Glossa, n° 61 (22-29), 1998,

RÉSUMÉ :

Le traitement de corpus décrivant le langage, normal ou pathologique, du jeune enfant est un travail de longue haleine qui a depuis longtemps été facilité par l'existence de programmes sur ordinateur. Ceux-ci permettent de rationaliser les procédures, de retrouver et consulter rapidement les données, les diffuser aisément, etc. Toutefois, les traitements linguistiques qu'ils réalisent sont en général élémentaires ou absents, tout au plus trouve-t-on des tentatives de décomposition morphologique semi-automatique. Or, les recherches actuelles en linguistique automatique offrent de plus en plus de possibilités. L'étiquetage morphosyntaxique est notamment largement développé et son usage dans les bases de données textuelles s'accroît. Cet étiquetage consiste à déterminer automatiquement la classe lexicale des mots (en particulier des homographes) en fonction du contexte dans lequel ils se trouvent. Le traitement, totalement automatique, est basé sur un apprentissage des probabilités de successions de classes lexicales. La procédure d'analyse automatisée qui suit permet d'accélérer notablement le traitement des corpus textuels (même avec une vérification ou un contrôle a posteriori des résultats). Par ailleurs, la remarquable adéquation de la morphosyntaxe avec les caractéristiques du langage de l'enfant suggère, non seulement des applications pratiques pour l'évaluation, mais aussi des développements d'ordre théorique. Des exemples d'analyses et de traitements d'un corpus réel de langage d'enfants de deux à quatre ans sont fournis en conclusion.

MOTS-CLÉS :

Morphosyntaxe - Informatique - Linguistique - Petit enfant.

Christophe PARISSE
Chargé de Recherche
INSERM
et
Marie-Thérèse LE NORMAND
Directeur de Recherche
INSERM
Pavillon Claude Bernard
Hôpital de la Salpêtrière
47, Boulevard de l'Hôpital
75651 PARIS CEDEX 13
FRANCE

TRAITEMENT AUTOMATIQUE DE LA MORPHOSYNTAXE CHEZ LE PETIT ENFANT*

par Christophe PARISSE et Marie-Thérèse LE NORMAND

** Cette étude a bénéficié du soutien financier d'un contrat de recherche INSERM (40009B)*

SUMMARY : *Automatic processing of morphosyntax in young children*

The study of corpora describing the language, normal or pathological, of young children is a long term work which has long been made easier by computer programs. They rationalise procedures, speed up data recovery and consultation, facilitate its circulation, etc. Their linguistic processing are however usually either very simple or simply lacking ; at most, one may find attempts at semi-automatic morphological decomposition. Present research in computer linguistics offer more and more possibilities, with for instance, part of speech labelling being developed and increasingly in use in textual databases. This labelling consists in automatically providing the lexical tag of words (especially homographs) depending on the context in which the words are found. This fully automatic procedure is based on a lexical tag succession probabilities training. The automatic analysis procedure which follows speeds up the processing of textual corpora, even with a posteriori verification or control of results. Furthermore, the striking appropriateness of morphosyntax with the child language characteristics suggests concrete applications for assessment and further theoretical developments. Examples of analyses and computation of a real corpus of language of children aged from two to four is given at the end.

KEY WORDS :

Morphosyntax - Computer science - Linguistics - Young child.

INTRODUCTION

L'analyse morphosyntaxique consiste à rechercher la classe syntaxique des mots et leur décomposition en morphèmes. Les classes utilisées dans ce type d'analyse correspondent à celles que l'on trouverait dans un lexique. Il s'agit en particulier de la classe du mot hors contexte sémantique ou pragmatique. Toutefois, un mot peut correspondre à plusieurs classes lexicales, c'est-à-dire être un homophone ou un homographe. On dira dans ce cas qu'il est ambigu. Plusieurs entrées lexicales partagent alors la même entrée phonétique ou graphémique. L'analyse doit utiliser le contexte pour déterminer si la chaîne de caractères "porte" correspond au substantif féminin singulier "porte" ou au verbe conjugé "porter" au présent, soit à la première, soit à la troisième personne. L'analyse morphosyntaxique est particulièrement bien adaptée à l'étude de l'apprentissage du langage du jeune enfant. En effet, cette analyse se fonde sur deux éléments-clés pour celui-ci : la morphologie des mots et l'analyse des concurrences de classes syntaxiques.

