HAL open science

Classification des catégories grammaticales sur deux corpus longitudinaux d'enfants

Andrea Briglia, J. Sauvage, Giovanni Pirrotta, Massimo Mucciardi

To cite this version:

Andrea Briglia, J. Sauvage, Giovanni Pirrotta, Massimo Mucciardi. Classification des catégories grammaticales sur deux corpus longitudinaux d'enfants. 2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT), 2020, Montrouge, France. pp.28-38. hal-03047149

HAL Id: hal-03047149

https://hal.science/hal-03047149v1

Submitted on 3 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Classification des catégories grammaticales sur deux corpus longitudinaux d'enfants

Andrea Briglia[1, 2] Jérémi Savage[1] Giovanni Pirrotta[2] Massimo Mucciardi[2]
(1) Université Paul Valéry, Montpellier, France
(2) Université de Messine, Messine, Italie
prenom.nom@univ-montp3.fr, prenom.nom@unime.it

RÉSUMÉ

Cet article analyse deux suivis longitudinaux de deux enfants du projet CoLaJE: une annotation automatique des parties du discours a été appliquée à chaque énoncé (15'000 en total) en adoptant le standard des « Universal Dependencies » comme référence et « stanza », un librairie Python, comme outil d'analyse. L'âge et le taux d'erreur ont servi comme base pour la création de neuf strata: réduire la dimension du corpus nous permet de rendre intérpretables les groupements crées avec une méthode non-supervisée, EM clustering. Regrouper en clusters les énoncés des enfants annotés en parties du discours aide à mieux cibler le développement des catégories grammaticales au cours du temps: deux exemples concernants le développement de la cohérence morphosintaxique sont proposés, ainsi que deux exemples concernants l'évolution de la relation entre l'usage de pronoms et des noms. Une discussion nale des résultats et des limites de cette recherche est ensuite proposée.

ABSTRACT

Classi cation of grammatical categories in two longitudinal chlidren corpora. This article analyses two child spoken language longitudinal corpora from the CoLaJE project: a parts of speech automatic annotation was applied to each sentence (15'000 in total) using « Universal Dependencies » as a standard of reference and "stanza", a Python library, as an analysis tool. Age and error rate were used as criteria for the creation of nine strata: reducing the size of the corpus helps to make more easily interpretable clusters created with EM, an unsupervised method. Aim of the article is to propose a way to target the development of grammatical categories over time: two examples concerning the development of morphosintactic coherence are proposed, as well as two examples concerning the evolution of the relationship between the use of pronouns and nouns. A final discussion of the preliminary results and limitations of this research is then proposed.

MOTS-CLÉS : corpus d'enfants ; acquisition du français L1 ; développement syntaxique ; clusterisation EM

 ©2020 CNRS Attribution 4.0 International.

KEYWORDS: child spoken language corpora ; first language acquisition ; syntax development ; EM clustering

1 Objectifs et hypothèse de recherche

Le projet ANR « CoLaJE » (Morgenstern & Parisse, 2012) consiste en sept corpora d'enfants francophones filmés une heure par mois, tous les mois, dès l'âge d'un an jusqu'à environ 5 ans. L'ensemble de données est disponible en libre accès et fait partie de la branche française de CHILDES[1] . Nous avons choisi cette base de données parce que – à ce jour – elle est la plus complète sur les plans qualitatif et quantitatif. Par ailleurs, nous estimons que l'échantillonnage effectué dans la collecte mensuelle des données est conforme aux indications de fiabilité énoncées par Tomasello et Stahl (2004). Chaque corpus a été codé en CHAT et transcrit en pho (ce que l'enfant prononce) et – pour certains corpora dont les deux qu'on utilise en cette étude – mod (ce que l'enfant aurait dû prononcer selon la norme de prononciation standard), ce qui nous permet d'uniformiser les données phonético/phonologiques, de les contextualiser pour mieux les interpréter et, enfin, de pouvoir y appliquer des traitements automatiques.

