PRAXILING CJC-Praxiling

NumeRev

actes n° 2022 | 2023
LE DISCOURS HORS-NORME(S) OU L'ATYPIE DU DISCOURS

La norme et la variation dans le cadre du Traitement Automatique du Langage

Représenter la diversité des parcours d'acquisition du langage à l'aide de modèles informatiques

Andrea BRIGLIA *Chercheur PostDoc*
GIPSA
CRISSP
Grenoble Alpes University
Massimo MUCCIARDI

Pirrotta GIOVANNI

Édition électronique :
URL :
https://cjc-praxiling.numerev.com/articles/actes-2022/2407-la-norme-et-la-variation-dans-le-cadre-du-traitement-automatique-du-langage
DOI : numerev_1937
Date de publication : 16/10/2023

CertiScience© *Certifié évalué par les pairs*

Pour **citer cette publication** : BRIGLIA, A., MUCCIARDI, M., GIOVANNI, P. (2023) La norme et la variation dans le cadre du Traitement Automatique du Langage. *CJC-Praxiling*, (actes n°2022). https://doi.org/10.34745/numerev_1937

Résumé : Cet article pose la problématique du statut de la norme et de la variation en TAL en proposant des exemples tirés des recherches précédentes concernant des modèles informatiques employés pour représenter l'acquisition de langue française. Deux cas d'étude exemplifient le choix autour de l'axe norme-variation : le calcul automatique d'une distribution de fréquence et la reconnaissance de motifs séquentiels. Que le niveau d'analyse soit le mot (premier exemple) ou le phonème (deuxième exemple), des obstacles et compromis reviennent d'une manière analogue. Le choix – souvent difficile et contraint - entre la précision de la description du langage et la nécessité d'avoir des données uniformes pour que la machine puisse les traiter aisément. Les biais évitables et inévitables, les précautions à prendre en amont, ainsi que les avantages et les inconvénients de ce type de modèles seront discutés. L'article se termine en dessinant les contours des futures complémentarités possibles entre méthodes qualitatives et quantitatives.

Abstract : This article deals with the problem of the status of norm and variation in NLP by proposing examples drawn from previous research concerning computer models used to represent French language acquisition. Two case studies illustrate the choice around the norm-variation axis: the automatic computation of a frequency distribution and the recognition of sequential patterns in words containing specific syllable sequences that are hard to learn due to their inner phonetic difficulty. Whether the level of analysis is the word (first example) or the phoneme (second example), obstacles and trade-offs come up in a similar way. The choice - often difficult and constrained - between the accuracy of the language description and the need to have uniform data for the machine to be easily handled. The avoidable and unavoidable biases, the precautions to be taken beforehand, as well as the advantages and disadvantages of these types of NLP models will be discussed. The article ends by outlining the possible future complementarities between qualitative and quantitative methods in current linguistics.

Keywords : first language acquisition; NLP, French; variation; norm

---

Mots-clés :
Variation, TAL, Acquisition du langage, Français L1, Norme

Dans cet article, le but est de proposer un état de l'art de la variation en sciences du langage dans la perspective du TAL.

Si la définition de ce qu'est la norme pose déjà nombre de problèmes en linguistique (Siouffi & Steuckardt, 2007), dans le domaine du TAL, le défi d'établir un contour précis de norme et – par la suite – de ce qui est à considérer comme variation autour de ladite norme assume des formes différentes qui s'expriment sur d'autres niveaux d'analyse.

Dans cet article il n'est pas sujet de retracer l'histoire des définitions du concept de « norme » en linguistique, toutefois il est pertinent de noter comment les débats autour de la norme (ainsi qu'autour de ses variations) pivotent souvent autour du noyau épistémologique qui suit :

« Devra-t-on décrire la langue à partir de faits linguistiques observables, c'est-à-dire les performances diverses et variées auxquelles on est exposés dans la vie quotidienne ou bien penser la langue à partir de compétences idéalisées ? » (Barge, 2009)

Que l'on veuille rendre compte des diversités dialectales, diachroniques, sociolinguistiques ou pas ; que l'on milite en faveur d'un usage prescriptif et évaluatif de la langue ou bien que l'on accepte tout type de variation linguistique - pourvu qu'elle puisse toujours garantir la transmission du sens ainsi que sa compréhension mutuelle sans défaillance - la richesse de la langue française pose déjà une quantité de « variations normées » non négligeables. Par cette expression l'auteur de cet article voudrait définir tout phénomène linguistique qui – à l'oral comme dans sa forme écrite – ne suit pas la règle, c'est-à-dire ce qui est usuellement prévu pour le même élément dans le même contexte.

Parmi ces « variations normées » à l'oral on trouve entre autres le hiatus, les différentes formes de liaisons, les verbes irréguliers. Alors qu'à l'écrit ces variations se multiplient : l'orthographe du français étant opaque, le nombre d'homographes/homophones ou bien d'homophones non homographes (ou bien encore son inverse) ne sont que la pointe de l'iceberg d'une multitude de « variations normées ».

Mais alors, qu'est-ce qu'est la norme ? Est-ce qu'il s'agit exclusivement d'un usage non conforme qui diffère en fonction du dialecte, du temps, de la classe sociale ou de l'ethnie ? Ou peut-on considérer la variation comme toute déviation d'un ensemble de critères logiques sur lesquels une langue naturelle devrait se baser ?

Effectivement, si l'on adopte la définition suivante de norme « Tout ce qui est d'usage commun et courant dans une communauté linguistique ; la norme correspond alors à l'institution sociale que constitue la langue » (Dubois et al., 1973, p 342), on pourrait répondre à la première question posée dans le paragraphe ci-dessus.

La réponse à la deuxième question est bien plus difficile, et le seul fait de formuler cette question ouvre déjà la voie à plusieurs niveaux d'analyse. Tout d'abord, l'orthographe du français ne suit pas une logique (Hoedt & Piron, 2016) : par exemple, si l'on prend un nouveau mot qui n'existe pas mais qui respecte les règles phonotactiques du français, i.e le mot / kÊefisjÉÌ / (Hoedt & Piron, 2016), comment pourrait-on le transcrire de manière à respecter les normes de l'orthographe du français ?

« Krefision » ou « krefisiont » ? Certes, mais aussi « crephission » ou bien « crefition » ou « chraisfiscion » devraient être considérés comme des candidats conformes.

Toutes ces formes sont possibles selon l'orthographe du français, aucune ne pourrait être jugée comme étant hors-norme ou atypique. Un algorithme programmé pour cette finalité – grâce à un calcul combinatoire qui prend en compte toutes les lettres et/ou syllabes homophones non homographes – a produit comme output le nombre total de transcriptions possible du mot inventé / kÊefisjÉì / : elles sont 240 (Hoedt & Piron, 2016). Il est clair qu'il est difficile parler de norme et de variation quand la norme orthographique ne dérive – au moins dans un bon nombre de cas – que d'une association majoritairement arbitraire reliant un phonème à son/ses graphème(s) correspondant(s).

Les auteurs de cet ouvrage se demandent pourquoi « l'esprit critique s'arrête aux seuils de l'orthographe » (Hoedt & Piron, 2016). Le manque d'univocité dans la relation entre graphème et phonème donne à l'orthographe du français un caractère particulier, qui est commun à d'autres langues (par exemple l'anglais ou l'allemand). Les langues qui ont une orthographe totalement claire sont relativement peu, comme l'espagnol ou le turc par exemple (à noter que l'alphabet latin à été introduit dans le XXème siècle en Turquie, et qu'il a fait l'objet d'une adaptation de haut en bas : l'usage s'est défini une fois que la norme avait été déjà établie par la nouvelle forme étatique).

