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 UMR 8094, [2] Université Paris 3 - 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 syntactic 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-syntactic labels). We show that the distinct ages can be characterized by variations of proportions of morpho-syntactic labels, which can also be seen 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, en particulier la manière dont les constructions grammaticales sont peu à peu maîtrisées, reste en grande partie mystérieux. Pour aborder cette question avec des outils de traitement automatique des langues, on peut chercher à modéliser l'apprentissage d'une langue par des programmes (Alishali, 2010 ; Chater et al., 2006). Notre approche dans cet article est différente : elle consiste à étudier certaines propriétés morphosyntaxiques des 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). Quand les séquences extraites sont des portions contiguës de textes, cela revient à chercher les segments répétés de ces textes (Salem, 1986).

Pour les données textuelles, les éléments pris en compte peuvent être les mots eux-mêmes, leur lemme, ou leur catégorie syntaxique. L'utilisation de la fouille de données séquentielles à des textes 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 (ou patrons) d'étiquettes morpho-syntaxiques qui retiendront ici notre attention. Ils sont en effet plus généraux que les simples suites de mots et fourniront une caractérisation plus abstraite d'une tranche d'âge donnée. Nous chercherons en particulier à exhiber des patrons spécifiques « émergents » de chaque tranche d'âge.

La suite de l'article présente tout d'abord le mode de constitution de 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 la technique utilisée pour l'extraction de n-grammes d'étiquettes morpho-syntaxiques spécifiques de chacune des tranches d'âge, 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 métadonnées du corpus, nous avons constitué six corpus différents correspondant à six tranches d'âge : de « 1-2 ans » jusqu'à « 6-7 ans ».

2.2. Prétraitements

Dans ces corpus, les enfants et les parents communiquent par tours de parole. Chaque tour de parole est transcrit dans une ligne délimitée par un point. Par la suite, nous considérons que cette ligne correspond à une phrase. 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 donc 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. Et nous avons éliminé de nos données toutes les prises de parole des

[1] Centre National des Ressources Textuelles et Linguistiques : http://www.cnrtl.fr

[2] http://childes.psy.cmu.edu/

[3] http://childes.psy.cmu.edu/manuals/CHAT.pdf

adultes, pour nous concentrer uniquement sur les productions des enfants. 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. Pour équilibrer les corpus des différentes tranches d'âge, 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. Échantillonnage

Le plus petit corpus en termes de mots (celui des « 6-7 ans ») 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 cet échantillonnage, nous disposons donc de six nouveaux corpus, dont les propriétés sont données 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. Le nombre total de phrases de ces corpus a bien sûr diminué, mais nous remarquons que les tailles moyennes des phrases qu'ils contiennent progressent de manière similaire à celles des corpus de base : plus les enfants 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 et Chapman, 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 des mots. 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-Etiqueteur Markovien (SEM), qui résulte d'un apprentissage automatique à l'aide de CRF (Conditional Random Fields) (Tellier et al., 2012) appliqué sur le corpus arboré *French Treebank* (Abeillé et al., 2003). Le jeu d'étiquettes adopté dans SEM, similaire à celui de (Crabbé et al., 2008), comprend 30 catégories différentes. SEM intègre également une ressource externe le *LeFFF*, le Lexique des Formes Fléchies du Français (Clément et al., 2004) pour l'aider à réaliser 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) et les productions d'enfants sont souvent loin du français standard. Pour évaluer la qualité de SEM sur nos données, nous avons tiré aléatoirement 200 phrases de chacun de nos six corpus, les avons étiquetées avec SEM et 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 dans tous les cas des 97% atteints sur des données écrites proches de celles sur lesquelles il a été entrainé.

