Caractériser l'acquisition d'une langue avec des patrons d'étiquettes morpho-syntaxiques

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

[1] Lattice, [2]Université Sorbonne Nouvelle

makhlouf.zineb@gmail.com, yoa.dupont@gmail.com, isabelle.tellier@univ-paris3.fr

Abstract

In this paper, we want to characterize the various steps of the syntax acquisition of their native language by children. To this aim, we first build a corpus extracted from the French part of the CHILDES database, then we study the linguistic utterances of the children belonging to various ages with tools coming from natural language processing (morpho-syntactic labeling by supervised machine learning) and sequential data-mining (extraction of emerging patterns among the sequences of morpho-syntactc labels). We show that the distinct ages can be characterized by variations of proportions of morpho-syntactic labels, which are also visible inside the emerging patterns.

Résumé

Dans cet article, nous cherchons à caractériser les différentes phases de l'acquisition de la syntaxe de leur langue maternelle par les enfants. Pour cela, nous constituons tout d'abord un corpus extrait de la base CHILDES en français, puis nous proposons d'étudier les productions langagières d'enfants de différentes tranches d'âge à l'aide d'outils issus du traitement automatique des langues (l'annotation morpho-syntaxique par apprentissage automatique supervisé) et de la fouille de données séquentielle (l'extraction de motifs émergents parmi les séquences d'étiquettes morpho-syntaxiques). Nous montrons notamment que les différentes tranches d'âges peuvent se caractériser par des variations de la fréquence des étiquettes qui se retrouvent dans les patrons syntaxiques émergents.

Mots-clés : acquisition du langage, étiquetage morpho-syntaxique, CRF, fouille de données séquentielles, patrons syntaxiques.

1. Introduction

Le processus d'acquisition de leur langue maternelle par les enfants reste en grande partie mystérieux. La manière dont les constructions grammaticales sont peu à peu acquises est en particulier le lieu d'intenses débats. Pour aborder cette question avec des outils de traitement automatique des langues, on peut par exemple chercher à modéliser ce processus d'apprentissage par des programmes (Alishali, 2010, Chater et al., 2006). Notre approche dans cet article est différente : elle consiste à étudier les productions langagières d'enfants de différentes tranches d'âge avec des méthodes de fouille de données séquentielles.

La fouille de données séquentielles permet d'extraire des suites d'événements contenus dans de grandes masses de données qui suivent une relation d'ordre. Cette relation peut être de nature temporelle ; pour les textes c'est seulement l'ordre linéaire des mots dans les phrases. Les suites extraites prennent la forme de motifs (ou patrons) séquentiels, c'est-à-dire de séquences ou sous-séquences d'éléments répétées à plusieurs reprises dans les données. Ce

domaine a donné lieu à de nombreux travaux (Srikant et Agrawal, 1996, Zaki, 2001, Nanni et Rigotti, 2007).

Pour les données textuelles, les éléments pris en compte peuvent être les mots eux-mêmes, leur lemme, ou leur catégorie syntaxique. L'application de la fouille de données séquentielles à des textes est assez récente. Elle a notamment été appliquée dans (Nouvel et al., 2013) pour l'extraction d'entités nommées, dans (Charnois et al., 2009, Cellier et al., 2010, Béchet et al., 2012) pour la découverte de relations entre entités dans le domaine biologique et dans (Quiniou et al., 2012) pour l'étude des différences stylistiques entre genres textuels. Comme nous nous intéressons à l'émergence de constructions syntaxiques chez les enfants, ce sont principalement les n-grammes d'étiquettes morpho-syntaxiques qui retiendront ici notre attention.

La suite de l'article présente tout d'abord comment nous avons constitué notre corpus de productions d'enfants de différentes tranches d'âge. Puis, nous expliquons comment nous avons procédé à l'analyse morpho-syntaxique des énoncés qui y figurent. Constatant que les étiqueteurs disponibles pour le français courant font beaucoup d'erreurs sur nos données, nous en avons construit un nouveau par apprentissage automatique à partir de données corrigées manuellement. Enfin, la dernière partie expose comment les n-grammes d'étiquettes morpho-syntaxiques spécifiques de chacune des tranches d'âges ont été extraits, et propose une analyse quantitative et qualitative des motifs émergents obtenus.

2. Constitution d'un corpus d'acquisition d'une langue

2.1. Constitution de corpus de productions d'enfants par tranches d'âge

Plusieurs ressources de productions d'enfants existent en ligne comme celles disponibles dans le CNRTL[1]. Mais la base de données la plus connue et la plus utilisée est CHILDES[2] (Elman, 2001), corpus multilingue de transcriptions d'enregistrements d'interactions entre des adultes et des enfants. Nous nous intéressons dans le cadre de cet article uniquement aux données en français. Les enregistrements s'étendent sur plusieurs mois, voire plusieurs années, l'âge des enfants varie donc d'un enregistrement à l'autre. En nous appuyant sur le manuel de transcription[3], qui explicite les méta-données du corpus, nous avons regroupé les textes selon l'âge des enfants, qui varie de 1 à 7 ans. Nous avons ainsi constitué six corpus différents correspondant à six tranches d'âges : de « 1-2 ans » jusqu'à « 6-7 ans ».

