UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ANALYSE MULTIDIMENSIONNELLE DE L'ACQUISITION DES PARTICULES DE FIN DE PHRASE PAR LES ENFANTS LOCUTEURS DU CANTONAIS

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

POUR LA MAÎTRISE EN LINGUISTIQUE

PAR

DAVID VERGNAUD

JANVIER 2026

REMERCIEMENTS

Un mémoire est un travail de longue haleine, imprégné des phases de vie qu'il traverse, inspiré des personnes qu'il croise, et égayé de chacune des étincelles qui se sont allumées dans nos yeux pendant toute la durée de sa réalisation. C'est donc le produit d'une collaboration cosmique entre d'innombrables flux d'énergie vibratoire parcourant l'univers, dont certains peuvent, méritent et doivent être nommés pour que soient apposées les dernières touches de couleur au tableau astral dont ils sont les artisans primordiaux.

Mes premiers remerciements, et bien plus encore, vont à Jess, pour le soutien indéfectible qu'elle m'octroie, tant dans mes lubies et engouements personnels que dans mes ambitions académiques, et bien sûr plus béatement pour son ardente et patiente présence à mes côtés, nos discussions, nos rires, nos questionnements, nos projets, nos petits billards. C'est grâce à toi que je me suis rendu là, et avec toi que je veux continuer.

Le deuxième pilier de cette aventure, qui m'a insufflé l'idée et la motivation d'offrir mon corps et mon esprit à la science, est bien sûr Marie-Jo, sans qui la linguistique ne serait restée pour moi qu'un lointain concept ésotérique. D'enseignante à collègue et bien sûr amie, elle m'a aiguillé sur le chemin de la maîtrise, et bien malin qui aurait pu imaginer à l'époque que nous nous retrouverions un jour côté à côte sur les bancs d'école, à la poursuite de notre doctorat. Quel hasard également qu'au moment où je me prépare à remettre ce mémoire, tu t'apprêtes à valider toi aussi les nombreux crédits accumulés dans le cadre de ta vie familiale. Je te souhaite tout le bonheur du monde, et merci pour toute cette tranche de vie.

Je remercie chaleureusement ma directrice et mon directeur de recherche, Marine et Grégoire, pour m'avoir accompagné et encouragé durant ce parcours, et m'avoir fait bénéficier de leur expérience, leur expertise et leur sagesse, pour leur disponibilité ainsi que leurs conseils éclairés, et pour avoir fait en sorte que mes efforts puissent enfin se matérialiser en un document que j'espère digne des attentes qu'ils ont placées en moi.

Merci aux membres de mon jury, Naomi Nagy et Lyn Tieu de l'Université de Toronto, qui se sont prêtées de bonne grâce à cette expérience, ont cru en mon projet et m'ont permis de l'améliorer grâce à leurs

remarques et questions constructives; je leur souhaite de trouver dans la lecture de ce manuscrit un intérêt scientifique et des réponses éclairantes.

Pour leur accueil durant la phase finale de ce projet, pour m'avoir offert un espace de calme et de sérénité où j'ai pu m'absorber dans mes recherches, et pour leur soutien et leur compréhension tout au long de mon parcours vis-à-vis de mes choix et de mon travail, je remercie mes parents, Marie-Claude et Michel. Dire que je ne serais pas là sans eux serait bien sûr pléonastique, mais ne pas le dire serait ignorer l'importance du rôle qu'ils jouent dans ma vie et de l'inspiration qu'ils sont pour moi.

Offrir des étoiles n'est pas chose aisée, sauf peut-être à l'école primaire ou à la fin des matches de hockey, mais c'est bien ce qu'a fait le Parc National du Mont Mégantic en acceptant que je vienne y découvrir et y faire découvrir l'astronomie, et y commencer mon travail de recherche dans l'environnement le plus sauvage et le plus céleste dans lequel il m'a été donné de passer un été, le chalet La Voie Lactée. Merci à Sébastien, merci à Claude, merci à l'équipe, et merci au ciel étoilé.

Enfin, pour les moments où la pression doit s'évacuer, pour les soirées ou les *weekends* de relaxation volée où l'univers se réduit à une table en bois garnie de barres et de petits bonshommes, et aux personnes qui gravitent autour dans leur voie maltée, merci au monde du *babyfoot*, cette grande famille toujours prête à ouvrir les bras pour lever son verre et activer la petite balle jaune.

Merci à Antoine, Boris, Didier, Émilie, Fred, Iris, Steph, Maël, Magalie, Minou, Rachel, Vava pour les apéros, les jeux et le *fun* qu'on n'arrête jamais d'avoir.

Merci à Anna, Simon-Hubert, Liberty et leurs amis libres pour leur énorme travail, leur disponibilité, leur bienveillance, leur résilience, et leur insatiable soif de connaissances dont ils me permettent de profiter.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS, DES SIGLES ET DES ACRONYMES

A_M_A : structure interrogative produisant une phrase interrogative polaire, dans laquelle la particule négative 唔|m4 est insérée entre une réduplication de la première syllabe d'un verbe ou adjectif et la forme complète du même terme

ADS : *adult-directed speech*, discours adressé aux adultes

CC : corpus CANCORP

CC2012 : version 2012 du corpus CANCORP

CDS : *child-directed speech*, discours adressé aux enfants

CS : *child-speech*, discours produit par l'enfant

GLMM : *generalized linear mixed-effect model*, modèle linéaire généralisé à effets mixtes

HAI_M_HAI : structure interrogative figée produisant une phrase interrogative polaire, constituée du composant 係|hai6 réduplicqué et de la particule négative 唔|m4 insérée entre les deux occurrences

HKU : corpus HKU-70 de la *Hong Kong University*

LWL : version du corpus CANCORP retravaillée par Lee, Wong et Leung

MLU : *mean length of utterance*, longueur moyenne des énoncés

SFP : *sentence-final particle*, particule de fin de phrase

LISTE DES SYMBOLES ET DES UNITÉS

$\bar{a}$: tranche d'âge

$\bar{m}$: tranche de MLU

RÉSUMÉ

Les particules de fin de phrase (SFP) constituent un élément incontournable du cantonais oral, dans lequel elles sont utilisées très fréquemment par les adultes pour exprimer un riche éventail de nuances sémantiques et pragmatiques. Les SFP apparaissent relativement tôt dans le discours des enfants, mais ceux-ci mettent plusieurs années à s'approprier la totalité de l'inventaire utilisé par les locuteurs expérimentés. Les études ciblant l'acquisition des SFP ont permis de mettre en évidence l'existence d'une certaine régularité dans le processus d'émergence de ces particules, mais la variabilité constatée indique clairement l'influence de nombreux facteurs sur l'utilisation qu'ils en font, en particulier en ce qui concerne l'ordre d'apparition et la fréquence de production.

Les recherches antérieures nous ont permis d'identifier certains facteurs susceptibles d'avoir une influence significative sur l'acquisition des SFP, que nous avons regroupés en trois catégories : les facteurs individuels (genre, âge, niveau de développement langagier), les facteurs linguistiques (structures syntaxiques reconnaissables, phono-prosodie) et les facteurs extrinsèques (caractéristiques de l'interlocuteur, caractéristiques du discours adressé à l'enfant). Pour valider la pertinence de ces facteurs et mesurer leur impact, nous procédons au moyen d'un programme informatique à l'extraction systématique des occurrences de SFP et des contextes associés dans le discours d'enfants et de leurs interlocuteurs dans le cadre de deux corpus d'interactions enfants–adultes, le CANCORP et le HKU-70. Les annotations générées sont ensuite compilées sous forme d'analyses descriptives et utilisées pour l'ajustement de modèles statistiques.

Nos analyses montrent que la majorité des facteurs considérés jouent effectivement un rôle dans l'utilisation que les enfants font des SFP, et que ces rôles changent au cours du développement des enfants. Nous constatons par ailleurs que la production des SFP est largement influencée par le contexte discursif, tant linguistique (tendance à produire plus de SFP lorsque l'interlocuteur en utilise également, ou lors de la formulation d'énoncés interrogatifs) que pragmatique (tendance à produire plus de SFP vis-à-vis d'interlocuteurs étrangers au foyer et vis-à-vis de femmes). Ces résultats confirment que l'acquisition des SFP doit être considérée comme un processus ancré dans le développement cognitif de l'enfant, et qu'elle répond autant à une complexification naturelle du discours qu'à l'évolution des besoins pragmatiques.

Mots clés : particules de fin de phrase (SFP), acquisition du langage, cantonais.

SUMMARY

Sentence-final particles (SFP) constitute a central element of spoken Cantonese, in which they are used extensively by adults to express a rich variety of semantic and pragmatic nuances. Children start to use SFPs rather early in their discourse, but it takes them several years to integrate into their own speech the full inventory that experienced speakers use. Studies focusing on the acquisition of SFPs have shown the emergence process of these particles to follow some sort of regularity, but have also uncovered some variability that clearly indicates that numerous factors play a role in the use children make of them, particularly with regard to the order of appearance and frequency of production.

Previous research has allowed to identify some of the factors likely to have a significant impact on the acquisition process of SFPs. We have grouped these factors into three categories: individual factors (gender, age, level of linguistic development), linguistic factors (particular syntactic structures, phono-prosody) and extrinsic factors (characteristics of the interlocutor, characteristics of the child-directed speech). In order to validate the relevance of these factors and evaluate their impact, we used a computer program to systematically extract all occurrences of SFPs and their associated contexts from the speech of children and their interlocutor in two child–adult interaction corpora, CANCORP and HKU-70. The resulting annotations were then compiled in the form of descriptive analyses and used to fit statistical models.

Our analyses show that most of the factors considered do indeed play a role in the use that children make of SFPs, and that their roles change during the children's development. We also observe that the production of SFPs is largely influenced by the discursive context, both linguistic (tendency to produce more SFPs when the interlocutor uses them too, or when uttering interrogative sentences) and pragmatic (tendency to produce more SFPs when talking to strangers or to women). These results confirm that the acquisition of SFPs must be recognized as a process that is deeply anchored into the cognitive development of the child, and that it responds to a natural increase in the speech complexity as well as to an evolution of their pragmatic needs.

Keywords: sentence-final particles (SFPs), language acquisition, Cantonese.

INTRODUCTION

Le cantonais, langue sinitique parlée par plus de 75 millions de locuteurs à travers le monde, se démarque par son importante productivité en termes de particules de fin de phrase (*sentence-final particles*, dorénavant SFP). Ces unités linguistiques ponctuent la vaste majorité des énoncés du discours oral des adultes, et sont utilisées pour exprimer un large éventail de fonctions et nuances sémantiques et pragmatiques.

Bien que les SFP apparaissent dans le discours des enfants dès l'âge d'un an et demi, le processus d'acquisition s'étend sur de nombreuses années : la quantité importante de SFP différentes, ainsi que la finesse et la variété des valeurs qu'elles permettent d'exprimer, rendent le processus d'acquisition complexe pour les enfants, qui, vers l'âge de cinq ans, ne maîtrisent qu'une fraction de l'inventaire complet.

Notre objectif, par le présent travail de recherche, est d'évaluer le rôle joué par un certain nombre de facteurs dans le processus d'acquisition des SFP pendant les premières années de la vie des enfants cantonophones. À travers une analyse systématique du discours produit par et adressé aux enfants dans le cadre de deux corpus de référence, nous étudierons la dynamique d'émergence des particules produites par les enfants entre un an et demi et cinq ans et demi, et proposerons une interprétation des effets des facteurs considérés ainsi que des interactions qui les lient au cours de cette période.

Notre premier chapitre présentera de manière approfondie le cadre de notre recherche; nous y évoquerons tout d'abord le rôle et la diversité des SFP en cantonais et dans d'autres langues, ainsi que l'historique des travaux effectués en lien avec leur acquisition. Nous expliciterons par la suite, à travers une revue de littérature exhaustive, les différents facteurs individuels, linguistiques et extrinsèques susceptibles d'avoir un impact sur ce processus.

Ces considérations nous permettront, au chapitre 2, d'introduire notre question de recherche et de formuler des hypothèses sur les résultats de nos analyses.

Les trois chapitres suivants décrivent la méthodologie appliquée pour mener à bien ce projet. Nous présenterons tout d'abord au chapitre 3 un certain nombre de choix techniques et pratiques en lien avec

notre méthodologie. Le chapitre 4 décrira en détail les corpus utilisés pour notre étude, et le chapitre 5 présentera de manière exhaustive le processus de génération des annotations extraites des corpus.

Les résultats de nos analyses seront ensuite présentés, tout d'abord sous forme d'un portrait global des SFP dans notre corpus et de leur production en rapport avec les différents facteurs considérés (chapitre 6), puis par le biais de l'ajustement de modèles statistiques (chapitre 7). Nous analyserons ensuite de manière approfondie le comportement de huit des particules les plus présentes dans le discours des enfants au chapitre 8.

Enfin, le chapitre 9 présentera, dans une première partie, l'interprétation de ces résultats en lien avec chacun des facteurs d'influence évoqués dans le chapitre 1. Nous effectuerons ensuite dans une partie de discussion un retour critique sur les hypothèses de recherche que nous avons formulées au chapitre 2. Nous terminerons par une analyse des différentes limitations qui se sont manifestées au cours de la réalisation de ce projet.

CHAPITRE 1

Présentation du cadre de la recherche

Ce premier chapitre offre une description détaillée du cadre de notre recherche ainsi que de la littérature ayant trait à nos objectifs. Nous évoquerons tout d'abord le rôle des particules de fin de phrase dans le discours et dans l'acquisition des notions pragmatiques, et décrirons leur usage et leur identité dans le cadre spécifique du cantonais. Nous présenterons ensuite l'état actuel des connaissances sur le processus d'acquisition des SFP par les enfants, en tissant des liens avec d'autres langues où ces particules sont également employées. Par la suite, nous proposerons une liste de différents facteurs susceptibles de jouer un rôle dans ce processus, en nous basant sur les études antérieures et en justifiant leur pertinence dans le cadre spécifique de l'acquisition des SFP. Nous commencerons par les facteurs individuels regroupant des caractéristiques propres à l'enfant, puis décrirons les facteurs linguistiques associés à l'environnement discursif. Enfin, nous présenterons les facteurs extrinsèques liés en particulier à l'identité de l'interlocuteur, et proposerons une réflexion préliminaire à propos de l'impact attendu du discours adressé aux enfants sur l'acquisition dans le cadre d'une analyse de corpus.

1.1 Les SFP comme outil pragmatique

Les SFP entrent dans la catégorie des marqueurs pragmatiques, dont les fonctions incluent, entre autres, la transmission de la force illocutoire, le marquage des tours de parole, l'évidentialité ou encore l'humeur du locuteur (Foolen, 1997). Ces marqueurs sont cruciaux pour étendre le sens d'un énoncé au-delà de la valeur sémantique des mots qui le composent, et servent à orienter l'interlocuteur dans le processus d'interprétation (Aijmer, 2014). Si les linguistes s'accordent à reconnaître des éléments ayant cette fonctionnalité dans de très nombreuses langues, la nature de ceux-ci est toutefois difficile à définir précisément, puisqu'ils prennent souvent la forme de termes ou locutions ayant également d'autres fonctions dans le discours (Foolen, 1997). Leur rôle syntaxique ne s'inscrit par ailleurs pas toujours précisément dans la structure des phrases qu'elles complètent, ce qui est susceptible d'avoir un effet sur la stabilité des usages qui en sont faits, et sur la saillance qui leur est attribuée. Dans certaines langues comme l'allemand (particules modales) ou le finnois (particules enclitiques), elles adoptent un rôle structurel dans la formation des phrases (Luke, 1990 ; Tulling, 2017).

Les SFP se retrouvent quant à elles dans plusieurs langues d'Asie de l'Est et du Sud-Est, comme le mandarin, le japonais, le vietnamien ou encore le tagalog (Nomoto et McCready, 2023). Elles se distinguent

des autres types de particules pragmatiques par leur ancrage dans la syntaxe d'un énoncé, où elles apparaissent toujours en position finale (voir §1.4.1 ci-dessous pour des précisions sur la notion d'énoncé). Elles sont généralement décrites comme des morphèmes liés (ne pouvant constituer un énoncé de manière autonome), dépourvus de contenu véric onditionnel, et servent à exprimer la modalité, l'intention, l'émotion, l'évidentialité et autres nuances pragmatiques.

Elles forment par ailleurs une classe syntaxique propre et constituent un ensemble de marqueurs pragmatiques grammaticalisés apparaissant fréquemment dans le discours; cette particularité les rend particulièrement saillantes, et donc très accessibles aux locuteurs débutants, qui en les identifiant dans le discours qu'ils perçoivent peuvent les associer aux contextes dans lesquels elles sont utilisées. En fournissant ainsi un balisage dans le processus d'interprétation, elles jouent un rôle important dans le développement pragmatique des enfants, en particulier dans l'établissement du positionnement épistémique par rapport au discours (Chor, 2018). Leur émergence dans le langage des enfants est donc en partie corrélée avec l'évolution de leurs besoins pragmatiques, à la fois comme outil et comme incitation à exprimer leur rapport à une situation d'interaction par le biais d'éléments discursifs. Cette forte connexion psycholinguistique entre le développement des compétences pragmatiques et les SFP a par ailleurs été observée en japonais par Matsui et Miura (2009), qui ont constaté que les notions épistémiques étaient comprises plus tôt par les enfants lorsqu'elles étaient véhiculées par des SFP (en l'occurrence *yo*, indiquant la certitude, et *kana*, indiquant l'incertitude) que par des verbes.

1.2 Les SFP en cantonais

Le cantonais est la langue maternelle de plus de 75 millions de locuteurs natifs à travers le monde, principalement en Chine dans la ville de Hong Kong et la province du Guangdong (The Editors of Encyclopaedia Britannica, 2024). Réputée pour être une langue à tradition majoritairement orale, elle présente des différences significatives entre ses formes parlée et écrite. L'une de ces différences est l'utilisation des SFP, très importante à l'oral et pratiquement nulle à l'écrit (Matthews et Yip, 2011).

Les SFP constituent un élément incontournable du cantonais parlé, étant présentes dans plus de 60% des énoncés produits par les adultes (jusqu'à 85% selon Gibbons, 1980). Elles peuvent être placées à la fin des énoncés, mais aussi de structures linguistiques de plus bas niveau comme des propositions ou syntagmes (voir §1.4.1 ci-après), et servent à en marquer la forme (interrogative ou déclarative) ou à en modifier la dimension pragmatique. Le cantonais possède un inventaire de SFP particulièrement important : on en

compte de 30 à plus d'une centaine, selon les auteurs et les critères de distinction; comparativement, le mandarin et le japonais, où les SFP sont clairement identifiées et également très présentes, n'en possèdent qu'une dizaine. Le cantonais est d'ailleurs considéré par certains comme la langue en ayant l'inventaire le plus riche : « Cantonese utterance particles far outnumber their Mandarin counterparts, or those of any other language that I know of. » (Luke, 1990, p. 1).

La grande variabilité du catalogue des SFP du cantonais est due à différents facteurs : d'une part, il n'en existe pas de liste officielle, et les différents linguistes effectuent un choix personnel des éléments qu'ils intègrent dans leurs descriptions (ainsi, Matthews et Yip (2011) mentionnent des particules comme ho2 ou lei4, absentes de l'étude de Kwok (1984) – voir §1.3 ci-après pour une explication du système de romanisation utilisé pour transcrire le cantonais); d'autre part, certaines SFP peuvent être considérées comme des combinaisons de particules atomiques (par exemple, la particule gaa3 est fréquemment analysée comme la fusion de *ge3* et *aa3*), et ne pas être rapportées comme particules autonomes dans certains travaux; enfin, certaines séquences de SFP peuvent également se voir attribuer par certains linguistes le statut de particules autonomes : Kwok (1984) décrit la séquence laa3wo3 comme une particule dissyllabique ayant une valeur différente de la succession de laa3 et wo3, alors que Matthews et Yip (2011) ne considèrent que les deux particules laa3 et wo3 de manière indépendante.

Leur omniprésence s'explique en particulier par l'importante variété de fonctions sémantiques et pragmatiques (*i.a.* marquage épistémique, force illocutoire, interrogation, modalité) qu'elles permettent d'exprimer (Kwok, 1984 ; Matthews et Yip, 2011 ; Yau, 1965). Certaines recherches suggèrent que leur usage est en partie associé au caractère tonal du cantonais, qui a un impact sur l'utilisation de l'intonation comme marqueur pragmatique; le rôle des SFP a en conséquence largement été comparé à celui joué par l'intonation dans les langues non tonales (Gibbons, 1980 ; Wakefield, 2011). La diversité de leurs utilisations est illustrée en (1) (exemples adaptés de Ko, 2000).

Jusqu'à quatre SFP peuvent être juxtaposées à la fin d'un énoncé pour en combiner les fonctions; de telles associations sont toutefois soumises à des règles de composition qui déterminent leur acceptabilité

(Matthews et Yip, 2011, p. 395)[1]. Dans l'exemple (2), les quatre SFP suivent un ordre canonique et ne pourraient être positionnées différemment. La notation jyutping utilisée pour la transcription est décrite à la section suivante.

(1) a.

佢	肯	做.	
koei5	hang2	zou6	
il	vouloir	faire	
Il veut le faire.			(neutre)

b.

佢	肯	做	**嘞**.	
koei5	hang2	zou6	**laak3**	
il	vouloir	faire	**SFP**	
Il veut le faire (dans tous les cas).				(certitude)

c.

佢	肯	做	**嗎**?	
koei5	hang2	zou6	**maa3**	
il	vouloir	faire	**SFP**	
Est-ce qu'il veut le faire?				(question oui/non)

d.

佢	肯	做	**喎**.	
koei5	hang2	zou6	**wo4**	
il	vouloir	faire	**SFP**	
Je suis surpris qu'il veuille le faire.				(mirativité)

(2)

佢	擺咗	第	一	名	**添**	**嘅**	**喇**	**喎**.
koei5	lo2zo2	dai6	jat1	ming4	**tim1**	**ge3**	**laa3**	**wo3**
il	prendre-PFV	numéro	un	place	**SFP**	**SFP**	**SFP**	**SFP**

Et il a aussi obtenu la première place, tu sais.

Bien que certaines d'entre elles ne puissent être associées qu'à certains types d'énoncés (p. ex. *lo4* : déclaratif, *aa4* : interrogatif; voir p. ex. Kwok (1984) pour une description exhaustive des contextes d'utilisation), la contribution des SFP aux énoncés où elles apparaissent ne peut en règle générale pas être résumée à leur association avec une forme locutoire. Comme se sont efforcées de le démontrer plusieurs analyses approfondies (p. ex. Leung, 2008 ; Luke, 1990), les valeurs pragmatiques portées par les SFP sont discursivement et culturellement complexes et transcendent l'acte de langage en tant que tel.

[1] Fung (2000) suggère que des séquences plus longues seraient également possibles, en particulier grâce à la fusion de SFP. Ainsi, la séquence tim1+ge3+aa3+ze1+aa3+aa1maa3+aa5 serait prononcée tim1+gaa3+zaa3+aa1+maa5; certains auteurs considèrent gaa3 ou zaa3 comme étant des SFP à part entière. Une telle séquence (dans sa forme compacte comme dans sa forme étendue) ne respecte toutefois pas les contraintes de positionnement mentionnées par Matthews et Yip (2011).

1.3 Petit aparté : la notation jyutping

Le cantonais est traditionnellement écrit à l'aide de caractères chinois. Bien que la majorité de ces caractères soient partagés avec le mandarin, le cantonais possède également son inventaire propre de caractères, qui ne sont pas utilisés en chinois écrit standard.

À des fins d'accessibilité, le cantonais peut être romanisé. Il existe pour cela différents systèmes, dont les plus courants sont le système de Yale et, plus récemment, le jyutping, que nous utiliserons tout au long de ce mémoire.

La notation jyutping a été proposée en 1993 par la *Linguistic Society of Hong Kong* et a fait l'objet d'une publication (Tang *et al.*, 2002), elle est également décrite succinctement sur le site internet de la LSHK https://lshk.org/jyutping-scheme/. Dans ce système, chaque syllabe (qui correspond de manière générale à un caractère chinois) est transcrite au moyen d'une partie segmentale (composée d'un noyau obligatoire et d'une attaque et d'une coda optionnelles), à laquelle est associée un chiffre (1–6) indiquant le contour tonal associé à la syllabe[2]. Ainsi, le caractère 係, qui peut signifier « être », est romanisé *hai6*, qui indique que le ton bas (6) est associé à la partie segmentale hai (`/hei/`). Rappelons que le cantonais est une langue dite tonale, c'est-à-dire que les tons y ont une valeur lexicale et font partie intégrante de l'identité d'une syllabe ou d'un mot.

1.4 Aspects linguistiques liés aux SFP

1.4.1 Notions de phrase et d'énoncé

Dans le cadre de l'étude des SFP, il est important de noter que la notion de « phrase » mentionnée dans leur nom (*sentence-final particle*) doit être manipulée avec précaution. Luke (1990, p. 7–10) explique d'une part que les composants linguistiques susceptibles d'être ponctués par des SFP peuvent être des structures plus petites que des phrases grammaticales complètes (comme des propositions, des syntagmes ou même des mots seuls), et d'autre part qu'il n'est pas pertinent de limiter les possibilités d'expansion de tels éléments à de simples phrases, puisque c'est leur dimension pragmatique que la SFP vient moduler. En conséquence, il recommande l'utilisation du terme « énoncé » (*utterance*) pour se référer au contenu correspondant à la portée d'une SFP et le terme *utterance particles* pour celles-ci. Nous conserverons dans

[2] Les indices des tons correspondent à haut (1), moyen montant (2), moyen (3), bas descendant (4), bas montant (5) et bas (6).

ce mémoire l'acronyme SFP pour sa fréquence plus élevée dans la littérature, mais parlerons d'énoncés, puis de segments (voir §3.2.6), pour nous référer aux structures linguistiques englobantes.

1.4.2 Association à une forme linguistique

De manière générale, des SFP sont susceptibles d'être associées à des formes déclaratives ou interrogatives. Cependant, certaines caractéristiques d'un énoncé contraignent l'utilisation des SFP qui peuvent y être utilisées.

1.4.2.1 SFP ambiguës

La majorité des SFP peuvent être utilisées dans des énoncés déclaratifs ou interrogatifs; leur valeur sémantico-pragmatique peut alors varier selon le cas (voir Kwok, 1984 pour une description exhaustive). Elles n'ont cependant pas d'impact sur la forme elle-même, qui est régie par d'autres éléments de la phrase. Ainsi, un énoncé non marqué est généralement interprété comme déclaratif, et peut être ponctué de SFP qui apportent alors le sens ou la nuance associés à ce contexte.

Certaines structures linguistiques marquent formellement un énoncé comme interrogatif. Il s'agit des pronoms interrogatifs comme 點樣|*dim2joeng2* (« comment ? ») ou 幾時|*gei2si4* (« quand ? »), regroupés par la suite sous l'appellation « mots QU », et de structures réduplicatives articulées autour de la particule négative 唔|*m4* produisant des interrogations polaires : le marqueur 係唔係|*hai6m4hai6* (ou sa forme abrégée 係咪|*hai6mai6*) assimilable à une *tag question* (ci-après HAI_M_HAI), et les constructions A_M_A, dans lesquelles la particule négative 唔|m4 est insérée entre une réduplication (préposée) de la première syllabe d'un verbe ou adjectif et la forme complète du même terme, avec en particulier la locution 有冇|*jau5mou4* (forme compacte de la locution « avoir–non avoir ») (voir Yin, 1996 pour une description exhaustive des types de phrases interrogatives). La majorité des SFP ambiguës peuvent accompagner ces constructions pour moduler la dimension pragmatique de la question résultante. Une description exhaustive de ces structures est donnée en §3.2.4.

Notons que certains énoncés non marqués au niveau lexico-syntaxique peuvent également être utilisés comme des questions par le biais de l'intonation. De tels énoncés ne peuvent cependant être ponctués par des SFP (Wong et Ingram, 2003).

1.4.2.2 SFP interrogatives

Certaines SFP ont quant à elles un rôle strictement interrogatif : leur adjonction à la fin d'une phrase déclarative a pour fonction de transformer celle-ci en une phrase interrogative. Il s'agit des SFP *aa4*, *aa5*, *gaa4*, *laa4*, *me1* et *zaa4* (Kwok, 1984 ; Sybesma et Li, 2007). Ces SFP sont rarement utilisées conjointement avec d'autres SFP.

Notons que deux SFP, *ne1* et *ge2*, ont un statut variable : elles peuvent être utilisées comme des particules interrogatives autonomes, auquel cas leur simple ajout permet de transformer une phrase déclarative en phrase interrogative; elles peuvent également ponctuer des phrases déclaratives ou accompagner des structures syntaxiques interrogatives, auquel cas elles n'ont pas d'impact sur la forme de la phrase. Leur signification varie alors également en fonction du contexte.

1.5 Acquisition des SFP

L'acquisition des SFP par les enfants a déjà été étudiée dans différentes langues, bien que relativement peu d'écrits à ce sujet aient été publiés. Cette section décrit les travaux à ce sujet en cantonais, mais aussi en mandarin et en japonais, où les SFP sont également très présentes.

1.5.1 Notions préliminaires

1.5.1.1 Âge et développement langagier

L'étude de l'acquisition d'une langue, en particulier la langue maternelle, cherche à décrire et expliquer l'émergence de structures et comportements langagiers chez les enfants locuteurs de cette langue; elle s'intéresse donc à des tranches d'âges pendant lesquelles le langage est en cours de construction, depuis la production des premiers sons jusqu'à la combinaison de structures complexes. S'il est indéniable que ce développement n'est pas complètement aléatoire et fait montre d'une certaine part de régularité parmi les enfants, il est cependant clair que de nombreux facteurs, tant individuels qu'environnementaux, jouent un rôle dans ce processus et influent le rythme et la façon dont ils s'approprient le langage (Huttenlocher *et al.*, 2010 ; Kidd et Donnelly, 2020), et que l'âge ne constitue qu'un seul des nombreux éléments à considérer pour tenter d'expliquer les phénomènes liés à l'acquisition.

La notion de longueur moyenne des énoncés (*mean length of utterance*, ci-après MLU) a été proposée comme indicateur du niveau de développement langagier d'un enfant (Brown, 1973 ; Rice *et al.*, 2010). Si

cette valeur est de manière générale fortement corrélée avec l'âge, elle a toutefois été reconnue comme étant un meilleur prédicteur que celui-ci pour l'explication de différents aspects de l'acquisition d'un langage, en particulier le développement du lexique et de la morphosyntaxe (DeThorne *et al.*, 2005 ; Miller et Chapman, 1981). Cette mesure jouera un rôle clef dans le cadre de notre étude, et fait l'objet d'une discussion complémentaire à la section 3.3.

1.5.1.2 Restrictions à la notion d'acquisition

Dans le cadre de l'étude de l'acquisition des SFP, les auteurs se basent généralement sur le fait qu'une particule spécifique est activement produite par l'enfant pour la considérer acquise (Ko, 2000 ; Lee *et al.*, 1996 ; Lim, 2018 ; Zhang *et al.*, 2019), suivant ainsi le principe de productivité proposé par Bloom (1991). Il est toutefois nécessaire de garder à l'esprit qu'un terme peut également être compris (et donc correspondre à un certain degré d'acquisition) avant que l'enfant ne soit en mesure de le produire activement. En contrepartie, une évaluation du niveau global d'acquisition d'un élément du langage, lorsque celui-ci est activement produit, doit également prendre en considération la pertinence des utilisations, ce qui requiert une analyse contextuelle critique de chaque occurrence.

Cette discrépance entre acquisition en compréhension et acquisition en production est particulièrement notable avec des concepts concrets comme les couleurs, où il est possible d'évaluer la compréhension en observant la réaction de l'enfant à un stimulus oral; on constate alors que les enfants sont en mesure de reconnaître certains noms de couleurs lorsque ceux-ci sont utilisés pour désigner des objets, mais qu'ils sont en revanche incapables d'utiliser ces mêmes noms pour décrire les objets (voir, par exemple, Andrick et Tager-Flusberg, 1986).

L'évaluation de la compréhension est bien entendu plus complexe pour des éléments linguistiques se référant à des aspects abstraits du discours, comme les nuances pragmatiques exprimées par certaines SFP. Il serait néanmoins envisageable d'estimer la compréhension de certaines particules comme *aa4* (SFP interrogative utilisée pour formuler une question polaire à partir d'un énoncé déclaratif) en observant la réponse de l'enfant à une telle interrogation, afin de s'assurer qu'elle correspond bien au motif attendu (oui ou non); cette mesure peut alors être mise en relation avec la compétence de l'enfant à effectivement produire la particule *aa4* pour formuler des questions, ce qui permet d'obtenir deux mesures distinctes d'acquisition, soit l'acquisition en compréhension et l'acquisition en production. Un protocole semblable

a été mis en place par Lee et Law (2000) pour tester la compréhension de certaines particules épistémiques.

Dans le cadre de cette étude, nous avons fait le choix de limiter notre analyse de l'acquisition à l'observation de la production active des SFP. Ainsi, notre utilisation du terme « acquisition » se réfère uniquement à l'acquisition en production, sans prendre en compte la pertinence des occurrences observées vis-à-vis du contexte d'utilisation.

1.5.2 Acquisition des SFP en cantonais

L'acquisition du cantonais a été étudiée sous différents aspects, comme la phonologie (So et Dodd, 1995 ; Wang et Wong, 2024), la syntaxe (Leung et Li, 2015 ; Tsang et Stokes, 2001 ; Tse *et al.*, 2002) ou le lexique (Klee *et al.*, 2004 ; Stokes et Fletcher, 2000). L'émergence des SFP n'a, quant à elle, fait l'objet que de peu d'études dans les dernières décennies. À la suite de la constitution du corpus CANCORP, Lee *et al.* (1996) se sont penchés sur la dimension syntaxique de plusieurs aspects du langage des enfants pendant le processus d'acquisition, parmi lesquels l'utilisation des SFP par deux enfants avant l'âge de 2 ans. Ko (2000) s'est appuyé sur la production des huit enfants observés dans ce même corpus pour effectuer une analyse détaillée des relations qui unissent les formes et fonctions des SFP produites. Lim (2018) a effectué une étude similaire en se basant sur les données transversales du corpus HKU-70 (Fletcher *et al.*, 2000). D'autres travaux se sont focalisés sur certains groupes de SFP en particulier, comme par exemple les particules épistémiques (Tang, 2018) ou évidentielles (Lee et Law, 2000).

La période d'émergence des SFP mentionnée dans la littérature est généralement celle rapportée par Lee *et al.* (1996), qui déclarent qu'elles sont clairement identifiées dans le discours (des deux seuls enfants considérés dans leur étude) à 1;08 et 1;09. Comme le rapporte Ko (2000) dans son analyse des huit enfants observés lors de la constitution du corpus, l'utilisation des SFP est également identifiable dans le discours du plus jeune des enfants âgé de moins d'un an et demi (1;05.22).

Leur fréquence et leur diversité pendant les mois qui suivent leur apparition dans le discours restent cependant extrêmement limitées : Ko (2000) rapporte que seules deux particules, *aa3* et *aa1*, peuvent être considérées comme acquises lorsque la MLU (voir §1.5.1.1) atteint 2.2. Lim (2018), quant à elle, évoque que seuls 50% des enfants de deux ans sont capables de produire spontanément deux particules distinctes. Si le rythme d'acquisition et la fréquence d'utilisation augmentent par la suite, la compétence

dont font preuve les enfants de cinq ans reste encore largement inférieure à celle des adultes, puisque seules une dizaine de SFP peuvent généralement être considérées comme acquises à cet âge. Toutefois, la fréquence avec laquelle les enfants de cinq ans utilisent les SFP se rapproche de celle retrouvée chez les adultes, atteignant une fréquence d'utilisation de 62% des énoncés (Lim, 2018).

Bien que l'ordre d'acquisition varie selon les individus, on retrouve certaines tendances générales : Ko (2000) et Lim (2018) constatent dans deux corpus différents que *aa3*, *aa1*, *gaa3*, *laa3*, *lo1* et *ne1* sont généralement parmi les premières à être produites, ce qui permet d'envisager que cette régularité est le résultat d'une combinaison de facteurs propres à ces particules ou aux contextes dans lesquels elles sont utilisées, en particulier par les adultes. Cependant, les variations observées entre les enfants impliquent que certains facteurs individuels et contextuels doivent également être pris en compte.

Ko (2000) met par ailleurs en avant la relation qui existe entre l'émergence des formes de surface produites par les enfants (p. ex. « aa3 », « lo1 ») et la fonction illocutoire associée à chacune des occurrences. Les fonctions sémantico-pragmatiques de chaque SFP ont été décrites en détail et selon différents angles par Yau (1965), Gibbons (1980) ou encore Matthews et Yip (2011), et couvrent différents aspects comme l'association aux actes de langage, la modalité, l'humeur ou l'évidentialité; Matthews et Yip avertissent toutefois qu'il est très difficile de fournir une description globalement applicable de l'apport de chaque SFP, et que cette valeur dépend largement du contexte d'utilisation. Les observations de Ko sont en accord avec les conclusions de Law et Lee (1998), selon lesquelles une SFP pouvant servir à exprimer différentes fonctions pragmatiques est, dans un premier temps, utilisée avec une seule de ces fonctions, pour ensuite se diversifier. Lim (2018) constate pour sa part une concomitance des usages déclaratifs et interrogatifs de aa3, suggérant que cette acquisition incrémentale des fonctions n'est pas systématique. L'analyse des associations entre forme et fonction constitue un défi majeur dans l'analyse des conversations et de l'utilisation des SFP, puisqu'elle requiert d'être en mesure d'évaluer, pour chaque occurrence d'une SFP, la fonction illocutoire qui lui est associée (au-delà de la simple distinction entre usage déclaratif et interrogatif), en prenant en compte l'imprécision de la transcription ou les erreurs commises par les enfants. Pour des raisons de temps et de moyens, nous avons fait le choix de ne pas considérer cet aspect dans nos analyses.

1.5.3 Acquisition des SFP dans d'autres langues

Comme il a été mentionné, le cantonais se distingue des autres langues utilisant des SFP par le fait qu'elles y sont beaucoup plus nombreuses, et permettent ainsi l'expression de nuances sémantiques et pragmatiques beaucoup plus fines. En termes d'acquisition, cela suggère également une plus grande variabilité dans le rythme et l'ordre d'émergence dans le langage des enfants, du fait d'une part de l'exposition nécessairement plus dispersée à chacune (ou au moins à la majorité) des SFP dans le langage environnant, et d'autre part en raison de la subtilité des différences existant entre certaines SFP. Il est toutefois intéressant d'évoquer les résultats d'études menées sur l'acquisition de ces particules dans des langues où elles jouent un rôle comparable, comme le mandarin et le japonais.

Notons que tout comme en cantonais, les publications sur l'acquisition des SFP dans ces langues restent rares et n'offrent donc qu'une compréhension limitée des processus en jeu.

1.5.3.1 Mandarin

L'étude de Zhang *et al.* (2019) constitue, à notre connaissance, la seule publication disponible en anglais focalisée sur l'acquisition des SFP en mandarin[3], même si certains travaux plus généraux y consacrent également une section (Paul et Yan, 2022 ; Yang, 2022). Elle est axée sur le suivi longitudinal de quatre enfants âgés de 1 à 4 ans, et étudie l'émergence des SFP dans leur discours en fonction de leur moment d'apparition, de la MLU de l'enfant, et des fréquences d'utilisation par les enfants et les parents.

Le mandarin compte une dizaine de SFP, réparties en deux catégories : d'une part les particules modales typiques (*typical modal particles*), fréquemment utilisées, et elles-mêmes classifiées en particules tonales (*tone particles*), qui ne jouent pas de rôle dans la structure syntaxique de la phrase et informent sur l'attitude du locuteur, et particules fonctionnelles (*functional particles*), qui ont une fonction syntaxique; et les particules modales générales (*general modal particles*), plus rares. L'étude a permis de constater que les particules tonales apparaissaient systématiquement (chez les quatre enfants) avant les autres, et que les particules fonctionnelles apparaissaient globalement avant les particules modales générales. Les raisons de ce motif d'émergence n'ont quant à elles pas pu être associées avec des éléments extérieurs spécifiques. Les auteurs constatent une corrélation entre la fréquence d'utilisation dans le CDS (*child-*

[3] Plusieurs travaux universitaires au niveau maîtrise se sont également intéressés à cette question, mais ces travaux n'ont pas été publiés et ne sont pas disponibles en ligne.

directed speech, discours adressé à l'enfant) et l'ordre d'acquisition par les enfants, mais lui attribuent un impact faible. La conclusion générale de l'étude est que les SFP du mandarin sont organisées en une hiérarchie abstraite qui contraint leur acquisition et qui fait partie du savoir linguistique intrinsèque des enfants, et que leur émergence dans le discours des enfants ne correspond pas à un processus d'apprentissage.

1.5.3.2 Japonais

Le japonais compte également une dizaine de SFP dont les fonctions sémantiques et pragmatiques (i. a. expression d'évidentialité ou de doute, recherche de confirmation) sont très similaires à certaines SFP du cantonais (Shirai *et al.*, 2000). Les SFP apparaissent très tôt dans le discours des enfants, et le processus d'acquisition est considéré comme complété lorsque leur MLU atteint environ 3.0 (Fujimoto, 2008 ; Shirai *et al.*, 2000).

Les facteurs d'influence avancés pour décrire l'acquisition des SFP sont variés : Shirai *et al.* (2000) concluent de leur étude du processus d'acquisition chez quatre enfants (observés pendant environ un an et demi) que l'ordre d'acquisition est lié au développement cognitif des enfants, ce qu'ils associent aux compétences relatives à la Théorie de l'esprit. Fujimoto (2008), pour sa part, ayant étudié l'émergence de toutes les particules en japonais chez trois enfants âgés d'un an et trois mois à trois ans, conclut que leur ordre d'apparition est lié au niveau de développement langagier tel qu'évalué par la mesure de MLU. Toutefois, elle constate que cet ordre ne correspond pas aux fréquences d'utilisation des SFP dans le discours adressé aux enfants lors des sessions d'enregistrement, et conclut que l'émergence des particules dans leur discours est dictée par une « machinerie grammaticale innée » (*innate grammatical machinery*).

1.6 Facteurs d'influence

Les recherches effectuées sur l'acquisition des SFP, quelle que soit la langue, permettent de mettre en avant certains facteurs en lien avec ce processus. Par ailleurs, l'influence de certains facteurs sur l'acquisition d'autres aspects linguistiques a également été constatée, et leur pertinence dans le cadre des SFP est décrite ci-dessous. Les facteurs que nous considérons sont regroupés en trois catégories : facteurs individuels, facteurs linguistiques et facteurs extrinsèques.

1.6.1 Postulat innéiste

Dans le contexte de l'acquisition, plusieurs auteurs accordent un rôle important, voire central, à des dispositions linguistiques innées dont sont dotés les enfants avant même leurs premières expositions au langage en général, et à leur langue maternelle en particulier. Dans leur étude du processus d'acquisition des SFP en mandarin, Zhang *et al.* (2019, p. 16-17) relativisent radicalement l'effet imputable au CDS :

> The abstract classification structure and the dominant order of sentence-final particles use are part of the children's inherent and intrinsic linguistic knowledge and do not need to be learned. […] The role of adult discourse input frequency cannot exceed this principle of acquisition. It can only have a partial or individual corresponding influence on children's language acquisition.

Cette affirmation permet *a minima* de largement simplifier l'analyse des facteurs extérieurs susceptibles d'avoir un impact sur l'acquisition. On retrouve cette considération dans l'analyse de Lee *et al.* (1996), à l'origine du corpus CANCORP utilisé dans notre étude, qui débute par ces mots : « We take is as axiomatic that children come to the language learning task with a rich set of linguistic universals. »

Il nous semble cependant difficile de prendre pour acquise l'existence innée de ces connaissances chez les enfants, quand de nombreux éléments n'ont à ce jour pas été étudiés de manière suffisamment approfondie pour concevoir l'étendue du rôle qu'ils peuvent jouer dans le processus d'acquisition en général. Nous nous efforcerons donc d'analyser de manière objective l'impact que ces autres facteurs, internes et externes à l'enfant, sont susceptibles d'avoir, sans présumer de l'ampleur de leur contribution.

1.6.2 Facteurs individuels

1.6.2.1 Niveau de développement langagier et âge

Comme l'ont observé Ko (2000) et Lim (2018), l'acquisition des SFP semble suivre un schéma chronologique, selon lequel certaines particules sont généralement acquises par les enfants avant certaines autres. Ces deux études diffèrent cependant dans les marqueurs chronologiques qu'elles utilisent : Lim utilise l'âge biologique des enfants pour caractériser l'acquisition d'une SFP, alors que Ko se réfère à des « stades » (*stages* en anglais) définis comme des intervalles de MLU.

La MLU a été largement adoptée comme indicateur de développement langagier chez les enfants, en particulier du fait de sa corrélation avec la complexité syntaxique des énoncés (Brown, 1973 ; Rice *et al.*, 2010). Elle semble donc incontournable dans l'étude de l'acquisition des SFP, dont l'interaction avec la

syntaxe de la phrase a été expliquée plus haut (§1.4), et est d'ailleurs utilisée dans la majorité des études sur l'acquisition des SFP dans différentes langues (Fujimoto, 2008 ; Ko, 2000 ; Shirai *et al.*, 2000 ; Zhang *et al.*, 2019) et sur l'acquisition de différents aspects du cantonais (p. ex. Leung et Li, 2015 ; To *et al.*, 2010 ; Tse *et al.*, 2002).

L'âge biologique reste toutefois également largement utilisé comme critère de différenciation (Tang, 2018 ; Tse *et al.*, 2007). Du fait de sa forte corrélation avec la MLU, il constitue un prédicteur valable pour certains processus d'acquisition. De plus, la corrélation entre l'âge et l'utilisation de verbes décrivant des états mentaux a été confirmée pour des enfants d'origines linguistiques et culturelles différentes, et en particulier pour les enfants cantonophones (Tardif et Wellman, 2000), chez lesquels on constate une chronologie stable entre l'utilisation de verbes en lien avec les désirs, et ceux en lien avec les croyances. Ce développement de la perception des états mentaux est associé à la constitution de la Théorie de l'esprit, dont Meltzoff (1999) prévoit l'émergence aux alentours de 18 mois. Law (2004) suggère que l'acquisition tardive de certaines SFP, en particulier en lien avec l'évidentialité, pourrait être due à une immaturité de la Théorie de l'esprit. Lee et Law (2000) ont d'ailleurs testé la compréhension de certaines particules épistémiques (wo4, indiquant la surprise, et aa1maa3, indiquant la justification) et constaté que des enfants de 6 ans n'étaient en majorité pas capables de constamment saisir le sens lié à leur utilisation. Tang (2018, p. 29) décrit pour sa part le lien qui unit l'acquisition de la modalité épistémique, et par conséquent des SFP qui servent à l'exprimer, à la Théorie de l'esprit. Elle associe l'utilisation tardive de certaines particules épistémiques (p. ex. gwaa3, wo5, gaak3) au fait qu'elle implique la capacité de percevoir les croyances d'autres personnes, qui se développe à partir de trois ans et plus.

1.6.2.2 Genre de l'enfant

Plusieurs études ont constaté une influence du genre sur différents aspects du cantonais, parmi lesquels l'utilisation des SFP (Chan, 2000, 2002 ; Winterstein *et al.*, en préparation). La forte asymétrie observée entre les discours masculin et féminin confirme l'existence d'importantes interactions; ces différences sont de plus particulièrement marquées pour plusieurs des particules acquises généralement assez tôt par les enfants, comme *aa3*, *lo1* ou *gaa3*. Tse *et al.* (2002) ont pour leur part montré que le développement syntaxique du discours des enfants entre 3 et 5 ans est globalement plus rapide chez les filles que chez les garçons, et Li *et al.* (2013) ont noté un effet significatif du genre sur la longueur des questions posées par les enfants entre 3 et 5 ans, ainsi que sur le type de ces questions. Notons toutefois que Klee *et al.* (2004), en voulant reproduire les résultats de Tse *et al.*, n'ont trouvé aucun effet significatif du genre sur la MLU

ou la diversité lexicale dans une étude transversale effectuée sur des enfants cantonophones âgés de 2 à 5 ans. Santos *et al.* (2015) rapportent quant à eux qu'une différence entre garçons et filles n'est observable que jusqu'à 3 ou 4 ans. Considérant que l'acquisition des SFP peut être observée dès un an et demi, il est donc pertinent de s'interroger sur le rôle que le genre des enfants joue dans ce processus.

1.6.3 Facteurs linguistiques

1.6.3.1 Éléments syntaxiques

Les SFP peuvent être divisées en deux catégories : non-interrogatives et interrogatives (§1.4.2). La dimension interrogative d'une SFP ne réside cependant pas dans sa compatibilité avec une phrase interrogative, mais dans sa capacité à transformer, par sa seule présence, une affirmation en question, comme c'est le cas par exemple pour *aa4* et *me1* (p. ex. Leung, 2008). D'autres particules, comme *aa3* ou *ne1*, peuvent être ajoutées à des énoncés qui sont grammaticalement déjà marqués comme questions, par exemple par la présence de marqueurs interrogatifs explicites (voir ci-dessous) et y apporter des nuances pragmatiques, par exemple épistémiques ou modales.

L'acquisition de ces différents types d'interrogations a été étudiée par Wong et Ingram (2003), qui ont pu constater que leur émergence dans le discours des enfants suit un ordre globalement stable : les enfants forment d'abord des questions à l'aide de particules ou par intonation, puis utilisent des mots QU, puis les structures plus complexes A_M_A. Le lien syntaxique qui unit les SFP à l'utilisation de ces constructions implique que ce développement dans la formulation de phrases interrogatives a nécessairement des répercussions sur l'utilisation des SFP dans le discours des enfants. Il convient toutefois de prendre en considération le fait que les SFP ambiguës, susceptibles d'être utilisées dans des phrases déclaratives ou avec des structures interrogatives, peuvent avoir des schémas d'acquisition différents selon le type d'environnement où elles sont utilisées.

Lee *et al.* (1996) analysent les interactions entre les SFP utilisées avant 2 ans et les types de constructions et d'actes de langage produits par les adultes dans les énoncés précédents. Bien que les données décrites ne concernent que deux individus, les auteurs concluent que, dès l'âge de 1;09[4], les enfants sont en

[4] Nous utiliserons dans ce mémoire la notation courante pour indiquer les âges, soit `années;mois(.jours)`. Ainsi, 1;09 indique un âge d'un an et neuf mois.

mesure d'associer les particules aux environnements syntaxiques adaptés, et donc de différencier les SFP ambiguës des SFP strictement interrogatives.

La reconnaissance de ces constructions syntaxiques permet d'évaluer leur rôle dans la fréquence et la régularité d'utilisation de certaines SFP.

1.6.3.2 Phono-prosodie et phonologie

La structure phonologique a déjà été reconnue comme jouant un rôle dans la construction des énoncés (Breiss et Hayes, 2020). En cantonais, le nombre de syllabes influence la forme de certains syntagmes nominaux (Winterstein *et al.*, 2023b). Il semble donc pertinent d'observer l'effet que cette donnée peut avoir sur la production de SFP des enfants, particulièrement dans le cas d'énoncés courts. De même, l'acquisition progressive des tons (So et Dodd, 1995) est susceptible d'avoir un impact sur la capacité des enfants à produire certaines SFP; cet effet ne serait cependant probablement observable qu'au tout début du processus d'acquisition, puisque So et Dodd rapportent que la majorité des enfants ont complètement assimilé le système tonal du cantonais vers 2 ans.

1.6.4 Facteurs extrinsèques

1.6.4.1 Caractéristiques de l'interlocuteur

Comme le remarque Lim (2018), le fait que le corpus HKU-70 ne contienne que des interactions entre expérimentateurs et enfants peut avoir pour effet de modifier le discours de ceux-ci, par un effet de non-familiarité et de distance sociale. Comparativement, les interlocuteurs enregistrés pour le corpus CANCORP sont plus variés et incluent, en plus des expérimentateurs, les parents, grands-parents ou même les frères et sœurs des enfants. Harris (2007) explique que le lien affectif qui existe entre un enfant et ses parents est de nature à créer une impression de « disponibilité intellectuelle » pour l'enfant, formant un lien épistémique avec eux et leur donnant la confiance de poser davantage de questions; cela implique un impact sur la production de SFP, différente dans les phrases déclaratives et interrogatives. On peut d'autre part envisager que l'impact de ce lien affectif et de la confiance qui en découle s'exprime également dans les choix langagiers des enfants. Il est par ailleurs naturel d'étendre l'effet de cette proximité affective à la famille proche. Cette donnée devra en conséquence dans tous les cas être prise en compte.

Nous intègrerons également à l'analyse le genre de l'interlocuteur, bien que l'effet de celui-ci puisse être ambigu : dans le cadre strictement familial (et dans le contexte restreint de nos corpus), il est bien entendu

largement équivalent aux rôles de père et mère, qui constituent en eux-mêmes des statuts susceptibles d'être à l'origine de variations dans le discours que l'enfant adresse à ses parents. Hors du contexte familial (c'est-à-dire dans un cadre expérimental, en présence d'un expérimentateur ou d'une expérimentatrice), il serait concevable d'envisager un effet du genre de l'interlocuteur sur le discours produit par les enfants. Cependant, comme il sera expliqué à la section 5.6, cette information n'est pas systématiquement disponible dans les corpus et ne pourra donc être intégrée que de manière parcellaire.

1.6.4.2 Discours adressé à l'enfant

Le rôle joué par le CDS dans le processus d'acquisition a été souligné à de nombreuses reprises (p. ex. Clark, 2017 ; Saxton, 2009), même si de nombreuses études dans des environnements non occidentaux ont relativisé certaines de ses caractéristiques comme l'attention conjointe ou le ciblage direct de l'enfant (p. ex. Brown et Gaskins, 2014 ; Casillas *et al.*, 2020 ; Gaskins, 2006). Son importance a entre autres été mise en avant dans le rapport que l'enfant entretient vis-à-vis des actes de langage, intimement liés aux SFP : Nikolaus *et al.* (2021) ont avancé que leur compréhension est corrélée à la qualité des signes linguistiques fournis par les parents. L'enfant s'appuie sur ce discours et les variations du contexte dans lequel il est utilisé pour bâtir son expérience des relations interpersonnelles; les paramètres contextuels, comme le rapport entre son propre statut et celui de son interlocuteur, l'humeur associée à la situation, ou encore son état physique interne (faim, fatigue etc.), sont autant de facteurs qu'il est en mesure d'intégrer dans l'évaluation de la pertinence d'une forme de surface dans une situation pragmatique donnée.

Zhang *et al.* (2019) ont confirmé l'influence du CDS sur l'ordre d'acquisition des SFP en mandarin, tout en lui accordant une importance très mineure. Wong et Ingram (2003) ont pu, quant à eux, constater que la forme des questions posées dans le CDS en cantonais a un impact sur l'acquisition des structures interrogatives par les enfants. Il semble donc inévitable que le CDS joue également un rôle sur le rythme d'acquisition des SFP du cantonais.

Il est toutefois important de considérer le CDS sous deux angles différents, que l'on pourrait qualifier de global et ponctuel. L'environnement linguistique global, constitué des interactions que l'enfant a de manière régulière avec ses proches ou dans des environnements de sociabilisation (école, garderie), est réputé pour être influent, et peut être décrit par de nombreuses variables comme le niveau d'éducation des parents, le niveau socio-économique de la famille, et les caractéristiques du discours des proches

comme la complexité syntaxique ou la diversité lexicale (Kidd et Donnelly, 2020). Dans ce sens, les habitudes langagières des proches (comme la fréquence des mots qu'ils utilisent) sont reconnus pour influencer le développement de l'enfant. Notons par ailleurs que le CDS dans ce contexte doit également être considéré comme une entité dynamique, du fait de l'adaptation (consciente ou non) des interlocuteurs au niveau de développement langagier de l'enfant (Spinelli *et al.*, 2023). Wang et Wong (2024) ont d'ailleurs constaté en cantonais une évolution des dimensions intonative et tonale du CDS en fonction de l'âge de l'enfant.

L'environnement ponctuel est celui auquel l'enfant est exposé lors d'une conversation, typiquement lors d'une interaction enregistrée dans le cadre d'une expérimentation ou de la constitution d'un corpus. Si cette interaction peut bien entendu être écologique et représentative du type de discours auquel l'enfant est régulièrement soumis (cadre, interlocuteurs et activités habituels), elle risque néanmoins de présenter des différences clefs; l'interaction entre un enfant et un expérimentateur, par exemple, est très peu révélatrice de l'environnement linguistique classique de l'enfant, puisqu'elle l'expose à un discours différent de celui auquel il est habituellement soumis, avec un interlocuteur envers lequel il n'a peu ou pas tissé de liens affectifs – et qui n'a potentiellement pas la même expérience d'interaction avec les enfants que les parents. Dans ce contexte, il reste envisageable de trouver des effets ponctuels du discours de l'adulte sur celui de l'enfant. Par exemple, Lee *et al.* (1996, p. 171) ont noté l'influence de la structure syntaxique des énoncés produits par les interlocuteurs sur les SFP utilisées dans les réponses des enfants. Dans un contexte de conversation, il est clair que les énoncés s'inscrivent dans une séquence construite où chacun dépend du ou des précédents. En tant qu'acte conséquent, un énoncé qui contient une SFP entretient donc un lien pragmatique avec son prédécesseur; l'acte de langage exprimé par celui-ci, ainsi que toute SFP qu'il pourrait contenir le cas échéant, est donc susceptible d'influencer le choix de SFP du locuteur. Il est toutefois difficile d'envisager qu'un lien étroit entre la fréquence d'utilisation des SFP dans le CDS dans le cadre d'interactions ponctuelles avec des personnes étrangères (comme c'est majoritairement le cas dans les deux corpus que nous avons à disposition, le HKU-70 et le CANCORP, décrits au chapitre 4) et l'émergence de ces SFP dans le discours des enfants puisse être constaté : ces interactions constituent probablement des expériences trop éphémères pour qu'un transfert notable ait lieu. Il est donc nécessaire de relativiser l'importance attendue de ce CDS sur le processus d'acquisition.

La dimension CDS doit donc être intégrée à notre analyse non pas comme un tout (description globale de l'utilisation des SFP), mais plutôt comme une donnée segmentée et dynamique : le contexte de chaque

énoncé devra être décrit et intégré à notre modèle de manière exhaustive (en y intégrant, comme indiqué plus haut, les caractéristiques de l'interlocuteur ainsi que celles de l'énoncé précédent) pour capturer au mieux les interactions à l'œuvre.

CHAPITRE 2

Question et hypothèses de recherche

Nous présentons ici notre question de recherche, ainsi que les hypothèses formulées sur la base des éléments présentés dans le chapitre 1.

2.1 Question de recherche

La revue de littérature des chapitres précédents nous permet d'anticiper que de nombreux éléments linguistiques, individuels et extrinsèques sont susceptibles de jouer un rôle dans le processus d'acquisition des SFP par les enfants cantonophones. L'objectif de ce projet de recherche sera donc d'apporter des réponses à la question suivante :

Quelles interactions entre les facteurs individuels, linguistiques et extrinsèques l'analyse des échanges entre enfants et adultes permet-elle de mettre en évidence dans le cadre du processus d'acquisition des particules de fin de phrase entre un an et demi et cinq ans et demi?