Dans la présente contribution, nous nous focalisons sur les corpora d'« Adrien » et « Madeleine » car ils sont les plus complets : nous avons extrait chaque ligne en format .csv, ensuite nous avons choisi de commencer par la transcription no 8 (1an 11mois; 14jours) pour Adrien et la no 3[2] (1 ;01 ;10) pour Madeleine, puisque pour les précédentes il était difficile de distinguer entre les mots et les simples suites de syllabes correspondantes à l'étape du « babillage canonique » et du « babillage diversifié » dans le développement de la production de la parole du jeune enfant (Sauvage, 2015). Nous avons au total 26 enregistrements et 8214 énoncés pour Adrien et 25 enregistrements et 7168 énoncés pour Madeleine. Nous avons choisi le « Universal Dependencies » (de Marneffe et al., 2006, 2008, 2014) comme modèle de référence d'analyse du langage en parties du discours, principalement parce que nous avions déjà eu recours à ce modèle (Briglia et al., 2020). Ce choix nous a conduit à adopter "stanza", un outil d'analyse du langage majoritairement entraîné en utilisant les UD. « stanza » est une des bibliothèques de TAL disponible en langage Python, développée par l'Université de Stanford: puisque le système d'annotation automatique ne reconnaît pas les caractères spéciaux de l'API (Alphabet Phonétique International), nous l'avons appliqué sur le tiers CHI (transcription orthographique): ce choix implique une forte confiance envers l'interprétation des transcripteurs : il est néanmoins possible de consulter – énoncé par énoncé – toutes les différences entre CHI – pho –mod. La qualité de l'annotation produite par « stanza » est élevée et, pour la plupart des tâches, son score est meilleur que celui de ses concurrents (e.g UDPipe, spaCy), comme le montre le tableau numéro 2 « Neural pipeline performance comparisons on the Universal Dependencies (v2.5) test treebanks" (Qi et al., 2020).

Puisque le langage de l'enfant se caractérise par une forte variabilité et reste imprévisible à court et moyen termes et puisque UD et « stanza » ont été conçus pour le langage des adultes, il nous a

[1] https://childes.talkbank.org/access/French/

[2] https://ct3.ortolang.fr/data/colaje/madeleine/; https://ct3.ortolang.fr/data/colaje/adrien/

semblé nécessaire d'opérer un contrôle manuel de quatre-vingts énoncés pour chaque enfant (ce qui représente environ 1% du total) équitablement répartis au fil du temps, afin de comprendre l'effective fiabilité de l'outil pour cette application. Nous avons remarqué que certaines répétitions du même mot - typiques lorsque l'enfant cherche de cibler l'apprentissage d'un mot donné – étaient parfois codés comme NOUN – ADJ, alors qu'il s'agissait soit de deux noms communs, soit de deux adjectives (par exemple, l'énoncé « des grands grands arbres » était codé comme « DET-ADJ-NOUN-NOUN »), « ouais » était codé comme SYM et, de façon plus générale, les nombreuses exclamations des enfants (par exemple « ah ! », « oh ») ou des surnoms comme « papi » ou « mémé » étaient codés différemment selon leur contexte. Mais ces exceptions sont tout à fait faciles à comprendre dans leur contexte et, en tout cas, ne représentent qu'un faible pourcentage du total des productions. En fait, pour effectuer une analyse syntaxique pertinente sur un corpus longitudinal d'enfant il est indispensable de comprendre avant tout si ce que l'enfant dit – que ce soit au niveau phonético/phonologique ou syntaxique - est conforme ou non aux normes linguistiques des adultes ou pas. C'est pourquoi nous nous sommes appuyés sur une précédente étude où le SPVR (Sentence Phonetic Variation Rate) avait été calculé (Briglia A. et al., 2020). Le taux obtenu sera le résultat d'une comparaison entre les tiers « pho » et « mod » : pour ce faire, nous avons mis en place un algorithme indiquant si le premier est équivalent au deuxième ou non, en donnant comme réponse 0 ou 1 et la distance de Levenshtein relative.