Après cette petite digression, il faut noter que pour l'ordinateur les variations sont toujours les mêmes puisqu'elles posent constamment le même problème : l'ambigüité (Kraif & Ponton, 2007 ; Jusoh, 2018).

Tout ce qui sort du cadre d'une logique déterminable et prévisible devient difficile pour un ordinateur : calculer le rapport entre "type"/"token" (nombre de mots différents divisé par nombre de mots total) de l'intégralité de l'Encyclopédie de Diderot et d'Alembert est une tâche simple, alors que mettre sur le même niveau ces deux expressions « je ne peux pas », « je peux pas » devient plus compliqué. Le pourquoi - on le sait bien – se trouve dans la déductibilité des règles à appliquer et les exceptions à accorder à ces règles en fonction du contexte : si on a appris à un programme à reconnaître la négation avec cette structure (sujet + ne + verbe + pas), il sera compliqué de lui faire détecter la même entité dans un contexte où un élément manque. Il sera encore plus difficile de le rendre capable de reconnaître que dans certains contextes sociaux la première forme est obligatoire alors que dans d'autres contextes sociaux les deux formes sont acceptables. Une chaine de caractère ne donnant pas d'information sur les locuteurs, il est difficile que la machine puisse mettre en contexte et faire des inférences pragmatiques.

Ces problèmes de multiplicité de transcriptions, d'alignement, de désambiguïsation en fonction du contexte sont présents dans toutes les branches de la linguistique qui utilisent le TAL pour automatiser des tâches répétitives, pour vérifier des hypothèses ou bien pour proposer des représentations des grandes bases de données.

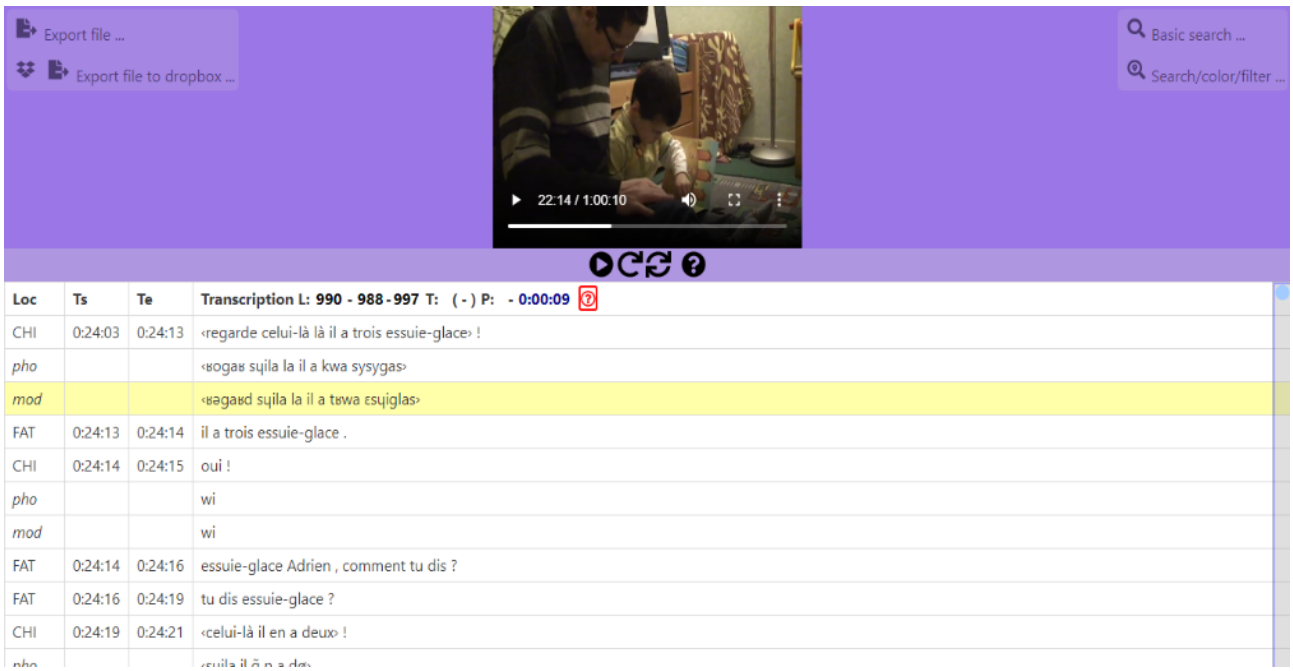
Dans les deux parties de cet article, deux études de cas seront proposées : la première porte sur un calcul de fréquence d'occurrences de mots et montrera comment la variation lexicale de l'enfant a été modélisée pour faciliter l'automatisation d'une tâche. Dans la deuxième étude de cas, plusieurs outils et manipulations seront présentés dans le cadre d'un essai visant à uniformiser le traitement des variations phonétiques/phonologiques chez l'enfant, dans le but ultime de dégager son parcours d'acquisition des phonèmes.

Ces exemples montrent que le TAL est devenu un outil incontournable dans le domaine de la linguistique grâce à sa puissance de calcul et à sa rapidité d'exécution. Cependant, son utilisation peut se révéler insidieuse puisque la nature intrinsèquement ambigüe et polysémique du langage implique un nombre non négligeable de biais et d'exceptions aux règles. Comme il sera détaillé dans les deux parties, le TAL nous amène à des décisions importantes, souvent dans la forme d'un compromis ou d'une balance qu'il faut étalonner soigneusement. Par exemple : est-il mieux de privilégier l'efficience en dépit de la précision, ou bien est-il mieux de choisir de laisser passer un biais dans le codage initial afin d'éviter des problèmes de traitement de catégories par la suite, ou à l'inverse est-il mieux de rendre compte de toute variation lors du codage, pour ensuite avoir des catégories ayant des contours flous ?

Première étude de cas : estimer l'évolution de la distribution de Zipf chez l'enfant

Le corpus CoLaJE (Morgenstern, 2012) est la base de cette étude sur l'acquisition du français L1. Il est composé de sept suivis longitudinaux d'enfants qui ont été enregistrés une heure par mois, tous les mois, dès l'âge d'un an jusqu'à cinq ans environ. Le corpus respecte les standards de représentativité statistique demandés dans ce domaine (Stahl, 2004 ; Yamaguchi, 2018).

Pour chaque enfant il y a environ 8000 énoncés et 20000 mots avec une longueur moyenne d'énoncé (Mean Length of Utterance, Mac Whinney, 2000) de trois mots. Le langage adressé à l'enfant a également été enregistré et il est transcrit en utilisant les lignes FAT et MOT. Chaque transcription est soumise à une relecture par un pair, afin que les interprétations des expressions ambigües des enfants soient concordées par plusieurs chercheur.ses dans un souci de fiabilité et rigueur.



Loc	Ts	Te	Transcription L: 990 - 988 - 997 T: (-) P: - 0:00:09
CHI	0:24:03	0:24:13	‹regarde celui-là là il a trois essuie-glace› !
pho			‹ʁɔgaʁ sɥila la il a kwa sysygas›
mod			‹ʁəgaʁd sɥila la il a tɯwa ɛsɥiglas›
FAT	0:24:13	0:24:14	il a trois essuie-glace .
CHI	0:24:14	0:24:15	oui !
pho			wi
mod			wi
FAT	0:24:14	0:24:16	essuie-glace Adrien , comment tu dis ?
FAT	0:24:16	0:24:19	tu dis essuie-glace ?
CHI	0:24:19	0:24:21	‹celui-là il en a deux› !

