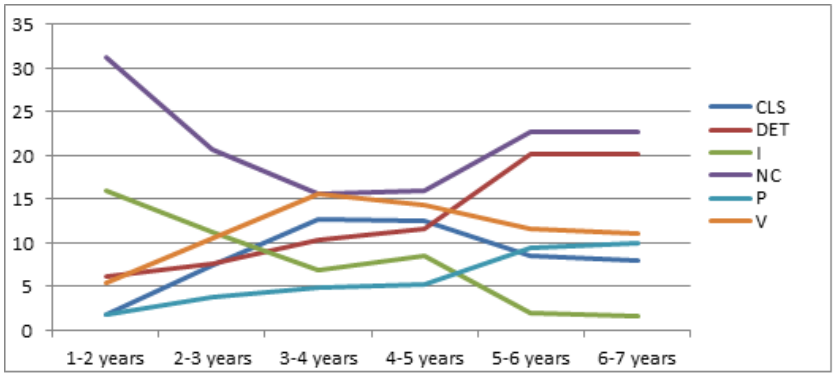

Fig. 5. Proportions of each label for each age group

categories. We thus decided to use sequential data mining techniques on our data to explore them further.

4 Sequential Patterns Extraction

4.1 General Definitions

Many studies have focused on the analysis of texts seen as sequential data. For example, the notion of *repeated segment* is used in textometrics [Sal86] to characterize a contiguous sequence of items appearing several times in a text. Sequential data mining [AS95] generalizes such concept, with notions like sequential patterns of itemsets. In our case, itemsets can be composed of words and POS labels. A sequence of itemsets is an ordered list of itemsets. An order relation can be defined on such sequences: a sequence $S_1 = \langle I_1, I_2, ..., I_n \rangle$ is included into a sequence $S_2 = \langle I'_1, I'_2, ..., I'_m \rangle$, which is noted $S_1 \subseteq S_2$, if there exist integers $1 \leq j_1 \leq j_2 \leq ... \leq j_n \leq m$ such that $I_1 \subseteq I'_{j_1}, I_2 \subseteq I'_{j_2}, ..., I_n \subseteq I'_{j_n}$ (in the classical sense of itemset inclusion). The table of Figure 6 provides examples of sequences of itemsets found in our corpus labelled with the re-trained tagger.

The support of a sequence S, denoted $sup(S)$, is equal to the number of sentences of the corpus containing S. For example, in the table of Figure 6, $sup(\langle(\text{ADJ})\ (\text{NC})\rangle) = 2$. The *relative support* of a sequence S is the proportion of sequences containing S in the base of initial sequences. It is worth $\frac{1}{2}$ for the sequence in our example, because this sequence is present in 2 out of the 4 sequences of the database. Algorithms mining sequential patterns are based on a minimum threshold for extracting frequent patterns. A *frequent pattern* is thus a sequence for which the support is greater than or equal to this threshold. Other concepts are also useful to limit the number of extracted patterns.

seq. id	sequence
1	⟨(le, DET) (petit, ADJ) (chat, NC)⟩ ("the little cat")
2	⟨(le, DET) (grand, ADJ) (arbre, NC)⟩ ("the big tree")
3	⟨(le, DET) (chat, NC)⟩ ("the cat")
4	⟨(tombé, VPP) (et, CC) (cassé, VPP)⟩ ("fallen and broken")

Fig. 6. Examples of sequences of itemsets (word, POS label)

4.2 Extraction of Sequential Patterns under constraints

In [YHA03], was introduced the notion of closed patterns that allows to eliminate redundancies without loss of information. A frequent pattern S is closed, if there is no other frequent pattern S' such $S \subseteq S'$ and $sup(S) = sup(S')$. In our example, if we fix minsup=2, the frequent pattern ⟨(DET) (NC)⟩, extracted from Figure 6, is not closed because it is included in the pattern ⟨(le, DET) (NC)⟩ and they both have a support equals to 3. But the pattern ⟨(DET) (small, ADJ) (NC)⟩ is closed. A length constraint can also be used. It defines the minimum and maximum number of items contained in a pattern [BCCC12].