1.1 Corpus recueilli et méthodologie d'analyse

Notre but est de fournir un outil d'évaluation du développement de la syntaxe basé sur des associations et des distributions. Afin de savoir comment les catégories syntaxiques évoluent pendant le temps et le taux d'erreur de ces dernières, nous divisons l'ensemble des données en 9 strata (LL, LM, LH, ML, MM, MH, HL, HM, HH) selon trois classes temporelles successives (représentées par la première lettre) et trois classes d'erreur (représentée par la deuxième lettre). Par exemple, LL veut dire que le strata représente la première tranche d'âge et le taux d'erreur le plus bas (c'est-à-dire Low < 33.3%).

code	STRATA	TIME (age)	SPVR
1	LL	1.01 - 2.09	<=33%
2	LM	1.01 - 2.09	>33% and <=66%
3	LH	1.01 - 2.09	>66%
4	ML	2.10 - 2.61	<=33%
5	MM	2.10 - 2.61	>33% and <=66%
6	MH	2.10 - 2.61	>66%
7	HL	2.70 - 3.53	<=33%
8	HM	2.70 - 3.53	>33% and <=66%
9	HH	2.70 - 3.53	>66%

Figure 1: Madeleine

code	STRATA	TIME (age)	SPVR
1	LL	1.97 - 2.64	<=33%
2	LM	1.97 - 2.64	>33% and <=66%
3	LH	1.97 - 2.64	>66%
4	ML	2.71 - 3.39	<=33%
5	MM	2.71 - 3.39	>33% and <=66%
6	MH	2.71 - 3.39	>66%
7	HL	3.46 - 4.33	<=33%
8	HM	3.46 - 4.33	>33% and <=66%
9	HH	3.46 - 4.33	>66%

Figure 2: Adrien

Ainsi, nous voyons que la 1 ère tranche d'âge représente le parcours de Madeleine à partir de l'âge d'un an jusqu'à deux ans et 1 mois. Cette tranche compte 12 enregistrements, pour un total de 1956 énoncés analysés. La 2 ème tranche d'âge est constituée de 6 enregistrements (c'est-à-dire six mois consécutifs) et 2765 énoncés. Enfin, la 3 ème et dernière tranche d'âge est composée de 7 enregistrements et 2447 énoncés. La première tranche compte deux fois plus d'enregistrements que les suivantes parce que l'enfant à cet âge parle très peu et la quasi-totalité de ce qu'il dit sont des mots isolés (il s'agit du stade « holophrastique » de son développement syntaxique, où un mot comme « eau » peut vouloir signifier un énoncé entier comme « je veux boire »). Les six premiers enregistrements ne comptent que 66 énoncés, alors que le 7 ème enregistrement (1 an et six mois) en compte 187. En effet, à partir de 18 mois et jusqu'à 30 mois on assiste à une période d'explosion lexicale où l'enfant va apprendre plusieurs mots par jour. En particulier, de 18 mois à 24 mois l'enfant développe un lexique de plus en plus riche mais qui n'arrive pas encore à combiner au niveau syntaxique (on parle de langage « télégraphique », d'une syntaxe construite autour d'un mot pivot, par exemple dans les énoncés à deux mots), alors qu'à partir de 24 mois jusqu'à 36 mois on assiste au développement de la grammaire, avec une phase dite d'« explosion grammaticale » à partir de 30 mois environ (Sekali M., 2012).

Pour la première tranche d'âge d'Adrien, on compte 1709 énoncés repartis en 9 enregistrements, pour la 2 ème tranche d'âge on compte 2623 énoncés en 9 enregistrements et la 3 ème tranche compte 3882 énoncés en 8 enregistrements. Nous obtenons donc un total de 25 enregistrements et 7168 énoncés pour Madeleine, et 27 enregistrements et 8214 énoncés pour Adrien. C'est en cherchant un équilibre entre les étapes du développement et les enregistrements disponibles que nous avons choisi de diviser en trois parties le nombre total d'enregistrement des deux enfants pour mieux organiser l'analyse. On peut remarquer que les enregistrements d'Adrien commencent à presque deux ans et que les transcriptions en tier « mod » se terminent pour Madeleine à l'âge de 3;5 ans alors que pour Adrien elles continuent jusqu'à l'âge de 4;4 ans : nous avons ainsi choisi de décaler le début d'analyse pour Adrien, ce qui nous a permis d'obtenir des tranches temporelles plus proches et, en conséquence, plus comparables.