Figure 1. Extrait de CoLaJE. ADRIEN-33-4_02_15

L'étude en question porte sur le développement de la distribution de la fréquence des mots chez les enfants du corpus CoLaJE visant à évaluer comment leur production lexicale soit liée à une distribution standard de la fréquence des mots : la loi de Zipf, qui est présente dans toutes les langues connues (Zipf, 1949 ; Piantadosi, 2014). Dans le détail, cette étude prend comme exemple des travaux précédents sur l'évolution de cette distribution de fréquence de mots qui avaient déjà été effectués sur plusieurs langues (Baixeries et al., 2013) en l'appliquant pour la première fois sur la langue française (Briglia et al., 2022).

La distribution de Zipf est considérée comme un équilibre d'efficience dans la communication humaine (Lestrade, 2017) : une langue doit pouvoir véhiculer le sens de manière précise tout en évitant de rendre cette tâche trop coûteuse pour les locuteurs. Le principe du moindre effort (Zipf, 1949) est fait de telle manière que la proportion entre "types" et "tokens" dans une forme de langage donnée (orale ou écrite) suffise pour atteindre le but communicatif : si par exemple un auteur d'un article peut s'assurer de se faire comprendre en utilisant une gamme de 70 mots différents, il n'y aura aucune raison pour qu'il en utilise plus puisque la valeur communicative des mots qui excèdent par rapport à la constante de Zipf ne vaut pas plus que le coût cognitif de les traiter. La constante de la loi de Zipf est considérée selon certains auteurs (Lestrade, 2017) comme un compromis implicite entre les locuteurs qui s'articule aux niveaux sémantique et syntaxique. Cette loi s'applique à l'oral tout comme dans le texte, avec des variations négligeables entre les deux formes (Piantadosi, 2014)

L'intérêt de vérifier comment cette constante se développe au cours de l'acquisition de la langue maternelle est donc celui de comprendre comment le langage de l'enfant en

évolution se rapproche d'une norme adulte d'efficience dans la communication. Pour prouver cette hypothèse, il a fallu opérer un choix méthodologique commun au sein du TAL. La production langagière des enfants du corpus CoLaJE qui ont été pris en examen se compose par trois lignes (voir exemple en Figure 1) : "pho" représente ce que l'enfant dit en API (Alphabet Phonétique International), "mod" représente ce que l'enfant aurait dû prononcer selon la norme adulte en API, et "CHI" représente ce que l'enfant aurait dû prononcer selon la norme adulte en orthographe standard. Avant de calculer la distribution de fréquence de mots dans un enregistrement, il faut d'abord comprendre ce qu'un mot est pour un enfant (Vihman & McCune, 1994). Par exemple, pour le mot cible « comprendre », Adrien[1] à l'âge de 4 ans et 3 mois (4_03_26) prononce les variations suivantes :

/ pÊopÊÉìd / et / kÉìpÊÉìd /.

Etant donné que le contexte est le suivant et que le papa voulait lui faire faire des exercices de lecture de lettres, il est clair que les deux formes variées ci-dessus se réfèrent à la même entité (e.g. le verbe « comprendre »). Il y a de nombreux cas analogues à celui-ci (par exemple le mot « tracteur » ou « pourquoi ») qui conduisent à un choix obligé : si l'on prend en compte chaque variation phonétique/phonologique de l'enfant, on ne pourra jamais étudier le développement de la constante de Zipf dans ce corpus, puisque le fait de considérer toute variation va entraîner un nombre d'occurrences très élevé alors que le signifié est toujours le même. En d'autres termes, il y aura plusieurs "types" différents alors qu'il n'y a – selon une certaine perspective – que plusieurs "tokens" différents qui se réfèrent au même "type".

Comme il a été remarqué par les porteurs du projet CoLaJE : « Cette distance entre formes et transcriptions se réduit à mesure que l'enfant grandit mais ne disparaît jamais. On est donc face à des choix théoriques importants dans la mesure où ils induisent les résultats des recherches menées sur les transcriptions. De quelle nature doit être la transcription ? Phonétique, phonologique, lexicale, orthographique ? » (Morgenstern, 2007, p56)

En outre, un mot donné peut être prononcé de plusieurs manières différentes avec des degrés de variation différents, ce qui rend les calculs complexes : il est difficile d'établir avec certitude si un enfant donne à un mot le même sens qu'un adulte lui attribue, par exemple des différences dues à des erreurs de sous-extensions ou de sur-extension par les enfants (Thomson & Chapman, 1977) peuvent être à l'œuvre sans que l'on puisse en être conscients. Il est difficile d'établir quand un mot signifie ce qu'il était censé signifier pour un enfant, et dans quelle mesure différentes formes variées se réfèrent à la même entité, notamment au cours des premiers âges (Vihman, 1994).

Il a donc été décidé – dans le but d'homogénéiser le corpus et rendre les comparaisons inter-enfants possibles – de baser la modélisation TAL sur le signifié/référent sans tenir compte des différentes images acoustiques qui indiquaient ce dernier. Ce choix a impliqué l'acceptation de biais potentiels liés au choix des transcripteurs qui pour les premiers avaient interprété la parole de l'enfant. Ces biais sont difficiles à estimer étant

donnée la taille importante du corpus. Au niveau du TAL, il s'agit de rassembler un ensemble de variations sous une catégorie unique liée au référent. Cela a permis de pouvoir traiter de manière automatique une grande quantité de données issus des enfants de CoLaJE afin de dégager l'évolution de la constante de la loi de Zipf au cours du temps (Briglia et al., 2022, p6-7). Il pourrait être résumé que le fait de renoncer à une variation à un niveau d'analyse (celui du mot) a permis de pouvoir analyser le rôle de la variation à un niveau supérieur (celui du lexique), selon une perspective temporelle qui met en relief les différences inter-enfants relativement à la variabilité intra-enfant.

La constante estimée est le paramètre exponentiel de la distribution de la fréquence des mots ("alpha") pour chaque enfant, ainsi que pour le langage des parents. Nous montrons comment les valeurs de "alpha" tendent à converger vers la valeur de 1 au cours du développement, ce qui est cohérent avec l'état de l'art (Baixieries et al., 2013). Le choix entre variation et norme expliqué ci-dessus a également permis de rapprocher le langage de l'enfant et celui de l'adulte, en établissant ainsi les bases pour une comparaison entre l'exposant "alpha" du langage des enfants et l'exposant "alpha" des adultes : le "rho" de "Spearman" montre une corrélation positive (p-value < 0.05) entre l' "alpha" de l'enfant et l' "alpha" des parents au cours de tous les âges, qui augmente à un âge plus avancé (Briglia et al., 2022, p184). Cela indique clairement que l' "input" parental joue un rôle de plus en plus important dans la structuration de l' "output" de l'enfant (Goodman et al., 2008).

Les trois graphes ci-dessous montrent la variation de l'exposant "alpha" au cours du temps. On pourrait considérer "alpha = 1" comme étant la norme puisqu'il a été démontré que cette valeur pour cet exposant donne le nombre optimal pour décrire combien de mots différents un extrait (écrit ou oral) d'une taille donnée a en moyenne à l'issue d'un compromis implicite atteint par les locuteurs (Zipf, 1949 ; Piantadosi, 2014). Si l'on compare les trois graphes on peut remarquer que les trois courbes ne sont pas isomorphes, et pourtant elles semblent graviter en dessous ou au-dessus de la valeur 1 au cours du temps (c'est-à-dire au cours du développement), ce qui expliquerait une tendance implicite du langage humain à atteindre l'équilibre décrit par la formule de Zipf (1949).