4.3 Algorithm

There are several available tools for extracting sequential patterns such as GSP [SA96] and SPADE [Zak01]. CloSpan [YHA03] and BIDE [WH04] are able to extract frequent closed sequential patterns. SDMC[5], used here, is a tool based on the method proposed in [PHMA+01]. It extracts several types of sequential patterns, where items can correspond to simple words, lemma and/or their morpho-syntactic category (the tagger is parameterized, which allowed us to use our tagger). In this work, we wanted to characterize grammatical constructions, and we thus focused only on sequences of POS labels. The algorithm of SDMC implements the *pattern growth* technic; it is briefly discussed in [BCCC12]. It allows to extract sequential patterns under several constraints.

4.4 Emerging Patterns

[DL99] introduced the concept of *emerging pattern*. A frequent sequential pattern is called *emerging* if its relative support in a set of data set is significantly higher than in another set of data. Formally, a sequential pattern P of a set of data D_1 is emerging relatively to another set of data D_2 if $GrowthRate(P) \geq \rho$,

[5] https://sdmc.greyc.fr, login and password to be asked

with $\rho > 1$. The growth rate function is defined by:

$$\begin{cases} \infty & if\ support_{D_2}(P) = 0 \\ \frac{supp_{D_1}(P)}{supp_{D_2}(P)} & otherwise \end{cases}$$

where $supp_{D_1}(P)$ (respectively $supp_{D_2}(P)$) is the relative support of the pattern P in D_1 (respectively D_2). Any pattern P whose support is zero in a set is neglected.

5 Experiments

5.1 Parameters

The corpora used in our experiments are those described in section 2.3. We are interested here in sequences of itemsets restricted to POS labels without any gap (thus corresponding to n-grams, or repeated segments of labels), under some constraints (such as having a support strictly greater than a given threshold or pruning non-closed patterns), to limit their number. To set the lengths of sequences, we took account of the average size of sentences. So, we have decided to select patterns of length between 1 and 10. The minsup threshold is set to 2 and $\rho = 1.001$. To find the emerging patterns of a certain age group, we do as [QCCL12] did for literary genres: each age group (D_1) is compared to the set of every other age groups (D_2).

5.2 Quantitative Results

Figure 7 shows the number of frequent and emerging patterns obtained under our constraints for each age group. For example, for the age of 4-5 years, there are 1933 frequent patterns but only 842 emerging ones (42.6%). A serious reduction has occurred, which will make the observation easier. The number of emerging patterns is relatively stable across ages from 3-4 years and is important in each age group. As these emerging patterns are defined relatively to every other age group, this suggests the existence for each age group of characteristic phases of grammatical acquisitions.

Figure 8 shows the average size of the frequent and emerging patterns for each age group. The curves are very similar, suggesting that emerging patterns have properties which are similar to frequent patterns. In both cases, the length is increasing and reaches its maximum at the age of 5-6 years old. This parameter seems very correlated to the one of sentence length (see Figure 1): not only utterances become longer as the children grow up, but also the grammatical patterns they instantiate.

Figures 9 and 10 show the distributions of the main morpho-syntactic labels in frequent and emerging patterns respectively for each age group. These results are consistent with those obtained on the entire corpus (cf. Figure 5).