Plus spécifiquement, nous porterons notre attention sur le genre, l'âge et le niveau de développement langagier de l'enfant (facteurs individuels), la syntaxe, la phono-prosodie et le type des énoncés contenant des SFP (facteurs linguistiques), et le statut et le genre des interlocuteurs, ainsi que l'utilisation des SFP dans le discours adressé à l'enfant (facteurs extrinsèques).

2.2 Hypothèses de recherche

En réponse à notre question de recherche, les travaux effectués par le passé nous permettent d'émettre les hypothèses de recherche suivantes, dont les effets seront très probablement combinés en un schéma complexe :

- (H1) La MLU est un meilleur prédicteur que l'âge biologique pour l'utilisation des SFP (fréquence et diversité) dans le discours des enfants (§1.5.1.1); toutefois, l'âge biologique pourrait s'avérer jouer un rôle dans l'émergence de certaines SFP à valeur épistémique (gwaa3, wo4, aa1maa3) dans le discours.
- (H2) À âge égal, les filles ont une utilisation des SFP (fréquence, diversité) plus développée que les garçons (§1.6.2.2); l'impact de cette hypothèse doit cependant être modulé par le fait que les filles

ont, à âge égal, une MLU plus élevée que les garçons, ce qui signifie que le genre n'est sans doute pas le facteur principal de cette différence. L'impact du genre sur l'utilisation des SFP individuelles sera observé, mais il est envisageable que la différenciation observée dans les études antérieures ne s'exprime qu'à un âge plus avancé.

- (H3) Le rôle de l'interlocuteur, qui caractérise le degré de familiarité avec l'enfant, a un impact sur la production de questions et donc sur l'utilisation des SFP qui y sont associées. On attend également un effet sur l'utilisation globale des SFP, bien que la nature de celui-ci (inhibiteur ou facilitateur) soit difficile à anticiper (§1.6.4.1). Le genre de l'interlocuteur, quant à lui, sera également observé, mais son impact sur l'utilisation des SFP à cet âge reste hypothétique.
- (H4) Le CDS (dans le cadre de nos corpus) exerce une influence locale et ponctuelle sur la production de SFP; en particulier, les formes des segments produits par les interlocuteurs, ainsi que les structures syntaxiques qu'ils y utilisent, favoriseront l'utilisation de certaines SFP.
- (H5) La fréquence et le choix des SFP utilisées dans les phrases interrogatives évolue en relation avec l'émergence de l'utilisation de structures syntaxiques interrogatives (§1.6.3.1).

Bien que l'impact du CDS sur le CS (*child-speech*, discours produit par l'enfant) soit globalement attesté dans le cadre d'interactions familiales soutenues (§1.6.4.2), il risque de n'être que difficilement observable dans le contexte d'interactions ponctuelles avec des personnes en partie étrangères au foyer, telles qu'elles sont menées dans les corpus que nous prévoyons d'étudier dans notre étude (voir chapitre 4). Ainsi, l'absence de corrélation forte entre l'utilisation des SFP dans le CS et dans le CDS dans ces corpus ne saurait remettre en question l'apport des interactions quotidiennes entre enfants et locuteurs expérimentés dans le processus d'acquisition des SFP.

CHAPITRE 3

Choix méthodologiques

Ce chapitre décrit le processus et l'environnement technique que nous utilisons pour effectuer l'analyse des facteurs influençant l'acquisition des SFP par les enfants cantonophones.

3.1 Utilisation de corpus

Comme nous l'avons expliqué à la section 1.5.1.2, l'étude de l'acquisition d'une construction langagière est souvent associée à l'analyse de sa production par les individus considérés. Ainsi, pour étudier l'acquisition des SFP par les enfants, nous nous limitons à étudier la production de SFP dans le discours oral d'enfants en phase d'acquisition.

La collecte et la préparation de données utilisables pour l'étude de l'oral constituent un défi de taille, puisque ces processus impliquent de recréer des contextes de production aussi écologiques que possible, de stimuler la production d'une quantité suffisamment importante de discours pour atteindre une significativité statistique, de pérenniser les paroles des participants par le biais d'enregistrements et de transcriptions, et de rendre ces données accessibles à des procédés de traitement automatisé. L'analyse du discours de jeunes enfants, dans une langue peu documentée et largement endémique comme le cantonais, rend la constitution d'un corpus encore plus complexe. Dans ce contexte, l'existence de deux corpus de discours de jeunes enfants, collectés pendant les phases précoces de leur acquisition du cantonais comme langue maternelle, a constitué un élément crucial à l'élaboration de cette étude, et à l'étude des processus d'acquisition du cantonais en général : les deux corpus que nous utilisons ici, le CANCORP et le HKU-70 (tous deux présentés en détail au chapitre 4), ont été utilisés pour la majorité des études sur ce sujet depuis leurs publications, respectivement en 1998 et 2000. Un nouveau corpus, le MeHKCC (*Multi-ethnic Hong Kong Cantonese Corpus*)[5], collecté entre 2019 et 2021, constituerait une source de données appréciable pour sa récence et son caractère écologique (interactions mère–enfant), mais il n'a pas encore été rendu disponible pour la recherche universitaire internationale.

Dans le but d'améliorer la représentativité de notre étude, nous avons fait le choix d'assembler les deux corpus (CANCORP et HKU-70), et d'effectuer nos analyses sur le corpus global résultant. Cette combinaison

[5] https://ccds.edu.hku.hk/

a nécessité d'importantes manipulations afin de normaliser les données, de manière à permettre un traitement homogène des transcriptions et conventions propres à chacun des corpus, et la constitution de modèles statistiques globaux. Les étapes des traitements appliqués lors de cette phase de normalisation sont décrites au chapitre 4.

3.2 Choix linguistiques

3.2.1 Description des SFP prises en compte dans notre étude

Selon les auteurs, le nombre de SFP en cantonais varie de 30 (Kwok, 1984) à une centaine (Yau, 1965 ; Luke, 1990). Pour les besoins de notre étude, nous avons constitué une liste de 46 SFP dont le statut fait l'objet d'un consensus, en nous basant sur différents travaux (en particulier Kwok, 1984 ; Matthews et Yip, 2011). Si cette liste peut paraître plus exhaustive que ne le requiert l'étude des particules acquises par les enfants vers l'âge de cinq ans, il nous a paru important de viser la couverture la plus large possible dans la perspective d'obtenir un aperçu complet de l'utilisation des SFP dans le CS et dans le CDS, ainsi que de l'importance relative de chacune des SFP. Les versions jyutping (voir notre courte description de la notation jyutping en §1.3) de ces 46 SFP sont présentées dans le Tableau 1.

aa1	aa1maa3	aa3	aa4	aa5	aa6	aak3
bo3	gaa1maa3	gaa2	gaa3	gaa4	gaak3	ge2
ge3	gwaa3	haa2	ho2	laa1	laa3	laa3haa2
laa4	laak3	le4	le5	lei4[6]	lei4gaa3	lo1
lo3	lo4	lok3	maa3	me1	ne1	sin1
tim1	waa2	wo3	wo4	wo5	zaa1maa3	zaa3
zaa3haa2	zaa4	ze1	zek1			

Tableau 1. SFP considérées dans notre étude

Les analyses envisagées requièrent également des informations plus détaillées pour décrire chacune des SFP, à savoir :

- sa ou ses notations jyutping alternatives, le cas échéant (allophonie, allomorphie ou variations de ton);
- sa ou ses notations en caractères chinois;

[6] *lei4* en tant que particule autonome est rapportée et décrite dans plusieurs travaux (entre autres Fung, 2000 ; Matthews et Yip, 2011), mais ignorée dans d'autres (Kwok, 1984 ; Yau, 1965). Toutefois, son homophonie avec le verbe

- son statut de particule strictement interrogative (§1.4.2);
- sa position dans une séquence de SFP (§1.2).

Les informations concernant chacune des SFP de notre liste sont présentées en Annexe A.

3.2.2 SFP interrogatives

Comme mentionné à la section 1.4.2, certaines SFP ont un statut spécifiquement interrogatif, en ce sens qu'elles sont utilisées de manière autonome (en l'absence de structure lexico-syntaxique interrogative) pour conférer à l'énoncé la forme interrogative. Les SFP auxquelles nous avons attribué ce statut sont 呀|aa4[7], 吓|aa5, 㗎|gaa4, 嗱|laa4, 咩|me1 et 喳|zaa4 (Kwok, 1984 ; Sybesma et Li, 2007)[8]. Ces SFP marquent de manière univoque un énoncé comme étant interrogatif.

Les particules 呢|ne1 et 哿|ge2, dont le statut est variable (elles peuvent se trouver dans des contextes déclaratifs, accompagner des structures interrogatives et être utilisées de manière autonome pour former des interrogations), ont été assimilées à des particules ambivalentes et ne sont donc pas associées à une forme d'énoncé spécifique.

3.2.3 Groupes de SFP

Comme mentionné à la section 1.2, les SFP peuvent généralement être associées à la fin d'un énoncé, ce qui a pour effet de combiner leurs apports sémantico-pragmatiques. La production de telles associations est toutefois contrainte par des règles de positionnement (Matthews et Yip, 2011).

Ko (2000) rapporte que les groupes sont cependant très peu utilisés par les enfants. De notre côté, même si nous prenons en compte la possibilité que plusieurs SFP soient combinées pour former un groupe à la fin d'un énoncé, nous n'avons pas cherché à analyser spécifiquement la production des groupes par les

[7] Le statut de particule interrogative de aa4 est remis en question par Sybesma et Li (2007), mais il semble que la majorité des sources lui confèrent tout de même ce statut.

[8] Bien que Kwok (1984) catégorise 嗎|maa3 comme strictement interrogative, Sybesma & Li (2007) mentionnent également un usage non interrogatif apportant une notion d'évidence. Au vu des données du corpus, où elle est souvent associée à une marque de ponctuation affirmative (p. ex. « 開 得 倒 **maa3** ! » : « C'est enfin ouvert! »), nous avons décidé ne pas lui donner le statut d'interrogative stricte.

enfants, et ne leur appliquons donc pas de traitement particulier. L'existence des groupes, et les règles auxquelles ils obéissent, sont utilisées uniquement pour statuer sur la validité de certaines occurrences de SFP ambiguës.

3.2.4 Constructions lexico-syntaxiques reconnaissables

Les constructions lexico-syntaxiques utilisées comme marqueurs de forme interrogative (cf §1.4.2) sont constituées de composants mono- ou plurisyllabiques. Plusieurs constructions connaissent une certaine allomorphie, comme la construction « quand…? » qui peut être réalisée au moyen de 幾點|gei2dim2 ou de 幾時|gei2si4. Les différences sémantiques ou pragmatiques entre ces formulations ne sont toutefois pas pertinentes pour notre étude, nous regroupons donc les différents allomorphes dans les catégories sémantiques présentées dans le Tableau 2.

Signification/emploi	Séquences
« Comment ? »	點|dim2 樣|joeng2
« Où ? »	邊|bin1 度|dou6
« Pourquoi ? »	點|dim2 解|gaai2
« Quand ? »	幾|gei2 點|dim2 幾|gei2 時|si4
« Qui ? »	邊|bin1 個|go3
« Quoi ? »	點|dim2 算|syun3 乜|mat1 嘢|je5 咩|me1 嘢|je5 咩|me1
Interrogatif générique	几|gei2 邊|bin1 點|dim2 乜|mat1
hai-m-hai (tag question) – question oui–non	係|hai6 唔|m4 係|hai6 係|hai6 咪|mai6 (forme compacte)
A-not-A – question oui–non	A_1 唔|m4 A (réduplication de la première syllabe de l'unité A avec insertion du négatif 唔|m4)

	有|jau5 冇|mou4 (forme réduite de la locution 有|jau5 唔|m4 有|jau5, « avoir non avoir », couramment utilisée)

Tableau 2. Constructions lexico-syntaxiques interrogatives

Les composants des constructions interrogatives sont de manière générale transcrits à l'aide de caractères chinois, dont le choix semble très stable parmi les transcripteurs. On trouve cependant quelques occurrences de transcriptions en jyutping. Nous avons donc également ajouté ces versions comme critères de reconnaissance.

3.2.5 Reconnaissance des SFP

Pour détecter dans un texte les entités décrites ci-dessus (SFP et constructions lexico-syntaxiques), la reconnaissance de la séquence d'unités syllabiques qui la composent s'avère globalement satisfaisante. Les graphies des mots QU sont pour leur part majoritairement monosémiques (à l'exception de 咩|me1, qui peut également être une SFP interrogative). La structure A-not-A, quant à elle, est également univoque puisqu'une telle séquence est toujours interprétée comme une interrogation polaire sur le thème verbal ou adjectival A.

Les séquences syllabiques correspondant à des SFP sont quant à elles également majoritairement monofonctionnelles, avec toutefois quelques exceptions notables comme 呢|ne1 (particule de mise à jour du thème), 咩|me1 (mot interrogatif signifiant « quoi ») et 嘅|ge3 (marqueur possessif). Cependant, les critères grammaticaux d'utilisation des SFP (position en fin de clause, rang dans la séquence de SFP) permettent en règle générale de reconnaître les occurrences utilisées comme SFP. Il va sans dire que l'utilisation de ces règles comme critères de désambiguïsation implique de considérer tous les usages comme conventionnels, ce qui constitue bien entendu une simplification de la problématique liée à l'analyse de corpus, particulièrement en rapport avec le CS.

Les versions originales des corpus utilisés mises à disposition du public (voir les descriptions des corpus au chapitre 4) contiennent les transcriptions des énoncés sous forme segmentée, c'est-à-dire que des espaces ont été insérées afin de délimiter des « mots » et d'en faciliter le traitement automatique. Le cantonais est toutefois traditionnellement transcrit sans séparation entre les constituants des propositions, et il n'existe pas de distinction claire entre les notions de morphème et de lexème (Lam *et al.*, 2024). La

segmentation en cantonais constitue donc un problème complexe concernant lequel il n'existe pas de consensus; dans la transcription d'un corpus oral, elle peut par conséquent varier selon les directives du corpus ou l'interprétation des transcripteurs, et ne constitue pas une information stable. On observe par exemple dans nos corpus une différence dans la manière de transcrire les qualificatifs de couleur, illustrée en (3) :

(3) a. INV: 藍色，靚 唔 靚 架？ (hku-040515)
laam4sik1 leng3 m4 leng3 gaa3
bleu beau NEG beau SFP
Est-ce que le bleu est joli?

b. CHI: 駿駿 藍 色 釣 到． (cc-ccc020507)
zeon3zeon3 laam4 sik1 diu3 dou2
Zeon-DIM bleu couleur pêcher COMPL
Zeon a pêché un poisson bleu.

c. CHI: 最 鍾意 食 白色 粥． (hku-040021)
zeoi3 zung1ji3 sik6 baak6sik1 zuk1
le.plus aimer manger blanc gruau
Je préfère manger le gruau blanc.

d. CHI: 我 鍾意 呢啲 白 色 貓貓． (cc-mhz020618)
ngo5 zung1ji3 ni4di1 baak6 sik1 maau1maau1
moi aimer ces blanc couleur chat-DIM
J'aime ces petits chats blancs.

Ces exemples illustrent la variation systématique entre la présence et l'absence, selon le corpus, d'une espace entre la racine du mot de couleur (粉紅 : rose, 黃 : jaune) et le suffixe morphémique 色 utilisé pour mentionner la couleur dans certaines configurations adjectivales.

Comme nous l'avons expliqué plus haut, le niveau d'analyse requis par les objectifs que nous nous sommes fixés (identification des SFP dans les énoncés, reconnaissance de certaines structures syntaxiques particulières, évaluation du type d'énoncés) peut donc, de manière globalement satisfaisante, être atteint au moyen d'une décomposition basée sur la reconnaissance d'unités linguistiques simples représentées par des caractères chinois uniques ou des séquences jyutping minimales (association d'une partie

segmentale et d'un ton), correspondant phonologiquement à une syllabe[9]. Cette approche nous permet de nous affranchir de traitements additionnels et potentiellement sources de variabilité, outre la segmentation, comme les analyses morphologique et syntaxique.

L'utilisation de l'unité syllabique comme base d'analyse a également un impact sur d'autres aspects de l'analyse, comme les calculs de MLU (cf. §3.3 ci-dessous) ou de fréquence d'utilisation des SFP.

3.2.6 Types d'énoncés et décomposition en segments

Comme nous l'avons décrit à la section 1.4.2, il existe un lien étroit entre la forme d'un énoncé (déclarative ou interrogative) et les SFP qui sont susceptibles d'y être utilisées – et ainsi un lien potentiel entre les formes des énoncés et l'acquisition des SFP par les enfants. L'évaluation de la forme interrogative d'un énoncé peut être effectuée grâce à la reconnaissance d'éléments interrogatifs s'y trouvant (particules, locutions ou constructions interrogatives), comme nous l'avons décrit plus haut. Nous avons également déjà évoqué au chapitre 1 le fait que certaines SFP (呀|aa4, 吓|aa5, 㗎|gaa4, 嗱|laa4, 咩|me1 et 喳|zaa4) ont une fonction strictement interrogative et transforment un énoncé déclaratif en interrogation lorsqu'elles lui sont apposées. Enfin, dans le cadre d'une transcription, la ponctuation finale (p. ex. point d'interrogation) est également susceptible de fournir une indication sur la forme d'un énoncé, dans les cas où celui-ci ne contient pas d'autre élément formel. L'évaluation de ces différents éléments permet d'assigner à chaque énoncé la forme déclarative ou interrogative.

Nous avons également évoqué à la section 1.4.1, en nous appuyant sur l'analyse de Luke (1990), que les contextes grammaticaux dans lesquels les SFP peuvent effectivement être utilisées ne se limitent pas aux phrases graphiques complètes (délimitées par une marque de ponctuation finale), mais correspondent plutôt à des parties de ces phrases, comme des propositions ou des groupes de mots. Dans le cadre de la transcription d'un corpus oral, de tels regroupements, produits en séquence constituant une prise de parole complète, sont susceptibles d'être délimités au moyen de virgules (pauses orales). Pour l'analyse des SFP (particules de ***fin*** de ***phrase***), nous avons donc pris le parti de conserver l'aspect « final » dans notre processus de reconnaissance en ne considérant que les occurrences constituant la séquence finale

[9] Notons que ces unités ne peuvent être assimilées à des morphèmes : certaines unités lexicales sont formées d'unités syllabiques n'apparaissant jamais individuellement (c'est le cas par exemple de « 甲由|gaat6zaat6 » signifiant « cafard », dont les composants n'existent que dans ce lexème); d'autres peuvent être analysées comme des combinaisons d'unités (p. ex. gaa4 = ge3 + aa4; Kwok, 1984, p. 87 ; Matthews et Yip, 2011, p. 391).

d'une structure grammaticale, mais d'en adapter l'aspect « phrase » en considérant comme structure englobante non pas les phrases graphiques délimitées par des marques de ponctuation finale, mais plutôt les propositions internes à ces phrases, délimitées par la présence d'une virgule ou d'une ponctuation finale dans le discours transcrit. Les exemples en (4), extraits de notre corpus, montrent que des SFP apparaissent effectivement à la fin de ces entités.

Pour la suite de notre rapport, nous désignerons ces structures, qui constituent ainsi les structures englobantes de base pour la reconnaissance de séquences de SFP, au moyen du terme « segment ».

(4) a. CHI: 梗 係 **啦** , 係 咪 **呢** , 聽 haa5 人 講 嘢 **aa3** ? (hku-050023)
gang2 hai6 **laa1**, hai6 mai1 **ne1**, teng1 haa5 jan4 gong2 je5 **aa3**
C'est sûr, n'est-ce pas? Écoute ce que disent les gens, d'accord?

b. INV: 唔 係 **aa3** , 鐘 鐘 嚟 gaa3 , 你 睇 **吓** . (cc-ckt020108)
m4 hai6 **aa3**, zung1 zung1 **lei4gaa3**, nei5 tai2 **haa2**.
Non, il est temps, regarde.

3.3 Calcul de la MLU

La MLU (*Mean Length of Utterance*) est couramment utilisée comme un indicateur du niveau de développement langagier d'un enfant (Brown, 1973 ; Rice *et al.*, 2010), et peut à ce titre servir à identifier des tendances dans le processus d'acquisition. La définition proposée par Brown (1973), et généralement reprise lors de l'analyse pour les langues indo-européennes, se base sur le nombre de morphèmes produits dans les énoncés. Toutefois, comme nous l'avons expliqué plus haut, le manque de cohérence dans les processus de segmentation appliqués dans les corpus, dû à l'absence de séparation claire entre morphème et lexème en cantonais, a justifié que nous utilisions les unités syllabiques comme composants langagiers de base. En conséquence, la MLU que nous calculons correspond à la longueur moyenne en termes de syllabes, que Tse *et al.* (2002) rapportent comme ayant déjà été utilisée pour évaluer le niveau de développement langagier dans les langues chinoises. Notons que cette mesure, si elle ne correspond pas à la définition « canonique » basée sur le décompte des morphèmes, a cependant le mérite d'être objective et de ne pas dépendre du niveau d'analyse appliqué aux énoncés, ou de la portée du concept de morphème, encore mal défini en cantonais (Sybesma et Li, 2007 ; voir également à ce sujet la discussion de Lam *et al.*, 2024).

La notion d'énoncé évoquée dans le concept de MLU (*utterance* en anglais) doit également être clarifiée pour effectuer le calcul. Comme nous l'avons expliqué plus haut, et en accord avec la définition de Luke (1990), la présence de SFP à la fin des segments (§3.2.6) justifie de les considérer comme entités assimilables à des énoncés, et de les utiliser comme construction langagière dont la longueur est représentative du niveau de développement langagier.

Enfin, à la différence de Brown (1973, p. 54) qui suggère de ne prendre en compte que les 100 premiers énoncés pour effectuer le calcul de MLU, nous avons pris le parti d'utiliser l'ensemble des segments produits par l'enfant. La MLU que nous associons à une conversation est donc calculée en divisant le nombre total de syllabes produites par un enfant au cours de cette conversation par le nombre total de segments non vides (c'est-à-dire des segments contenant au moins un élément oral, même s'il n'est pas reconnu comme un lexème ou un morphème existant).

3.4 Génération des annotations

Le processus de génération des annotations utilisées pour les analyses statistiques est décrit de manière exhaustive au chapitre 5. Pour maximiser les possibilités d'analyse par des modèles statistiques, le choix a été fait de ne pas agréger les données (par exemple en effectuant des décomptes globaux de production des SFP dans les conversations), mais plutôt de générer des descriptions exhaustives de tous les contextes de production. Ainsi, chaque occurrence de SFP dans notre corpus donne lieu à la création d'une annotation complète contenant les éléments listés dans le Tableau 3 :

Caractéristiques personnelles de l'enfant	genre
	âge
	niveau de développement langagier (MLU)
	Fréquence d'utilisation des SFP
	Diversité des SFP utilisées
Caractéristiques personnelles de l'interlocuteur	genre (si disponible)
	rôle (statut vis-à-vis de l'enfant) et appartenance à la famille
	tranche d'âge
Environnement langagier	SFP produite
	position de la SFP dans la séquence
	nombre de SFP dans le segment
	longueur du segment (syllabes)
	forme du segment (déclarative ou interrogative)
	construction lexico-syntaxique

	forme du segment précédent (dernier segment produit par l'interlocuteur)
	SFP produite dans le segment précédent
Données propres à la SFP	fréquence d'utilisation par l'enfant (dans la conversation)
	fréquence d'utilisation par les interlocuteurs (dans la conversation)

Tableau 3. Données d'annotation de chacune des occurrences de SFP dans le corpus

Une description fonctionnelle de tous ces champs, ainsi que de champs complémentaires à usage utilitaire, est donnée à la section 5.10.

Ces annotations sont générées non seulement pour chaque occurrence de SFP dans le discours des enfants, mais également pour chaque occurrence de SFP dans le discours de leurs interlocuteurs. En effet, comme il a été évoqué aux chapitres 1 et 2, l'influence du CDS sur l'acquisition du langage par les enfants est largement acceptée et doit être étudiée.

De plus, des annotations sont également générées pour les segments ne contenant aucune SFP (pour le CS comme pour le CDS). Pour chaque segment ne contenant pas de SFP, une unique annotation est produite; cette annotation contient exactement les mêmes informations que celles décrivant la production d'une SFP. Cela permet de comparer les contextes de production et ceux de non-production, afin d'évaluer les facteurs favorisant l'utilisation des SFP de manière globale, ou de comparer des groupes de segments partageant certaines caractéristiques pour étudier l'impact de différentes combinaisons de facteurs. Comme nous le verrons au chapitre 8, l'utilisation de valeurs binaires indiquant la production de SFP permet l'ajustement de modèles statistiques linéaires généralisés, grâce auxquels l'importance relative et les interactions entre facteurs peuvent être estimées.

CHAPITRE 4

Données utilisées dans le cadre du projet

Notre étude de l'utilisation des SFP dans le discours des enfants est basée sur le contenu de deux corpus d'interactions enfant–adulte : le HKU-70 (Fletcher *et al.*, 2000) et le CANCORP (Lee *et al.*, 1996), disponibles publiquement au format CHAT décrit ci-dessous. En plus de permettre l'analyse du discours des enfants, ces corpus donnent également accès aux énoncés de leurs interlocuteurs (adultes ou enfants), ce qui permet d'étudier les interactions entre le CS et le CDS. Nous avons de plus utilisé le corpus d'interactions entre adultes *Hong Kong Cantonese corpus* (Luke et Wong, 2015) pour effectuer des comparaisons des propriétés du discours expert de locuteurs natifs adultes (ADS, *adult-directed speech*, discours adressé aux adultes) au CS et au CDS. Dans ce chapitre, nous décrivons tout d'abord le format CHAT utilisé pour la transcription des corpus; nous présentons par la suite chacun des corpus individuellement, puis donnons finalement un aperçu global des données analysées.

4.1 Description du format CHAT

Les corpus utilisés sont structurés selon le format CHAT (MacWhinney, 2000). La conception de ce format s'est inscrite dans la perspective de mettre à disposition des chercheurs un écosystème permettant la compilation et le partage de transcriptions d'interactions entre individus de compétences langagières très différentes (i. a. parent–enfant, médecin–patient). La structure et la flexibilité du format CHAT constituent des éléments fondamentaux de la base de données CHILDES, dont l'un des objectifs premiers est de permettre l'étude de la compétence et du développement langagiers des enfants.

Un fichier au format CHAT correspond à une séance d'interaction entre plusieurs interlocuteurs, dont l'un peut être spécifiquement ciblé (typiquement un enfant), constituée d'une série de prises de parole. Des métadonnées comme l'âge ou le genre peuvent être indiquées pour chacun des participants, en particulier pour l'enfant ciblé par l'observation, ce qui permet une documentation précise du contexte d'interaction. Ces métadonnées se présentent sous la forme présentée en (5).

On y retrouve en particulier l'âge et le genre de l'enfant, ainsi que la liste des personnes participant à la conversation.

(5)
```
@Begin
@Name of CHI:        Yiu Chun Li
@Birth of CHI:       14-Jun-1989
@Sex of CHI:         Female
@Tape location:      22
@Date:               30-Dec-1992
@Time duration:      10:00-11:00
@Age of CHI:         3;6.16
@Participants:       CHI Child, INV Investigator, MOT Mother, SIS sister, SER servant
@Transcriber:        Kitty Ka-sinn Szeto
```

Dans les corpus que nous utilisons, chaque prise de parole fait l'objet d'une transcription lexicale, qui peut être accompagnée de données supplémentaires (p. ex. romanisation, analyse morphologique ou syntaxique, gestes et actions associés à l'énoncé). Grâce à sa syntaxe élaborée et systématique, un fichier CHAT peut être parcouru et traité par des outils informatiques comme CLAN (MacWhinney, 2018), ou à l'aide de scripts écrits en langage python (§5.1), ce qui permet d'en extraire les informations nécessaires à l'analyse des structures langagières utilisées.

Voici un exemple d'énoncés au format CHAT (hku-050525) :

(6)
```
*INV:   唔 驚 架 擺 喺度 .
%mor:   cconj|唔 verb|驚-Inf-S noun|架 verb|擺-Inf-S noun|喺度 .
%gra:   1|4|CC 2|4|XCOMP 3|4|NSUBJ 4|5|ROOT 5|4|OBJ 6|4|PUNCT
*CHI:   點解 aa3 ?
%mor:   verb|點解-Inf-S x|aa3 ?
%gra:   1|2|ROOT 2|1|OBJ 3|1|PUNCT
```

Les lignes commençant par ***INV** et ***CHI** correspondent aux transcriptions (lexicales) des énoncés produits respectivement par l'expérimentateur (INV – *investigator*) et l'enfant (CHI – *child*). Les lignes commençant par **%mor** et **%gra** correspondent respectivement aux analyses morphologique et syntaxique. Notons que ces analyses, intégrées aux fichiers de la version du HKU-70 mise à disposition sur le site du projet CHILDES, ont été générées automatiquement selon le standard *Universal Dependencies* (de Marneffe *et al.*, 2021 ; Liu et MacWhinney, 2024); comme expliqué dans la section 3.2.5, nous avons cependant fait le choix de ne pas utiliser de données morphosyntaxiques pour l'identification des SFP, ces couches d'analyses sont donc ignorées par la suite.

4.2 Le corpus HKU-70

4.2.1 Aperçu

Le corpus HKU-70[10] (ci-après HKU) a été collecté entre 1998 et 2000 à Hong Kong (Fletcher *et al.*, 2000). Il est composé des transcriptions d'enregistrements de 70 séances d'interaction avec des enfants (tous différents) âgés entre 2 ans et demi et 5 ans et demi; à chaque tranche de 6 mois correspondent dix enregistrements d'enfants (5 filles et 5 garçons). Les rencontres sont menées exclusivement par des membres de l'équipe de recherche, avec lesquels les enfants ont toutefois pu faire connaissance au préalable, et visent à axer les conversations sur des situations de la vie quotidienne pour favoriser la prise de parole et inciter les enfants à s'exprimer spontanément sur différents thèmes[11].

Notons que les enregistrements audios originaux sont également disponibles pour l'intégralité des transcriptions, bien que les énoncés n'aient pas été annotés pour être alignés avec les segments correspondants. L'accès aux enregistrements a permis de déterminer le genre des locuteurs et locutrices pour lesquels cette information n'était pas disponible formellement (§5.6). Cette information sera utilisée dans notre analyse pour étudier l'impact du genre de l'interlocuteur sur l'utilisation des SFP.

4.2.2 Caractéristiques des transcriptions

L'archive du HKU disponible en ligne sur CHILDES contient l'ensemble des fichiers de transcriptions. Pour chacun des énoncés, la transcription est majoritairement composée de caractères chinois, mais des séquences en jyutping sont également utilisées pour certaines unités syllabiques. Pour une très large majorité, les séquences transcrites en jyutping correspondent à des SFP. Toutefois, des séquences jyutping sont également utilisées pour d'autres unités linguistiques. De plus, certaines SFP sont systématiquement transcrites en caractères chinois, d'autres systématiquement en jyutping, d'autres encore sont transcrites en utilisant une notation hybride (caractère chinois + ton). Voici quelques exemples extraits du corpus (les graphies des énoncés, incluant la segmentation et les romanisations, sont celles de la version originale; les lignes de romanisation ainsi que les gloses ont été produites manuellement) :

[10] Le corpus HKU-70 est accessible sur le site du projet CHILDES (*Child Language Data Exchange System*) à l'adresse suivante : https://childes.talkbank.org/access/Chinese/Cantonese/HKU.html.

[11] Le rapport de Fletcher *et al.* n'est pas disponible en ligne et n'a pas pu être consulté. Les informations fournies ici sont celles indiquées sur la page internet d'accès au corpus de la banque de données CHILDES, dont l'adresse est donnée plus haut.

(7) a. CHI: 唔好 嘈醒 佢 **aa3** ! (hku-020530)
m4hou2 cou4seng2 keoi5 **aa3**
ne.pas réveiller lui **SFP**
Ne le réveille pas!

b. CHI: 瀡 牙膏 落 **囃** ! (hku-050121)
zit1 ngaa4gou1 lok6 **bo3**
presser dentifrice déposer **SFP**
Étale le dentifrice!

c. INV: 果汁 要 倒 **噦 3**. (hku-050525)
gwo2zap1 jiu3 dou2 **ge3**
jus devoir verser **SFP**
Il faut verser le jus.

Notons toutefois que les règles de transcription sont globalement stables et homogènes : les mots inscrits en jyutping le sont de manière systématique et, à quelques exceptions près, correspondent généralement soit à des SFP, soit à des interjections ou non-mots. D'autre part, bien que ce ne soient pas les seules, les SFP transcrites en jyutping sont majoritairement celles pour lesquelles les caractères chinois utilisés sont potentiellement ambigus (ce qui est particulièrement le cas avec la famille aa*).

4.3 Le corpus CANCORP

4.3.1 Aperçu

Le corpus CANCORP a été collecté aux cours des années 1991 et 1992 (Lee *et al.*, 1996). Il s'agit d'un corpus longitudinal qui contient les transcriptions de 171 enregistrements effectués sur une période d'un an auprès de huit enfants (4 filles et 4 garçons), âgés d'entre 1 an et demi et 2 ans et demi au début du projet. La plupart des conversations de ce corpus impliquent d'autres participants que les expérimentateurs et l'enfant cible, aussi bien des membres de l'entourage proche de celui-ci que des personnes extérieures. Chaque enfant a participé à 21.375 conversations en moyenne (σ=4.34). Les durées séparant les conversations successives pour un enfant sont très irrégulières, pouvant varier d'un jour à plus de deux mois pour un même enfant (WBH). Le Tableau 4 donne un aperçu de la distribution des entrevues par enfant.

Au vu de ces données, on peut constater que les rythmes appliqués aux différents enfants sont très irréguliers. Lors de l'utilisation du corpus, il conviendra de s'interroger sur la pertinence de considérer les conversations comme des entités indépendantes ou de prendre également en compte l'identité de l'enfant pour créer un lien entre les conversations.

	Âge min.	Âge max.	Nb convs	Écart min.	Écart max.	Moyenne	Écart type
CCC	01;10.08	02;10.27	22	7	54	18.29	9.72
CGK	01;11.01	02;09.09	19	3	68	17.39	18.01
CKT	01;05.22	02;07.02	25	7	41	16.86	8.69
HHC	02;04.08	03;04.14	16	6	52	24.73	12.01
LLY	02;08.10	03;08.09	20	5	39	19.16	9.33
LTF	02;02.10	03;02.18	16	12	41	24.87	9.20
MHZ	01;07.00	02;08.06	26	6	48	16.04	8.67
WBH	02;03.23	03;04.08	27	1	65	14.62	14.90

Tableau 4. Distribution des conversations par enfant dans le corpus CANCORP

Après sa publication en 1996, le corpus a divergé en deux versions différentes : la version « CHILDES » (ci-après LWL pour Lee–Wong–Leung, d'après les noms des trois chercheurs responsables), disponible sur le site internet de CHILDES[12], et la version 2012 (*2012 updated standard version of CANCORP*, ci-après CC2012), disponible depuis le site du *Language Acquisition Laboratory* de la *Chinese University of Hong Kong*[13]. Les deux versions ont fait l'objet de traitements et améliorations différents, sous l'initiative d'équipes indépendantes.

Une analyse préliminaire des deux versions, effectuée par nos soins, a révélé qu'elles ne sont pas compatibles entre elles, n'étant plus alignées – en particulier, certains énoncés « vides » (dépourvus de contenu verbal, mais correspondant par exemple à une action : **CHI: 0 [=! nods].*) et segments inintelligibles ont été supprimés de LWL; certains fichiers y sont manquants (en particulier pour l'un des enfants, WBH, où seuls 16 fichiers sur 27 sont disponibles), d'autres ont été tronqués. De plus, l'analyse lexicale préliminaire des transcriptions de cette version dans l'utilitaire CLAN (MacWhinney, 2018) a rapporté un grand nombre de lexèmes inconnus. Enfin, si l'analyse morphologique de CC2012 correspond à l'analyse manuelle correspondant au modèle décrit dans Lee *et al.* (1996) lors de la création du corpus original, LWL a fait l'objet d'une nouvelle analyse automatisée. Voici un exemple de différence entre CC2012 (8) et LWL (9) (cc-ccc021027) :

[12] Le corpus CANCORP est accessible sur le site du projet CHILDES à l'adresse suivante : https://childes.talkbank.org/access/Chinese/Cantonese/LeeWongLeung.html

[13] https://www.arts.cuhk.edu.hk/~lal/corpora.html

(8) *INV: 去 邊度 &aa3 駿駿?
%rom: heoi3 bin1 dou6 &aa3 zeon3 zeon3 ?
%mor: dir wh sfp nnpp.

(9) *INV: 去 邊度 &~a3 駿駿 ?
%mor: verb|去-Inf-S verb|邊度-Inf-S verb|駿駿-Inf-S ?
%gra: 1|3|ADVCL 2|3|ADVCL 3|3|ROOT 4|3|PUNCT

Si l'information du niveau morphologique (ligne %mor) présente dans LWL semble plus complète (il s'agit de l'analyse effectuée selon le standard *Universal Dependencies*), elle est toutefois clairement erronée (邊度 est effectivement un mot interrogatif signifiant « où », comme indiqué par l'étiquette wh dans CC2012, et n'existe *a priori* pas sous forme de verbe; 駿駿 est une apostrophe réduplicuée, et donc un nom propre, en accord avec l'étiquette nnpp). De plus, la particule *aa3* (recodée &~a3) est ignorée dans la ligne morphologique. Comparativement, l'analyse de CC2012, qui utilise les marqueurs morphologiques décrits dans Lee *et al.* (1996), correspond tout à fait au sens de la phrase (« Où vas-tu, Zeon? »).

À beaucoup d'égards, la version LWL semble donc souffrir d'importantes imperfections. Notre choix s'est donc porté sur la version CC2012 (ci-après CC), à laquelle nous avons appliqué un traitement d'uniformisation (§5.1).

Les fichiers de transcription ne fournissent aucune information relative aux personnes expérimentatrices, et les enregistrements originaux ne sont pas disponibles dans la base de données CHILDES. Dans ce contexte, il n'a malheureusement pas été possible de déterminer le genre de ces personnes pour l'ajouter aux annotations des SFP, comme nous avons pu le faire pour le HKU. Toutefois, le corpus CANCORP contient également des interactions avec des personnes de l'entourage proche des enfants (parents, grands-parents, …), pour lesquelles il est possible d'extrapoler cette information.

4.3.2 Caractéristiques des transcriptions

À la différence du HKU qui bénéficie d'une certaine stabilité dans les méthodes de transcription, le CC souffre pour sa part d'une forte variabilité dans les graphies utilisées pour transcrire les SFP. On y trouve ainsi des différences importantes de styles de romanisation (longueur de la voyelle *a*, transcription de la consonne *z*); certaines SFP sont par ailleurs transcrites parfois en jyutping, parfois en caractères chinois.

Voici quelques exemples d'occurrences de la SFP aa3, extraits du corpus original (les gloses ne sont pas disponibles dans le corpus, et ont été ajoutées par nos soins[14]) :

(10) a. MOT: 妳 有 冇 呀 ? (cc-cgk011108)
nei5 jau5 mou5 aa3
tu avoir avoir-NEG SFP
Est-ce que tu en as?

b. CHI: 晌 度 笑 &aa3 . (cc-cgk011101)
hoeng2 dou6 siu3 aa3
temps QT rire SFP
J'ai beaucoup ri.

c. AUN: 食 唔 食 &a3 ? (cc-ckt020009)
sik6 m4 sik6 aa3
manger NEG manger SFP
Est-ce que tu manges?

d. GRM: 唔該 邊個 啊 ? (cc-wbh020406)
m4goi1 bin1go3 aa3
s'il.vous.plaît qui SFP
Excuse-moi, qui est-ce?

On observe dans les exemples en (10) que deux caractères chinois différents ont été utilisés en a. et d. (呀 et 啊), et que la notation romanisée en c. ne correspond pas à la norme jyutping, selon laquelle le *a* long doit être doublé. De telles variations existent pour plusieurs SFP et ont requis un important travail de normalisation pour regrouper, de la manière la plus fiable et la plus exhaustive possible, les occurrences de toutes les SFP.

4.4 Hong Kong Cantonese corpus

Le *Hong Kong Cantonese corpus* (ci-après HKC) est un corpus composé de 58 séances d'interactions en cantonais, collecté à Hong Kong entre 1997 et 2002, impliquant une centaine de locuteurs adultes (15–60 ans) (Luke et Wong, 2015)[15]. Il vise à mettre à disposition des chercheurs un ensemble de conversations

[14] Les gloses et traductions d'énoncés en cantonais cités dans ce mémoire seront, autant que possible, ceux fournis dans les corpus ou documents d'où ils auront été extraits. Dans les cas où ces informations ne sont pas disponibles, nous fournirons des gloses obtenues sur le dictionnaire en ligne CantoDict accessible à l'adresse https://www.cantonese.sheik.co.uk/, en nous basant sur les correspondances entre caractères et prononciations, ou des traductions obtenues sur Microsoft Bing https://www.bing.com/translator?ref=TThis&from=yue&to=en. Elles sont dans ce cas uniquement données à titre indicatif et ne doivent pas être considérées comme fiables.

[15] Le *Hong Kong Cantonese corpus* est disponible en ligne à l'adresse suivante : https://github.com/fcbond/hkcancor.

non dirigées permettant l'analyse du cantonais dans des contextes pseudo-authentiques : les conversations ont été enregistrées dans des locaux de recherche, mais les locuteurs étaient des personnes proches (famille, amis), les thèmes des discussions étaient entièrement libres, et la personne expérimentatrice quittait la pièce pendant l'enregistrement.

Les conversations sont transcrites presque exclusivement en caractères chinois, les seules exceptions étant des interjections, non-mots ou marqueurs d'hésitation. La couche d'analyse morphologique présente dans les fichiers originaux, dans laquelle les SFP sont identifiées grâce à l'étiquette *y*, a facilité l'association entre les SFP et les caractères chinois correspondants, ce qui a permis d'appliquer à ce corpus les mêmes traitements qu'aux deux autres, et ainsi de générer les annotations selon le même format (cf. chapitre 5). Il a ainsi été possible d'effectuer les mêmes analyses et de générer des modèles statistiques à des fins de comparaison entre le discours entre adultes, le discours des adultes envers les enfants et le discours des enfants.

4.5 Description de notre corpus de travail

Nous présentons ici les caractéristiques de notre corpus d'interactions enfant–adulte, constitué de la combinaison du HKU et du CC.

4.5.1 Population

Notre corpus est constitué d'un ensemble de 241 conversations mettant en jeu de deux à sept participants, réparties comme indiqué sur la Figure 1.

Les caractéristiques de chacune des conversations sont données en Annexe B. Les âges des enfants vont de 1;05.22 à 05;08.00 (1.47–5.66 ans; μ=2.928, σ=0.969), leur répartition est illustrée sur la Figure 2. 117 conversations ciblent des filles et 124 ciblent des garçons. Toutefois, comme le CC est une étude longitudinale mettant en jeu huit enfants (quatre filles et quatre garçons) sur une période d'un an, le nombre effectif de filles est égal au nombre effectif de garçons, soit 39 (4 + 35). La répartition des conversations par genre et par tranche d'âge est présentée sur la Figure 3.

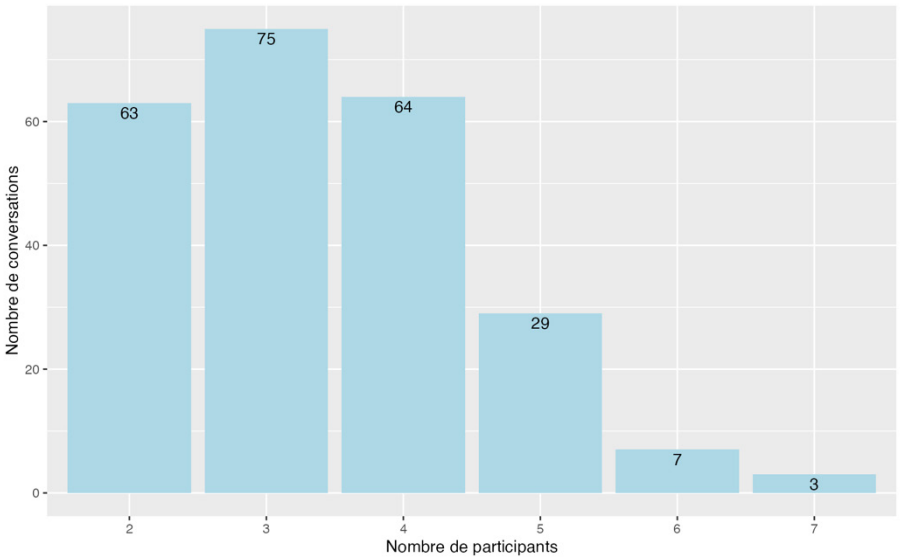

Figure 1.Effectifs de conversations par nombre de participants[16]

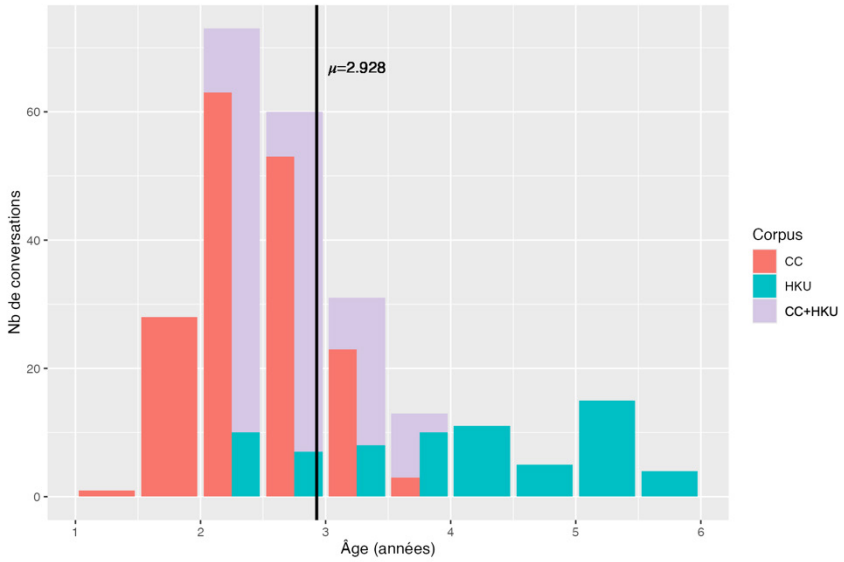

Figure 2.Répartition des âges des enfants cibles des conversations, ventilés par corpus

[16] Diagramme généré avec le module `ggplot2` (Wickham, 2016)

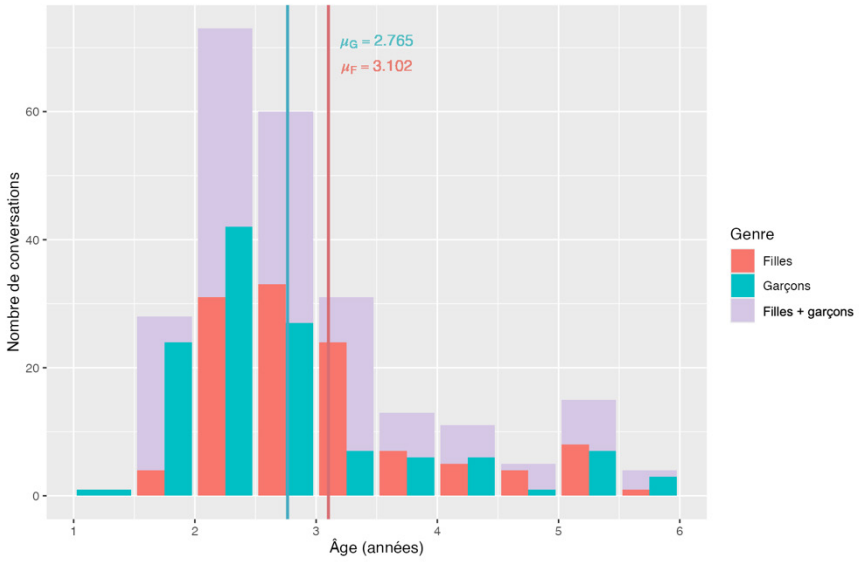

Figure 3.Répartition des genres des enfants par tranche d'âge

Si les deux corpus se différencient significativement par les tranches d'âge qu'ils recouvrent, ils sont toutefois comparables pour le nombre d'observations de chaque genre (χ^2=0.0833; df=1; p=0.7729; φ=0.0186). Le test t de Student nous indique cependant que les moyennes d'âge pour les conversations avec les garçons et celles avec les filles sont globalement significativement différentes (μ_F=3.102, μ_G=2.765; t=2.7432; df=237.68; p=0.0065).

4.5.2 Prises de parole et segments

Le format CHAT utilisé pour la transcription des conversations permet de rapporter, outre les mots et sons produits par les participants, certaines actions en rapport avec la séquence d'interaction et non nécessairement accompagnées de verbalisation. Dans les corpus originaux, ces actions sont codées au moyen d'un codage spécial illustré en (11) :

(11) a. *CHI: 0 [=! facial expression showing discontent]. (cc-wbh020609)
b. *SIS: 0 [=! sneezes]. (cc-lly030113)
c. *CHI: 0 [=! feeds the doll and uses the spoon to@s hit the doll]. (hku-030701)

De telles descriptions, si elles peuvent clairement être utilisées pour l'évaluation de la dimension pragmatique d'une situation, ne sont cependant pas directement exploitables dans le cadre de notre projet, qui se focalise sur la production verbale. Ainsi, bien qu'elles soient conservées dans la version unifiée de notre corpus sous une forme simplifiée (afin de conserver l'alignement avec les corpus originaux), elles ne sont pas considérées comme des énoncés, ne sont pas décrites par une annotation et

ne n'entrent donc pas dans les statistiques et les différents calculs comme celui de la MLU. Elles ne sont pas prises en considération dans les décomptes ci-dessous.

Notre corpus est constitué d'un total de 284231 prises de paroles, c'est-à-dire de lignes correspondant à une séquence de mots produite par un participant et assimilable à une phrase grammaticale unique ou un fragment de phrase. Comme nous l'avons expliqué à la section 3.2.6, il est toutefois indispensable de considérer individuellement les parties de ces prises de paroles séparées par des virgules dans la transcription comme des entités autonomes du point de vue de l'utilisation des SFP, puisque de nombreuses SFP sont utilisées à la fin de tels segments; cela revient à interpréter chaque prise de parole comme une série de segments (potentiellement réduite à un unique segment). En effet, sur 9741 segments ne se trouvant pas en fin d'énoncé, 2502 sont effectivement ponctués par des SFP. Cette segmentation des énoncés en fragments supportant l'utilisation de SFP se retrouve par ailleurs dans les deux corpus. Il est important de noter que l'approche que nous adoptons, bien qu'elle semble indispensable aux fins d'identifier les SFP et d'analyser leurs contextes de production de manière exhaustive, n'est pas mentionnée dans les autres études de corpus. Il en résulte une différence de traitement ayant un impact inévitable sur différents aspects de notre analyse, comme le calcul de la MLU ou encore les taux d'utilisation des SFP.

Au total, le corpus compte 104646 segments produits par des enfants (63223 par des garçons, 41423 par des filles), dont la répartition selon l'âge et le genre est présentée à la Figure 4.

Outre les enfants cibles, les conversations mettent en jeu différents types d'interlocuteurs. Les conversations du HKU (N=70) sont menées exclusivement par des expérimentateurs, seuls ou en dyades. Les conversations du CC (N=171) impliquent également généralement un expérimentateur, elles se passent toutefois dans des environnements plus familiaux et incluent la plupart du temps au moins un des parents, ainsi que d'autres personnes de l'entourage comme les grands-parents, frères ou sœurs, employées de maison, ou encore des personnes étrangères à la famille. Ainsi, notre corpus comprend 189111 segments produits par des expérimentateurs, mais également 35865 produits par des membres de l'entourage proche (famille, employées de maison). De même, 93771 des énoncés des enfants sont adressés à des personnes étrangères et 10875 aux personnes de l'entourage proche. Bien que le rôle joué par les expérimentateurs soit globalement nettement plus important que celui des personnes proches, ces données nous permettront tout de même d'observer les potentielles différences liées au degré de

familiarité de l'interlocuteur. À des fins d'analyse, nous avons catégorisé les interlocuteurs selon un statut `household` (« foyer »), et assigné à ce statut la valeur `TRUE` pour les personnes considérées comme appartenant à l'entourage proche de l'enfant, soit MOT (*mother*), FAT (*father*), AUN/MAN/WAI (*aunt*), CUS (*cousin*), GDM/GMO/GRA/GRM (*grandmother*), GRF (*grandfather*), UNC (*uncle*), BRO (*brother*), SIS (*sister*). Nous y avons également ajouté les employées domestiques SER (*servant*) et HEL (*caretaker*).

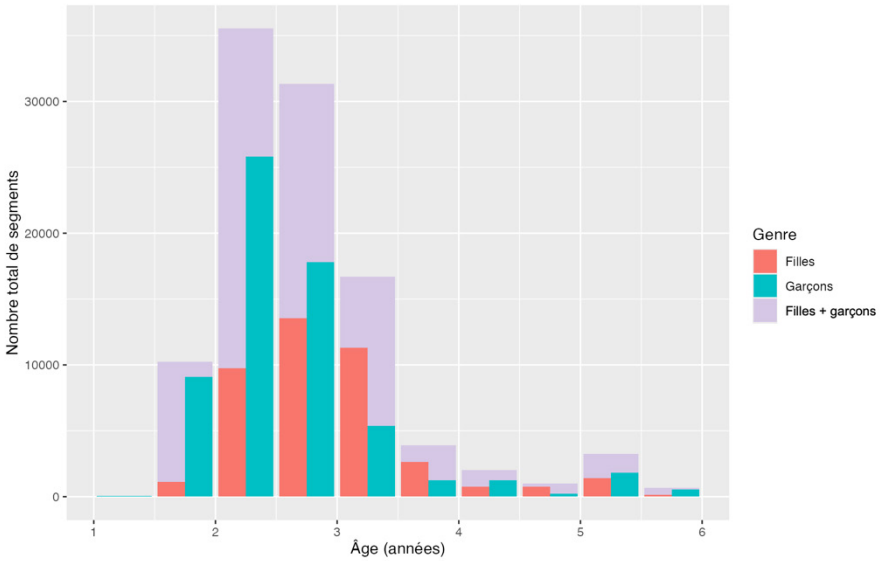

Figure 4.Répartition par tranche d'âge du volume de segments produits par les enfants

4.5.3 Développement langagier

4.5.3.1 MLU

Les MLU (dont le calcul est décrit en §3.3) s'échelonnent entre 1.45 et 6.851 (μ=2.912, σ=0.795). La forte corrélation entre la MLU et l'âge (considéré sous forme de logarithme, r=0.72) est très clairement illustrée sur la Figure 5 (la ligne bleue correspond à la régression linéaire liant le logarithme de l'âge à la MLU, avec son intervalle de confiance).

Par ailleurs, le test t de Student nous indique que, comme pour les moyennes des âges, les moyennes des MLU des garçons et des filles sont significativement différentes dans notre corpus (μ_F=3.220, μ_M=2.622; t=6.3001; df=238.86; p<0.001), ce qui est cohérent avec la corrélation observée entre l'âge et la MLU. Nous avons pu confirmer par ailleurs un effet significatif du genre de l'enfant donnant lieu à une MLU globalement plus élevée chez les filles à âge égal : l'ajustement d'un modèle linéaire *mlu~log(age)+sex* [R^2=0.5701; F(2, 238)=157.8, p<0.001] montre que le facteur *sex* est significatif (p<0.001) et que l'effet de

la valeur *MALE* sur la MLU est estimé à –0.359. Cette différence constatée entre garçons et filles, illustrée en Figure 6, est en accord avec les conclusions d'études antérieures selon lesquelles, en cantonais comme dans d'autres langues, les filles ont un développement langagier plus précoce que les garçons (p. ex. Bornstein *et al.*, 2004 ; Gayraud *et al.*, 2025 ; Tse *et al.*, 2002). Là encore, les lignes rose (filles) et cyan (garçons) correspondent aux régressions linéaires liant le logarithme de l'âge à la MLU.

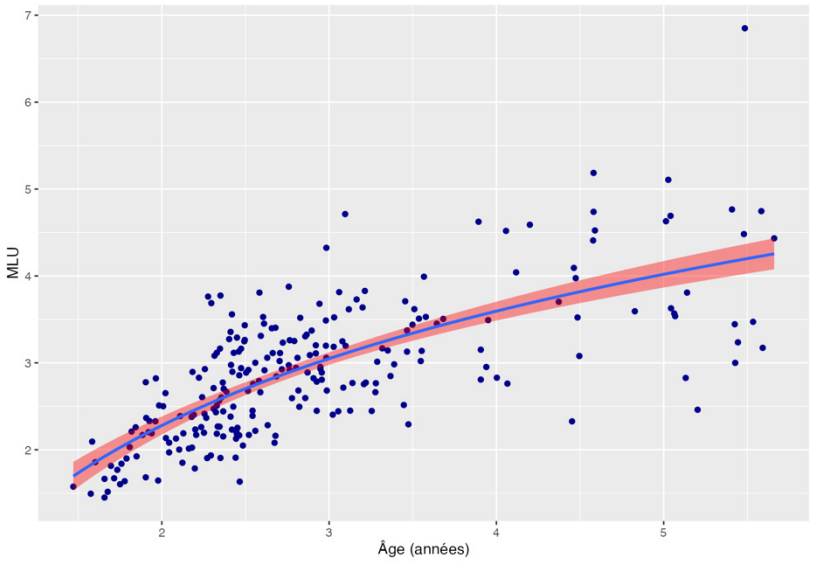

Figure 5.Évolution de la MLU en fonction de l'âge dans le corpus, avec régression linéaire (log)

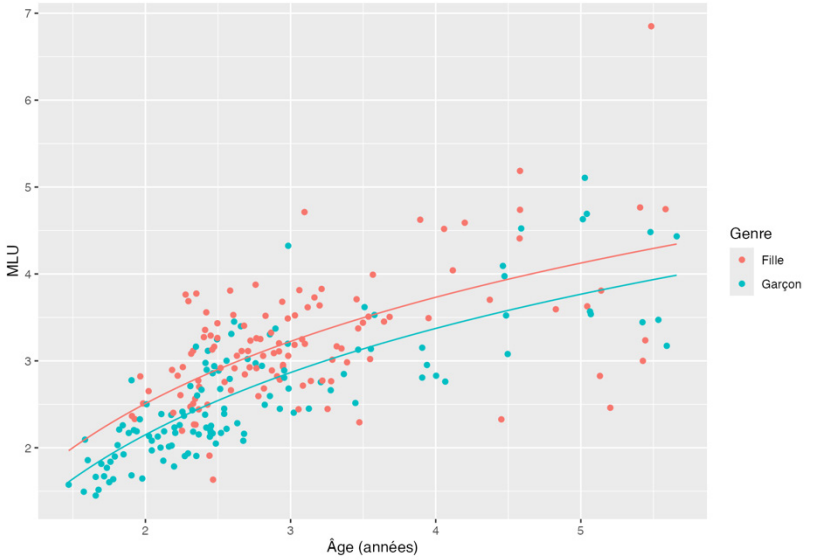

Figure 6. MLU en fonction de l'âge pour les filles et les garçons

Comme cela avait été suggéré par Brown (1973), et repris par Ko (2000) dans le cadre de l'acquisition du cantonais, la MLU est souvent utilisée sous forme d'intervalles pour définir des stades de développement langagier. Cependant, les stades définis par Brown en fonction de l'acquisition des structures syntaxiques de l'anglais ne peuvent s'appliquer pour le cantonais. Ko a pour sa part défini trois stades de développement en fonction des valeurs de $MLU_{morphèmes}$ du corpus CC (intervalles de 0.6 à partir de la valeur seuil de 1.6)[17]. Les valeurs de $MLU_{syllabes}$ dans notre corpus vont de 1.45 à 6.851, leur distribution peut être observée sur la Figure 6; les conversations sont clairement concentrées dans les intervalles 2–3 et 3–4, et les intervalles 0–2 et 4–7 rassemblent les valeurs extrêmes. Il semble donc pertinent de définir nos tranches de MLU comme suit :

Tranche	1	2	3	4
Intervalle	MLU < 2	2 ≤ MLU < 3	3 ≤ MLU < 4	4 ≤ MLU
Nombre	24	115	83	19

Tableau 5. Distribution des conversations selon les intervalles de MLU et les tranches associées

Dans la suite de ce mémoire, le symbole $\bar{m}$ sera utilisé pour désigner une valeur de tranche de MLU.