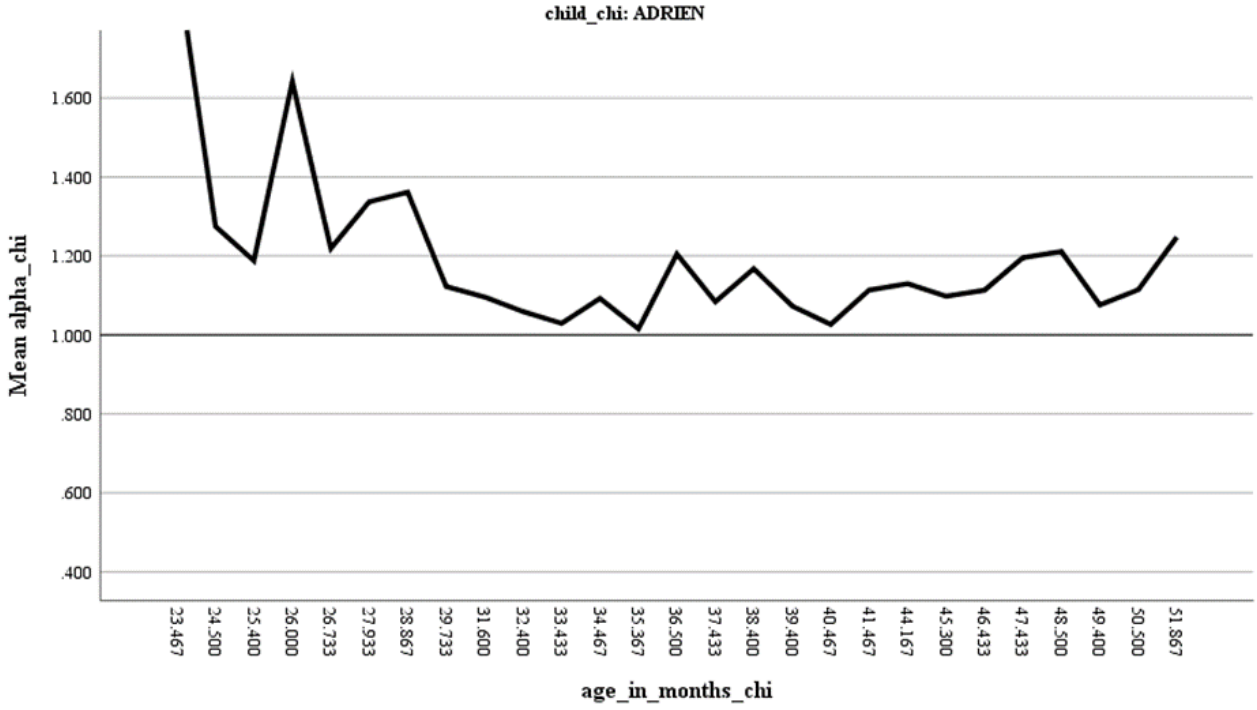


Figure 2. Evolution de l'exposant alpha pour Adrien

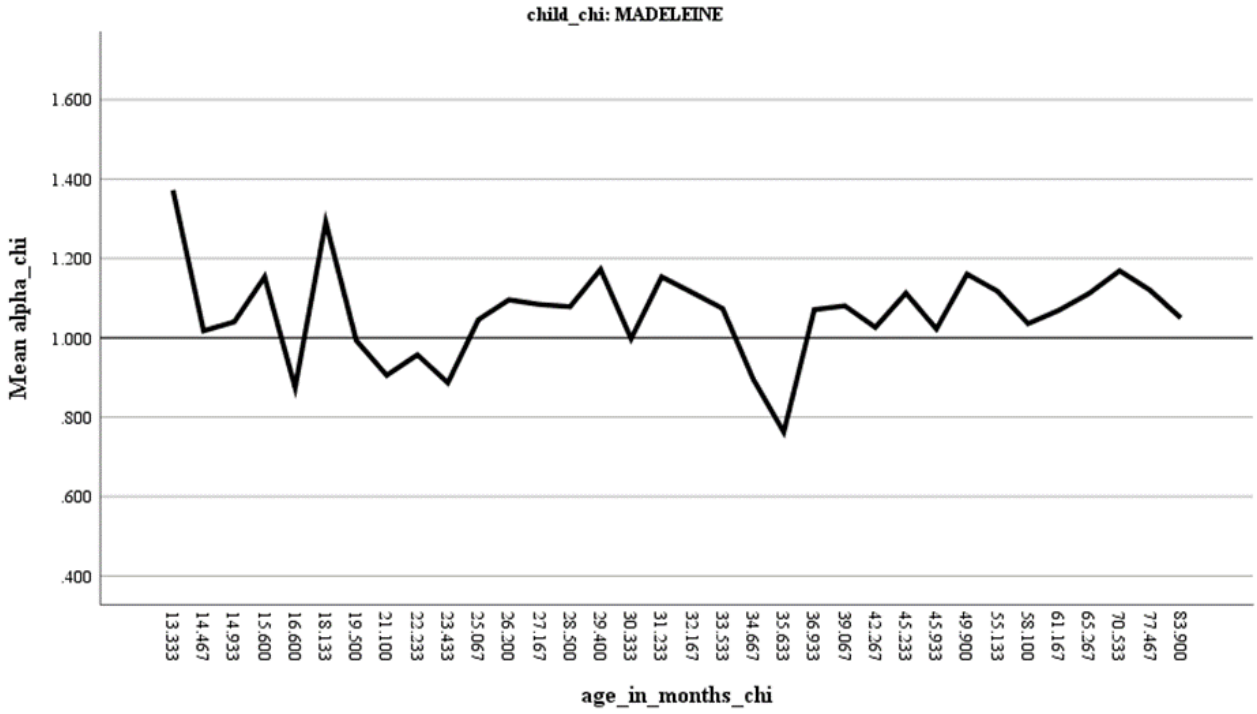


Figure 3. Evolution de l'exposant alpha pour Madeleine

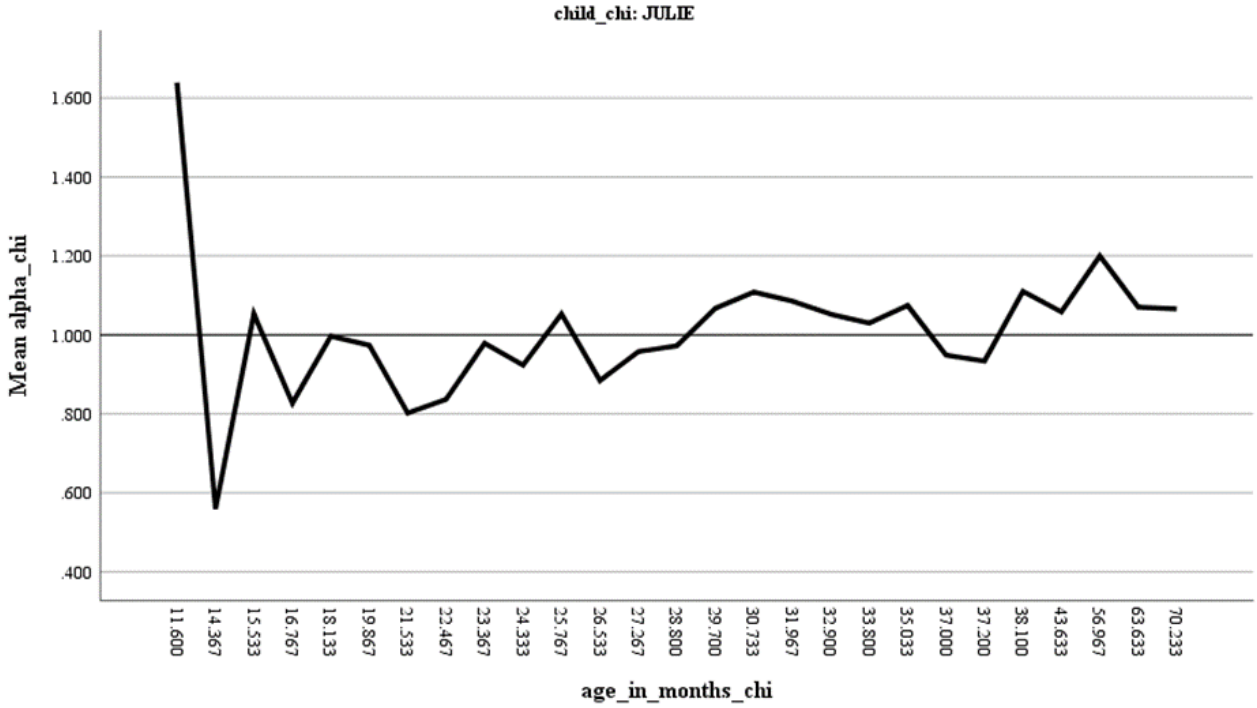


Figure 4. Evolution de l'exposant "alpha" pour Julie

La variation inter-enfants est également présente dans le cadre de l'acquisition des parties du discours : comme il a été montré dans une étude précédente sur le corpus CoLaJE (Mucciardi et al., 2022), Adrien et Madeleine présentent deux parcours d'acquisition des catégories grammaticales à la fois similaire (dans les proportions entre mots lexicaux et mots fonctionnels au cours des mêmes âges) et différent (dans le rythme d'acquisition)[2]. Le lien en bas de page montre efficacement ce parcours et permet de saisir le degré de différence et ressemblance entre les deux enfants. La modélisation graphique se révèle donc importante pour mieux apprécier la variation chez différents sujets.

Deuxième étude de cas : le statut de la norme et de la variation phonétique/phonologique

La variation est au cœur de l'acquisition du langage de l'enfant (Hickmann et al., 2018), elle influence toutes les étapes de ce processus, tant sur le plan de la perception que sur le plan de la production, ainsi que sur les différents niveaux d'analyse, en allant de la phonétique jusqu'à la pragmatique. On pourrait dire que le seul dénominateur commun de l'acquisition de la langue maternelle est la variation, puisqu'elle est présente tant au niveau inter-individuel qu'au niveau intra-individuel. Comme il est souligné par Bates : « il est nécessaire de relativiser cette apparente uniformité en soulignant la très grande variabilité intra et inter- individuelle qui caractérise cette acquisition » (Bates et al., 1995).

L’importance de la variation chez les enfants du corpus CoLaJE est bien représentée par les graphes qui montrent l’évolution de plusieurs indices linguistiques proposés par les chercheurs qui ont réalisé le corpus CoLaJE (Morgenstern, 2012). Avant d’atteindre la maîtrise de leur langue maternelle et de pouvoir parler comme un adulte, c’est-à-dire avant d’être capable au niveau perceptif et articulatoire de prononcer la forme cible (i.e. la norme sociale) d’un mot, les enfants passent à travers plusieurs étapes. La première est la reconnaissance du niveau suprasegmental, qui joue « un rôle important dans la mise en place des premières constructions grammaticales, notamment au moment de l’apparition des premiers mots et des premières combinaisons de mots, dans la période qui suit la période du mot isolé (stade holophrastique) » (Martel & Dodane, 2012, p13). La prosodie n’a pas été considérée dans cette étude pour des raisons de faisabilité, le focus étant sur lexique d’une part et la phonétique d’autre part. Cependant, les enfants basent leur acquisition sur la prosodie afin de détecter les pauses, les intonations et les accentuations qui les aident à visualiser la frontière entre mots ainsi que les relations de dépendance syntaxique. En fait « il semble bien que les caractéristiques prosodiques soient utilisées par l’enfant pour poser les fondements des futures constructions grammaticales, mais que celles-ci se manifestent différemment au moment des premiers mots (gabarit temporel des proto-mots et des premiers mots) et des premières combinaisons de mots (contours unitaires qui permettent d’assurer la cohésion des différentes unités au sein d’une unité plus grande) » (Martel & Dodane, 2012, p32-33).