3.2. 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 que pour obtenir un étiqueteur performant par apprentissage automatique supervisé, il était plus efficace de disposer d'un corpus source annoté petit mais similaire aux données cibles qu'un corpus plus grand mais trop différent. 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 nos données. Nous avons pour cela reconduit les expériences qui avaient permis l'apprentissage de SEM, en utilisant les CRF introduits par (Lafferty et al., 2001). Les CRF sont des modèles graphiques ayant largement fait leur preuve dans le domaine de l'annotation par apprentissage automatique supervisé (Tsuruoka et al., 2009 ; Tellier et al., 2012). Ils permettent d'attribuer la meilleure séquence d'annotations y à une séquence observable x. Pour nous, les éléments de x sont les mots auxquelles sont associés des attributs 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, équitablement réparties dans les différents corpus. Les valeurs de l'exactitude des deux étiqueteurs (avant et après réapprentissage) sur chaque corpus pour les données test sont fournies 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 le réapprentissage apporte une amélioration en moyenne de l'ordre de 10 % d'exactitude. Il est donc très bénéfique, malgré un corpus d'apprentissage limité. Ceci peut s'expliquer par le fait que le vocabulaire employé dans ces textes est relativement limité et redondant : peu de données suffisent donc à obtenir un étiqueteur efficace sur nos corpus (il le serait évidemment moins sur d'autres types de données). Nous utilisons dans la suite SEM réappris pour étiqueter l'ensemble des six corpus.

3.3. Analyse des étiquettes POS

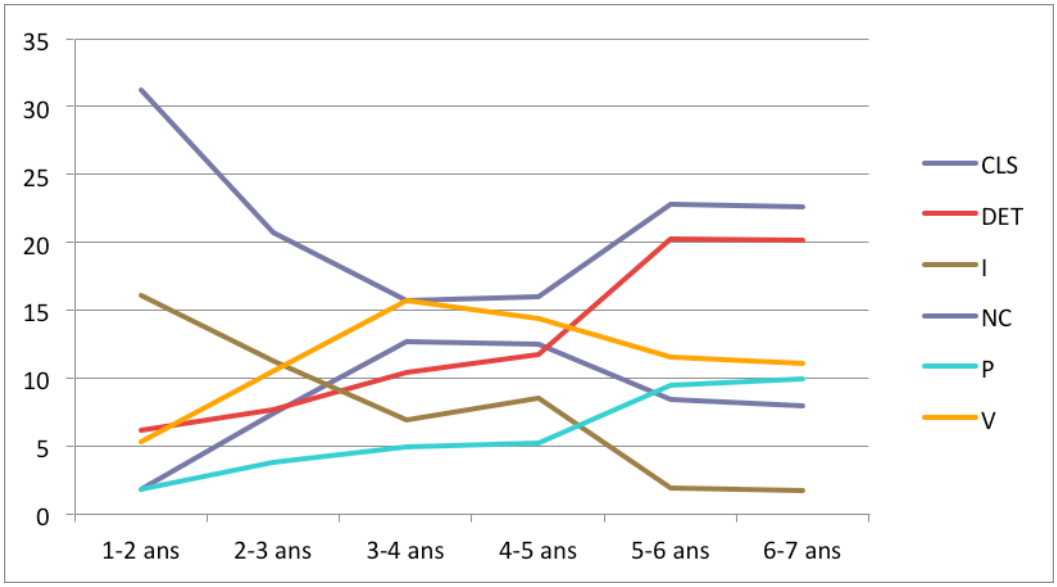

Figure 1. Proportion des différentes étiquettes dans les corpus

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. Les erreurs résiduelles de l'étiqueteur ré-appris incitent à nuancer ces observations. Mais il est évident que certains des phénomènes observés ici n'auraient pas pu l'être sans ré-apprentissage : les interjections, par exemple, sont les catégories les plus mal reconnues par le SEM original, car elles sont très peu présentes dans les articles journalistiques. Or, leur production semble être un indice important de la tranche d'âge de l'enfant. Par exemple, nous avons remarqué dans l'un des corpus les phrases suivantes « ah maman » et « euh voilà » étiquetées respectivement comme « ADJ NC » et « ADV V », avec le SEM initial. Après ré-apprentissage, les étiquettes deviennent « I NC » et « « I V », ce qui est un peu plus correct. 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 aller plus loin dans notre exploration.

4. Extraction de patrons d'étiquettes

4.1. Présentation des motifs (ou patrons) séquentiels

De nombreux travaux se sont intéressés à l'analyse de données textuelles, notamment (Salem, 1986). Ils s'appuient sur une relation d'ordre total existant entre les données pour découvrir des régularités appelées *segments répétés*. Ces derniers sont des suites d'unités appelées *formes*. La fouille de données séquentielles, introduite dans (Srikant et Agrawal, 1995), se fonde sur le même principe et la notion de segments répétés y est connue sous le nom de *séquences* ou *motifs (ou patrons) séquentiels*. Cependant, les formes (ou items, dans le vocabulaire de la fouille de données) peuvent correspondre à d'autres types d'unités que de simples mots, par exemple des couples (mot, étiquette POS). La table 4 montre des exemples de séquences de tels couples présents dans nos corpus après étiquetage avec SEM ré-appris. Elle nous servira à illustrer les différentes notions introduites.