4.5.3.2 Types de segments

Notre corpus comprend 104646 segments produits par les enfants cibles et 225172 segments produits par leurs interlocuteurs (adultes ou enfants de l'entourage). Le corpus de discours adulte contient pour sa part 27275 segments. Chaque segment est associé à un type (déclaratif ou interrogatif). Les types de segments sont répartis comme suit :

	Déclaratifs	Interrogatifs	Total
CS	93.9% (98243)	6.1% (6403)	104646
CDS	63.2% (142206)	36.8% (82966)	225172
ADS	83.8% (22862)	16.2% (4413)	27275

Tableau 6. Distribution des types d'énoncés dans le CS, le CDS et l'ADS

Ces données nous montrent que les segments produits par les enfants sont majoritairement déclaratifs, tendance qui se retrouve par ailleurs en partie chez les adultes lorsqu'ils parlent entre eux, alors qu'ils utilisent beaucoup plus d'interrogatives lorsqu'ils s'adressent aux enfants. Ce déséquilibre est mis en

[17] La justification de la sélection de ces trois stades n'est cependant pas clairement exposée; les trois intervalles ne couvrent par ailleurs pas la plage complète des conversations du corpus, que Ko rapporte comme allant de 1.125 à 4.170.

évidence dans la visuation du tableau de contingence présentée en Figure 7[18] (χ^2=36382; df=2; p<0.001; φ=0.319) :

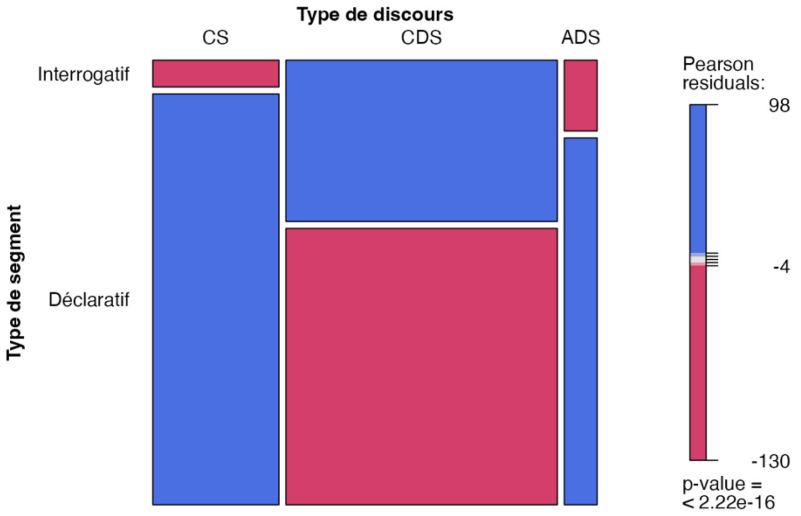

Figure 7.Tableau de contingence des types de segments dans les discours d'enfants, adressés aux enfants et entre adultes, avec résidus de Pearson[19]

On peut voir que cette tendance évolue également avec le niveau de développement langagier des enfants tel que mesuré par la MLU, et que la proportion de segments interrogatifs, si elle reste globalement largement inférieure, augmente au détriment des segments déclaratifs (χ^2=1943; df=3; p<0.001; φ=0.136) :

	Déclaratifs	Interrogatifs	Total
m̄=1	98.6% (9099)	1.4% (125)	9224
m̄=2	95.7% (56113)	4.3% (2539)	58652
m̄=3	91.5% (29285)	9.5% (3093)	32478
m̄=4	84.9% (3646)	15.1% (646)	4292

Tableau 7. Distribution des types d'énoncés selon la tranche de MLU

Le tableau de contingence avec résidus de Pearson est présenté en Figure 8 :

[18] Les tableaux de contingence ainsi que la méthode de visualisation utilisée sont décrits de manière succincte à la section 6.3.1.

[19] Diagramme généré avec le module `vcd` (Meyer *et al.*, 2006, 2024 ; Zeileis *et al.*, 2007)

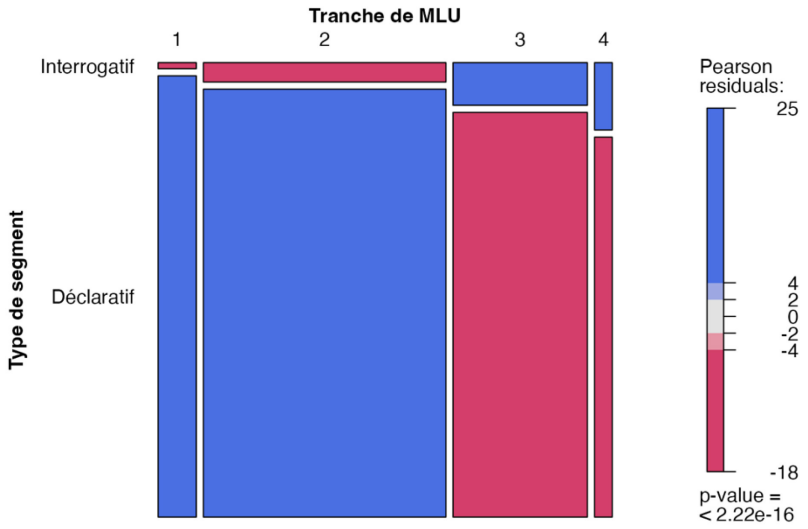

Figure 8.Distribution des types de segments pour les différentes tranches de MLU, avec résidus de Pearson

Enfin, on note également une différence significative dans l'utilisation des types de segments entre les filles et les garçons, par laquelle les filles utilisent plus d'interrogations que les garçons (χ^2=546.6; df=1; p<0.001; φ=0.072). Il est toutefois nécessaire de garder à l'esprit que les moyennes d'âge et de MLU des filles sont plus élevées que celles des garçons dans notre corpus, ce qui empêche, dans un premier temps, de tirer une conclusion sur un éventuel effet du genre de l'enfant sur les types de phrases qu'il a tendance à employer.

CHAPITRE 5

Préparation des annotations

Cette section décrit le processus complet de génération des annotations liées aux énoncés, depuis l'uniformisation des données brutes jusqu'à la production de fichiers exploitables en R.

5.1 Normalisation et unification des corpus

L'extraction du contenu des fichiers sous forme d'annotations a été effectuée en langage python. La liste des fichiers CHAT constituant un corpus, ainsi que l'accès aux informations propres à chaque conversation (données relatives aux participants) et le contenu textuel de chaque énoncé sont obtenus par l'intermédiaire de la bibliothèque logicielle PyLangAcq v. 0.16.2[20]. Le processus de traitement des données textuelles elles-mêmes a été entièrement conçu et programmé spécifiquement pour ce projet, le code source est disponible en ligne sur https://github.com/david-uqam/sfp-acquisition.

5.1.1 Segmentation, étiquetage et ponctuation

Pour notre étude, nous avons donc pris le parti de nous baser sur les unités syllabiques individuelles pour reconnaître les entités pertinentes à l'analyse des SFP et de leurs contextes d'occurrence (§3.2.5). Ainsi, l'une des étapes de notre processus de normalisation a consisté à déconstruire, dans les trois corpus, les regroupements de caractères chinois ou séquences jyutping introduits par les transcripteurs (interprétation des « mots »), en séparant toutes les unités par des marqueurs (espaces), et de les aligner ainsi sur le même plan. Les marques de ponctuation ont quant à elles été conservées : les virgules, apparaissant à l'intérieur des énoncés, ont été interprétées comme séparateurs de propositions, et ont servi à la séparation des énoncés en segments, qui constituent notre unité contextuelle de base pour l'analyse des SFP.

Les points d'interrogation ont été conservés comme marque de l'interprétation d'un énoncé comme interrogation (§3.2.6); si certaines interrogations peuvent être reconnues grâce à leur composition lexico-syntaxique ou la présence de SFP interrogatives, d'autres ne sont marquées dans le discours oral que par l'intonation (Kwok, 1984 ; Wong et Ingram, 2003) et nécessitent donc l'appréciation directe du

[20] Librairie développée par Jackson L. Lee, disponible sur https://pylangacq.org.

transcripteur (voir la section 5.4 pour une discussion à ce sujet). Les points simples et d'exclamation ont été conservés comme marqueurs déclaratifs.

5.1.2 Normalisation de la graphie des SFP

Comme il a été mentionné plus haut, le cantonais est une langue de tradition majoritairement orale (Bauer, 2018 ; Snow, 2004). Il en résulte une certaine variabilité de l'orthographe, qui n'a pas été soumise aux mêmes processus de normalisation que dans d'autres langues utilisées à l'écrit depuis plus longtemps. Cette variabilité se retrouve particulièrement dans les graphies utilisées pour retranscrire les SFP, qui ne sont que rarement utilisées dans le discours écrit.

Les exemples présentés dans ce mémoire ont déjà permis d'illustrer la variation des méthodes utilisées pour la transcription des SFP, non seulement selon les corpus et les SFP elles-mêmes, mais pour certaines également selon les transcripteurs au sein des corpus et selon les énoncés au sein des transcriptions (cf. en particulier les sections 4.2.2 et 4.3.2). Les variations s'expriment tant au niveau du choix du système d'écriture pour transcrire les SFP (romanisation ou caractères chinois) que du respect des normes du jyutping, ainsi que des caractères chinois utilisés ou même de l'utilisation d'une notation hybride. Ces différences de notation sont potentiellement liées au fait que l'association entre SFP et caractères chinois n'est pas univoque : certaines SFP peuvent être représentées par plusieurs caractères (p. ex. *gaa3* peut être transcrite comme 㗎, 嚟 ou 嫁), et certains caractères peuvent être associés à plusieurs SFP (p. ex. 喎 peut être utilisé pour transcrire *wo3*, *wo4*, *wo5*). Il existe également une forme d'allomorphie en lien avec la notation jyutping : par exemple, la particule 呢|*ne1* peut également être prononcée/transcrite *le1*, et la particule 來|*lai4* peut être prononcée/transcrite *lei4*. Des exemples de cette variabilité, extraits du corpus original CC (MHZ/100800.cha), sont présentés en (12).

La liste des différentes notations (caractères chinois et forme jyutping) associées aux SFP que nous considérons est donnée en Annexe A.

Une partie importante du traitement des données brutes a donc consisté à produire une version uniforme des corpus dans laquelle les SFP sont transcrites de manière normalisée. Dans les cas où une désambiguïsation totale n'était pas possible, les versions les plus fréquentes (selon la littérature et l'avis d'une locutrice native du cantonais) ont été priorisées.

				Remarque
(12)	a.	INV:	係 唔 係 **呀** ?	Le caractère 呀 est généralement associé à *aa3* (Matthews et Yip, 2011) ou *aa4* (Kwok, 1984).
			hai6 m4 hai6 **aa1** ?	La version jyutping fournie avec le corpus indique *aa1*, bien que ce soit le caractère 吖 qui est généralement associé à *aa1* (Kwok, 1984 ; Matthews et Yip, 2011), et que la fréquence élevée d'occurrences du caractère 呀 dans le corpus soit clairement associable à *aa3*.
	b.	MOT:	你 又 講 **&a1** .	La notation conventionnelle en jyutping est *aa1*.
			nei5 jau6 gong2 **aa1** .	
	c.	INV :	郁 &ga3 喎, 係 唔 係 **&a3** ?	Cette notation, en plus de ne pas correspondre à la norme jyutping, diffère de la notation 呀 majoritairement utilisée dans le reste du corpus.
			juk1 gaa3 wo3, hai6 m4 hai6 **aa3** ?	
	d.	MOT:	唔 飲 **呀 4** ?	Cette notation « hybride » permet de désambiguïser le caractère 呀 en y associant le ton 4 de manière univoque.
			m4 jam2 **aa4** ?	

5.2 Reconnaissance des SFP

5.2.1 Positionnement et agrégation

Pour être identifiée comme une SFP, une unité syllabique (voire une séquence de deux ou trois unités) doit se trouver dans notre liste (comme caractère(s) chinois ou en notation jyutping), et faire partie d'une séquence ininterrompue de SFP potentielles à la fin d'un segment (par définition, les SFP ne peuvent apparaître qu'en fin de segment). Toutefois, la présence du marqueur HAI_M_HAI est tolérée après une séquence de SFP (Kwok, 1984, p. 74).

De plus, comme les SFP ne peuvent constituer un segment par elles-mêmes, toute séquence de SFP doit nécessairement être précédée d'au moins une autre unité syllabique. Comme nous l'avons précisé à la section 3.2.6, nous considérons comme segment toute séquence délimitée par des signes de ponctuation.

Certaines SFP sont constituées de deux unités syllabiques (*aa1maa3*, *gaa1maa3*, *lei4gaa3*, *zaa1maa3*), dont certaines peuvent elles-mêmes être des SFP autonomes. Nous avons dans ce cas donné la priorité à la séquence dissyllabique. Ainsi, dans l'exemple suivant, le deuxième segment contient une seule SFP lei4gaa3, plutôt que deux SFP *lei4* et *gaa3*.

(13) CHI: 呢個 先生 **嚟 gaa3** . (cc-lly030809)
ni4go3 sin1saang1 **lei4gaa3**
celui-ci enseignant **SFP**
Lui, c'est l'enseignant.

5.2.2 Séquences acceptables

L'ordre dans lequel les SFP peuvent apparaître au sein d'une séquence est stable et a été décrit précisément (Matthews et Yip, 2011, p. 395). Il est ainsi possible d'assigner à chaque SFP sa position théorique dans une séquence au moyen d'un indice allant de 1 (première position, *sin1* et *tim1*) à 4 (en particulier les SFP *aa** et les SFP interrogatives). Cette information est mise à profit pour valider les séquences de SFP, en imposant que l'indice d'une SFP soit toujours strictement supérieur à celui de la SFP précédente. Cela permet de résoudre certaines ambiguïtés (p. ex. le comportement de *me1*, qui peut être particule ou mot QU). En contrepartie, cette contrainte ne permet de considérer que les SFP employées correctement selon l'ordre canonique, ce qui constitue bien entendu une limitation puisqu'il est envisageable que les enfants (et éventuellement les adultes) commettent parfois des erreurs en agençant les SFP. Il nous a toutefois paru préférable de ne pas prendre le risque de surinterpréter des situations potentiellement ambiguës.

Chaque SFP reconnue dans la séquence finale d'un segment est décrite par une annotation complète, dans laquelle sa position au sein de la séquence, le cas échéant, est clairement indiquée – un unique segment peut ainsi théoriquement engendrer la création d'une à quatre annotations. Un segment ne contenant aucune SFP sera pour sa part décrit par une unique annotation indiquant l'absence de SFP.

5.2.3 Validité d'une mention de SFP

Sur la base des différents critères mentionnés ci-dessus, nous comptabilisons chaque occurrence de SFP comme une production valide, sans évaluer la pertinence sémantico-pragmatique du contexte dans lequel la SFP est produite. Ko (2000) considère une série de limitations sur la base desquelles une production peut être exclue, en particulier le fait qu'un énoncé est une imitation de l'énoncé précédent, ou qu'il est

disfluent (sans préciser les critères d'appréciation). Nous avons choisi de considérer que toute occurrence d'un élément linguistique, même s'il s'agit d'une imitation ou d'une protoforme, mérite d'être prise en considération comme une potentielle utilisation volontaire.

5.2.4 Déroulement

Considérant les critères ci-dessus, la reconnaissance des SFP dans chaque énoncé s'effectue selon la séquence suivante :

- découpage de l'énoncé en sous-énoncés (segments) séparés par des virgules;
- pour chaque segment, suppression, le cas échéant, du marqueur final HAI_M_HAI au moyen d'une expression régulière (décrite à la section suivante);
- reconnaissance des SFP suivant l'ordre canonique au moyen d'une expression régulière construite automatiquement sur la base de la liste de SFP (l'expression régulière complète est décrite en Annexe C);
- reconnaissance des structures lexico-syntaxiques interrogatives.

Ainsi, l'énoncé présenté en (14) est découpé en deux segments. Pour chacun des segments 1 et 2, les SFP sont reconnues au moyen de notre expression régulière : le segment 1, qui ne contient pas de SFP, donne lieu à une unique annotation indiquant l'absence de SFP. Le segment 2, qui contient les deux SFP *lei4* et *aa3*, donne lieu à la production de deux annotations.

(14) CHI: 嘩 , 佢 又 跌落 **lei4** **aa3** . (hku-020600)

waa3	keoi5	jau6	dit3lok6	**lei4**	**aa3**	
oh	il	encore	tomber	**SFP**	**SFP**	

Oh, il est encore tombé.

segment 1

texte	嘩 ,
SFP	Ø

segment 2

texte	佢 又 跌 落 lei4 aa3 .
SFP-1	lei4
SFP-2	aa3

5.3 Reconnaissance des structures lexico-syntaxiques interrogatives

La présence de structures lexico-syntaxiques reconnaissables (§3.2.4) est également évaluée au moyen d'expressions régulières appliquées de manière successive à chacun des segments. Ces structures permettent de reconnaître la qualité interrogative d'un segment, et ainsi d'observer l'utilisation conjointe de SFP et de constructions interrogatives explicites telles que décrites par Wong et Ingram (2003) ou Li *et al.* (2013). Nous effectuons une reconnaissance systématique de ces structures dans notre corpus, afin de pouvoir évaluer leur degré de cooccurrence avec les SFP produites par les enfants.

La reconnaissance de ces structures syntaxiques est effectuée au moyen d'expressions régulières appliquées de manière séquentielle au texte d'un énoncé, incluant les graphies jyutping et chinoises. Par exemple, la reconnaissance du marqueur HAI_M_HAI est effectuée grâce aux expressions régulières suivantes (la syntaxe utilisée ici est celle des expressions régulières standard de python) :

```
(?:係|hai6)\s*(?:唔|m4)\s*(?:係|hai6)
(?:係|hai6)\s*(?:咪|mai6)
```

La reconnaissance d'une structure A_M_A, qui nécessite la répétition d'un terme arbitraire, est possible grâce à l'utilisation d'une référence arrière permettant d'imposer l'occurrence de la première syllabe du verbe ou adjectif utilisé avant la particule négative 唔|*m4* :

```
(\w+)\s*(?:唔|m4)\s*(\1)
```

Dans certains cas, ces structures peuvent être incompatibles avec les SFP qui les accompagnent : en effet, les SFP strictement interrogatives (p. ex. *aa4*, ou *me1*), dont la fonction est de transformer une phrase déclarative en phrase interrogative, ne peuvent être combinées avec de telles constructions. Toutefois, des exemples de telles cooccurrences ont été trouvés dans notre corpus, comme illustré dans l'exemple (15) :

(15) a. INV: 車頭 係咪 **aa4** ? (cc-hhc020503)
ce1tau4 hai6mai6 **aa4**
avant.de.la.voiture Y.N **SFP**
Est-ce que c'est à l'avant de la voiture?

b. CHI: 邊個 **aa4** ? (hku-050121)
bin1go3 **aa4**
qui **SFP**
Qui est-ce?

Il est intéressant de noter que ces cas d'erreurs sont toutefois particulièrement rares, même chez les enfants : nous en avons dénombré 44 dans le CS (sur 104648 segments, soit 0.04%), 264 dans le CDS (sur 225172 segments, soit 0.12%) et 13 dans l'ADS (sur 27275 segments, soit 0.05%). Ces cas n'ont pas fait l'objet d'un traitement ou d'une correction, et les SFP qui s'y trouvent ont été comptabilisées telles quelles, puisque de telles erreurs peuvent être le résultat d'ambiguïtés lors de la transcription, d'une mauvaise interprétation du texte ou simplement également de fautes commises par les locuteurs.

Le segment présenté en (16) donne ainsi lieu à la production d'une unique annotation pour la SFP *gaa3*, pour laquelle la structure syntaxique reconnaissable est A_M_A (la locution 有冇 correspond à la contraction de 有唔有, soit « avoir–NEG–avoir ») et le type, imposé par la structure syntaxique, est interrogatif.

(16) CHI: 有 冇 白雪公主 **gaa3** ? (cc-ltf020824)
jau5 mou5 baak6syut3gung1zyu2 **gaa3**
avoir avoir.NEG Blanche-Neige **SFP**
Est-ce qu'il y a Blanche-Neige?

texte	有 冇 白 雪 公 主 gaa3 ?
SFP	gaa3
Structure	A_M_A
Type	Interrogatif

5.4 Évaluation du type des segments

Même lorsqu'aucune structure particulière n'est identifiable, il est généralement possible d'inférer le type d'un énoncé (§3.2.6) en se basant sur les SFP qu'il contient, puisque certaines SFP ne peuvent théoriquement apparaître que dans des énoncés interrogatifs (*aa4*, *aa5*, *gaa4*, *laa4*, *me1* et *zaa4*). Lorsque l'une de ces SFP est identifiée dans un segment, celui-ci est donc systématiquement marqué comme interrogation. De tels cas sont illustrés en (17) : les deux exemples sont constitués d'une question formée avec une particule interrogative, et d'une réponse reprenant les composants de la question à l'exception de la SFP. Dans l'exemple (17.a), la particule interrogative utilisée par l'enfant est *aa4*, et le même énoncé, repris par la grand-mère sans aucune particule, est une déclaration. Dans l'exemple (17.b), c'est la particule interrogative *laa4* qui donne la forme interrogative à la phrase de l'expérimentateur, mais l'énoncé de l'enfant est déclaratif avec *laak3*.

(17) a. CHI: 唔 著 鞋 **aa4** ? (cc-ccc020624)
m4 zoek3 haai4 **aa4**
NEG porter chaussure **SFP**
Ne portes-tu pas de chaussures?
GRM: 唔 著 鞋 .
m4 zoek3 haai4
NEG porter chaussure
Je ne porte pas de chaussures.

b. INV: 去 完 廁所 **laa4** ? (cc-mhz020309)
heoi3 jyun4 ci3so2 **laa4**
aller finir toilettes **SFP**
As-tu été aux toilettes?
CHI: 去 完 廁所 **laak3** .
heoi3 jyun4 ci3so2 **laak3**
aller finir toilettes **SFP**
J'ai été aux toilettes.

Les transcripteurs de nos corpus ont également utilisé des points d'interrogation pour marquer le caractère interrogatif de certains énoncés. Cette information est conservée et mise à profit pour les énoncés dans lesquels aucun autre indice explicite (dans la transcription) ne permet de statuer; c'est le cas pour les interrogations marquées uniquement par l'intonation (Wong et Ingram, 2003). Kwok (1984, p. 70) indique que les questions uniquement intonatives ne peuvent être ponctuées d'une SFP, mais on trouve dans le corpus des énoncés marqués comme interrogatifs et ne contenant aucune structure ou SFP interrogative, comme dans (18). Dans de tels cas, nous n'avons invalidé ni les SFP présentes ni le choix du transcripteur, et avons donc marqué ces énoncés comme interrogatifs :

(18) INV: 啤啤 夠 未 **aa3** ? (hku-031026)
bi4bi1 gau3 mei6 **aa3**
bébé assez avoir.NEG **SFP**
Est-ce que le bébé n'a pas eu assez?
CHI: 你 夠 未 **aa3** ?
nei5 gau3 mei6 **aa3**
tu assez avoir.NEG **SFP**
Est-ce que tu n'as pas eu assez?

Notons toutefois que l'utilisation correcte des points d'interrogation comme marqueurs fiables d'énoncés interrogatifs a parfois dû être remise en question. Nous avons en effet trouvé un nombre important d'énoncés (4915, soit environ 2.7% des énoncés contenant au moins une SFP) dans lesquels une particule strictement interrogative est associée à un signe de ponctuation affirmatif (typiquement un point), et plus de 500 occurrences du même problème avec la structure HAI_M_HAI (sous forme *hai6 m4 hai6* ou *hai6 mai6*), strictement interrogative. De tels cas sont illustrés en (19) :

(19) a. CHI: 有 電 **gaa4** . (hku-040705)
jau5 din6 **gaa4**
il-y-a électricité **SFP**
Est-ce qu'il y a de l'électricité?
INV: 有 電 **㗎** .
jau5 din6 **gaa3**
il-y-a électricité **SFP**
Il y a de l'électricité.

b. CHI: 係咪 有 燈 **aa3** . (cc-ckt020400)
hai6mai6 jau5 dang1 **aa3**
Y-N il.y.a lumière **SFP**
Est-ce qu'il y a une/de la lumière?

Pour la particule *me1*, une telle contradiction peut dans certains cas être attribuée à une interprétation erronée du statut de SFP du fait de sa polysémie : *me1* est également un mot interrogatif signifiant « quoi », pouvant être utilisé dans le cadre d'une interrogation indirecte.

Le cas de *aa4* est également intéressant, puisque cette particule apparaît couramment juste après un nom (propre ou commun) isolé et semble être utilisée pour désigner un objet ou comme vocatif, comme dans les exemples en (20) :

(20) a. INV: 駿駿 **aa4** . (cc-ckt010522)
zeon3zeon3 **aa4**
Zeon-DIM **SFP**
Zeon.
INV: 喂, 駿駿 **aa4** .
wai3 zeon3zeon3 **aa4**
hey Zeon-DIM **SFP**
Hey, Zeon.

b. CHI: 老鼠. (cc-hhc020503)
lou5syu2
souris

Une souris.

INV: 米奇老鼠 **aa4** .
mai5kei4lou5syu2 **aa4**
Mickey.Mouse **SFP**
Mickey Mouse.

c. INV: Bernard **呀 4**, 嗰 度 有 架 車車. (cc-mhz010800)
Bernard **aa4** go2 dou6 jau5 gaa3 ce1ce1
Bernard **SFP** ça LOC il.y.a CL voiture-DIM
Bernard, il y a une voiture là-bas.

Il est toutefois possible que cette notation constitue un choix du transcripteur, puisque, bien que nombreuses, ces occurrences sont concentrées dans un nombre restreint de conversations. Nous n'avons par ailleurs pas pu trouver de confirmation de cette utilisation de aa4 dans la littérature.

Les autres SFP strictement interrogatives ne sont pour leur part pas sujettes à une telle polysémie et devraient en conséquence toujours imposer l'interprétation de l'énoncé comme interrogatif. En conséquence, nous avons fait le choix de nous fier à la transcription des unités syllabiques plutôt qu'à celle de la ponctuation, et de donner ainsi la priorité aux types d'énoncés inférés par les structures et particules. Cette décision implique que, dans ces cas conflictuels au moins, les usages des structures et particules sont considérés comme conventionnels. Il s'agit là d'un choix délibéré : en effet, bien qu'il semble inévitable que les enfants (et probablement les adultes) commettent des erreurs, la reconnaissance et l'interprétation de celles-ci sortent de la portée de ce mémoire.

5.5 Calcul de la MLU

La MLU calculée pour la conversation (selon la méthode décrite en §3.3) est intégrée dans l'annotation correspondant à chacune des SFP extraites. Pour réduire la complexité de l'ajustement de modèles statistiques, nous ajoutons également la tranche de MLU m̄, nombre entier variant de 1 (pour les valeurs de MLU inférieures à 2) à 4 (pour les valeurs de MLU supérieures ou égales à 4), comme décrit en §4.5.3.1.

5.6 Genre

Le genre des participants (locuteurs et interlocuteurs) est susceptible d'avoir une influence sur la production des SFP. Si cette information est systématiquement présente pour les enfants cibles des conversations des corpus, elle n'est malheureusement généralement pas indiquée pour leurs interlocuteurs. Il est toutefois parfois possible de l'inférer quand la personne est identifiée de manière

formelle (père, mère, marraine…). Cependant, un grand nombre de conversations (en particulier dans le HKU-70) mettent uniquement en jeu des personnes extérieures à la famille.

Les enregistrements audio étant disponibles pour le HKU, nous avons pu tenter d'estimer le genre des interlocuteurs adultes en nous basant sur leurs voix. Il convient cependant d'admettre que cette technique n'est clairement pas infaillible et doit être manipulée avec précaution. Dans le cadre d'une analyse statistique, elle nous a tout de même semblé présenter un degré de fiabilité suffisant pour justifier l'intégration de ces données dans nos modèles.

Du fait de l'absence d'enregistrements accessibles pour le CC, il ne nous a pas été possible d'obtenir le genre pour les interlocuteurs extérieurs aux familles. L'information est donc incomplète et sa puissance statistique s'en trouve amoindrie.

L'information relative au genre de l'enfant, ainsi, quand elle pouvait être inférée, au genre de l'interlocuteur, a été ajoutée à chacune des annotations.

5.7 Âge

L'âge des enfants est indiqué sous forme textuelle dans les corpus, et sert généralement d'identifiant pour les conversations dont ils sont la cible. Il est typiquement présenté comme une chaîne de caractères indiquant le nombre d'années, de mois et de jours : 01;05.22 indique donc que l'enfant a un an, cinq mois et 22 jours.

L'information n'est pas directement exploitable sous cette forme : elle est trop granulaire pour être utilisée de manière catégorielle, et ne suit pas une progression linéaire. Afin de pouvoir l'utiliser, nous avons généré une valeur numérique à partir de cette chaîne, en calculant une approximation du nombre d'années sous forme décimale. Pour ce faire, nous considérons que les mois font 30 jours et les années 365 jours, et effectuons le calcul suivant :

```
age = (365 * nb_années + 30 * nb_mois + nb_jours) / 365
```

Considérons par exemple la conversation cc-lly030422 du corpus CANCORP, mettant en jeu l'enfant LLY à l'âge de 3 ans, 4 mois et 22 jours, comme indiqué dans le nom de fichier `LLY/030422.cha` et dans l'en-tête du fichier au format CHAT, présenté en (21) :

(21) (cc-lly030422)

```
@UTF8
@Begin
@Participants:   CHI Child, INV Investigator, MOT Mother, SER servant
@Name of CHI:    Yiu Chun Li
@Birth of CHI:   14-Jun-1989
@Sex of CHI:     Female
@Tape location:  18
@Date:           5-Nov-1992
@Time duration:  16:00-17:00
@Age of CHI:     3;4.22
@Transcriber:    Kitty Ka-sinn Szeto
```

Notre calcul `(365 * 3 + 30 * 4 + 22) / 365` génère la valeur 3.389, qui sera utilisée comme approximation de l'âge auquel cette conversation a eu lieu. S'il est clair que ce calcul n'est pas fondamentalement précis, il permet toutefois d'obtenir une valeur numérique sur une échelle pseudo-continue.

Pour réduire la complexité de l'ajustement de modèles statistiques, nous utilisons également la tranche d'âge $\bar{a}$ correspondant à la valeur entière de l'âge, c'est-à-dire le nombre entier d'années (arrondi vers le bas) de l'enfant.

En plus de l'utilisation de l'âge sous forme numérique, nous produisons également une information catégorielle indiquant la classe d'âge d'un interlocuteur, soit adulte ou enfant. Cette information est susceptible de mettre à jour des différences de comportement langagier. Notons toutefois que l'indication n'est généralement pas disponible textuellement dans les corpus, mais que nous l'avons inférée des descriptions des participants. Ainsi, des participants comme CUS *cousin*, BRO *brother* ou STU *piano student of the child's mother* ont été étiquetés comme enfants, et d'autres comme AUN *aunt*, PBY *passerby* ou SER *servant* ont été catégorisés comme adultes.

5.8 Fréquences

De manière à pouvoir comparer l'utilisation des différentes SFP et observer leur évolution dans le discours des enfants, nous calculons différentes fréquences descriptives :

- fréquence absolue dans le CS : fréquence d'utilisation d'une SFP donnée dans le discours de l'enfant, soit le rapport entre le nombre d'occurrences de la SFP et le nombre total de syllabes prononcé par l'enfant dans la conversation;
- fréquence relative dans le CS : fréquence d'utilisation d'une SFP par rapport aux autres SFP dans le discours de l'enfant, soit le rapport entre le nombre d'occurrences de la SFP et le nombre total de SFP utilisées dans la conversation;

- fréquences absolue et relative dans le CDS : mêmes définitions que précédemment, mais calculées dans le discours complet des interlocuteurs pour cette conversation. Précisons que, pour des raisons de complexité, nous ne calculons pas ces valeurs pour chacun des interlocuteurs de l'enfant, mais seulement une valeur globale pour chacune des conversations.

5.9 Informations relatives à l'interlocuteur

5.9.1 Informations personnelles

Chaque annotation contient des informations concernant la personne à laquelle l'enfant s'adresse lorsqu'il parle. Comme la plupart des conversations mettent en jeu plusieurs interlocuteurs (en plus de l'enfant), et à défaut d'avoir accès à des enregistrements vidéo des conversations, nous avons fait le choix de considérer comme interlocuteur la dernière personne à avoir pris la parole avant l'énoncé considéré. Ainsi, pour chacun des énoncés, nous avons accès au rôle de l'interlocuteur (son statut social ou familial vis-à-vis de l'enfant), son genre (voir §5.6 ci-dessus) et sa classe d'âge (§5.7).

5.9.2 Informations linguistiques

Nous ajoutons également des données relatives au dernier énoncé produit : type (déclaratif ou interrogatif), présence d'une SFP, et le cas échéant la SFP produite et s'il s'agit ou non de la même SFP que celle produite par l'enfant dans l'énoncé courant.

5.10 Production des annotations

Les données décrites ci-dessus constituent la base des annotations générées à partir du corpus. Afin de permettre une analyse statistique complète du comportement langagier des enfants et de leurs interlocuteurs, le choix a été fait de produire une annotation pour chaque énoncé produit, qu'il contienne ou non une SFP; cette exhaustivité nous permet de décrire d'une part les environnements dans lesquels les SFP sont effectivement produites, mais également ceux où aucune n'est utilisée, ce qui permet d'analyser l'évolution de l'utilisation des SFP de manière globale, et d'effectuer des comparaisons entre les contextes de production et les contextes de non-production.

De plus, les annotations sont générées non seulement pour les énoncés des enfants, mais également ceux de leurs interlocuteurs. Cela permet d'effectuer des comparaisons entre le comportement langagier des enfants et le discours qui leur est adressé, et d'observer l'évolution de celui-ci par rapport au niveau de

développement langagier des enfants. Les annotations générées pour le CDS sont structurellement identiques à celles générées pour le CS.

Nous avons également généré ces annotations pour le corpus HKC (§4.4), qui ne contient que des discussions entre adultes, afin de pouvoir effectuer des comparaisons entre nos analyses du CS et du CDS et le comportement observé dans l'ADS. Dans ces annotations de l'ADS, toutes les données relatives aux enfants cibles sont bien entendu remplacées par des valeurs factices, et ne sont jamais utilisées dans les calculs ou analyses.

Chaque annotation contient donc les données suivantes :

Champ	Type	Exemples de valeur	Description
`id`	chaîne de caractères	`010522_cc_ckt_0003_1_01`	Identifiant unique de l'annotation, construit sur la base de l'identifiant unique de la conversation, l'indice de l'énoncé, l'indice du segment et l'indice de la SFP dans l'énoncé (ou 01 si l'énoncé n'en contient pas)
`cid`	chaîne de caractères	`010522_cc_ckt`	Identifiant unique de la conversation, basée sur le nom du corpus, l'âge du participant et éventuellement son nom (pour le CC)
`tid`	chaîne de caractères	`030218_cc_ltf_1668`	Identifiant unique de l'énoncé (prise de parole) produit, susceptible d'être décomposé en segments
`sid`	chaîne de caractères	`010522_cc_ckt_0019_1`	Identifiant unique du segment (sous-énoncé), soit une partie d'un énoncé terminée par une marque de ponctuation (virgule ou ponctuation finale)
`role`	valeur catégorielle	`TARGET, INVESTIGATOR, MOTHER, …`	Rôle de la personne qui produit l'énoncé. L'enfant cible de la conversation est indiqué par TARGET
`sfp_found`	valeur booléenne	`Yes, No`	Statut de la production d'une SFP dans ce segment

sfp	valeur catégorielle	aa3, lo1, me1…	SFP produite le cas échéant, ou rien si aucune SFP n'a été produite
sfp_count	valeur numérique	1, 2, 3	Nombre de SFP dans le segment courant
sfp_idx	valeur numérique	1, 2, 3	Position de la SFP dans la séquence finale du segment
sex	valeur catégorielle	MALE, FEMALE	Genre de la personne qui a produit l'énoncé, ou rien si l'information n'est pas disponible (typiquement pour les interlocuteurs)
age	valeur numérique	1.4712	Âge de l'enfant cible, exprimé en années
age_group	valeur catégorielle	1, 2, 3, 4, 5	Tranche d'âge de l'enfant cible, correspondant au nombre entier d'années (notée $\bar{a}$ dans ce mémoire)
mlu	valeur numérique	1.5753	MLU de l'enfant cible, calculée comme le nombre moyen de syllabes par segment
mlu_group	valeur catégorielle	1, 2, 3, 4	Intervalle de MLU de l'enfant cible (noté $\bar{m}$ dans ce mémoire)
sfp_rate	valeur numérique	0.137	Fréquence d'utilisation des SFP par l'enfant, soit le nombre moyen de SFP par segment
sfp_div	valeur numérique	3, 12, 25	Nombre de SFP différentes utilisées par l'enfant au cours de la conversation
syllables	valeur numérique	2, 4, 27	Nombre de syllabes du segment courant
syntax	valeur catégorielle	NONE, A_M_A, WH_HOW_MANY, …	Structure syntaxique particulière reconnue dans le segment
utt_type	valeur catégorielle	STATEMENT, INTERROGATION	Type du segment
prev_utt_type	valeur catégorielle	STATEMENT, INTERROGATION	Type du dernier segment du tour de parole précédent
role_inter	valeur catégorielle	INVESTIGATOR, MOTHER, SIBLING, …	Rôle de la personne ayant émis l'énoncé précédent (interlocuteur courant), du point de vue de l'enfant cible

sex_inter	valeur catégorielle	MALE, FEMALE, NA (not available)	Genre de l'interlocuteur courant, lorsqu'il est connu
age_range_inter	valeur catégorielle	ADULT, CHILD, NA	Classe d'âge de l'interlocuteur courant (valeur NA utilisée pour les premiers énoncés des conversations, lorsqu'il n'y a pas encore d'interlocuteur)
household	valeur booléenne	Yes, No	Appartenance de l'interlocuteur courant au foyer de l'enfant
abs_freq_sfp_cs	valeur numérique	0.01560	Fréquence absolue d'utilisation de la SFP courante dans le CS
rel_freq_sfp_cs	valeur numérique	0.1136	Fréquence relative d'utilisation de la SFP courante dans le CS
abs_freq_sfp_cds	valeur numérique	0.01030	Fréquence absolue d'utilisation de la SFP courante dans le CDS
rel_freq_sfp_cds	valeur numérique	0.07280	Fréquence relative d'utilisation de la SFP courante dans le CDS
prev_sfp_found	valeur booléenne	Yes, No	Présence d'une SFP dans le dernier segment du tour précédent (SFP précédente)
prev_sfp_same	valeur booléenne	Yes, No	Congruence entre la SFP courante et la SFP précédente
prev_sfp	valeur catégorielle	aa3, lo1, me1…	SFP précédente, ou vide si aucune n'a été produite
abs_freq_psfp_cds	valeur numérique	0.04	Fréquence absolue dans le CDS de la SFP précédente
rel_freq_psfp_cds	valeur numérique	0.29	Fréquence relative dans le CDS de la SFP précédente

Tableau 8. Liste des éléments composant une annotation pour une occurrence de SFP

Les annotations sont produites sous la forme de fichiers texte `tsv` (*tab-separated values*), qui peuvent être chargés directement dans des scripts en langage R pour effectuer des analyses statistiques. Les annotations du CS, du CDS et de l'ADS sont sauvegardées dans des fichiers séparés pour permettre le chargement et la manipulation indépendants des données. Ces fichiers contiennent respectivement 105644 (CS), 231361 (CDS) et 28336 (ADS) annotations, pour un total de 365341 observations.

CHAPITRE 6

Portrait des SFP dans notre corpus

L'analyse des annotations générées permet de dresser un premier portrait de l'utilisation des SFP dans notre corpus d'interactions enfant–adulte. Après avoir présenté dans ce chapitre un décompte des principales SFP, nous y décrirons deux aspects clefs : le taux d'utilisation et la diversité des SFP, ainsi que l'évolution de cette utilisation selon le développement langagier. Nous analyserons ensuite les corrélations simples entre la production des SFP et nos différents facteurs d'influence. Nous décrirons enfin plus en détail l'évolution parallèle des SFP dans le CS.

Bien que les deux corpus soient différents dans leur conception (transversale pour le HKU, longitudinale pour le CC), nous considérons dans un premier temps chaque conversation comme une observation unique pour pallier cette hétérogénéité. Ainsi, les critères objectifs associés à chaque conversation seront utilisés pour nos analyses, sans tenir compte du fait que plusieurs conversations peuvent cibler le même enfant à des âges différents.

6.1 Décompte des SFP

6.1.1 Présence des SFP dans les segments

Comme indiqué à la section 3.2.1, nous avons constitué une liste de 46 SFP susceptibles d'être utilisées par les enfants, et que nous avons activement identifiées dans notre corpus. Comme cela avait été observé lors des recherches antérieures (Lee *et al.*, 1996 ; Ko, 2000 ; Lim, 2018), et comme nous allons le voir dans ce chapitre selon différents axes d'observation, l'apparition des SFP se fait de manière très progressive dans le discours de enfants. De plus, les SFP n'émergent pas « aléatoirement », mais semblent suivre un rythme relativement régulier : si on observe bien entendu des variations importantes selon les individus, il reste clair que certaines particules apparaissent de manière globale plus tôt que d'autres, et sont plus utilisées que d'autres. Le Tableau 9 donne un premier aperçu du schéma d'acquisition en indiquant la fréquence d'apparition dans les énoncés des 12 particules les plus courantes (et l'absence de SFP) dans le CS (à $\bar{m}$=4) selon les tranches de MLU, ainsi que leurs décomptes dans le CDS et l'ADS.

# segments	9224	58652	32478	4292	225172	27275
SFP \ $\bar{m}$	1	2	3	4	CDS	ADS
Ø	84.64% (7807)	67.63% (39669)	52.11% (16924)	43.71% (1876)	40.09% (90268)	49.81% (13586)
aa3	13.12% (1210)	21.10% (12376)	22.42% (7280)	14.21% (610)	18.33% (41265)	12.88% (3514)
gaa3	0.28% (26)	1.43% (840)	4.57% (1485)	8.25% (354)	4.13% (9309)	5.26% (1434)
laa1	0.08% (7)	0.93% (545)	2.06% (670)	4.50% (193)	4.95% (11144)	5.57% (1518)
lo1	0%	1.06% (621)	2.76% (898)	3.05% (131)	1.60% (3597)	5.05% (1377)
ge2	0.03% (3)	0.22% (130)	0.78% (253)	2.82% (121)	0.75% (1700)	1.24% (337)
lei4gaa3	0.01% (1)	0.38% (224)	1.11% (360)	2.73% (117)	1.94% (4374)	0.45% (123)
ne1	0.04% (4)	0.72% (424)	2.13% (693)	2.73% (117)	2.31% (5193)	6.57% (1793)
laa3	0.24% (22)	0.93% (543)	1.76% (571)	2.68% (115)	3.10% (6981)	1.95% (531)
sin1	0.03% (3)	0.32 % (189)	1.04% (338)	2.63% (113)	2.04% (4594)	0.47% (128)
aa1	0.39% (36)	1.24% (730)	2.13% (692)	1.65% (71)	3.99% (8985)	1.08% (295)
aa4	0.24% (22)	0.48% (284)	0.82% (265)	1.54% (66)	4.82% (10846)	1.18% (321)
zaa3	0.04% (4)	0.14% (80)	0.91% (295)	1.30% (56)	0.16% (358)	0.44% (121)

Tableau 9. Fréquence d'apparition des 12 SFP les plus courantes dans le CS, le CDS et l'ADS

Ce tableau permet d'ores et déjà d'observer l'augmentation de l'utilisation des SFP dans le discours des enfants (leur taux de présence passe de 15.36% des segments à $\bar{m}$=1 à plus de 55% à $\bar{m}$=4). On peut également observer que, si le taux de présence de la particule aa3 est globalement très élevé (> 10%) dans tous les types de discours, l'utilisation des autres SFP débute à des taux très faible mais augmente sensiblement avec le niveau de développement langagier pour atteindre des valeurs comparables à celles dans le discours adulte.

6.1.2 Ordre d'apparition des SFP dans le discours des enfants

Notre corpus permet d'observer l'utilisation de SFP multiples dans la très grande majorité des conversations, même à des âges et des MLU encore précoces – comme on peut le voir dans l'Annexe B, certains des premiers enregistrements montrent l'utilisation de plusieurs SFP, allant jusqu'à 10 ou 11 avant 2 ans ou pour une MLU inférieure à 2 (p. ex. cc-mhz010918 : âge=01;09.18, mlu=1.9; cc-cgk011129 : âge=01;11.29, mlu=2.51). Il apparaît donc clairement que les SFP émergent très tôt dans le discours des enfants et qu'elles y occupent rapidement un rôle important.

On remarque également que certaines SFP apparaissent de manière globale plus tôt que d'autres, ou sont largement plus utilisées que d'autres. C'est en particulier le cas pour la particule *aa3*, utilisable avec tous types d'énoncés, et qui est généralement décrite comme ayant une fonction d'atténuation, visant à rendre l'énoncé « moins abrupt » (Kwok, 1984, p. 45). Elle est, dans le discours des enfants comme dans le discours des adultes, la plus largement utilisée (Winterstein *et al.*, 2023a), ce que confirme notre analyse du corpus adulte HKCancor selon laquelle *aa3* représente 26% de toutes les occurrences de SFP; elle est

également généralement l'une des premières à faire son apparition dans le discours des enfants, et est présente dans pratiquement 100% des conversations à tous les stades de développement[21]. D'autres SFP se retrouvent également chez plus de la moitié des enfants dès m̄=2, et la diversité augmente bien entendu avec le niveau de développement. La Figure 9 présente l'évolution de m̄=1 à m̄=4 de la couverture (proportion de conversations où apparaît une SFP) pour les 12 SFP les plus fréquentes (par conversation) à m̄=4.

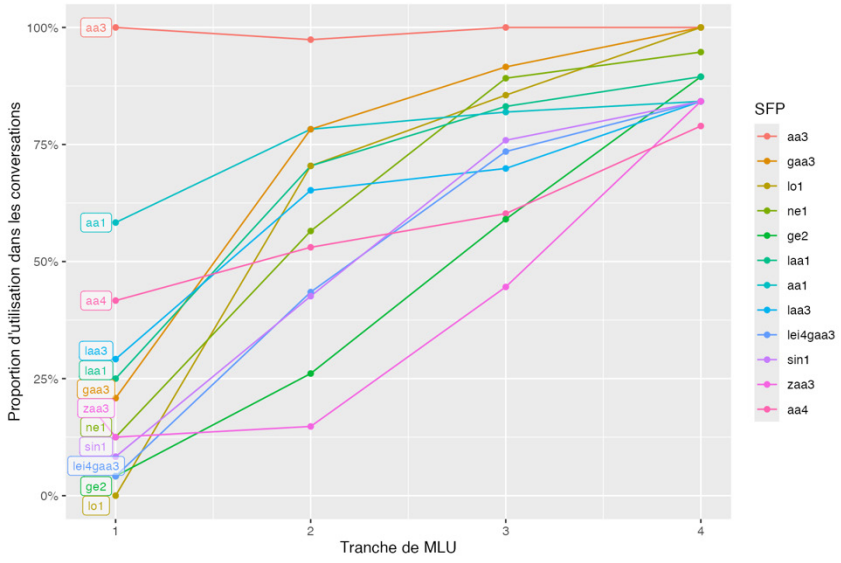

Figure 9. Évolution de la proportion d'utilisation des 12 SFP présentes dans le plus de conversations à m̄=4

Il est important de préciser que les statistiques présentées sur la Figure 9 ont été calculées sur la base des conversations individuelles, et non pas des enfants eux-mêmes – pour rappel, le CC suit 8 enfants pendant environ un an, à des intervalles allant d'une à plus de huit semaines (§4.3). Dans ce contexte, il n'est pas pertinent de rassembler toutes les conversations correspondant à une tranche de MLU pour un même enfant : cela constitue une quantité d'énoncés trop importante et couvrant une variation du niveau de langage trop large par rapport aux conversations individuelles du HKU, dans laquelle les SFP ont beaucoup plus de chance d'apparaître au moins une fois. Comparativement, chaque conversation, même dans le cadre d'un suivi, constitue un contexte d'interaction spécifique (thèmes, activités, interlocuteurs)

[21] Nous rapportons ici uniquement le fait que la particule est présente au moins une fois dans le discours de l'enfant, sans chercher à évaluer la validité de l'utilisation, et sans considération sur le fait que la particule peut être considérée comme « acquise », critère sur lequel certaines études imposent comme contrainte qu'une structure doit être produite un certain nombre de fois dans la conversation (p. ex. Bloom, 1991).

présentant ses propres incitations à utiliser différentes SFP. Nous avons donc considéré ici les conversations comme des objets d'étude indépendants, présentant chacun son propre ensemble de SFP.

La Figure 9 permet de constater que le rythme d'intégration des SFP au langage des enfants n'est pas linéaire et varie selon les SFP. En particulier, on peut constater que, si *aa3* conserve tout au long de l'enfance le statut de SFP la plus largement maîtrisée, d'autres jouent un rôle plus variable dans le discours des enfants : par exemple, *lo1*, utilisée pour signaler l'aspect évident d'un fait (Kwok, 1984), qui est l'une des SFP les mieux maîtrisées à $\bar{m}$=4, n'est utilisée par aucun enfant à $\bar{m}$=1, mais connaît une forte progression à $\bar{m}$=2, et est présente dans toutes les conversations à $\bar{m}$=4. Comparativement, *aa1* (insistance dans les questions et requêtes) et *aa4* (particule interrogative neutre) sont parmi les plus largement utilisées à $\bar{m}$=1, mais connaissent une évolution assez lente et ne sont utilisées que dans respectivement 84% et 79% des conversations à $\bar{m}$=4. Les statistiques globales de couverture des SFP sont présentées en Annexe F.

6.1.3 Fréquences d'utilisation des SFP

En parallèle de l'ordre dans lequel les SFP sont attestées dans le discours des enfants, il est pertinent de s'intéresser à la fréquence avec laquelle ceux-ci les utilisent. Cette quantité peut s'exprimer selon deux critères : la fréquence d'utilisation relative, qui correspond à la fréquence d'utilisation d'une SFP par rapport aux autres SFP, soit le rapport entre le nombre d'occurrences d'une SFP et le nombre total de SFP dans la conversation; et la fréquence d'utilisation absolue, qui correspond à la fréquence d'utilisation d'une SFP par rapport au discours global, soit le rapport entre le nombre d'occurrences d'une SFP et le nombre total d'unités syllabiques dans le discours.

Là encore, *aa3* se distingue fortement des autres SFP, puisque, présente dans 20.52% de la totalité des segments produits par les enfants (21476 occurrences / 104646 segments), elle a une fréquence relative de 54.56% (21476 occurrences / 39368 occurrences de SFP) et une fréquence absolue de 7.38% (21476 occurrences / 290989 unités syllabiques). Pour comparaison, ses fréquences relative et absolue sont respectivement 29.25% et 4.34% dans le CDS et 25.90% et 2.11% dans l'ADS. La fréquence relative de *aa3* dans le CS évolue bien entendu avec le niveau de développement langagier et au fur et à mesure que la diversité augmente, comme illustré sur la Figure 10 :

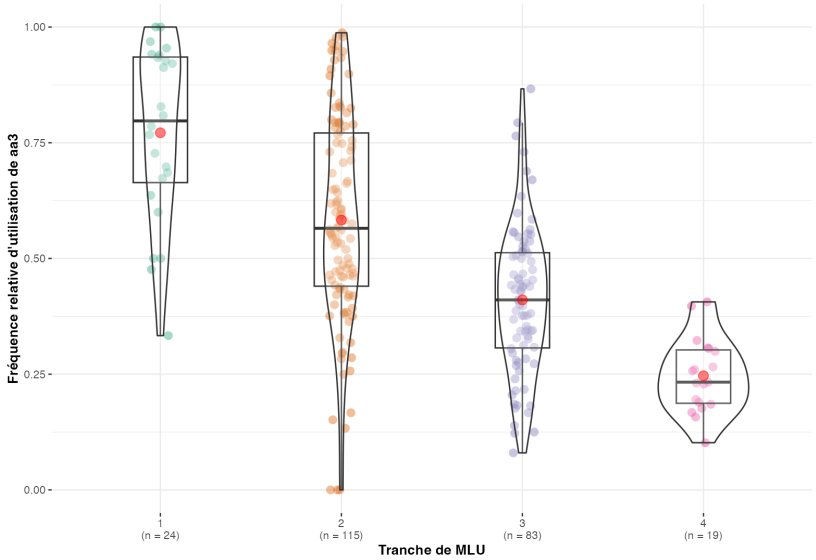

Figure 10. Fréquence relative de l'utilisation de aa3 selon la tranche de MLU

La fréquence absolue d'utilisation de *aa3* varie quant à elle beaucoup moins (Figure 11) :

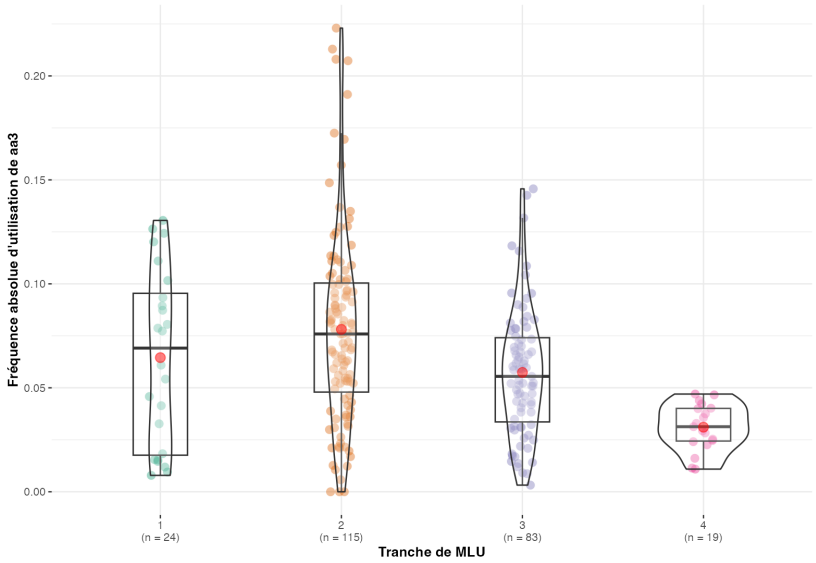

Figure 11. Fréquence absolue d'utilisation de aa3 selon la tranche de MLU

Notons que les fréquences relative (24.66%) et absolue (3.11%) d'utilisation de *aa3* à $\bar{m}$=4 sont très proches des valeurs observées dans l'ADS.

L'omniprésence de *aa3* dans le discours des enfants rend difficile l'appréciation visuelle de l'évolution des autres particules, comme illustré sur la Figure 12. En la retirant du graphique, on obtient un meilleur aperçu du comportement des autres SFP (Figure 13) :

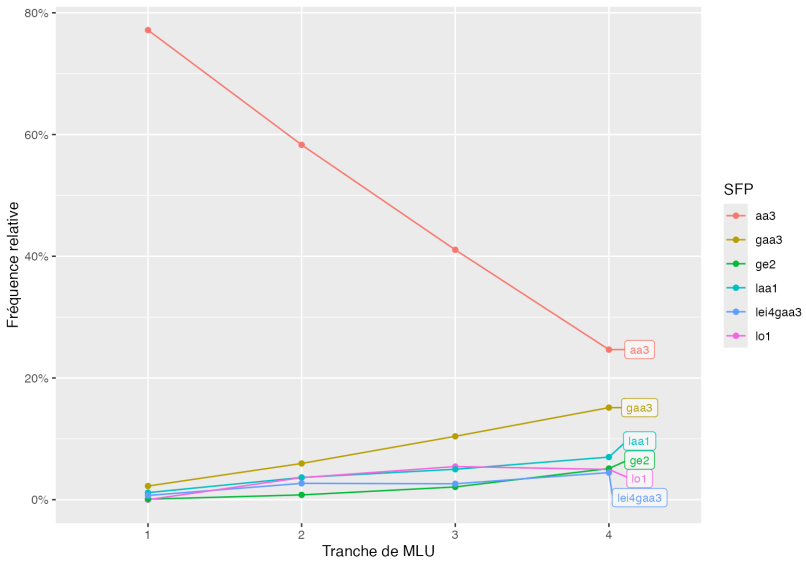

Figure 12. Évolution des fréquences relatives d'utilisation des 6 SFP les plus employées à $\bar{m}$=4

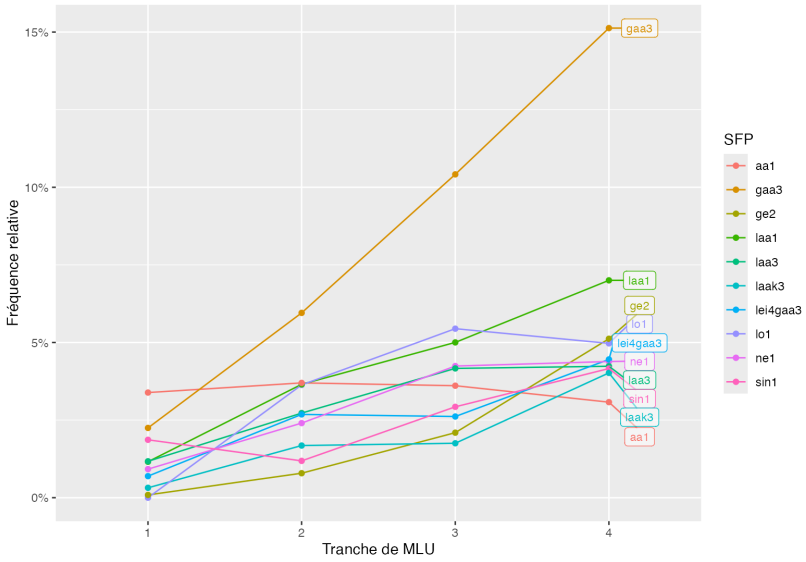

Figure 13. Fréquences relatives d'utilisation des 10 SFP les plus employées à $\bar{m}$=4, excluant *aa3*

Si la Figure 12 permet de visualiser la chute rapide de *aa3* au fur et à mesure que l'éventail de SFP s'élargit avec le niveau de développement langagier, la Figure 13 permet quant à elle de mieux visualiser certaines tendances comme la montée rapide de gaa3 ou laa1, ou encore la stagnation de *aa1*.