2.2. Prétraitements

Dans ces corpus, les transcriptions sont annotées et souvent suivies par des informations supplémentaires dans un format (semi-)standard : des éléments de la situation (par exemple, quels objets sont dans la scène) peuvent ainsi être décrits. Nous avons effectué une étape de prétraitement pour nous concentrer sur les seules productions langagières. Nous avons supprimé tous les caractères spéciaux liés aux normes de transcription, ainsi que toutes les informations de nature phonétique, qui ne sont pas utiles à l'analyse des constructions syntaxiques et empêchent le fonctionnement de l'étiqueteur qui sera employé par la suite. Comme déjà mentionné, ces corpus correspondent à des interactions parents-enfants, chaque prise de parole étant transcrite dans une ligne. Nous avons éliminé de nos données toutes les

1 Centre National des Ressources Textuelles et Linguistiques : www.cnrtl.fr.
2 http://childes.psy.cmu.edu/
3 http://childes.psy.cmu.edu/manuals/CHAT.pdf,

prises de parole des adultes, pour nous concentrer uniquement sur les productions des enfants. La fin de chaque transcription est marquée par une ponctuation, nous l'assimilerons donc par la suite à une phrase.

Les caractéristiques de chacun des corpus obtenus sont présentées dans la Table 1. Il existe des différences entre les tranches d'âges : le « 6-7 ans » est le corpus de plus petite taille, il représente 2.93 % et 6.62 % respectivement de la taille de celui des « 2-3 ans » en termes de phrases et de mots. Pour équilibrer les corpus des différentes tranches d'âges, nous les avons échantillonnés selon leur nombre de mots (indice plus fiable que celui du nombre de phrases).

Corpus	**Nombre de phrases**	**Nombre de mots**	**Nombre de mots #**	**Taille moyenne des phrases**
1-2 ans	41786	63810	3019	1.23
2-3 ans	115114	324341	8414	2.15
3-4 ans	60317	243244	8479	4.62
4-5 ans	16747	74719	4465	4.71
5-6 ans	4542	29422	938	6.96
6-7 ans	3383	21477	841	6.88

Table 1 : Caractéristiques des corpus des différentes tranches d'âges

2.3. Echantillonnage

Le plus petit corpus en termes de mots est celui de la tranche d'âges « 6-7 ans ». C'est celui qui a servi de référence pour échantillonner les autres tranches d'âge. Nous avons donc choisi de prendre 20000 mots par corpus, avec un taux de tolérance de 0.01%. Pour construire de nouveaux corpus à partir de ceux créés précédemment, nous avons tiré les phrases aléatoirement jusqu'à ce que la somme de tous les mots de toutes les phrases soit égale à cette taille. Après l'échantillonnage, nous avons obtenu six nouveaux corpus comparables en termes de mots, dont nous résumons les caractéristiques dans la Table 2.

Corpus	**Nombre de phrases**	**Nombre de mots**	**Nombre de mots #**	**Taille moyenne des phrases**
1-2 ans	14284	20348	1086	1.42
2-3 ans	9075	20504	1427	2.26
3-4 ans	5043	21051	1575	4.17
4-5 ans	4433	20949	1806	4.73
5-6 ans	3047	20514	805	6.73
6-7 ans	3147	20525	819	6.52

Table 2 : Caractéristiques des corpus échantillonnés

Les corpus ont maintenant des tailles comparables en termes de mots. Bien que le nombre de phrases ait diminué, nous remarquons que les tailles moyennes des phrases de ces nouveaux corpus suivent la même évolution que celles des corpus de base. Cette taille moyenne augmente en fonction de l'âge des enfants : plus ils grandissent, plus les phrases qu'ils

produisent sont longues. On retrouve là une propriété bien connue des spécialistes de l'acquisition du langage (Brown R.W., 1973, Miller J.F. et Chapman R.S., 1981). Pour aller plus loin dans notre exploration, nous devons maintenant étiqueter les productions des enfants avec des catégories morpho-syntaxiques.

3. Étiquetage morpho-syntaxique (Part Of Speech)

3.1 Ré-utilisation d'un étiqueteur existant

Comme nous souhaitons caractériser l'acquisition de constructions syntaxiques, nous avons besoin de plus d'informations que les simples transcriptions. Nos expériences dans la suite de cet article s'appuient sur un étiquetage morpho-syntaxique des productions des enfants : nous devons donc attribuer à chaque mot du corpus une étiquette correspondant à sa catégorie grammaticale. Plusieurs outils sont disponibles pour annoter du texte brut en français avec des étiquettes Part Of Speech (POS), tels que le TreeTagger. Dans notre travail nous avons utilisé le Segmenteur-Étiqueteur Markovien (SEM), qui a été appris automatiquement par des CRF (Tellier et al., 2012) à partir du French Treebank (Abeillé et al., 2003). Le jeu d'étiquettes adopté dans SEM est celui présenté dans (Crabbé et al., 2008), il comprend 30 étiquettes différentes. SEM intègre également une ressource externe le *LeFFF*, le Lexique des Formes Fléchies du Français (Clément et al., 2004) pour fournir un meilleur étiquetage.