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 de formes (mot, étiquette POS)

Le *support* d'une séquence S_1, noté $sup(S_1)$, est égal au nombre de phrases du corpus la contenant. Par exemple, dans la table 4, sup(<(ADJ)(NC)>) = 2. Le *support relatif* d'une séquence S_1 vaut quant à lui la *proportion* de séquences contenant S_1 dans la base de séquences initiales. Elle vaut ½ pour <(ADJ)(NC)> dans notre exemple, car cette séquence est présente dans 2 des 4 séquences de départ. 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 <(le DET)(NC)> et ils ont tous les deux un support de 3 ; par contre le motif <(DET)(petit ADJ)(NC)> est clos. La contrainte de longueur peut également être utilisée. Elle définit le nombre minimal et le nombre maximal de formes contenus dans un patron (Béchet et al., 2012).

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). Il permet d'extraire plusieurs types de motifs séquentiels, dans lesquels les items (ou formes) peuvent correspondre aux simples mots, à leur lemme et/ou à leur catégorie morpho-syntaxique. Nous nous intéressons ici uniquement aux catégories grammaticales, car elles sont plus générales que les mots. L'algorithme utilisé est brièvement discuté dans (Béchet et al., 2012), il permet d'extraire des motifs séquentiels sous plusieurs contraintes.

4.4. Motifs séquentiels émergents

(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 émergent par rapport à un autre ensemble de données D_2 si on a $GrowthRate(P) \geq \rho$, avec $\rho \geq 1$, le taux de croissance (*growth rate*) étant défini par :

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

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 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, ou encore des segments répétés 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 : nous avons donc décidé de prendre une longueur de motifs qui varie de 1 à 10. Le support minimal *minsup* à atteindre pour être considéré comme un motif fréquent est fixé à 2 et le seuil d'extraction des motifs 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

La figure 2 compare les motifs fréquents et émergents, et 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 1 933 motifs fréquents mais seulement 842 émergents (42,6%). Les ensembles de patrons émergents sont plus petits, donc plus faciles à analyser, et 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.

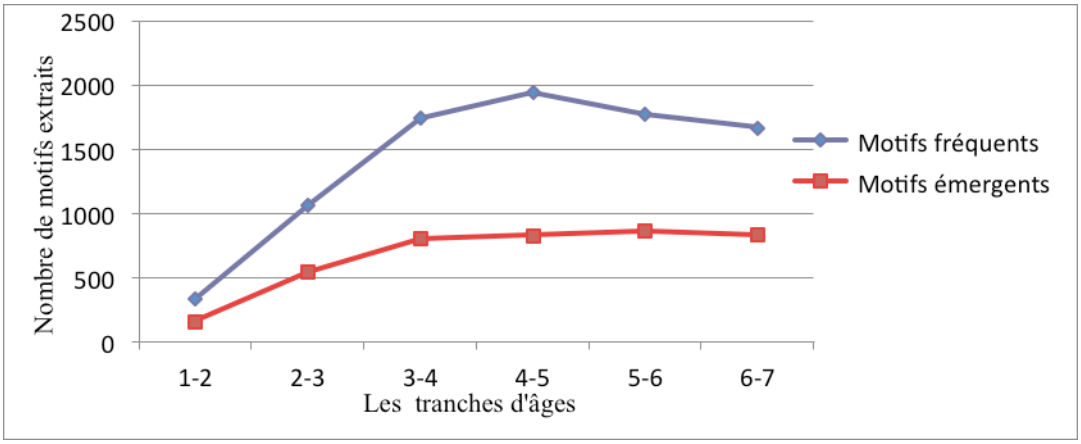

Figure 2. Nombre de motifs fréquents et émergents des différentes tranches d'âges

La figure 3 montre quant à elle que la taille moyenne des motifs est croissante en fonction de l'âge des enfants. Elle atteint une valeur maximale (environ six 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 contiennent des patrons eux-mêmes de plus en plus grands.