Les fréquences absolues d'utilisation présentent un schéma globalement similaire, avec toutefois quelques variations illustrées sur les Figures 14 et 15 :

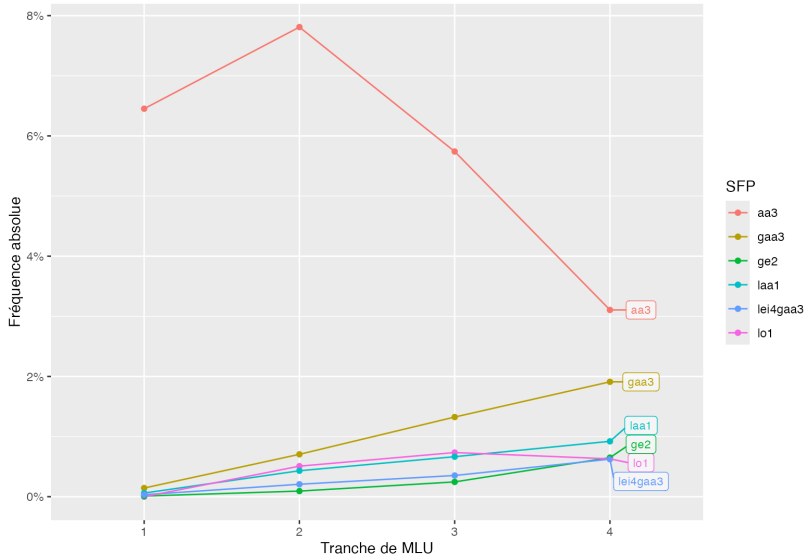

Figure 14. Fréquences absolues d'utilisation des 6 SFP les plus employées à $\bar{m}$=4

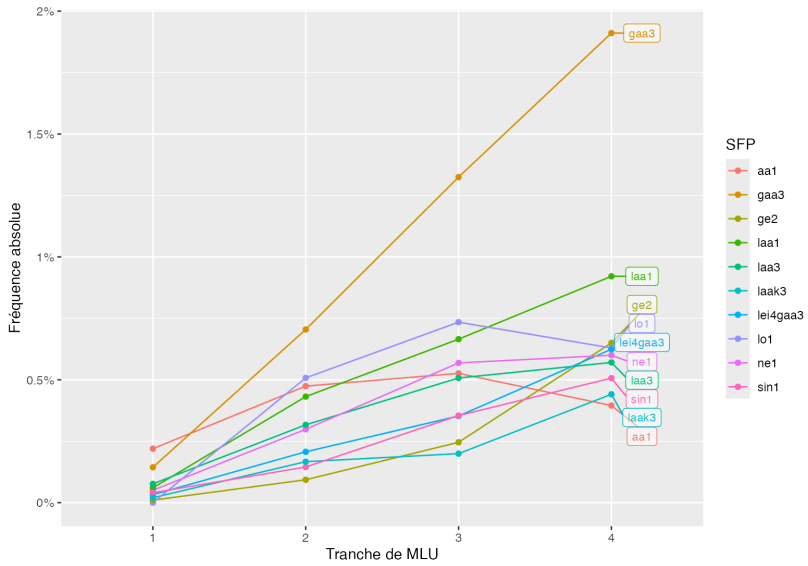

Figure 15. Fréquences absolues d'utilisation des 10 SFP les plus employées à $\bar{m}$=4, excluant aa3

On note sur la Figure 14, incluant *aa3*, que la baisse d'utilisation de celle-ci au niveau global du discours est plus tardive, c'est-à-dire que, si son utilisation par rapport à celle d'autres SFP a tendance à baisser, l'explosion des SFP dans le discours des enfants lui confère tout de même une part prépondérante. La Figure 15 permet quant à elle de mieux visualiser l'évolution du comportement des SFP autres que *aa3*.

6.2 Données relatives à l'utilisation globale des SFP dans une conversation

Dans le cadre du processus de génération des annotations décrivant les contextes d'occurrence des SFP, des statistiques relatives à l'utilisation globale des SFP dans chacune des conversations sont calculées afin d'obtenir un aperçu du comportement individuel de chacun des enfants. En particulier, le taux d'utilisation ainsi que la diversité des SFP sont calculés.

6.2.1 Taux d'utilisation des SFP

Le taux d'utilisation correspond au nombre moyen de SFP par segment produit par un enfant au cours d'une conversation, il constitue donc une constante pour chacune des conversations. Ce taux peut théoriquement varier de 0 (aucune SFP utilisée au cours d'une conversation) à des valeurs supérieures à 1, puisque plusieurs SFP sont susceptibles d'être utilisées ensemble à la fin d'un segment. En pratique, les taux d'utilisation dans notre corpus vont de 0.0066 (hku-050214, mlu=2.46) à 0.7519 (hku-031026, mlu=4.62). Les taux d'utilisation associés à chaque conversation sont présentés en Annexe B. Nous pouvons comparer ces valeurs au taux d'utilisation global de 0.6265 dans le CDS et 0.5401 dans l'ADS, bien que ces taux globaux soient peu représentatifs du fait qu'ils sont calculés sur la base des corpus entiers, sans distinguer les locuteurs entre eux.

6.2.2 Diversité des SFP

La diversité des SFP dans une conversation correspond au nombre de SFP différentes utilisées par un interlocuteur, elle constitue donc également une donnée constante pour chacune des conversations. S'il est clair que l'ensemble des SFP produites au cours d'une conversation est susceptible de ne représenter qu'une partie du lexique à la disposition de l'enfant, sa richesse constitue toutefois une bonne approximation de la variété de valeurs sémantiques ou pragmatiques que l'enfant est en mesure d'exprimer par le biais des SFP. Dans notre corpus, les diversités s'échelonnent de 1 à 27[22], elles sont présentées en Annexe B.

[22] La comparaison de ces valeurs à leurs équivalents chez les adultes (CDS ou ADS) requerrait un décompte par locuteur pour chacune des conversations, puisque la cumulation de SFP produites par différents locuteurs ne donnerait pas d'indication sur les comportements langagiers. Nous n'avons pas étendu notre programme pour calculer ces informations, qui seraient néanmoins intéressantes pour l'étude des facteurs influençant l'utilisation des SFP dans le langage adulte.

6.2.3 Corrélations

L'étude des corrélations de Pearson entre âge, MLU, taux d'utilisation et diversité permet de confirmer que des liens existent entre ces mesures :

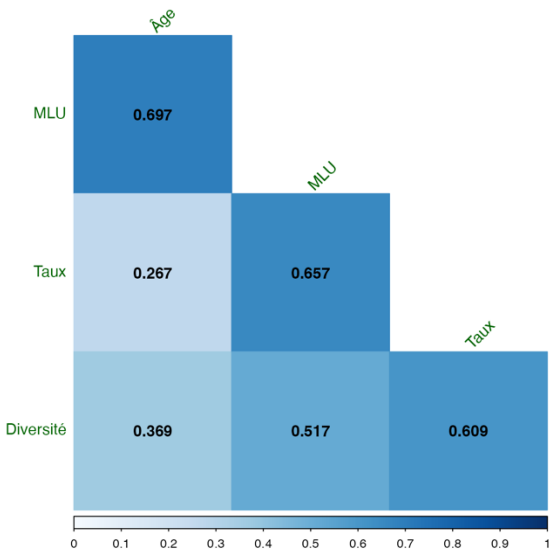

Figure 16. Table de corrélation de Pearson entre âge, MLU, taux d'utilisation et diversité[23]

Ces valeurs confirment d'une part que le taux d'utilisation des SFP ainsi que leur diversité sont fortement corrélés avec le niveau de développement langagier exprimé par la MLU ($r_{\text{MLU–taux}}$=0.657, $r_{\text{MLU–div}}$=0.517), et d'autre part qu'ils ne sont en contrepartie que faiblement corrélés avec l'âge des enfants ($r_{\text{âge–taux}}$=0.267, $r_{\text{âge–div}}$=0.369). Cela suggère que l'âge, malgré sa corrélation également forte avec la MLU, n'est pas un prédicteur fiable pour les aspects de la compétence langagière liés à l'utilisation des SFP en général.

L'évolution du taux d'utilisation et de la diversité des SFP selon les tranches de MLU peut être observée sur les diagrammes suivants[24] :

[23] Diagramme généré avec le module `corrplot` (Wei et Simko, 2024).

[24] Diagrammes réalisés avec le module `ggstatsplot` (Patil, 2021).

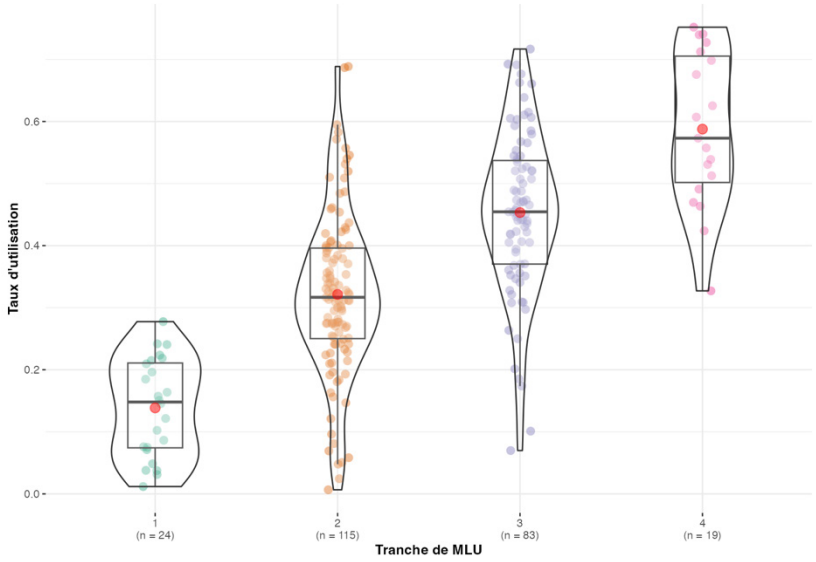

Figure 17. Taux d'utilisation des SFP par tranche de MLU

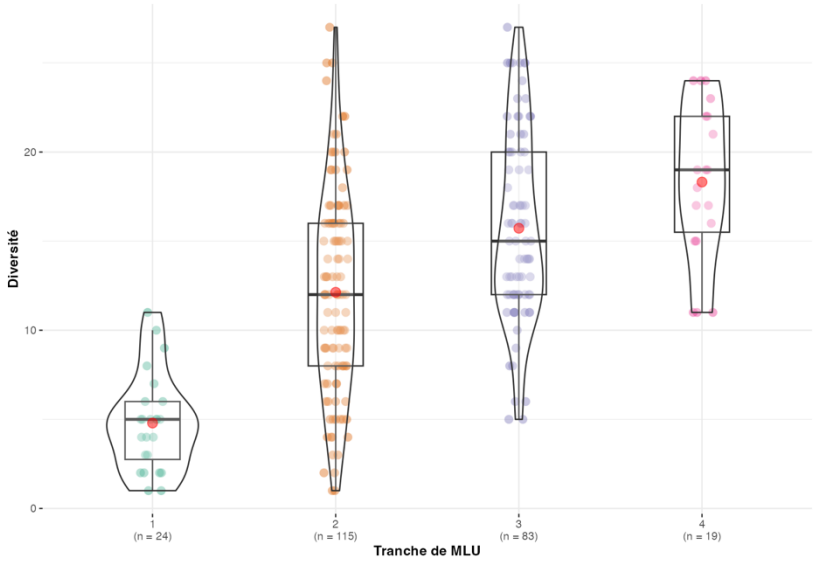

Figure 18. Diversité de SFP par tranche de MLU

Les diagrammes en Figure 17 et Figure 18 illustrent clairement l'augmentation globale du taux d'utilisation et de la diversité selon la tranche de MLU.

6.2.4 Influence du genre sur le taux d'utilisation et la diversité des SFP

Comme nous l'avons vu à la section 4.5.3.1, l'effet significatif du genre de l'enfant sur le rythme de développement langagier est confirmé dans notre corpus – à âge égal, les filles ont une MLU globalement plus élevée que les garçons, ce qui est interprété comme un développement plus précoce. Cette

observation explique au moins en partie le fait que le taux d'utilisation, de même que la diversité, sont significativement différents chez les filles et les garçons dans notre corpus (taux d'utilisation : μ_F=0.417, μ_M=0.325; t=4.3795; df=238.99; p<0.001; diversité : μ_F=14.419, μ_M=11.903; t=3.2052; df=238.59; p=0.001534). Ces résultats ne permettent toutefois pas d'affirmer que le genre à lui seul a un effet sur la propension à utiliser des SFP.

L'influence du genre sur le taux d'utilisation (*sfp rate*) des SFP peut être estimée par la comparaison de modèles de régression linéaire. Dans un premier modèle *sfp_rate~mlu* [R^2=0.432; $F(1, 239)$=181.4, p<0.001], le coefficient *mlu* est significatif (p<0.001). En ajoutant le genre (*sex*) comme facteur, on obtient un deuxième modèle *sfp_rate~mlu+sex* [R^2=0.432; $F(2, 238)$=90.59, p<0.001], dans lequel le coefficient *mlu* est toujours significatif, mais le coefficient *sex* ne l'est pas (p=0.590). Ce résultat est confirmé lors de la comparaison par anova, qui indique que les deux modèles ne diffèrent pas significativement (p=0.590). La comparaison des taux d'utilisation pour les filles et les garçons par tranche de MLU permet de confirmer visuellement que les deux populations sont sensiblement similaires :

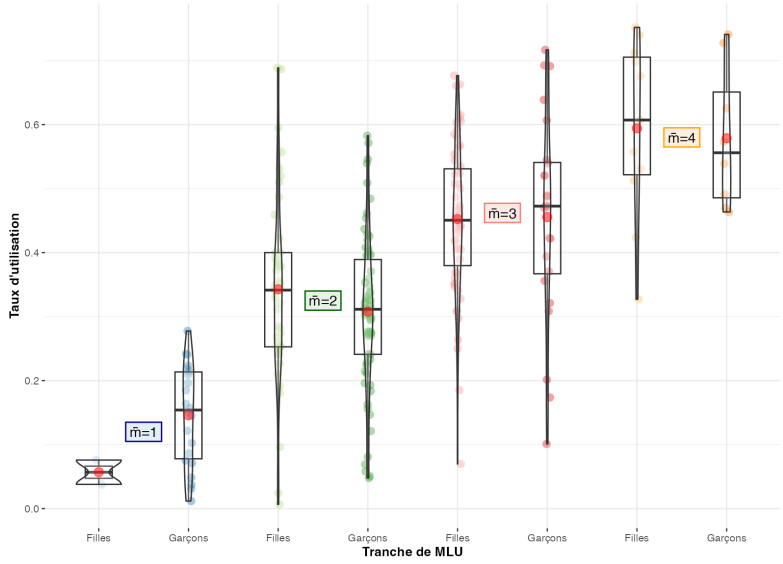

Figure 19. Comparaison des taux d'utilisation des SFP entre filles et garçons par tranche de MLU

Des résultats semblables sont obtenus pour la diversité, quoique les effets observés soient moins marqués.

Cette différence d'âge et surtout de MLU entre les populations de filles et de garçons dans notre corpus devra être prise en compte lors des analyses suivantes, pour éviter de confondre un effet de la MLU avec un effet du genre.

6.3 Analyse de la production des SFP

Les annotations générées décrivent, pour chaque occurrence de SFP, l'environnement dans lequel elle se trouve (données globales de la conversation ainsi que données spécifiques au segment lui-même). Les segments dans lesquels aucune SFP n'apparaît sont également décrits par des annotations similaires. À l'aide de ces données, il est possible d'analyser les environnements langagiers dans lesquels les SFP sont susceptibles d'être plus ou moins produites. Les sections suivantes décrivent les effets de différents facteurs sur la production de SFP vue comme une variable binaire (vrai ou faux).

6.3.1 Tableaux de contingence et résidus de Pearson

Tout au long de cette section, nous illustrerons les relations entre les différentes variables considérées au moyen de représentations graphiques de tableaux de contingence. Ces tableaux permettent de détecter les irrégularités dans la distribution des variables, signes de potentielles dépendances entre les valeurs prises par celles-ci.

Un tableau de contingence décrit la répartition numérique des valeurs prises par deux variables[25]. Ainsi, on peut décrire la répartition des segments produits par les enfants selon leur genre (garçon/fille) et le corpus (CC/HKU) dans le Tableau 10 :

genre \ corpus	CC	HKU
fille	35608	5815
garçon	56603	6620

Tableau 10. Tableau de contingence de la répartition des segments produits selon le genre et le corpus

La distribution des combinaisons des valeurs des variables peut être visualisée graphiquement en associant à chaque paire de valeur une surface correspondant à la proportion des observations qu'elle représente, comme illustré à la Figure 20[26] :

[25] Bien qu'il soit possible de créer des tableaux de contingence mettant en relation un nombre plus élevé de variables, nous n'utilisons dans ce projet que des tableaux à deux dimensions.

[26] Les commandes R pour la création du tableau de contingence ainsi que la génération des visualisations graphiques sont données en annexe H.

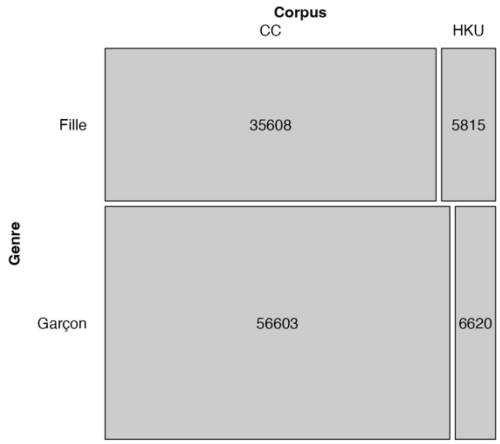

Figure 20. Visualisation du tableau de contingence du nombre de segment selon le genre et le corpus

L'indépendance des deux variables, qui correspond à l'hypothèse nulle, est vérifiée au moyen d'un test d'indépendance dit du *khi carré*. Pour notre table, le test du khi carré indique que les deux variables ne sont pas indépendantes (χ^2=304.14; df=1; p<0.001; φ=0.054) et que leurs valeurs ne sont pas distribuées de manière homogène. L'ampleur de la différence entre les valeurs observées et les valeurs attendues (correspondant à une répartition homogène des valeurs lorsque celles-ci sont indépendantes) est exprimée par les résidus de Pearson. Ces résidus peuvent être visualisés sur le même diagramme que précédemment en indiquant la déviation pour chaque paire de valeurs, comme illustré à la Figure 21 :

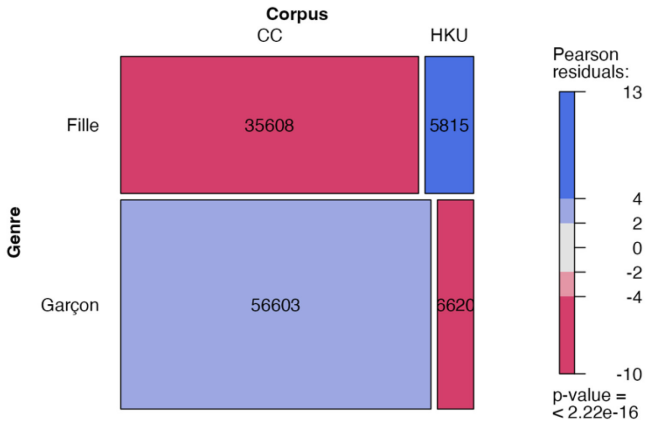

Figure 21. Visualisation du tableau de contingence de la répartition des segments avec résidus de Pearson

Les couleurs associées aux résidus de Pearson permettent une interprétation rapide des données, indiquant entre autres que les filles sont significativement plus prolixes dans le HKU que dans le CC, et significativement plus prolixes que les garçons dans le HKU.

Dans la suite du chapitre, les valeurs numériques correspondant à chaque association de valeurs seront omises des représentations visuelles des tableaux de contingence pour en améliorer la lisibilité.

6.3.2 Facteurs individuels

6.3.2.1 Effet de la MLU sur la production de SFP

Comme il a été montré à la section précédente, la MLU est fortement corrélée avec le taux d'utilisation des SFP, ce qui est bien sûr pratiquement équivalent à dire que la probabilité de présence d'une SFP à la fin d'un segment augmente avec la MLU (la différence entre ces deux énoncés est que le taux d'utilisation est calculé en considérant l'ensemble des SFP produites lors d'une conversation, sachant que plusieurs peuvent apparaître à la fin d'un même segment, alors que la probabilité de présence correspond à la probabilité qu'au moins une SFP soit présente à la fin d'un segment). La variation de la probabilité de présence selon les tranches de MLU est illustrée sur la Figure 22 ; la significativité de ces différences est confirmée par un test d'indépendance (χ^2=4743.9; df=3; p<0.001; φ=0.213).

La transition constatée entre $\bar{m}$=2 et $\bar{m}$=3, que nous avions évoquée plus tôt, se retrouve ici clairement exposée : le diagramme montre le basculement significatif qui s'opère entre ces deux valeurs. Cette observation permet d'envisager l'utilisation ultérieure d'une variable binaire `mlu_bin` qui portera une grande partie de la puissance prédictive de `mlu_group` mais qui permettra de simplifier les modèles statistiques.

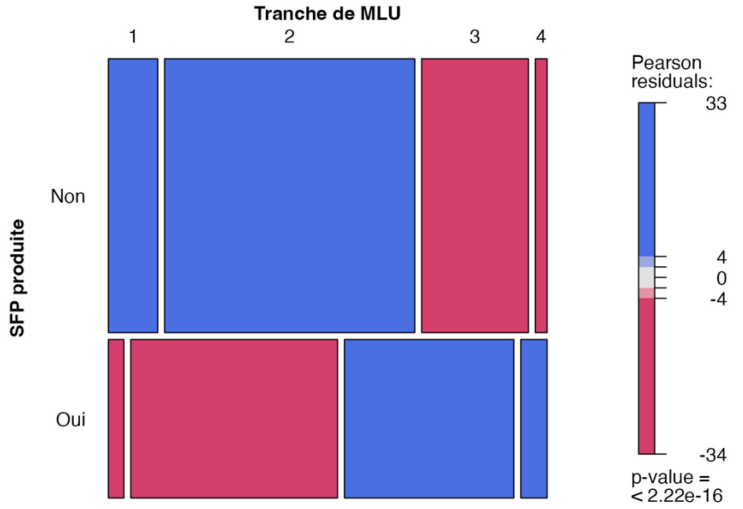

Figure 22. Tableau de contingence de tranches de MLU et production de SFP

6.3.2.2 Âge

Nous avons déjà pu constater que l'âge biologique était un moins bon prédicteur que la MLU pour les aspects liés aux SFP en général. Bien qu'il ne soit pas exclu que l'âge joue un rôle dans le processus d'acquisition de certaines SFP spécifiques en lien avec leur dimension pragmatique, nous nous limitons dans ce chapitre à une analyse de haut niveau. Nous reviendrons sur les effets de l'âge à la section 7.2.2.2.

6.3.2.3 Genre

Tout comme pour l'âge, le genre ne semble pas avoir d'effet direct sur la production des SFP en général en lien avec la MLU. Toutefois, il n'est pas exclu que le genre commence à influencer le choix des SFP utilisées pendant la phase d'acquisition. Nous reviendrons sur cet aspect au chapitre 8.

6.3.3 Facteurs linguistiques

6.3.3.1 Effet du type de segment sur la production de SFP

Du fait de la relation étroite que les SFP entretiennent avec les types de phrases (en particulier le fait que certaines interrogations sont formées spécifiquement par l'ajout de SFP exclusivement interrogatives), il paraît inévitable que la présence d'une SFP à la fin d'un segment dépende en partie du type de l'énoncé. Cette hypothèse est étayée par nos données, qui montrent que, si les enfants produisent, dans notre corpus, beaucoup moins de segments interrogatifs (N_{int}=6403) que de segments déclaratifs ($N_{\text{décl}}$=98243), l'utilisation de SFP y est beaucoup plus fréquente (74.5% vs 25.5%). Cette tendance est confirmée par le test d'indépendance (χ^2=4203.1; df=1; p<0.001; φ=0.200) ainsi que l'analyse des résidus :

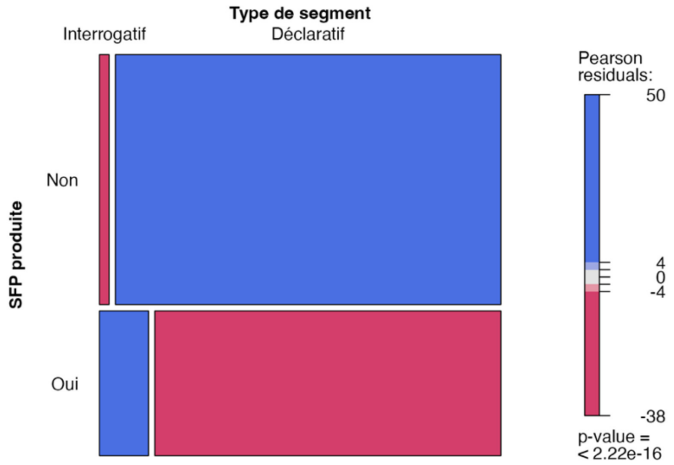

Figure 23. Tableau de contingence des types de segments et production de SFP

Il convient toutefois de noter que ce déséquilibre se retrouve, sans surprise, dans les analyses du CDS (χ^2=17358; df=1; p<0.001; φ=0.278) et de l'ADS (χ^2=618.71; df=1; p<0.001; φ=0.151). On retrouve d'ailleurs des proportions similaires de production de SFP pour les segments interrogatifs (CDS : 77.7%; ADS : 67.3%).

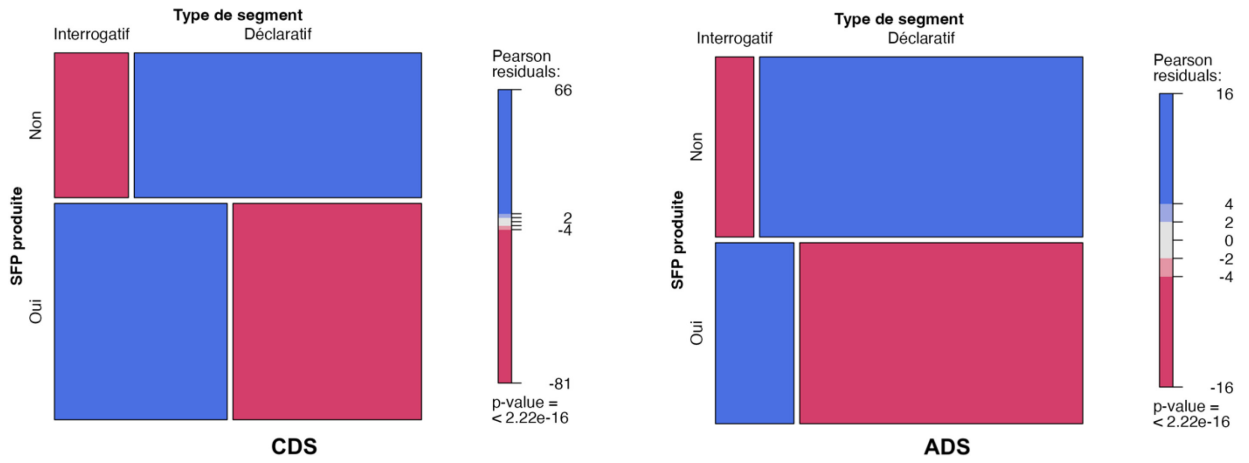

Figure 24. Tableaux de contingence de la production de SFP et des types de segments, dans le CDS et dans l'ADS

On observe comme précédemment que les interrogatives sont plus fréquentes dans le CDS que dans l'ADS. Cependant, la tendance à utiliser plus souvent des SFP dans les segments interrogatifs que dans les segments déclaratifs se retrouve dans tous les discours.

6.3.3.2 Structures syntaxiques

Nous avons présenté à la section 3.2.4 le rapport qui unit certaines structures syntaxiques aux types de segments, en particulier les interrogations, qui peuvent être formées par l'utilisation des constructions A_M_A et HAI_M_HAI, ou par l'utilisation de mots QU. L'effet de l'acquisition de ces structures remarquables sur l'utilisation des SFP est difficile à analyser de manière globale. En effet, de même que nous avons pu constater que la fréquence de production de SFP varie avec le niveau de développement langagier de l'enfant, la production de ces structures syntaxiques, dont certaines sont complexes (p. ex. HAI_M_HAI ou A_M_A), entretient un lien avec la compétence langagière. De plus, les structures syntaxiques que nous considérons sont toutes en lien avec la production de segments interrogatifs, pour lesquels le choix de SFP joue également un rôle prépondérant. Nous allons donc nous concentrer ici uniquement sur les segments interrogatifs produits par les enfants (N=6403). La Figure 25 montre

l'évolution de l'utilisation des différentes structures syntaxiques dans les segments interrogatifs par tranche de MLU.

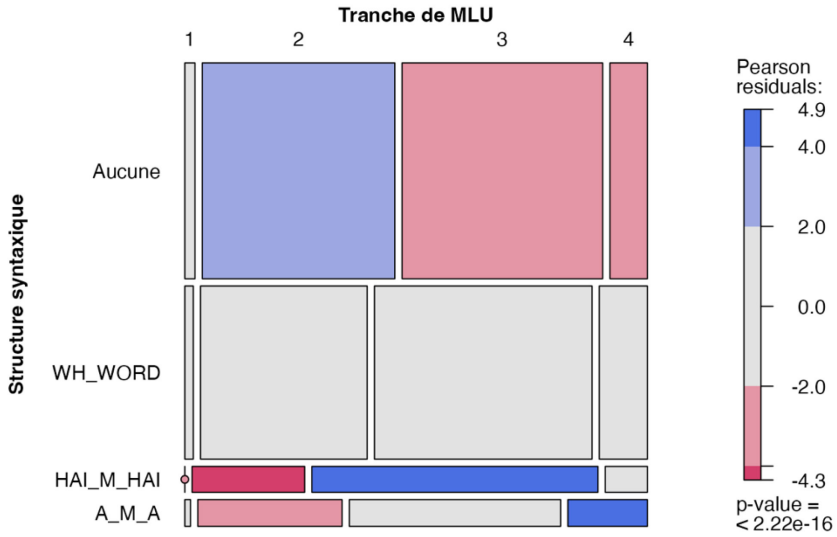

Figure 25. Tableau de contingence de l'utilisation des structures syntaxiques dans les segments interrogatifs par tranche de MLU

Ce diagramme montre que les interrogations sont tout d'abord largement exprimées sans l'aide des constructions syntaxiques, ce qui est bien entendu tout à fait logique du fait que les constructions en question représentent au moins une unité syllabique (et généralement plusieurs) au sein d'un énoncé, et qu'un enfant dont la MLU est trop réduite produit généralement des énoncés trop courts. Par la suite, les moyens utilisés pour formuler une question se diversifient, et on constate tout d'abord une utilisation accrue des mots QU, suivie par l'émergence de la structure multimorphémique figée HAI_M_HAI et enfin de la structure plus complexe (avec réduplication) A_M_A. Ces résultats sont cohérents avec les observations de Wong et Ingram (2003) sur l'ordre d'acquisition des différents types de questions.

De manière globale (toutes MLU confondues), les données extraites de notre corpus, présentées sur la Figure 26, semblent indiquer que l'utilisation des structures syntaxiques n'a globalement que peu d'impact sur l'utilisation des SFP :

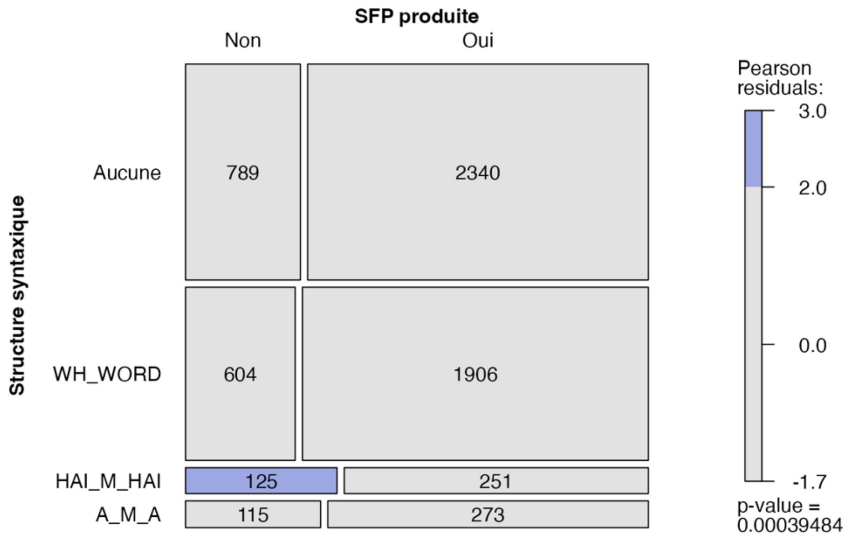

Figure 26. Tableau de contingence entre structures syntaxiques et production de SFP

Ce tableau nous confirme que les interrogations sont plus fréquemment construites avec des SFP, quelle que soit la structure syntaxique associée ou lorsqu'aucune n'est utilisée (ce qui correspond soit à l'utilisation d'une SFP strictement interrogative, soit à la formation d'une question par intonation). Mais, mis à part les interrogations construites avec la construction HAI_M_HAI, qui semblent légèrement inhiber l'utilisation des SFP, les autres constructions ont un comportement largement homogène.

Toutefois, en observant les associations entre les types de constructions interrogatives et les SFP selon les tranches de MLU, on peut constater que les moyens mis en œuvre par les enfants évoluent :

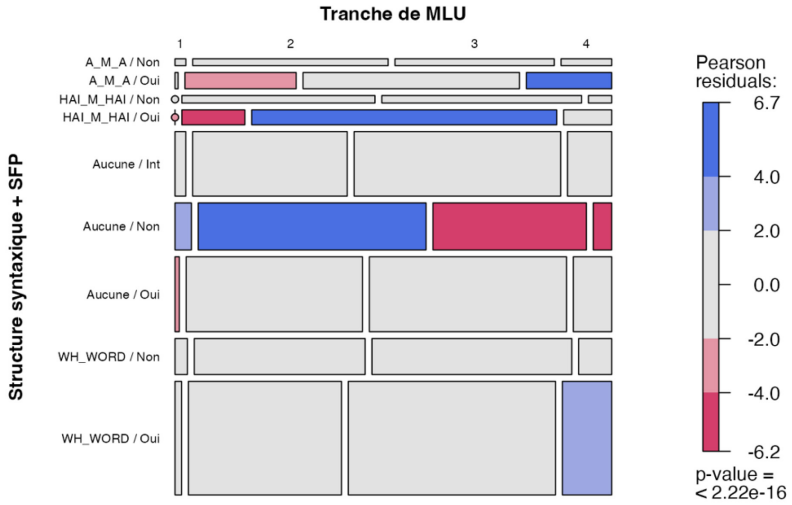

Figure 27. Tableau de contingence des combinaisons de structures syntaxiques avec SFP selon les tranches de MLU[27]

L'axe vertical sur la Figure 27 montre les combinaisons possibles entre les structures syntaxiques habituelles et une SFP générique (« Oui »), une SFP spécifiquement interrogative (« Int ») ou aucune SFP (« Non »). La tendance qui apparaît sur ce diagramme est que les enfants construisent d'abord ($\bar{m}$=2) les interrogations sans l'aide de structures syntaxiques particulières (« Aucune »), et qu'ils utilisent pour cela majoritairement l'intonation seule (« Aucune / Non ») ou des SFP génériques (« Aucune / Oui »)[28]. Ils utilisent également des mots interrogatifs simples (mots QU), déjà accompagnés d'une SFP. Les SFP semblent sous-utilisées à $\bar{m}\leq 2$ dans les interrogations utilisant des structures remarquables, en partie par manque d'espace. À $\bar{m}$=3, les structures plus complexes (A_M_A, HAI_M_HAI) sont plus fréquemment utilisées, au détriment des interrogations sans structures ni SFP. L'utilisation des SFP devient majoritaire pour l'ensemble des constructions interrogatives.

[27] Nous ignorons ici à dessein les conflits pouvant exister entre l'utilisation de SFP spécifiquement interrogatives et l'utilisation de structures interrogatives, théoriquement incompatibles. Nous avons dénombré 44 conflits de ce type dans le CS, 266 dans le CDS et 7 dans l'ADS. D'après ces chiffres et la qualité globale des corpus, il est envisageable que ces conflits soient liés à des erreurs ou ambiguïtés dans la transcription, ou qu'ils correspondent à des erreurs commises par les locuteurs.

[28] Kwok (1984) indique que les questions strictement intonatives ne peuvent être ponctuées d'une SFP. En revanche, certaines SFP, qui ne sont pas spécifiquement interrogatives, peuvent également être utilisées pour former des questions à partir de déclaratives; il s'agit de ge2, ne1, le5, aa1maa3 et laa3wo3. En pratique, on retrouve dans notre corpus également d'autres SFP non interrogatives dans des énoncés marqués comme questions (p. ex. aa3, ge3 ou lei4). Nous n'avons pas analysé ces cas de manière plus approfondie.

On peut comparer ces tendances avec les associations entre SFP et structures syntaxiques dans le CDS et dans l'ADS :

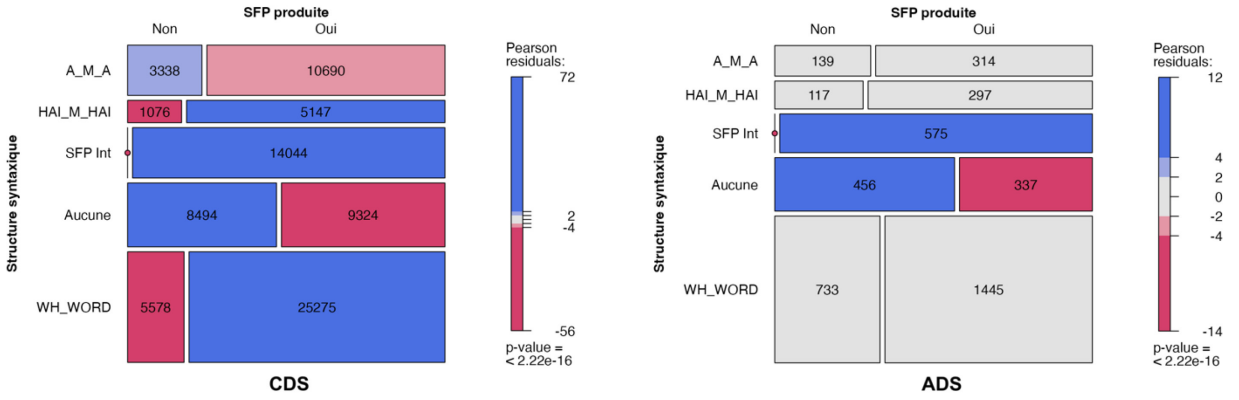

Figure 28. Tableaux de contingence des associations entre structures syntaxiques et production de SFP dans le CDS et l'ADS

On constate que les contrastes sont beaucoup plus marqués dans le CDS, alors que l'utilisation des SFP dans l'ADS est plus homogène. De manière globale, les adultes utilisent plus fréquemment des SFP dans les constructions interrogatives, ce que l'on retrouve dans l'évolution du comportement des enfants avec une utilisation de plus en plus fréquente des SFP en combinaison avec des structures syntaxiques.

6.3.3.3 Phono-prosodie

Dans la section 1.6.3.2, nous avions évoqué la possibilité que le nombre de syllabes d'un énoncé puisse jouer un rôle dans la production des SFP, pour des raisons de parité. Les segments produits par les enfants dans notre corpus font de 1 à 27 syllabes, leur distribution est présentée à la Figure 29 :

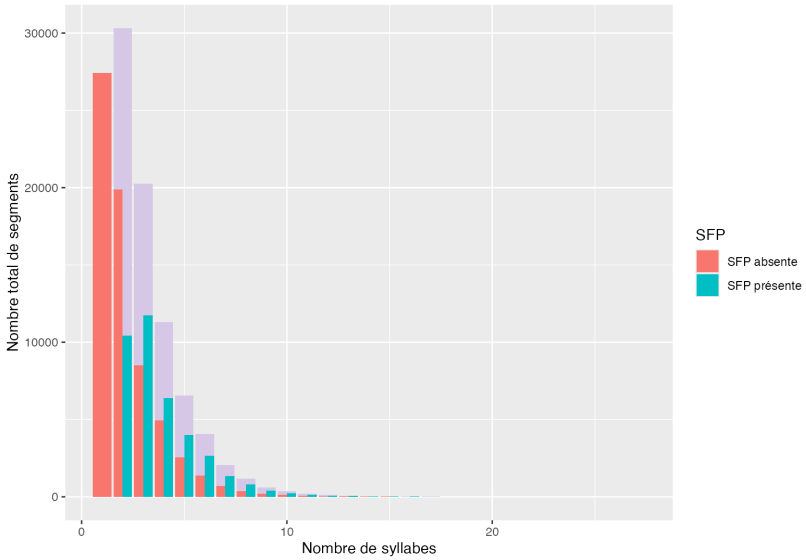

Figure 29. Distribution de la longueur des segments en nombre de syllabes

Cette distribution montre tout d'abord, sans surprise, que la quasi-totalité des segments font 10 syllabes ou moins (99.5%). D'autre part, excepté les segments de longueur 1, *de facto* trop courts pour accueillir une SFP (une SFP ne peut constituer un énoncé à elle seule), et ceux de longueur 2 (voir ci-dessous), les autres longueurs de segments semblent présenter un profil assez homogène. Cette tendance est confirmée par le tableau de contingence en Figure 30 :

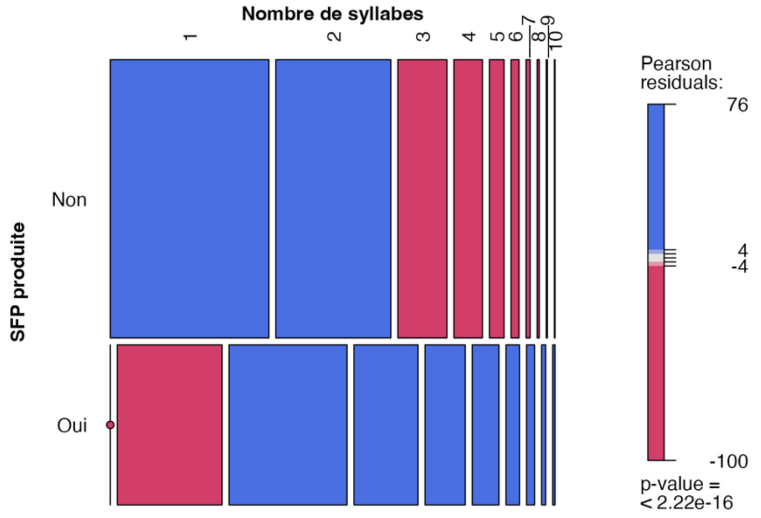

Figure 30. Tableau de contingence de l'association de la production de SFP et du nombre de syllabes

Il apparaît ainsi que la parité du nombre de syllabes ne joue aucun rôle dans la tendance à utiliser une SFP à la fin d'un énoncé (une telle tendance se serait traduite par une alternance de résidus positifs et négatifs selon la parité, ce qui n'apparaît ni dans ce graphique ni dans des représentations focalisées sur les énoncés plurisyllabiques). Le nombre de syllabes joue pour sa part un rôle évident (n_{syll}≤2 ou n_{syll}>2), mais son impact pour les valeurs supérieures n'est pas clair, au-delà du fait qu'un nombre plus élevé de syllabes correspond à une MLU plus élevée, qui est elle-même fortement corrélée à la production de SFP.

Le cas des segments de longueur 2 est particulier puisque ceux-ci se différencient significativement des segments de longueur plus importante, sans toutefois connaître la limitation des segments de longueur 1. L'analyse comparée de la production des SFP dans les segments déclaratifs et interrogatifs de longueur 2 n'a pas permis de constater une dépendance marquée. La différence de production est potentiellement liée au fait que la majorité des segments de longueur 2 sont produits à m̄=2, tranche dans laquelle les enfants produisent également comparativement moins de SFP – la proportion trouvée à m̄=2 (18983 / 58662, soit 32.4%) correspond à celle trouvée dans les segments de longueur 2 (10426 / 30325, soit 34.4%). Il est toutefois intéressant de constater que cette caractéristique des segments de longueur 2 se retrouve également dans le CDS, où la tranche de MLU n'entre pas en compte de la même façon, alors que dans l'ADS, c'est en revanche dans les énoncés de longueur 3 qu'on observe une production significativement moins importante de SFP.

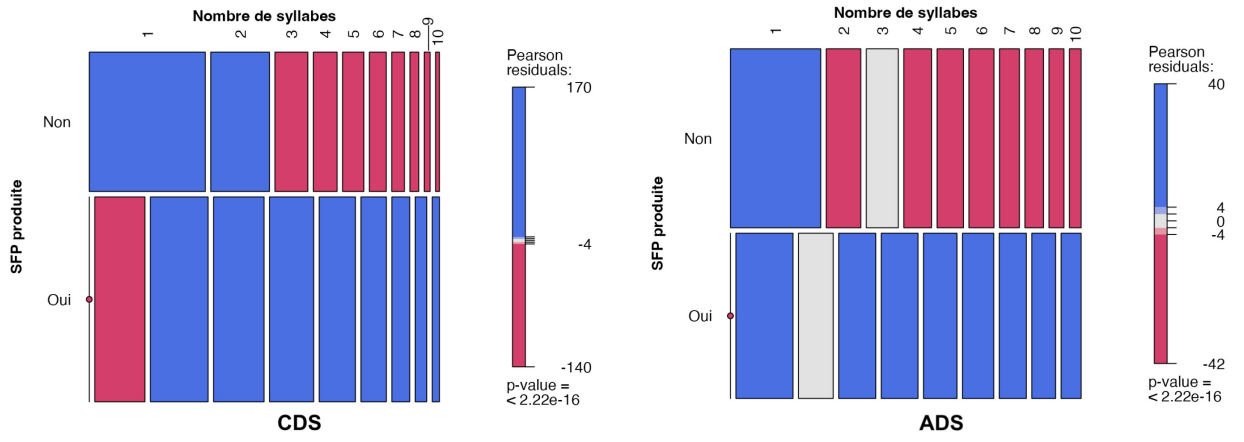

Figure 31. Table de contingence de la production de SFP et du nombre de syllabes dans le CDS (à gauche) et l'ADS (à droite)

6.3.4 Facteurs extrinsèques

6.3.4.1 Genre de l'interlocuteur

Comme il avait été évoqué à la section 1.6.4.1, le genre de l'interlocuteur est susceptible d'avoir un impact sur la manière dont les personnes, et en particulier les enfants, s'adressent à lui – une sorte d'effet miroir des observations de Winterstein *et al.* (en préparation) selon lesquelles les hommes et les femmes utilisent certaines SFP de manière préférentielle. Dans notre corpus, nous identifions comme interlocuteur de l'enfant pour un énoncé donné la dernière personne ayant pris la parole avant cet énoncé. Comme nous l'avons mentionné à la section 5.6, nous avons pu déterminer avec une bonne fiabilité le genre des expérimentateurs dans le HKU, et une partie des interlocuteurs du CC sont décrits de manière à pouvoir inférer leur genre également. Ces données (nous n'utilisons ici que les énoncés pour lesquels le genre de l'interlocuteur est effectivement connu) nous permettent d'analyser l'impact que cette information peut avoir sur la production de SFP des enfants. Les résultats obtenus sont significatifs (χ^2=93.821; df=1; p<0.001; φ=0.064) :

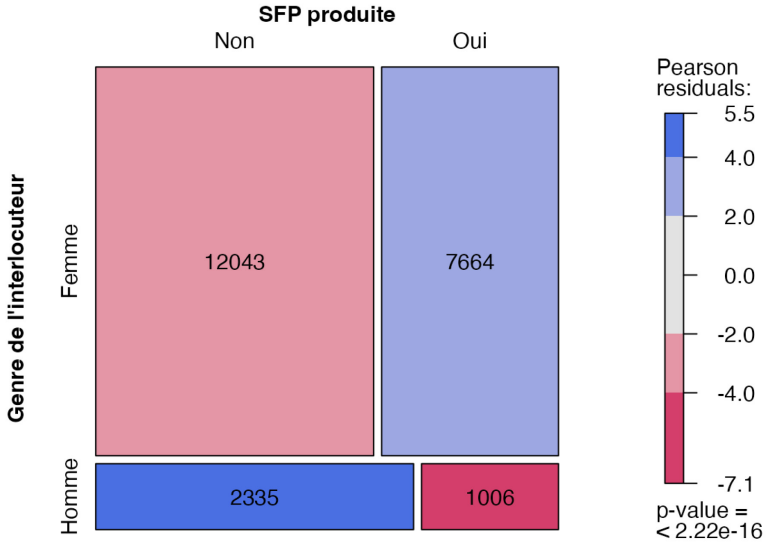

Figure 32. Tableau de contingence de la production de SFP et du genre de l'interlocuteur

Ce diagramme nous montre que les enfants ont tendance à utiliser des SFP plus fréquemment lorsqu'ils s'adressent à des femmes que lorsqu'ils s'adressent à des hommes. Il est intéressant de constater que cette propension se retrouve de façon beaucoup moins marquée dans le CDS (χ^2=17.655; df=1; p<0.001; φ=0.009), et qu'elle est pratiquement absente, mais toujours significative, dans l'ADS (χ^2=11.058; df=1; p<0.001; φ=0.022).

Notons que si la quantité d'énoncés répondant à des interlocuteurs masculins (expérimentateurs ou membres de la famille, adultes ou enfants) est largement inférieure à celle d'énoncés répondant à des femmes, des hommes sont tout de même présents dans 77 conversations, ce qui limite un potentiel effet dû à une interaction particulière entre un enfant et un interlocuteur masculin spécifique.

Il est également possible d'étendre cette observation aux effets des interactions homme–femme en analysant la production de SFP dans les différents contextes (χ^2=162.74; df=3; p<0.001; φ=0.084) :

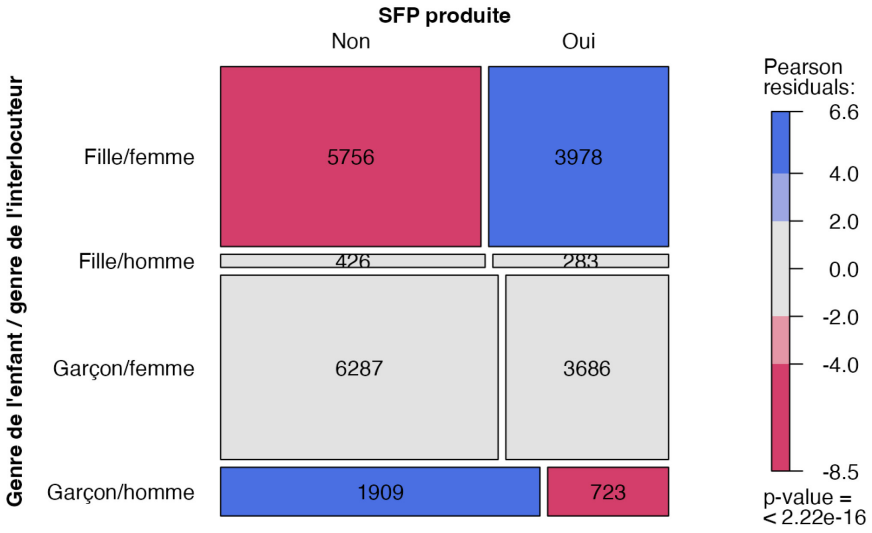

Figure 33. Tableau de contingence de la production de SFP dans les différentes combinaisons de genres

Ce diagramme indique que les filles ont tendance à utiliser plus de SFP envers les femmes, et les garçons moins envers les hommes, mais que ces différences sont moins marquées lors des relations entre les genres.

6.3.4.2 Rôle de l'interlocuteur

Nous avons décrit en 4.5.2 la façon dont nous avons regroupé les rôles des interlocuteurs (expérimentateurs, parents, employées de maison…) selon une dimension d'appartenance à l'entourage proche. Ainsi que nous l'avions suggéré en 1.6.4.1, les données de notre corpus confirment que le niveau de familiarité a un effet significatif sur la production de SFP par les enfants (χ^2=131.97; df=1; p<0.001; φ=0.036) :

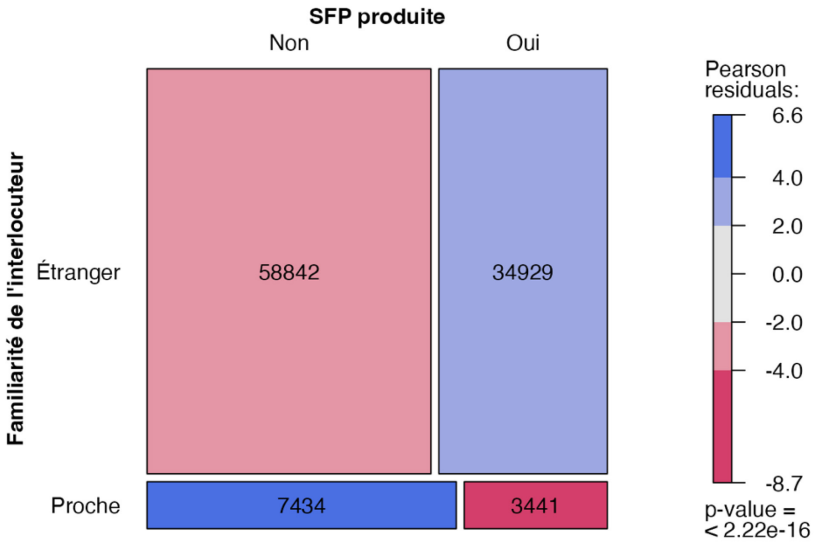

Figure 34. Tableau de contingence de la production de SFP et de la familiarité de l'interlocuteur

Ce diagramme montre que les enfants utilisent plus de SFP lorsqu'ils s'adressent ou répondent à une personne étrangère, et moins lorsqu'ils parlent à leurs proches.

Notons d'ailleurs que le pendant de ce schéma dans le CDS est également significatif : les personnes étrangères au foyer utilisent eux aussi plus fréquemment des SFP lorsqu'ils s'adressent aux enfants :

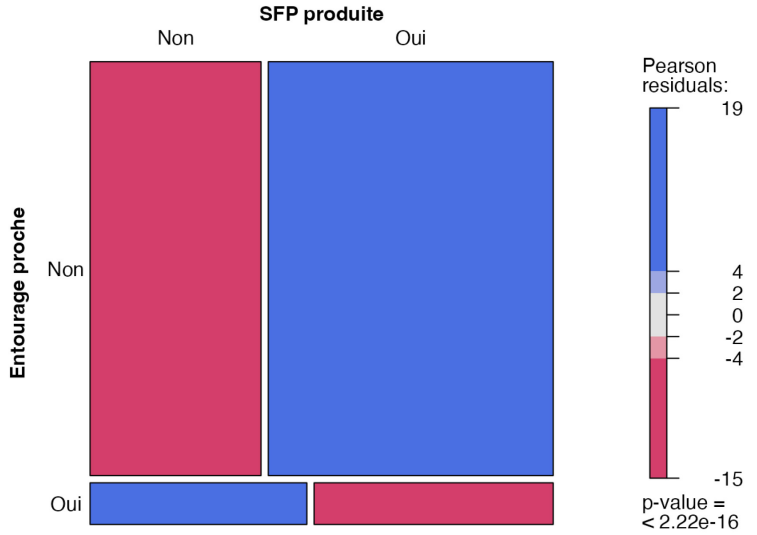

Figure 35. Tableau de contingence de la production de SFP envers les enfants et de la proximité de l'interlocuteur

Cette constatation est particulièrement pertinente en lien avec la section suivante, qui concerne l'influence du CDS sur la production des enfants.

Ajoutons également qu'on observe une différence significative dans la fréquence des interrogations adressées aux enfants entre les personnes proches et les personnes étrangères au foyer : ces dernières utilisent significativement moins de phrase interrogatives (nous ne considérons ici que les énoncés précédant directement une prise de parole de l'enfant), comme illustré sur la Figure 36 (χ^2=1221.7; df=1; p<0.001; φ=0.108) :

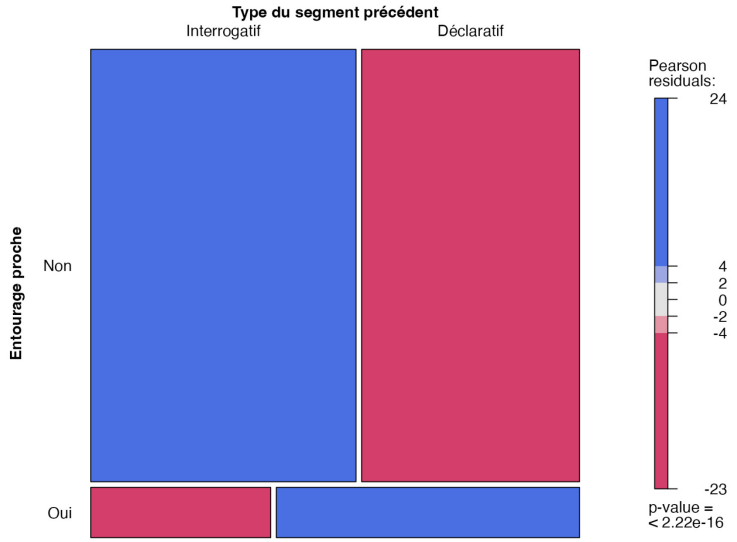

Figure 36. Tableau de contingence du type d'énoncé adressé aux enfants selon la proximité de l'interlocuteur

6.3.4.3 Influence du CDS

À ce stade de notre analyse, nous pouvons évaluer l'impact ponctuel du CDS en évaluant la fréquence de production d'une SFP par l'enfant selon la présence d'une SFP dans l'énoncé précédent. Le test d'indépendance montre une tendance significative à suivre le schéma de production de l'interlocuteur (χ^2=358.96; df=1; p<0.001; φ=0.059).

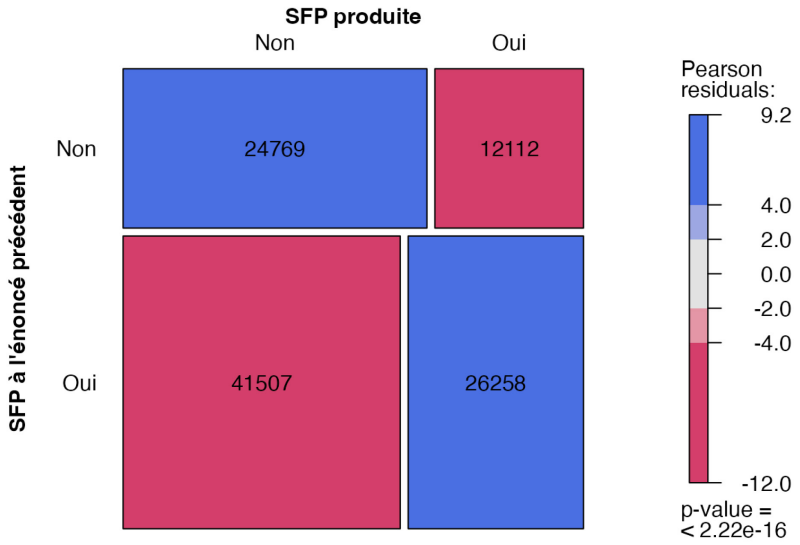

Figure 37. Tableau de contingence de la production de SFP à l'énoncé courant et à l'énoncé précédent

Ce diagramme montre que les enfants ont globalement plus tendance à utiliser une SFP si leur interlocuteur en a utilisé une juste avant, et à ne pas en utiliser si leur interlocuteur s'est également abstenu.

