Usages des corpus oraux pour l'étude de l'acquisition du français langue maternelle

Caroline Masson, Christine da Silva, Emmanuelle Canut, Stéphanie Caët

Collection **MSH Lorraine**

Les corpus en sciences humaines et sociales

Sous la direction de
Christophe Benzitoun & Manuel Rebuschi

Ouvrage publié grâce au soutien de la MSH Lorraine

PUN - Éditions Universitaires de Lorraine

1
USAGES DES CORPUS ORAUX POUR L'ÉTUDE DE L'ACQUISITION DU FRANÇAIS LANGUE MATERNELLE

Caroline Masson[1,3], Christine da Silva Genest[2,3], Emmanuelle Canut[4] & Stéphanie Caët[4]

[1]Université Sorbonne Nouvelle–Paris 3, CLESTHIA (EA 7345), [2]Université de Lorraine, DevAH (EA 3450), [3]ATILF (UMR 7118), [4]Université de Lille, STL (UMR 8163)

Pour étudier le langage d'enfants à développement typique ou atypique entre 0 et 6 ans et rendre compte de ce développement *in situ*, le recueil de données orales en situations spontanées représente l'une des méthodologies les plus pertinentes. Plusieurs bases de données audio ou vidéo existent, tout comme différents logiciels de transcription et d'annotation qui sont fonction des objectifs de recherche (analyse des productions orales et/ou gestuelles, des interactions, développement lexical et/ou syntaxique...). Nous proposons dans ce chapitre un état des lieux de l'usage des corpus oraux pour traiter de l'acquisition du français chez l'enfant. Il s'agira de présenter les spécificités (et les écueils) du recueil, de la transcription et de l'analyse du langage des jeunes enfants dans les développements typique et atypique, afin de montrer l'intérêt des données naturelles et écologiques pour compléter ou suppléer les données expérimentales et leur pertinence pour la formation des professionnels de l'enfance.

1 Caractéristiques des corpus oraux en acquisition du langage

Un corpus en sciences du langage peut être défini comme une collection de données langagières que l'on peut interroger à l'aide de techniques et technologies dédiées, dont la taille peut varier du simple énoncé à un ensemble de discours. Identifier ce que sont et ce que contiennent les corpus oraux en acquisition langue première n'est pas chose aisée. En effet, le domaine des études sur l'acquisition première du langage présente une grande hétérogénéité qui est liée à la diversité des contextes épistémologiques, théoriques et méthodologiques ainsi qu'aux différentes disciplines concernées : psychologie, linguistique, psycholinguistique, sciences cognitives, sciences de l'éducation ou encore didactique. Si les chercheurs s'accordent sur la nécessité de travailler sur des données attestées pour observer le développement typique ou atypique du langage, en complément ou non de données expérimentales, les caractéristiques des corpus recueillis dépendent des choix opérés : nombre et âges des enfants, environnement (école, maison, laboratoire), nombre et type d'interlocuteurs (parents, enseignants, expérimentateurs...), situations (spontanée, avec ou sans tâche et support prédéterminés, expérimentale...), taille du corpus (durée des enregistrements, nombre de productions langagières...), type de recueil (transversal, longitudinal), etc., qui sont liés aux objectifs de l'étude visée et qui ont un impact sur le choix du média (notes manuscrites, audio, vidéo), des conventions de transcription et des logiciels de traitement des données (Behrens 2008, Canut & Vertalier 2008, Demuth 2008, da Silva Genest & Masson 2017).

1.1 Historique

En retraçant l'histoire des études en acquisition, on peut distinguer deux grands types d'approches, qualitatif et quantitatif, qui peuvent en partie expliquer la tendance actuelle à vouloir utiliser des corpus oraux de grande taille pour étudier le langage de l'enfant. Selon Ingram (1989), trois périodes sont à considérer :

— De 1876 à 1926 : période des monographies parentales, qui sont des collectes manuscrites quasi quotidiennes réalisées par les chercheurs observant le langage de leurs propres enfants (Taine 1876, Darwin 1877, Preyer 1887, Stern & Stern 1907, Cohen 1925). Même si de nouvelles orientations méthodologiques voient

le jour à la fin des années 1920, ces études qualitatives ne disparaissent pas complètement (par exemple Weir 1962, Dromi 1988).

— De 1926 à 1957 : période des études expérimentales, quantitatives, comparatives, transversales, en appui sur les théories behavioristes, fondées sur le recueil d'échantillons de langage de nombreux sujets d'âges différents. Le langage est testé par des séries d'épreuves standardisées, portant sur l'articulation, la répétition de mots ou de logatomes, la discrimination de syllabes, le développement du vocabulaire ou de la phrase afin de déterminer des normes de développement (McCarthy 1946). Ce type de recueil annonce le début d'un développement important des études sur des corpus dits « transversaux » qui se poursuit à l'heure actuelle.

— De 1957 à la fin des années 1990 : période de recueils d'échantillons de langage dits « représentatifs » dans les corpus longitudinaux de quelques enfants en interaction avec l'expérimentateur ou un membre de leur famille. Des enregistrements sont réalisés à intervalles réguliers et pour une durée prédéfinie (en fonction du type d'étude et de l'âge de l'enfant), avec l'appui éventuel de journaux parentaux. Les travaux de Braine (1963), Brown (1973) et Bloom (1970) font figure de pionniers, et ce type de recueil s'est généralisé dans les études en acquisition.

Depuis une trentaine d'années, une quatrième période peut être envisagée, liée à l'essor de la linguistique de corpus et à l'accessibilité de données en nombre plus conséquent : à l'heure actuelle, les chercheurs en acquisition peuvent s'appuyer sur des corpus plus étendus (comportant un plus grand nombre d'enfants dans chaque tranche d'âge, des situations d'échanges plus diversifiées, sur plusieurs heures de recueil effectuées à une même période ; Lieven *et al.* 2009), et ainsi décrire plus finement le langage de l'enfant à différents stades de développement, mais aussi identifier et généraliser des processus d'apprentissage en lien avec l'étayage proposé par les adultes. Les avancées technologiques reprennent de façon novatrice le travail précurseur des monographies du début du XX^{e} siècle, avec des possibilités de vérifications et de compilation de multiples corpus rendant possible le passage d'études qualitatives à des études quantitatives sur un grand nombre de données *in situ* et de vérifier le cas échéant les données issues de recherches plus expérimentales.

1.2 Du recueil au traitement des données

Recueillir et transcrire la langue parlée ne sont pas chose aisée et posent de multiples questions juridiques et techniques (Blanche-Benveniste & Jeanjean 1987). Avec des enfants de moins de 6 ans, dans le cadre de développements typique et atypique, d'autres contraintes et spécificités sont à prendre en compte d'un point de vue juridique, technique et méthodologique. L'extraction et le codage des données posent notamment des problèmes particuliers liés au fait que le langage de l'enfant n'est pas stabilisé, qu'il est en constante évolution, et qu'il n'est donc pas toujours possible de s'appuyer sur les contraintes repérées pour le langage de l'adulte.

1.2.1 Contraintes éthiques et juridiques

Les questions d'éthique liées à la création de corpus sont devenues centrales. Un guide des « bonnes pratiques » pour la constitution, l'étude et la conservation de données orales (Baude 2006) a d'ailleurs été élaboré pour répondre, entre autres, à ces enjeux. Depuis le 16 novembre 2016, la loi Jardé régit le cadre des recherches impliquant la personne humaine (RIPH), qu'elles soient interventionnelles, inscrites dans le cadre d'une pratique clinique habituelle ou observationnelles. Quel que soit le type, toutes les recherches devraient être présentées à un Comité de Protection de la Personne (CPP). Le décret n° 2017-884 du 9 mai 2017 simplifie cependant certaines dispositions règlementaires de la Loi Jardé en précisant le champ des RIPH soumises à l'avis des CPP. Ainsi, ne sont pas considérées comme des RIPH les recherches qui n'ont pas pour objet le développement des connaissances biologiques ou médicales et celles qui, bien que réalisées dans le domaine de la santé, visent à réaliser des expérimentations en sciences humaines et sociales.

Néanmoins, toute collecte de données doit être réalisée dans le respect de la loi *Informatique et libertés* en recueillant le consentement éclairé des participants et en informant les participants du cadre de la recherche, de l'utilisation des données et de leurs droits notamment d'accès et de rétractation. Dans le cadre des recherches en acquisition, le consentement est donné par les parents mais doit être redemandé aux enfants, une fois devenus majeurs.

Ces dispositions se trouvent renforcées depuis la mise en application du règlement général sur la protection des données (RGPD) du 25 mai

2018 qui assure aux personnes concernées d'être (mieux) informées sur la raison de la collecte des données, la façon dont elles seront traitées et sur leurs droits vis-à-vis de celles-ci.