SEM a été appris sur des phrases extraites d'articles du journal « Le Monde ». Nos textes de productions d'enfants présentent des propriétés bien différentes et il faut donc s'attendre à beaucoup d'erreurs d'annotation. En effet, les corpus de CHILDES sont des transcriptions de l'oral, dont les conventions diffèrent de celles de l'écrit (notamment au niveau de la ponctuation). De plus, nos corpus ne sont constitués que de productions d'enfants de 1 à 7 ans, qui n'utilisent pas encore le français standard.

Pour évaluer la qualité de SEM sur nos données, nous avons tiré aléatoirement 200 phrases dans chacun des 6 sous-corpus construits précédemment, nous les avons étiquetées avec SEM et nous avons corrigé manuellement les erreurs d'étiquetage en nous basant sur les conventions d'annotation du *French Treebanck*. L'exactitude (accuracy) de SEM sur cet échantillon (cf. Table 3) varie de 70% (2-3 ans) à 87% (6-7 ans), bien loin des 97% atteints sur des données proches de celles sur lesquelles il a été appris.

3.1 Ré-apprentissage d'un étiqueteur spécifique

Si nous voulons effectuer des mesures statistiques se basant sur les étiquettes morpho-syntaxiques, il convient de réduire au maximum les erreurs d'étiquetage. Dans (Tellier et al., 2013), il a été montré qu'il n'était pas nécessaire d'utiliser un grand corpus d'apprentissage, si ce dernier est comparable au corpus d'application, pour obtenir de meilleurs résultats par rapport à un corpus plus grand, mais inadapté. Nous avons donc décidé d'utiliser les phrases annotées et corrigées manuellement pour l'évaluation de SEM comme données d'apprentissage pour apprendre un étiqueteur du français qui soit plus adapté à notre corpus.

Nous avons pour cela reconduit les expériences qui avaient permis l'apprentissage de SEM, en utilisant les champs markoviens conditionnels ou CRF introduits par (Lafferty et al., 2001). Les CRF sont des modèles graphiques ayant largement fait leur preuve dans le domaine du l'annotation par apprentissage automatique supervisé (Tsuruoka et al., 2009, Tellier et al., 2012). Ils permettent d'attribuer une séquence d'annotations y à une séquence observable x. Pour nous, les éléments de x sont les unités lexicales auxquelles sont associés des traits

endogènes (casse, présence de chiffres, etc.) ou exogènes (propriétés associées dans LeFFF par exemple), tandis que *y* est la séquence d'étiquettes morpho-syntaxiques correspondante.

Nous avons appris le nouvel étiqueteur grâce aux 200*6=1200 phrases initiales corrigées à la main, et nous l'avons testé sur 50*6=300 nouvelles phrases corrigées. Nous avons calculé l'exactitude de l'étiqueteur avant et après le réapprentissage sur ces données, les résultats obtenus sont résumés dans la Table 3.

Corpus	SEM	SEM réappris
1-2 ans	82 %	85 %
2-3 ans	70 %	80 %
3-4 ans	73 %	88 %
4-5 ans	75 %	90 %
5-6 ans	80 %	92 %
6-7 ans	87 %	90 %

Table 3 : L'impact du réapprentissage de l'étiqueteur SEM

Nous constatons que l'étiquetage est nettement amélioré par le réapprentissage, avec un gain en moyenne de l'ordre de 10 % d'exactitude. Ces bons résultats peuvent s'expliquer par le fait que le vocabulaire employé dans ces textes est relativement limité. Notre étiqueteur ré-appris est adapté au corpus d'étude, il le serait sans doute moins sur d'autres types de données. Nous utilisons dans la suite SEM réappris pour étiqueter l'ensemble des 6 corpus.