En se focalisant sur les segments où l'interlocuteur et l'enfant ont tous deux produit une SFP (N=26258), on peut observer (Figure 38) la propension qu'ont les enfants à réutiliser la même SFP selon leur niveau de développement langagier (χ^2=474.07; df=3; p<0.001; φ=0.134).

Ce tableau nous montre que les enfants ont plus tendance à réutiliser la même SFP que leur interlocuteur pour les tranches de MLU 1 et 2 que pour les tranches supérieures, même s'ils utilisent majoritairement des SFP différentes. Cela peut être associé à une dynamique d'imitation, mais également au fait que certaines des SFP les plus utilisées par les adultes (typiquement *aa3*, *gaa3* ou *lo1*) correspondent aux premières SFP acquises par les enfants (décrites au chapitre 8), et que le fait d'utiliser la même peut alors être lié en partie au hasard.

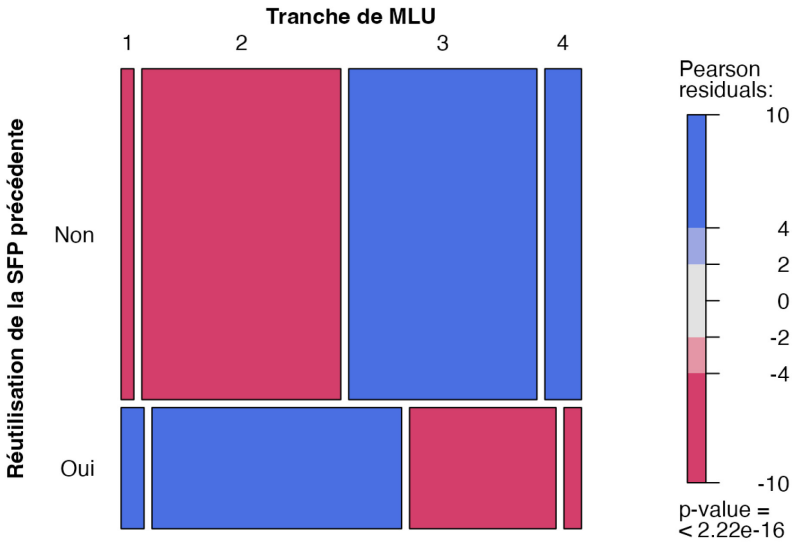

Figure 38. Tableau de contingence de la réutilisation de la SFP précédente et de la tranche de MLU

Notons que cette tendance à utiliser plus volontiers une SFP lorsque l'interlocuteur en a également utilisé une va dans une certaine mesure à l'encontre du fait que les enfants ont tendance à favoriser l'alternance des types de segments, c'est à dire qu'ils produisent plus facilement des segments déclaratifs en réponse à des segment interrogatifs, et plus facilement des segments interrogatifs en réponse à des segments déclaratifs, mais qu'ils utilisent plus fréquemment des SFP dans les segments interrogatifs, tout comme les adultes :

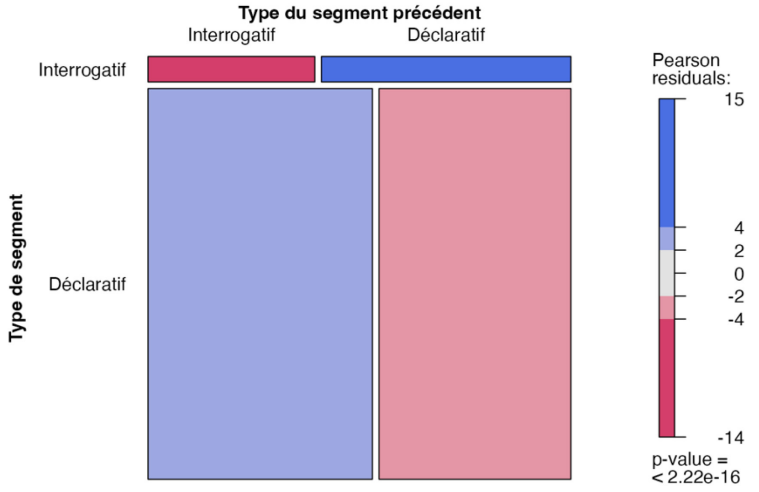

Figure 39. Table de contingence des types d'énoncés produits et types des énoncés précédents

Il sera donc utile d'essayer de combiner ces données pour comprendre les rôles joués par ces différents aspects du CDS dans les choix effectués par les enfants. Il faudra également prendre en compte le fait que le type du segment précédent joue lui aussi un rôle sur la production des SFP : les enfants en utilisent plus en réponse à une phrase interrogative qu'en réponse à une phrase déclarative :

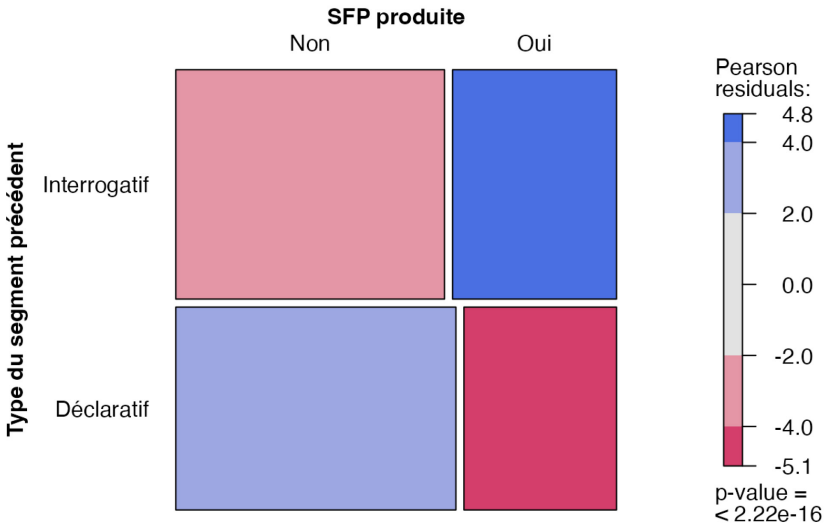

Figure 40. Table de contingence de la production de SFP par les enfants selon le type du segment précédent

Dans le chapitre suivant, nous combinerons ces premières observations dans un modèle plus global visant à évaluer l'importance de l'impact des différents facteurs sur la production des SFP par les enfants.

CHAPITRE 7

Élaboration et ajustement de modèles

7.1 Modèles linéaires généralisés mixtes

Les modèles linéaires généralisés mixtes (GLMM pour *Generalized Linear Mixed-effect Models*) sont des modèles mathématiques de régression linéaire qui permettent d'évaluer les effets de différentes variables prédictives sur une variable réponse, dont les valeurs de sortie sont comprises entre 0 et 1 et représentent une probabilité (ce qui correspond à la *généralisation* des modèles linéaires standard dont les valeurs cibles ne sont pas bornées). Aux variables indépendantes à effets fixes s'ajoutent une ou plusieurs variables à effet aléatoire (ce qui apporte au modèle son caractère *mixte*), qui permettent d'intégrer dans le modèle une variation intrinsèque due à des conditions d'expérience particulières – par exemple, le contexte spécifique d'une conversation ou les caractéristiques propres d'un participant. Pour de plus amples informations sur les GLMM et leurs utilisations en linguistique, voir par exemple Schäfer (2020).

7.2 Modélisation de la production globale de SFP

7.2.1 Constitution du modèle de base

Le premier modèle que nous ajustons vise à expliquer les facteurs jouant un rôle dans la production brute de SFP, c'est-à-dire le fait qu'au moins une SFP, quelle qu'elle soit, soit produite ou non à la fin d'un segment. Cette donnée booléenne (vrai ou faux) est disponible dans nos annotations, pour chacun des segments produits par les enfants. Nous pouvons donc ajuster un modèle en utilisant les annotations associées aux segments. Notons que ces données diffèrent légèrement des annotations brutes générées par notre programme, puisque celles-ci concernent chacune des SFP produites – ce qui peut donc aboutir à la production de plusieurs annotations pour un même segment, si celui-ci contient plusieurs SFP. Afin d'éviter une redondance de données qui biaiserait le modèle, nous avons donc effectué un simple filtre en R pour obtenir une annotation unique pour chacun des segments.

Les analyses de corrélations simples présentées au chapitre précédent (§6.2.3) permettent d'anticiper que certains de nos facteurs auront *a priori* un effet sur la production des SFP : tranche de MLU, type de segment (déclaratif ou interrogatif), nombre de syllabes, degré de familiarité avec l'interlocuteur, production d'une SFP à l'énoncé précédent. Nous avions vu au chapitre précédent que le genre de l'interlocuteur a également un effet significatif sur la production des SFP, mais la quantité limitée

d'énoncés pour lesquels cette donnée est connue est trop contraignante pour l'ajustement modèle; cette donnée a donc été laissée de côté dans un premier temps, mais nous reviendrons sur ses effets dans la suite de ce rapport. Les constructions syntaxiques remarquables, pour leur part, jouent uniquement un rôle dans le cadre des segments interrogatifs, et sont analysées dans la deuxième partie de ce chapitre à l'aide d'un modèle dédié.

De plus, nous partons du principe que l'individualité d'un enfant joue un rôle important dans son processus d'acquisition des SFP, nous ajoutons donc à nos modèles un facteur aléatoire correspondant à l'identité de l'enfant : pour les conversations du HKU, où toutes les conversations sont menées avec des enfants différents, cela correspond simplement à l'identifiant de la conversation elle-même (p. ex. `021106_hku` dans notre système). Pour les conversations du CC, toutefois, où huit enfants sont suivis sur un an, nous utilisons donc le code associé à chaque enfant (`ccc`, `cgk`, `ckt`, `hhc`, `lly`, `ltf`, `mhz`, `wbh`), sans différencier les conversations elles-mêmes.

Ainsi, notre premier GLMM[29] intègre tous ces paramètres sous la forme suivante (présentée en syntaxe R) :

(22)
```
glmer(data=segments_cs, formula=sfp_found~mlu_group + utt_type + syllables + household +
prev_sfp_found + (1|chid), family="binomial", control=glmerControl(optimizer="Nelder_Mead"))
```

La fonction utilisée, `glmer`, est la fonction dédiée à l'ajustement de GLMM dans le module `lme4`. `segments_cs` est le nom de la structure de données (*data frame*) dans laquelle les annotations relatives à chaque segment du CS sont stockées. Les paramètres de ce modèle, décrits ci-dessous, correspondent aux éléments des annotations décrits dans la Tableau 8 à la section 5.10.

La formule du modèle, `formula=sfp_found ~ mlu_group + utt_type + syllables + household + prev_sfp_found + (1|chid)`, indique que nous cherchons à exprimer la valeur booléenne indiquant la production ou non d'au moins une SFP (`sfp_found` : 0/non, 1/oui) en fonction de la tranche de MLU (`mlu_group` : catégorielle, 1, 2, 3 ou 4), le type de segment (`utt_type` : catégorielle, `STATEMENT` ou `INTERROGATION`), le nombre de syllabes (`syllables` : numérique), le degré de familiarité (`household` : booléenne, oui/familier ou non/étranger) et la présence ou non d'une SFP dans l'énoncé précédent du

[29] Modèle généré avec le module R `lme4` (Bates *et al.*, 2015)

dernier interlocuteur (`prev_sfp_found` : booléenne). Nous y ajoutons également la définition de notre effet aléatoire `(1|chid)` représentant l'identité de l'enfant (*child ID*).

Le paramètre `family="binomial"` indique que nous voulons effectuer une régression logistique, pour laquelle les valeurs cibles sont toujours 0 (aucune SFP produite) ou 1 (production d'une SFP). Le paramètre `control=glmerControl(optimizer="Nelder_Mead")`, quant à lui, permet de faire appel à un algorithme d'optimisation pour améliorer le choix des valeurs initiales lors de l'ajustement du modèle, et ainsi faciliter la convergence. Le modèle converge effectivement, ce qui indique que la procédure d'ajustement par régression a atteint un ensemble de valeurs stables. La description du modèle obtenu nous indique que tous nos paramètres ont bien un effet significatif :

(Intercept)	-2.30 (0.13)***
m̃=2	0.46 (0.03)***
m̃=3	0.70 (0.04)***
m̃=4	1.09 (0.24)***
Énoncé déclaratif	-1.09 (0.03)***
Nombre de syllabes	0.50 (0.00)***
Personne de l'entourage	-0.23 (0.02)***
SFP à l'énoncé précédent	0.29 (0.02)***
AIC	114002.4
BIC	114088.5
Vraisemblance Log	-56992.3
Nombre d'observations	104646
Nombre d'enfants (effet aléatoire)	78
Intercept (enfant)	0.75

***p < 0.001; **p < 0.01; *p < 0.05

Figure 41. Tableau des estimations des coefficients du GLMM pour la modélisation de la production de SFP

Les appels R correspondants sont donnés en Annexe E.

Nous avons encadré en rouge les valeurs de significativité (p) pour chacun des paramètres fournis au modèle. Tous les coefficients sont significatifs pour la valeur seuil α=0.05, ce qui valide leur emploi dans le modèle. Leurs valeurs, ainsi que les intervalles de confiance associés, peuvent être visualisés de manière plus naturelle sur la Figure 42 :

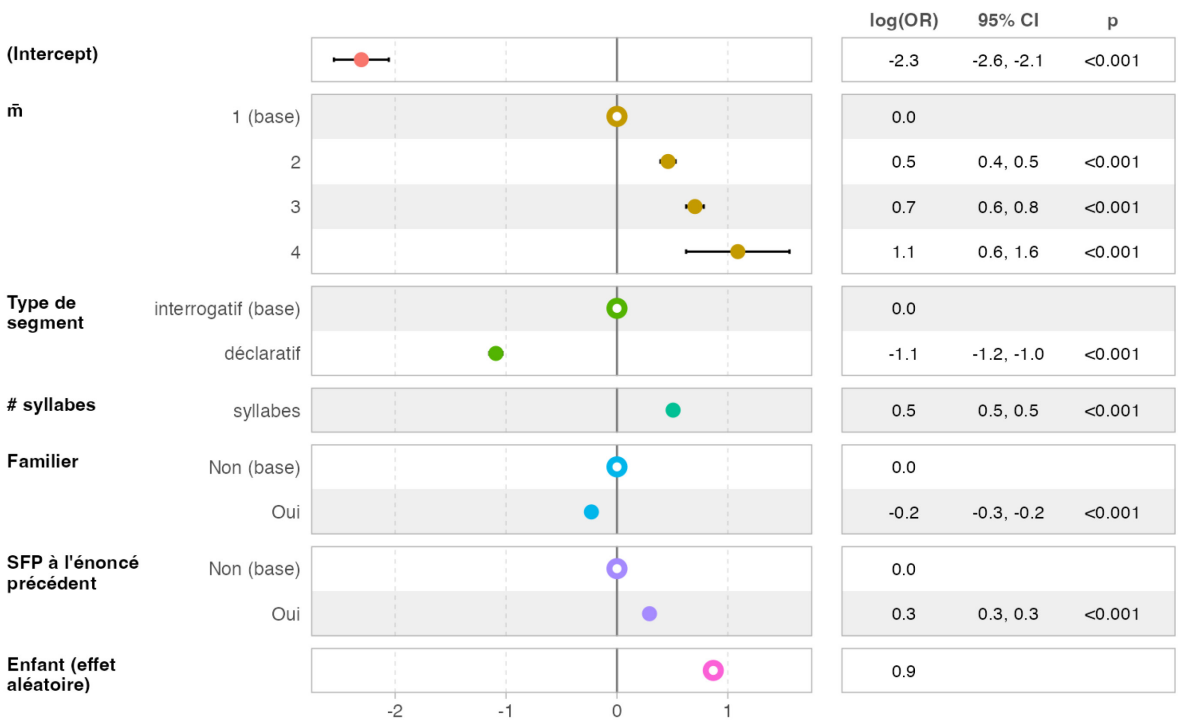

Figure 42. Estimations graphiques des coefficients pour le GLMM de modélisation de la production de SFP

Ce diagramme nous permet d'interpréter plus facilement les estimations des coefficients. On reconnaît l'effet de la MLU sur la production de SFP : plus la tranche de MLU ($\bar{m}$) est élevée, plus la tendance à utiliser une SFP augmente (coefficients positifs : 0.5 à $\bar{m}$=2, 0.7 à $\bar{m}$=3, 1.1 à $\bar{m}$=4). De même, nous avons vu que les SFP étaient plus largement utilisées dans les segments interrogatifs, ce que confirme la valeur négative du coefficient associé au type déclaratif (-1.1). On retrouve également l'effet facilitateur du nombre de syllabes (plus de SFP pour les syllabes au-delà de la deuxième : 0.5), l'effet inhibiteur du degré de familiarité de l'interlocuteur (-0.2), ainsi que l'effet facilitateur de l'utilisation d'une SFP à l'énoncé précédent (0.3). On constate néanmoins qu'une grande variabilité est liée à l'identité de l'enfant (effet aléatoire : 0.9), ce qui suggère que les paramètres précédents n'expliquent qu'une partie des valeurs rencontrées. Ce modèle est donc en accord avec les observations effectuées sur les variables individuelles.

Il est clair que ce modèle n'est pas en mesure d'expliquer la totalité des choix effectués par les enfants concernant la production de SFP. Une estimation de la puissance prédictive du modèle est obtenue grâce au coefficient de détermination R^2, qui indique théoriquement la proportion de la variance expliquée par le modèle. Une adaptation du calcul de ce coefficient pour les modèles mixtes a été proposée par Nakagawa et Schielzeth (2013) et Nakagawa *et al.* (2017). Le module R `performance` (Lüdecke *et al.*, 2021) permet d'obtenir les valeurs du R^2 marginal (variance attribuée aux effets fixes uniquement) et du R^2

conditionnel (variance attribuée aux effets fixes et aléatoires). Pour notre modèle, ces valeurs sont de 0.244 (R^2 marginal) et 0.385 (R^2 conditionnel), ce qui confirme que le facteur aléatoire (identité de l'enfant) joue un rôle important dans la variance observée pour la production des SFP. La valeur 0.385, que l'on interprète comme le fait que 38.5% de la variance du système est expliquée par notre modèle, constitue une base probante pour notre recherche.

7.2.2 Extension du modèle

Bien que notre premier modèle permette déjà de reconnaître différents facteurs clefs du processus d'intégration des SFP dans le discours des enfants, nous pouvons tenter de l'améliorer en y ajoutant d'autres paramètres ou interactions.

7.2.2.1 Type de l'énoncé précédent

À la section 6.3.4.3, nous avions noté d'une part que les enfants avaient plus tendance à produire une SFP lorsque l'énoncé précédent en contenait une également, et d'autre part qu'ils avaient plus tendance à répondre aux énoncés interrogatifs par des énoncés déclaratifs, et aux déclaratifs par des interrogatifs. Il est donc pertinent d'évaluer l'impact du type du dernier énoncé à notre modèle, ainsi que, comme ils sont potentiellement liés, l'impact de la combinaison de ces deux facteurs (type de l'énoncé précédent et présence d'une SFP dans l'énoncé précédent), en introduisant dans notre modèle une interaction entre eux. Les formules de ces nouveaux modèles correspondent à :

(23)
```
sfp_found~mlu_group + utt_type + syllables + household + prev_sfp_found + prev_utt_type +
(1|chid)
```

(24)
```
sfp_found~mlu_group + utt_type + syllables + household + prev_sfp_found * prev_utt_type +
(1|chid)
```

Nous avons ajouté le type de l'énoncé précédent (`prev_utt_type` : catégoriel, `STATEMENT` ou `INTERROGATION`). Dans la formule (24), l'interaction est symbolisée par l'astérisque, sous la forme `prev_sfp_found * prev_utt_type`. Les deux modèles convergent, et on peut observer la variation des coefficients précédents due à l'ajout de ces nouvelles données :

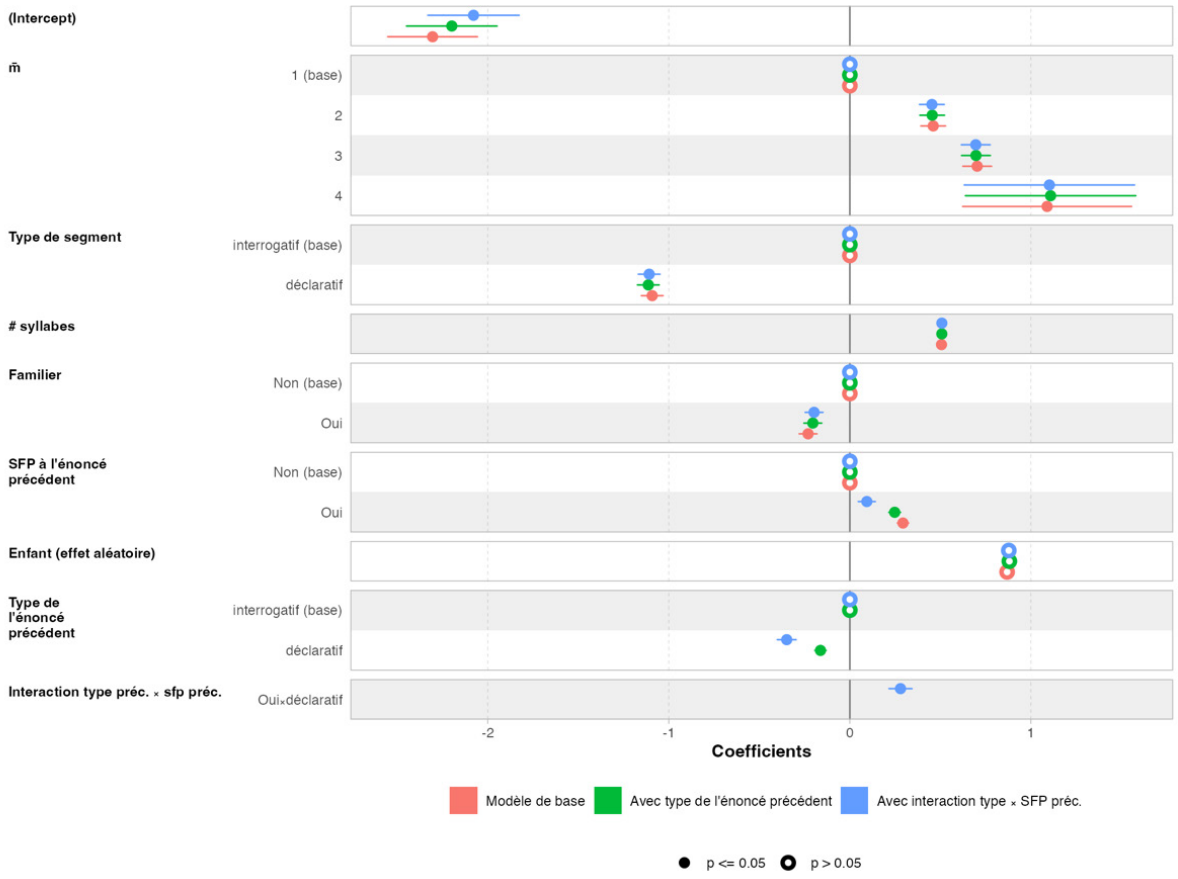

Figure 43. Comparaison des GLMM avec et sans le type de l'énoncé précédent, avec et sans interaction

On voit dans ce graphique un effet inhibiteur (significatif) du type déclaratif seul à l'énoncé précédent, mais un effet facilitateur dans le cas où l'énoncé précédent est déclaratif et qu'il est ponctué par une SFP (ligne `Interaction type préc. × sfp préc.`). Les performances des trois modèles peuvent être comparées par ANOVAs successives, tout d'abord entre les trois modèles[30], puis entre les modèles significativement plus performants. Les extraits de code R correspondants sont donnés en Annexe F.

Les résultats de la première ANOVA nous indiquent que les nouveaux modèles sont tous deux significativement plus performants que le modèle de base, et que le type de l'énoncé précédent a donc un effet significatif sur la prédiction de la production des SFP. Les résultats de la deuxième ANOVA nous indiquent que le modèle avec interaction entre le type de l'énoncé précédent et la présence d'une SFP

[30] Afin de pouvoir comparer les modèles, le premier a dû être réévalué en écartant les annotations ne contenant pas d'information sur l'énoncé précédent, c'est-à-dire quand l'enfant prend la parole en premier lors d'une conversation. Cette situation s'applique pour 50 segments, sur l'ensemble original de 104646 segments.

dans l'énoncé précédent est également significativement plus performant que le modèle sans interaction. Il est donc important d'interpréter dans ce contexte le lien entre le type d'énoncé et le fait qu'il contienne une SFP.

7.2.2.2 Âge

Nous avons vu à la section 6.2.3 que l'âge et la MLU sont fortement corrélés, mais que la MLU était en contrepartie beaucoup plus fortement corrélée avec des mesures d'utilisation des SFP comme le taux d'utilisation ou la diversité des SFP. Considérant qu'il est généralement préférable, pour l'entraînement d'un modèle de régression, de ne pas inclure de prédicteurs partageant une corrélation trop forte pour éviter les problèmes de colinéarité, nous avons fait l'hypothèse dans notre modèle de base que les effets liés à l'âge seraient largement couverts par la MLU.

L'ajustement d'un modèle intégrant l'âge sous forme de tranches d'âge (1–5) montre cependant que l'âge a également un effet significatif sur la production des SFP, sans que la MLU perde de significativité. Ainsi, le résultat de l'ajustement du modèle en (25) est présenté sur la Figure 44, et comparé avec les coefficients du modèle précédent :

(25)
```
sfp_found~age_group + mlu_group + utt_type + syllables + household + prev_sfp_found *
prev_utt_type + (1|chid)
```

Les estimations des valeurs du coefficient de tranche d'âge sont toutes significatives, et semblent avoir un effet inhibiteur sur la production de SFP. Cependant, l'ajout de ce facteur augmente les effets facilitateurs des valeurs de tranches de MLU. L'effet de la tranche d'âge sur la performance du nouveau modèle est cependant confirmé par la comparaison entre les deux modèles par ANOVA qui indique une amélioration significative (χ^2=132.65; $df=4$; p<0.001), confirmée par la valeur de AIC ($AIC_{\neg âge}$=113472, $AIC_{âge}$=113348). L'évaluation des R^2 indique quant à elle que le R^2 conditionnel passe de $R^2_{cond.(\neg âge)}$=0.392 à $R^2_{cond.(âge)}$=0.358 alors que le R^2 marginal passe de $R^2_{marg.(\neg âge)}$=0.249 à $R^2_{marg.(âge)}$=0.253, ce qui constitue une (faible) amélioration de l'explication de la variabilité due aux effets fixes. On note par ailleurs que l'ajout de l'âge n'a eu aucun impact notable sur les estimations des autres prédicteurs.

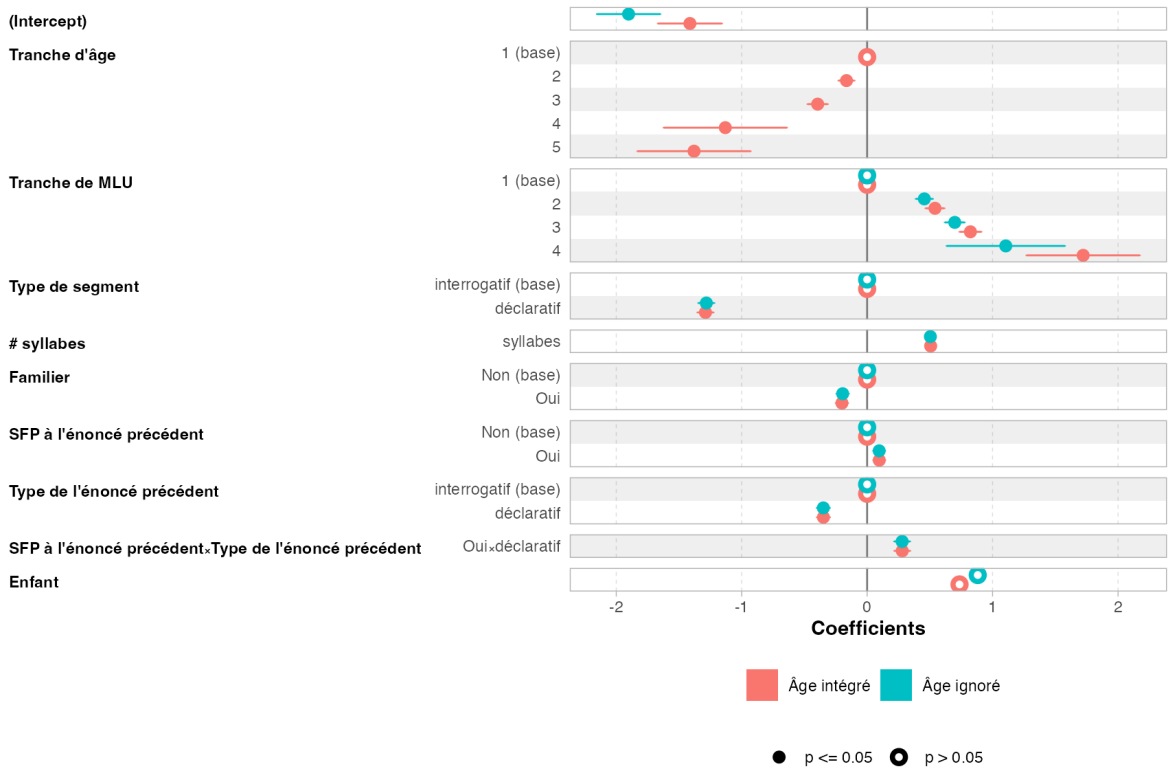

Figure 44. Comparaison des modèles avec et sans la tranche d'âge comme prédicteur de la production de SFP

L'ajout d'une interaction entre âge et MLU n'a pas permis d'ajuster un modèle convergent, elle n'a donc pas été conservée.

7.2.2.3 Autres effets

Nous observons également un effet significatif de l'interaction entre le type de segment (`utt_type`) et le degré de familiarité (`household`). La meilleure performance du modèle intégrant l'interaction est effectivement confirmée par ANOVA (p<0.001), mais la taille d'effet associée à l'interaction semble pour sa part très réduite, n'ayant qu'un impact mineur sur le R^2 conditionnel et le R^2 marginal.

Nous savons d'autre part que le genre de l'interlocuteur a un effet significatif, mais que nous n'avons pas pu l'observer pour le genre de l'enfant. L'interaction entre les deux devient pour sa part significative. Toutefois, l'ajustement de ce modèle requiert de ne considérer que les annotations pour les segments où le genre de l'interlocuteur est connu, ce qui ne représente que 22% de nos données. Aussi, nous n'avons dans un premier temps pas approfondi nos recherches dans cette direction.

7.3 Modélisation de la production de SFP dans les segments interrogatifs

Nous avons vu à plusieurs reprises que les productions des segments déclaratifs et interrogatifs constituent des processus distincts, en raison d'une part de l'évolution des besoins pragmatiques des enfants, mais également des spécificités syntaxiques de certaines constructions interrogatives. Pour affiner la modélisation de la production des SFP dans le cadre des énoncés interrogatifs, nous procédons à l'ajustement d'un modèle différent, entraîné spécifiquement sur les segments interrogatifs de notre corpus (N=6943). Les coefficients pour ce modèle sont présentés à la Figure 45 ci-dessous.

Ainsi que nous l'avions vu à la section 6.3.3.2, la production des structures syntaxiques interrogatives est fortement liée au niveau de développement langagier. Bien que la combinaison de ces deux facteurs (présence d'une structure syntaxique et tranche de MLU) engendre une certaine redondance dans le modèle, elle reste pertinente pour tenter de modéliser les variations de production de SFP au fil du développement.

Le nombre de syllabes est également pertinent pour justifier que certains segments sont trop courts pour contenir une SFP (p<0.001). Une interaction avec le type de structure syntaxique serait également envisageable, du fait de leur longueur, mais n'a pu être testée pour des raisons de complexité de modèle et de non-convergence.

L'effet de la présence d'une SFP à l'énoncé précédent a déjà été constaté dans notre premier modèle et est susceptible de jouer un rôle pour les interrogations également. Nous envisageons de plus une interaction avec le type de l'énoncé précédent, du fait de la tendance à l'alternance entre énoncés déclaratifs et interrogatifs (p=0.012).

L'effet du genre de l'interlocuteur a été rapporté comme étant significatif (p=0.007). L'intégration du genre de l'enfant, de son âge et de la familiarité de l'interlocuteur a été testée mais leurs effets ne sont pas significatifs.

L'utilisation de l'identifiant de l'enfant comme facteur aléatoire, comme au premier modèle, a posé des problèmes de convergence. Nous avons donc utilisé l'identifiant des conversations elles-mêmes, ce qui revient à ignorer le fait qu'un grand nombre de conversations ont été menées avec les mêmes enfants à des âges différents.

Le modèle obtenu avec les facteurs mentionnés (structure syntaxique, interaction entre SFP à l'énoncé précédent et type de l'énoncé précédent, nombre de syllabes, genre de l'interlocuteur) est présenté à la Figure 45 :

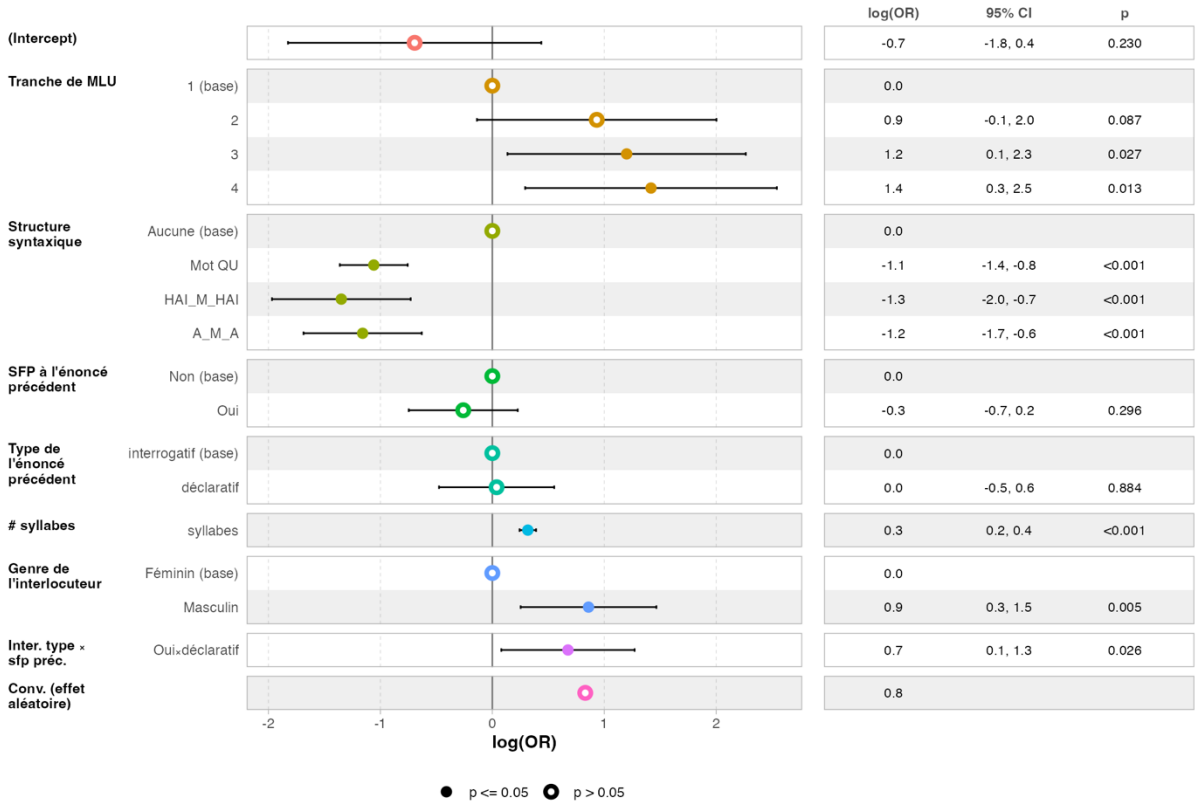

Figure 45. Estimations des coefficients pour le GLMM de la production de SFP dans les segments interrogatifs

La Figure 45 nous indique que les paramètres choisis sont effectivement significatifs – à l'exception de la présence d'une SFP à l'énoncé précédent et du type de l'énoncé précédent, pour lesquels l'interaction est pour sa part significative. Ces valeurs montrent que l'utilisation de structures syntaxiques (mot QU, HAI_M_HAI et A_M_A) a tendance à inhiber la production de SFP. Les énoncés plus longs (nombre de syllabes plus élevé) favorisent quant à eux l'utilisation d'une particule, ce qui correspond à nos observations antérieures. Notons également l'effet du genre de l'interlocuteur, selon lequel le fait que l'interlocuteur soit un homme a un effet facilitateur significatif sur la production d'une SFP.

Il est intéressant de comparer ce modèle à ceux ajustés pour le CDS et l'ADS, selon les mêmes critères (prise en compte des segments interrogatifs uniquement, mêmes formule et optimiseur), pour avoir un

aperçu des éléments qui jouent un rôle sur la production des SFP dans les énoncés interrogatifs chez les adultes :

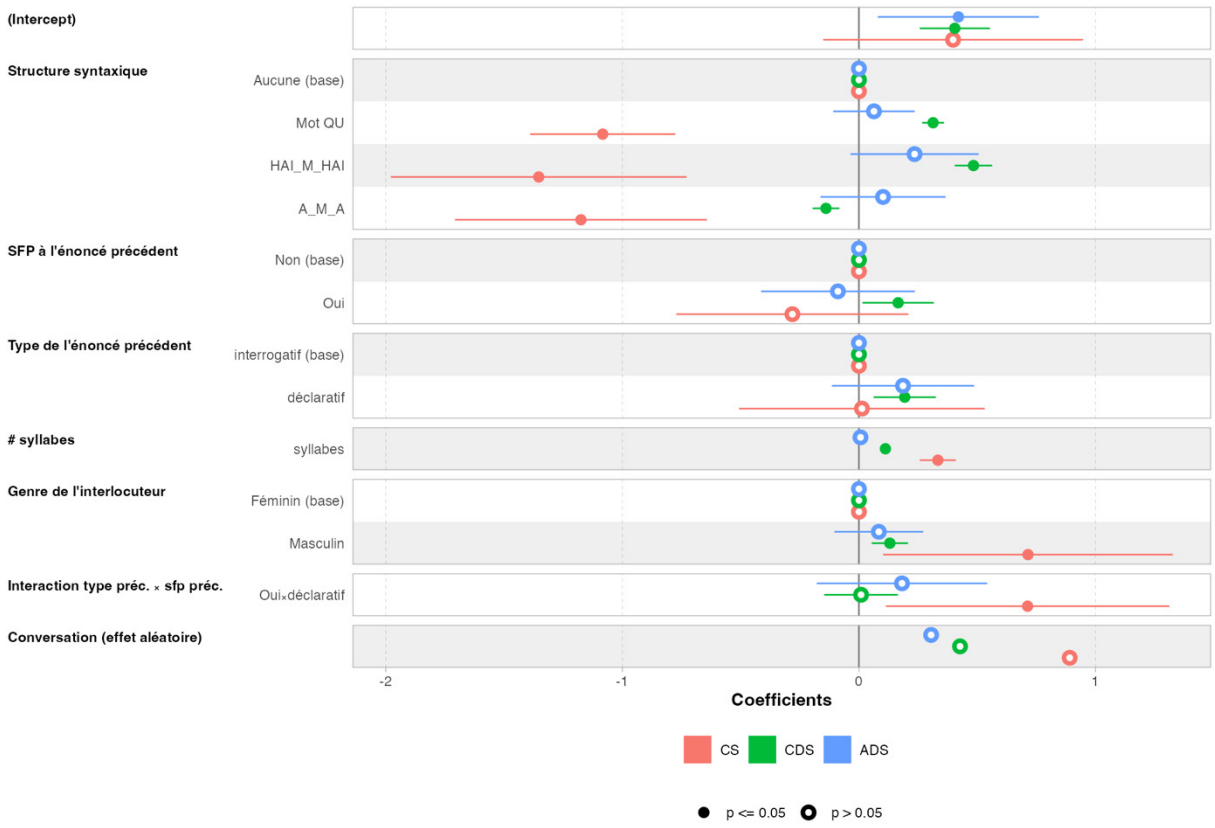

Figure 46. Comparaison des modèles de production de SFP dans les énoncés interrogatifs pour le CS, le CDS et l'ADS[31]

Si le modèle ajusté pour le CDS affiche une majorité de coefficients significatifs, l'ajustement du modèle pour l'ADS n'a pour sa part pas permis d'atteindre un bon potentiel prédictif.

La différence principale entre le CS et le CDS que l'on observe dans cette comparaison est le fait que l'utilisation des structures syntaxiques interrogatives, dont l'effet est inhibiteur chez les enfants, a chez les adultes un effet plutôt facilitateur. Notons également que l'effet du genre de l'interlocuteur reste significatif dans le CDS, même si la taille de l'effet est moindre.

[31] Pour que la comparaison soit valide, nous avons retiré du modèle pour le CS le facteur MLU. La seule différence notable avec le modèle précédent qu'engendre cette modification est que la valeur de l'intercept, de négative (-0.7), est devenue positive (0.398). Les autres coefficients ont gardé des estimations très similaires.

CHAPITRE 8

Aperçu détaillé de quelques SFP

Ce chapitre décrit l'évolution de huit SFP parmi les plus utilisées par les enfants. Sept d'entre elles (*aa3*, *gaa3*, *lo1*, *ne1*, *aa1*, *laa1* et *ge2*) sont celles bénéficiant de la couverture la plus importante à $\bar{m}$=4. La dernière, *aa4*, bien que légèrement moins utilisée (12ème SFP la plus couramment utilisée à $\bar{m}$=4), est intéressante par son statut de particule exclusivement interrogative. Nous décrivons ces particules et exposons certains facteurs qui semblent influencer leur émergence dans le discours des enfants. Des statistiques globales sur toutes les SFP observées dans notre étude sont fournies en Annexes G et H.

8.1 aa3

La particule *aa3* est, de très loin, la SFP la plus courante, tant dans l'ADS que dans le CDS ou le CS. Elle est couramment décrite comme une particule d'atténuation, permettant de rendre le contenu d'un énoncé moins « abrupt ».

aa3 est utilisée par la quasi-totalité des enfants observés. Elle n'est absente que dans trois conversations du HKU (hku-030523, hku-040515 et hku-050214), dans laquelle les enfants, respectivement de 3, 4 et 5 ans, démontrent un niveau de développement langagier particulièrement bas, avec des MLU autour de 2.4, et des valeurs de diversité de SFP de 3 ou moins, ce qui est très peu pour ce niveau de développement (la moyenne de diversité à $\bar{m}$=2 est μ=14, σ=5.24).

Elle est compatible avec différents types d'énoncés, en particulier les énoncés déclaratifs et les énoncés interrogatifs que nous observons dans cette étude. Quelques exemples d'utilisation de *aa3* par les enfants sont présentés en (26).

La Figure 47 illustre l'évolution de sa fréquence d'utilisation relative pour les différents stades de développement des enfants, ainsi que dans le CDS et l'ADS. Nous y indiquons également les valeurs de significativité p (notées $p_{\text{unadj.}}$ pour « non ajustées ») des tests t de Student effectués par paire pour vérifier l'indépendance des populations.

(26) a. 唔 好 **aa3** . (cc-ckt020317)
m4 hou2 **aa3**
NEG bien **SFP**
Pas bien/pas envie.

b. 乜嘢 **aa3** ? (cc-ccc020110)
mat1je5 **aa3**
quoi **SFP**
Qu'est-ce que c'est?

c. 熄咗 佢 好唔好 **啦** ? (cc-cgk020207)
sik1-zo2 keoi5 hou2m4hou2 **aa3**
éteindre.PERF ça bien-pas-bien **SFP**
Il faut l'éteindre?

d. 洗 完 頭 跟住 點樣 **aa3** . (hku-040703)
sai2 jyun4 tau4 gan1zyu6 dim2joeng2 **aa3**
laver fini cheveux et-ensuite comment **SFP**
J'ai terminé de laver les cheveux, quoi faire ensuite.[32]

[32] Cet énoncé est ponctué d'une marque déclarative (point final) dans le corpus original, mais le mot 點樣 |*dim2joeng2* « comment » impose une interprétation interrogative.

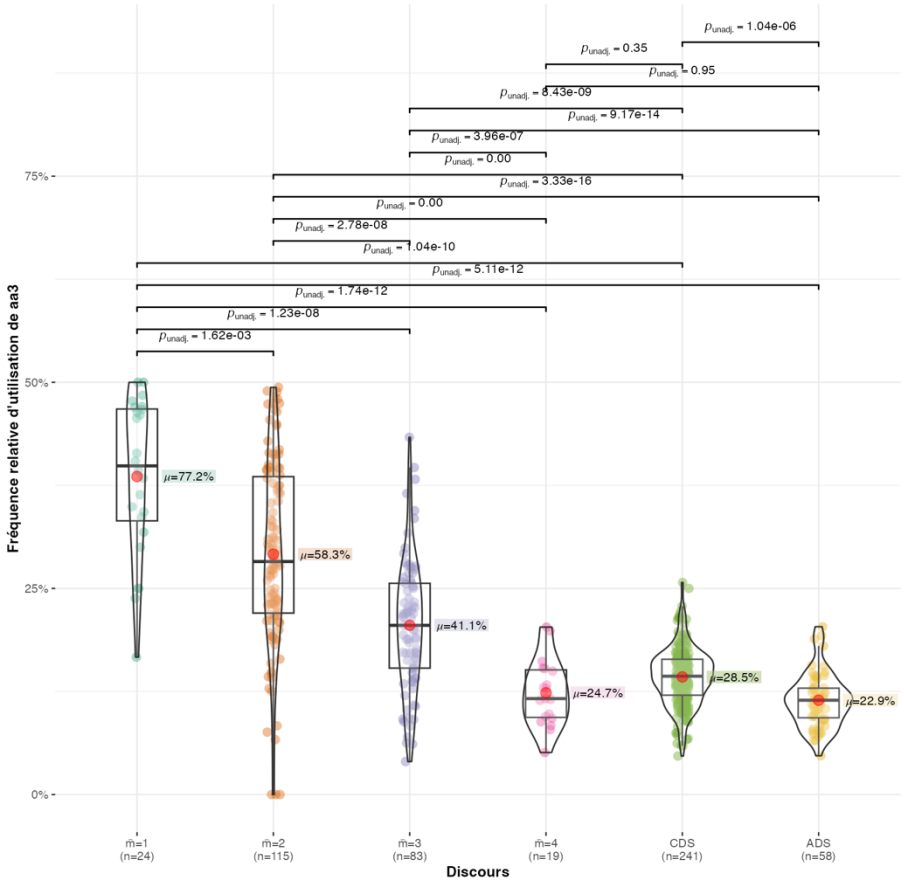

Figure 47. Fréquence relative d'utilisation de *aa3* pour les différentes tranches de MLU, le CDS et l'ADS

On remarque que les tranches de MLU 1, 2 et 3 sont caractérisées par une variabilité très importante de ces valeurs, mais que cette variabilité diminue avec le niveau de développement langagier. Le profil d'utilisation semble se stabiliser à mesure que le niveau de développement augmente, et s'approcher de celui retrouvé dans l'ADS, ce qui est confirmé par les tests *t* de Student effectués par paires : les populations m̄=4 et ADS ne se différencient pas de manière significative (p=0.95).

L'utilisation prédominante de *aa3* justifie également de comparer son utilisation à l'absence de SFP et à la présence d'une autre SFP, quelle qu'elle soit. Comme illustré sur la Figure 48, on constate une évolution de ce rapport avec le niveau de développement langagier :

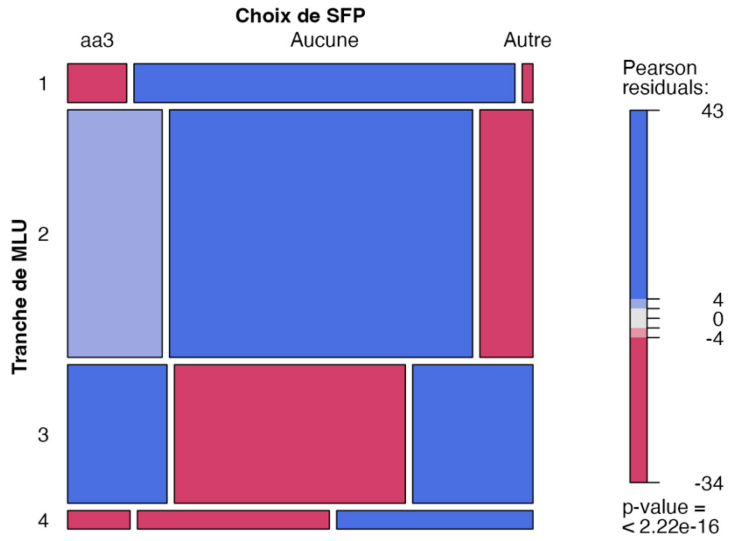

Figure 48. Table de contingence du choix de SFP (*aa3* vs Aucune/Autre) et de la tranche de MLU

Ce graphique nous confirme que, si les enfants ont bien entendu tendance à majoritairement ne pas utiliser de SFP à m̄=1, ils utilisent tout de même *aa3* plus que toutes les autres SFP confondues. L'utilisation de *aa3* trouve son maximum à m̄=3, où elle est toutefois déjà rendue minoritaire par rapport au reste des SFP. À m̄≥3, l'émergence des nombreuses autres SFP réduit fortement l'utilisation de *aa3* ainsi que le nombre d'énoncés dépourvus de SFP.

Comme nous l'avons mentionné, *aa3* peut être utilisée pour ponctuer des énoncés aussi bien déclaratifs qu'interrogatifs, avec dans tous les cas une valeur d'atténuation. On constate sur la Figure 49 que l'utilisation qu'en font les enfants change au cours de l'acquisition :

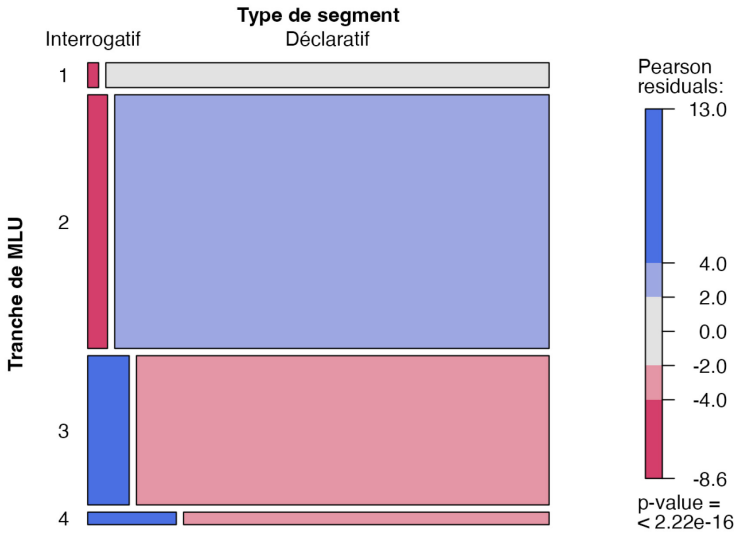

Figure 49. Répartition de l'utilisation de *aa3* selon le type de segment, en fonction de la tranche de MLU

aa3 suit l'augmentation du nombre de segments interrogatifs dans le discours des enfants, ce qui suggère que son utilisation dans ces énoncés reste persistante. Cette tendance est confirmée par la comparaison avec l'utilisation d'autres SFP selon le type de segment au cours du développement, présentée sur la Figure 50, qui montre que son rôle dans le cadre des segments déclaratifs change plus fortement que pour les segments interrogatifs :

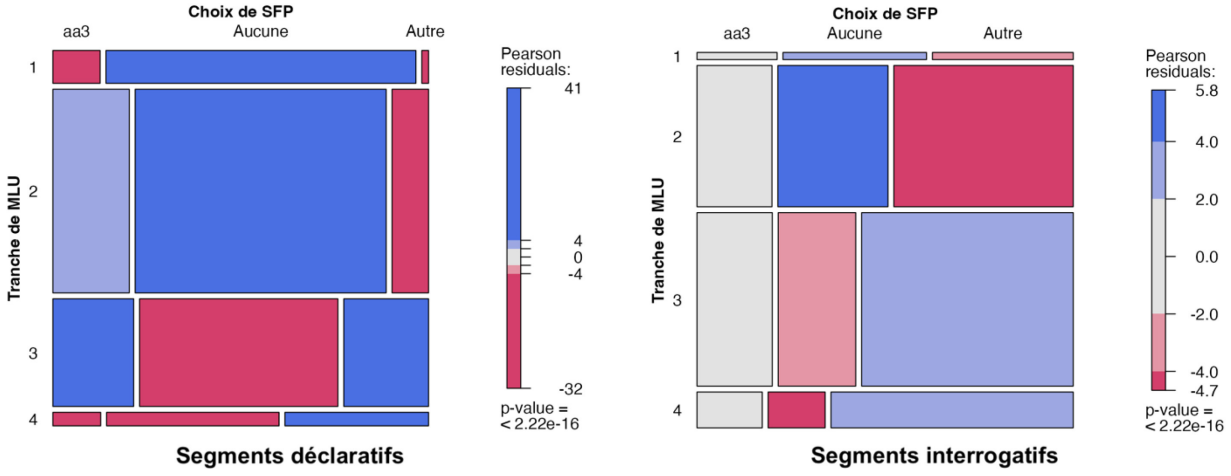

Figure 50. Comparaison de l'évolution de *aa3* par tranche de MLU pour les énoncés déclaratifs (à gauche) et interrogatifs (à droite)

On note que l'utilisation de *aa3* dans les énoncés déclaratifs est en tous points similaire à son utilisation globale (rappelons que les énoncés déclaratifs constituent la grande majorité des énoncés produits par les

enfants, soit 93.9%), et qu'on y retrouve donc exactement les mêmes variations, et en particulier la baisse due à la diversification du lexique des enfants. En revanche, le profil d'utilisation dans les énoncés interrogatifs s'en différencie fortement : *aa3* semble être utilisée de manière globalement stable, alors que la proportion d'interrogations où n'apparaît aucune SFP diminue au profit de l'utilisation d'autres SFP. Rappelons également que ces particules peuvent être soit strictement interrogatives, auquel cas elles ne peuvent être combinées avec d'autres SFP, soit génériques, pouvant être associées à des structures interrogatives (mot QU, HAI_M_HAI, A_M_A) : *aa3* n'est théoriquement compatible qu'avec les structures interrogatives, que leur complexité rend moins accessibles aux enfants au début du processus d'acquisition.

8.2 gaa3

La particule *gaa3* est généralement décrite comme la fusion de *ge3* et *aa3*, tant en termes de sémantique – selon Sybesma et Li (2007, p. 1745), « it may be seen as softening *ge3* a bit » – que de phonologie (Fung, 2000, p. 168 ; Matthews et Yip, 2011, p. 391). Son rôle est donc d'affirmer, de manière légèrement moins appuyée que *ge3* (et avec parfois une valeur de rappel), la confiance que le locuteur a dans la pertinence du fait qu'il énonce. On la retrouve ainsi dans des énoncés tels que :

(27) a. CHI: 食 得 晒 **gaa3** . (cc-mhz020226)
sik6 dak1 saai3 **gaa3**
manger pouvoir tout **SFP**
Je peux manger n'importe quoi!

b. CHI: 但係 有 好多 羹羹 **㗎** ? (hku-040021)
daa6hai6 jau5 hou2do1 gang1gang1 **gaa3**
mais il.y.a beaucoup soupe **SFP**
Mais il y a encore beaucoup de cuillers, n'est-ce pas?

Il est toutefois intéressant de constater que *gaa3* est beaucoup plus présente que *ge3* dans le discours des enfants : elle apparaît dans 21% des conversations à m̄=1, contre 13% pour *ge3*, et dans 100% des conversations à m̄=4, contre 79% pour *ge3*; elle apparaît donc avant *ge3* chez un certain nombre d'enfants, et n'en est donc clairement pas perçue comme une atténuation. Sa fréquence relative moyenne d'utilisation est également nettement plus élevée que *ge3* (15.1% à m̄=4 contre 2.7% pour *ge3*). Ces valeurs en font la deuxième SFP la plus utilisée par les enfants pendant la majorité du processus d'acquisition, tant en termes de couverture (proportion des enfants qui l'utilisent) qu'en termes de fréquence relative moyenne.

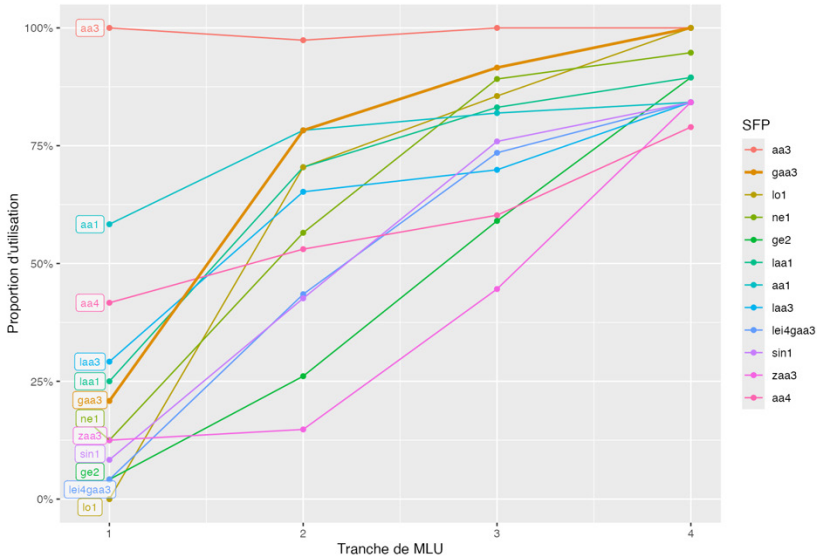

Figure 51. Évolution de la couverture de *gaa3* dans le discours des enfants, comparée aux autres SFP

Pour comparaison, *gaa3* a une fréquence relative moyenne d'utilisation de 6.95% dans le CDS et 9.93% dans l'ADS :

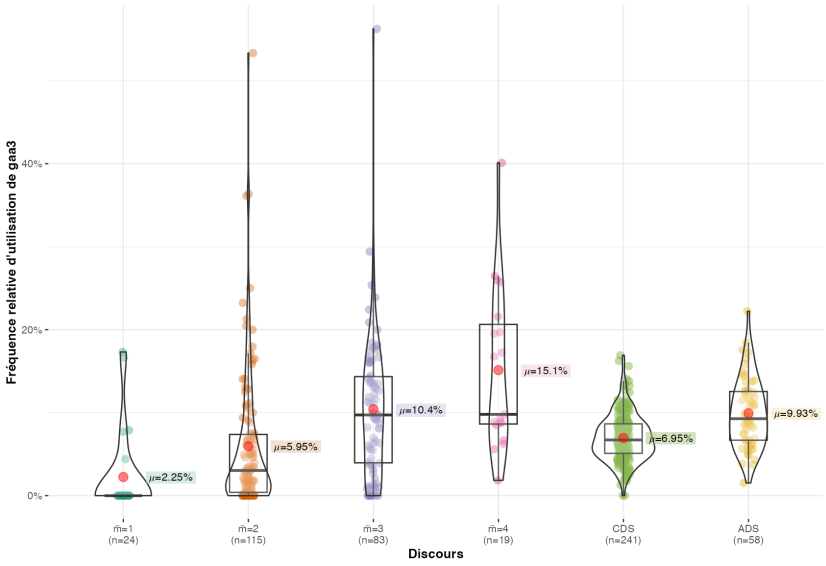

Figure 52. Évolution de la fréquence relative d'utilisation de *gaa3* dans le CS, ainsi que dans le CDS et l'ADS

Il est intéressant de constater que *gaa3* est de plus en plus utilisée par les enfants au fur et à mesure que leur MLU augmente, alors que sa fréquence est particulièrement basse dans le CDS. Tang (2018) rapporte à ce sujet qu'elle apparaît très souvent dans les énoncés scolaires à l'école maternelle (sans malheureusement donner d'exemple), ce qui signifie que les enfants y sont exposés de manière très

précoce. Dans notre corpus, *gaa3* est cependant attestée chez trois enfants à ā=1; pour deux d'entre eux (CCC et CGK), elle apparaît d'ailleurs dès la première séance d'interaction (respectivement à 1;10.08 et 1;11.01), ce qui suggère qu'elle était potentiellement déjà acquise avant le début de l'expérience. Le troisième, MHZ, observé depuis 1;07.00, commence à l'utiliser à 1;09.08.