1.2.2 Diffusion des corpus

Enregistrer et surtout transcrire sont une opération coûteuse et chronophage, en particulier dans le cas de jeunes enfants dont le langage est en développement. En effet, les productions ne sont pas toujours orthographiables et nécessitent très souvent le recours à la phonétique et à la dimension multimodale pour rendre les propos compréhensibles. Pour exemple, la constitution et la transcription des 160 premières heures d'enregistrements du corpus CoLaJE (Morgenstern & Parisse 2012) ont nécessité près de cinq mille heures de travail.

La mise à disposition des corpus à la communauté des chercheurs est précieuse et c'est pourquoi plusieurs initiatives de diffusion ont vu le jour. La base de données la plus importante à ce jour est regroupée sous le projet international CHILDES [1] (MacWhinney & Snow 1985, MacWhinney 2000). En 2014, elle comportait des corpus dans plus de 32 langues pour un total de plus de 44 millions de mots, téléchargeables gratuitement. Pour le français langue première, étaient archivés les corpus longitudinaux et transversaux d'enfants à développement typique pour un total d'environ 2,8 millions de mots (enfants et adultes compris), ce qui représente environ un millier d'heures de langage transcrit. En revanche, il existe très peu, dans cette base de données, de corpus d'enfants francophones à développement atypique. En effet, seules les transcriptions de trois corpus sont accessibles (Le Normand 1997, Bang & Nadig 2015, Foudon *et al.* 2007) et les enregistrements audio ne sont pas tous disponibles. De plus, elles ne concernent qu'un nombre limité de pathologies (troubles du spectre autistique et troubles spécifiques du développement du langage).

En France, d'autres initiatives, comme les projets CoLaJE et TCOF (André & Canut 2010), archivés sur le réservoir de données

1. https://childes.talkbank.org.

ORTOLANG[2], ont vu le jour. Le volume de corpus y est encore modeste, relativement aux données de CHILDES, mais sont en pleine expansion[3].

1.2.3 Taille des corpus oraux

Malgré la quantité *a priori* conséquente des données, les corpus oraux sont en réalité souvent de taille réduite (Parisse 2014). En effet, le nombre relativement faible d'enfants et la grande variabilité des situations observées ne permettent pas toujours de comparer les données. Or, pour que les études sur corpus puissent permettre de dégager des tendances générales, il est nécessaire de vérifier la représentativité des corpus et de déterminer la quantité de données suffisante à une généralisation. Si le nombre de mots est une mesure souvent utilisée, elle n'est pas toujours celle qui convient le mieux. En effet, selon le phénomène linguistique étudié, un « petit » corpus bien ciblé peut être pertinent pour certains types d'analyses (Koester 2010). La fréquence, la durée et le nombre d'enfants des échantillons de langage sélectionnés sont également à questionner (Rowland *et al.* 2008, Tomasello & Stahl 2004). Par exemple, en disposant d'un enregistrement par mois, le chercheur n'aura que trois chances sur trente d'observer un phénomène qui se produit trente fois par jour (sachant qu'un enfant peut potentiellement avoir des échanges dix heures par jour en moyenne). Le problème s'accroît avec l'étude de phénomènes rares ou atypiques, comme les emplois de surgénéralisation (« *j'ai prendu* »), le découpage de la chaine sonore (« *parmi nous* » segmenté « *par minou* »), les liaisons (« *des n-ours* »), les créations lexicales (« *cancer de la fumation* »), etc. Les corpus spontanés n'étant pas toujours assez importants pour observer ces phénomènes, le recours à des situations semi-expérimentales permettent de les provoquer (Dugua *et al.* 2017).

1.2.4 Techniques de recueil

Les enregistrements audio sont massivement représentés dans les corpus en acquisition. La raison principale est qu'il est plus simple, moins coûteux (en temps) et moins intrusif et perturbant pour les adultes et les

2. www.ortolang.fr.

3. Nous proposons en annexe un répertoire de corpus disponibles d'enfants francophones monolingues au développement typique et atypique qui se veut le plus exhaustif possible.

enfants d'être enregistrés que filmés. La vidéo est cependant de plus en plus utilisée car elle offre en supplément la possibilité d'étudier les aspects multimodaux du langage, et permet ainsi d'interpréter plus aisément les productions des participants en levant certaines ambiguïtés grâce au contexte. L'exemple ci-dessous, extrait d'une interaction orthophoniste/enfant, illustre l'importance d'avoir recours à la vidéo : sans elle, tout le travail gestuel déployé par l'orthophoniste qui soutient le discours de l'enfant ne pourrait pas être intégré dans le traitement des données.

ORT 34	–	donc la fille a mis la chaise : + ((fait le geste "sur" avec ses mains))
ETE 33	–	dans .
ORT 35	–	dans ? = ((en faisant le geste "dans" avec ses mains))
ETE 34	–	((fait un geste de négation de la tête)) [dy]
ORT 36	–	s :: + ((fait le geste "sur" avec ses mains))
ETE 35	–	sous .
ORT 37	–	nan . pas sous . = ((met ses mains sous la table)) mais su ::r .= ((fait le geste "sur" avec ses mains))
ETE 36	–	sur .
ORT 38	–	sur :: + ((pointe du doigt la table sur la table))
ETE 37	–	la table

Exemple 1 – Extrait d'une interaction orthophoniste-enfant (Corpus da Silva, 2014)

1.2.5 Transcription des corpus oraux en acquisition

Tout corpus oral nécessite par définition une transcription, c'est-à-dire une manière de rendre lisible par un ensemble de codes prédéfinis ce qui est purement audible (le son) et visible (l'image). Or, selon les objectifs d'analyse, les transcriptions renvoient à des niveaux linguistiques différents et n'utilisent pas les mêmes conventions, allant de la transcription orthographique à la transcription phonétique ou phonologique intégrale, et pouvant ajouter l'intonation ou encore la description des gestes, des centres d'attention, de la situation, des objets manipulés, etc.

Se pose par ailleurs la question de la transcription des formes non typiques produites par l'enfant dont la représentation transcrite peut être problématique et sujet à variabilité (Morgenstern & Parisse 2007) comme c'est le cas pour la transcription de formes grammaticales transitoires

dont on ne peut faire correspondre aucune unité orthographique précise (ex. « [e] peux faire »).

Les logiciels utilisés pour transcrire sont eux aussi divers : de la simple transcription à l'alignement texte-son (par exemple avec Transcriber ; Barras *et al.* 2001) ou à l'alignement texte-image-son (par exemple avec CLAN, logiciel intégré à la plateforme de CHILDES ; (MacWhinney 2000), ou encore des utilisations de logiciels centrés sur des niveaux d'analyse spécifiques par exemple PHON (Rose *et al.* 2006) pour la phonétique, Praat (Boersma & Weenink 2009) pour l'intonation, ELAN (Wittenburg *et al.* 2006) pour le visuo-gestuel...

1.2.6 Codage

Le choix de l'unité servant de base à l'analyse est une première étape : délimiter des propositions, des temps de parole, des énoncés ou des séquences indique déjà un objectif de segmentation du discours qui a des implications dans l'analyse qui sera réalisée.

Le codage relève d'un choix arbitraire du chercheur pour identifier des éléments du corpus et ne constitue pas en soi une analyse de ce que l'enfant maîtrise ou est en passe de maîtriser car l'étiquetage de ces unités n'est pas toujours fiable pour le langage en construction de l'enfant. Par exemple, étiqueter des éléments produits par le jeune enfant en noms et en verbes pourrait laisser supposer qu'il maitrise ces catégories. Néanmoins, cela permet parfois de réaliser des comparaisons entre niveaux d'âge (un enfant jeune et un enfant plus âgé) et entre l'enfant et l'adulte.

1.2.7 Outils d'analyse

Plusieurs logiciels permettent de réaliser des analyses linguistiques automatiques et d'extraire un grand nombre d'informations. Par exemple, les outils d'exploitation de corpus comme TXM (Heiden *et al.* 2010) ou Antconc (Anthony 2005) permettent notamment de répertorier dans un ou plusieurs corpus un même élément (ou une chaîne de caractères) et d'afficher les occurrences en contexte, ce qui représente un gain de temps non négligeable et une certaine fiabilité au regard du relevé manuel. Néanmoins, pour les corpus en acquisition, ils présentent des limites puisqu'ils ne peuvent pas extraire les formes en gestation/essai chez l'enfant, c'est-à-dire les lemmes non répertoriés en français parlé

chez l'adulte et non identifiables d'un point de vue orthographique et les formes en émergence non réalisées verbalement, sauf à les connaître par avance ou à les découvrir par hasard. Ils ne peuvent pas non plus à l'heure actuelle cibler les phénomènes interactionnels de manière fine et ne permettent pas toujours aux utilisateurs de revenir aux données sonores ou vidéo.

2 Utiliser les corpus oraux en acquisition : pour quoi faire ?

Si le recours à de « grands » corpus oraux et aux outils informatiques récemment développés pour étudier le langage de jeunes enfants est de plus en plus fréquent, il n'en reste pas moins que les objectifs de recherche peuvent varier, selon que l'on cherche à déterminer des étapes de développement ou un processus d'apprentissage, impliquant une exploitation différente des données.