3.3 Analyse des étiquettes POS

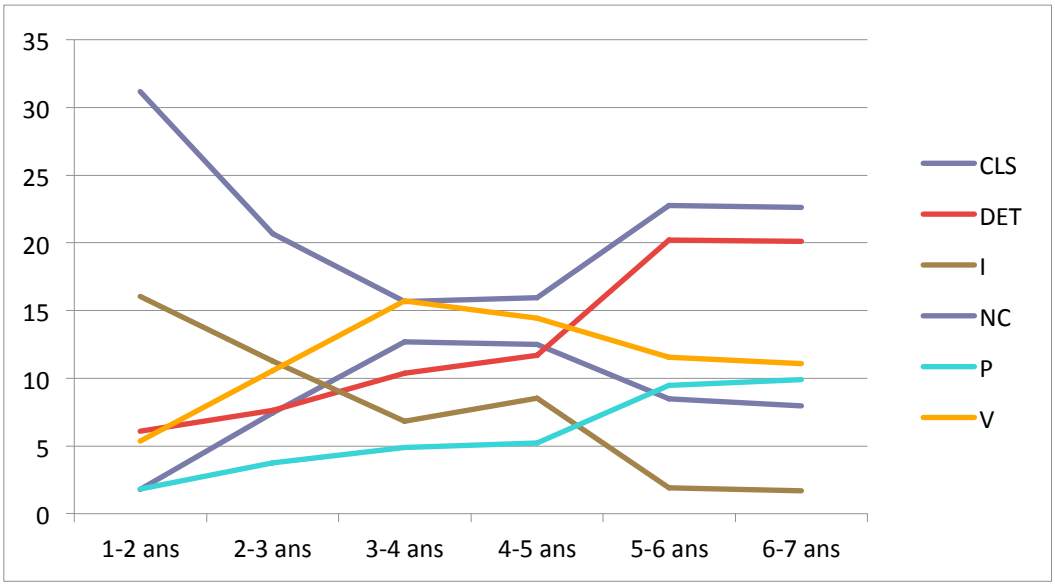

Figure 1 : Proportion des différentes étiquettes dans les corpus

La Figure 1 montre comment se répartissent les principales catégories morpho-syntaxiques dans les différentes tranches d'âge. Nous voyons notamment que la courbe de l'étiquette I (Interjection) est décroissante : il semble que les enfants utilisent de moins en moins d'interjections dans leurs productions. En revanche, celle de l'étiquette P (Préposition) est croissante, ce qui va dans le sens d'une acquisition de constructions syntaxiques de plus en

plus sophistiquées. Les courbes des étiquettes CLS (CLitique Sujet) et V (Verbe) suivent des variations similaires : elles sont croissantes jusqu'à l'âge de 4 ans, décroissantes entre 4 et 6 ans, puis se stabilisent à partir de 6 ans. Pour ce qui est des étiquettes DET (DETerminant) et NC (Nom Commun), nous remarquons que jusqu'à l'âge de 4 ans, les NC sont les étiquettes les plus fréquentes, mais sans être encore systématiquement associées à des DET. Ce n'est qu'à partir de 4 ans que les deux courbes deviennent parallèles (la plupart des NC étant sans doute alors précédés de DET). Nous constatons finalement qu'à partir de l'âge de 5 ans, les proportions des différentes étiquettes se stabilisent.

Bien que nous puissions déjà tirer quelques constats intéressants de ces courbes, nous ne pouvons pas caractériser l'acquisition syntaxique chez les enfants à partir de simples catégories isolées. Nous avons donc décidé d'utiliser des techniques de fouille de données séquentielles pour extraire des n-grammes d'étiquettes caractérisant les constructions syntaxiques des enfants selon différentes tranches d'âge. Nous présentons la méthode utilisée dans la section suivante.

4. Extraction de patrons d'étiquettes

4.1 Présentation des motifs (ou patrons) séquentiels

La fouille de données séquentielle a été introduite dans (Srikant et Agrawal, 1995). Elle s'appuie sur une relation d'ordre total existant entre certaines données pour découvrir des régularités apparaissant sous forme de séquences. La Table 4 montre des exemples de séquences de couples (mot, étiquette) présents dans nos corpus. Elle nous servira à illustrer différentes notions utiles par la suite.

sid	séquences
1	< (le DET) (petit ADJ) (chat NC) >
2	< (le DET) (grand ADJ) (arbre NC) >
3	< (le DET) (chat NC) >
4	<(tombé VPP) (et CC)(cassé VPP) >

Table 4 : Exemple de séquences d'itemsets (mot, étiquette)

Un *itemset I*, noté I = ($i_1 \ldots i_n$) est un ensemble de littéraux appelés *items*. Le processus d'étiquetage avec SEM permet en effet d'associer à chaque mot sa catégorie morpho-syntaxique. Ainsi, (chat NC) est un itemset contenant deux items *chat et NC*. Une *séquence S*, notée s = <$I_1 \ldots I_m$> est une liste ordonnée d'itemsets. Une séquence S = <$I_1 \ldots I_n$> est *incluse* dans une autre séquence S_2 = <$I'_1 \ldots I'_m$> s'il existe des entiers $1 \leq j_1 < \ldots < j_n \leq m$ tels que $I_1 \subseteq I'_{j1}, \ldots, I_n \subseteq I'_{jn}$. La séquence S_1 est appelée *sous-séquence* de S_2, et on note $S_1 \leq S_2$. Par exemple <(DET) (ADJ)> est incluse dans <(le DET)(petit ADJ)(NC)>.

Une séquence S_1 est incluse dans S si $S_1 \leq S$. Le *support* d'une séquence S_1 dans une base de séquences *SDB*, noté *sup(S_1)*, est le nombre de séquences contenant S_1 dans la base. Par exemple, dans la Table 4, sup(<(ADJ)(NC)>) = 2. Un autre type de support est également utilisé, c'est le *support relatif*, défini selon la formule suivante :

$$sup(S_1) = \frac{|\{(sid, S) \mid (sid, S) \in SDB \wedge (S_1 \preceq S)\}|}{|SDB|}$$

Les algorithmes de fouille de motifs séquentiels s'appuient sur un seuil minimal pour extraire les motifs fréquents. Un *motif fréquent,* est donc une séquence dont le support est supérieur ou égal au seuil fixé *minsup*. Outre le seuil du *minsup*, d'autres notions sont utiles pour limiter le nombre de motifs extraits.