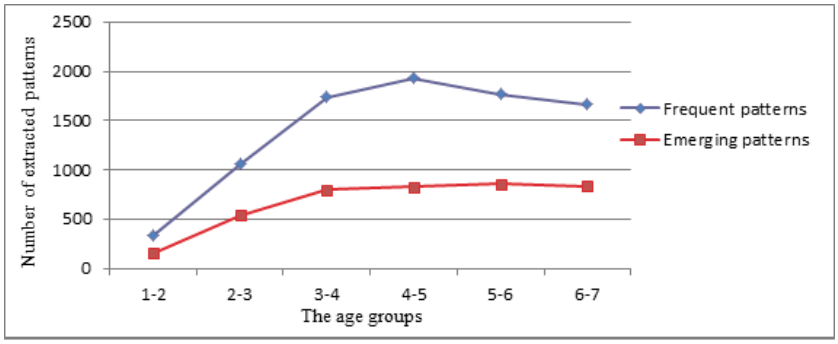

Fig. 7. Number of frequent versus emerging patterns for each age group

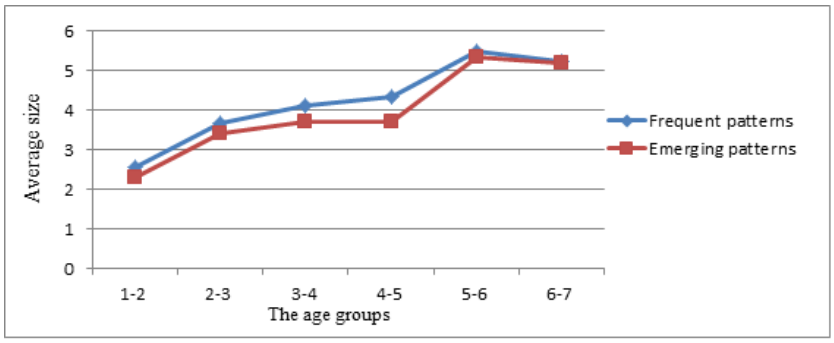

Fig. 8. Average length of frequent versus emerging patterns for each age group

The proportion of interjections still regularly decreases, while the one of prepositions increases, which is consistent with syntactic constructions of increasing complexity. We also note that the CLS and V curves are parallel and that, before the age of 4 years, the NC label is very frequent without being associated with the label DET. These curves show that the proportions of labels in the frequent and emerging patterns of each age group are similar to those of the corpus. In this sense, these patterns seem to be *representative* of the different age groups.

5.3 Qualitative Results

The table of Figure 11 provides examples of emerging patterns of each age group, and some corresponding sentences. These examples show that a single pattern can correspond to various sentences, and that they have increasing complexity. We note that even before the age of 2, children can produce sentences with a CN preceded by a DET. We also note, for example, that the patterns "(DET) (NC)" and "(DET) (NC) (CLS) (V) (VINF)" respectively extracted of the age

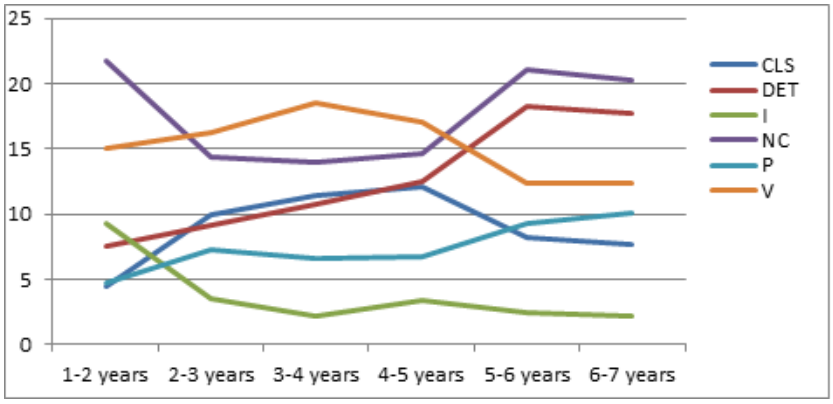

Fig. 9. Proportions of distinct labels in frequent patterns for each age group

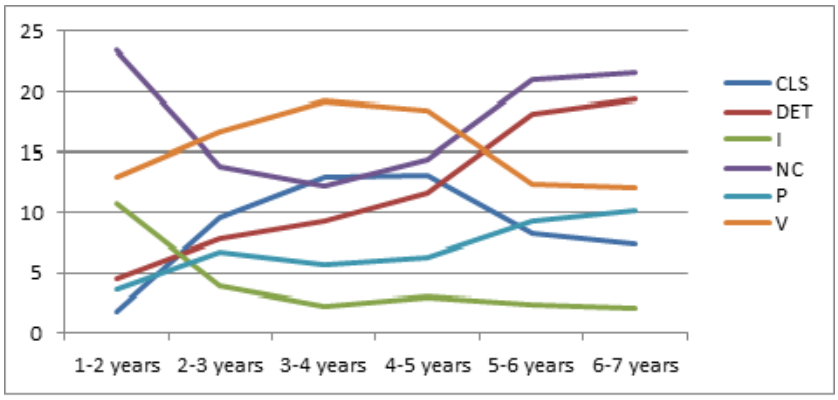

Fig. 10. Proportions of distinct labels in emerging patterns for each age group

"1-2 years" and "4-5 years are included in "(P) (DET) (NC) "and" (DET) (NC) (CLS) (V) (VINF) (DET) (NC)" respectively, of the following age group. This is consistent with a gradual acquisition of complex syntactic constructions.