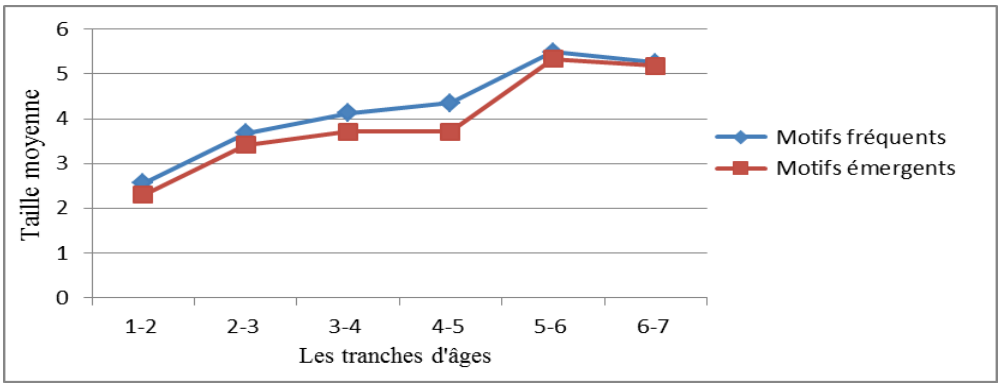

Figure 3. Taille moyenne des motifs fréquents et émergents des différentes tranches d'âges

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 sur l'ensemble du corpus (*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 des étiquettes dans les patrons sont similaires à celles de l'ensemble du corpus : en ce sens, les patrons sont donc bien *représentatifs* des différentes tranches d'âge.

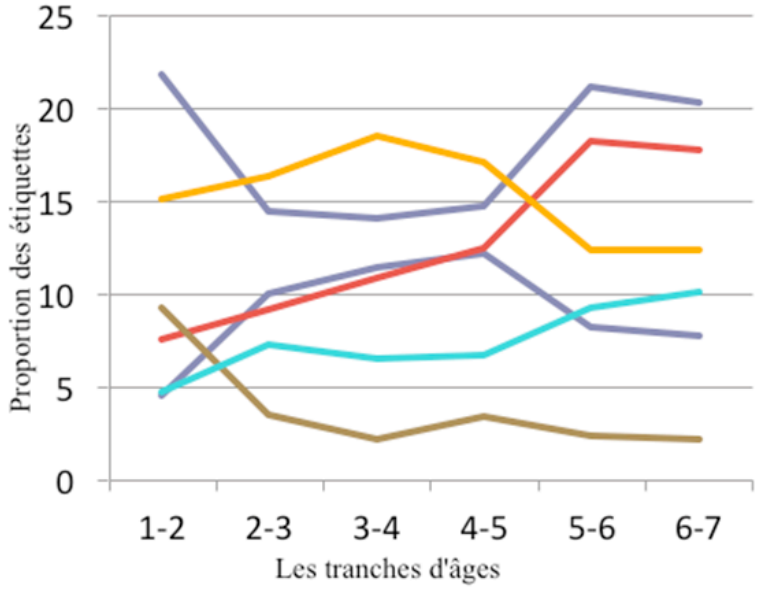

Figure 4. Proportion des étiquettes dans les motifs fréquents

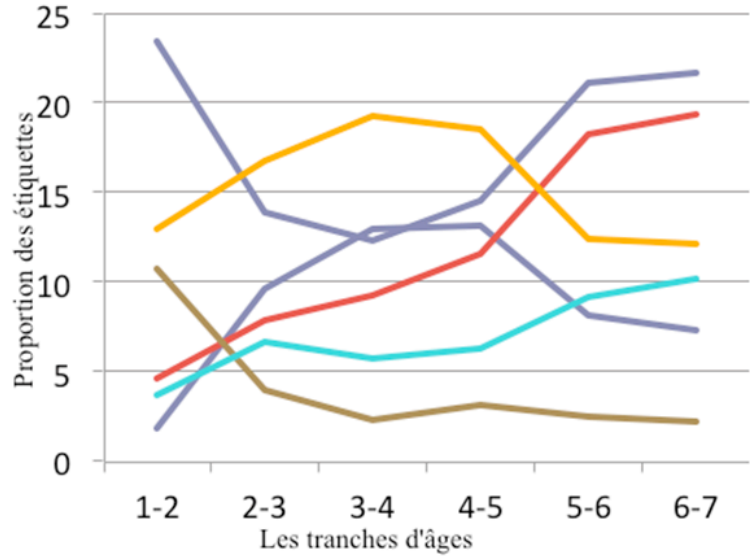

Figure 5. Proportion des étiquettes dans les motifs émergents

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'âge « 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'âge suivantes. C'est cohérent avec une acquisition progressive de constructions syntaxiques complexes.