2.1 Des corpus pour étudier les étapes de développement du langage de l'enfant

Pour relever les processus de développement, typiques ou atypiques, les données peuvent être issues de corpus longitudinaux ou transversaux, recueillis dans des situations d'interactions dyadiques (adulte/enfant ou, dans une moindre mesure, entre pairs) ou polyadiques. Combinés, les deux types de recueil (longitudinal et transversal) permettent de vérifier et généraliser les étapes de développement ou de dégager des propriétés générales grâce notamment à l'analyse quantitative et statistique.

2.1.1 Langage spontané *vs* langage induit

Les termes « spontané », « écologique » ou « naturel » renvoient au recueil de données dans une situation de production non contrôlée, par opposition aux situations expérimentales qui induisent et provoquent des comportements verbaux et non verbaux particuliers. D'autres types de corpus peuvent aussi porter sur des productions de langage induit (ou provoqué), c'est-à-dire qu'ils offrent la possibilité de contrôler ou faire apparaître certains phénomènes, peu fréquents ou dont les conditions d'apparition dépendent de variables spécifiques. En ce sens, les données

recueillies dans ce cadre se rapprochent de celles obtenues en contexte expérimental.

Une distinction nette entre « spontané » et « provoqué » reste néanmoins difficile à établir. En effet, en proposant des activités langagières particulières (lors de jeux symboliques ou de société, de narration, etc.), le chercheur provoque des prises de parole spécifiques en imposant une thématique et en proposant au(x) locuteur(s) de s'inscrire dans un *jeu de langage*. Toutefois, dans ce cas, la forme de leurs productions n'est pas induite et les attentes discursives sont peu contraignantes (jouer avec le matériel ou s'aider du support). On peut désigner ces situations comme relevant du semi-spontané. Elles peuvent cibler une activité monologale (raconter une histoire, un récit d'expériences personnelles) ou dialogale (jeux entre pairs, entre un adulte et un enfant, etc.) permettant de collecter des données comparables (conditions de passation similaires, consignes et supports identiques, etc.).

2.1.2 Étude des productions spontanées : intérêt et apport

Le recueil de données dans des situations spontanées ou semi-spontanées est privilégié par les chercheurs qui ont pour objectif d'accéder aux compétences linguistiques des enfants et à la façon dont ils les mobilisent (Le Normand & Clouard 2014). En outre, il permet de s'intéresser à l'ensemble des niveaux linguistiques si les conditions de recueil le permettent car, dans ce type d'échanges, les participants ont une certaine « liberté créative » (de Weck 2003).

Dans le domaine du développement atypique, l'analyse du langage spontané permet de rendre compte également des stratégies d'évitement des enfants ou des adolescents (Tuller *et al.* 2012) ou d'apprécier l'adaptation et l'évolution de compétences dans le temps qui n'étaient pas observables dans les résultats obtenus à un test de langage (da Silva 2015).

Alors que les données de productions induites visent essentiellement à montrer les compétences acquises ou non acquises, les situations dites spontanées rendent compte des compétences en construction dans le système du locuteur en relevant tout ce que l'enfant réussit à accomplir grâce à l'aide de son interlocuteur ainsi que ses tentatives de productions langagières. Les deux types d'évaluation sont donc complémentaires.

2.1.3 Décrire les étapes du développement langagier

Pour mettre en évidence les étapes d'acquisition de phénomènes linguistiques, plusieurs études s'appuient sur des données spontanées. Par exemple, à partir du corpus CoLaJE, l'étude de Martel & Dodane (2012) met en lien le développement de la prosodie et celui de la syntaxe en analysant les formes mélodiques des productions d'une enfant entre 11 mois et 1 an 11 mois. Sur le même corpus, Sekali (2012) s'est attachée à décrire la mise en place des énoncés complexes entre 10 mois et 4 ans 1 mois en relevant l'émergence et le développement des constructions causatives. Dans le cadre du projet DIAREF [4], l'analyse statistique des productions de 28 enfants entre 1 an 7 mois et 2 ans 6 mois en situation de dialogue en milieu familial montre qu'en plus des facteurs grammaticaux, d'autres facteurs (principalement des facteurs pragmatico-discursifs) interviennent dans le développement des expressions référentielles (da Silva Genest *et al.* à paraître).

D'autres études s'appuient sur des données semi-spontanées pour décrire les compétences d'enfants avec et sans troubles. C'est le cas du projet Subjela (Kugler-Lambert & Préneron 2013), portant sur 235 enfants et visant à comparer le développement typique et atypique des conduites narratives. Les données issues du projet forment le corpus Narracom, constitué de récits monologaux de 235 enfants de 6 à 11 ans (163 enfants tout-venant de 6 à 11 ans, 50 enfants dyslexiques et 22 enfants dyscalculiques de 8 à 12 ans). Dans le champ des troubles du spectre autistique, Foudon *et al.* (2007) ont réalisé une étude longitudinale sur deux ans des productions induites de neuf enfants de 3 ans à 9 ans pour montrer leur évolution lexicale et syntaxique en comparaison avec les normes de développement de l'enfant tout-venant en général.

2.1.4 Établir des mesures du développement langagier

Les épreuves contrôlées et les tests standardisés ont longtemps été considérés comme les méthodes les plus fiables pour mesurer les étapes et les décalages du développement langagier typique et atypique. Le développement d'études s'intéressant aux données spontanées et l'utilisation d'outils de description et d'analyse issus de la linguistique de corpus permettent à l'heure actuelle de valider la pertinence de ce

4. http://www.univ-paris3.fr/anr-diaref-37421.kjsp.

type de données pour l'évaluation des capacités langagières des enfants (Rondal 2003, de Weck & Marro 2010).

Des recherches se sont attachées ces dernières décennies à élaborer de nouvelles méthodes ou à créer et adapter des mesures existantes pour évaluer le langage spontané des enfants avec ou sans troubles du langage. Par exemple, au niveau phonologique, pour mesurer le degré d'intelligibilité de l'enfant notamment, il est maintenant fréquent d'analyser les productions spontanées des jeunes enfants en considérant le pourcentage de mots (PWPC), de syllabes, de phonèmes (PPC), de consonnes (PCC) et de voyelles phonologiquement correctes (PVC) (Parisse & Maillart 2007, da Silva 2015). À un autre niveau linguistique, Le Normand (2007) a proposé de combiner plusieurs mesures déjà existantes : longueur moyenne des énoncés, diversité lexicale (via le programme VOCD[5] par exemple), que l'on peut obtenir avec les commandes automatiques du logiciel CLAN.

Pour évaluer les compétences morphosyntaxiques des productions spontanées des enfants entre 1 an et demi et 4 ans, le LARSP (Language Assesment Remediation and Screening Procedure ; Ball *et al.* 2012) est un outil de référence. L'adaptation française (Parisse *et al.* 2012) a été réalisée à partir d'un corpus transversal regroupant 341 enregistrements d'interactions adulte/enfant afin de mettre en évidence les profils morphosyntaxiques des enfants en fonction de stades de développement.

Les productions spontanées recueillies dans le cadre d'interactions familiales peuvent également servir à la création d'outils de mesure. Ainsi, les items de vocabulaire du questionnaire « Développement du Langage de Production en Français » (Bassano *et al.* 2005) ont été élaborés en partie avec des données issues de la base « Corpus français de productions langagières précoces » (PLPNat ; Bassano 2007) et de CHILDES.

D'autres auteurs proposent également des grilles d'observation issues de l'analyse des données obtenues. Par exemple, l'étude des récits produits par 198 enfants entre 4 et 8 ans à partir du support de la *Frog story* (Mayer 1969) a permis à Hilaire-Debove & Kern (2013) d'élaborer une grille d'évaluation de la macrostructure narrative des récits d'enfants, ou encore l'élaboration d'une grille d'analyse en catégories syntaxiques d'énoncés issue d'un corpus longitudinal portant sur une soixantaine d'enfants âgés de 3 à 6 ans (Canut & Vertalier 2010).

5. Le programme VOCD (McKee *et al.* 2000) permet de calculer la mesure « D » de diversité lexicale, indépendante de la taille des corpus.

2.2 Des corpus pour étudier l'*input*

Les corpus d'interactions permettent également l'analyse du langage entendu par l'enfant, qu'il lui soit adressé ou non (l'*input*), et à partir duquel un certain nombre de chercheurs postulent qu'ils reconstruisent le système de leur langue et apprennent à en faire usage. Si les productions de l'enfant peuvent être étudiées sans faire appel à des corpus, en adoptant une démarche expérimentale, il est impossible de s'intéresser à l'*input* autrement qu'à l'aide de corpus [6] (Cameron-Faulkner *et al.* 2003).