Toutefois, malgré son intérêt, l'étiquetage morphosyntaxique de corpus n'est pas si souvent réalisé exhaustivement car cela nécessite beaucoup de temps. Ainsi, l'enfant de 2 ans ou plus utilise en français le mot/la/sous trois formes : 'la'article féminin singulier, 'la'pronom féminin singulier, 'là'adverbe de lieu. Si les corpus traités ont été transcrits de leur forme orale à leur forme écrite, il n'y a plus à traiter les ambiguïtés à l'oral (entre 'la'et'là', de même prononciation [LA]). Néanmoins, l'ambiguïté sur la forme écrite subsiste et doit être levée. Avec des corpus importants, le nombre de cas d'ambiguïtés peut être très grand. Ainsi, sur les 95000 mots recueillis chez des enfants de deux à quatre ans, et traités en exemple plus loin, 2900 occurrences de 'la'doivent être désambiguïsées entre article ou pronom (cas de l'écrit), 4600 occurrences devraient l'être entre article, pronom et adverbe (cas de l'oral). Pour ce corpus de 95000 mots, un étiquetage manuel prenant 10 secondes par mot prendrait 7 semaines de travail ininterrompu (sans aucune pose). On peut imaginer qu'en réalité il faudrait au moins un an. Si de telles réalisations ne sont pas impossibles, elles ne peuvent avoir qu'un caractère exceptionnel et limité. L'automatisation partielle du traitement morphosyntaxique semble nécessaire et offrirait des perspectives nouvelles d'exploitation des corpus. Nous proposons dans cet article l'utilisation d'un analyseur automatique utilisant les chaînes de Markov* pour aider le travail du linguiste et donnons un exemple de l'utilisation d'un tel programme.

Andreewsky, 1973 et Andreewsky, Delibi, Dehli, Fluhr, 1980

Etat de la question

Il existe de nombreux programmes sur ordinateur destinés à aider le chercheur, linguiste, psychologue ou psycholinguiste dans son travail, et en particulier dans l'étude des corpus linguistiques. Ces outils sont généralistes et peuvent être complétés par d'autres outils plus spécifiques du traitement linguistique, comme des outils pour découper en mots un texte, décomposer les mots en morphèmes, réaliser l'analyse grammaticale complète d'une phrase, qu'il s'agisse d'une relativement « simple » analyse morphosyntaxique ou d'une plus complète décomposition en groupes ou en arbres. Si une partie de ces outils existe déjà, comme l'attestent les travaux de Miller et Chapman*, de MacWhinney et Snow**, de Baker-Van den Goorbergh***, de Rondal et al*, il n'y a pas encore beaucoup de systèmes qui exécutent un travail linguistique automatique. Les seuls exemples sont la décomposition en morphèmes* pour les calculs de longueur moyenne d'énoncés ou pour d'autres types de statistiques. Les évaluations linguistiques plus complexes comme le « Language Assessment, Remediation and Screening Procedure » (LARSP) initié par Crystal, Fletcher et Garman* doivent être réalisées manuellement. Dans cet article, les termes «manuel» et «manuellement» signifient qu'il faut qu'un opérateur réalise la tâche de manière consciente, réfléchie et souvent fastidieuse en utilisant des outils de Bureautique traditionnels. A l'inverse, la notion de «traitement automatique» signifie que l'ordinateur réalise la totalité des traitements, y compris ceux qui nécessitent usuellement la décision d'un expert humain. Dans certains cas, ce traitement

** 1983 ** 1985 *** 1994 * 1987*

**Miller et Chapman 1983 ; Cappelli et coll. 1991*

** 1976*

automatique exige une supervision de l'opérateur humain (confirmation en direct de la validité des choix) ou une vérification a posteriori. Les recherches en linguistique automatique, et en particulier l'analyse morphosyntaxique, sont plus évoluées que le tableau des applications disponibles actuellement ne le laisse paraître. C'est pourquoi il est possible de laisser l'ordinateur faire tout le travail. Les taux d'erreurs sont en effet suffisamment faibles (autour de 3 % selon le type de corpus) pour que les résultats obtenus automatiquement soient utilisés directement par le linguiste, l'orthophoniste ou le médecin.