Le but de l’exemple proposé est de modéliser la structuration des variations phonético-phonologiques au cours du temps, ainsi que d’estimer le degré de variabilité intra-enfant et inter-enfants. Des études précédentes (Dos Santos, 2007 ; Yamaguchi, 2012 ; Morgenstern, 2012) ont montré qu’il n’y a pas un parcours « typique » dans l’acquisition, mais plutôt des contraintes d’ordre phonétique et phonologique qui définissent les contours possibles du cheminement vers la norme adulte. Chaque variation semblerait être influencée par la variation précédente qui, à son tour, exercerait une influence sur la variation suivante (Sauvage, 2015). Dans d’autres termes, les variations ne seraient pas dues au hasard, mais elles seraient contraintes par plusieurs facteurs comme le lieu d’articulation, le mode d’articulation, ainsi que la fréquence d’occurrences d’une cible dans l’input parental (Ambridge et al., 2015).

Il y a essentiellement deux théories qui pourraient être adoptées afin de rendre compte des parcours d’acquisition : la théorie de l’optimalité (connue sous les termes anglais d' "optimality theory", Prince & Smolensky, 2004) et la théorie des traits phonologique (Clements, 1985). Ces théories font respectivement partie du courant innéiste et constructiviste. Dans cette étude, la théorie de Clements a été adoptée pour différentes raisons : l’auteur de l’article est convaincu qu’elle a un pouvoir explicatif plus profond et exhaustif de la théorie concurrente ; de plus, la majorité des références bibliographiques citées dans cet article adoptent le constructivisme (ou "usage-based theory" en anglais) comme point de départ des analyses. Cependant, le focus n’est pas sur la capacité de cette théorie à rendre compte de toutes les variations possibles dans les parcours d’acquisition des consonnes et des voyelles du français. Notre contribution vise plutôt à comprendre comment un algorithme de reconnaissance de motifs séquentiels ("pattern mining") peut nous aider à fouiller une grande base de suivis

longitudinaux qu'il serait impossible de traiter manuellement. La versatilité de cet algorithme pourrait fournir les bases pour comprendre quels sont les facteurs les plus importants dans l'acquisition des phonèmes parmi le lieu d'articulation, le mode d'articulation et la fréquence d'occurrences d'une cible dans l'input parental. Il est en fait difficile de pouvoir quantifier précisément quelle est la proportion entre ces facteurs.

Le corpus CoLaJE offre déjà – via la plateforme numérique Ortolang – un outil de requête précieux qui aide à cibler des mots précis et la possibilité de saisir des expressions régulières[3]. Les résultats proposés ont constitué le point de départ de notre travail, puis une analyse plus détaillée a été menée en utilisant la librairie « pylangacq » en langage Python[4] (Lee et al., 2016), ainsi que l'ensemble des algorithmes présents dans une autre librairie Python, appelée « pymining »[5]. Les exemples ci-dessous ont été choisis pour leur représentativité en fonction de plusieurs critères : la présence de plusieurs suites consonantiques, le fait d'avoir au moins deux syllabes, la présence de consonnes qui sont acquises relativement tard (le /Ê/ par exemple), leur fréquence élevée dans le corpus en question (c'est-à-dire, plusieurs occurrences différentes à des âges différents pour plusieurs enfants différents, ce qui permettrait de poser les bases pour une éventuelle généralisation d'un parcours typique).