8.3 lo1

La particule *lo1* est largement utilisée dans le discours adulte (présente dans 5% des segments dans l'ADS, mais seulement 1.6% des conversations dans le CDS). Elle est utilisée pour indiquer que le contenu de l'énoncé qu'elle ponctue relève de l'évidence et ne saurait être remis en question (p. ex. Matthews et Yip, 2011, p. 405). Lee et Law (2000, p. 134) rapportent qu'elle est activement utilisée par les enfants qu'ils ont observés, et que l'utilisation qu'ils en font correspond en tous points à celle faite par les adultes. On retrouve des utilisations comme celles illustrées en (28) :

(28)	a.	CHI:	要	啲	番 鹼	囉！		(hku-050010)
			jiu3	di1	faan1gaan2	**lo1**		
			vouloir	PART	savon	**SFP**		
			Il me faut du savon!					
	b.	CHI:	好靚	嗰啲	床	仔	囉！	(hku-040702)
			hou2leng3	go2di1	cong4	zai2	**lo1**	
			très-joli	ces	lit	DIM	**SFP**	
			Ces lits sont très jolis!					

Dans notre corpus, *lo1* est utilisée par l'ensemble des enfants à m̄=4. Il est toutefois intéressant de noter qu'elle n'apparaît dans aucune des conversations (dans le CS) à m̄=1, ce qui n'est le cas que pour 20 des 47 SFP que nous observons. À m̄=2, elle est déjà présente dans 70% des conversations, ce qui constitue la progression la plus forte à ce stade du développement. Sa fréquence relative d'utilisation passe de 0 (m̄=1) à 4.97% (m̄=4), son évolution est présentée sur la Figure 53.

On constate de plus que *lo1* n'est utilisée que par deux enfants de moins de 2 ans (ā=1), dans les deux cas à 1;11. L'un deux (MHZ), observé depuis l'âge de 1;07, l'utilise 3 fois lors d'une conversation à 1;11.06 (MLU=2.2). L'autre (CGK), observé depuis l'âge de 1;11.01, l'utilise 4 fois lors d'une conversation à 1;11.29 (quatrième séance; MLU=2.5), dont deux fois dans une répétition de l'énoncé précédent. À ā=2, *lo1* est attestée dans 76% des conversations.

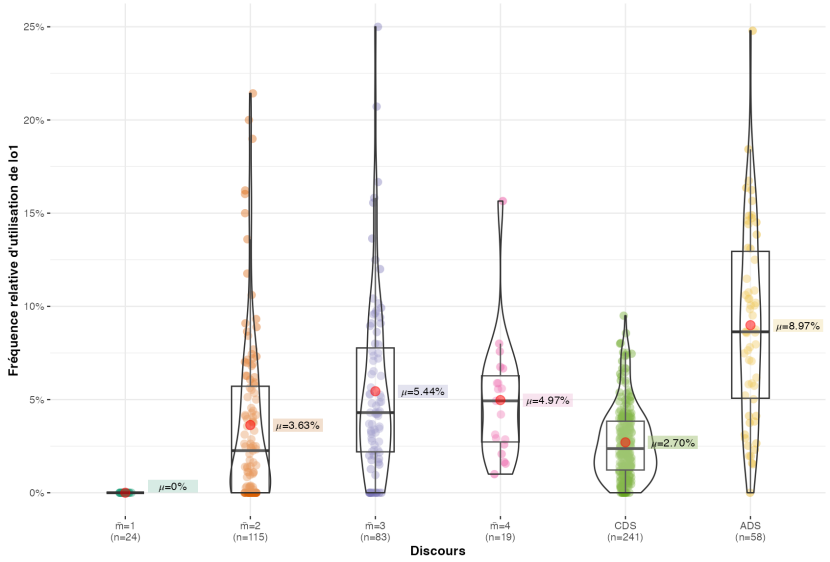

Figure 53. Évolution de la fréquence relative d'utilisation de *lo1* dans le CS, ainsi que dans le CDS et l'ADS

À $\bar{m}$=4[33], le test d'indépendance indique un effet du genre sur la production de *lo1* (χ^2=17.742; df=1; p<0.001; φ=0.065). La Figure 54 montre que les filles ont tendance à plus utiliser cette particule que les garçons; bien que faible, cette tendance est significative.

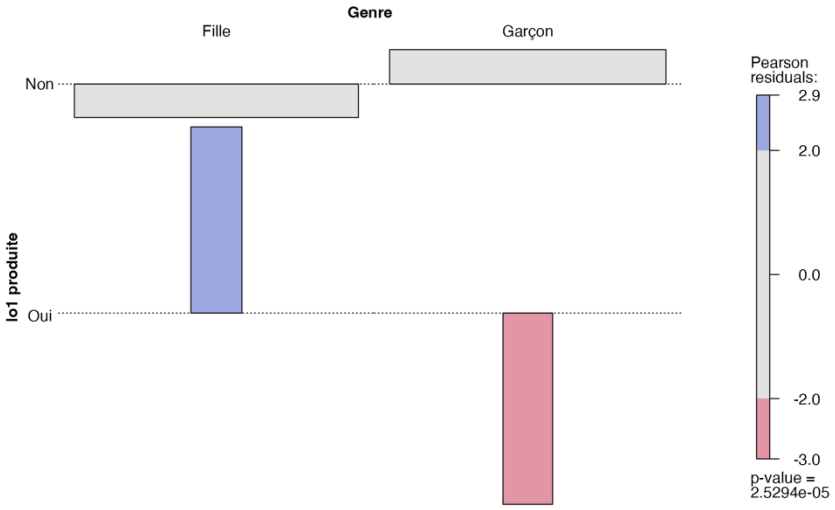

Figure 54. Résidus de Pearson de la table de contingence de la production de *lo1* par genre à $\bar{m}$=4

[33] Nous appliquons cette analyse à $\bar{m}$=4 plutôt qu'à l'ensemble du corpus pour pallier le fait que celui-ci est biaisé, en ce que les moyennes d'âge et de MLU sont globalement plus élevées pour les filles.

On trouve également, sur l'ensemble du corpus, une tendance significative à moins utiliser *lo1* envers les interlocuteurs masculins (χ^2=20.394; df=1; p<0.001; φ=0.030) :

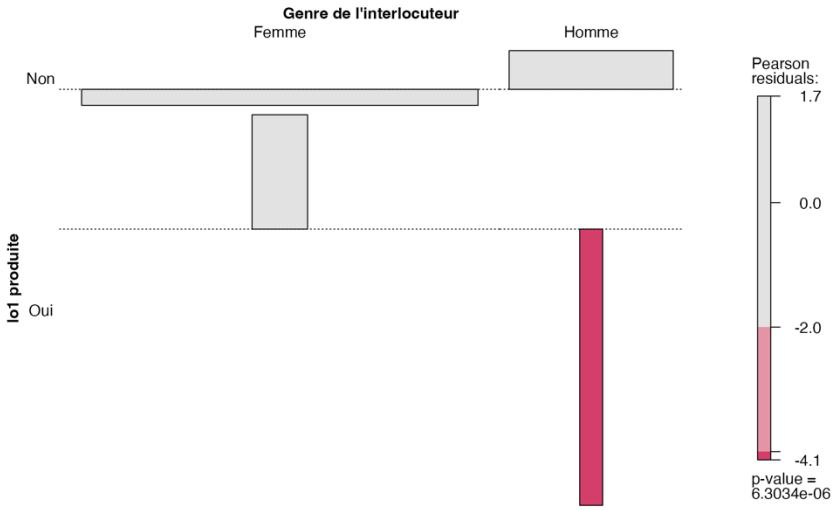

Figure 55. Résidus de Pearson de la table de contingence de la production de *lo1* en fonction du genre de l'interlocuteur

8.4 ne1

Tout comme *aa3*, *ne1* peut être utilisée dans des segments déclaratifs et interrogatifs. Toutefois, à la différence de *aa3*, elle peut aussi bien accompagner une question marquée par une structure interrogative (mot QU, A_M_A) que transformer un segment déclaratif en segment interrogatif (Kwok, 1984 ; Sybesma et Li, 2007, p. 1758). Sa signification est différente selon le type d'énoncé où elle est utilisée : dans les énoncés déclaratifs, elle marque le fait que le locuteur attire l'attention sur un élément de l'environnement. Dans le cadre des énoncés interrogatifs, elle apporte une notion d'hésitation lorsqu'elle accompagne une structure interrogative, et une notion de comparaison lorsqu'elle est utilisée de manière autonome (Kwok, 1984, p. 73, 91).

Comme pour les autres SFP, son utilisation varie au cours du processus d'acquisition; la Figure 56 montre sa répartition selon les trois types d'utilisation mentionnés plus haut, soit déclaratif, interrogatif avec structure interrogative ou interrogatif sans structure (autonome).

(29) a. INV: 即是 咁樣 剝 呢 . (cc-cgk020711)
zik1si6 gam2joeng2 bok1 ne1
précisément ainsi éplucher SFP
C'est comme ça qu'il faut l'éplucher.

b. CHI: 你 估 有 冇 巫婆 呢 . (cc-ltf020907)
nei5 gu5 jau5 mou5 mou4po4 ne1
tu deviner avoir avoir-NEG sorcière SFP
Tu penses qu'il y a des sorcières?

c. CHI: 幪面 超人 呢？ (cc-cgk020207)
mung4min2 ciu1jan4 ne1
masqué superhéros SFP
Qu'en est-il du superhéros masqué?

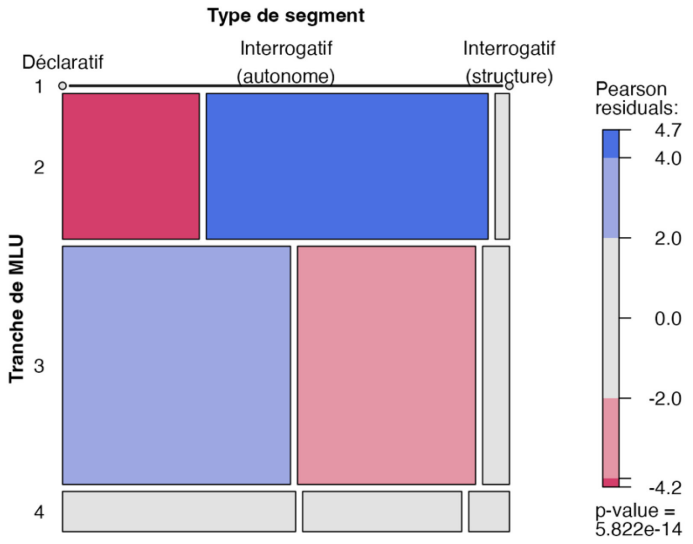

Figure 56. Répartition de l'utilisation de *ne1* selon le type d'énoncé et la tranche de MLU

On voit sur ce diagramme que son utilisation devient plus importante pour les segments déclaratifs au cours du développement. L'utilisation autonome a tendance à perdre de l'importance au profit de l'utilisation avec structures interrogatives, à mesure que celles-ci s'établissent dans le discours.

ne1 n'est utilisée que dans 12.5% des conversations à m̄=1, mais sa couverture augmente rapidement (56.5% à m̄=2, 89.2% à m̄=3, 94.7% à m̄=4). Sa fréquence relative d'utilisation moyenne (moyenne des fréquences relatives de toutes les conversations) est de 4.39% à m̄=4, proche de la moyenne des fréquences d'utilisation dans le CDS (4.19%) mais bien en dessous des 14% observés dans l'ADS.

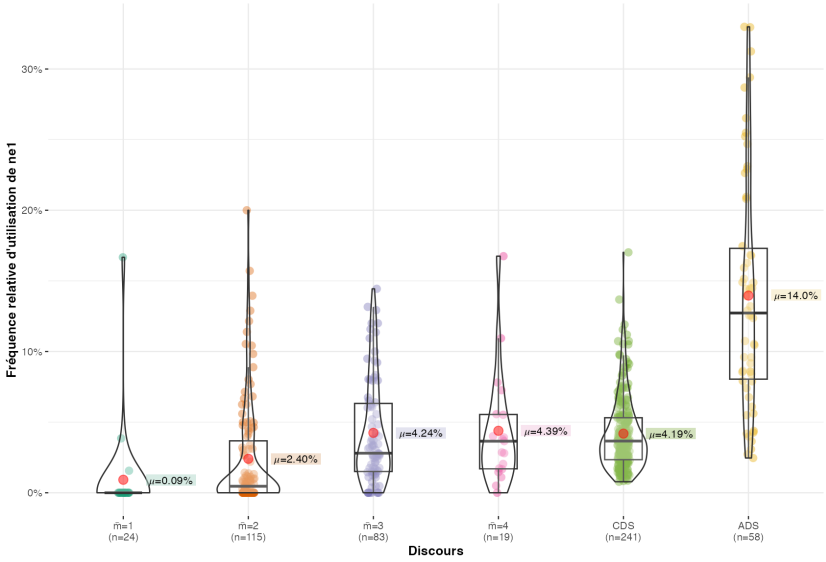

Figure 57. Évolution de la fréquence relative d'utilisation de *ne1* dans le CS, ainsi que dans le CDS et l'ADS

ne1 est attestée aussi bien à m̄=1 qu'à ā=1 (10.3% des conversations).

Comme pour *lo1*, on note une légère préférence d'utilisation chez les filles à m̄=4 (χ^2=10.109; df=1; p=0.001475; φ=0.049).

8.5 aa1

La particule *aa1* est utilisée dans les énoncés déclaratifs à valeur d'ordres et de requêtes, ainsi que dans les énoncés interrogatifs. Dans ces contextes, elle est comparable à *aa3*, mais rend l'énoncé plus « vivant » (« It is said to make the utterance more "lively" » : Sybesma et Li, 2007), et indique une implication émotionnelle plus importante.

(3C a. CHI: 我 幫 你 梳頭 **aa1** . (cc-ltf020720)
ngo5 bong1 nei5 so1tau4 **aa1**
moi aider toi peigner **SFP**
Je vais t'aider à la/te peigner.

b. CHI: 仲 有 冇 嘢 食 **aa1** ? (hku-050525)
zung6 jau5 mou5 je5 sik6 **aa1**
encore avoir avoir-NEG quelque chose manger **SFP**
Est-ce qu'il y a encore quelque chose à manger?

Son schéma d'évolution, par rapport aux autres SFP, est assez particulier : elle apparaît très tôt dans le discours des enfants avec une couverture de 58.3% à m̄=1, ce qui en fait la deuxième SFP la plus utilisée à

ce stade. Toutefois, elle semble par la suite stagner, et sa couverture à m̄=4 n'est que de 84.2%. On peut constater cette évolution atypique sur la Figure 58 :

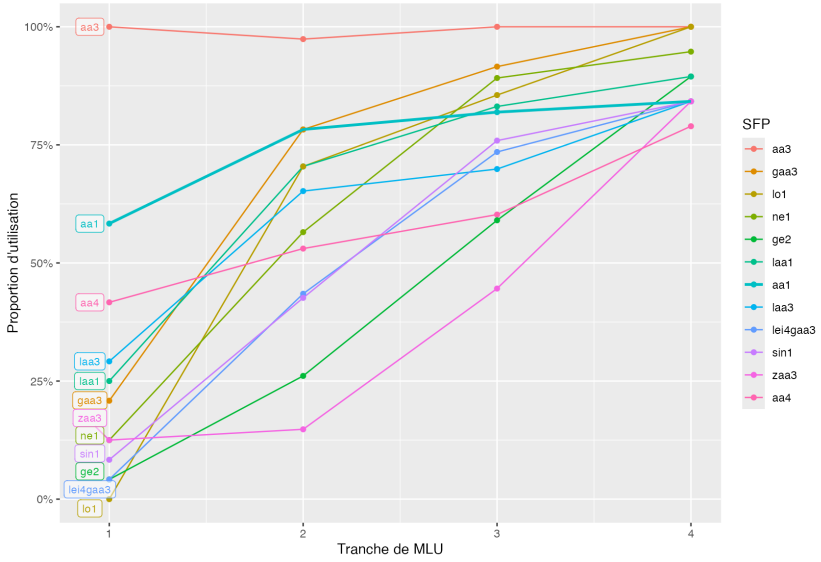

Figure 58. Évolution de la couverture de *aa1* dans le discours des enfants, comparée aux autres SFP

En termes de fréquence relative moyenne d'utilisation, elle reste également très stable et connaît même une récession aux stades plus avancés du développement :

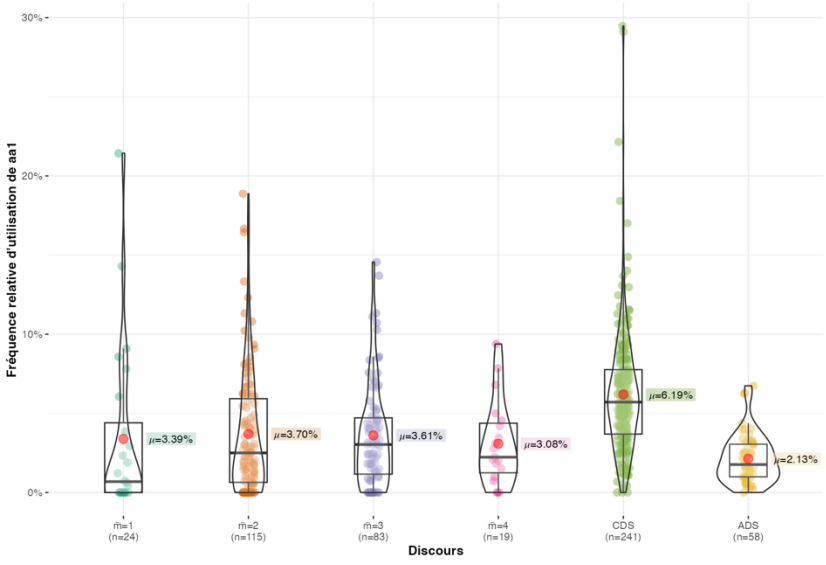

Figure 59. Évolution de la fréquence relative d'utilisation de *aa1* dans le CS, ainsi que dans le CDS et l'ADS

Ce diagramme nous montre la baisse de fréquence et de variabilité de la particule *aa1*, qui semble évoluer de manière linéaire vers le profil dans l'ADS.

8.6 laa1

La particule *laa1* peut être utilisée comme SFP dans les segments déclaratifs, pour ponctuer les affirmations ou les ordres et requêtes (elle a également une utilisation non finale, non considérée ici). En fin d'affirmation, elle sert à indiquer un manque d'exhaustivité, dans le sens où les items énumérés avant ne constituent pas la totalité des options. Quand elle est utilisée avec un ordre ou une requête, elle constitue la particule par défaut et ne porte pas de signification particulière (nous n'effectuons toutefois pas cette distinction dans notre étude).

(31) a. CHI: 我 個 家姐 都 有 **laa1** . (cc-lly030725)
ngo5 go3 gaa1ze1 dou1 jau5 **laa1**
moi CL grande-sœur aussi avoir **SFP**
J'ai aussi une grande sœur.

b. CHI: 食 月餅 **啦** . (cc-wbh021023)
sik6 jyut6beng2 **laa1**
manger gâteau-de-lune **SFP**
Mangeons des gâteaux de lune.

Son évolution en termes de couverture et de fréquence relative d'utilisation est très régulière :

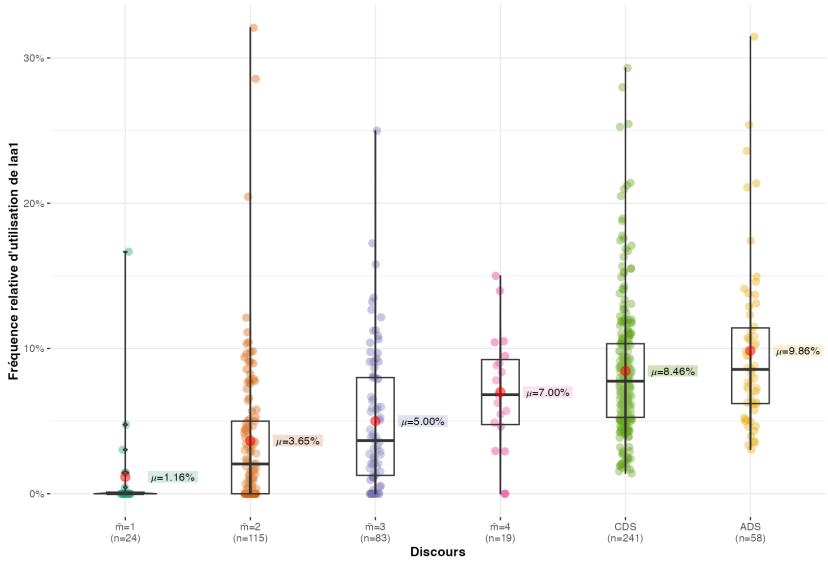

Figure 60. Évolution de la fréquence relative d'utilisation de *laa1* dans le CS, ainsi que dans le CDS et l'ADS

Elle est également attestée assez tôt en termes d'âge, avec une présence dans 24.1% des conversations à $\bar{a}$=1 et 76% à $\bar{a}$=2.

On constate à nouveau une légère préférence des filles pour cette SFP à $\bar{m}$=4 (χ^2=9.6674, df=1, p=0.001876, φ=0.0480), ainsi qu'une tendance générale à moins l'utiliser vis-à-vis d'interlocuteurs masculins (χ^2=22.507, df=1, p<0.001, φ=0.0314).

8.7 ge2

La particule *ge2*, comme *ne1*, peut être utilisée à la fin des énoncés déclaratifs, mais également à la fin des énoncés interrogatifs, en combinaison avec une structure interrogative ou de manière autonome (Kwok, 1984). Sa signification est en lien avec une dose de réserve ou d'incertitude dans les énoncés déclaratifs, et de remise en question ou d'incrédulité dans les questions.

(32 a. CHI: 不過 冇 筷子 **ge2** ! (hku-040021)
bat1gwo3 mou5 faai3zi2 **ge2**
mais avoir-NEG baguettes **SFP**
Mais il n'y a pas de baguettes!

b. CHI: 點解 冇 著 底褲 **ge2** ? (hku-050800)
dim2gaai2 mou5 zoek3 dai2fu3 **ge2**
pourquoi avoir-NEG porter sous-vêtements **SFP**
Pourquoi tu ne portes pas de sous-vêtements?

Elle est très peu présente à $\bar{m}$=1 (une seule conversation sur les 24 enregistrées), mais sa couverture augmente fortement après $\bar{m}$=2 pour atteindre 89.5% à $\bar{m}$=4 (Figure 61).

On constate également que les enfants l'utilisent globalement plus que les adultes, puisqu'elle représente 5.12% des SFP produites à $\bar{m}$=4, contre 1.05% dans le CDS et 2.17% dans l'ADS (Figure 62).

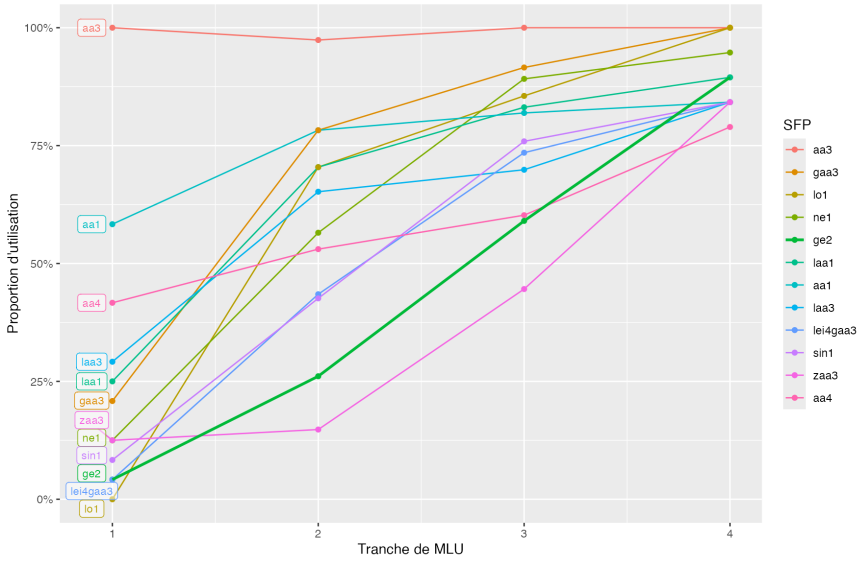

Figure 61. Évolution de la couverture de *ge2* dans le discours des enfants, comparée aux autres SFP

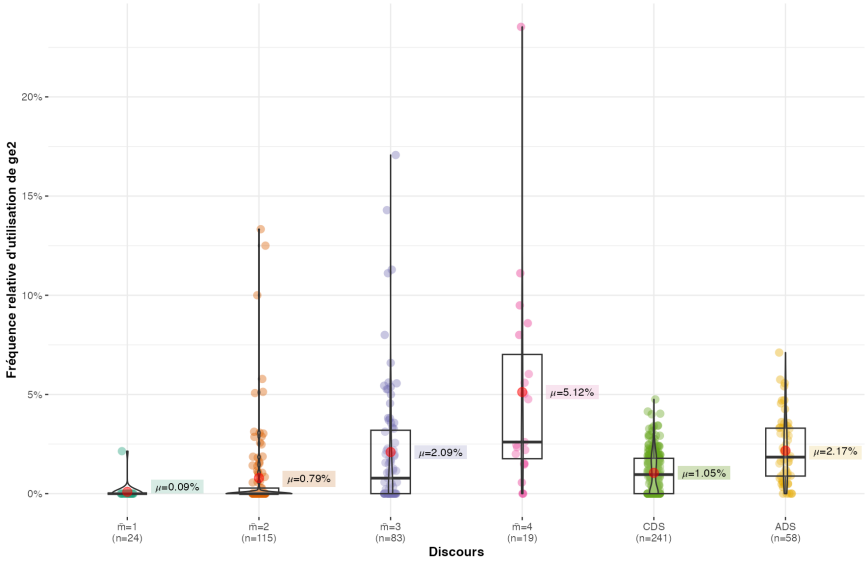

Figure 62. Évolution de la fréquence relative d'utilisation de *ge2* dans le CS, ainsi que dans le CDS et l'ADS

ge2 n'est attestée dans aucune conversation à $\bar{a}$=1. À $\bar{a}$=2, on l'observe dans 29% des conversations : huit conversations sur seize dans le HKU et sept enfants sur huit dans le CC. Le huitième enfant (CGK), observé de 1;11.01 (MLU=2.4) à 2;09.09 (MLU=2.9), ne la produit pour sa part jamais.

À la différence de *ne1*, pour laquelle on pouvait constater une claire évolution selon la MLU, son utilisation dans les différents types de segments (déclaratifs, interrogatifs avec et sans structure interrogative) reste

quant à elle plutôt stable, avec une légère baisse constatée pour les segments interrogatifs autonomes à m̄=4 :

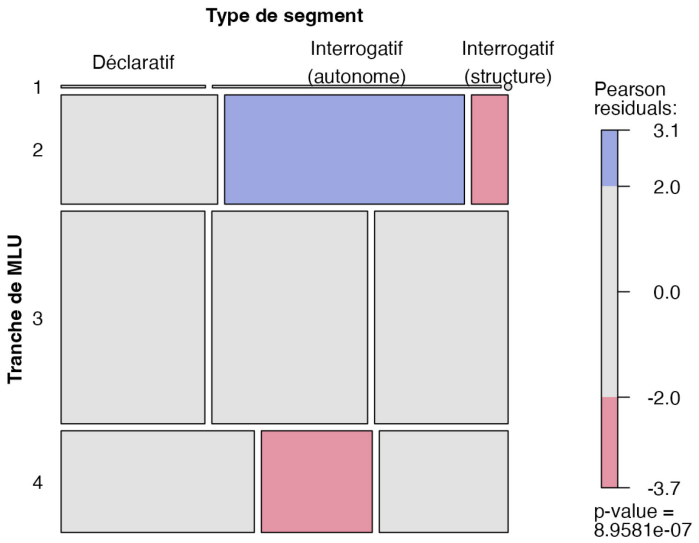

Figure 63. Répartition de l'utilisation de *ge2* selon le type d'énoncé et la tranche de MLU

8.8 aa4

La dernière particule que nous analysons en détail, *aa4*, est particulière du fait qu'elle est la seule, dans les 12 SFP les plus utilisées par les enfants à m̄=4, à être strictement interrogative : elle ne peut théoriquement pas être utilisée dans un énoncé déclaratif, puisqu'elle le transforme automatiquement en énoncé interrogatif requérant une réponse par oui ou par non. En tant que particule interrogative, *aa4* est considérée neutre et ne présuppose pas de préjugé de la part du locuteur (Hara, 2023).

(33) a. CHI: 你 鍾意 紫色 **aa4** ? (hku-050800)
ni5 zung1ji3 zi2sik1 **aa4**
toi aimer violet **SFP**
Est-ce que tu aimes le violet?

b. CHI: 全部 係 超人 **aa4** ? (cc-ccc020624)
cyun4bou6 hai6 ciu1jan4 **aa4**
tout être superhéros **SFP**
Est-ce que tous sont des superhéros?

Comme nous l'avons mentionné, le cantonais dispose de plusieurs mécanismes de formation d'énoncés interrogatifs, mutuellement exclusifs : l'intonation, qui n'est pas compatible avec les SFP; l'utilisation de

structures interrogatives (HAI_M_HAI, A_M_A) ou de mots interrogatifs (mots QU), qui peuvent être accompagnés de SFP; et l'utilisation de particules spécifiquement interrogatives, dont la présence transforme un énoncé déclaratif en énoncé interrogatif, et qui ne sont pas compatibles avec d'autres SFP. Comme nous l'avons vu à la section 6.3.3.2, les enfants, aux premiers stades du développement, ont tendance à former les questions sans utiliser de structures interrogatives complexes. La distribution de ces énoncés, mettant en avant l'utilisation de *aa4*, est présentée sur la Figure 64 :

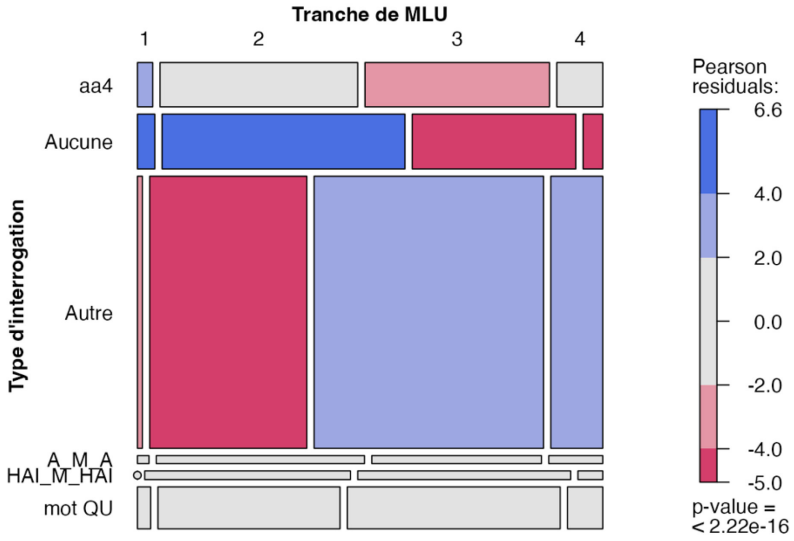

Figure 64. Distribution des formes d'interrogations par tranche de MLU

Ce diagramme montre que *aa4* est largement utilisée pour la formation des énoncés interrogatifs à $\bar{m}$=1, mais que son importance diminue par la suite pour laisser la place à d'autres SFP interrogatives. Cette interprétation est en partie confirmée par le diagramme de la Figure 65, qui présente l'évolution de la proportion d'utilisation des SFP interrogatives pour la formation de segments interrogatifs sans structure interrogative.

Avant toute chose, il est important de préciser que deux de ces SFP, *aa3* et *ge3*, ne sont pas des SFP interrogatives. Cependant, le fait qu'elles apparaissent dans le corpus avec des proportions non négligeables justifie de les mentionner – leur présence dans nos données peut avoir plusieurs explications : il peut s'agir d'erreurs de transcription, d'erreurs commises par les enfants, ou bien d'erreurs de traitement (pour rappel, certaines des notations utilisées dans les transcriptions sont ambiguës et pourraient mener à ce type de confusion).

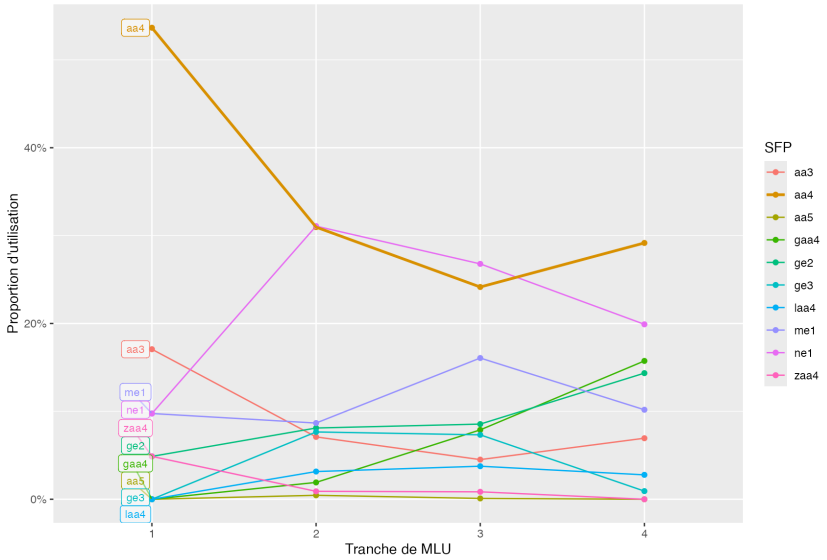

Figure 65. Évolution de la proportion d'utilisation des SFP interrogatives pour la formation de segments interrogatifs

Pour les SFP attestées comme strictement interrogatives (*aa4*, aa5, *gaa4*, *laa4*, *me1*, *zaa4*) ou épisodiquement interrogatives (*ne1*, *ge2*) (Kwok, 1984), le diagramme montre bien la prédominance de *aa4* à m̄=1. On note cependant la forte montée de *ne1* à m̄=2, bien que ces deux particules aient des utilisations différentes. Les autres particules (*me1*, *gaa4*, *ge2*) prennent également de l'importance à partir de m̄=2 et surtout m̄=3. La baisse de me1 et *ne1* à m̄=4 devra être approfondie.

La fréquence relative d'utilisation de *aa4* dans le CS reste toutefois stable au fur et à mesure de l'augmentation de la MLU, comme le montre la Figure 66. On y retrouve d'ailleurs dans le CDS la tendance accrue des adultes à formuler des questions à l'égard des enfants, en particulier au moyen de particule *aa4*.

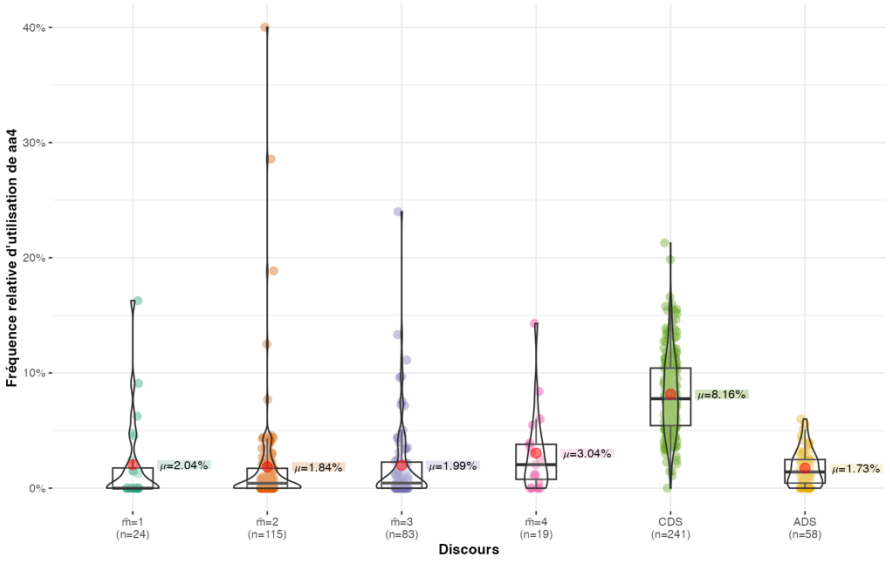

Figure 66. Évolution de la fréquence relative d'utilisation de *aa4* dans le CS, ainsi que dans le CDS et l'ADS

CHAPITRE 9

Interprétation et discussion

Les analyses présentées aux chapitres précédents, si elles ne font bien entendu qu'effleurer la surface de la complexité du processus d'acquisition des SFP par les enfants cantonophones, permettent tout de même d'entrevoir des éléments intéressants sur le rapport qui unit ces particules à l'environnement dans lequel les enfants développent leur compétence langagière. En particulier, nous pouvons constater que des facteurs issus des trois espaces que nous avions évoqués (individuel, linguistique et extrinsèque) cohabitent et interagissent dans l'émergence des différentes SFP, et que celles-ci entretiennent entre elles des rapports également complexes.

De nombreux aspects très spécifiques à l'environnement de l'acquisition des SFP en cantonais en rendent l'étude particulièrement délicate. Rappelons que le cantonais est une langue dont la forme écrite n'est pas enseignée officiellement et n'a pas été soumise à un processus de normalisation (Bauer, 2018). Il résulte de cette situation une variabilité de l'orthographe particulièrement manifeste dans la transcription des SFP. Celles-ci constituent par ailleurs une famille nombreuse de plusieurs dizaines d'éléments (dont la composition exacte ne fait cependant pas l'objet d'un consensus; une large proportion des éléments monosyllabiques, ainsi que certaines combinaisons dissyllabiques, sont toutefois attestés dans les principaux travaux), ce qui rend leur rôle au sein du cantonais très particulier et impossible à comparer avec celui des SFP présentes dans d'autres langues. Cette abondance est justifiée par les spécificités syntaxiques, sémantiques ou pragmatiques propres à chaque SFP, peu accessibles aux non-locuteurs. Enfin, le caractère peu écologique des conversations qui constituent nos corpus, l'austérité des données (accès limité aux enregistrements originaux, pauvreté des métadonnées, manque d'uniformité dans les transcriptions) ainsi que dans une certaine mesure leur vétusté (les 30 ans écoulés depuis la constitution des corpus que nous utilisons ne permettent pas d'avoir un aperçu réellement contemporain de l'état de la langue) rendent difficile une analyse précise des rôles joués par les différents facteurs dans le processus d'acquisition.

La place importante que prennent les SFP dans le discours des enfants au fil de leur développement langagier permet toutefois de suivre leur évolution et de voir se dessiner des schémas d'utilisation, grâce auxquels nous pouvons tenter d'expliquer certains aspects du processus. Dans ce chapitre, nous essayons

d'interpréter et de mettre en relation les différents résultats et observations issus de nos analyses pour proposer des réponses à la question de recherche que nous avons soumise (cf. chapitre 2) :

> Quelles interactions entre les facteurs individuels, linguistiques et extrinsèques l'analyse des échanges entre enfants et adultes permet-elle de mettre en évidence dans le cadre du processus d'acquisition des particules de fin de phrase en cantonais entre un an et demi et cinq ans et demi?

Nous évaluons également dans quelle mesure ces résultats permettent de statuer sur la validité des hypothèses de recherche que nous avons proposées au chapitre 2 :

- (H1) La MLU est un meilleur prédicteur que l'âge biologique pour l'utilisation des SFP (fréquence et diversité) dans le discours des enfants (§1.5.1.1); toutefois, l'âge biologique pourrait s'avérer jouer un rôle dans l'émergence de certaines SFP à valeur épistémique (gwaa3, wo4, aa1maa3) dans le discours.
- (H2) À âge égal, les filles ont une utilisation des SFP (fréquence, diversité) plus développée que les garçons (§1.6.2.2); l'impact de cette hypothèse doit cependant être modulé par le fait que les filles ont, à âge égal, une MLU plus élevée que les garçons, ce qui signifie que le genre n'est sans doute pas le facteur principal de cette différence. L'impact du genre sur l'utilisation des SFP individuelles sera observé, mais il est envisageable que la différenciation observée dans les études antérieures ne s'exprime qu'à un âge plus avancé.
- (H3) Le rôle de l'interlocuteur, qui caractérise le degré de familiarité avec l'enfant, a un impact sur la production de questions et donc sur l'utilisation des SFP qui y sont associées. On attend également un effet sur l'utilisation globale des SFP, bien que la nature de celui-ci (inhibiteur ou facilitateur) soit difficile à anticiper (§1.6.4.1). Le genre de l'interlocuteur, quant à lui, sera également observé, mais son impact sur l'utilisation des SFP à cet âge reste hypothétique.
- (H4) Le CDS (dans le cadre de nos corpus) exerce une influence locale et ponctuelle sur la production de SFP; en particulier, les formes des segments produits par les interlocuteurs, ainsi que les structures syntaxiques qu'ils y utilisent, favoriseront l'utilisation de certaines SFP.
- (H5) La fréquence et le choix des SFP utilisées dans les phrases interrogatives évolue en relation avec l'émergence de l'utilisation de structures syntaxiques interrogatives (§1.6.3.1).

Les analyses que nous avons effectuées et dont nous avons décrit les résultats aux chapitres précédents ont eu pour cibles deux objectifs distincts : d'un côté, la description du processus d'intégration des SFP en

tant que catégorie dans le discours des enfants (chapitre 7); et d'un autre côté, l'étude du comportement spécifique de certaines SFP en rapport avec nos axes d'observation (chapitre 8). Ces deux approches, bien qu'ayant des objets d'étude différents, peuvent toutefois être combinées pour affiner nos conclusions concernant le rôle des facteurs considérés.

Dans ce chapitre, nous effectuerons tout d'abord une interprétation pratique des résultats obtenus, organisée selon les différents types de facteurs considérés. Nous proposerons ensuite un exemple de représentation graphique permettant une visualisation globale des effets de ces facteurs. Nous procèderons par la suite à une discussion sur les différentes interactions à l'œuvre entre ces facteurs, et apporterons ainsi des réponses à notre question de recherche. Nous terminerons par une analyse des difficultés rencontrées au cours du projet et des limitations susceptibles d'avoir un impact sur la précision de nos conclusions.

9.1 Interprétation des résultats

9.1.1 Facteurs individuels

9.1.1.1 Âge et MLU

Les facteurs intrinsèques que nous avons observés étaient l'âge, le niveau de développement langagier tel qu'exprimé par la MLU, et le genre des enfants.

La MLU, considérée sous forme de tranches correspondant à une répartition homogène du corpus, était envisagée comme un fort prédicteur de la fréquence d'utilisation des SFP du fait des études antérieures (p. ex. Ko, 2000 ; Tse *et al.*, 2002) et du lien attendu entre l'utilisation de SFP et le niveau global de compétence langagière des enfants. La corrélation entre la MLU et l'utilisation des SFP a été confirmée dans toutes nos analyses, tant pour la propension des enfants à utiliser des SFP que pour l'observation des SFP individuelles dans le discours des enfants. Il nous a ainsi été possible de confirmer que, si de nombreuses SFP différentes peuvent effectivement être utilisées par les enfants très tôt dans leur processus d'acquisition ($\bar{m}$=1), leur diversité ainsi que la fréquence d'utilisation augmentent avec la MLU.

Nous avons pu observer que, bien que l'âge et la MLU soient fortement corrélés, il existe des différences importantes dans le rôle qu'ils jouent vis-à-vis de l'acquisition des SFP. L'ajustement de notre modèle générique décrivant la production des SFP a montré non seulement que l'utilisation de la tranche de MLU comme prédicteur est indispensable à l'ajustement d'un modèle pertinent, mais également que l'ajout de

la tranche d'âge mène lui aussi à l'obtention d'effets significatifs (§7.2.2.2). Les coefficients obtenus par ajustement du modèle intégrant la MLU et l'âge sont rappelés en Figure 67.

Figure 67. Estimations des coefficients du GLMM intégrant la tranche d'âge ($\bar{a}$) et la tranche de MLU ($\bar{m}$) pour la prédiction de la production de SFP

Ces résultats suggèrent que l'âge biologique joue effectivement un rôle dans le processus d'acquisition des SFP. Toutefois, l'hypothèse que nous avions émise sur l'effet de l'âge était à l'origine basée en particulier sur les conclusions de plusieurs études (Law, 2004 ; Lee et Law, 2000 ; Tang, 2018) concernant l'acquisition de la modalité épistémique en lien avec la Théorie de l'Esprit (§1.6.2.1). L'effet observé dans notre modèle ne peut être mis en relation avec les propositions de ces recherches (et ne les remet pas en question), qui se référaient spécifiquement à l'utilisation de certaines particules épistémiques (p. ex. *lo1*, exprimant l'évidence, ou *wo4*, exprimant la surprise). Nos résultats, qui mettent en évidence un effet inhibiteur de l'âge pour la production des SFP en général (coefficients négatifs), ne peuvent s'interpréter qu'en combinaison avec le développement langagier : si l'on observe effectivement une augmentation de la probabilité de produire des SFP parallèlement à l'augmentation de la MLU, l'effet inhibiteur de l'âge contrebalance cette corrélation dans une certaine mesure : les coefficients associés aux tranches de MLU sont plus élevés lorsque l'âge est intégré au modèle que lorsqu'il ne l'est pas. On note également que les coefficients associés aux tranches d'âge croissent, ce qui signifie que l'effet inhibiteur augmente avec l'âge.

Nous pouvons interpréter ces résultats en mettant en relation l'âge et la MLU dans nos observations de production de SFP, par exemple en considérant les taux d'utilisation des SFP (proportion de segments contenant une SFP) par âge et MLU. Ces observations sont présentées sur la Figure 68 :

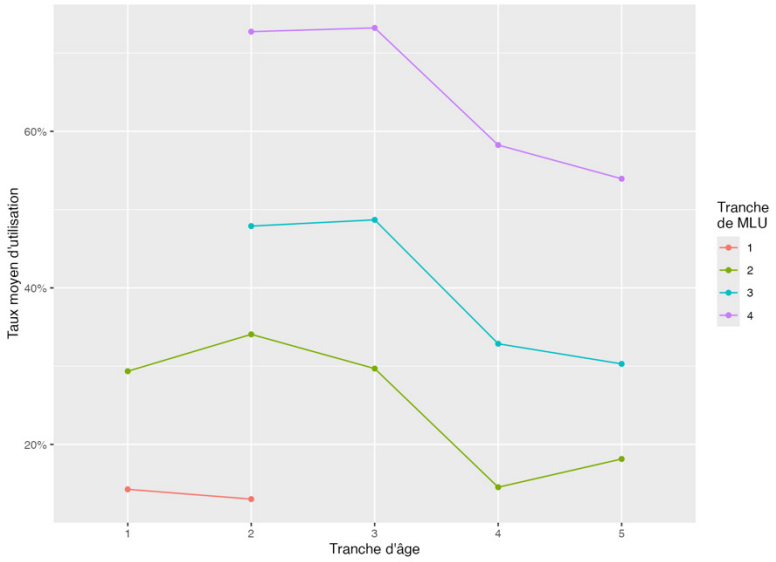

Figure 68. Taux moyen d'utilisation de SFP par tranche d'âge et tranche de MLU

Ce diagramme montre très clairement que, bien que la MLU et le taux d'utilisation des SFP soient fortement corrélés, les enfants atteignant la même tranche de MLU à un âge plus avancé ont tendance à utiliser moins de SFP[34]. Ces résultats confirment l'effet inhibiteur de l'âge mis en avant par notre modèle, et indiquent donc que la MLU n'est pas un prédicteur suffisant pour expliquer l'utilisation des SFP par les enfants. Il est intéressant de constater que ce résultat ne se retrouve que partiellement dans la diversité des SFP utilisées par les enfants (Figure 69).

L'effet globalement inhibiteur de l'âge est également présent, mais ne s'exprime pas de manière aussi systématique que pour le taux d'utilisation des SFP.

[34] Notons que les corpus semblent présenter un biais, selon lequel le taux de SFP dans le HKU est inférieur à celui du CC à MLU égales; la tendance constatée reste cependant la même si on ne considère que le HKU.

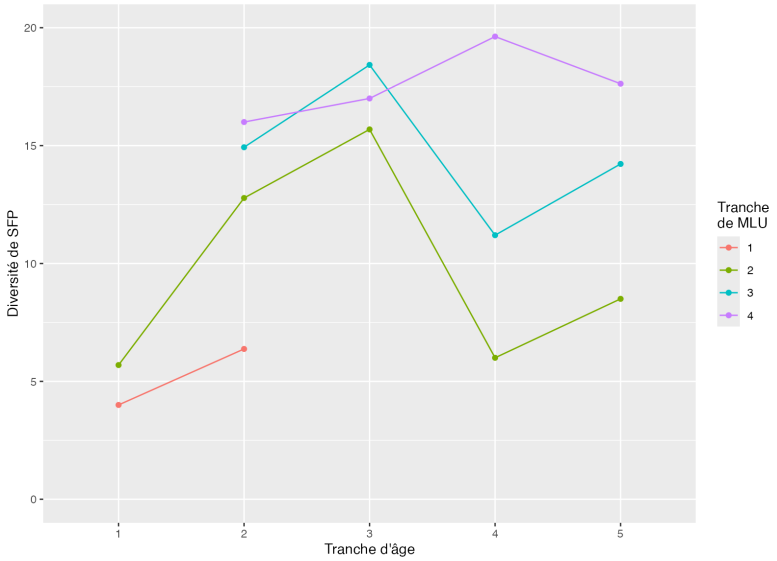

Figure 69. Diversité moyenne de SFP par tranche d'âge et tranche de MLU

Ces résultats suggèrent que l'utilisation des SFP relève non seulement du niveau de développement langagier tel qu'exprimé par la MLU, mais aussi fortement d'aptitudes individuelles liées au rythme de ce développement; une propension accrue à l'utilisation des SFP pourrait être un facteur d'augmentation de la MLU, de même que la tendance à faire des phrases plus complexes pourrait inciter le recours à des SFP. Tse *et al.* (2002) suggèrent pour leur part que l'augmentation de la MLU entre 3 et 5 ans serait plutôt liée à la complexification syntaxique.

L'effet de l'âge sur l'acquisition des SFP individuelles n'a, en contrepartie, pas pu être observé de manière systématique. Les modèles dédiés que nous avons ajustés pour leur analyse et incluant l'âge (sous forme de tranche d'âge, $1 \leq \bar{a} \leq 5$) en plus de la MLU n'ont pas permis d'obtenir de résultats probants permettant de confirmer un effet significatif de l'âge s'ajoutant à celui de la MLU. Cela peut s'expliquer par une trop grande variabilité dans les données, ou une insuffisance de données lorsqu'on considère les SFP individuellement.

9.1.1.2 Genre

Le genre des enfants a pour sa part également des effets mitigés. Bien que nous ayons pu confirmer que la MLU est significativement plus élevée, à âge égal, chez les filles que chez les garçons (§4.5.3.1), la production de SFP en lien avec le niveau de développement ne semble pour sa part pas affectée. Ainsi, on ne retrouve pas de manière significative une tendance des filles à utiliser plus de SFP à MLU égale, bien

qu'elles atteignent théoriquement cette MLU à un âge plus précoce que les garçons. La Figure 70 présente l'évolution comparée des taux moyens d'utilisation de MLU pour filles et garçons, selon la tranche de MLU et la tranche d'âge. Comme nous l'avions observé à la section 6.2.4, les taux d'utilisation à MLU égale restent comparables. On retrouve la tendance globale à la baisse de l'utilisation de SFP, mais elle semble s'appliquer de manière similaire chez les filles et les garçons :

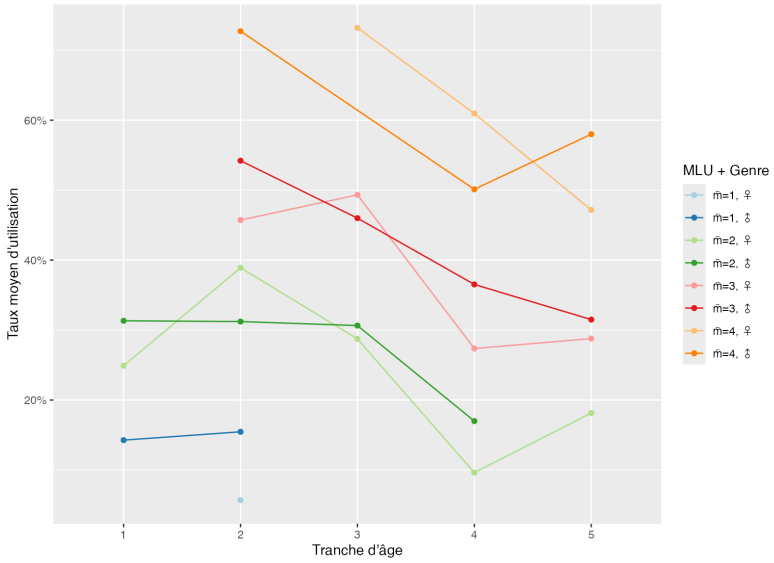

Figure 70. Comparaison des taux moyens d'utilisation des SFP par genre, MLU et âge

9.1.2 Facteurs linguistiques

9.1.2.1 Types de segments

Nous avons pu constater de manière très univoque que les SFP étaient utilisées avec des fréquences nettement plus élevées pour les segments interrogatifs que pour les segments déclaratifs, tant chez les adultes que chez les enfants (§6.3.3.1). Cette tendance est illustrée pour les différentes tranches de MLU, ainsi que pour le CDS et l'ADS, sur la Figure 71.

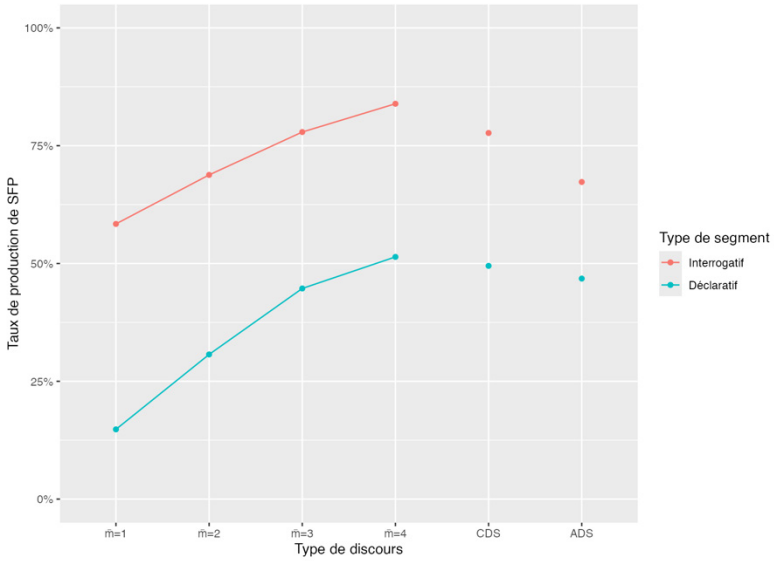

Figure 71. Évolution de la fréquence d'utilisation des SFP pour les segments interrogatifs et déclaratifs, par tranche de MLU et pour le CDS et l'ADS

Le fait que cette asymétrie se retrouve de manière si marquée très tôt dans le discours des enfants, et très proche du début de leur utilisation des SFP, est probablement surprenant, d'autant plus que les segments interrogatifs ne représentent à ce stade du développement langagier qu'une partie très minime du discours (1.4% à m̄=1). Il convient donc de se demander si l'utilisation des SFP et les énoncés interrogatifs seraient susceptibles de se développer de manière concomitante dans le discours des enfants. Il est intéressant de noter que la particule *aa3*, dont on sait qu'elle est utilisée également très tôt et très fréquemment par les enfants (utilisation par 100% des enfants observés à m̄=1, âgés de 1;05.22 à 2;05.20), caractérise elle aussi cette asymétrie précoce : bien qu'elle représente 87.5% des occurrences de particules dans des segments déclaratifs (1181 sur 1350) contre 36% dans les énoncés interrogatifs (29 sur 80) à m̄=1, son utilisation correspond à 22% de l'ensemble des segments interrogatifs, contre seulement 13% des segments déclaratifs : on constate donc dès les premières phases de l'acquisition une discrimination dans le discours des enfants. La particule *aa4*, particule strictement interrogative servant à transformer un énoncé déclaratif en énoncé interrogatif total, représente pour sa part 27.5% des occurrences de particules dans les segments interrogatifs (22 sur 80).

Il est également pertinent de s'interroger sur l'origine de cette asymétrie. Le CDS joue indéniablement un rôle important dans l'appropriation des questions par les enfants, et des SFP qui les ponctuent; on trouve d'ailleurs une plus grande proportion de questions dans le CDS que dans l'ADS (36.8% contre 16.2%; cf. §4.5.3.2), ainsi qu'une utilisation plus fréquente des SFP (77.7% contre 67.3%). *aa3* y est elle aussi utilisée

plus fréquemment, avec 45.4% des occurrences de SFP dans les segments interrogatifs, contre 36% dans l'ADS; elle est de plus largement plus fréquente dans le CDS que l'absence de SFP (35.3% du total des segments interrogatifs contre 22.3% sans SFP), ce qui ajoute à la différence entre le CDS et l'ADS (*aa3* : 24.2%, aucune SFP : 32.7%). Enfin, *aa4*, première particule strictement interrogative maîtrisée par les enfants, est également plus présente dans le CDS (16.8% des occurrences de particules dans les segments interrogatifs) que dans l'ADS (10.5%), et la deuxième SFP la plus utilisée dans les segments interrogatifs après *aa3*. L'exemple (34) montre une situation dans laquelle l'enfant reprend clairement la formulation interrogative de l'adulte, mais l'intègre à sa propre phrase :

(34)

INV:	佢	話	凍	唔	凍	aa3 ?		(hku-030606)
	keoi5	waa2	dung3	m4	dung3	aa3		
	il/elle	dire	froid	NEG	froid	SFP		
	Il/elle demande s'il fait froid?							
CHI:	凍	唔	凍	aa3 ,	唔	凍	喎！	
	dung3	m4	dung4	aa3	m4	dung3	wo3	
	froid	NEG	froid	SFP	NEG	froid	SFP	
	S'il fait froid, bien sûr qu'il ne fait pas froid!							