Il est possible d'appliquer le principe de l'analyse morphosyntaxique à toutes les langues, ce qui a d'ailleurs été déjà largement réalisé, en particulier pour formaliser la structure lexicale des mots d'une langue. Le traitement automatique de la morphosyntaxe subit un grand essor depuis une dizaine d'années car, d'une part un tel type d'analyse est souvent utilisé comme point de départ pour des traitements linguistiques ultérieurs plus complexes et, d'autre part, on sait de mieux en mieux réaliser des analyseurs automatiques de ce type. Ce type d'analyseur, appelé « part-of-speech tagger » peut se trouver sur Internet, comme par exemple le Xerox's Part of Speech Tagger (ftp ://parcftp. xerox. com/pub/tagger/) ou le Brill's Transformation Based Tagger (ftp ://blaze. cs. jhu. edu/pub/brill/Programs/). Il existe une certaine gradation des types de catégories qui sont couvertes par la morphosyntaxe dans ses applications automatiques. Cette gradation tire son origine, soit du type d'application ou d'usage envisagé, soit de la nature lexicale de la langue traitée. Ainsi certains analyseurs ne traitent pas les accords en genre et en nombre car, soit l'application envisagée ne nécessite pas cette distinction, soit cette distinction n'est pas marquée dans le lexique de la langue ou ne modifie pas la structure des phrases (les deux remarques s'appliquent au cas du genre en anglais, la deuxième s'applique presque toujours au cas du nombre en français). Par ailleurs, l'identification du cas grammatical à l'intérieur d'une même classe syntaxique - sujet, objet direct ou objet indirect (agent) ne sont marqués en français que par leur position relative - est rarement traitée par la morphosyntaxe, sauf si la langue analysée présente un système de marque de cas qui résoud directement ce problème. Dans les langues n'utilisant pas de système de cas, l'ordre des mots est en général strict et permettrait à première vue d'aborder cette question. En réalité, hormis quelques structures simples, la complexité structurelle fait sortir la détermination du sujet et de l'objet du champ de la morphosyntaxe dans ce type de langue. On touche ici les limites de ce type de syntaxe qui fonctionne en général en contexte très limité et n'utilise aucune connaissance sémantique. Toutefois, la morphosyntaxe reste un système linguistique complet qui entremêle traitements lexicaux et syntaxiques.

MÉTHODOLOGIE

L'analyseur morphosyntaxique que nous avons utilisé a été mis au point pour traiter automatiquement différentes langues européennes à structure positionnelle ou semi-positionnelle comme le français, l'anglais, l'allemand, le néerlandais, l'italien, l'espagnol, le grec et pour aider à réduire la complexité des traitements de reconnaissance automatique de l'écrit*. Son but est d'étiqueter automatiquement des textes suite à un apprentissage aussi réduit que possible. Il est fondé sur un modèle markovien de succession de règles de résolution de biclasses ambiguës. Lorsqu'il y a un choix entre plusieurs classes suivi d'un autre choix entre d'autres classes – par exemple, « la porte » : 'article' ou 'pronom' suivi de 'substantif' ou 'verbe conjugué'– il y a biclasse ambiguë. Il faut résoudre cette ambiguïté, c'est à dire décider de quel biclasse il s'agit – dans l'exemple précédent, seul 'article/substantif'et 'pronom/verbe conjugué' sont possibles. Il est ainsi bien adapté pour résoudre les problèmes d'ambiguïté lexicale, très nombreux en français et dans toutes les langues d'une manière générale.