4.2 Extraction des motifs séquentiels sous contraintes

Les travaux réalisés dans (Yan et al. (2003)) ont introduit la notion de motifs *fermés* (ou *clos*), qui permettent d'éliminer les redondances sans pertes d'information. Un motif fréquent S est *clos*, s'il n'existe aucun motif fréquent S' tel que S ≤ S' et sup(S) = sup(S'). Par exemple, si on fixe minsup à 2, le motif fréquent <(DET) (NC)> extrait de la Table 4 n'est pas clos car il est inclus dans le motif séquentiel <(le le DET)(NC)> et ils ont tous les deux un support de 3 ; par contre le motif <(DET)(petit petit ADJ)(NC)> est clos.

Une contrainte de longueur, basée sur un nombre minimal et un nombre maximal d'itemsets contenus dans un patron (Béchet et al., 2012), est également possible.

4.3 Algorithme pour extraire les motifs séquentiels

Il existe dans la littérature plusieurs algorithmes permettant d'extraire des motifs séquentiels tels que GSP (Srikant et Agrawal, 1996), SPADE (Zaki, 2001) ou encore, pour extraire des motifs séquentiels clos, CloSpan (Yan et al., 2003) et BIDE (Wang et al., 2004). L'outil SDMC[4] utilisé ici s'appuie sur la méthode proposée dans (Pei et al., 2001). L'algorithme est fondé sur la notion de *pattern growth* et est brièvement discuté dans (Béchet et al., 2012), il permet d'extraire des motifs séquentiels sous plusieurs contraintes.

4.4 Motifs séquentiels émergents

L'approche proposée dans (Dong et Li, 1999) a introduit la notion de motifs *émergents*. Un motif séquentiel sera dit émergent si son support relatif dans un ensemble de données est significativement plus haut que dans un autre ensemble de données. Formellement, un motif séquentiel *P* d'un ensemble de données D_1 est un motif émergent par rapport à un autre ensemble de données D_2, si *GrowthRate(P)* ≥ ρ, avec ρ ≥ 1. Le taux de croissance (growth rate) est défini comme suit :

$$GrowthRate(P) = \begin{cases} \infty, & \text{if } sup_{D_2}(P) = 0 \\ \frac{sup_{D_1}(P)}{sup_{D_2}(P)}, & \text{otherwise} \end{cases}$$

où $sup_{D1}(P)$ (respectivement $sup_{D2}(P)$), est le support relatif du motif P dans D_1 (respectivement dans D_2). Tout motif P dont le support est nul dans un ensemble de donnée est négligé.

4.4 Expérimentation

4.4.1 Paramètres pour l'extraction des motifs séquentiels d'items

Les corpus utilisés dans nos expériences sont ceux présentés en section 2.3. Nous nous intéressons aux motifs d'items restreints aux étiquettes POS (correspondant donc à des n-grammes d'étiquettes), sous contraintes, afin d'en limiter le nombre. Pour fixer le nombre d'items contenus dans une séquence, nous nous sommes basés sur la taille moyenne des phrases qui varie de 1 à 7 : nous avons donc décidé de prendre une longueur de motifs qui varie de 1 à 10. Le support minimal minsup est fixé à 2 et le seuil d'extraction des motifs

[4]https ://sdmc.greyc.fr/, login et mot de passe sont à demander aux auteurs

émergents ρ à 1.001. Pour déterminer quels sont les patrons émergents d'une certaine tranche d'âge, nous procédons comme (Quiniou et al., 2012) l'a fait pour étudier les différences entre genres littéraires : nous comparons cette tranche d'âge à l'ensemble de toutes les autres.

4.4.2 Résultats quantitatifs

Cette section présente les résultats obtenus lors de l'extraction des motifs fréquents et émergents dans les corpus des différentes tranches d'âges. La Figure 2 montre que la notion de motifs émergents permet d'éliminer un grand nombre de motifs fréquents. Ainsi, dans la tranche d'âge des « 4-5 ans », il y a 1933 motifs fréquents mais seulement 842 émergents (42,6%). Les ensembles de patrons émergents sont plus réduits donc plus faciles à analyser et sont plus caractéristiques de la tranche d'âge. L'interprétation de ces courbes et des patrons eux-mêmes reste délicate, et devra être soumise à des spécialistes de l'acquisition des langues.

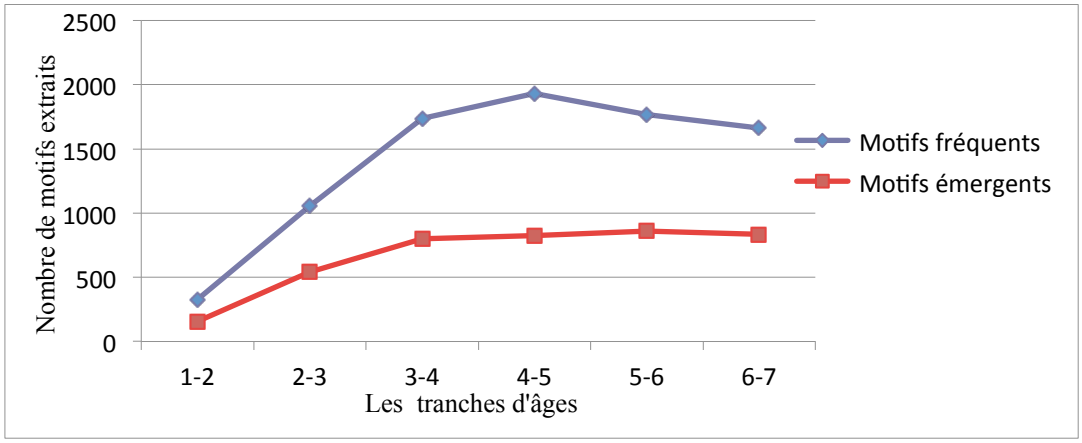

Figure 2 : Nombre de motifs fréquents et émergents des différentes tranches d'âges