Deux perspectives peuvent être adoptées sur cet *input* : une perspective socio-interactionniste, dans la lignée des travaux de Bruner (1983), où l'analyse porte sur l'imbrication dans la séquentialité du dialogue entre les énoncés de l'adulte et les énoncés de l'enfant ou une perspective fonctionnaliste-constructiviste (Tomasello 2003), où l'analyse porte principalement sur les relations statistiques entre les productions de l'adulte et celles de l'enfant. Ces deux perspectives se rejoignent en ce qu'elles s'inscrivent dans une approche « basée sur l'usage », où l'enfant est acteur de son développement langagier puisqu'il traite statistiquement et pragmatiquement les usages du langage dans l'interaction.

Suivant cette seconde perspective, l'*input* peut renvoyer à plusieurs objets. Les études peuvent ainsi porter sur le langage adressé à l'enfant (dont les caractéristiques et les effets ont fait l'objet d'études importantes dans les années 1970-1980 ; Gallaway & Richards 1994) par des adultes (parents, enseignants, orthophonistes), des enfants plus âgés ou des pairs. Elles peuvent également s'intéresser au langage qui n'est pas directement adressé à l'enfant mais qui constitue pour l'enfant une source d'*input* (Akhtar 2005).

Différentes méthodes peuvent être utilisées pour collecter les données, en fonction des objectifs de la recherche. Si le chercheur souhaite étudier l'*input* que l'enfant reçoit habituellement, il l'enregistrera alors à son domicile par exemple, dans le cadre de ses activités quotidiennes (cf. corpus CoLaJE ou corpus Dîners familiaux ; Morgenstern *et al.* 2015). S'il souhaite étudier l'*input* que peut recevoir l'enfant lors d'une activité (langagière) donnée, il peut alors enregistrer l'interaction entre l'enfant et son/ses interlocuteur/s autour d'un support qu'il aura lui-même déterminé (par exemple un album, une situation de jeu, cf.

6. On peut toutefois manipuler de manière expérimentale les caractéristiques de l'input pour étudier la façon dont l'enfant traite l'*input* (Richards 1994).

corpus TCOF). C'est davantage la question posée que le type d'étude (longitudinale *vs* transversale) qui déterminera la méthodologie adoptée par le chercheur.

2.2.1 Travailler sur l'input : quelques contraintes méthodologiques

Travailler sur l'*input*, et en particulier celui de l'adulte, soulève d'importantes questions méthodologiques du fait de la nécessité éthique d'informer les participants des objectifs de la recherche. L'adulte est en effet non seulement conscient de l'enregistrement mais il a également connaissance des objectifs de la recherche, ce qui peut le conduire (au début de l'enregistrement en tout cas) à surveiller ses productions langagières ou à vouloir faire parler son enfant ou que son enfant « parle bien ». Comme dans la plupart des études sur la langue parlée, ces biais tendent à s'amoindrir si des tâches non orientées vers l'objet d'étude sont proposées par l'observateur, si des activités du quotidien comme préparer le repas doivent être menées ou si l'enregistrement dure suffisamment longtemps. Par ailleurs, si l'observateur est présent lors de l'enregistrement, il est indispensable de travailler à ce que les participants, notamment les adultes, comprennent à la fois la démarche ethnographique du linguiste qui constitue le corpus et le positionnement humble et bienveillant de l'observateur, qu'ils « l'acceptent, et même l'accueillent » (Morgenstern 2016).

Comme pour l'étude des productions des enfants, la question du volume des données et des situations à enregistrer pour avoir un aperçu représentatif de l'*input* des enfants se pose. Ceci d'autant plus qu'on ne s'intéresse aux productions des interlocuteurs de l'enfant que pour le lien qu'elles entretiennent avec celles-ci, qu'on ne transcrit jamais les productions des interlocuteurs de l'enfant sans transcrire celles de l'enfant et que l'étude des productions des interlocuteurs de l'enfant nécessite donc des ressources beaucoup plus importantes. Un système de détection automatique de la parole de l'enfant et de son environnement a été développé pour pallier certaines de ces difficultés. Il s'agit du système LENA™(Language ENvironment Analysis System), qui permet la réalisation, la segmentation et l'analyse automatiques d'un enregistrement audio d'une durée pouvant aller jusqu'à 16h par jour. C'est un système d'enregistrement qui ne nécessite pas la présence de l'observateur et qui est portatif. Les analyses du langage portent sur le nombre de mots de l'adulte, le nombre de vocalisations de l'enfant et le

nombre de tours de parole. Canault *et al.* (2016) ont réalisé une validation pour le français de ce système, initialement conçu pour l'anglais, mais le taux de corrélation entre les résultats obtenus par le système et par une analyse manuelle du nombre de mots produits par l'adulte sur les mêmes données est assez faible. Les perspectives d'évolution techniques (traitement d'autres langues que l'anglais, du signal sonore en contexte de bruit ou d'interactions polyadiques) présagent toutefois d'études riches sur l'*input* dans le développement typique ou atypique du langage.

2.2.2 Études de corrélations entre les productions des adultes et des enfants

Les études sur les corrélations entre les productions des adultes et celles des enfants sont anciennes et ont d'abord principalement porté sur le niveau (morpho)syntaxique (par exemple Furrow *et al.* 1979). Plus récemment, des données appliquées au français, à d'autres champs de la linguistique, à différents types d'interactions et/ou à des corpus plus importants, suggèrent que la question est encore très actuelle. Ainsi, Chabanal *et al.* (2015) étudient le rapport entre la réalisation de la liaison et de l'élision dans le discours adressé et non adressé à l'enfant et l'acquisition de cette variation phonologique par l'enfant à partir d'un corpus dense (corpus ALIPE, 2 fois 5 heures d'enregistrement audio pour chacun des deux enfants de l'étude). Dans le domaine de la référence, Caët & Morgenstern (2015) suggèrent, suite à l'étude d'un sous-corpus de CoLaJE, que la façon dont les enfants font usage de leur prénom pour se désigner ou de « maman » pour désigner leur interlocutrice est liée à la manière dont les parents se désignent et réfèrent à l'enfant lorsqu'ils s'adressent à lui. De la même manière, le rôle de l'*input* est souligné dans diverses études s'intéressant, par exemple, au développement de la négation (Beaupoil-Hourdel 2015), à la valeur illocutoire des énoncés de mères et d'enseignants adressés aux enfants dans des situations de lecture de livres (Vinel 2014) ou à la détermination nominale (Bassano & van Geert 2018).

Les recherches portant sur le développement atypique ou pathologique du langage ne sont pas sans s'intéresser à l'*input* et en particulier au langage adressé à l'enfant, notamment lorsque ces recherches sont menées par des orthophonistes qui tentent de développer la guidance parentale (Maillart *et al.* 2011). Les recherches récentes adressent ces questionnements dans le domaine du non verbal, dans différents types

d'activités langagières ou différents types d'interactions, notamment polyadiques. Ainsi, de Weck et ses collaborateurs comparent l'*input* de mères d'enfant tout-venant et d'enfants dysphasiques en considérant la façon dont les mères emploient des gestes dans un jeu de devinette (de Weck & Marro 2010), du discours représenté (Rezzonico *et al.* 2012), des reformulations du discours des enfants dans diverses activités langagières (Rezzonico *et al.* 2014). Dans le champ de l'autisme, Dascalu (2014) analyse les formes standard et non standard de référence à soi et à l'enfant employées par les mères de deux enfants avec autisme et à leur relation avec l'usage de telles formes par les enfants. Dans le domaine de la surdité, Bessaguet *et al.* (2016) analysent les interactions familiales lors des repas dans des familles avec un enfant sourd pour interroger la participation de ces enfants à ce type d'interactions.

Dans le cadre des études portant sur le développement atypique, il est toutefois fréquent qu'une étude de l'*input* soit associée à une étude du déroulé des interactions, à la façon dont les parents suivent l'attention de leur enfant et à la manière dont ils reprennent et modèlent leurs énoncés en regard des énoncés de leur enfant (Salazar Orvig 2017).

3 Utiliser les corpus oraux pour questionner les processus d'acquisition en interaction

Depuis les travaux précurseurs de Bruner (1983), François *et al.* (1984) et Lentin (1984), les recherches qui visent à mettre au jour des processus d'acquisition tentent de comprendre comment l'enfant s'appuie sur les données qu'il reçoit de son entourage et de caractériser les modalités d'interactions verbales et non verbales favorables au développement des capacités cognitivo-langagières. Elles prennent notamment appui sur les notions de « zone proximale de développement » (Vygotski 1934) et de « schèmes syntaxiques » (Canut & Vertalier 2014).