** Parisse, 1989*

Spécificité

Avant toute analyse, il faut disposer d'un corpus étiqueté permettant au programme d'analyse automatique de réaliser un entraînement initial. Pour cela, on doit étiqueter manuellement quelques centaines de mots qui serviront de départ à une procédure cyclique. Ces mots forment le corpus d'apprentissage. Grâce à celui-ci, il est possible de réaliser un entraînement du programme puis d'analyser un corpus plus grand que l'on contrôlera manuellement et qui deviendra le nouveau corpus d'apprentissage. De proche en proche, on va être amené à contrôler et disposer de corpus de taille de plus en plus grande. Cette procédure peut toutefois mener à un écueil. Lorsque le corpus d'apprentissage devient très long, le travail est alors très fastidieux et tend à aller à l'encontre du but initial qui est d'économiser du travail et d'accélérer les temps de traitements. Même si une vérification est plus rapide qu'un étiquetage réalisé à partir de zéro, cela peut prendre beaucoup de temps. Pour éviter de tomber dans ce travers, il faut à un certain point donné décider que la qualité d'étiquetage réalisée est suffisante et ne plus vérifier les résultats de l'analyse de manière exhaustive, mais seulement là où l'analyseur est confronté à des difficultés, c'est à dire dans les situations grammaticales nouvelles. Ces situations apparaissent lorsque : 1) une biclasse ambiguë n'a jamais été rencontrée, 2) la suite des résolutions d'ambiguïtés ne permet pas d'obtenir de solution, 3) cette suite laisse plusieurs solutions possibles. Le cas 1) n'est plus très fréquent après un apprentissage important. Le cas 2) nécessite une vérification plus complète, mais il est souvent possible de corriger les erreurs de manière systématique une fois repérées manuellement. Dans ce cas, il y a manque d'apprentissage et les erreurs de ce type sont toujours les mêmes. Il suffit de rechercher tous les contextes identiques et une correction globale est alors possible. Le cas 3) est courant mais ne nécessite pas forcément de vérification exhaustive car les diverses solutions sont triées dans l'ordre de leurs probabilités et les résultats obtenus sont assez bons pour que la qualité générale soit considérée comme satisfaisante.

L'utilisation de textes d'apprentissage autorise la mise en place d'un système grammatical spécifique d'un corpus, d'une personne, d'un niveau de langue. Inversement, un apprentissage réalisé sur un type de corpus donné n'est pas toujours suffisant pour aborder toute situation dans de nouveaux textes. En particulier, les corpus des enfants présentent de nombreux mots isolés, ce qui n'était pas le cas des textes écrits dont nous disposions comme base morphosyntaxique initiale. Ces mots isolés se trouvent également dans des corpus adultes de dialogue et le problème de détermination de la classe lexicale des mots isolés est le même pour les enfants et pour les adultes. On peut facilement comprendre qu'il est impossible de générer des règles contextuelles sophistiquées pour des énoncés ne comprenant qu'un seul mot. Le seul contexte étant ponctuation à gauche et à droite, aucune règle ne peut venir résoudre les ambiguïtés. Nous avons décidé de nous conformer dans ces cas à l'analyse que ferait un adulte. Les cas d'ambiguïté entre verbe et substantif ont été résolus à la main au coup par coup en fonction du contexte, les ambiguïtés entre substantif et adjectif ont été résolues comme substantifs, celles entre substantif et interjection comme interjection. Dans les cas où il pourrait y avoir ambiguïté entre un mot fonctionnel et un mot à forte valeur sémantique, l'étiquette sera celle du mot plein. Un exemple est celui de 'un'. En français, 'un' représente le chiffre 1 ainsi que l'article indéfini. Nous avons considéré que, en tant que mot isolé, il s'agissait du chiffre. En fait, en regardant manuellement toutes les circonstances où 'un' est isolé, on s'aperçoit que dans un cas, il s'agit bien de l'article que l'adulte utilise pour suggérer un mot à l'enfant. Cette circonstance exceptionnelle est exemplaire à plusieurs titres : elle montre que l'analyse automatique ne peut remplacer complètement l'étude manuelle et qu'il faut revenir à cette dernière pour étudier certaines situations syntaxiques précises très localisées ; elle montre qu'il existe toujours des énoncés « agrammatiques » que seules justifient les circonstances pragmatiques du discours et qu'aucun programme automatique ne saura traiter dans un avenir proche.