La Figure 3 montre quant à elle que la taille moyenne des motifs est croissante en fonction de l'âge des enfants. La taille moyenne atteint une valeur maximale (environ 6 items par motifs) à l'âge de 5 ans et se stabilise ensuite. Ces paramètres semblent corrélés à la taille des phrases produites dans les mêmes tranches d'âge : non seulement les phrases deviennent de plus en plus longues, mais elles représentent des patrons eux-mêmes de plus en plus grands.

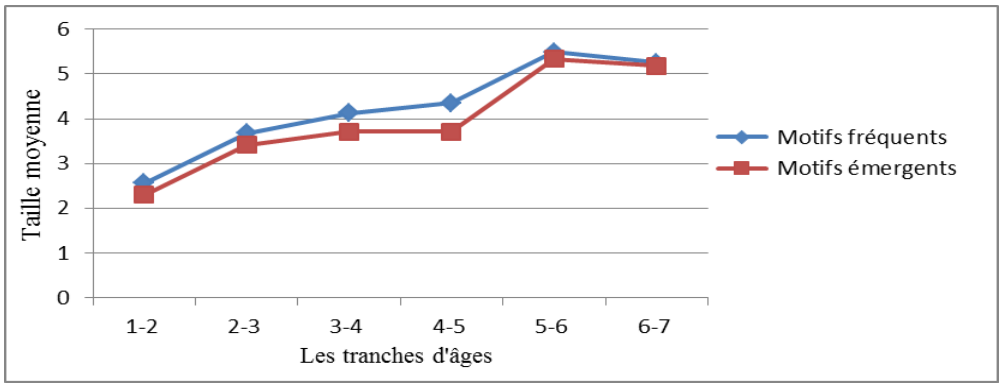

Figure 3 : Taille moyenne des motifs fréquents et émergents des différentes tranches d'âges

Les Figures 4 et 5 montrent la répartition des principales étiquettes morpho-syntaxiques dans les motifs extraits (fréquents et émergents) selon les différentes tranches d'âges. Ces résultats

sont cohérents avec ceux obtenus en étiquetant l'intégralité des corpus avec *SEM* réappris (cf. Figure 1). La proportion d'interjections diminue continument, alors que celle des prépositions augmente, ce qui est cohérent avec des constructions syntaxiques de plus en plus complexes. Nous remarquons également que les courbes CLS et V sont parallèles et que, jusqu'à l'âge de 4 ans, l'étiquette NC est très fréquente sans être associée à l'étiquette DET. Ces courbes indiquent que les proportions d'étiquettes dans les patrons sont similaires à celles de l'ensemble du corpus : en ce sens, les patrons sont donc bien *représentatifs*.

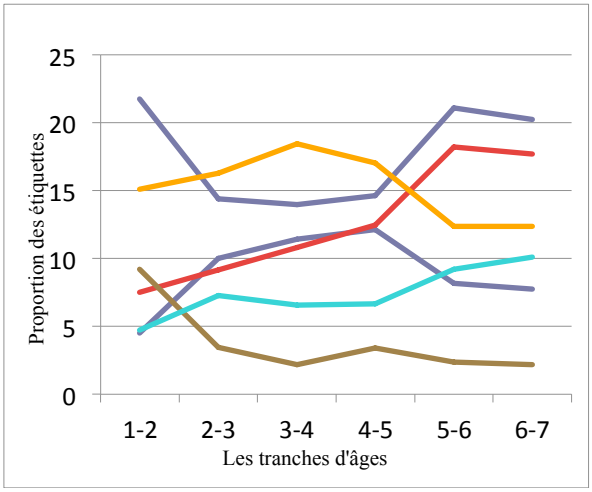

Figure 5 : Proportion des étiquettes dans les motifs fréquents

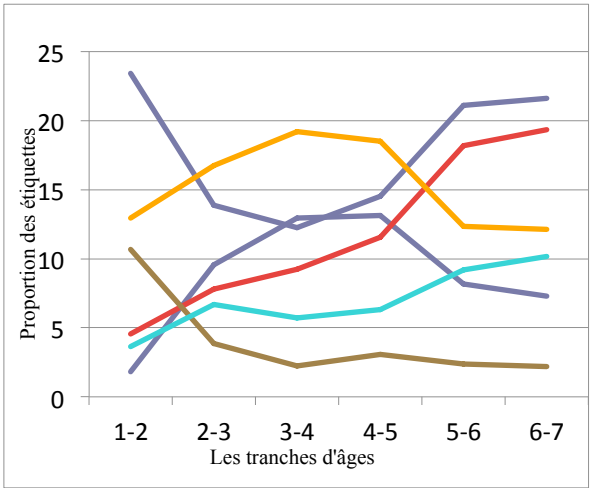

Figure 6 : Proportion des étiquettes dans les motifs émergents

4.4.3 Résultats qualitatifs

La Table 5 donne des exemples de motifs émergents et de phrases correspondantes. Ces motifs ont été sélectionnés pour montrer l'intérêt des patrons d'étiquettes, qui « couvrent » plusieurs phrases, et pour montrer l'évolution des productions syntaxiques d'un âge à l'autre