Voici deux exemples d'application :

Premier exemple : pour le mot cible 'tracteur', /tÊaktoeÊ/, qui a une structure syllabique du type CCVCCVC, on liste toutes les variations phonético-phonologiques observées dans les transcriptions des enfants du projet CoLaJE (la valeur numérique correspondant à l'âge sur le modèle années_mois_jours) :

/kÊiktœÊ/ Antoine 2_02_27 /kÊatœÊ/ Antoine 2_02_27 /kÊatœÊ/ Antoine 2_03_05

/kÊakœÊ/ Antoine 2_04_03 /tatoÊ/ Théophile 2_10_28 /taktÉ/ Adrien 3_09_09

/toktÉÊ/ Adrien 4_00_15 /taktÉÊ/ Adrien 4_00_15 /taktœÊ/ Adrien 4_02_15

/taktœÊ/ Adrien 4_02_15 /saktœÊ/ Julie 1_06_04 (BRO) /ÊaktœÊ/ Julie 1_07_26

/tatø/ Julie 1_07_26 /tÊaktœÊ/ Julie 2_09_24 /tÊaktœÊ/ Julie 2_09_24

On observe que les variations autour de la norme (ou cible phonétique/phonologique) /tÊaktœÊ/ varient en fonction de l'âge et de l'enfant.

Deuxième exemple : mot cible crayon / kÊÉjÉÌ/ , structure syllabique CCVCV

/kÊÉjÉÌ/ Antoine 2_06_24 /kÊejÉÌ/ Antoine 2_06_24 /kÊejÉÌ/

Antoine 2_06_24

/tÊ□É□jÉ□Ì□/ Antoine 2_06_24 /kÉ□Ê□ejÉ□Ì□/ Théophile 3_02_00 /crejÉ□Ì□Ë□/ Théophile 3_04_10

/kÊ□É□jÉ□Ì□/ Théophile 3_05_11 /kÊ□É□jÉ□Ì□/ Théophile 3_07_08 /kÊ□É□jÉ□Ì□/ Théophile 4_03_29

/kÊ□É□jÉ□Ì□/ Théophile 4_09_07 /kijo / Adrien 4_01_12 /kÊ□É□jÉ□Ì□/ Julie 2_03_08

/kÊ□É□jÉ□Ì□/ Julie 2_11_01 /kÊ□É□jÉ□Ì□/ Julie 3_04_21 /kÊ□É□Ê□jÉ□Ì□/ Julie 3_04_21

/kÊ□É□jÉ□Ì□/ Julie 3_04_21 /kejÉ□Ì□/ Anae 2_00_26 /tejÉ□Ì□/ Anae 2_00_26

/tijÉ□Ì□/ Anae 2_00_26 /tijÉ□Ì□/ Anae 2_00_26 /kÊ□ejÉ□Ì□/ Anae 2_06_27

/jÊ□ajÉ□Ì□/ Anae 2_08_24 /kÊ□ejÉ□Ì□/ Anae 5_10_30

Le dénominateur commun entre Anaë et Julie est qu'elles semblent – autour de l'âge 2 ans et demi/trois ans – avoir appris une fois pour toutes la forme correcte du mot cible, puisqu'elles arrivent à bien l'articuler à des intervalles de temps successifs. Cependant, elles produisent une variation qu'elles n'avaient jamais produit auparavant. Ce phénomène, bien qu'il soit contre-intuitif – il est assez commun en acquisition L1 (Sauvage, 2015, p125, en particulier le phénomène de régression). Il faut remarquer qu'il pourrait s'agir également d'une variation qui n'avait simplement pas été collectée par la densité d'échantillonnage d'une heure par mois prévue par le projet CoLaJE, (voir les réflexions de Yamaguchi autour de la représentativité - Yamaguchi, 2018).

La procédure pour repérer et analyser les variations est la suivante :

i) chercher le mot désiré via la requête du projet Ortolang

ii) avoir accès aux transcriptions des enfants CoLaJE par le biais de la librairie Pylangacq

iii) mettre en place un algorithme du type "if-then" et vérifier que le mot prononcé n'est pas différent du mot cible

iv) S'il l'est, détecter sa structure syllabique via Pymining et si non, la ligne du code se termine ainsi.

La partie la plus difficile consiste en la définition de la variation, c'est-à-dire qu'une fois que la variation a été détectée, il faudrait apprendre à la machine à la classer dans une des catégories ci-dessous, qui à leur tour se basent sur plusieurs critères (lieu et mode

d’articulation, voisement, permutation dans l'ordre des syllabes, ajout ou suppression d’une syllabe/phonème, etc.) :

1) Omission

2) Substitution

3) Assimilation

4) Réduction

5) Duplication

6) Epenthèse

7) Métathèse

Cet essai s'est arrêté à la structure syllabique car il a été difficile de programmer la partie concernant les 7 variations phonétiques possibles : trop de variables et trop d’étapes conséquentielles étaient présentes. Par exemple, après avoir détecté une substitution, il aurait fallu aussi trouver un moyen de classer cette substitution en fonction du phonème remplacé : une substitution de fricatives par des occlusives n’est pas équivalente à une substitution de liquides par des semi-voyelles. Un autre exemple encore plus complexe : dans l’assimilation, deux sons deviennent semblables au niveau du lieu d’articulation, du mode d’articulation ou du voisement, mais l’on voit bien qu’il ne serait pas rigoureux de mettre sur le même plan ces trois critères. Il aurait peut-être fallu concevoir une hiérarchie, mais laquelle ?

Prenons un dernier exemple, le cas des métathèses. L’écueil principal a été le nombre et la variété de ces dernières : aéroport → [aÊeopÉÊ] n’est pas identique au cas suivant : toboggan → [togobÉÌ]. Dans le premier cas il s’agit d’une métathèse entre une consonne et une voyelle, dans le deuxième cas, d’une métathèse entre deux consonnes.

On pourrait également ajouter une autre difficulté : les variations liées au processus phonétique/phonologique énumérées ci-dessus peuvent avoir lieu en début de mot, au milieu ou à la fin, et elles peuvent concerner une seule consonne, une seule voyelle ou bien une syllabe.

Durant la réflexion autour de la multiplicité de ces variations, des questions ont été récurrentes : puisqu’il y a des variations de nature différente, est-ce qu’il faut attribuer un poids différent selon la nature de la variation ? Quels critères pourrait-t-on adopter afin d’attribuer ce poids ?

Malheureusement, il n’a pas été possible de prendre en compte toutes ces possibles variations, trop de facteurs concurrents étaient en jeu et les compétences de l’auteur ne sont pas à la hauteur d’une tâche algorithmique si complexe. Néanmoins, certains

travaux ont conduit à un travail analogue, par exemple le réseau neurones qui prend en compte à la fois l'aspect phonétique et phonologique du logiciel PRAAT (Boersma et al., 2020), qui propose des pistes qui pourraient répondre aux questionnements ci-dessus.

Il est difficile de dégager un parcours typique à partir de ces variations : le nombre et la nature des variations est relativement trop grand. Le premier obstacle est d'ordre purement statistique et concerne la relation entre échantillon et population : malheureusement, il n'y avait pas moyen d'avoir une occurrence de chaque mot pour chaque enregistrement mensuel et pour chaque enfant du corpus CoLaJE. Même les mots les plus fréquents peuvent parfois manquer, notamment aux plus jeunes âges lorsque les enfants parlent relativement peu. La deuxième difficulté est de comprendre pourquoi une variation s'est produite à la place d'une autre.

Par exemple, pourquoi /toktÉ□Ê□/ Adrien 4_00_15 et /taktÉ□Ê□/ Adrien 4_00_15 ? Il serait difficile de croire que l'enfant à 4 ans ne soit pas capable de percevoir et articuler la différence entre les voyelles /o/ et /a/.