Catégories lexicales

Le choix des classes lexicales traitées est celui de classes très générales (voir tableau 1 – sont exclus de cette table les ponctuations de fin d'énoncés). Aucun travail d'étiquetage n'a été fait concernant le genre et le nombre car ceux-ci ne sont pas très

difficiles à traiter sur les corpus des enfants de 2 à 3 ans et leur étude ne nécessite pas d'emblée d'analyse syntaxique sophistiquée. Bien que très général, ce choix des classes reflète l'analyse distributionnelle qui peut être faite du français. Ainsi, la différence entre les différents pronoms, génériques, démonstratifs ou relatifs reflète les différences des contextes dans lesquels ils s'insèrent. Si l'on voulait utiliser un nombre de catégories plus important, il faudrait alors veiller à ce que les nouvelles catégories créées puissent se différencier par leur contexte.

***Tableau 1* : liste des 25 catégories morphosyntaxiques utilisées pour l'enfant de 2 à 4 ans**

Code de la classe	Nombre d'occurrences à 2 ans	Nombre d'occurrences de 2 à 4 ans	Description de la classe
A	87	2753	Verbe Avoir
ADJ	107	2475	Adjectif
ADV	159	4023	Adverbe
ADV-l	273	3585	Adverbe de lieu
ADV-n	130	3255	Adverbe de négation
ART	181	8977	Article
ART-g	9	1171	Article généralisé
COJ	31	1915	Conjonction
E	246	4745	Verbe être
I	241	2207	Interjection
I-e	253	1590	Interjection exclamative
NB	6	503	Nombre
NP	21	509	Nom propre
PP	198	2323	Participe passé
PREP	15	3657	Préposition
PREP-a	42	1599	Préposition article
PRN	303	14980	Pronom
PRN-d	122	2411	Pronom démonstratif
PRN-r	72	2148	Pronom relatif ou interrogatif
S	847	13909	Substantif
V	169	8999	Verbe
V-inf	115	4558	Infinitif
V-ppre	—	25	Participe présent
VOICI	100	838	Locution voici, voilà
Y	38	1009	Pronoms Y, EN
Nombre total d'occurrences	3765	94164	

Matériel d'étude

Une évaluation systématique de la formation des catégories lexicales chez le jeune enfant a été réalisée à l'aide d'une base de donnée créée en utilisant la technique d'observation directe des comportements*. Il s'agit de recueil de la parole spontanée effectué au cours d'un jeu symbolique, dans la même situation standard, par enregistrement sur magnétoscope, au vu et au su de l'enfant et réalisé toujours par le même observateur. Les enregistrements ont été pris dans cette situation de jeu afin de permettre à l'enfant de commenter ses actes, de raconter des événements vécus ou imaginaires et de dialoguer avec un partenaire adulte. Le matériel strictement standardisé est constitué de la « Maison de famille » Fisher-Price qui comprend cinq personnages (2 figurines adultes, 2 figurines enfants, 1 bébé), un chien, onze éléments appartenant au mobilier (2 tables, 4 chaises, 2 fauteuils et 3 lits) et cinq éléments figuratifs (escalier avec porte mobile, garage avec porte coulissante, sonnette de la porte d'entrée).

* *Le Normand, 1986*

Pour le recueil des données, la technique de l'échantillonnage complet des comportements a été utilisée et le discours des enfants a été segmenté en énoncés selon les critères définis par Rondal et coll.*, permettant une transcription standard puis le calcul de paramètres linguistiques considérés qui sont décrits au terme de l'analyse des corpus par programme sur micro-ordinateur, système CLAN (Child Language Analysis, version 2.01, MacWhinney et Snow, 1985, 1995). Les corpus présentés ici couvrent l'âge de 2 ans 0 mois à quatre ans 0 mois. Les caractéristiques des corpus sont données dans le tableau 2 ci-dessous.