Pour réaliser ce type d'étude, les modes de recueil, de transcription et d'extraction des données sont semblables à ceux déjà mentionnés. L'analyse qualitative de corpus d'interactions adulte/enfant est privilégiée dans ce cadre, et explore, sans s'y limiter, les dimensions suivantes :

— les reprises immédiates et différées, identiques ou différentes, d'un locuteur à l'autre ;
— les reformulations avec extensions et expansions[7] de l'adulte ;

7. Les expansions sont des reprises de l'énoncé de l'enfant par l'adulte dans un énoncé organisé syntaxiquement. Les extensions sont des reprises de la signification

— toutes les autres formes d'aide à la production verbale, qu'elles soient sous forme interrogative ou assertive ;
— les « essais », c'est-à-dire les tentatives de l'enfant pour réinvestir des éléments ;
— l'évolution dans le temps des productions de l'enfant, des formes émergentes aux formes stabilisées, en fonction des modalités d'étayage proposées par l'adulte ;
— la variabilité individuelle dans la construction du langage de chaque enfant.

En outre, appliquée à des données multiples, l'analyse qualitative peut déboucher sur des généralisations. Dès lors, les bases de données de « grands » corpus représentent une source inestimable, sinon indispensable, pour passer de l'étude du processus d'apprentissage d'un enfant en particulier à celle de l'enfant en général, qu'il présente un développement typique ou atypique.

3.1 Explorer les processus interactionnels d'acquisition

3.1.1 Appropriation de constructions syntaxiques en interaction : un exemple dans les corpus de TCOF

Afin d'illustrer les différentes étapes de l'analyse des processus interactionnels d'acquisition, nous prendrons pour exemple l'étude du discours indirect (DI, par la suite) dans le corpus longitudinal d'une enfant, Léa, entre 3 et 6 ans, issu de la base de données TCOF (Canut 2012), dans une situation de co-narration avec le chercheur.

L'extraction des données à partir du logiciel Jconc[8] a permis de répertorier séparément les productions de DI chez l'adulte et l'enfant et de relever 49 occurrences de DI chez l'enfant et 39 chez l'adulte, et de voir la diversité des formes de DI utilisées par les deux locuteurs (Tableau 1).

En codant les productions de l'enfant par âge (10 enregistrements sur chaque période de 10 à 12 mois), on peut mettre en évidence leur évolution et observer ainsi chez Léa une augmentation sensible des énoncés comportant du DI à partir de 5 ans (par rapport au

supposée de l'énoncé de l'enfant avec ajouts d'éléments nouveaux dans des champs sémantiques proches et avec des précisions lexicales et syntaxiques (Veneziano 2000).

8. Logiciel libre et gratuit développé sous la direction de Christophe Benzitoun (Université de Lorraine & UMR 7118 ATILF).

	Dire que	Dire de	Autres verbes
Enfant (49 occurrences)	89,8 % (44)	10,2 % (5)	0 % (0)
Adulte (39 occurrences)	74,35 % (29)	17,94 % (7)	7,7 % (3)

Tableau 1 – Formes de discours indirect (DI) chez l'enfant et l'adulte

nombre total d'occurrences), avec une production plus importante d'essais (constructions de discours indirect incomplètes, partiellement incompréhensibles ou imbriquées à du discours direct) que de formes stabilisées. Ainsi, l'enfant produit seulement quatre DI et six essais de cette construction avant 5 ans. À 5 ans, le nombre d'occurrences s'accroit avec 15 occurrences et 24 essais. On observe également que le pourcentage de DI chez l'adulte est plus élevé que celui de l'enfant et augmente quasiment dans les mêmes proportions que celui des essais de l'enfant, selon une évolution parallèle (Figure 1).

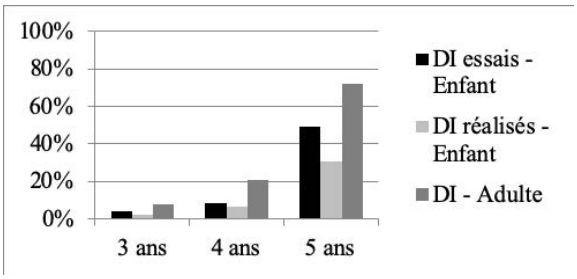

FIGURE *1 – Proportion des constructions de discours indirect (DI) dans les énoncés de l'adulte et de l'enfant à chaque période d'âge*

La comparaison des productions de l'adulte et de l'enfant a pour objectif d'identifier le processus d'apprentissage qui a présidé à l'évolution repérée. Pour ce faire, les occurrences de l'adulte et de l'enfant sont étudiées conjointement, en particulier les types de reprises et de reformulations, comme par exemple les reformulations des essais (Tableau 2).

L'analyse qualitative de l'ensemble des séquences d'interaction impliquant le DI permet d'observer un réinvestissement de la construction par l'enfant à partir des propositions antérieures de l'adulte, à l'image de l'exemple suivant :

L13 – elle dit **que j'ai oublié** ma toutoune

A14 – ah oui Julien **dit qu'il a oublié** sa toutoune

[...]

L31 – après **elle dit qu'**aujourd'hui j(e) vais pas vous ramener dans la classe

Corpus 21 (5 ans 3 mois)

Formes de DI en essai *Pronoms référents (X, Y ou Y') présents mais pronoms référents inappropriés dans le contexte de production*	**Valeur sémantique supposée** *Reformulations de l'adulte avec d'autres pronoms référents et une autre configuration si nécessaire*
réf(X) + "dire que" + réf(Y) (1) E – elle dit que j'allais j'allais + vers mes amis (1') E – il dit que tu vas aller se laver (2) E – elle dit que que que tu te réveilles (2') E – elle dit que tu t'es bien amusé (3) E – là sa maman e(lle) dit que t(u) as un(e) bosse	**réf(X) + "dire que" + réf(X)** (1) A – elle dit qu'elle va aider ses amis (1') A- il dit qu'il va aller se laver **réf(X) + "dire/demander (à Y) de"** **réf(X) + "demander (à Y) si" + réf(Y)** (2) A – elle dit à Ludovic de se réveiller (2') A – elle lui demande s'il ne s'est pas ennuyé **réf(X) + "dire que" + réf(Y')** (3) A – elle dit qu'elle [Magali] a une bosse
réf(X) + "dire à (Y) que" + réf(Y ?) (5) E – il dit à son papa que j'a pas j'a pas *** occupé **réf(X) + "dire à (Y) que" + réf(Y)** (6) E – elle m'a dit que je vais me soigner	**réf(X) + "dire que" + réf(X)** (5) A – le papa de Petit Ours Brun dit qu'il est occupé **Discours direct** (6) Livre- Sophie dit à Jérôme : "je vais me faire soigner"

Tableau 2 – Reformulations des essais de DI dans le corpus de Léa entre 4 et 5 ans

L'adulte reformule immédiatement l'essai L13 de l'enfant portant sur la configuration réf(X) + « dire que » + réf(X) (« *Julien dit qu'il a oublié sa toutoune* »). Un peu plus tard, dans le même corpus, on observe une autre tentative chez l'enfant (en L31) : on peut faire l'hypothèse que la reformulation en A14 a pu être une source de réactivation de ce nouvel essai chez Léa.

3.1.2 Des généralisations possibles

Une étude élargie à l'ensemble de la base TCOF – soit 87 enregistrements différents d'enfants entre 2 et 6 ans en conversation avec un adulte et 10 corpus longitudinaux d'une durée moyenne de 6 mois par enfant en situation de narration dialoguée (13141 énoncés pour les enfants et 13473 pour les adultes) – montre une évolution globale et des procédures d'interaction similaires à celles identifiées chez Léa.

Groupe d'âge	**2 ans**	**3 ans**	**4 ans**	**5 ans**
Moyenne	0,08	0,15	0,29	0,74
N (pondéré) : 206	24	59	65	58
Écart-type	0,28	0,48	0,60	0,98

Tableau 3 – Moyennes d'occurrence du DI et écarts-types selon le groupe d'âge

Si l'on étudie globalement la production du DI chez les adultes qui interagissent avec les enfants, les résultats montrent que le langage adressé à l'enfant est en relation significative avec celui produit par les enfants et que le nombre d'occurrences de DI dans les énoncés des adultes (193 occurrences) est un peu plus élevé que dans les énoncés des enfants (69 occurrences) et suit la même évolution dans les groupes d'âge. Les reprises par l'enfant de verbalisations de DI antérieures de l'adulte (au nombre de 48) constituent près de 70 % de la production des formes de DI chez l'enfant dont près de la moitié sont des reprises identiques et plus de 68 % des reprises différées (Tableau 4).

Ces résultats tendent à conforter l'idée que lorsque l'enfant commence à tester la construction du DI, il s'appuie sur l'organisation syntaxique et lexicale proposée par l'adulte et en reproduit le schème à l'identique avant de l'utiliser de manière plus autonome dans d'autres cotextes. Réciproquement, les propositions et reformulations de l'adulte des essais de l'enfant ont un effet catalyseur, qui offrent à l'enfant la possibilité de faire des hypothèses (non conscientes) sur les différentes configurations du DI.