Il reste une différence remarquable entre le développement langagier des deux enfants : celui d'Adrien pourrait être considéré comme normé, alors que celui de Madeleine est sans doute plus

rapide que la moyenne (Morgenstern & Parisse, 2012). De façon plus générale, il est couramment accepté que les filles parlent souvent mieux que les garçons à un même âge, que ce soit au niveau qualité ou quantité de parole. Mais ces derniers rattrapent cet écart entre 4 et 6 ans. Pour résumer, nous montrerons de façon simplifiée l'évolution du nombre des mots par énoncé et l'évolution du nombre de mots différents par énoncé ci-dessous (en bleu Madeleine et en rouge Adrien). Ensuite, nous discuterons si l'analyse en clusters conduite avec EM peut être mise en relation avec ces graphes. Il faut se rappeler que dans ces graphes les trois points temporels sur les abscisses représentent un décalage de presque un an: malgré elle soit plus jeune, Madeleine montre un meilleur développement, donc si les deux lignes auraient été temporellement alignées, l'écart aurait été majeur (ce qui est conforme aux résultats de l'étude principale sur ce corpus par Morgenstern & Parisse, 2012).

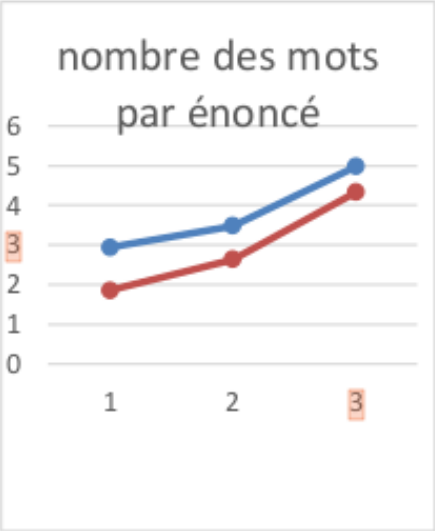

Figure 3

Figure 4

1.2 Traitement automatique

Nous calculons le F (Fisher) et le p-value relatif pour chaque partie du discours (POS tag) en supposant que la distribution des occurrences discrètes sous-jacente est de type Poisson (nous avions d'abord essayé avec une distribution de type Gauss mais sans résultat convaincant). Ensuite le nombre d'énoncés est divisé en plusieurs groupes grâce à la méthode non-supervisée EM (Expectation Maximization). Nous avons choisi cette méthode parmi d'autres parce qu'elle s'adapte bien, selon nous, à la quantité et à la typologie de nos données. Son fonctionnement est le suivant : le nombre optimal de groupes est obtenu en suivant les deux étapes d'un algorithme itératif qui termine son calcul lorsque l'ajout ultérieur d'un groupe aboutit à une amélioration négligeable dans la fonction de vraisemblance logarithmique (le seuil a été fixé à 5% de la vraisemblance à n-1). En sélectionnant les parties du discours plus significatives pour les regroupements, EM est influencé en

amont par une distinction fondamentale entre les classes ouvertes (ADJ, ADV, INTJ, NOUN, PROPN, VERB) et les classes fermées (ADP, AUX, CCONJ, DET, NUM, PART, PRON, SCONJ) : les premières contiennent beaucoup plus d'éléments que les deuxièmes, pour lesquelles le nombre est fixé. Pourtant, la fréquence entre les deux classes ne varie pas autant, ce qui est confirmé dans le développement de la parole (voir les valeurs des variables mean associées à chauqe POS tag dans les tableaux ci-dessous). Toute étude concernant la distribution des fréquences des mots et leurs interactions devrait en fait garder à l'esprit l'importance de l'équilibre entre syntaxe et sémantique comme principe organisateur du langage (Zipf, 1949 ; Lestrade, 2017).