* *1985*

Tableau 2 : **liste des caractéristiques des corpus par âge**

Age Ans Mois	Nb. enfants	LME* moyenne	LME* min.	LME* max.	Nb. énoncés	Moy. nb. énoncés	Nb. mini. énoncés	Nb. maxi énoncés
2.0	27	1.63	1.10	2.88	2157	79	27	187
2.3	24	2.04	1.15	3.71	2156	89	46	161
2.6	30	2.62	1.28	3.79	3149	104	41	283
2.9	24	3.33	1.67	4.74	3300	137	41	567
3.0	19	3.72	1.67	4.98	2085	109	52	220
3.3	23	3.82	2.68	4.66	3450	150	48	305
3.6	23	4.11	1.88	6.88	2884	125	50	260
3.9	20	4.42	3.49	5.47	2192	109	34	217
4.0	28	5.39	2.60	10.55	4024	143	25	603

**LME = Longueur Moyenne de l'Enoncé*

RÉSULTATS

Le premier corpus étiqueté a été celui des enfants de deux ans. L'ensemble des mots triés par ordre alphabétique a d'abord été étiqueté à l'aide d'un dictionnaire informatisé classique puis tous les mots inconnus ont été traités manuellement (en général, il s'agissait d'interjections ou de formes enfantines déformées). Un premier apprentissage syntaxique a été réalisé sur un corpus dont nous disposions déjà, issu des travaux précédents*. Enfin, le corpus des enfants de deux ans a été étiqueté automatiquement puis corrigé manuellement. Lorsque le nombre de situations syntaxiques nouvelles a atteint une valeur assez grande (notamment à cause des mots isolés ou des petites phrases exclamatives ou incomplètes spécifiques des débuts du langage), il a été procédé à un complément d'apprentissage sur la partie corrigée et à une nouvelle analyse automatique. La procédure suivie a été la même pour chaque tranche d'âge des enfants si ce n'est que l'utilisation de l'analyseur tenait compte des résultats obtenus aux âges inférieurs.

* *Parisse, 1989*

Même sur des corpus aussi simples (en nombre de termes lexicaux, pour le moins), il y a beaucoup de cas d'ambiguïtés à résoudre. Le tableau 3 présente ce nombre d'ambiguïtés, d'une part de manière relative au vocabulaire des enfants de l'âge concerné (c'est-à-dire qu'un mot n'est ambigu à un âge donné que s'il apparaît avec plusieurs classes différentes à cet âge), d'autre part de manière relative à l'ensemble de la langue française (c'est à dire comme un adulte ou un linguiste considérerait le mot AVANT de connaître les caractéristiques lexicales d'une tranche d'âge donnée).

Tableau 3 **: Nombre de mots ambigus par rapport au nombre total de mots en fonction de l'âge, soit de manière relative, soit de manière absolue**

Age Ans Mois	Nb. mots	Ambiguïté relative aux enfants du même âge		Ambiguïté absolue (relative à l'adulte)	
		Nb. ambigus	Nb. ambigus par mot	Nb. ambigus	Nb. ambigus par mot
2.0	5950	534	1.09	2441	1.81
2.3	6701	725	1.11	3145	1.96
2.6	11649	1572	1.14	5897	2.02
2.9	14844	2417	1.16	7980	2.10
3.0	10011	1365	1.14	5335	2.06
3.3	16690	2310	1.14	8861	2.08
3.6	15122	2410	1.16	8049	2.08
3.9	12023	2184	1.19	6492	2.07
4.0	27552	5515	1.21	14951	2.14
Adulte	156201	63923	1.68	88968	2.28

L'analyse détaillée de tous les cas d'ambiguïté à deux ans 0 mois est encore assez facile à réaliser car le nombre de mots ambigus n'est pas trop grand et une vérification très fine de l'analyse syntaxique peut être faite. La liste détaillée des ambiguïtés rencontrées à 2 ans est fournie dans le tableau 4. On trouve dans cette liste trois types

de situations : les ambiguïtés classiques (déjà fonctionnelles comme « l' », balbutiantes comme « la »), les ambiguïtés non classiques relatives au principe de la morphosyntaxe (« c'est qui » vs. « qui c'est » où « qui » n'a pas la même fonction, « au dodo » et « dodo ! » où les structures distributionnelles sont différentes), et les « erreurs » ou cas litigieux (notés entre guillemets dans le tableau 4). Ces derniers cas sont délicats dans la mesure où il peut s'agir de phrases incomplètes ou incorrectes. Il sont toutefois très peu nombreux et la « prise de décision » de l'analyseur permet de pointer du doigt ce type de problème.