Nous remarquons que même avant l'âge de 2 ans, les enfants produisent des phrases avec des NC précédés par des DET. Nous constatons par exemple, que le motif «{DET}{NC}» et le motif «{DET}{NC}{CLS}{V}{VINF}» extraits respectivement des tranches d'âges « 1-2 ans » et « 4-5 ans » sont inclus respectivement dans les motifs «{P}{DET}{NC}» et «{DET}{NC}{CLS}{V}{VINF}{DET}{NC}» des tranches d'âges suivantes. C'est cohérent avec une acquisition progressive de constructions syntaxiques complexes.

Corpus	**Patrons d'étiquettes**	**Exemples**
« 1-2 ans »	{P}{NC}	**- à maman**. - sac **à dos.**
	{DET}{NC}	**- le ballon** ! **- des abeilles**.
« 2-3 ans »	{P}{DET}{NC}	**- de la tarte.** - poissons **dans l'eau**.
	{ADVWH}{CLS}{V}	**- où il est** ? **- comment il marche** ?

« 3-4 ans »	{ADV}{CLS}{V}	**- non il est** par terre**.** **- ici il pourra** passer**.**
	{ADVWH}{CLS}{CLO}{V}	**- comment on le voit ?** **- pourquoi tu y vas ?**
« 4-5 ans »	{ADV}{CLS}{CLO}{V}	**- alors tu m'as** vue ? **- oui j'en fais** souvent.
	{DET}{NC}{CLS}{V}{VINF}	**- les lapins ils vont rentrer**. **- le chat il veut attraper** l'oiseau.
« 5-6 ans »	{DET}{NC}{CLS}{V}{VINF}{DET}{NC}	**- l'enfant il va chercher le chat**. **-le monsieur il va chercher les cerises**.
	{CC}{DET}{NC}{CLS}{V}{DET}{NC}	**-** la maman **et le papa ils regardaient le garçon**. **- et le chat il mange les cerises.**
« 6-7 ans»	{P}{VINF}{DET}{NC}	-les oiseaux les aident **à ramasser les cerises**. -il y a un chat qui essaye **de chasser des oiseaux**.
	{DET}{NC}{PROREL}{V}{DET}{NC}{P}{DET}{NC}	**-** il y a **un chat qui suit la fille avec son panie**r. **-** et aussi **un monsieur qui ramasse des cerises dans un arbre.**

5. Conclusion

Dans cet article, nous avons appliqué des techniques issues du TAL, de l'apprentissage automatique et de la fouille de données séquentielles pour étudier l'évolution de productions d'enfants de différentes tranches d'âge. La phase d'annotation morpho-syntaxique a ainsi nécessité l'apprentissage d'un étiqueteur spécifique, adapté à nos données. C'était un préalable indispensable car les étiqueteurs standards ne traitent pas correctement les transcriptions orales : les interjections, par exemple, très spécifiques de l'oral, auraient été très mal analysées sans ré-apprentissage, or leur fréquence apparaît comme un indice important pour caractériser la tranche d'âge d'un enfant.

L'approche utilisée dans ce travail pour l'extraction des motifs, inspirée de (Quiniou et al., 2012), est non supervisée. Elle permet d'extraire des motifs fréquents et émergents d'items sous contraintes qui prennent la forme de *patrons grammaticaux*. Nous nous sommes restreints pour le moment aux n-grammes d'étiquettes mais un travail plus poussé pourrait bien sûr exploiter des itemsets plus riches du type (mot, lemme, étiquette). Nos mesures semblent confirmer que les patrons extraits sont représentatifs de la tranche d'âges d'où ils

proviennent. Les exemples fournis suggèrent en outre que non seulement la taille des patrons augmente quand l'enfant avance en âge, mais aussi que les patrons des tranches d'âges croissantes sont inclus les uns dans les autres, allant dans le sens d'une sophistication grammaticale. Ce phénomène reste à confirmer par des mesures précises.

L'analyse fine des patrons obtenus reste également à faire, mais ils constituent à n'en pas douter des outils précieux pour l'étude des phases d'acquisition du langage.

Remerciement

Nous remercions Christophe Parisse pour ses conseils et avis.

Références

Agrawal R. and Srikant R. (1995). Mining sequential patterns. In Int. Conf. on Data Engineering : IEEE.

Alishahi, A (2010). Computational modeling of human language acquisition (Synthesis lectures on human language technologies). San Rafael: Morgan & Claypool Publisher.

Abeillé A., Clément L. , Toussenel (2003). Building a treebank for French, in Abeillé, A., éditeur: Treebanks. Kluwer, Dordrecht.