	Adultes	Enfants
% DI[1]	6,81	4,07
Reprises du DI (%)		69,56
– Reprises identiques		49,27
– Reprises différées		68,11

[1] Les pourcentages sont calculés par rapport au nombre total de constructions complexes (c'est-à-dire les complétives, circonstancielles, relatives, quantitatives et clivées) présentes dans les énoncés de l'adulte et de l'enfant, soit 1692 occurrences chez les enfants et 2834 occurrences chez les adultes (Canut 2012).

Tableau 4 – Pourcentage de DI dans les énoncés de l'adulte et l'enfant et types de reprises

Ces résultats sont le reflet de ceux obtenus pour d'autres constructions (Canut 2014, Diessel 2004) tendant ainsi à confirmer le rôle du langage adressé à l'enfant et soulignant la similarité des procédures d'appropriation chez les enfants (voir Nelson *et al.* 2001, Ervin-Tripp 2005).

4 Apports de l'étude des corpus sur les pratiques et la formation des professionnels

La recherche fondamentale en acquisition du langage ne peut totalement s'abstraire de la réalité des pratiques langagières effectives dans le cadre éducatif (familial, scolaire) et clinique (orthophonie), ce qui constitue le fondement même des sciences sociales (Bronckart 2001) : ne pas se confronter aux corpus *in situ*, c'est se priver d'une source féconde pour la réflexion théorique et méthodologique. Réciproquement, les professionnels se nourrissent des apports de la recherche pour comprendre leurs pratiques et les faire évoluer.

Avoir accès et travailler sur des données recueillies en situation de production réelle présente un intérêt principal, celui de passer d'une approche déductive à une approche empirique. Ces corpus de données attestées contribuent ainsi à une meilleure connaissance des développements typique et atypique et à l'amélioration des prises en charge, notamment orthophoniques, et à la mise en œuvre d'actions éducatives ciblées, en particulier dans le cadre scolaire. Si les professionnels privilégient les

batteries de tests de langage pour situer le niveau d'un enfant par rapport à une norme établie, les compétences ne peuvent pas se réduire aux résultats obtenus par ce biais. En effet, les tests ne tiennent pas compte des différences socioculturelles et des variantes langagières des sujets, des usages de la langue ou des compétences réelles des sujets (*vs* leurs déficits ; de Weck & Marro 2010, Alary & Brun 2018). Les professionnels doivent donc avoir recours à des méthodes complémentaires, telles que celle de l'évaluation du langage en situations dites « spontanées », pour considérer l'ensemble des compétences d'un locuteur.

Par ailleurs, si les professionnels disposent d'outils leur permettant d'évaluer les performances langagières d'un sujet, ces outils ne mettent pas en relation l'évolution des capacités de l'enfant et le type de stimulation dont il bénéficie, et ne disent rien sur les modalités à adopter pour stimuler l'apprentissage ou y remédier. Le recours aux corpus de langage spontané peut en revanche être tout à fait approprié pour amener les professionnels à orienter leurs pratiques (da Silva Genest & Masson 2017). L'observation de corpus oraux leur permet de mieux cerner les caractéristiques mêmes du langage oral, bien souvent méconnu et dévalorisé par la référence à la norme de l'écrit. En parallèle, elle les conduit à prendre conscience des modalités d'interaction proposées et à déterminer celles qui sont les plus efficientes pour aider les enfants à apprendre à parler. Dans le cadre de recherches-actions-formations, l'enregistrement et la transcription par les professionnels sont même prônés car ils constituent la base d'une possible transformation des pratiques, par la mesure des écarts entre les objectifs visés et les pratiques langagières effectivement réalisées (Canut *et al.* 2013). Les études mentionnées précédemment, centrées sur la façon dont les enfants s'approprient les éléments du langage qui leur est adressé, permettent d'apprécier à la fois la nature des modalités de l'étayage et son efficience. Dès lors, il est possible de proposer aux professionnels des conduites langagières au plus près des zones proches de développement de l'enfant et de les aider à repérer une évolution dans le langage de chaque enfant, allant des formes émergentes vers des formes stabilisées, en lien avec leurs propres productions (Canut *et al.* 2018).

5 Conclusion

L'évolution actuelle des recherches en acquisition liée à l'essor de la linguistique de corpus et leur apport possible pour la formation

des professionnels engagent à poursuivre le partage de données et à mettre à disposition de nouveaux corpus. Si les bases CHILDES et TCOF sont régulièrement alimentées, les besoins sont encore nombreux : corpus avec des pathologies variées, corpus d'enfants tout-venant dans des situations d'interaction et avec des interlocuteurs plus diversifiés, corpus plus étendus et plus denses (Liégeois 2014) pour pouvoir étudier des phénomènes rares, etc. La tâche est encore vaste mais pleine de promesses.

Nom	Âge des enfants	Nb	Dvpt typique/ atypique	Longitudinal/ transversal	Activité	Interaction/ monologue	Interlocuteur principal de l'enfant	Formats trans-crip-tions	Formats média	Base de données
CoLaJE/ Paris Corpus	0–7 ans	7	typique	longitudinal	conversation	interaction	parent	.cha, .phon, .tei, .xml	vidéo .mov, .mp4	Ortolang CHILDES
TCOF enfants	3–7 ans	24	typique	longitudinal	narration et conversation	interaction	adulte non parent	.trs, .xml	au-dio.wav	Ortolang
	1–10 ans	170	typique	transversal	narration et conversation	interaction	adulte non parent	.trs, .xml	audio .wav	
	6–10 ans	64	typique	transversal	conversation	interaction	adulte non parent	.trs, .xml	audio .wav	
PLPnat	1–4 ans	3	typique	longitudinal	conversation	interaction	parent	.doc, .rtf	–	Ortolang CHILDES (corpus de Pauline)
	5–8 ans	16	typique et atypique (TSA)	transversal	non renseigné	non renseigné	non renseigné	.docx (pour 4 enfants ty-piques)	–	Ortolang
Narracom	6–11 ans	163	typique	transversal	narration	monologue	expérimenta-teur	.csv	audio .ogg	Ortolang
ALIPE	2;04–5;04	3	typique	longitudinal	conversation	interaction	parent	.cha, .xml	audio .mp3	Ortolang
Champaud	1;09–2;05	1	typique	longitudinal	conversation	interaction	parent	.cha	–	CHILDES
Geneva	1;08–2;06	1	typique	longitudinal	conversation	interaction	parent	.cha	audio .mp3	CHILDES

GoadRose	1–4 ans	2	typique	longitudinal	conversation	interaction	parent	.cha, .phon	audio .mp3	CHILDES
Hammelrath	3;06–5;06	258	typique et typique (retard de langage)	transversal	conversation	interaction	clinicien	.cha	–	CHILDES
Hunkeler	1;06-2;06	2	typique	longitudinal	conversation	interaction	parent	.cha	vidéo .mp4	CHILDES
Kern French	0;07–2;01	4	typique	longitudinal	conversation	interaction	parent	.cha, .phon	audio .mp3	CHILDES
Leveillé	2;01–3;03	1	typique	longitudinal	conversation	interaction	parent	.cha	–	CHILDES
Lyon	1–3 ans	5	typique	longitudinal	conversation	interaction	parent	.cha, .phon	vidéo .mp4	CHILDES
MTLN	2–4 ans	56	typique	transversal	description et conversation	monologue	–	.cha	–	CHILDES
Palasis	2;05–4;00	22	typique	longitudinal et transversal	conversation et narration	interaction	adulte non parent	.cha	vidéo et audio .mp4, .mp3	CHILDES
Stanford French	0;09–1;07	6	typique	longitudinal	non précisé	monologue	–	.cha, .phon	audio .mp3	CHILDES
Vion Colas	7, 9 et 11 ans	191	typique	transversal	narration	monologue	–	.cha	–	CHILDES
York	1;09–4;03	3	typique	longitudinal	conversation	interaction	parent	.cha	–	CHILDES
French-MTLN	4–10 ans	14	typique	transversal	narration	interaction	expérimenta-teur	.cha	audio .mp3	CHILDES
French-Lyon	5,6,7,10 et 21 ans	40	typique	transversal	narration	interaction	expérimenta-teur et parent	.cha	audio .mp3	CHILDES

Foudon-Reboul	3;09–9;02	9	atypique (TSA)	longitudinal	conversation	interaction	éducateur	.cha	–	CHILDES
Le Normand	4–5 ans	7	atypique (épilepsie et troubles du dvpt langagier)	transversal	description et conversation	monologue	–	.cha	audio .mp3	CHILDES
Nadig	3–6 ans	26	atypique (TSA)	longitudinal	conversation	interaction	parent	.cha	–	CHILDES

Bibliographie

Akhtar, N. (2005). The robustness of learning through overhearing. *Developmental Science*, *8*(2), 199–209.