Tableau 4 : **exemples d'ambiguïtés à deux ans 0 mois**

mot	classes ambiguës	nb. occur.	exemples cas 1	cas 2	
autre	S, ADJ	17, 3	autre chaise	l'autre	
"balance"	S, V	3, 1	"balance le cheval"	"la balance"	?
boum	S, I, ADJ	2, 26, 2	oh boum	un autre boum c'est boum	
bébé	S, NP	24, 80	le bébé	oh bébé !	
dodo	S, I	25, 47	au dodo	dodo !	
faitV, PP	7, 4	fait dodo	c'est fait		
l'	PRN, ART	55, 18	où l'est	l'école	
la	PRN, ART	1, 54	ouvrir la porte	c'est la dame	
le	PRN, ART	4, 65	y a le chien	le voilà	
maman	S, NP	2, 35	la maman	maman !	
"petit"	S, ADJ	2, 15	"tout petit"	"les petits enfants"	
"place"	S, V	2, 1	"le chien place"	"place"	
quiPRN, PRN-r	3, 3	c'est qui	qui c'est		
tout	PRN, ART-g	3, 6	c'est tout	tout ça	
un PRN, ART	3, 21	un lit	encore un		

CONCLUSION

L'utilisation d'un analyseur morphosyntaxique pour traiter le langage des enfants est tout à fait fonctionnelle. La relative aisance avec laquelle nous avons mis en place cette analyse suggère d'ailleurs des discussions sur les fondements théoriques de l'acquisition de la morphosyntaxe chez l'enfant qui ne peuvent pas être rapidement traitées ici car cela dépasse le cadre de cette présentation et exige des études complémentaires. Soulignons simplement que nos premières analyses, effectuées à partir d'un apprentissage initial réalisé sur des textes adultes de langue écrite ont très bien fonctionné, comme si la syntaxe de l'enfant ne présentait pas de différences majeures avec celle de l'adulte. Seuls des cas particuliers comme ceux mis entre guillemets dans le tableau 4, le traitement des abondantes exclamations et celui des énoncés réduits à un mot ont imposé mises au point et apprentissages complémentaires. Certes, l'analyseur que nous utilisons est particulièrement adapté au traitement de fragments de phrases ou énoncés paraissant incomplets à l'adulte, mais cela signifie aussi que le langage de l'enfant est souvent constitué de parties correctes du langage adulte. Egalement, l'analyseur permet aisément de pointer un à un les énoncés réellement « agrammatiques » – et d'évaluer cet agrammatisme, c'est à dire de voir quels éléments sont manquants dans le discours de l'enfant.

Le nombre de mots ambigus progresse doucement avec l'âge. Cette progression est à mettre à l'actif, aussi bien de la progression du lexique de l'enfant que de la taille des corpus. Mais pour le linguiste ou le psychologue, même l'étude du corpus d'un enfant de deux ans nécessite un travail attentif puisqu'il a déjà un potentiel de 40 % de mots ambigus. Ainsi, par exemple, même si « porte » n'est pas encore ambigu dans notre corpus, le mot doit être traité comme si il l'était et soulève d'ailleurs deux cas délicats : « ça porte » et « porte là ». Ces deux cas, mis en évidence par l'analyse syntaxique qui propose un verbe, vont pouvoir être contrôlés en détail avec, si possible, un retour aux enregistrements originaux. L'utilisation de l'étiquetage automatique ne supprime donc pas toute intervention du scientifique, mais elle

accélère le travail de manière non négligeable. Ainsi, en reprenant l'exemple présenté plus haut de 95000 mots à traiter, le temps nécessaire peut être réduit, la durée dépendant du type de corpus (variable ou régulier) et des besoins de l'utilisateur (statistiques générales ou analyse de détail).