Le troisième obstacle réside dans l'interprétation de la cause de la variation, c'est-à-dire d'identifer les motivations d'un enfant à prononcer telle variation ou une autre. Par exemple, une stratégie d'évitement qui porte les enfants à omettre ou à réduire une consonne cible qui demande trop d'effort articulatoire, comme dans le cas suivant : /tatoÊ□/ Théophile 2_10_28, ou bien une assimilation, quand l'enfant tend à préférer les suites syllabiques qui ont un point d'articulation en commun, comme dans : /kÊ□akœÊ□/ Antoine 2_04_03.

Après avoir essayé plusieurs combinaisons d'algorithmes pour plusieurs mots différents, nous nous sommes confrontés aux limites de l'approche informatisée. Il n'est possible que de confirmer les tendances d'acquisition qui ont déjà été confirmées par la littérature existante (Dos Santos, 2007 ; Yamaguchi, 2012), notamment l'ordre d'acquisition des voyelles ou des consonnes, ainsi que les variations les plus fréquentes et les moins fréquentes. Mais pour ce qui concerne la prédiction avec un degré de précision acceptable, il a été difficile d'envisager la compréhension des suites de variations au cours du temps : quelle variation suivra en fonction des deux variations précédentes ? Cette question reste sans réponse.

La combinaison d'algorithmes s'est révélée être une méthode infructueuse, la variabilité intra-enfant et inter-enfants étant trop grande. Une autre piste possible pourrait être de se focaliser sur un sujet plus restreint, par exemple explorer les variations syllabiques analogues comme les occlusives-liquides. On pourrait commencer en dressant une liste suffisamment représentative de mots qui contiennent ce type de syllabe et procéder étape par étape (cf les 4 étapes listées ci-dessus). Ce focus devrait permettre de réduire considérablement le nombre de variations possible et rendre par la suite la tâche de programmation plus simple.

Pour conclure, ces résultats montrent comment il est a priori bénéfique de modéliser les multiples variations phonétiques/phonologiques à l'aide d'outils de TAL : on s'aperçoit

que – malgré le fait que la nature des variations soit multiforme et leur nombre élevé – elles peuvent être inclues dans un seul modèle qui pourrait rendre compte des règles qui régissent les parcours possibles de leur évolution. Comme il a déjà été dit, les résultats présentés dans cette étude n'ont qu'une valeur anecdotique : ils s'accordent de manière globale à des études de cas qui ont été menées sur le même corpus (Yamaguchi, 2012) ou sur d'autres enfants francophones collectés avec des méthodes comparables (Dos Santos, 2007).

Ce travail contient une partie de nombreux travaux de fouille et modélisation du corpus CoLaJE qui ont été produits lors d'une collaboration entre linguistes et informaticiens de l'Université Paul Valéry Montpellier3 (pour la précision, il s'agit des data scientists du master MIASHS guidé.e.s par S. Bringay) pendant l'année académique 2019-2020. Pour un aperçu de ces travaux, veuillez suivre le lien en bas de page[6].

Conclusion

Le but de cet article était de mener une réflexion autour de l'utilisation de modèles et techniques de TAL pour mettre en relief la relation entre norme et variation dans le cadre de l'acquisition du français langue première.

Deux cas d'étude ont été proposés : dans le premier, la variation avait une double articulation au niveau lexical et au niveau du vocabulaire de l'enfant. Les résultats d'une étude précédente (Briglia et al., 2022) ont montré comment la création d'un modèle unifié de la catégorie de mot (conçu comme une unité composée par trois constituants : signifiant-signifié-référent) permet de rassembler plusieurs variations phonétiques/phonologiques sous une même catégorie afin de faciliter l'analyse d'un autre type de variation, celle de l'exposant alpha, un indice qui représente comment la distribution de fréquence des mots dans le vocabulaire de l'enfant varie respectivement :

i) au cours du temps (intra-enfant)

ii) entre les enfants (inter-enfants)

iii) entre les enfants et leurs parents (corrélation de Spearman).

Pour cette dernière analyse, le codage des transcriptions CHI-FAT-MOT, la mise au point de critères pour unifier les variations sous un seul ensemble ainsi que le calcul des fréquences d'occurrence et des corrélations a été fait automatiquement en langage Python.

Alors que dans le deuxième exemple on a pu apprécier la rapidité des algorithmes de reconnaissance de motifs séquentiels et comprendre comment la prise en compte de toutes les variations phonétiques-phonologiques autour de la norme adulte est

théoriquement faisable, dans la pratique il est difficile d'attribuer la bonne place et le juste poids aux différents critères articulatoires qui définissent les variations.

L'application de modèles, de techniques et de référentiels issus de l'informatique dans le domaine de la linguistique est croissant et permet la vérification d'hypothèses de manière fiable, reproductible et rapide. De plus, la plupart des logiciels pour l'analyse de corpus (Antconc, TXM, Iramuteq), de la parole (PRAAT, PHON) ou de la gestualité (ELAN) sont en libre accès et "open source", ce qui représente un véritable atout.

Malgré ces avantages, l'adoption des techniques de TAL ne doit pas être interprétée comme un passepartout qui se fait a priori sur une connaissance approfondie de la langue elle-même ou du phénomène linguistique (l'acquisition de la L1 par exemple). La rapidité et la puissance de calcul doivent être dirigées par des assomptions, des hypothèses, des cadres théoriques que – à l'heure d'aujourd'hui – seuls les intelligences humaines peuvent maîtriser.

D'autres outils de TAL développés au sein de la communauté francophone qui pourraient être utilisés afin d'évaluer l'acquisition du français langue première chez l'enfant sont par exemple l'iPhocomp (Lee et al., 2014) et l'ISC (Index de Complexité Syntaxique, Szmrecsanyi, 2004). En effet, lorsque l'on dispose de suivis longitudinaux disponibles sous plusieurs formats différents comme pour le corpus CoLaJE, on a par conséquent l'opportunité d'obtenir un score pour chaque mot et/ou énoncé prononcé par l'enfant en automatisant – par le biais d'un langage de programmation comme Python - la tâche de calcul de ces scores pour chaque ligne, qu'il s'agisse de la ligne CHI, PHO ou MOD, et quel que soit son format (csv, CHAT ou TEI, pour ne citer que les formats présents sur CoLaJE-Ortolang). Une étude récente a montré la validité de l'emploi de ces deux scores pour la prédiction de l'acquisition de certaines catégories grammaticales sur une étude de cas (sur l'enfant Adrien) tiré du corpus CoLaJE (Briglia et al., 2022).

Au cours de ces dernières années, la technologie de TAL qui semblerait être la plus complète et exhaustive, le BERT (acronyme pour "Bidirectional Encoder Representations from Transformes") a été améliorée (en termes de performance pour la langue française) grâce à la prise en compte des particularités de la langue visée. C'est ainsi que CamemBERT (Martin et al., 2020) a pu voir le jour.

On pourrait craindre que cette augmentation constante de la présence de l'informatique dans le champ d'investigation qui a traditionnellement fait partie de la linguistique cause – dans un avenir proche ou lointain – un déclassement de cette dernière. Ces craintes sont vraisemblables, pourtant il est à noter que tout système d'annotation automatique en parties du discours ("POS tagging" en anglais), classification de texte, plongement de mots en allant jusqu'aux dernières technologies d'apprentissage par la machine (BERT ou, plus généralement, les réseau de neurones, qu'ils soient supervisés ou pas), ne peut pas être conçu sans une connaissance linguistique préalable. De plus, bien que l'intelligence artificielle soit toujours plus raffinée dans ses prédictions et ses inférences sur le langage, elle présente des problèmes récurrents au niveau de la

coarticulation (les technologies "speech-to-text" et "text-to-speech"), la synonymie et la polysémie, ainsi que pour ce qui concerne la signification en contexte (i.e le niveau pragmatique). En d'autres mots, tout ce qui relève de la compréhension des différents accents ou des différentes acceptions, du style, de la nuance, de variation en fonction du contexte, d'ambiguïtés ou bien de sous-entendus reste encore difficile à détecter pour les machines. La souplesse, ainsi que la créativité, sembleraient devoir rester des compétences mieux maîtrisées par les intelligences humaines.