9.1.2.2 Structures syntaxiques

Dans le cadre des segments interrogatifs, nous avons pu constater l'existence d'une interaction entre les structures ou éléments syntaxiques remarquables (structures HAI_M_HAI et A_M_A, mots QU) et la production de SFP par les enfants : le GLMM que nous avons ajusté spécifiquement pour ce contexte a montré l'effet inhibiteur de ces composants linguistiques sur la production des SFP chez les enfants (§7.3). En contrepartie, on voit sur la comparaison des modèles pour le CDS et l'ADS que l'utilisation de ces structures n'a que peu ou pas d'impact sur la production de SFP chez les adultes, qui eux sont en mesure de combiner la création d'une forme interrogative avec l'expression d'une composante sémantique ou pragmatique.

L'absence de structures syntaxiques interrogatives entretient bien entendu également un rapport étroit avec les SFP puisqu'elle valide l'utilisation des SFP strictement interrogatives *aa4*, *aa5*, *gaa4*, *laa4*, *me1*, *zaa4*, ainsi que l'utilisation interrogative de *ne1* et *ge2*. Ces particules sont pour certaines utilisées par une large proportion d'enfants à $\bar{m}$=4, comme le montre la Figure 72 :

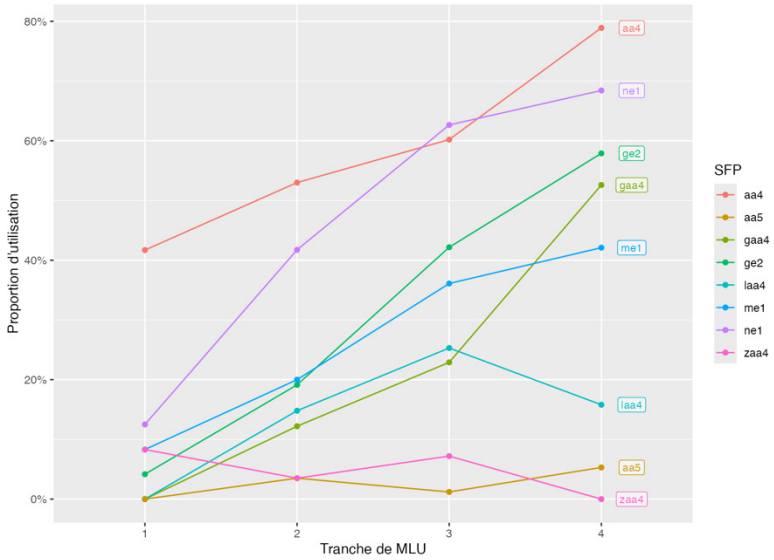

Figure 72. Évolution de la couverture des particules interrogatives selon la tranche de MLU[35]

On note sur ce diagramme l'évolution quelque peu singulière de *zaa4* (particule indiquant une remise en question de la portée sémantique de l'énoncé déclaratif transformé en question; Sybesma et Li, 2007), que l'on trouve chez 8.3% des enfants à m̄=1 mais chez 0% à m̄=4. Cela est très probablement lié au fait que *zaa4* est très peu présente dans les transcriptions du HKU (seulement 4 occurrences chez les enfants et 28 chez les adultes sur l'ensemble du corpus), mais peut également avoir pour cause des erreurs de transcription ou d'identification de la SFP.

Ces SFP interrogatives restent toutefois globalement rares dans le discours, puisqu'elles ne représentent qu'une fraction de l'utilisation des SFP : mis à part *aa4* dans le CDS, qui totalise presque 8% des occurrences, la majorité des particules ne dépassent pas 2% de fréquence relative, et certaines n'atteignent pas 0.1%. On note toutefois (Figure 73) que plusieurs fréquences d'utilisation dans le CS à m̄=4 (en particulier celles de *ne1*, *gaa4* et *ge2*) paraissent très élevées comparées à celles observées dans l'ADS, et semblent se rapprocher plus de celles observées dans le CDS. Comme nous avons déjà pu le constater, l'émergence des SFP et celle des interrogations dans le discours des enfants semblent liées, il est donc possible que les enfants s'inspirent plus fortement des schémas alliant ces deux aspects dans le discours entendu.

[35] Ce graphique et le suivant ne tiennent compte que des usages de *ne1* et *ge2* comme particules interrogatives autonomes, c'est-à-dire sans cooccurrence d'une structure interrogative (A_M_A ou HAI_M_HAI) ou d'un mot QU.

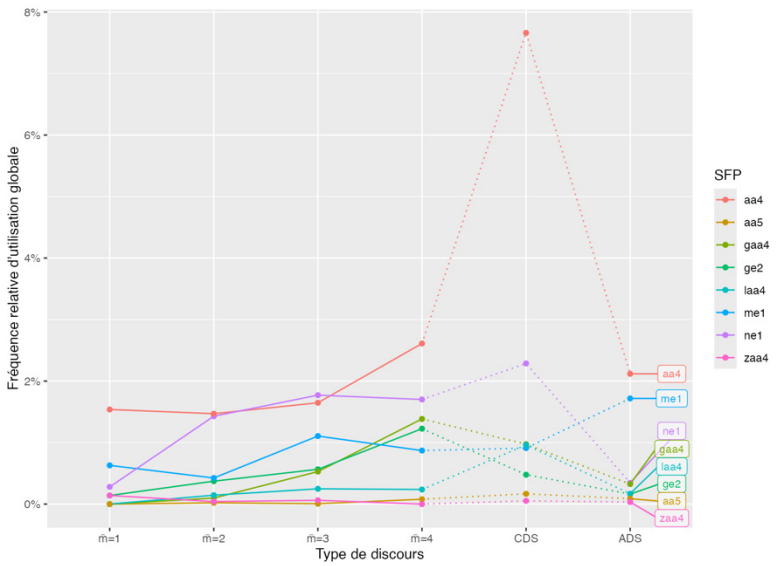

Figure 73. Évolution de la fréquence relative d'utilisation moyenne des SFP interrogatives par tranche de MLU, ainsi que dans le CDS et l'ADS

9.1.2.3 Phono-prosodie

Comme nous l'avons vu au chapitre 7 dans les différents modèles de régression que nous avons ajustés, le nombre de syllabes d'un énoncé joue un premier rôle crucial mais trivial dans le processus de production des SFP : comme celles-ci ne peuvent que ponctuer un énoncé, mais pas en constituer un à elles seules, un énoncé d'une seule syllabe ne peut contenir de SFP. De ce fait, n'importe quel modèle de prédiction de la production de SFP doit intégrer ce facteur.

L'effet facilitateur du nombre de syllabes se retrouve également, bien que dans une moindre mesure, dans le CDS. Nous avons par ailleurs ajusté le même modèle qu'au chapitre 7, mais restreint tout d'abord aux segments du CS de plus d'une syllabe, puis aux segments de plus de deux syllabes. Dans les deux cas, le nombre de syllabes restait significatif et facilitateur (est.$_{>1syll}$=0.201, p<0.001; est.$_{>2syll}$=0.05, p<0.001). Ainsi, bien que la longueur des segments soit clairement directement corrélée à la MLU, et que ces deux valeurs partagent donc une partie de leur effet sur la production des SFP, le fait de produire des segments plus longs semble avoir un impact propre, potentiellement lié à l'effet facilitateur constaté pour l'augmentation précoce de la MLU.

La question de la sous-production de SFP pour les segments de longueur 2 (et, dans l'ADS, pour les segments de longueur 3), constatée à la section 6.3.3.3, reste quant à elle ouverte. De manière attendue, ils offrent *de facto* peu d'espace pour exprimer un contenu en jeu lorsqu'une syllabe est déjà occupée par

une SFP. Cependant, et sans même prendre en compte le fait que beaucoup de morphèmes sont toutefois monosyllabiques en cantonais, il semblerait logique qu'une limitation en lien avec la complexité du contenu exprimé s'applique également, de manière inversement proportionnelle, aux segments de taille 3 ou supérieure, ce qui ne semble pas être le cas (c'est-à-dire que des segments de 3 syllabes seraient moins propices à accueillir une SFP en plus du contenu que des segments de longueur 4 ou plus). Les observations effectuées montrent par ailleurs que la distribution des SFP est sensiblement similaire pour les segments de taille 2 et les segments de taille supérieure (mis à part les SFP dissyllabiques, peu utilisées par les enfants à ce niveau). Il serait donc pertinent d'analyser le contenu hors SFP exprimé dans ces énoncés, pour évaluer sa spécificité vis-à-vis des SFP. Une piste de recherche pourrait être que la proportion de segments interrogatifs (pour lesquels les SFP sont largement plus utilisées que pour les segments déclaratifs) semble être également moins importante, dans tous les types de discours, pour les segments de taille 2 (et, pour l'ADS, également pour les segments de longueur 3). Toutefois, les segments interrogatifs étant largement minoritaires dans le CS, la différence ne serait pas suffisante pour justifier entièrement le manque de SFP.

9.1.3 Facteurs extrinsèques

Nos modèles et observations de contingences ont permis de mettre en avant plusieurs facteurs extrinsèques jouant un rôle important dans la production de SFP chez les enfants.

9.1.3.1 Rôle du CDS

L'un des facteurs extrinsèques importants est bien entendu le discours adressé aux enfants. Il a été largement rapporté comme ayant un impact majeur sur le comportement langagier des enfants, même si certaines études relativisent son importance par rapport à un discours ambiant, non spécifiquement dirigé vers les enfants. Pour notre étude, nous avons analysé aussi bien le CS que le CDS, et en particulier les éléments du CDS qui semblent avoir un impact sur le CS dans le cadre de la production des SFP.

En particulier, nous avons pu observer que l'utilisation d'une SFP dans le dernier segment produit par l'interlocuteur précédent (la personne ayant pris la parole juste avant l'enfant) avait un impact significatif sur la tendance de l'enfant à utiliser lui-même une SFP. Ce comportement indique clairement l'influence du discours entendu sur le discours produit par l'enfant. Il est intéressant d'observer que cette tendance change au cours du développement – de très significative à $\bar{m}=2$, elle devient moyennement significative

à m̄=3 et non significative à m̄=4 (la production est trop faible à m̄=1 pour que les effets soient observables) :

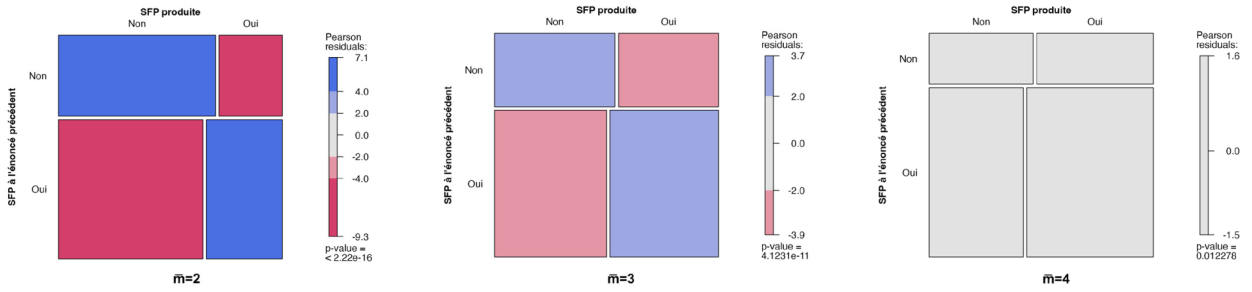

Figure 74. Évolution de la contingence entre l'utilisation d'une SFP à l'énoncé précédent et la production d'une SFP dans l'énoncé courant, par tranche de MLU

L'influence du CDS sur les SFP individuelles est pour sa part difficile à cerner. S'il est globalement clair que les particules les plus utilisées par les adultes sont également celles que l'on finit par retrouver le plus dans le discours des enfants, la symétrie ne s'applique cependant totalement ni aux fréquences d'utilisation, ni au rythme auquel les SFP apparaissent dans le discours des enfants. La Figure 75 présente l'évolution des fréquences relatives d'utilisation moyennes de sept des particules décrites au chapitre précédent (du fait de ses fréquences d'utilisation nettement plus élevées, nous avons placé *aa3* dans un encart pour permettre une meilleure représentation des variations des fréquences des autres SFP).

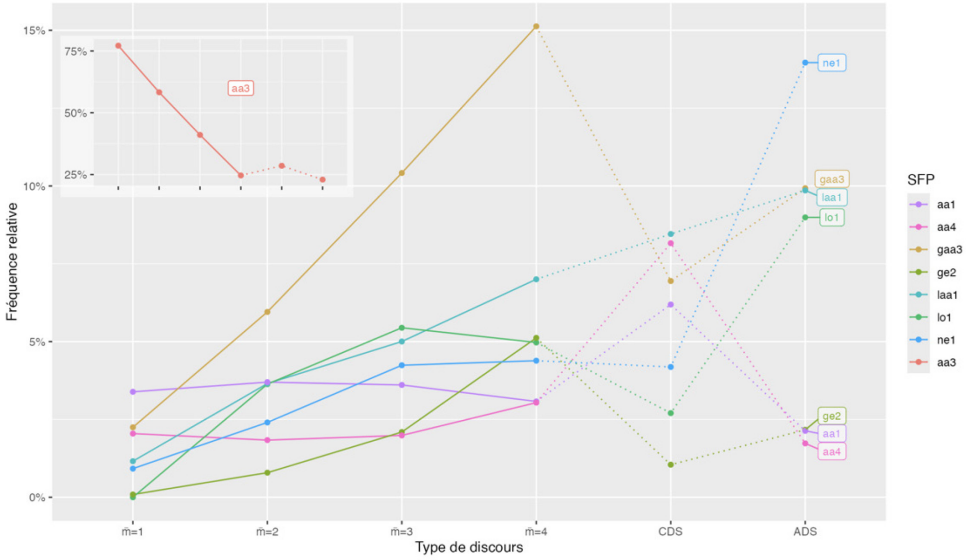

Figure 75. Évolution des fréquences relatives d'utilisation moyennes de 8 particules, par tranche de MLU ainsi que dans le CDS et l'ADS

Les huit SFP les plus utilisées par les enfants et les adultes sont sensiblement les mêmes (les principales particules utilisées fréquemment par les adultes mais peu par les enfants sont wo3, indiquant le caractère informatif et intéressant d'une information, et *ge3*, indiquant le haut degré de confiance du locuteur dans la pertinence d'une information). On remarque toutefois des différences majeures dans leur distribution.

Par exemple, *gaa3* (analysée par certains comme une version « atténuée » de *ge3*) est beaucoup plus utilisée par les enfants que par les adultes, que ce soit dans le CDS ou l'ADS. Cette observation indique que le besoin pragmatique couvert par *gaa3*, en lien avec l'expression de la confiance, est différent chez les adultes et les enfants, peut-être parce que ceux-ci se lancent dans des processus de négociation. Le choix de *gaa3* plutôt que *ge3* pourrait être dû au fait que *ge3* est globalement moins utilisée que *gaa3* dans le CDS comme dans l'ADS.

aa4 (particule interrogative neutre servant à former des interrogations totales) est utilisée de manière similaire dans le CS et dans l'ADS, mais beaucoup plus dans le CDS, ce qui pourrait correspondre à un comportement des adultes à poser plus fréquemment des questions aux enfants, par exemple pour les inciter à parler, clarifier ou reformuler des énoncés incorrects. Les enfants ne semblent dans ce cas pas s'approprier ce comportement et leur utilisation de *aa4*, qu'ils commencent à maîtriser très tôt (41.7% de couverture à $\bar{m}$=1), reste assez stable au cours de leur développement.

À l'instar de *aa4*, *aa1* (de valeur similaire à *aa3*, mais de connotation plus incitative) a également une utilisation très stable tout au long du développement, et très comparable à l'utilisation qu'en font les adultes entre eux; elle jouit cependant d'un statut particulier dans le CDS où les adultes, en s'adressant aux enfants, l'utilisent de manière plus soutenue. On peut imaginer que cette utilisation correspond à un effort de la part des adultes (peut-être particulièrement de la part des expérimentateurs) pour stimuler les enfants et les inciter à s'exprimer. Les enfants, quant à eux, ne semblent pas reproduire cette dynamique particulière dans le contexte des conversations enregistrées.

ne1 est beaucoup plus présente dans l'ADS que dans le CS ou le CDS, mais sa polyvalence (elle peut être utilisée pour ponctuer des énoncés déclaratifs et des énoncés interrogatifs, ou être ajoutée à un énoncé déclaratif pour en faire une interrogation) rend son analyse complexe. On remarque d'emblée que ses acceptions sont réparties de manière très différente entre les discours : dans l'ADS (14% de fréquence relative moyenne), 79.9% des occurrences accompagnent des segments déclaratifs, où elle sert à attirer l'attention sur un élément de l'environnement; dans le CDS, 76.4% se trouvent dans des segments

interrogatifs, parmi lesquels 81.3% sont des utilisations autonomes utilisées pour interroger sur un aspect de comparaison; dans le CS, elle est utilisée de manière assez équitable, avec 54.6% d'utilisation dans des énoncés interrogatifs, mais 90% de ceux-ci correspondent également à une utilisation autonome. Il serait intéressant de suivre de manière plus granulaire l'évolution de chacune des acceptions à travers le CS et jusqu'à l'ADS pour comprendre ces répartitions.

lo1 (marqueur d'évidence) apparaît assez tard dans le CS (couverture nulle à $\bar{m}$=1) mais y prend par la suite rapidement de l'importance (5% à $\bar{m}$=4). Elle est par ailleurs peu présente dans le CDS (2.7%), mais très fréquente dans l'ADS (9%). On peut suggérer une évolution du rapport épistémique de l'enfant vis-à-vis de son environnement, qui l'incite à exprimer sa notion de la réalité avec plus d'insistance, alors que des adultes s'adressant à des enfants ont sans doute moins tendance à pointer les évidences comme telles qu'à les indiquer de manière informative.

Notons également le comportement de *ge2*, qui jouit de la même polyvalence que *ne1*, mais dont la signification comporte toujours une notion d'incertitude. Avec 5.1% d'utilisation à $\bar{m}$=4, elle est largement plus présente chez les enfants que dans le CDS (1.0%) ou dans l'ADS (2.2%). Là encore, on peut imaginer que cette utilisation est en lien avec le fait que les enfants recherchent l'approbation de leurs interlocuteurs adultes, en particulier en répondant à leurs questions, et que les adultes s'adressant aux enfants ont pour leur part rarement de raison d'exprimer une incertitude.

Ces observations, pour exploratoires qu'elles soient, suggèrent en tous les cas que, même si les caractéristiques du CDS (et très probablement de l'ADS) tiennent une place importante dans le schéma d'acquisition des SFP par les enfants, les aspects pragmatiques propres à chacune des particules jouent un rôle primordial dans l'usage qu'ils en font. Ces deux aspects sont nécessairement liés : exposés aux particules utilisées par leurs interlocuteurs, les enfants se les approprient et semblent les remobiliser rapidement dans les contextes pertinents pour répondre à leurs propres besoins pragmatiques, différents de ceux des adultes.

9.1.3.2 Caractéristiques de l'interlocuteur

Comme nous l'avons évoqué aux chapitres précédents, certaines caractéristiques de l'interlocuteur influencent soit la production de SFP de manière globale, soit l'utilisation de certaines SFP.

Le genre de l'interlocuteur semble avoir un impact significatif, mais variable, sur le fait que les enfants décident ou non d'utiliser une SFP : comme l'indiquait notre table de contingence à la section 6.3.4.1, les enfants utilisent globalement plus souvent des SFP lorsqu'ils s'adressent à des femmes que lorsqu'ils s'adressent à des hommes (ce résultat s'inverse cependant lorsqu'on n'observe que les segments interrogatifs, pour lesquels l'effet du genre masculin rapporté par notre modèle en §7.3 est facilitateur).

Ainsi que nous l'avons vu à la section 6.3.4.2, les enfants utilisent également plus fréquemment des SFP lorsqu'ils interagissent avec des personnes étrangères au foyer, qui pour leur part en utilisent également plus que les personnes du foyer lorsqu'ils s'adressent aux enfants.

Notons que les observations en lien avec les interlocuteurs et interlocutrices extérieurs au foyer sont nécessairement imprécises du fait de l'absence d'information à ce sujet, ainsi que du fort déséquilibre entre femmes et hommes dans les conversations où cette information a été inférée. Ainsi, il serait utile d'effectuer une analyse plus ciblée sur l'utilisation des SFP en lien avec ces facteurs dans un environnement mieux contrôlé pour pouvoir vérifier nos hypothèses.

Ces deux effets semblent en tous les cas confirmer l'influence du CDS sur la production de SFP des enfants, au moins de manière instantanée. Une conclusion secondaire qui en découle est que les enfants auront tendance à utiliser moins de SFP vis-à-vis des membres masculins de leur foyer, comme par exemple leur père. Cette hypothèse se trouve confirmée dans nos données :

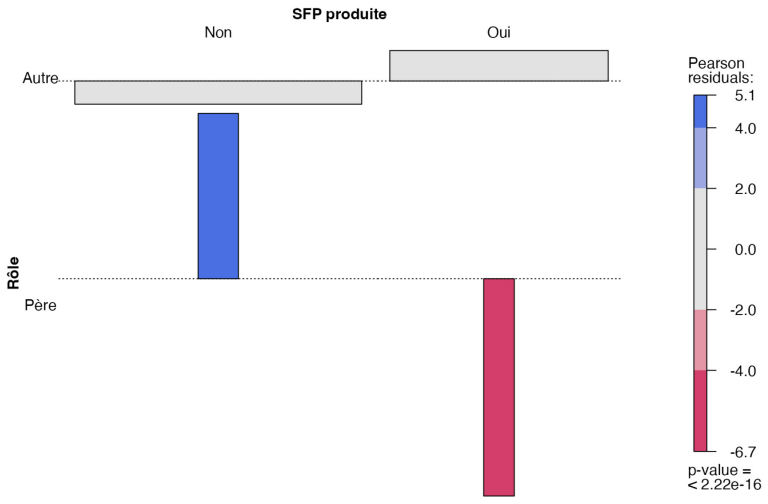

Figure 76. Résidus de Pearson de la table de contingence de la production de SFP et du rôle de l'interlocuteur (père vs autre)

9.2 Visualisation

Les modèles statistiques sont des objets relativement abstraits, dont les effets sont complexes à appréhender du fait du nombre potentiellement important de dimensions dont ils sont composés. La visualisation *break-down* permet de visualiser ces effets en appliquant le modèle à des situations spécifiques. Prenons par exemple la situation présentée en (35) :

(35) MOT: 呢 咩 嚟㗎 ? (cc-011008-ccc)
ni1 me1 **lei4gaa3**
et quoi **SFP**
Qu'est-ce que c'est?
CHI: 單車 **aa3** .
daan1ce1 **aa3**
vélo **SFP**
C'est un vélo.

Lors de cette conversation, l'enfant CCC (corpus CC) a un an, dix mois et huit jours, et sa MLU (calculée sur la base de la conversation entière) est de 2.26. L'enfant (CHI) répond à la question de sa mère (MOT), qui contient une SFP (*lei4gaa3*). Le modèle décrit en 7.2.2 (dont les coefficients sont rappelés à la section 9.1.1.1 de ce chapitre), incluant la tranche d'âge en plus de la tranche de MLU, peut être appliqué à cette situation spécifique, et les coefficients associés aux valeurs des différents facteurs peuvent être représentés au moyen du graphique *break-down* présenté en Figure 77[36].

Cette représentation permet de visualiser les effets des différents facteurs intégrés à notre GLMM pour le cas spécifique proposé. Les écarts rouges indiquent les effets inhibiteurs sur la probabilité de produire une SFP, les écarts verts les effets facilitateurs. On observe alors que, pour l'enfant CCC, la tranche d'âge $\bar{a}$=1 a un effet fortement positif sur la production de SFP pour la tranche de MLU $\bar{m}$=2 (effet facilitateur dû à la précocité). On retrouve également les effets inhibiteurs associés aux segments déclaratifs et à la proximité familiale de l'interlocutrice, ainsi que les effets facilitateurs du nombre de syllabes, de la présence d'une SFP à l'énoncé précédent et du type interrogatif de l'énoncé précédent. Le coefficient associé à l'identificateur de l'enfant, intégré au modèle comme effet aléatoire, indique que l'enfant CCC montre une forte tendance à utiliser des SFP (+0.178). La prédiction négative associée à cette combinaison de facteurs (−0.284) nous indique que, selon notre modèle et dans la situation proposée, l'enfant aura une légère tendance à ne pas utiliser de SFP. Notons que, dans ce cas précis, la prédiction ne correspond pas au fait

[36] Graphique réalisé au moyen de la bibliothèque R DALEX (Biecek, 2018).

que l'enfant a tout de même produit la SFP *aa3*. Il faut bien entendu garder à l'esprit que les modèles statistiques n'ont pas pour objectif de fournir des résultats correspondant exactement aux cas observés – une précision trop élevée est généralement le résultat d'un surajustement (*overfitting*), indiquant que le modèle n'est pas en mesure de s'adapter à la variabilité des facteurs.

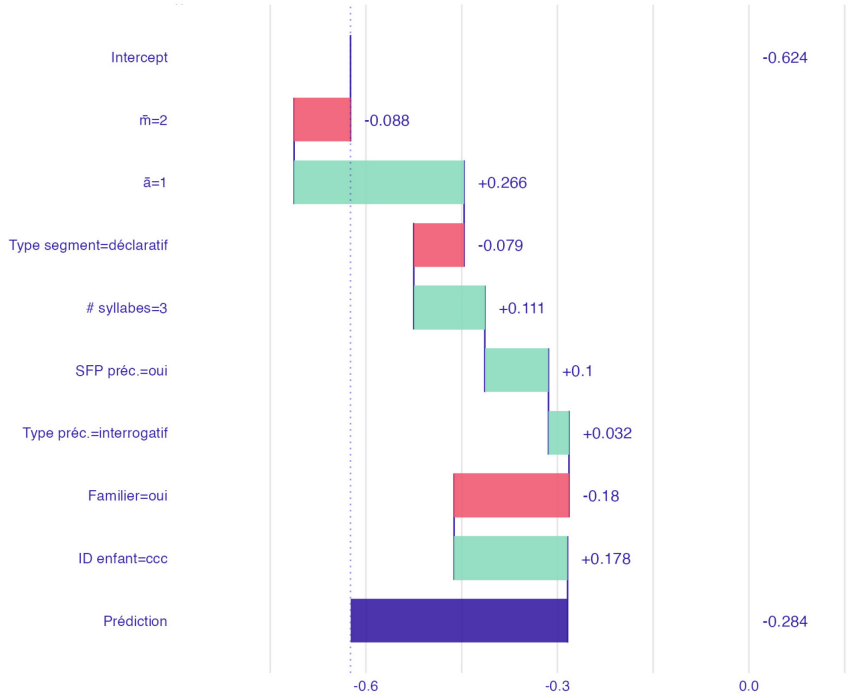

Figure 77. Visualisation *break-down* des effets des facteurs de notre GLMM pour l'énoncé en (35)

9.3 Discussion

9.3.1 Hypothèse 1 : supériorité de la MLU sur l'âge comme prédicteur pour la production de SFP

L'interprétation de nos résultats nous permet de statuer sur la validité de nos hypothèses de recherche. En premier lieu, nous avons pu effectivement vérifier la corrélation plus forte qu'avec l'âge de la mesure du développement langagier (MLU) avec la fréquence globale d'utilisation des SFP, ainsi qu'avec la diversité des SFP utilisées. Si elle correspond à un résultat attendu, cette conclusion confirme cependant que l'intégration des SFP dans le discours s'inscrit essentiellement dans un processus de construction de la compétence langagière, et non de maturation cognitive, comme cela aurait été le cas si la corrélation avec l'âge avait été plus forte. Cette constatation justifie de prioriser la MLU dans les études de phénomènes langagiers.

Nous avons toutefois constaté une interaction intéressante entre l'âge et la MLU, selon laquelle l'augmentation de la production de SFP en rapport avec la MLU décroît avec l'âge; ainsi, deux enfants de MLU équivalente et d'âges différents ne produisent pas les SFP avec la même fréquence, et la précocité de l'augmentation de la MLU semble être corrélée avec une propension accrue pour l'utilisation des SFP. Cette observation, si elle est confirmée, aurait un impact majeur sur la recherche en acquisition : en effet, la majorité des études se focalisent sur l'une ou l'autre de ces mesures (MLU et âge) comme critère de catégorisation ou comme prédicteur. La possibilité d'une interaction entre ces deux facteurs justifierait de revoir les modèles d'acquisition pour y intégrer les deux, et ainsi effectuer une analyse plus contextualisée des schémas de développement. Notre interprétation doit toutefois être modérée par deux considérations : d'une part, la forte corrélation qui unit l'âge et la MLU (observée dans notre étude et confirmée par de nombreux travaux antérieurs) est de nature à biaiser la pondération de ces facteurs dans l'adaptation d'un modèle statistique. Il est donc nécessaire de pallier cette difficulté pour pouvoir évaluer les effets indépendants de chacun des facteurs, ce que nous n'avons pas eu le temps de faire. D'autre part, le fait que nous ayons unifié deux corpus différents doit être considéré comme potentielle source d'imprécisions, puisque les deux corpus émanent de sources différentes; si nous avons fait notre possible pour les aligner au niveau des notations, il est malgré tout possible que d'autres différences existent entre les deux, en particulier en raison du manque de formalisme concernant le cantonais (règles de transcription, segmentation, inventaire de SFP).

Nous avons également observé que certaines SFP très fréquentes à $\bar{m}$=4 n'apparaissent dans le discours des enfants qu'à une période assez tardive, bien qu'il soit difficile de déterminer les rôles respectifs de l'âge biologique et du niveau de développement langagier. En particulier, *lo1*, exprimant une notion d'évidence et de certitude, et *ge2*, exprimant une notion d'incertitude, ne sont pour ainsi dire pas du tout utilisées à $\bar{a}$=1 et $\bar{m}$=1, mais connaissent une progression importante à $\bar{a}$≥2 et $\bar{m}$≥2. Cette observation semble compatible avec un phénomène de maturation de fonctions cognitives, comme le suggère Tang (2018), qui associait l'acquisition des fonctions épistémiques à l'émergence de la théorie de l'esprit. Elle explique par ailleurs l'évolution de l'utilisation d'autres SFP comme *aa1* ou *ge3*, présentes très tôt dans le discours des enfants, par l'acquisition progressive des fonctions auxquelles elles sont associées : ainsi, *aa1*, attestée dans la majorité des conversations à $\bar{a}$=1 et $\bar{m}$=1, serait dans un premier temps utilisée comme particule d'invitation, pour acquérir par la suite les fonctions de confrontation, puis de défi. Il est intéressant de constater que, dans notre corpus, la fréquence d'utilisation de *aa1* par rapport aux autres SFP devient relativement stable à partir de $\bar{m}$=2, ce qui peut sembler surprenant si les enfants l'associent

à de plus nombreuses fonctions. N'ayant pas accès aux interprétations fonctionnelles des SFP utilisées dans le corpus, nous n'avons pas pu vérifier l'hypothèse selon laquelle les fonctions sont acquises de manière incrémentale, et si certaines fonctions sont corrélées à l'âge plutôt qu'au niveau de développement langagier. Notons toutefois que les observations de Tang, qui avaient été effectuées sur des enfants de 3 à 5 ans, n'incluaient pas de mesure de MLU, et ne pouvaient donc discriminer entre un effet de l'âge et un effet du développement pour l'explication de ces phénomènes.

9.3.2 Hypothèse 2 : effet du genre sur la production de SFP

Tse *et al.* (2002) ont observé un développement syntaxique plus précoce chez les filles que chez les garçons entre 3 et 5 ans, qu'ils ont associé à une MLU plus élevée. Cette observation a été confirmée par Li *et al.* (2013) dans leur étude sur le développement des formes interrogatives chez les enfants. Nous avons également pu constater dans notre corpus des valeurs de MLU plus élevées chez les filles à âge égal, ce qui résulte *de facto* en une production de SFP plus importante chez les filles que chez les garçons, et confirme donc notre hypothèse de recherche. Toutefois, il nous semble important de rappeler que cette différence disparaît lorsque les enfants sont groupés par MLU et non par âge. Cela semble confirmer que la MLU est un bon descripteur du niveau de développement langagier. Cela suggère également que les différences observées entre les garçons et les filles sont liées principalement au rythme de développement, plus lent chez les garçons, et ne peuvent donc pas être intrinsèquement associées au genre lui-même. Cette constatation est toutefois en contradiction avec les effets combinés de l'âge et de la MLU sur la production de SFP, dont nous avons observé qu'elle était moins importante pour les enfants plus âgés à MLU égale : les filles, étant plus jeunes à MLU égale, devraient, en accord avec cette observation, produire également plus de SFP, ce que les analyses statistiques focalisées sur le genre n'ont pas permis de mettre en avant. Il conviendra d'approfondir les recherches dans cette direction pour comprendre cette discontinuité.

Nous avons par ailleurs pu constater un effet du genre sur l'utilisation des particules *lo1*, *ne1* et *laa1* à $\bar{m}$=4, selon lequel les filles ont tendance à utiliser ces particules plus fréquemment. Il est intéressant de noter que *lo1* (qui souligne l'aspect évident d'un énoncé) et *laa1* (qui indique l'existence d'alternatives non mentionnées) sont décrites par Winterstein *et al.* (en préparation) comme étant significativement associées avec le genre féminin dans le discours adulte. Cette tendance n'est toutefois pas observée pour *ne1* (pour laquelle le biais constaté par les auteurs est masculin), et nous n'avons constaté dans nos données aucune différence significative pour d'autres particules courantes comme *aa3*, *laa3* ou *aa4*, qui

présentent également un biais (masculin ou féminin) dans le discours adulte. Winterstein *et al.* suggèrent que ces différences seraient liées à l'autorité épistémique à laquelle peuvent prétendre les locuteurs et locutrices, laquelle serait généralement moins accordée aux femmes qu'aux hommes. S'il est envisageable que cette asymétrie chez les adultes prenne racine dès l'enfance, par exemple par le biais des rapports que les enfants ont avec leurs parents ou avec d'autres adultes soumis à ce biais social, il serait nécessaire d'effectuer une étude longitudinale plus approfondie sur l'évolution de l'utilisation de ces particules de l'enfance à l'âge adulte pour pouvoir confirmer que les effets observés vers 5 ans sont bien les précurseurs de ceux observés chez les adultes. Il serait également intéressant d'analyser les biais liés au genre dans le CDS; il serait cependant dans ce cas important de différencier les environnements (p. ex. milieu familial ou école), en rapport avec les différences d'utilisation des SFP liées à la proximité entre les interlocuteurs.

9.3.3 Hypothèse 3 : influence des caractéristiques de l'interlocuteur sur la production de SFP

Le genre joue également un rôle comme caractéristique significative de l'interlocuteur : comme nous l'avons vu, les enfants utilisent plus de SFP lorsqu'ils s'adressent à des femmes que lorsqu'ils s'adressent à des hommes. Notons que cette tendance ne se retrouve pas dans l'ADS, où le genre de l'interlocuteur ne semble pas jouer de rôle significatif pour la production de SFP. On note en revanche dans l'ADS une tendance des femmes à utiliser plus souvent des SFP que les hommes, ce qui correspond aux observations de Chan (2002) ou Winterstein *et al.* (en préparation). On remarque également cet effet dans le CDS, quoique de manière moins marquée – rappelons toutefois que nos informations à ce sujet sont parcellaires et ne permettent pas une analyse statistique approfondie. La combinaison des faits que les femmes adultes semblent utiliser plus de SFP que les hommes et que les enfants ont tendance à produire plus de SFP lorsque leur interlocuteur vient d'en produire une est compatible avec cette propension des enfants à utiliser des SFP plus fréquemment lorsqu'ils s'adressent à des femmes. L'effet d'interaction entre le genre de l'enfant et celui de l'interlocuteur, selon lequel les filles utiliseraient encore plus de SFP à l'égard des femmes, et les garçons encore moins à l'égard des hommes, bien que plausible au regard des considérations précédentes, devra être examiné de manière plus approfondie pour pouvoir s'assurer de son bien-fondé.

En plus du genre, le degré de proximité de l'interlocuteur a lui aussi été analysé comme facteur significatif pour la production des SFP par les enfants : elle est plus élevée lorsque l'adulte est étranger au foyer, dans le CS et dans le CDS. S'il est logique de penser là encore à un effet d'émulation basé sur le fait que les enfants s'inspirent du schéma qui leur est présenté, il serait en tous les cas intéressant d'analyser cet effet

dans l'ADS pour voir si l'utilisation des SFP entre adultes constitue dans une certaine mesure un outil d'expression de distance sociale. Il est intéressant toutefois de noter que ce comportement est partiellement contradictoire avec le fait que les personnes étrangères au foyer posent globalement moins de questions aux enfants que les proches de ceux-ci (§6.3.4.2), et que les questions incitent plus les enfants à utiliser des SFP. Cela indique potentiellement que l'utilisation des SFP comme marqueurs de distance sociale prendrait alors le pas sur la propension des enfants à utiliser moins de SFP en réponse à un énoncé déclaratif.

Nous avons également pu constater que le fait que l'interlocuteur soit masculin avait tendance à inhiber l'utilisation des SFP *laa1* et *lo1*. Nous n'avons effectué cette analyse que sur les SFP les plus utilisées, et les effets calculés, bien qu'ils soient significatifs, étaient d'une ampleur très limitée. Ces analyses requièrent donc d'être approfondies pour confirmer si possible les résultats, à moins qu'il ne se soit agi d'un effet particulier dû à la constitution du corpus. Une telle analyse serait par ailleurs également pertinente pour l'ADS : elle permettrait d'observer si, en plus du fait que certaines SFP sont plutôt associées au discours des femmes et d'autres au discours des hommes, le genre de l'interlocuteur joue également un rôle sur les SFP utilisées pour s'adresser à eux (et, bien sûr, si des interactions existent entre les genres des deux interlocuteurs).

9.3.4 Hypothèse 4 : impact du CDS ponctuel sur la production de SFP

Comme nous l'avions évoqué au début de ce mémoire (§1.6.4.2), les interactions à l'œuvre entre le CDS et le CS dans le cadre de l'analyse de conversations ponctuelles sont plus susceptibles d'être liées à l'environnement discursif local qu'à des caractéristiques générales du CDS produit au cours d'une séance, trop éphémère et anecdotique (dans le processus d'acquisition) pour avoir un effet durable mesurable.

De manière générale, on constate d'ailleurs que les fréquences d'utilisation de certaines SFP chez les enfants sont plus proches de celles observées dans l'ADS que dans le CDS, comme le montrent les différents diagrammes présentés au chapitre 8. Ainsi, parmi les huit SFP observées, les fréquences relatives d'utilisation de *aa3*, *aa1* et *aa4* à $\bar{m}$=4 sont extrêmement proches des valeurs de l'ADS, alors qu'elles ne sont que partiellement comparables (*aa3*) à totalement différentes (*aa4*) dans le CDS. D'autres SFP (*gaa3*, *lo1*) ont également une fréquence d'utilisation plus proche de l'ADS que du CDS, avec toutefois une différence significative. Dans la majorité des cas, on constate cependant que l'évolution de l'utilisation en fonction des tranches de MLU s'uniformise vers le schéma observé dans l'ADS, avec une fréquence

moyenne similaire et une moins grande variabilité, ce qui suggère que les enfants acquièrent, dès m̄=4, un comportement langagier comparable à celui des adultes (au regard de ces SFP).

Nous avons pu constater que certaines caractéristiques des énoncés produits par les interlocuteurs des enfants au tour de parole précédent avaient un impact tangible sur la tendance de ceux-ci à utiliser des SFP dans leurs propres réactions. En particulier, l'utilisation d'une SFP par un adulte semble constituer une incitation pour l'enfant à produire une SFP – cette tendance est observée tant dans les tables de contingence (§6.3.4.3) que dans le coefficient associé à ce facteur dans nos différents modèles (p. ex. §7.2.1). Si cette observation semble très cohérente avec le fait que l'enfant effectue un transfert du discours entendu vers ses propres énoncés, elle ne doit toutefois pas nécessairement être interprétée comme de l'imitation : la tendance des enfants à utiliser la même SFP que la dernière produite par leur interlocuteur reste minoritaire (18.5% des segments contenant une SFP, 6.8% au total) et son évaluation doit prendre en compte le fait que l'inventaire des SFP disponibles est tout d'abord limité et correspond en partie aux SFP les plus utilisées par les adultes, ce qui explique que les SFP produites par les uns et les autres se recoupent parfois Notons que l'utilisation de la même SFP que l'interlocuteur est nettement moins fréquente dans l'ADS (7.2% des segments contenant une SFP et 3.7% au total), où la possibilité d'une influence du discours entendu semble moins envisageable. Lee *et al.* (1996) confirment pour leur part que le très faible taux de répétitions à l'identique permet d'exclure l'hypothèse d'un simple phénomène d'imitation. On peut ainsi émettre l'hypothèse que les enfants développent très tôt une notion typologique de ce que sont les SFP, et que ce qu'ils reproduisent correspond au concept linguistique; cette hypothèse est (au niveau empirique au moins) compatible avec l'analyse de Lee *et al.*, qui affirment que les enfants, avant l'âge de deux ans, réanalysent les SFP comme une catégorie (autre que celle de particule interrogative, en accord avec leurs observations).

Nous avons également noté les diverses influences du type de l'énoncé produit par l'interlocuteur sur la réponse de l'enfant : ceux-ci ont tendance à alterner les formes d'énoncés (répondre à une phrase interrogative par une phrase déclarative et vice versa), à utiliser plus de SFP dans leurs propres phrases interrogatives (tendance qui se retrouve chez les adultes), et à utiliser plus de SFP en réponse aux phrases interrogatives. Il est bien entendu que les conditions d'application de ces différentes tendances, constatées statistiquement sur l'ensemble du corpus, sont complexes, et qu'elles ne peuvent s'expliquer sans prendre en compte tous les autres paramètres que nous avons considérés jusqu'à présent. Il est toutefois envisageable que l'effet inhibiteur du contexte déclaratif observé chez les adultes (en termes de

production de SFP) fasse l'objet d'un transfert par les enfants; l'interaction entre la forme déclarative de l'énoncé précédent et l'utilisation d'une SFP dans cet énoncé pourrait alors être interprété comme une mise en saillance de cette SFP, ce qui apparaîtrait dans nos modèles comme l'effet facilitateur rapporté (§7.2.2.1).

9.3.5 Hypothèse 5 : interaction entre structures interrogatives et production de SFP

L'existence d'interactions fortes entre la production des structures interrogatives et la production des SFP a été confirmée à plusieurs reprises au cours de notre étude. Il convient toutefois de s'intéresser également à l'évolution de ces interactions au cours du processus d'acquisition.

L'observation de la structure des énoncés interrogatifs dans notre corpus montre que l'augmentation de la fréquence d'utilisation et la diversification des structures interrogatives (mots QU, HAI_M_HAI, A_M_A) est dans un premier temps corrélée avec une sous-utilisation des SFP dans le contexte interrogatif, ce qui se traduit dans notre modèle statistique par un effet inhibiteur de la production des SFP en lien avec l'utilisation de ces structures (§7.3). Cet effet peut s'expliquer par le fait que, comme nous l'avons vu à la section 6.3.3.2, ces structures sont elles aussi en phase d'émergence dans le discours : à un stade où les enfants produisent des énoncés de 2 à 3 syllabes en moyenne ($2 \leq \bar{m} \leq 3$), ils y intègrent des éléments d'une ou plusieurs syllabes leur permettant d'exprimer la notion d'interrogation. Comme l'ont observé Wong et Ingram (2003), la formulation des questions par les enfants suit un schéma de complexité syntaxique croissante, dans lequel la combinaison d'une structure syntaxique et d'une SFP correspond à une construction avancée.

Il est également pertinent de prendre en compte que ces structures ne disposent que d'un registre pragmatique très limité : les structures HAI_M_HAI et A_M_A n'introduisent pas de biais dans la question formulée, celui-ci pouvant être fourni le cas échéant par une SFP non interrogative; *aa3* est la particule la plus utilisée par les enfants avec les structures interrogatives (§8.1), et ne porte pas non plus de biais. La particule *aa4*, qui est généralement la première SFP interrogative utilisée productivement par les enfants (cf. §8.8), permet de formuler une question oui–non à partir d'un énoncé déclaratif et est réputée apporter une notion de surprise, bien que Kwok (1984, p. 86–87) indique que cette dimension mirative n'est pas essentielle dans sa signification; Hara (2023) décrit les questions formulées au moyen de *aa4* comme des « questions pures dénuées de contenu assertif ». Ces propriétés paraissent cohérentes avec l'utilisation que peuvent en faire les jeunes enfants : ils posent leurs premières questions sans y ajouter de nuance

épistémique, ou au maximum une nuance mirative, avec comme objectif l'obtention d'une valeur de vérité (oui ou non). Il serait donc concevable que l'association des structures syntaxiques avec des SFP portant une charge épistémique résulte d'un besoin pragmatique, lequel ne s'exprimerait toutefois pas dans les premières phases de l'acquisition, et ne justifierait donc pas l'utilisation de SFP au-delà du *aa3* d'« adoucissement ». Cette hypothèse serait par ailleurs compatible avec le fait que les mêmes structures syntaxiques ne sont pas associées à cet effet inhibiteur chez les adultes, comme le suggèrent nos modèles de production des SFP dans les segments interrogatifs pour le CDS et l'ADS : de même qu'ils diversifient leur éventail de particules interrogatives pour y intégrer des nuances pragmatiques (p. ex. *me1* est inquisitrice, *ge2* exprime un doute, *ne1* une comparaison), ils associent également des notions plus variées aux constructions interrogatives neutres pour enrichir les messages transmis.

9.4 Limitations

Ce projet a constitué une formidable opportunité non seulement d'approfondir nos connaissances sur de nombreux aspects de la recherche en linguistique et de l'analyse de corpus, mais aussi de découvrir de nouveaux domaines comme les langues chinoises, l'utilisation de processus langagiers radicalement différents de ceux généralement rencontrés dans les langues indo-européennes, ainsi que l'ajustement de modèles statistiques complexes. Toutefois, certaines limitations doivent être prises en compte dans l'évaluation des résultats auxquels nous sommes parvenus.

En premier lieu, l'étude d'une langue inconnue (le cantonais), en relation avec le manque de normalisation dont il fait l'objet (orthographe, segmentation, romanisation), a constitué un frein important à l'analyse critique que nous avons pu appliquer aux données que nous manipulions. En particulier, il nous a été difficile de porter un jugement éclairé sur certains choix effectués par les auteurs des deux corpus utilisés, en particulier lorsque ces choix semblaient contradictoires ou introduisaient des ambiguïtés (§4.2 et §4.3). L'asymétrie dont font preuve les deux corpus a nécessité de prendre des décisions parfois insuffisamment étayées, dont les répercussions tardives ont pu avoir des impacts négatifs sur la qualité des analyses effectuées.

Par ailleurs, l'étude d'un phénomène aussi complexe et multifactoriel que l'acquisition des SFP, particulièrement en cantonais où elles forment une classe présentant de nombreuses particularités (en particulier leur nombre et la pluralité des nuances qui y sont associées), requiert de prendre en compte un nombre conséquent de facteurs – l'identification de certains de ces facteurs, ainsi que l'étude de leurs

interactions, ont constitué nos principaux objectifs. Cependant, dans le cadre d'un mémoire de maîtrise, il nous a été impossible de considérer l'ensemble des paramètres mis en jeu; en particulier, l'aspect fonctionnel des SFP, très important dans le cadre de l'analyse d'un phénomène fortement ancré dans la dimension pragmatique du langage, n'a pu être intégré à notre étude. Ce travail est donc essentiellement axé sur la forme, et ne peut que formuler des hypothèses sur les rapports qui unissent l'acquisition des SFP et l'évolution des besoins pragmatiques des enfants.

Enfin, la mise à disposition d'autres facteurs pertinents concernant l'environnement de vie des enfants (par exemple niveau d'éducation des parents, statut socio-économique, genre des interlocuteurs), ainsi que la mise en place de conditions d'enregistrement plus écologiques, auraient constitué de précieux apports dans le déroulement de cette étude.

CONCLUSION

L'acquisition des particules de fin de phrase en cantonais représente un défi majeur pour les jeunes locuteurs, qui y sont exposés dès leur plus jeune âge. Du fait de leur multiplicité, de leur omniprésence et de la variété des fonctions sémantiques et pragmatiques qu'elles permettent d'exprimer, les SFP constituent un élément central mais complexe du cantonais parlé, dont la maîtrise est cruciale au développement des compétences langagières observées chez les locuteurs adultes.

Par le biais du traitement automatisé de deux corpus d'interactions entre enfants et adultes (le CANCORP et le HKU-70), nous avons procédé à une description systématique des environnements dans lesquels les SFP sont produites par les enfants et leurs interlocuteurs. Ces données nous ont permis d'étudier avec précision l'impact de facteurs individuels (âge, niveau de développement langagier, genre), linguistiques (présence de structures syntaxiques particulières, phono-prosodie) et extrinsèques (genre et rôle de l'interlocuteur, caractéristiques du discours adressé à l'enfant), ainsi que des interactions pouvant exister entre ces différents facteurs, au moyen d'analyses de corrélations et d'ajustements de modèles statistiques.

Nos analyses ont permis de confirmer que l'émergence des SFP dans le discours des enfants est avant tout liée au développement langagier de ceux-ci (tel qu'exprimé par la MLU), et que l'âge biologique ne joue pour sa part qu'un rôle secondaire. Nous avons toutefois observé une interaction inattendue entre l'âge et la MLU, selon laquelle les enfants au développement plus précoce avaient tendance à utiliser plus de SFP que des enfants plus vieux de niveau de développement comparable. Cette constatation, si elle est confirmée par des analyses complémentaires, devra être exploitée pour vérifier si l'effet constaté sur la production des SFP s'exprime également sur d'autres aspects du discours des enfants, éventuellement dans d'autres langues. Elle justifierait de plus de toujours considérer la MLU et l'âge ensemble comme descripteurs du niveau de développement langagier.

Nous avons également pu mettre en avant l'importance du contexte tant discursif que pragmatique dans le processus de production des SFP; on constate ainsi un effet d'émulation par lequel les enfants ont plus tendance à utiliser des SFP lorsque leurs interlocuteurs en utilisent. Cette interaction se retrouve dans la propension qu'ont les enfants à utiliser plus de SFP à l'égard des femmes, mais aussi à l'égard de personnes étrangères à leur foyer. Il convient alors de s'interroger sur l'origine de ces différences, et en particulier

sur le potentiel rôle des SFP comme marqueurs de distance sociale, qui pourrait alors s'observer également chez les adultes.

Enfin, nous avons noté une interaction très forte entre la production de phrases interrogatives et la production de SFP, ce que le rôle joué par les SFP dans la formation de tournures interrogatives laissait présager. La production de SFP est par ailleurs largement plus fréquente dans les phrases interrogatives que dans les phrases déclaratives. Toutefois, l'émergence tardive des SFP interrogatives autres que *aa4* suggère que les nuances pragmatiques portées par celles-ci ne constituent pas un besoin majeur dans le développement des interrogations chez les enfants.

Ces résultats et observations confirment que l'émergence des SFP dans le discours des enfants doit être interprétée comme un processus multidimensionnel, et que les facteurs qui l'influencent présentent des interactions complexes; ils justifient d'être analysés de manière contextuelle, et montrent que le développement du langage chez les enfants, ainsi que l'utilisation qu'il en fait, sont fortement influencés par des éléments situés tant à l'intérieur qu'en dehors de la sphère discursive. L'une des forces de notre projet a d'ailleurs été de pouvoir intégrer ces facteurs extrinsèques à nos analyses, ajoutant ainsi une dimension pragmatique cruciale à l'étude des SFP. Ainsi, même s'il semble logique d'envisager qu'un rôle important doit être assigné au discours ambiant quotidien (cadre familial, école) pour justifier de l'émergence d'éléments langagiers, et que s'il est vrai que ce rôle n'a pas pu être évalué dans le cadre de notre projet faute de données réellement écologiques illustrant la qualité globale du CDS, il est important de constater que les conversations hors des environnements familiers apportent elles aussi des renseignements cruciaux sur les rapports que les enfants entretiennent avec le monde qui les entoure, et sur leur utilisation du langage comme outil servant à exprimer ces rapports.

Ce projet s'inscrit dans une initiative de description analytique des processus par lesquels les enfants mettent à profit leur environnement social et langagier pour s'approprier les éléments constitutifs de leur langue maternelle et les compétences langagières caractéristiques des locuteurs adultes. En se focalisant sur l'étude de particules fortement ancrées dans la dimension pragmatique, il permet aussi de mettre en avant le lien étroit qui unit la langue à l'environnement, et l'évolution dont ce lien fait l'objet au cours du développement de l'enfant. L'étude du cantonais, une langue dont les schémas se différencient de ceux des langues indo-européennes sur de nombreux aspects fondamentaux, permet par ailleurs d'élargir notre vision globale du domaine de la linguistique en exposant nos théories à de nouveaux paradigmes, et

d’intégrer de nouvelles observations à nos réflexions sur les processus cognitifs à l’œuvre dans la production et l’interprétation du langage.

ANNEXE A

Liste des SFP considérées dans notre étude

Cette table présente la liste des 47 SFP que nous reconnaissons dans les corpus. Chaque SFP est décrite au moyen de sa ou ses notations en jyutping, sa ou ses notations en caractères chinois, son statut de particule exclusivement interrogative, et sa position dans une séquence de SFP.

La forme canonique d'une SFP admettant plusieurs notations jyutping correspond à la concaténation des premières options de chacun de ses composants. Ainsi, la forme canonique de « zaa1|ze1|zi1 maa3 » est *zaa1maa3*.

jyutping	caractères chinois	SFP interrogative	position
aa1	吖	Non	4
aa1|aa2|aa3|aa4|aa5|aa6 maa3	吖|呀|啞|啊 嘛|嗎	Non	4
aa3	啊|呀|呀 3	Non	4
aa4	呀	Oui	4
aa5	呃|啞|呀	Oui	4
aa6	噯	Non	4
aak3	呃	Non	3
bo3	嘩	Non	4
gaa1|gaa2 maa3	加|㗎|嗰 嘛	Non	2
gaa2	架|嘅	Non	2
gaa3	㗎|㗎3|嚟	Non	2
gaa4	㗎|㗎4	Oui	2
gaak3	嗝	Non	2
ge2	咖|嘅 2|嗱	Non	2
ge3	嘅|嘅 3	Non	2
gwaa3	掛|啩	Non	4
haa2	吓|吓 2	Non	4
ho2	呵	Non	4
laa1	啦|喇|啦 1	Non	3
laa3	喇|啦 3|嚇	Non	3
laa3 haa2	嚇|喇 吓	Non	3

laa4	嗱|啦 4	Oui	4
laak3	嘞	Non	3
le4	嘞	Non	3
le5	哩	Non	3
lei4	嚟	Non	2
lei4 gaa3	嚟 喫	Non	2
lo1	囉|囉 1	Non	2
lo3	囉|囉 3|嘍	Non	3
lo4	囉	Non	3
lok3	咯	Non	3
maa3	嗎|嘛	Non	4
me1	咩	Oui	4
ne1|le1	呢	Non	4
sin1	先	Non	1
tim1	添|嗮|啙	Non	1
waa2|aa2[37]	話	Oui[38]	4
wo3	喎	Non	4
wo4	唲|喎|喎 4	Non	4
wo5	喎|喎 5	Non	4
zaa1|ze1|zi1 maa3	咋|啫|之 嘛	Non	3
zaa3	咋|咋 3|喳	Non	3
zaa3 haa2	咋 吓	Non	3
zaa4	喳|咋 4	Oui	4
ze1	啫|姐 1	Non	3
zek1	唧	Non	3

[37] La forme aa2 n'est répertoriée que dans Yau (1965) (qui par ailleurs ne mentionne pas waa2 sous forme autonome), mais on la trouve étiquetée comme SFP dans la version originale du CC 2012. Après analyse par une locutrice native, il a été décidé que certaines occurrences (me1aa2) correspondaient en fait à la particule waa2. La valeur des autres occurrences n'a pas pu être identifiée de manière univoque.

[38] Si waa2 n'est pas à proprement parler une particule interrogative au sens où elle ne peut pas être utilisée pour transformer une affirmation en interrogation, elle est en revanche systématiquement utilisée dans des énoncés interrogatifs (Sybesma et Li, 2007).

ANNEXE B

Liste des conversations du corpus

NB : les noms des fichiers sont censés être basés sur les âges des enfants aux dates des enregistrements. Toutefois, il y a parfois des incohérences entre ces noms de fichiers et les métadonnées listées dans les en-têtes des transcriptions. Dans la grande majorité des cas, l'âge indiqué par le nom du fichier est correct au regard des autres métadonnées (date de naissance de l'enfant et date de l'enregistrement), nous nous sommes donc basés sur celui-ci pour nos analyses. Seul le fichier `{HKU}/3/031120.cha`, dont le nom indique un âge théorique de 3 ans, 11 mois et 20 jours, a été enregistré, selon les métadonnées, à l'âge de 4 ans, 0 mois et 1 jour (`@Birth of CHI: 14-JUN-1994; @Date: 14-JUN-1998`). Cette donnée a été corrigée dans les analyses.

Dans la table suivante, la colonne « Enfant » indique l'enfant concerné pour les conversations du CC, dans lequel huit enfants sont suivis sur une période d'un an. Les colonnes « Taux SFP » et « Div. SFP » correspondent respectivement au taux d'utilisation des SFP (nombre d'énoncés contenant une SFP) et à la diversité de SFP (nombre de SFP différentes utilisées par l'enfant au cours d'une conversation), notions décrites à la section §6.2. La colonne « Participants » indique la liste des personnes présentes lors de la conversation, selon les métadonnées de chaque conversation. L'enfant cible est toujours indiqué en tant que « CHI » (*child*).