6 Conclusion

In this article, we have applied techniques from Natural Language Processing, machine learning and sequential Data Mining to study the evolution of children's utterances of different ages. The phase of morpho-syntactic labeling required the learning of a specific tagger, adapted to our data. It was a necessity, considering that current available taggers do not properly handle oral transcriptions, and even less those of children: interjections, for example, which are very specific of

1-2 years	(P) (NC) (DET) (NC)	- **à maman** ("to mom") - sac **à dos** ("backpack") - **le ballon** ("the ball") - **des abeilles** ("some bees")
2-3 years	(P) (DET) (NC) (ADVWH) (CLS) (V)	- **de la tarte** ("some pie") - poissons **dans l'eau** ("fishes in the water") - **où il est ?** ("where it is ?") - **comment il marche ?** ("how it works ?")
3-4 years	(ADV) (CLS) (V)	- **non il est** par terre (" no it is on the floor") - **ici il pourra** passer ("here it will be able to pass")
4-5 years	(ADV) (CLS) (CLO) (V) (DET) (NC) (CLS) (V) (VINF)	- **alors tu m'as** vue ? ("so you saw me ?") - **oui j'en fais** souvent ("yes I do some often") - **les lapins ils vont rentrer** ("the rabbits they will come in") - **le chat il veut attraper** l'oiseau ("the cat it wants to catch the bird")
5-6 years	(DET) (NC) (CLS) (V) (VINF) (DET) (NC) (CC) (DET) (NC) (CLS) (V) (DET) (NC)	- **l'enfant il va chercher le chat** ("the child he goes and fetch the cat") - **le monsieur il va chercher les cerises** ("the man he goes and catch the cherries") - la maman **et le papa ils regardaient le garçon** ("the mommy and the daddy they watched the boy") - **et le chat il mange les cerises** ("and the cat it eats the cherries")
6-7 years	(P) (VINF) (DET) (NC) (DET) (NC) (PROPEL) (V) (DET) (NC) (P) (DET) (NC)	- les oiseaux les aident **à ramasser les cerises** ("the birds help them to pick up the cherries") - il y a un chat qui essaie **de chasser des oiseaux** ("there is a cat trying to catch birds") - il y a **un chat qui suit la fille avec son panier** ("there is a cat which follows the girl with a basket") - et aussi **un monsieur qui ramasse des cerises dans un arbre** ("and a man picking up cherries in a tree")

Fig. 11. Examples of emerging patterns in each age group

oral productions, would have been poorly recognized without re-learning. This is crucial, as the curves of label proportions show that their frequency appears as an important way to characterize a child's age group.

We currently restricted our research to n-grams of POS labels but further work could use richer itemsets of the type (word, lemma, POS tag). Our exploration seems to confirm that the extracted emerging patterns are representative of the age group in which they arise. The provided examples further confirm the intuition that (at least some of) the patterns of increasing age groups are included into each other, going in the direction of a grammatical sophistication.

As far as we know, these kinds of analyses had never been performed before. Of course, a detailed analysis of the patterns obtained remains to be done by specialists of language acquisition. They could for example allow to characterize typical evolutions of grammatical knowledge, or help to diagnose pathological evolution of a child's productions. We hope that they will provide valuable tools for the study of language acquisition phases.

7 Aknowlegment

This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083).
The authors acknowledge Christophe Parisse, for his advice.

References

[ACT03] A. Abeillé, L. Clément, and F. Toussenel. Building a treebank for french. In A. Abeillé, editor, *Treebanks*. Kluwer, Dordrecht, 2003.

[Ali10] A. Alishahi. *Computational modeling of human language acquisition (Synthesis lectures on human language technologies)*. San Rafael: Morgan and Claypool Publisher, 2010.

[AS95] R. Agrawal and R. Srikant. Mining sequential patterns. In *Int. Conf. Data Engineering: IEEE*, 1995.

[BCCC12] N. Béchet, P. Cellier, T. Charnois, and B. Crémilleux. Discovering linguistic patterns using sequence mining. In *proceedings of CICLing'2012*, pages 154–165, 2012.