Béchet, N., Cellier, P., Charnois T., et Crémilleux B. (2012). Discovering linguistic patterns using sequence mining. In proceedings of CICLing'2012, pp.154–165.

Biber D. (2009), A corpus-driven approach to formulaic language in English. International Journal of Corpus Linguistics, 14(3).

Brown, R. W. (1973). *A first language: the early stages*. Cambridge, Mass.: Harvard University Press.

Cellier, P., Charnois, T., Plantevit, M. (2010). Sequential Patterns to Discover and Characterise Biological Relations. In: Gelbukh, A. (ed.) CICLing 2010. LNCS, vol. 6008, pp. 537–548. Springer, Heidelberg.

Charnois T., Plantevit M., Rigotti C., & Crémillieux B. (2009). Fouille de données séquentielles pour l'extraction d'information. Traitement Automatique des Langues, 50(3).

Chater, N., and Manning C. D (2006). Probabilistic models of language processing and acquisition. Trends in Cognitive Science, 10(7) 335-344

Clément L., Sagot B., Lang (2004). Morphology based automatic acquisition of large-coverage lexica (LREC 2004), Lisbonne.

Crabbé B., Candito M. (2008). Expériences d'analyse syntaxique du français, in Actes de TALN 2008 (Traitement automatique des langues naturelles), Avignon.

Dong G et Li J. (1999). Efficient mining of emerging patterns: Discovering trends and differences. In Proc. of SIGKDD'99.

Dong G and Pei J. (2007). Sequence Data Mining. Springer. Hunston S. et Francis J. (2000). Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English. Amsterdam/Philadelphia.

Elman, J (2001). Connectionism and langauage acquisition. In Essential readings in langauage acquisition. In Oxford : Blackwell.

Giuliano C., Lavelli A. and Romano L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. In Proc. of the Conf. of the European Chapter of the Association for Computational Linguistics : The Association for Computer Linguistics.

Hobbs J.R and Riloff E. (2010). Information extraction. In N. INDURKHYA & F. J. DAMERAU, Eds., Handbook of Natural Language Processing, Second Edition. Boca Raton, FL : CRC Press, Taylor and Francis Group.

Krallinger M., Leitner F. Rodriguez-Penagos C. and Valencia A. (2008). Overview of the protein-protein interaction annotation extraction task of BioCreative II. Genome Biology.

Lafferty, J., Mccallum, A. and Pereira, F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In Proceedings of ICML 2001, pages 282-289.

Miller, J. F. and Chapman, R. S. (1981). The relation between age and mean length of utterance in morphemes. Journal of Speech and Hearing Research, 24, 154–161.

Nanni, M. and C. Rigotti (2007). Extracting trees of quantitative serial episodes. In Proc. Of KDID'07, pp. 170–188.

Nouvel D., Antoine J-Y., Friburger N., Soulet A. (2013). Fouille de règles d'annotation partielles pour la reconnaissance d'entités nommées. TALN'13,

Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu,M.(2001). Prefixspan: Mining sequential patterns by prefix-projected growth. In: ICDE, pp. 215–224. IEEE Computer Society.

Renouf A. et Sinclair J. (1991). Collocational Frameworks in English. English Corpus Linguistics: Studies in Honour of Jan Svartvik. Longman.

Riloff, E. (1996). Automatically generating extraction patterns from untagged text. In: AAAI/IAAI .

Rinaldi F., Schneider G., Kaljurand K., Hess M. & Romacker M. (2006). An environment for relation mining over richly annotated corpora : the case of genia. BMC Bioinformatics, 7(S-3).

Srikant, R., Agrawal, R. (1996). Mining Sequential Patterns: Generalizations and Performance Improvements. In: Apers, P.M.G., Bouzeghoub, M., Gardarin, G. (eds.) EDBT 1996. LNCS, vol. 1057, pp. 3–17. Springer, Heidelberg.

Tellier I., Duchier D., Eshkol I., Courmet A., Martinet (2012). Apprentissage automatique d'un chunker pour le français, Traitement Automatique des Langues Naturelles, (TALN 2012, papier court), Grenoble.

Tellier I., Dupont Y., Eshkol I., Wang I. (2013). Adapt a Text-Oriented Chunker for Oral Data: How Much Manual Effort is Necessary?, The 14th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'2013), Special Session on Text Data Learning, LNAI, Hefei (Chine).

Tsuruoka, Y., Tsujii, J. et Ananiadou, S. S. (2009). Fast full parsing by linear-chain conditional random fields. In Proceedings of EACL 2009, pages 790–798.

Wang, J., Han, Bide J. (2004). Efficient mining of frequent closed sequences. In: ICDE, pp. 79–90. IEEE Computer Society.

Quiniou, Cellier P., Charnois T., Legallois D. (2012). Fouille de données pour la stylistique : cas des motifs séquentiels émergents. Proceedings of the 11th International Conference on the Statistical Analysis of Textual Data, Liege.

Yan, X., Han, J., Afshar, R.: Clospan (2003). Mining closed sequential patterns in large databases. In: Barbara, D., Kamath, C. (eds.) SDM. SIAM.

Zaki, M. J. (2001). SPADE: An efficient algorithm for mining frequent sequences. Machine Learning Journal 42(1/2), 31–60.