Alary, L., & Brun, G. (2018). Une approche située des atypies langagières : éléments pour un dialogue entre l'orthophonie et la philosophie. Dans C. Bogliotti, F. Isel, & A. Lacheret-Dujour (Eds.), *Atypies langagières de l'enfance à l'âge adulte : Apports de la psycholinguistique et des neurosciences cognitives*, Louvain-la-Neuve : De Boeck, (p. 5–26).

André, V., & Canut, E. (2010). Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement des Corpus Oraux en Français). *Pratiques*, *147–148*, 35–51.

Anthony, L. (2005). AntConc : design and development of a freeware corpus analysis toolkit for the technical writing classroom. Dans *IPCC 2005. Proceedings International Professional Communication Conference*, (p. 729–737).

Ball, M.J., Crystal, D., & Fletcher, P.E. (2012). *Assessing grammar – The Languages of LARSP*. Bristol : Multilingual Matters.

Bang, J., & Nadig, A. (2015). Learning language in autism : Maternal linguistic input contributes to later vocabulary. *Autism Research*, *8*(2), 214–223.

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber : Development and use of a tool for assisting speech corpora production. *Speech Communication*, *33*(1), 5–22.

Bassano, D. (2007). Émergence et développement du langage : enjeux et apports des nouvelles approches fonctionnalistes. Dans E. Dumont & M.N. Metz-Lutz (Eds.), *L'Acquisition du langage et ses troubles*, Marseille : SOLAL Éditeurs, (p. 13–46).

Bassano, D., Labrell, F., Champaud, C., Lemétayer, F., & Bonnet, P. (2005). Le DLPF : un nouvel outil pour l'évaluation du développement du langage de production en français. *Enfance*, *57*(2), 171–208.

Bassano, D., & van Geert, P. (2018). New perspectives on input-output dynamics : Example from the emergence of the noun category. Dans M. Hickmann, E. Veneziano, & H. Jisa (Eds.), *Sources of Variation in First Language Acquisition*, Amsterdam : John Benjamins, (p. 201–218).

Baude, O. (2006). *Corpus oraux : guide des bonnes pratiques*. Orléans : Presses Universitaires d'Orléans.

Beaupoil-Hourdel, P. (2015). *Acquisition et expression multimodale de la négation. Étude d'un corpus vidéo et longitudinal de dyades mère-enfant francophone et anglophone*. Thèse de doctorat, Université Paris 3 – Sorbonne Nouvelle.

Behrens, H. (2008). *Corpora in language Acquisition Research. History, methods, perspectives*. Amsterdam : Benjamins.

Bessaguet, S., Gorry, M., & Caët, S. (2016). L'enfant sourd en situation d'interactions polyadiques : accès au langage adressé et non adressé au cours de repas familiaux. Dans *Colloque de logopédie « De l'interagir à l'intervenir : quelles clés ? »*, Neuchâtel.

Blanche-Benveniste, C., & Jeanjean, C. (1987). *Le Français parlé : transcription et édition*. Paris : Didier Érudition.

Bloom, L. (1970). *Language Development : Form and function in emerging grammars*. Cambridge, Mass. : MIT Press.

Boersma, P., & Weenink, D. (2009). Praat : Doing phonetics by computer (Version 5.1 05). [Computer program]. Retrieved May 1, 2009.

Braine, M.D.S. (1963). The ontogeny of English phrase structure : The first phase. *Language*, *39*(1), 1–13.

Bronckart, J.P. (2001). S'entendre pour agir et agir pour s'entendre. Dans J.M. Baudoin & J. Friedrich (Eds.), *Théories de l'action et éducation*, Louvain-la-Neuve : De Boeck, (p. 133–153).

Brown, R. (1973). *A First Language : The early stages*. Cambridge, Mass. : Harvard University Press.

Bruner, J.S. (1983). *Comment les enfants apprennent à parler*. Paris : Retz.

Caët, S., & Morgenstern, A. (2015). First and second person pronouns in two mother-child dyads. Dans *The Pragmatics of Personal Pronouns*, Amsterdam : Benjamins, (p. 173–193).

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*(6), 843–873.

Canault, M., Le Normand, M.T., Foudil, S., Loundon, N., & Thai-Van, H. (2016). Reliability of the Language ENvironment Analysis system (LENA™) in European French. *Behavior Research Methods*, *48*(3), 1109–1124.

Canut, E. (2012). *L'Acquisition des constructions syntaxiques complexes chez l'enfant. Un processus interactionnel*. Habilitation à diriger des recherches non publiée, Université Stendhal – Grenoble 3.

Canut, E. (2014). Acquisition des constructions syntaxiques complexes chez l'enfant français entre 2 et 6 ans. *SHS Web of Conferences*, *8*, 1437–1452.

Canut, E., Espinosa, N., & Vertalier, M. (2013). Corpus et prise de conscience des processus interactionnels d'apprentissage du langage. *LINX*, *68–69*, 69–93.

Canut, E., Masson, C., & Leroy-Collombel, M. (2018). *Accompagner l'enfant dans son apprentissage du langage. De la recherche en acquisition à l'intervention des professionnels*. Paris : Hachette Éducation.

Canut, E., & Vertalier, M. (2008). Des données représentatives... de quoi en acquisition du langage ? Constitution de données à observer et objectifs d'analyse. *Verbum*, *30*(4), 299–312.

Canut, E., & Vertalier, M. (2010). Étudier la complexité syntaxique chez l'enfant de moins de six ans dans une perspective interactionnelle : choix d'une méthodologie qualitative. Dans J. Bernicot, E. Veneziano, M. Musiol, & A. Bert-Erboul (Eds.), *Interactions verbales et acquisition du langage*, Paris : L'Harmattan, (p. 239–260).

Canut, E., & Vertalier, M. (2014). Les schèmes sémantico-syntaxiques dans le processus interactionnel d'acquisition. Dans N. Espinosa, E. Canut, & M. Vertalier (Eds.), *Linguistique de l'acquisition du langage oral et écrit. Convergences entre les travaux fondateurs de Laurence Lentin et les problématiques actuelles*, Paris : L'Harmattan, (p. 85–116).

Chabanal, D., Liégeois, L., & Chanier, T. (2015). Acquisition de la variation phonologique et recueil de corpus d'interactions naturelles parents-enfants : nouvelle méthode, nouveaux enjeux. *Lidil*, *51*, 65–88.

Cohen, M. (1925). *Mélanges linguistiques offerts à M. J. Vendryes par ses amis et ses élèves*, chap. Sur les langages successifs de l'enfant. Paris : Champion, (p. 109–127).

da Silva, C. (2015). Compétences pragmatico-discursives des enfants dysphasiques en situation de jeu symbolique : évaluation et évolution. Dans *Colloque "Narration et interaction"*, Université Paris Descartes.

da Silva Genest, C., Marcos, H., Salazar Orvig, A., Caët, S., & Heurdier, J. (à paraître). The choice of referring expressions : early development. Dans A. Salazar Orvig, G. de Weck, R. Hassan, & A. Rialland (Eds.), *The Acquisition of Referring Expressions : A dialogic approach*, Amsterdam : Benjamins.

da Silva Genest, C., & Masson, C. (2017). Apport de la linguistique de corpus à l'étude des situations cliniques : utilisation de ressources écologiques pour évaluer les pratiques professionnelles. *Studii de lingvistică*, *7*, 89–112.

Darwin, C. (1877). I.– A Biographical Sketch of an Infant. *Mind*, *os-2*(7), 285–294.

Dascalu, C. (2014). *La Référence à soi chez les enfants atteints d'autisme. Perspectives sémantiques, pragmatiques et cognitives*. Thèse de doctorat, Université de la Sorbonne Nouvelle-Paris 3.

de Weck, G. (2003). Pratiques langagières, contextes d'interaction et genres de discours en logopédie/orthophonie. *TRANEL*, *38–39*, 25–48.

de Weck, G., & Marro, P. (2010). *Les Troubles du langage chez l'enfant. Description et évaluation*. Issy-les-Moulineaux : Elsevier Masson.

Demuth, K. (2008). Exploiting corpora for language acquisition research. Dans *Corpora in language Acquisition Research. History, methods, perspectives*, Amsterdam : Benjamins.

Diessel, H. (2004). *The Acquisition of Complex Sentences*. Cambridge : Cambridge University Press.

Dromi, E. (1988). *Early Lexical Development*. Cambridge : Cambridge University Press.

Dugua, C., Nardy, A., Liégeois, L., Chevrot, J.P., & Chabanal, D. (2017). L'acquisition des liaisons après les clitiques préverbaux est-elle spécifique ? Apport d'une expérimentation à grande échelle. *Journal of French Language Studies*, *27*(1), 73–86.

Ervin-Tripp, S. (2005). Impact du cadre interactionnel sur les acquisitions en syntaxe. *AILE*, *4*, 55–80. http://journals.openedition.org/aile/1253.

Foudon, N., Reboul, A., & Manificat, S. (2007). Language acquisition in autistic children : A longitudinal study. Dans *CamLing2007 : University of Cambridge Postgraduate Conference in Language Research*.