Corpus	Patrons d'étiquettes	Exemples
« 1-2 ans »	{P}{NC} {DET}{NC}	**- à maman**. - sac **à dos.** **- le ballon** ! **- des abeilles**.
« 2-3 ans »	{P}{DET}{NC} {ADVWH}{CLS}{V}	**- de la tarte.** **-** poissons **dans l'eau**. **- où il est** ? **- comment il marche** ?
« 3-4 ans »	{ADV}{CLS}{V} {ADVWH}{CLS}{CLO}{V}	**- non il est** par terre**.** **- ici il pourra** passer**.** **- comment on le voit ?** **- pourquoi tu y vas ?**
« 4-5 ans »	{ADV}{CLS}{CLO}{V} {DET}{NC}{CLS}{V}{VINF}	**- alors tu m'as** vue ? **- oui j'en fais** souvent. **- les lapins ils vont rentrer**. **- le chat il veut attraper** l'oiseau.
« 5-6 ans »	{DET}{NC}{CLS}{V}{VINF}{DET}{NC} {CC}{DET}{NC}{CLS}{V}{DET}{NC}	**- l'enfant il va chercher le chat**. **- le monsieur il va chercher les cerises**. **-** la maman **et le papa ils regardaient le garçon**. **- et le chat il mange les cerises.**
« 6-7 ans»	{P}{VINF}{DET}{NC} {DET}{NC}{PROREL}{V}{DET}{NC}{P}{DET}{NC}	-les oiseaux les aident **à ramasser les cerises**. -il y a un chat qui essaye **de chasser des oiseaux**. **-** 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, et encore moins celles des enfants : 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 pour l'extraction des motifs fréquents et émergents d'items sous contraintes est non supervisée. Nous nous sommes restreints pour le moment aux n-grammes

d'étiquettes mais un travail plus poussé pourrait bien sûr exploiter des formes plus riches du type (mot, lemme, étiquette POS). Nos mesures semblent confirmer que les patrons extraits sont représentatifs de la tranche d'âges d'où ils proviennent. Les exemples fournis confirment en outre l'intuition que les patrons des tranches d'âges croissantes sont inclus les uns dans les autres, en allant dans le sens d'une sophistication grammaticale. L'analyse fine des patrons obtenus reste à faire, mais ils constituent à n'en pas douter des outils précieux pour l'étude des phases d'acquisition du langage.

Remerciement

« Ce travail a bénéficié d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme Investissements d'Avenir portant la référence ANR-10-LABX-0083 ». Nous remercions Christophe Parisse pour ses conseils et avis.

Références

Agrawal R. et 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. et 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. et 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. et Crémillieux B. (2009). Fouille de données séquentielles pour l'extraction d'information. Traitement Automatique des Langues, 50(3).

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

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

Crabbé B. et 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 et 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 language acquisition. In Essential readings in langauage acquisition. In Oxford : Blackwell.

Giuliano C., Lavelli A. et 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 et 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. et Valencia A. (2008). Overview of the protein-protein interaction annotation extraction task of BioCreative II. Genome Biology.

Lafferty J., Mccallum A. et 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. et 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. et Rigotti C. (2007). Extracting trees of quantitative serial episodes. In Proc. Of KDID'07, pp. 170–188.

Nouvel D., Antoine J-Y., Friburger N. et 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. et 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. et Romacker M. (2006). An environment for relation mining over richly annotated corpora : the case of genia. BMC Bioinformatics, 7(S-3).

Salem André. Segments répétés et analyse statistique des données textuelles. In Histoire & Mesure, 1986 volume 1 - n°2. pp. 5-28.

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

Quiniou, Cellier P., Charnois T. et 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. et Afshar R. (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.