[Bro73] R. W. Brown. *A first language: the early stages. Cambridge, Mass.* Harvard University Press, Cambridge, Massashusetts, 1973.

[CC08] B. Crabbé and M. H. Candito. Expériences d'analyse syntaxique statistique du français. In *Actes de TALN'08*, 2008.

[CCP10] P. Cellier, T. Charnois, and M. Plantevit. Sequential patterns to discover and characterise biological relations. In A. Gelbukh, editor, *CICLing 2010. LNCS, vol. 6008*, pages 537–548, 2010.

[CM06] N. Chater and C. D. Manning. Probabilistic models of language processing and acquisition. In *Trends in Cognitive Science, 10(7)*, pages 335–344, 2006.

[CPRC09] T. Charnois, M. Plantevit, C. Rigotti, and B. Crémilleux. Fouille de données séquentielles pour l'extraction d'information. In *Traitement Automatique des Langues, 50(3)*, 2009.

[CSL04] L. Clément, B. Sagot, and B. Lang. Morphology based automatic acquisition of large-coverage lexica. In *LREC 2004, Lisbonne*, 2004.

[DL99] G. Dong and J. Li. Efficient mining of emerging patterns: Discovering trends and differences. In *Proc. of SIGKDD'99*, 1999.

[Elm01] J. Elman. Connectionism and language acquisition. In *Essential readings in language acquisition. In Oxford : Blackwell*, 2001.

[LCY10] Thomas Lavergne, Olivier Cappé, and François Yvon. Practical very large scale CRFs. In *Proceedings of ACL'2010*, pages 504–513. Association for Computational Linguistics, July 2010.

[LMP01] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 282–289, 2001.

[MC81] J. F. Miller and R. S. Chapman. The relation between age and mean length of utterance in morphemes. In *Journal of Speech and Hearing Research, 24*, pages 154–161, 1981.

[NAFS13] D. Nouvel, J-Y. Antoine, N. Friburger, and A. Soulet. Fouille de règles d'annotation partielles pour la reconnaissance d'entités nommées. In *TALN'13*, pages 421–434, 2013.

[NR07] M. Nanni and C. Rigotti. Extracting trees of quantitative serial episodes. In *Proc. of KDID'07*, pages 170–188, 2007.

[PHMA+01] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu. Prefixspan: Mining sequential patterns by prefix-projected growth. In *ICDE, IEEE Computer Society*, pages 215–224, 2001.

[QCCL12] S. Quiniou, P. Cellier, T. Charnois, and D. Legallois. Fouille de données pour la stylistique : cas des motifs séquentiels émergents. In *Proceedings of the 11th International Conference on the Statistical Analysis of Textual Data, Liege*, pages 821–833, 2012.

[SA96] R. Srikant and R. Agrawal. Mining sequential patterns: Generalizations and performance improvements. In *EDBT 1996. LNCS, vol. 1057*, pages 3–17, 1996.

[Sal86] A. Salem. Segments répétés et analyse statistique des données textuelles. In *Histoire & Mesure volume 1 - numéro 2*, pages 5–28, 1986.

[Sch94] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, 1994.

[TDE+12] I. Tellier, D. Duchier, I. Eshkol, A. Courmet, and M. Martinet. Apprentissage automatique d'un chunker pour le français. In *Actes de TALN'12, papier court (poster)*, 2012.

[TDEW13] I. Tellier, Y. Dupont, I. Eshkol, and I. Wang. Adapt a text-oriented chunker for oral data: How much manual effort is necessary? In *The 14th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'2013), Special Session on Text Data Learning, LNAI, Hefei (Chine)*, 2013.

[TTA09] Y. Tsuruoka, J. Tsujii, and S. Ananiadou. Fast full parsing by linear-chain conditional random fields. In *Proceedings of EACL 2009*, pages 790–798, 2009.

[WH04] J. Wang and J. Han. Bide: Efficient mining of frequent closed sequences. In *ICDE, IEEE Computer Society*, pages 79–90, 2004.

[YHA03] X. Yan, J. Han, and R. Afshar. Mining closed sequential patterns in large databases. In *SDM SIAM*, 2003.

[Zak01] M. J. Zaki. Spade: An efficient algorithm for mining frequent sequences. In *Machine Learning Journal 42(1/2)*, pages 31–60, 2001.