1.3 Résultats préliminaires

La cohérence morphosyntaxique est plus élevée dans les clusters en HL, HM par rapport à ceux dans les couches L et M, ce qui est conforme aux résultats d'une étude précédente (Parisse et al., 2010). On peut remarquer que les parties du discours PRON, VERB, SCONJ – qui pourraient être considérées comme marqueurs des phrases plus longues – augmentent leur importance (voir la variable mean) au fil du temps, il est également à noter que les groupes qui reconnaissent des énoncés syntaxiquement proches font aussi partie des classes d'erreurs et d'âge différentes, par exemple, dan le cas d'Adrien:

2452 escargot tout chaud

ɛskaʁɡo tu ʃo

didago to so

En MH

6746 une souris verte

yn suʁi vɛʁtə

yn ʦoji vatə

En HH On peut ensuite noter comme le NOUN et le PROPN – qui indiquent une personne ou une chose concrète et sont souvent utilisés dans le même contexte de la parole adulte (répétition du langage adressé à l'enfant) – sont au fur et à mesure remplacés par PRON (une catégorie grammaticale plus abstraite pour indiquer choses et personnes) en termes du nombre relatif d'occurrence (réf aux tableaux 1 et 2). L'âge à laquelle ce changement devient visible est environ 3 ans et correspond à l' âge individuée dans des études précédentes sur le développement de la capacité de produire des énoncés en forme transitive à partir des énoncés en forme intransitive (Childers &Tomasello, 2001). Par exemple, dans le cas d'Adrien,

1973 sait pas faire maman	sɛ pa fɛʁ mamɑ̃	te ta pa mamɑ̃
2915 est caché papa	ɛ kaʃe papa	e kaɕe papa
3674 il peut pas ouvrir la porte	i pø pa uvʁiʁ la pɔʁt	i pu pa uvij a pɔt

En MH , on peut remarquer le manque du pronom « je », alors que à l'âge 3_09_09 dans cet énoncé les pronoms sont bien utilisés

5339 moi j' aime plutôt celle-ci maman elle aime plutôt celle-ci

mwa ʒɛm plyto sɛlsi mamɑ̃ ɛl ɛm plyto sɛlsi

mwa zem pyto sɛsi mamɑ̃ ɛl øm pytosɛti

En observant la valeur de la variable « mean », on peut remarquer comme la valeur de PRON augmente au fil du temps en dépit de la valeur de NOUN : chez Madeleine ce développement est plus rapide que chez Adrien.

Madeleine présente un développement plus ordonné puisque dans les mêmes trois strata considérés pour Adrien, EM arrive dans ce cas à identifier la structure syntaxique la plus basique : sujet (ou pronom) – verbe. Les énoncés de Madeleine présentent constamment plus VERB and AUX que celles d'Adrien, sauf pour le troisième strata : ce qui colle avec les graphes montrés en Figure 3 et 4, où nous pouvons remarquer que les énoncés de la petite fille sont à la fois plus longs et lexicalement plus riches.

Par exemple, en HL, énoncé 186, cluster 2

euh ça je sais pas comment l'ouvrir

ø sa ʒə sɛ pa komɑ̃ luvʁiʁ

œ sa ʒə sɛ pa komɑ̃ luvʁiʁ

On peut remarquer comment le dernier strata HL de Adrien ressemble à celui de Madeleine : on en pourrait déduire que lorsque les quatre premières parties du discours classées comme plus significatives par EM contiennent NOUN, PRON, VERB et AUX alors l'enfant devrait présenter des énoncés syntaxiquement conformes à la norme adulte.

STRATA LL			STRATA ML			STRATA HL		
POS_tags	Mean	# sentences	POS_tags	Mean	# sentences	POS_tags	Mean	# sentences
INTJ	0,126	611	CCONJ	0,048	851	PRON	1,157	1762
DET	0,095	611	PRON	0,133	851	DET	0,321	1762
ADP	0,013	611	NOUN	0,220	851	VERB	0,788	1762
NOUN	0,468	611	AUX	0,052	851	NOUN	0,419	1762
SYM	0,023	611	VERB	0,157	851	SCONJ	0,149	1762
ADV	0,563	611	NUM	0,035	851	ADP	0,230	1762
PROPN	0,016	611	SYM	0,020	851	AUX	0,208	1762
PRON	0,025	611	ADV	0,828	851	ADV	0,727	1762
VERB	0,023	611	DET	0,086	851	ADJ	0,091	1762
X	0,020	611	PROPN	0,029	851	CCONJ	0,120	1762
CCONJ	0,023	611	ADP	0,034	851	SYM	0,022	1762
SCONJ	0,011	611	X	0,026	851	NUM	0,085	1762
AUX	0,007	611	INTJ	0,176	851	X	0,018	1762
NUM	0,103	611	ADJ	0,012	851	PROPN	0,033	1762
ADJ	0,000	611	SCONJ	0,011	851	INTJ	0,162	1762