	Identifiant	Enfant	Genre	Âge	MLU	Taux SFP	Div. SFP	Participants
1	010522_cc_ckt	CKT	masc.	01;05.22	1.5753	0.1507	4	CHI INV MOT FAT
2	010700_cc_mhz	MHZ	masc.	01;07.00	1.4941	0.0118	1	CHI INV MOT
3	010703_cc_ckt	CKT	masc.	01;07.03	2.0938	0.3281	4	CHI INV MOT FAT
4	010710_cc_ckt	CKT	masc.	01;07.10	1.8571	0.2418	2	CHI INV MOT FAT
5	010800_cc_ckt	CKT	masc.	01;08.00	1.45	0.1024	5	CHI INV FAT
6	010800_cc_mhz	MHZ	masc.	01;08.00	1.665	0.0485	4	CHI INV MOT
7	010807_cc_ckt	CKT	masc.	01;08.07	1.5166	0.1637	5	CHI INV FAT
8	010814_cc_mhz	MHZ	masc.	01;08.14	1.815	0.0751	1	CHI INV
9	010821_cc_ckt	CKT	masc.	01;08.21	1.6718	0.1574	5	CHI INV MOT FAT
10	010828_cc_mhz	MHZ	masc.	01;08.28	1.7692	0.1453	2	CHI INV MOT
11	010904_cc_mhz	MHZ	masc.	01;09.04	1.6036	0.0312	2	CHI INV MOT FAT
12	010907_cc_ckt	CKT	masc.	01;09.07	1.8392	0.2186	5	CHI INV MOT
13	010914_cc_ckt	CKT	masc.	01;09.14	1.6381	0.2095	3	CHI INV

14	010918_cc_mhz	MHZ	masc.	01;09.18	1.8995	0.1848	10	CHI INV MOT
15	010925_cc_mhz	MHZ	masc.	01;09.25	2.0305	0.0479	3	CHI INV MOT FAT
16	010929_cc_ckt	CKT	masc.	01;09.29	2.2094	0.4842	7	CHI INV MOT FAT
17	011008_cc_ccc	CCC	masc.	01;10.08	2.2584	0.4195	5	CHI INV MOT FRI
18	011010_cc_mhz	MHZ	masc.	01;10.10	1.9241	0.0864	6	CHI INV MOT FAT
19	011023_cc_mhz	MHZ	masc.	01;10.23	2.1701	0.0506	7	CHI INV MOT GRM GRF
20	011030_cc_ckt	CKT	masc.	01;10.30	1.6827	0.2406	6	CHI INV FAT
21	011100_cc_ccc	CCC	masc.	01;11.00	2.777	0.5827	2	CHI INV MOT
22	011101_cc_cgk	CGK	fém.	01;11.01	2.366	0.2268	8	CHI MOT INV
23	011106_cc_mhz	MHZ	masc.	01;11.06	2.2033	0.0809	9	CHI INV MOT
24	011108_cc_cgk	CGK	fém.	01;11.08	2.3306	0.2603	7	CHI MOT INV
25	011113_cc_ckt	CKT	masc.	01;11.13	2.1893	0.5089	5	INV CHI FAT
26	011121_cc_ccc	CCC	masc.	01;11.21	2.3278	0.3167	2	CHI INV MOT
27	011122_cc_cgk	CGK	fém.	01;11.22	2.8214	0.2679	4	CHI MOT INV
28	011127_cc_ckt	CKT	masc.	01;11.27	1.6459	0.2149	3	INV CHI FAT
29	011129_cc_cgk	CGK	fém.	01;11.29	2.5101	0.2412	11	CHI MOT INV
30	020003_cc_mhz	MHZ	masc.	02;00.03	2.5009	0.0583	12	CHI INV MOT
31	020008_cc_cgk	CGK	fém.	02;00.08	2.6517	0.2547	9	CHI MOT INV
32	020009_cc_ckt	CKT	masc.	02;00.09	2.1346	0.4369	8	CHI INV GRA AUN
33	020016_cc_ckt	CKT	masc.	02;00.16	2.0816	0.4539	4	INV CHI FAT
34	020016_cc_mhz	MHZ	masc.	02;00.16	1.9697	0.0375	5	CHI INV MOT FAT
35	020101_cc_mhz	MHZ	masc.	02;01.01	2.1289	0.0692	13	CHI INV SIV MOT
36	020108_cc_ckt	CKT	masc.	02;01.08	2.002	0.2751	13	CHI INV FAT
37	020110_cc_ccc	CCC	masc.	02;01.10	2.388	0.3003	10	CHI INV MOT GRA
38	020115_cc_mhz	MHZ	masc.	02;01.15	1.8508	0.0712	9	CHI INV MOT
39	020117_cc_ccc	CCC	masc.	02;01.17	2.1892	0.2282	6	CHI INV MOT
40	020129_cc_mhz	MHZ	masc.	02;01.29	2.0134	0.1562	9	CHI INV FAT
41	020205_cc_ckt	CKT	masc.	02;02.05	2.3782	0.2952	8	CHI INV GRA MAN WAI
42	020206_cc_ccc	CCC	masc.	02;02.06	2.0251	0.233	16	CHI INV MOT
43	020207_cc_cgk	CGK	fém.	02;02.07	2.8957	0.3791	10	CHI MOT INV
44	020210_cc_ltf	LTF	fém.	02;02.10	2.4024	0.2744	10	CHI INV MOT STU
45	020212_cc_mhz	MHZ	masc.	02;02.12	1.785	0.1215	7	CHI INV SIV MOT FAT
46	020213_cc_ccc	CCC	masc.	02;02.13	2.2332	0.2767	11	CHI INV MOT UNC GRM
47	020215_cc_ckt	CKT	masc.	02;02.15	2.1711	0.3088	12	CHI INV GRA SIV
48	020221_cc_cgk	CGK	fém.	02;02.21	2.8293	0.3213	13	INV CHI MOT
49	020226_cc_mhz	MHZ	masc.	02;02.26	2.261	0.2949	18	CHI INV FAT
50	020228_cc_cgk	CGK	fém.	02;02.28	2.6049	0.2098	12	CHI MOT INV

51	020302_cc_ltf	LTF	fém.	02;03.02	2.1959	0.1807	16	CHI INV MOT STU
52	020303_cc_ckt	CKT	masc.	02;03.03	2.4147	0.3253	10	CHI INV FAT
53	020304_cc_cgk	CGK	fém.	02;03.04	2.9284	0.3917	14	CHI MOT INV
54	020307_cc_ccc	CCC	masc.	02;03.07	2.3674	0.3707	15	CHI INV MOT GRM
55	020309_cc_mhz	MHZ	masc.	02;03.09	1.9042	0.2236	11	CHI INV FAT
56	020311_cc_cgk	CGK	fém.	02;03.11	3.7628	0.4416	11	CHI INV MOT
57	020317_cc_ckt	CKT	masc.	02;03.17	1.9346	0.1962	8	CHI INV FAT
58	020318_cc_cgk	CGK	fém.	02;03.18	3.6873	0.5444	11	CHI INV MOT
59	020323_cc_ccc	CCC	masc.	02;03.23	2.71	0.2702	15	CHI INV MOT UNC
60	020323_cc_wbh	WBH	fém.	02;03.23	2.4722	0.3519	5	CHI INV MOT
61	020325_cc_cgk	CGK	fém.	02;03.25	3.0825	0.3093	12	CHI INV VSI
62	020328_cc_mhz	MHZ	masc.	02;03.28	2.4318	0.273	16	CHI INV MOT FAT
63	020330_cc_ltf	LTF	fém.	02;03.30	3.1161	0.5238	10	CHI INV MOT FAT
64	020330_cc_wbh	WBH	fém.	02;03.30	2.5075	0.3843	13	INV CHI MOT FAT
65	020400_cc_ckt	CKT	masc.	02;04.00	2.1847	0.2753	5	CHI INV FAT GRA
66	020402_cc_wbh	WBH	fém.	02;04.02	2.2679	0.378	9	CHI INV MOT FAT
67	020405_cc_wbh	WBH	fém.	02;04.05	2.5638	0.5101	7	CHI INV
68	020406_cc_wbh	WBH	fém.	02;04.06	2.2655	0.3414	8	CHI MOT INV UNC GRM
69	020407_cc_mhz	MHZ	masc.	02;04.07	3.1635	0.4717	9	CHI INV MOT FAT
70	020408_cc_cgk	CGK	fém.	02;04.08	3.7745	0.5676	14	CHI MOT INV VSI
71	020408_cc_hhc	HHC	masc.	02;04.08	1.9058	0.2775	4	CHI INV MOT FAT SIS
72	020410_cc_ccc	CCC	masc.	02;04.10	2.6016	0.3132	12	CHI INV MOT UNC GRM
73	020413_cc_wbh	WBH	fém.	02;04.13	2.7737	0.3474	12	CHI MOT FAT
74	020414_cc_ckt	CKT	masc.	02;04.14	2.1535	0.241	17	CHI INV FAT
75	020414_cc_wbh	WBH	fém.	02;04.14	2.4416	0.3117	6	CHI INV MOT
76	020415_cc_wbh	WBH	fém.	02;04.15	2.7021	0.1915	4	CHI INV MOT
77	020421_cc_mhz	MHZ	masc.	02;04.21	2.6678	0.3023	15	CHI INV SIV FAT
78	020427_cc_ltf	LTF	fém.	02;04.27	3.2732	0.4466	11	CHI INV SIS MOT
79	020430_cc_cgk	CGK	fém.	02;04.30	3.3562	0.3525	16	CHI MOT INV
80	020500_cc_ckt	CKT	masc.	02;05.00	2.3809	0.2739	13	INV CHI IN2
81	020500_hku		masc.	02;05.00	2.9762	0.5714	9	CHI INV FAT MOT
82	020503_cc_cgk	CGK	fém.	02;05.03	3.5584	0.487	13	CHI MOT INV
83	020503_cc_hhc	HHC	masc.	02;05.03	2.2319	0.3121	9	CHI INV SIS SER MOT
84	020504_cc_mhz	MHZ	masc.	02;05.04	2.8973	0.3243	14	CHI INV FAT MOT
85	020506_cc_wbh	WBH	fém.	02;05.06	2.4961	0.3721	8	CHI MOT GRM INV
86	020507_cc_ccc	CCC	masc.	02;05.07	3.1154	0.3706	13	CHI INV MOT UNC GRM
87	020511_hku		fém.	02;05.11	1.9091	0.0379	2	INV CHI

88	020512_hku		masc.	02;05.12	2.1275	0.1961	9	INV CHI
89	020513_cc_hhc	HHC	masc.	02;05.13	2.2013	0.2144	12	CHI INV SIS SER
90	020514_cc_ckt	CKT	masc.	02;05.14	2.1782	0.317	17	INV CHI IN2
91	020514_hku		fém.	02;05.14	3.291	0.5075	13	CHI INV FAT
92	020515_hku		masc.	02;05.15	2.2526	0.3474	9	INV CHI
93	020518_cc_ltf	LTF	fém.	02;05.18	3.1299	0.4654	14	CHI INV MOT SER
94	020519_cc_hhc	HHC	masc.	02;05.19	2.1656	0.211	12	CHI INV SER SIS
95	020519_cc_mhz	MHZ	masc.	02;05.19	2.8571	0.3593	20	CHI INV FAT
96	020520_hku		fém.	02;05.20	1.6329	0.0759	5	INV CHI
97	020523_cc_ccc	CCC	masc.	02;05.23	2.9396	0.5396	17	INV CHI
98	020523_hku		fém.	02;05.23	3.1628	0.3895	16	CHI INV MOT GRM
99	020527_hku		masc.	02;05.27	2.0484	0.2984	8	INV CHI IN2
100	020530_hku		masc.	02;05.30	3.2483	0.6913	8	INV CHI IN2
101	020600_hku		fém.	02;06.00	3.4328	0.5851	19	INV CHI
102	020601_cc_ltf	LTF	fém.	02;06.01	3.2635	0.5655	22	CHI INV MOT ANC STU STM ACH
103	020604_cc_mhz	MHZ	masc.	02;06.04	2.8877	0.3209	15	CHI INV MOT
104	020608_cc_ccc	CCC	masc.	02;06.08	2.6768	0.3967	15	CHI INV MOT GRM
105	020609_cc_wbh	WBH	fém.	02;06.09	2.9185	0.4222	9	CHI MOT INV GRM
106	020610_cc_hhc	HHC	masc.	02;06.10	2.1706	0.2413	11	CHI INV SER SIS
107	020618_cc_ckt	CKT	masc.	02;06.18	2.4496	0.401	25	CHI INV FAT
108	020618_cc_mhz	MHZ	masc.	02;06.18	2.3906	0.2701	16	CHI INV SIV TIV FIV PBY
109	020619_cc_wbh	WBH	fém.	02;06.19	2.7571	0.4	5	CHI MOT INV
110	020624_cc_ccc	CCC	masc.	02;06.24	3.0015	0.4877	17	CHI INV MOT GRM
111	020624_cc_hhc	HHC	masc.	02;06.24	2.2179	0.3118	12	CHI INV MOT SER SIS SEC
112	020702_cc_ckt	CKT	masc.	02;07.02	2.7921	0.3966	24	CHI INV FAT
113	020703_cc_wbh	WBH	fém.	02;07.03	3.8084	0.6766	13	CHI INV MOT GRM
114	020705_cc_ltf	LTF	fém.	02;07.05	2.6619	0.2528	17	CHI INV SIS SER
115	020706_cc_ccc	CCC	masc.	02;07.06	3.3104	0.6924	15	CHI INV MOT BOY
116	020711_cc_cgk	CGK	fém.	02;07.11	3.5281	0.4419	12	CHI INV MOT
117	020713_cc_ccc	CCC	masc.	02;07.13	3.4513	0.7167	17	CHI INV MOT
118	020714_cc_wbh	WBH	fém.	02;07.14	2.9146	0.4591	15	CHI MOT GRM INV
119	020720_cc_ltf	LTF	fém.	02;07.20	3.0588	0.4542	17	CHI INV MOT
120	020721_cc_hhc	HHC	masc.	02;07.21	2.2828	0.3954	14	CHI INV MOT SER SIS
121	020800_cc_ccc	CCC	masc.	02;08.00	3.3981	0.5204	17	CHI INV MOT
122	020802_cc_ltf	LTF	fém.	02;08.02	3.113	0.418	18	CHI INV MOT SER SIS GDM
123	020806_cc_mhz	MHZ	masc.	02;08.06	2.0804	0.1836	16	CHI INV MOT

124	020808_cc_cgk	CGK	fém.	02;08.08	3.4043	0.3681	19	CHI MOT FAT INV
125	020808_cc_hhc	HHC	masc.	02;08.08	2.1608	0.2754	13	CHI INV MOT SER SIS GRA
126	020810_cc_lly	LLY	fém.	02;08.10	2.8443	0.6886	16	INV CHI SIS FAT MOT
127	020817_cc_ccc	CCC	masc.	02;08.17	3.0213	0.5447	17	CHI INV MOT UNC GRM
128	020818_cc_cgk	CGK	fém.	02;08.18	3.113	0.3466	17	CHI MOT INV
129	020822_cc_lly	LLY	fém.	02;08.22	2.9258	0.6868	16	INV CHI SIS MOT
130	020824_cc_ltf	LTF	fém.	02;08.24	3.2326	0.4505	20	CHI MOT INV STU
131	020907_cc_ccc	CCC	masc.	02;09.07	2.9727	0.4612	19	CHI INV MOT UNC GRM CUS
132	020907_cc_ltf	LTF	fém.	02;09.07	3.8764	0.5211	21	CHI INV MOT SER
133	020909_cc_cgk	CGK	fém.	02;09.09	2.9156	0.2844	16	CHI INV MOT
134	020909_cc_lly	LLY	fém.	02;09.09	3.2605	0.605	12	INV CHI SIS MOT
135	020914_cc_lly	LLY	fém.	02;09.14	2.5941	0.4	14	INV CHI SIS MOT
136	020919_cc_wbh	WBH	fém.	02;09.19	3.2514	0.4405	15	CHI MOT GRM INV
137	020923_cc_ccc	CCC	masc.	02;09.23	2.9424	0.4257	20	CHI INV MOT UNC GRM
138	020926_cc_wbh	WBH	fém.	02;09.26	3.0579	0.3079	14	CHI MOT INV
139	020928_cc_lly	LLY	fém.	02;09.28	2.6821	0.5197	22	CHI INV MOT SIS
140	020930_cc_hhc	HHC	masc.	02;09.30	2.4946	0.5458	12	CHI INV SIS GRM
141	021002_cc_wbh	WBH	fém.	02;10.02	3.5185	0.3407	12	CHI MOT FAT INV
142	021013_cc_ccc	CCC	masc.	02;10.13	3.3052	0.5396	23	CHI INV MOT UNC GRM
143	021013_cc_hhc	HHC	masc.	02;10.13	2.5954	0.4291	21	CHI INV SER SIS
144	021016_cc_wbh	WBH	fém.	02;10.16	3.3245	0.3511	15	CHI MOT FAT INV GRM
145	021018_cc_ltf	LTF	fém.	02;10.18	2.8897	0.3879	17	CHI INV MOT VSI SER STU LAD
146	021023_cc_wbh	WBH	fém.	02;10.23	3.089	0.4703	15	CHI GRM INV
147	021027_cc_ccc	CCC	masc.	02;10.27	3.3718	0.6066	27	INV CHI MOT ACQ
148	021101_cc_lly	LLY	fém.	02;11.01	2.8233	0.5948	25	INV CHI SIS MOT FAT SER
149	021106_cc_wbh	WBH	fém.	02;11.06	3.2025	0.4051	6	INV CHI
150	021106_hku		fém.	02;11.06	3.1092	0.5	12	CHI MOT INV VSI
151	021108_cc_hhc	HHC	masc.	02;11.08	2.4482	0.2768	13	CHI INV SER SIS MOT
152	021108_cc_lly	LLY	fém.	02;11.08	2.7841	0.4867	20	CHI INV MOT SIS ANN
153	021114_cc_wbh	WBH	fém.	02;11.14	3.6802	0.4569	12	CHI INV FAT GRM
154	021116_cc_ltf	LTF	fém.	02;11.16	2.953	0.3304	17	INV CHI
155	021116_hku		fém.	02;11.16	2.9321	0.5571	10	CHI INV MOT
156	021118_hku		masc.	02;11.18	2.8889	0.3111	8	INV CHI

157	021119_hku		masc.	02;11.19	2.806	0.2239	6	INV CHI
158	021128_cc_wbh	WBH	fém.	02;11.28	3.4877	0.4506	14	INV CHI
159	021128_hku		masc.	02;11.28	3.1971	0.3212	11	CHI INV MOT VSI GRM
160	021129_cc_lly	LLY	fém.	02;11.29	3.0579	0.4398	22	CHI INV
161	021129_hku		masc.	02;11.29	4.3239	0.7273	16	CHI INV MOT SER SIS GRA FAT
162	021130_hku		masc.	02;11.30	2.6812	0.3312	6	INV CHI
163	030008_cc_hhc	HHC	masc.	03;00.08	2.4039	0.336	17	CHI INV SER SIS
164	030010_cc_wbh	WBH	fém.	03;00.10	3.1847	0.4189	14	CHI MOT INV
165	030011_cc_lly	LLY	fém.	03;00.11	3.5234	0.5347	20	CHI INV MOT SER BRO
166	030020_cc_ltf	LTF	fém.	03;00.20	2.4428	0.251	19	CHI INV MOT GMO SIS
167	030022_cc_lly	LLY	fém.	03;00.22	3.8138	0.6607	22	CHI INV MOT SER
168	030029_hku		fém.	03;00.29	3.2486	0.5932	16	IN1 CHI IN2
169	030101_cc_wbh	WBH	fém.	03;01.01	2.7152	0.2818	10	CHI INV MOT GRM UNC
170	030105_hku		fém.	03;01.05	4.7123	0.7123	19	INV CHI
171	030106_hku		fém.	03;01.06	3.1959	0.2635	12	INV CHI
172	030113_cc_lly	LLY	fém.	03;01.13	3.6166	0.5798	23	CHI INV MOT BRO SIS
173	030116_cc_hhc	HHC	masc.	03;01.16	2.4499	0.3384	21	CHI INV SER SIS MOT
174	030121_cc_ltf	LTF	fém.	03;01.21	2.7683	0.3801	19	CHI INV MOT SIS FOM STU
175	030130_cc_lly	LLY	fém.	03;01.30	3.7295	0.6033	22	INV CHI SIS MOT BRO
176	030213_cc_lly	LLY	fém.	03;02.13	3.6378	0.6627	21	CHI INV MOT SIS BRO
177	030216_cc_hhc	HHC	masc.	03;02.16	2.7552	0.408	19	CHI INV SER
178	030218_cc_ltf	LTF	fém.	03;02.18	3.8278	0.4051	25	CHI INV MOT SIS FAT
179	030220_cc_wbh	WBH	fém.	03;02.20	2.7726	0.2493	15	CHI INV MOT GRM
180	030303_cc_wbh	WBH	fém.	03;03.03	2.4464	0.2422	15	CHI MOT FAT
181	030311_cc_hhc	HHC	masc.	03;03.11	2.6623	0.3255	22	INV CHI MOT SER SIS
182	030312_cc_wbh	WBH	fém.	03;03.12	2.7657	0.3392	16	CHI MOT
183	030315_cc_lly	LLY	fém.	03;03.15	3.0129	0.5271	25	CHI INV MOT SIS
184	030326_cc_lly	LLY	fém.	03;03.26	3.1669	0.5073	25	CHI INV MOT BRO
185	030408_cc_wbh	WBH	fém.	03;04.08	3.1424	0.4303	13	CHI INV MOT GRM
186	030414_cc_hhc	HHC	masc.	03;04.14	2.8482	0.4061	20	INV CHI SIS MOT SER
187	030422_cc_lly	LLY	fém.	03;04.22	2.9834	0.5312	27	CHI INV MOT SER
188	030513_hku		masc.	03;05.13	2.5152	0.3535	12	INV CHI
189	030516_hku		fém.	03;05.16	3.7089	0.3608	11	INV CHI
190	030520_cc_lly	LLY	fém.	03;05.20	3.3732	0.5535	25	INV CHI
191	030520_hku		masc.	03;05.20	3.1288	0.4735	16	CHI INV MOT

192	030523_hku		fém.	03;05.23	2.2927	0.0244	1	INV CHI HEL
193	030602_hku		fém.	03;06.02	3.439	0.4065	14	INV CHI IN2
194	030606_hku		masc.	03;06.06	3.6185	0.5202	16	INV CHI
195	030616_cc_lly	LLY	fém.	03;06.16	3.5088	0.4737	25	CHI INV MOT SIS SER
196	030620_hku		fém.	03;06.20	3.0205	0.3699	12	INV CHI IN2
197	030622_hku		masc.	03;06.22	3.1376	0.4893	20	INV CHI
198	030627_hku		fém.	03;06.27	3.9917	0.4545	11	INV CHI IN2
199	030701_hku		masc.	03;07.01	3.5288	0.3942	11	CHI INV
200	030725_cc_lly	LLY	fém.	03;07.25	3.4519	0.615	24	INV CHI MOT
201	030809_cc_lly	LLY	fém.	03;08.09	3.5061	0.6106	25	INV CHI MOT BRO
202	031026_hku		fém.	03;10.26	4.6241	0.7519	15	INV CHI
203	031100_hku		masc.	03;11.00	3.1512	0.4228	20	INV CHI
204	031101_hku		masc.	03;11.01	2.8074	0.163	8	INV CHI
205	031113_hku		masc.	03;11.13	2.9526	0.1211	10	INV CHI
206	031117_hku		fém.	03;11.17	3.4918	0.3279	11	INV CHI
207	031120_hku		masc.	04;00;01	2.8289	0.193	10	CHI INV
208	040021_hku		fém.	04;00.21	4.5185	0.6984	19	INV CHI
209	040024_hku		masc.	04;00.24	2.7615	0.1468	5	INV CHI
210	040112_hku		fém.	04;01.12	4.0413	0.7397	24	INV CHI
211	040113_hku		fém.	04;01.13	4.5894	0.4238	15	INV CHI IN2
212	040416_hku		fém.	04;04.16	3.7027	0.2973	12	INV CHI
213	040515_hku		fém.	04;05.15	2.3269	0.0962	3	INV CHI
214	040519_hku		masc.	04;05.19	4.0928	0.5389	19	INV CHI
215	040523_hku		masc.	04;05.23	3.9745	0.6387	20	INV CHI
216	040527_hku		masc.	04;05.27	3.5225	0.3559	13	INV CHI
217	040600_hku		masc.	04;06.00	3.0787	0.1011	5	INV CHI
218	040701_hku		fém.	04;07.01	4.4088	0.5126	23	INV CHI
219	040702_hku		fém.	04;07.02	5.1858	0.6757	24	CHI INV
220	040703_hku		fém.	04;07.03	4.7381	0.6071	11	INV CHI
221	040705_hku		masc.	04;07.05	4.5236	0.4635	22	INV CHI
222	041002_hku		fém.	04;10.02	3.5938	0.25	6	INV CHI
223	050005_hku		masc.	05;00.05	4.6299	0.5731	22	INV CHI
224	050010_hku		masc.	05;00.10	5.1065	0.6254	17	INV CHI
225	050015_hku		masc.	05;00.15	4.6918	0.491	17	INV CHI
226	050016_hku		fém.	05;00.16	3.6279	0.0698	5	INV CHI
227	050023_hku		masc.	05;00.23	3.568	0.4218	21	INV CHI
228	050025_hku		masc.	05;00.25	3.5374	0.3084	13	INV CHI
229	050118_hku		fém.	05;01.18	2.8265	0.3562	16	INV CHI
230	050121_hku		fém.	05;01.21	3.8084	0.4355	22	CHI INV

231	050214_hku		fém.	05;02.14	2.4605	0.0066	1	INV CHI
232	050429_hku		fém.	05;04.29	4.7653	0.5307	24	INV CHI
233	050505_hku		masc.	05;05.05	3.4444	0.4697	19	INV CHI IN2
234	050506_hku		fém.	05;05.06	3	0.1852	12	INV CHI
235	050512_hku		fém.	05;05.12	3.2364	0.4606	16	CHI INV
236	050525_hku		masc.	05;05.25	4.4819	0.7409	21	INV CHI
237	050527_hku		fém.	05;05.27	6.8505	0.3271	11	INV CHI
238	050615_hku		masc.	05;06.15	3.4722	0.1736	8	INV CHI
239	050703_hku		fém.	05;07.03	4.7459	0.5574	11	INV CHI
240	050706_hku		masc.	05;07.06	3.1727	0.2014	12	INV CHI
241	050800_hku		masc.	05;08.00	4.4332	0.4696	18	INV CHI

ANNEXE C

Expression régulière pour l'extraction des SFP

```
(((?:(?:\b(?:sin1|先)\b|\b(?:tim1|添|啱|嗮)\b) +)?)((?:(?:\b(?:gaa1|gaa2|㗎|架|嚟)\b \b(?:maa3|嘛)\b|\b(?:gaa2|架|嘅)\b|\b(?:gaa3|㗎|㗎3|噶)\b|\b(?:gaa4|㗎|㗎4)\b|\b(?:gaak3|嗰)\b|\b(?:ge2|咖|嘅 2|喺)\b|\b(?:ge3|嘅|嘅 3)\b|\b(?:lei4|嚟)\b|\b(?:lei4|嚟)\b \b(?:gaa3|㗎)\b)
+)?)((?:(?:\b(?:aak3|呃)\b|\b(?:laa1|啦|喇|啦 1)\b|\b(?:laa3|喇|啦 3|嘞)\b|\b(?:laa3|嘞|喇)\b
\b(?:haa2|吓)\b|\b(?:laak3|嘞)\b|\b(?:le4|嘞)\b|\b(?:le5|哩)\b|\b(?:lo1|囉|囉 1)\b|\b(?:lo3|囉|囉 3|嘍)\b|\b(?:lo4|囉)\b|\b(?:lok3|咯)\b|\b(?:zaa1|ze1|zi1|咋|啫|之)\b \b(?:maa3|嘛)\b|\b(?:zaa3|咋|咋 3|喳)\b|\b(?:zaa3|咋)\b \b(?:haa2|吓)\b|\b(?:ze1|啫|姐 1)\b|\b(?:zek1|唧)\b)
+)?)((?:(?:\b(?:aa1|吖)\b|\b(?:aa1|aa2|aa3|aa4|aa5|aa6|吖|呀|啞|啊)\b \b(?:maa3|嘛|嗎
)\b|\b(?:aa3|啊|呀|呀 3)\b|\b(?:aa4|呀|呀 4)\b|\b(?:aa5|啞|呀|呀 5|吼)\b|\b(?:aa6|呀|呀 6|啞
)\b|\b(?:bo3|囉)\b|\b(?:gwaa3|掛|啩)\b|\b(?:haa2|吓|吓 2)\b|\b(?:ho2|呵)\b|\b(?:laa4|嗱|啦
4)\b|\b(?:maa3|嗎|嘛)\b|\b(?:me1|咩)\b|\b(?:ne1|le1|呢)\b|\b(?:waa2|aa2|話)\b|\b(?:wo3|喎
)\b|\b(?:wo4|唲|喎|喎 4)\b|\b(?:wo5|喎|喎 5)\b|\b(?:zaa4|喳|咋 4)\b) +)?))[.,?!]$
```

Chaque SFP est décrite comme l'alternative (indiquée par la barre verticale |) de ses formes jyutping et chinoises. Les sous-expressions correspondant aux différentes positions (Matthews et Yip, 2011, p. 395) sont indiquées par des couleurs. Leur caractère optionnel est indiqué au moyen du marqueur ? situé à la fin de la sous-expression. Une seule SFP est permise pour chacune des positions, chaque position est donc une alternative de toutes les SFP qui peuvent s'y trouver. La ponctuation finale est couverte par le groupe `[.,?!]` entre crochets, et la fin de l'énoncé par le signe `$`.

ANNEXE D

Création d'une table de contingence et génération des visualisations

Code R pour la création d'une table de contingence des segments produits par les enfants selon leur genre (fille/garçon) et le corpus (CC/HKU).

```
> library(vcd) # pour mosaic et analyse des résidus de Pearson
> summary(segments_cs[c('sex', 'corpus')]) # résumé des colonnes 'sex' et 'corpus'
     sex        corpus
 FEMALE:41423   cc :92211
 MALE  :63223   hku:12435
> sexXcorpus_table <- table(sex=segments_cs$sex, corpus=segments_cs$corpus) # création du
tableau de contingence
> sexXcorpus_table
        corpus
sex         cc   hku
  FEMALE 35608  5815
  MALE   56603  6620
> mosaic(sexXcorpus_table, set_varnames=c(sex="Genre", corpus="Corpus"),
set_labels=list(sex=c("Fille", "Garçon"), corpus=c("CC", "HKU")), rot_label=c(0, 0, 0, 0),
offset_varnames=c(0, 0, 0, 3), just_label=c("center", "right"), pop=F) # création du diagramme
de contingence, sans résidus de Pearson
> labeling_cells(text = as.table(sexXcorpus_table), margin = 0)(as.table(sexXcorpus_table)) #
ajout des valeurs numériques associées à chaque paire de valeurs
>
> mosaic(sexXcorpus_table, shade=T, set_varnames=c(sex="Genre", corpus="Corpus"),
set_labels=list(sex=c("Fille", "Garçon"), corpus=c("CC", "HKU")), rot_label=c(0, 0, 0, 0),
offset_varnames=c(0, 0, 0, 3), just_label=c("center", "right"), pop=F) # création du diagramme
de contingence, avec résidus de Pearson
> labeling_cells(text = as.table(sexXcorpus_table), margin = 0)(as.table(sexXcorpus_table)) #
ajout des valeurs numériques
```

ANNEXE E

Ajustement d'un modèle de type GLMER en R

Code R pour l'ajustement du modèle de base intégrant comme effets fixes la tranche de MLU, le type d'énoncé, le nombre de syllabes, le degré de familiarité de l'interlocuteur et la présence d'une SFP à l'énoncé précédent, et comme effet aléatoire l'identifiant de l'enfant.

```
> model <- glmer(data=segments_cs, formula=sfp_found~mlu_group + utt_type + syllables +
household + prev_sfp_found + (1|chid), family="binomial",
control=glmerControl(optimizer="Nelder_Mead"))
> summary(model)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: binomial  ( logit )
Formula: sfp_found ~ mlu_group + utt_type + syllables + household + prev_sfp_found + (1|chid)
   Data: segments_cs
Control: glmerControl(optimizer = "Nelder_Mead")

      AIC       BIC    logLik -2*log(L)  df.resid
 114002.5  114088.5  -56992.3  113984.5    104637

Scaled residuals:
    Min      1Q  Median      3Q     Max
-396.87   -0.63   -0.42    0.85    6.26

Random effects:
 Groups Name        Variance Std.Dev.
 chid   (Intercept) 0.7547   0.8687
Number of obs: 104646, groups:  chid, 78

Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.30486    0.12624 -18.258  < 2e-16 ***
mlu_group2         0.46058    0.03461  13.307  < 2e-16 ***
mlu_group3         0.70343    0.04023  17.484  < 2e-16 ***
mlu_group4         1.08972    0.23786   4.581 4.62e-06 ***
utt_typeSTATEMENT -1.09169    0.03023 -36.114  < 2e-16 ***
syllables          0.50625    0.00497 101.865  < 2e-16 ***
householdTRUE     -0.23051    0.02474  -9.315  < 2e-16 ***
prev_sfp_foundYes  0.29350    0.01547  18.970  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANNEXE F

Comparaison de modèles au moyen de la fonction R anova

Extraits de code R montrant l'application des ANOVA à trois modèles : modèle de base (`modele_base`), modèle avec prise en compte du type de l'énoncé précédent (`modele_type_enonce_prec`), et modèle intégrant une interaction entre la présence d'une SFP à l'énoncé précédent et le type de celui-ci (`modele_interaction`).

```
> anova(modele_base, modele_type_enonce_prec, modele_interaction)
Data: subset(segments_cs, !is.na(prev_utt_type))
Models:
modele_base: sfp_found ~ mlu_group + utt_type + syllables + household + prev_sfp_found + (1 |
chid)
modele_type_enonce_prec: sfp_found ~ mlu_group + utt_type + syllables + household +
prev_sfp_found + prev_utt_type + (1 | chid)
modele_interaction: sfp_found ~ mlu_group + utt_type + syllables + household + prev_sfp_found *
prev_utt_type + (1 | chid)
                       npar    AIC    BIC logLik -2*log(L)   Chisq Df Pr(>Chisq)
modele_base               9 113938 114024 -56960    113920
modele_type_enonce_prec   10 113830 113926 -56905    113810 109.810  1  < 2.2e-16 ***
modele_interaction        11 113756 113861 -56867    113734  75.815  1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> anova(modele_type_enonce_prec, modele_interaction)
Data: subset(segments_cs, !is.na(prev_utt_type))
Models:
modele_type_enonce_prec: sfp_found ~ mlu_group + utt_type + syllables + household +
prev_sfp_found + prev_utt_type + (1 | chid)
modele_interaction: sfp_found ~ mlu_group + utt_type + syllables + household + prev_sfp_found *
prev_utt_type + (1 | chid)
                       npar    AIC    BIC logLik -2*log(L)  Chisq Df Pr(>Chisq)
modele_type_enonce_prec   10 113830 113926 -56905    113810
modele_interaction        11 113756 113861 -56867    113734 75.815  1  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANNEXE G

Couvertures des SFP pour les tranches de MLU de 1 à 4

Le tableau suivant contient les statistiques de couverture (proportion d'enfants utilisant une SFP) pour toutes les SFP considérées dans notre recherche, triées par valeur de couverture à $\bar{m}$=4. La couleur de fond de chaque case donne une indication visuelle de la couverture, de 0 (0%) à 1 (100%).

La présence de plusieurs SFP de couverture 0 en bas de la colonne pour $\bar{m}$=4 est due à des différences de codage dans les deux corpus, ainsi qu'à l'ensemble de SFP effectivement considérées. Par exemple, la particule *aak3*, largement documentée dans la littérature (Kwok, 1984 ; Matthews et Yip, 2011 ; Sybesma et Li, 2007 ; Tang, 2018), n'apparaît que deux fois dans HKU (une fois pour un expérimentateur dans hku-040705 et une fois pour l'enfant dans hku-031100), alors qu'elle est largement représentée dans CC.

$\bar{m}$ / SFP	1	2	3	4
aa3	100.0%	97.4%	100.0%	100.0%
gaa3	20.8%	78.3%	91.6%	100.0%
lo1	0.0%	70.4%	85.5%	100.0%
ne1	12.5%	56.5%	89.2%	94.7%
ge2	4.2%	26.1%	59.0%	89.5%
laa1	25.0%	70.4%	83.1%	89.5%
aa1	58.3%	78.3%	81.9%	84.2%
laa3	29.2%	65.2%	69.9%	84.2%
lei4gaa3	4.2%	43.5%	73.5%	84.2%
sin1	8.3%	42.6%	75.9%	84.2%
zaa3	12.5%	14.8%	44.6%	84.2%
aa4	41.7%	53.0%	60.2%	78.9%
gaa2	12.5%	40.0%	33.7%	78.9%
ge3	12.5%	50.4%	69.9%	78.9%
wo3	12.5%	46.1%	59.0%	78.9%
laak3	12.5%	46.1%	53.0%	73.7%
lei4	41.7%	67.8%	75.9%	73.7%
gaa1maa3	4.2%	0.9%	10.8%	52.6%
gaa4	0.0%	12.2%	22.9%	52.6%
tim1	0.0%	7.0%	26.5%	52.6%
aa1maa3	0.0%	14.8%	34.9%	47.4%

me1	8.3%	20.0%	36.1%	42.1%
maa3	4.2%	22.6%	34.9%	31.6%
ze1	4.2%	4.3%	3.6%	26.3%
bo3	12.5%	21.7%	24.1%	15.8%
laa4	0.0%	14.8%	25.3%	15.8%
aa5	0.0%	3.5%	1.2%	5.3%
haa2	4.2%	26.1%	27.7%	5.3%
lok3	0.0%	18.3%	6.0%	5.3%
waa2	0.0%	7.0%	3.6%	5.3%
wo5	0.0%	0.0%	1.2%	5.3%
zaa1maa3	0.0%	0.9%	2.4%	5.3%
zek1	0.0%	6.1%	7.2%	5.3%
aa6	0.0%	4.3%	2.4%	0.0%
aak3	8.3%	24.3%	34.9%	0.0%
gaak3	0.0%	2.6%	4.8%	0.0%
gwaa3	0.0%	0.9%	1.2%	0.0%
ho2	0.0%	1.7%	0.0%	0.0%
laa3haa2	0.0%	0.0%	1.2%	0.0%
le4	0.0%	0.0%	0.0%	0.0%
le5	0.0%	5.2%	3.6%	0.0%
lo3	4.2%	28.7%	22.9%	0.0%
lo4	0.0%	3.5%	4.8%	0.0%
wo4	12.5%	11.3%	14.5%	0.0%
zaa3haa2	0.0%	0.0%	0.0%	0.0%
zaa4	8.3%	3.5%	7.2%	0.0%

ANNEXE H

Fréquences relatives moyennes d'utilisation des SFP

La table suivante présente les fréquences relatives moyennes d'utilisation de chacune des SFP observées pour les tranches de MLU 1, 2, 3 et 4, triées selon les valeurs obtenues à m̄=4, ainsi que dans le CDS et l'ADS.

discours / SFP	m̄=1	m̄=2	m̄=3	m̄=4	CDS	ADS
aa3	77.2%	58.3%	41.1%	24.7%	28.6%	23.0%
gaa3	2.2%	6.0%	10.4%	15.1%	6.9%	9.9%
lo1	0.0%	3.6%	5.4%	5.0%	2.7%	9.0%
ne1	0.9%	2.4%	4.2%	4.4%	4.2%	14.0%
ge2	0.1%	0.8%	2.1%	5.1%	1.0%	2.2%
laa1	1.2%	3.7%	5.0%	7.0%	8.5%	9.9%
aa1	3.4%	3.7%	3.6%	3.1%	6.2%	2.1%
laa3	1.2%	2.7%	4.2%	4.2%	3.7%	3.7%
lei4gaa3	0.7%	2.7%	2.6%	4.5%	3.3%	0.8%
sin1	1.9%	1.2%	2.9%	4.2%	3.4%	0.8%
zaa3	0.9%	0.3%	1.8%	2.4%	0.3%	0.7%
aa4	2.0%	1.8%	2.0%	3.0%	8.2%	1.7%
gaa2	1.2%	1.2%	1.0%	2.2%	0.8%	0.1%
ge3	0.2%	1.9%	2.8%	2.7%	1.5%	4.5%
wo3	1.6%	1.0%	1.6%	2.0%	6.8%	5.2%
laak3	0.3%	1.7%	1.8%	4.0%	2.8%	3.0%
lei4	1.2%	3.0%	1.9%	1.0%	1.5%	0.0%
gaa1maa3	0.0%	0.0%	0.2%	0.7%	0.4%	1.4%
gaa4	0.0%	0.2%	0.6%	1.2%	0.9%	0.3%
tim1	0.0%	0.0%	0.4%	0.7%	0.2%	0.4%
aa1maa3	0.0%	0.2%	0.7%	0.6%	0.7%	1.8%
me1	0.2%	0.3%	0.8%	0.9%	0.9%	1.4%
maa3	0.1%	0.5%	0.6%	0.3%	0.3%	0.2%
ze1	0.8%	0.0%	0.0%	0.2%	0.1%	0.9%
bo3	1.6%	0.4%	0.4%	0.2%	0.3%	0.1%
laa4	0.0%	0.1%	0.3%	0.2%	1.1%	0.2%
aa5	0.0%	0.0%	0.0%	0.1%	0.1%	0.1%
haa2	0.0%	0.5%	0.6%	0.1%	1.0%	0.2%

lok3	0.0%	0.2%	0.1%	0.3%	0.3%	0.1%
waa2	0.0%	0.1%	0.0%	0.0%	1.3%	0.1%
wo5	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%
zaa1maa3	0.0%	0.0%	0.0%	0.0%	0.2%	0.5%
zek1	0.0%	0.1%	0.1%	0.0%	0.6%	0.5%
aa6	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%
aak3	0.1%	0.5%	0.6%	0.0%	0.4%	0.3%
gaak3	0.0%	0.1%	0.0%	0.0%	0.1%	0.2%
gwaa3	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%
ho2	0.0%	0.0%	0.0%	0.0%	0.1%	0.1%
laa3haa2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
le4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
le5	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
lo3	0.2%	0.5%	0.1%	0.0%	0.3%	0.2%
lo4	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
wo4	0.1%	0.1%	0.1%	0.0%	0.1%	0.1%
zaa3haa2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
zaa4	0.7%	0.1%	0.1%	0.0%	0.1%	0.0%

BIBLIOGRAPHIE

Aijmer, K. (2014). Pragmatic markers. Dans K. Aijmer et C. Rühlemann (dir.), *Corpus Pragmatics: A Handbook* (p. 195-218). Cambridge University Press.

Andrick, G. R. et Tager-Flusberg, H. (1986). The acquisition of colour terms. *Journal of Child Language*, *13*(1), 119-134. https://doi.org/10.1017/S0305000900000337

Bates, D., Mächler, M., Bolker, B. et Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Bauer, R. S. (2018). Cantonese as written language in Hong Kong. *Global Chinese*, *4*(1), 103-142. https://doi.org/10.1515/glochi-2018-0006

Biecek, P. (2018). DALEX: Explainers for complex predictive models in R. *Journal of Machine Learning Research*, *19*(84), 1-5. https://jmlr.org/papers/v19/18-416.html

Bloom, L. (1991). *Language development from two to three*. Cambridge University Press. https://archive.org/details/languagedevelopm0000bloo

Bornstein, M. H., Hahn, C.-S. et Haynes, O. M. (2004). Specific and general language performance across early childhood: Stability and gender considerations. *First Language*, *24*(3), 267-304. https://doi.org/10.1177/0142723704045681

Breiss, C. et Hayes, B. (2020). Phonological markedness effects in sentence formation. *Language*, *96*(2), 338-370. https://doi.org/10.1353/lan.0.0243

Brown, P. et Gaskins, S. (2014). Language acquisition and language socialization. Dans N. J. Enfield, P. Kockelman et J. Sidnell (dir.), *The Cambridge Handbook of Linguistic Anthropology* (1re éd., p. 187-226). Cambridge University Press. https://doi.org/10.1017/CBO9781139342872.010

Brown, R. (1973). *A first language: the early stages* (7. printing). Harvard Univ. Press.

Casillas, M., Brown, P. et Levinson, S. C. (2020). Early Language Experience in a Tseltal Mayan Village. *Child development*, *91*(5), 1819-1835. https://doi.org/10.1111/cdev.13349

Chan, M. K. M. (2000). Sentence-final particles in Cantonese: A gender-linked survey and study. Dans *Eleventh North American Conference on Chinese Linguistics (NACCL 11)* (p. 87-101). Cambridge: East Asian Language Programs. https://www.davidcong.net/wp-content/uploads/2012/05/naccl-11.pdf

Chan, M. K. M. (2002). Chinese. Gender-related use of sentence-final particles in Cantonese. Dans M. Hellinger et H. Bußman (dir.), *Gender Across Languages: The linguistic representation of women and men* (vol. 2, p. 57-72). John Benjamins Publishing Company. https://doi.org/10.1075/impact.10.08cha

Chor, W. (2018). Sentence final particles as epistemic modulators in Cantonese conversations: A discourse-pragmatic perspective. *Journal of Pragmatics*, *129*, 34-47. https://doi.org/10.1016/j.pragma.2018.03.008

Clark, E. V. (2017). Conversation and language acquisition: A pragmatic approach. *Language Learning and Development*, *14*(3), 170-185. https://doi.org/10.1080/15475441.2017.1340843

de Marneffe, M.-C., Manning, C. D., Nivre, J. et Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255-308. https://doi.org/10.1162/coli_a_00402

DeThorne, L. S., Johnson, B. W. et Loeb, J. W. (2005). A closer look at MLU: What does it really measure? *Clinical Linguistics & Phonetics*, *19*(8), 635-648. https://doi.org/10.1080/02699200410001716165

Fletcher, P. J., Leung, C.-S. S., Stokes, S. F. et Weizman, Z. (2000). *Cantonese pre-school language development: a guide*. Department of Speech and Hearing Science. https://api.semanticscholar.org/CorpusID:151342053

Foolen, A. (1997). Pragmatic particles. Dans J. Verschueren, J.-O. Östman, J. Blommaert et C. Bulcaen (dir.), *Handbook of Pragmatics* (p. 1-24). John Benjamins Publishing Company. https://doi.org/10.1075/hop.2.pra3

Fujimoto, M. (2008). *L1 acquisition of Japanese particles: A corpus-based study* [Thèse de doctorat, City University of New York]. CUNY Graduate Center. https://academicworks.cuny.edu/gc_etds/3931/

Fung, R. S. Y. (2000). *Final particles in standard Cantonese: semantic extension and pragmatic inference* [Thèse de doctorat, Ohio State University]. OhioLINK Electronic Theses and Dissertations Center.

Gaskins, S. (2006). Cultural perspectives on infant-caregiver interaction. Dans N. J. Enfield et S. C. Levinson (dir.), *Roots of Human Sociality: Culture, Cognition and Interaction* (p. 279-298). Berg Publishers. https://doi.org/10.4324/9781003135517-14

Gayraud, F., Lanoë, J.-L. et De Agostini, M. (2025). Factors influencing language performance in boys and girls at age 2 in the French ELFE birth cohort. *Brain research*, *1847*, 149305. https://doi.org/10.1016/j.brainres.2024.149305

Gibbons, J. (1980). A tentative framework for speech act description of the utterance particle in conversational Cantonese. *Linguistics*, *18*(9-10), 763-776. https://doi.org/doi:10.1515/ling.1980.18.9-10.763

Hara, Y. (2023). Cantonese question particles. Dans E. McCready et H. Nomoto, *Discourse Particles in Asian Languages Volume I* (1re éd., p. 112-142). Routledge. https://doi.org/10.4324/9781351057837-6

Harris, P. L. (2007). Commentary: Time for questions. *Monographs of the Society for Research in Child Development*, *72*(1), 113-120. https://doi.org/10.1111/j.1540-5834.2007.00421.x

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J. et Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive psychology*, *61*(4), 343-365. https://doi.org/10.1016/j.cogpsych.2010.08.002

Kidd, E. et Donnelly, S. (2020). Individual differences in first language acquisition. *Annual Review of Linguistics*, *6*(1), 319-340. https://doi.org/10.1146/annurev-linguistics-011619-030326

Klee, T., Stokes, S. F., Wong, A. M.-Y., Fletcher, P. et Gavin, W. J. (2004). Utterance length and lexical diversity in Cantonese-speaking children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*, *47*(6), 1396-1410. https://doi.org/10.1044/1092-4388(2004/104)

Ko, C. P. (2000). *Form and function of sentence final particles in Cantonese-speaking children* [Bachelor, University of Hong Kong]. The HKU Scholars Hub. https://hub.hku.hk/handle/10722/56558

Kwok, H. (1984). *Sentence particles in Cantonese*. Centre of Asian Studies, University of Hong Kong.

Lam, C., Lau, C. et Lee, J. L. (2024). Multi-tiered Cantonese word segmentation. Dans N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti et N. Xue (dir.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (p. 11993-12002). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1047/

Law, A. et Lee, T. H. T. (1998). Time, focus and evidentiality in the final particles of children in Cantonese. Dans *First Asia-Pacific Conference on Speech Language and Hearing*.

Law, Y. K. A. (2004). *Sentence-final focus particles in Cantonese* [Thèse de doctorat, University of London]. UCL Discovery. https://discovery.ucl.ac.uk/id/eprint/10101610/

Lee, T. H. et Law, A. (2000). Evidential final particles in child Cantonese. Dans E. V. Clark (dir.), *Thirtieth Annual Child Language Research Forum* (p. 131-138). Center for the Study of Language and Information.

Lee, T. H. T., Wong, C. H., Leung, S., Man, P., Cheung, A., Szeto, K. et Wong, C. S. P. (1996). *The development of grammatical competence in Cantonese-speaking children*. Report of RGC earmarked grant 1991-94.

Leung, M. et Li, H.-L. (2015). Hierarchical phrase-based grammatical analysis of language samples from Cantonese-speaking children with and without autism. *Clinical linguistics & phonetics*, *29*(8-10), 736-747. https://doi.org/10.3109/02699206.2015.1048529

Leung, W.-M. (2008). Promising approaches for the analysis of sentence-final particles in Cantonese: The case of [aa3]. *Asian Social Science*, *4*(5), 74-82. https://doi.org/10.5539/ass.v4n5p74

Li, H., Tse, S., Sin Wong, J. M., Mei Wong, E. C. et Leung, S. O. (2013). The development of interrogative forms and functions in early childhood Cantonese. *First Language*, *33*(2), 168-181. https://doi.org/10.1177/0142723713479422

Lim, L. Y. F. (2018). *A corpus-based study of the acquisition of sentence final particles in Cantonese speaking children aged 2;0 to 5;11* [Bachelor, University of Hong Kong]. The HKU Scholars Hub. https://hub.hku.hk/handle/10722/287522

Liu, H. et MacWhinney, B. (2024). *Morphosyntactic analysis for CHILDES*. https://doi.org/10.48550/arXiv.2407.12389

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P. et Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

Luke, K.-K. (1990). *Utterance particles in Cantonese conversation* (vol. 9). John Benjamins. https://doi.org/10.1075/pbns.9

Luke, K.-K. et Wong, M. L. (2015). The Honk Kong Cantonese corpus: Design and uses. *Journal of Chinese Linguistics Monograph Series*, *25*, 312-333. https://www.jstor.org/stable/26455290

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs* (3e éd., vol. 1). Lawrence Erlbaum Associates.

MacWhinney, B. (2018). *Tools for analyzing talk part 2: The CLAN program*. Lawrence Erlbaum Associates. https://doi.org/10.21415/T5G10R

Matsui, T. et Miura, Y. (2009). Children's understanding of certainty and evidentiality: Advantage of grammaticalized forms over lexical alternatives. *New Directions for Child and Adolescent Development*, *2009*(125), 63-77. https://doi.org/10.1002/cd.250

Matthews, S. et Yip, V. (2011). *Cantonese: A comprehensive grammar* (2nd éd.). Taylor & Francis. https://doi.org/10.4324/9780203835012

Meltzoff, A. N. (1999). Origins of theory of mind, cognition and communication. *Journal of communication disorders*, *32*(4), 251-269. https://doi.org/10.1016/s0021-9924(99)00009-x

Meyer, D., Zeileis, A. et Hornik, K. (2006). The Strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software*, *17*(3), 1-48. https://doi.org/10.18637/jss.v017.i03

Meyer, D., Zeileis, A., Hornik, K. et Friendly, M. (2024). *vcd: Visualizing categorical data*. https://CRAN.R-project.org/package=vcd

Miller, J. F. et Chapman, R. S. (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech, Language, and Hearing Research*, *24*(2), 154-161. https://doi.org/10.1044/jshr.2402.154

Nakagawa, S., Johnson, P. C. D. et Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, *14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Nakagawa, S. et Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133-142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Nikolaus, M., Maes, J. et Fourtassi, A. (2021). Modeling speech act development in early childhood: the role of frequency and linguistic cues. Dans *43rd Annual Meeting of the Cognitive Science Society.* https://doi.org/10.31234/osf.io/cgs5n

Nomoto, H. et McCready, E. (2023). *Discourse particles in Asian languages volume II: Southeast Asia*. Taylor & Francis. https://books.google.se/books?id=azzLEAAAQBAJ

Patil, I. (2021). Visualizations with statistical details: The « ggstatsplot » approach. *Journal of Open Source Software*, *6*(61), 3167. https://doi.org/10.21105/joss.03167

Paul, W. et Yan, S. (2022). Sentence-final particles in Mandarin Chinese. Syntax, semantics and acquisition. Dans A. E. Xabier Artiagoitia et S. Monforte (eds.) (dir.), *Discourse Particles. Syntactic, semantic, pragmatic and historical aspects.* (vol. 276, p. 179-206). John Benjamins Publishing Company. https://doi.org/10.1075/la.276.07pau

Rice, M. L., Smolik, F., Perpich, D., Thompson, T., Rytting, N. et Blossom, M. (2010). Mean length of utterance levels in 6-Month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research*, *53*(2), 333-349. https://doi.org/10.1044/1092-4388(2009/08-0183)

Santos, M. E., Lynce, S., Carvalho, S., Cacela, M. et Mineiro, A. (2015). Mean length of utterance-words in children with typical language development aged 4 to 5 years. *Revista CEFAC*, *17*(4), 1143-1151. https://doi.org/10.1590/1982-021620151741315

Saxton, M. (2009). The Inevitability of child directed speech. Dans S. Foster-Cohen (dir.), *Language Acquisition* (p. 62-86). Palgrave Macmillan. https://doi.org/10.1057/9780230240780_4

Schäfer, R. (2020). Mixed-effects regression modeling. Dans M. Paquot et S. Th. Gries (dir.), *A Practical Handbook of Corpus Linguistics* (p. 535-561). Springer International Publishing. https://doi.org/10.1007/978-3-030-46216-1_22

Shirai, J., Shirai, H. et Furuta, Y. (2000). Acquisition of sentence-final particles in Japanese. Dans M. Perkins et S. Howard (dir.), *New Directions In Language Development And Disorders* (p. 243-250). Springer US. https://doi.org/10.1007/978-1-4615-4157-8_23

Snow, D. (2004). *Cantonese as written language: The growth of a written Chinese vernacular*. Hong Kong University Press. https://books.google.nl/books?id=pFnP_FXf-lAC

So, L. K. H. et Dodd, B. J. (1995). The acquisition of phonology by Cantonese-speaking children. *Journal of Child Language*, *22*(3), 473-495. https://doi.org/10.1017/S0305000900009922

Spinelli, M., Suttora, C., Garcia-Sierra, A., Franco, F., Lionetti, F. et Fasolo, M. (2023). Editorial: Are there different types of child-directed speech? Dynamic variations according to individual and contextual factors. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.1056816

Stokes, S. F. et Fletcher, P. (2000). Lexical diversity and productivity in Cantonese-speaking children with specific language impairment. *International Journal of Language & Communication Disorders*, *35*(4), 527-41. https://doi.org/10.1080/136828200750001278

Sybesma, R. et Li, B. (2007). The dissection and structural mapping of Cantonese sentence final particles. *Lingua*, *117*(10), 1739-1783. https://doi.org/10.1016/j.lingua.2006.10.003

Tang, S.-W., Kwok, F., Lee, T. H.-T., Lun, C., Luke, K. K., Tung, P. et Cheung, K. H. (2002). *Guide to LSHK Cantonese romanization of Chinese characters*. Linguistic Society of Hong Kong.

Tang, T. Y. (2018). *The use of the epistemic modality in Cantonese sentence final particles for children from the age of three to five* [Mémoire de master, Université Paris III – Sorbonne Nouvelle].

Tardif, T. et Wellman, H. (2000). Acquisition of mental state language in Mandarin and Cantonese-speaking children. *Developmental psychology*, *36*(1), 25-43. https://doi.org/10.1037/0012-1649.36.1.25

The Editors of Encyclopaedia Britannica. (2024, 3 février). Cantonese Language. Dans *Encyclopedia Britannica*. https://www.britannica.com/topic/Cantonese-language

To, C. K.-S., Stokes, S. F., Cheung, H.-T. et T'sou, B. (2010). Narrative assessment for Cantonese-speaking children. *Journal of Speech, Language, and Hearing Research*, *53*(3), 648-669. https://doi.org/doi:10.1044/1092-4388(2009/08-0039)

Tsang, K. et Stokes, S. (2001). Syntactic awareness of Cantonese-speaking children. *Journal of Child Language*, *28*(3), 703-739. https://doi.org/10.1017/S0305000901004822

Tse, S. K., Chan, C., Kwong, S. M. et Li, H. (2002). Sex differences in syntactic development: Evidence from Cantonese-speaking preschoolers in Hong Kong. *International Journal of Behavioral Development*, *26*(6), 509-517. https://doi.org/10.1080/01650250143000463

Tse, S. K., Li, H. et Leung, S. O. (2007). The acquisition of Cantonese classifiers by preschool children in Hong Kong. *J Child Lang*, *34*(3), 495-517. https://doi.org/10.1017/s0305000906007975

Tulling, M. (2017). Finnish pragmatic enclitics, sentence final particle in second position? The 19th Annual SYNC Graduate Student Conference.

Wakefield, J. (2011). *The English equivalent of Cantonese sentence-final particles: A contrastive analysis* [Thèse de doctorat, Hong Kong Polytechnic University]. PolyU Electronic Theses. https://theses.lib.polyu.edu.hk/handle/200/6084

Wang, L. et Wong, P. C. M. (2024). Age-related changes in lexical tones and intonation in Cantonese infant-directed speech: A longitudinal study. *Journal of Child Language*, *52*(5), 1080-1103. https://doi.org/10.1017/S0305000924000333

Wei, T. et Simko, V. (2024). *R package « corrplot »: Visualization of a correlation matrix*. https://github.com/taiyun/corrplot

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Winterstein, G., Lai, R. et Luk, P. S. (en préparation). *Authority and gendered speech: Cantonese particles and Sajiao*.

Winterstein, G., Lai, R. et Luk, Z. P. (2023a). Softness, assertiveness and their expression via Cantonese sentence-final particles. Dans E. McCready et H. Nomoto (dir.), *Discourse Particles in Asian Languages Volume I* (1re éd., p. 143-170). Routledge. https://doi.org/10.4324/9781351057837-7

Winterstein, G., Vergnaud, D., Yu, H. H. T., Lupien, J., Laperle, S., Luk, P. S. et Davis, C. (2023b). An empirical, corpus-based, approach to Cantonese nominal expressions. Dans C.-R. Huang, Y. Harada, J.-B. Kim, S. Chen, Y.-Y. Hsu, E. Chersoni, P. A, W. H. Zeng, B. Peng, Y. Li et J. Li (dir.), *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation* (p. 436-445). Association for Computational Linguistics. https://aclanthology.org/2023.paclic-1.43

Wong, W. et Ingram, D. (2003). Question acquisition by Cantonese speaking children. *Journal of Multilingual Communication Disorders*, *1*(2), 148-157. https://doi.org/10.1080/1476967031000091024

Yang, X. (2022). Functional word acquisition in Mandarin Chinese. Dans *Oxford Research Encyclopedia of Linguistics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780199384655.013.919

Yau, S. C. (1965). *A study of the functions and of the presentations of Cantonese sentence particles* [Mémoire de master, Hong Kong University]. https://doi.org/10.5353/th_b3194643

Yin, W. K. (1996). A functional classification of questions in Cantonese. *Journal of Chinese Linguistics*, *24*(2), 355-390. http://www.jstor.org/stable/23753959

Zeileis, A., Meyer, D. et Hornik, K. (2007). Residual-based shadings for visualizing (conditional) independence. *Journal of Computational and Graphical Statistics*, *16*(3), 507-525. https://doi.org/10.1198/106186007X237856

Zhang, D., Xu, Q. et W. Elliott, R. (2019). The construction of acquisition pattern of sentence-final particles. *Language and Cognitive Science*, *5*(1), 1-20. https://doi.org/10.35534/LCS201905001