Le programme d'étiquetage présenté ici n'est pas une application industrielle. Il s'agit d'un programme directement issu de la recherche et destiné à des utilisateurs avertis. Toutefois, son insertion dans une procédure générale ouverte au profane en informatique ne pose pas de problème de fond. Il pourrait même probablement venir compléter des analyseurs d'échantillons de langage préexistants comme CLEAR* .

Pour démontrer son utilité, il a déjà été intégré sous le nom de POST - part-of-speech tagger - dans le système CHILDES comme un des programmes CLAN - cette option est disponible sous PC compatible seulement. Il permet de créer un champ "%ms:" qui contient l'énoncé original avec les classes syntaxiques de tous les mots de cet énoncé. Il devient alors beaucoup plus aisé de faire une évaluation précise de la structure lexicale et syntaxique des échantillons traités. L'utilisation d'outils d'analyse morphosyntaxique offre déjà une accélération du traitement de bases de données importantes, ce qui est le but de nombreuses applications linguistiques disponibles sur micro-ordinateur. De plus, elle permet d'offrir au linguiste, à l'orthophoniste ou au praticien une valeur ajoutée qui n'est pas purement quantitative mais aussi - ce qui n'est pas le cas des programmes actuels - qualitative et qui suggère des développements d'ordre fondamental grâce à son adéquation avec le langage de l'enfant.

* *Baker-Van den Goorbergh et Baker, 1991*

BIBLIOGRAPHIE

— ANDREEWSKY A. (1973). *Apprentissage, analyse automatique du langage, application à la documentation.* Collection : Documents de linguistique quantitative, 21, Paris : Dunod.

— ANDREEWSKY A., DEBILI F., DEHLI M., FLUHR C. (1980). Apprentissage, syntaxe, sémantique lexicale. *Revue du palais de la découverte*, 83, 17-40.

— BAKER-VAN DEN GOORBERGH L. (1994). Computers and language analysis : theory and practice. *Child Language Teaching and Therapy*, 10, 3, 329-348.

— BAKER-VAN DEN GOORBERGH L., BAKER K. (1991). *1991 : Computerised Language Error Analysis Report (CLEAR).* Kibworth, Leics : FAR Communications.

— CAPPELLI G., MACCARI A., PFANNER L. (1991). A system for semi-automatical treatment of child morphology. In *Actes de la 4th Annual Sentence Processing Conference (CUNY)*, Rochester.

— CRYSTAL D., FLETCHER P., GARMAN M. (1976). *The grammatical analysis of language disability.* London : Edouard Arnold.

— LE NORMAND M. T. (1986). A developmental exploration of language used to accompany symbolic play in young, normal children (2-4 years old), *Child, Care, Health and Development*, 12, 121-134.

— MACWHINNEY B., SNOW C.E. (1985). The Child Language Data exchange system. *Journal of Child Language*, 12, 271-296.

— MACWHINNEY B. (1995). *The CHILDES project : Tools for analyzing talk (2nd edition)*, Hillsdale : Lawrence Erlbaum Associates.

— MILLER J. F., CHAPMAN R. S. (1983). Using microcomputers to advance research in language disorders. *Theory Into Practice*, XXII, 4, 301-307.

— PARISSE C. (1989). *Reconnaissance de l'écriture manuscrite : analyse de la forme globale des mots et utilisation de la morpho-syntaxe.* Thèse de doctorat, Université de Paris-Sud, Orsay, France.

— RONDAL J.A., BACHELET J.F., PÉRÉE F. (1985). Analyse du langage et des interactions verbales adulte-enfant. *Bulletin d'Audiophonologie*, 5/6, 507-536.

— RONDAL J.A., PÉRÉE F., BACHELET J. F., THOORENS J. (1987). *ALGELO - Analyse lexico-grammaticale d'échantillons de langage par micro-ordinateur.* Nancy : Unadrio.