François, F., Hudelot, C., & Sabeau-Jouannet, E. (1984). *Conduites linguistiques chez le jeune enfant*. Paris : PUF.

Furrow, D., Nelson, K., & Benedict, H. (1979). Mothers' speech to children and syntactic development : some simple relationships. *Journal of Child Language*, *6*(3), 423–442.

Gallaway, C., & Richards, B.J. (1994). *Input and Interaction in Language Acquisition*. Cambridge : Cambridge University Press.

Heiden, S., Magué, J.P., & Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. Dans *10th International Conference on the Statistical Analysis of Textual-Data – JADT 2010*, t. 2, (p. 1021–1032).

Hilaire-Debove, G., & Kern, S. (2013). Évaluation & développement de la macrostructure du récit oral chez les enfants avec ou sans troubles du langage. *ANAE*, *124*, 306–315.

Ingram, D. (1989). *First Language Acquisition. Method, Description and Explanation*. Cambridge : Cambridge University Press.

Koester, A. (2010). Building small specialized corpora. Dans A. O'Keeffe & M. McCarthy (Eds.), *Routledge Handbook of Corpus Linguistics*, Londres : Routledge, (p. 66–79).

Kugler-Lambert, M., & Préneron, C. (2013). La syntaxe évaluative, source de cohérence : une interrogation à partir de récits d'enfants en difficulté d'apprentissage du langage écrit *vs* des mathématiques. *ANAE*, *124*, 316–324.

Le Normand, M. (1997). Early morphological development in French children. Dans A.S. Olofsson & S. Strömqvist (Eds.), *Cross-Linguistic Studies of Dyslexia and Early Language Development*, Luxembourg : Office for Official Publication of the European Communities, (p. 59–79).

Le Normand, M., & Clouard, C. (2014). De nouveaux outils pour l'évaluation de la parole, du langage et de la communication chez le jeune enfant. *Contraste*, *39*(1), 161–180.

Le Normand, M.T. (2007). Évaluation de la production spontanée du langage oral et de l'activité sémantique du récit chez l'enfant d'âge préscolaire. *Rééducation Orthophonique*, *231*, 53–72.

Lentin, L. (1984). *Recherches sur l'acquisition du langage*, t. 1. Paris : Presses de la Sorbonne Nouvelle.

Liégeois, L. (2014). *Usage des variables phonologiques dans un corpus d'interactions naturelles parents-enfant : impact du bain linguistique et dispositifs cognitifs d'apprentissage*. Thèse de doctorat, Université de Clermont-Ferrand 2.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances : A usage-based analysis. *Cognitive Linguistics*, *20*(3), 481–507.

MacWhinney, B. (2000). *The CHILDES Project : Tools for analyzing talk*. Mahwah : Lawrence Erlbaum.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*(2), 271–295.

Maillart, C., Leroy, S., Quintin, E., Ranc, L., Derouaux, F., D'Harcour, E., Al Mounajjed, M., Caët, S., Leroy-Collombel, M., & Morgenstern, A. (2011). Des interactions enrichies qui soutiennent le développement du langage : effets à court et moyen terme (6 mois) d'une guidance parentale logopédique. *ANAE*, *112–113*, 223–230.

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

Mayer, M. (1969). *Frog, Where Are You ?* New York : Dial Press.

McCarthy, D. (1946). Le développement du langage chez l'enfant. Dans L. Carmichael (Ed.), *Manuel de psychologie de l'enfant*, t. 2, Paris : PUF, (p. 751–916). 1952.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Digital Scholarship in the Humanities*, *15*(3), 323–338.

Morgenstern, A. (2016). Pratiques langagières et comportements du patient en milieu familial : apport des méthodes ethnographiques multimodales pour la recherche en médecine. *Ethics, Medicine and Public Health*, *2*(4), 641–649.

Morgenstern, A., Debras, C., Beaupoil-Hourdel, P., Le Mené, M., Caët, S., & Kremer-Sadlik, T. (2015). L'art de l'artichaut et autres rituels : transmission de pratiques sociales et alimentaires dans les dîners familiaux parisiens. *Anthropology of food*, *9*, en ligne. http://journals.openedition.org/aof/7836.

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

Morgenstern, A., & Parisse, C. (2012). The Paris Corpus. *Journal of French Language Studies*, *22*(1), 7–12.

Nelson, K.E., Welsh, J.M., Camarata, S.M., Tjus, T., & Heimann, M. (2001). A rare event transactional model of tricky mix conditions contributing to language acquisition and varied communicative delays. Dans K.E. Nelson, A. Aksu-Koç, & C.E. Johnson (Eds.), *Children's Language*, t. 11, Hillsdale : Lawrence Erlbaum Associates, (p. 165–195).

Parisse, C. (2014). Événements langagiers rares et acquisition du langage. *SHS Web of Conferences*, *8*, 1551–1562.

Parisse, C., & Maillart, C. (2007). Phonology and syntax in French children with SLI : A longitudinal study. *Clinical Linguistics & Phonetics*, *21*(11–12), 945–951.

Parisse, C., Maillart, C., & Tommerdahl, J. (2012). 13 F-LARSP : A Computerized Tool for Measuring Morphosyntactic Abilities in French. *Assessing Grammar : The languages of LARSP*, *7*, 230–244.

Preyer, W. (1887). *L'Âme de l'enfant*. Paris : Alcan. Titre original *Kinderseele*, 1882.

Rezzonico, S., Da Silva, C., Corlateanu, C., de Weck, G., & Salazar Orvig, A. (2012). Le discours représenté fictif dans des dialogues mère-enfant dysphasique et mère-enfant tout-venant. *ORALIA "Polifonía e intertextualidad en el diálogo"*, *6*, 199–214.

Rezzonico, S., de Weck, G., Salazar Orvig, A., da Silva Genest, C., & Rahmati, S. (2014). Maternal recasts and activity variations : A comparison of mother–child dyads involving children with and without SLI. *Clinical Linguistics & Phonetics*, *28*(4), 223–240.

Richards, B.J. (1994). Child-directed speech and influences on language acquisition : methodology and interpretation. Dans C. Gallaway & B.J. Richards (Eds.), *Input and Interaction in Language Acquisition*, Cambridge : Cambridge University Press, (p. 74–106).

Rondal, J.A. (2003). *L'Évaluation du langage*. Sprimont : Mardaga.

Rose, Y., MacWhinney, B., Byrne, R., Hedlund, G., Maddocks, K., O'Brien, P., & Wareham, T. (2006). Introducing Phon : A software solution for the study of phonological acquisition. Dans D. Bamman, T. Magnitskaia, & C. Zaller (Eds.), *Proceedings of the 30th Annual Boston University Conference on Language Development*, Somerville : Cascadilla Press, (p. 489–500).

Rowland, C.F., Fletcher, S.L., & Freudenthal, D. (2008). How big is big enough ? Assessing the reliability of data from naturalistic samples. Dans H. Behrens (Ed.), *Corpora in Language Acquisition Research. History, methods, perspectives*, Amsterdam : Benjamins, (p. 1–24).

Salazar Orvig, A. (2017). Dialogue et interaction au cœur de la réflexion sur l'acquisition du langage. *Revue Tranel*, *66*, 5–27.

Sekali, M. (2012). The emergence of complex sentences in a French child's language from 0 ;10 to 4 ;01 : causal adverbial clauses and the concertina effect. *Journal of French Language Studies*, *22*(1), 115–141.

Stern, C., & Stern, W. (1907). *Die Kindersprache*. Leipzig : Quelle Meyer.

Taine, H. (1876). Note sur l'acquisition du langage chez les enfants et dans l'espèce humaine. *Revue Philosophique de la France et de l'Étranger*, *1*, 5–23.

Tomasello, M. (2003). *Constructing a Language. A usage-based theory of language acquisition*. Cambridge : Harvard University Press.

Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech : how much is enough ? *Journal of Child Language*, *31*(1), 101–121.

Tuller, L., Henry, C., Sizaret, E., & Barthez, M.A. (2012). Specific language impairment at adolescence : Avoiding complexity. *Applied Psycholinguistics*, *33*(1), 161–184.

Veneziano, E. (2000). Interaction, conversation et acquisition du langage dans les trois premières années. Dans M. Kail & M. Fayol (Eds.), *L'Acquisition du langage. Le langage en émergence de la naissance à trois ans*, Paris : PUF, (p. 231–265).

Vinel, E. (2014). Comparaison des conduites discursives de mères et d'enseignants dans la co-construction de récits avec des enfants. *SHS Web of Conferences*, *8*, 1607–1625.

Vygotski, L.S. (1934). *Pensée et Langage*. Paris : La Dispute. 1997.

Weir, R. (1962). *Language in the Crib*. The Hague : Mouton.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN : A Professional Framework for Multimodality Research. Dans *Proceedings for the Fifth International Conference on Language Ressources and Evaluation*, (p. 1556–1559).