Ces différences nous montrent comment une synergie entre linguistes et informaticiens pourrait constituer le noyau d'une bonne partie des futures recherches dans le domaine du langage.

Bibliographie

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of child language*, 42(2), 239-273

Baixeries J., Elvevag B. and Ferrer-i-Cancho R. (2013). The Evolution of the Exponent of Zipf's Law in Language Ontogeny. PLoS ONE 8(3): e53227

Bates, E., Dale, P. S., & Thal, D. (1995). *Individual differences and their implications for theories of language development*. The handbook of child language, 30, 96-151

Boersma, P., Benders, T., & Seinhorst, K. (2020). Neural network models for phonology and phonetics. *Journal of Language Modelling Vol, 8*(1), 103-177

Briglia A. "Statistical and computational approaches to first language acquisition. Mining a set of French longitudinal corpora (CoLaJE"). Thèse Université Paul Valéry Montpellier 3; Università di Messina. 2021.

Briglia A., Mucciardi M., Pirrotta G. (2022). "A statistical model for predicting child language acquisition: unfolding qualitative grammatical development by using logistic regression model". In Salvati N., Perna C., Marchetti S., Chambers R. "Studies in Theoretical and Applied Statistics". Springer Proceedings in Mathematics & Statistics. PROMS, volume 406. SIS 2021, Pisa.

Briglia A., Mucciardi M., Pirrotta G. "The development of word frequency distribution in first language acquisition. An analysis on a spoken language corpus of French children". Vadistat Press. *Proceedings of the 16th International Conference on Statistical Analysis of Textual Data (JADT)*, 1 (16)

Clements, G. N. (1985). The geometry of phonological features. *Phonology yearbook* 2.225-252

Dos Santos, C. (2007). *Développement phonologique en français langue maternelle: une*

étude de cas (Doctoral dissertation, Université Lumière Lyon 2).

Dubois, J., Marcellesi, J-B., Méyel, J-P. & Giascamo, M. (1973). *Dictionnaire de linguistique*. Paris : Larousse

Goodman J., Dale P. and Li P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, 35(03), 515–531

Hickmann M.; Veneziano E.; Harriett J. (Eds) (2018). Sources of Variation in First Language Acquisition. *Languages, contexts and learners*. John Benjamins

Hoedt, A., & Piron, J. (2016). *La faute de l'orthographe*. Paris, *Textuel*

Barge J. S. Pour une nouvelle conception de la "norme" linguistique dans l'enseignement des langues. hal-00385090v2

Jusoh, S. (2018). A study on NLP applications and ambiguity. *Journal of Theoretical & Applied Information Technology*, 96(6)

Kraif O., Ponton C. (2007). Du bruit, du silence et des ambiguïtés : que faire du TAL pour l'apprentissage des langues ? In *Actes de la 14ème conférence sur le Traitement Automatique des Langues Naturelles*. Posters, pages 143–152, Toulouse, France. ATALA

Lee, H., Gambette, P., Barkat-Defradas, M. (2014). iPhocomp: calcul automatique de l'indice de complexité phonétique de Jakielski. *JEP 2014, XXXè édition des Journées d'Etudes sur la Parole*, Le Mans, France. pp.622-630, 2014, Actes de la XXXe édition des Journées d'Etudes sur la Parole.

Lee, Jackson L., Ross Burkholder, Gallagher B. Flinn, and Emily R. Coppess. (2016). Working with CHAT transcripts in Python. *Technical report TR-2016-02,* Department of Computer Science, University of Chicago.

Lestrade S. (2017). Unzipping Zipf's law. *PlosOne*

MacWhinney, B. (2000). The Childes Project: Tools for Analyzing Talk, Volume II: the Database (3rd ed.). *Psychology Press*

Martel, K., & Dodane, C. (2012). Le rôle de la prosodie dans les premières constructions grammaticales : étude de cas d'un enfant français monolingue. *Journal of French Language Studies*, 22(1), 13-35

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, É. V., Sagot, B. (2020). CamemBERT: a Tasty French Language Model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*

Morgenstern, A., & Parisse, C. (2007). Codage et interprétation du langage spontané d'enfants de 1 à 3 ans. *Corpus*, (6), 55-78

Morgenstern A.; Parisse C. (2012). The Paris Corpus. *French language studies* 22. 7-12. Cambridge University Press. Special Issue.

Mucciardi M., Pirrotta G., Briglia A., Sallaberry A. (2021). Visualizing cluster of words: a graphical approach to grammar acquisition. In Giovanni C. Porzio; Carla Rampichini; Chiara Bocci (Eds). *CLADAG 2021 BOOK OF SHORT PAPERS. 13th Meeting of the Classification and Data Analysis Group* - Firenze University Press

Piantadosi S. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychon Bull Rev*.; 21(5): 1112–1130

Prince, A., Smolensky P. (2004): Optimality Theory: Constraint Interaction in Generative Grammar. *Blackwell Publishers*.

Sauvage J. (2015). L'acquisition du langage : un système complexe. *L'Harmattan*, Louvain

Siouffi, G., & Steuckardt, A. (éds). (2007). *Les linguistes et la norme*. Berne : Peter Lang

Srikant R., Agrawal R. (1996). Mining Sequential Patterns: Generalizations and Performance Improvements. *Proceedings of the 5th International Conference on Extending Database Technology (EDBT'96)*. Avignon. France. p. 3-1

Szmrecsanyi, B. (2004). On operationalizing syntactic complexity, in: Purnelle, Gérard, Cédrick Fairon and Anne Dister (eds.), Le poids des mots. *Proceedings of the 7th International Conference on Textual Data Statistical Analysis. Vol. 2*. Louvain-la-Neuve, Presses Universitaires de Louvain.

Thomson, J. R., & Chapman, R. S. (1977). Who is daddy revisited: The status of two-year-olds' over-extended words in use and comprehension. *Journal of Child Language*, 4(3), 359–375

Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough ?. *Journal of child language*, 31(1), 101-121

Vihman, M. M. and McCune L. (1994). When is a word a word? *Journal of Child Language*, 21(3), 517–542

Yamaguchi N. (2012). Parcours d'acquisition des sons du langage chez deux enfants francophones. Phd thesis, Sorbonne Nouvelle University (Paris 3)

Yamaguchi N. (2018). What is a representative language sample for word and sound acquisition? *Revue canadienne de linguistique*. University of Toronto Press. 63 (04), pp.667-685

Zipf G.K. (1949). Human behaviour and the principle of least effort. *Addison-Wesley*. Cambridge (MA), USA

[1] Lien pour le point précis de l'enregistrement où se trouve le mot cible (utiliser la requête pour trouver d'autres mots) : https://ct3.ortolang.fr/tools/trjsbrowser/trjs.html?f=/data/colaje/adrien/ADRIEN-34-4_03_26/ADRIEN-34-4_03_26.tei_corpo.xml&m=/data/colaje/adrien/ADRIEN-34-4_03_26/ADRIEN-34-4_03_26-480p.mp4&time=1380.0&nowave

[2] http://advanse.lirmm.fr/EMClustering/

[3] https://ct3xq.ortolang.fr/ct3xq/interro

[4] https://pylangacq.org/

[5] https://github.com/bartdag/pymining

[6] https://marine27.github.io/TER/index.html