Tableau 1: Adrien

STRATA LL			STRATA ML			STRATA HL		
POS_tags	Mean	# sentences	POS_tags	Mean	# sentences	POS_tags	Mean	# sentences
NOUN	0,73	707	X	0,12	1452	NOUN	0,68	1171
VERB	0,38	707	NOUN	0,61	1452	PRON	0,95	1171
PRON	0,27	707	DET	0,45	1452	DET	0,53	1171
DET	0,28	707	VERB	0,70	1452	X	0,14	1171
ADP	0,16	707	PRON	0,81	1452	VERB	0,85	1171
X	0,07	707	AUX	0,20	1452	ADP	0,41	1171
AUX	0,04	707	ADV	0,60	1452	AUX	0,20	1171
ADV	0,34	707	ADP	0,30	1452	SCONJ	0,13	1171
NUM	0,03	707	SCONJ	0,07	1452	CCONJ	0,16	1171
SYM	0,02	707	ADJ	0,10	1452	ADV	0,55	1171
SCONJ	0,01	707	CCONJ	0,10	1452	PROPN	0,05	1171
ADJ	0,04	707	NUM	0,09	1452	NUM	0,12	1171
PROPN	0,04	707	PROPN	0,05	1452	INTJ	0,16	1171
CCONJ	0,02	707	SYM	0,01	1452	SYM	0,02	1171
INTJ	0,09	707	INTJ	0,08	1452	ADJ	0,11	1171

Tableau 2: Madeleine

Pour conclure, le regroupement en clusters pourrait être une façon d'améliorer la compréhension du développement de la syntaxe en proposant une meilleure visualisation de comment une partie du discours évolue dans le temps.

La prochaine étape de ce travail sera d'appliquer cette étude aux autres enfants du projet CoLaJE dans le but de vérifier si les généralisations proposées dans cet article puissent être confirmées pour le développement des autres enfants.

Références

Briglia A., Mucciardi M., Sauvage J. (2020). « Identifying the speech code through statistics: a data-driven approach ».

Childers J., Tomasello M. (2001). « The Role of Pronouns in Young Children's Acquisition of the English Transitive Construction » Developmental Psychology, Vol. 37. No. 6, 739-748.

Dempster A.P., Laird N.M., Rubin D.B. (1977). « Maximum likelihood from incomplete data via the EM algorithm ». Journal of the Royal Statistical Society. Series B: Methodological 39: 1-38.

Lestrade S. (2017). « Unzipping Zipf's law ». PLoS ONE 12(8).

MacWhinney B. (2000). « The CHILDES project: Tools for analyzing talk. 3rd edition ». Mahwah, NJ: Lawrence Erlbaum Associates

Morgenstern A., Parisse C. (2012). « The Paris corpus ». Journal of French language studies /Volume 22/ Special Issue. 7-12. Cambridge University Press (https://www.ortolang.fr/market/corpora/colaje)

Morgenstern A., Sekali M. « What can child language tell us about prepositions? ». Jordan Zlatev, Marlene Johansson

Falck, Carita Lundmark and Mats Andr en. Studies in Language and Cognition, Cambridge Scholars Publishing, 261-275

Parisse C., Le Normand M. T. (2000) « How children build their morphosyntax: The case of French ». Journal of Child Language, Cambridge University Press (CUP), 27, 267-292.

Qi P., Zhang Y., Bolton J., Manning C. D. (2020). « Stanza: A Python Natural Language Processing Toolkit for Many Human Languages ». Association for Computational Linguistics (ACL) System Demonstrations.

Sauvage J. (2015). L'acquisition du langage. Un système complexe. Louvain-la-Neuve : Academia.

Sekali M. (2012). « First language acquisition of French grammar (from 10 months to 4 years old) ». French Language Studies 22, 1-6.

Tomasello M., Stahl, D. (2004). « Sampling children's spontaneous speech: How much is enough? » Journal of Child Language, 31:101–121.