Outils et méthodes de recherche en acquisition du langage : de la complémentarité entre statistiques et analyse linguistique

Caroline Rossi, Aliyah Morgenstern

Outils et méthodes de recherche en acquisition du langage : de la complémentarité entre statistiques et analyse linguistique

Caroline Rossi[1], Aliyah Morgenstern[2]

[1]DDL (UMR 5594) – Université Lyon 2 – 14 avenue Berthelot
69363 Lyon Cedex 07 – France

[2]ENS-LSH – ICAR – 15, Parvis René Descartes – 69 367 Lyon cedex 07 – France

Abstract

Although somewhat neglected by the first developmental psycholinguists, the methodology of compilation, transcription, and presentation of data on child language development has been a topic of discussion at least from the late seventies onwards, since many variables are introduced in each stage of this process. They have become a new focus of attention since 1984, thanks to the CHILDES (Child Language Data Exchange System) project[1]. The ensuing use of similar tools and basic conventions (the CHAT format) by most researchers has certainly contributed to making transcriptions a more reliable "text", as well as one that can be shared and used for different purposes. The CLAN programs have also eased coding, morpho-syntactic analyses, frequency counts, etc.

But at the same time, interfering variables have not been eliminated altogether, and generalizations have become more and more difficult with the increasing amount and variety of data used (e.g. in cross-linguistic comparisons). We propose that one way of addressing these issues is to have statistics go hand in hand with detailed, linguistic analyses. In order to illustrate this view, we describe and discuss the creation and use of an enriched transcription format within the *Léonard project*[2]. We then provide one example of semantic coding (spatial vs. functional prepositions) and discuss the impact of categorization on statistical counts.

Résumé

Même si les premiers travaux de psycholinguistique développementale n'y prêtaient que peu d'attention, les méthodes de compilation, de transcription et de présentation des données ont été très discutées depuis la fin des années soixante dix, en raison des nombreuses variables intervenant à chaque étape du processus. Avec le lancement du projet CHILDES en 1984 ces méthodes ont été revues pour permettre l'adoption d'un format unique (CHAT), faisant des transcriptions un « texte » plus fiable, mais aussi réutilisable et partageable. Les programmes de CLAN ont aussi permis d'automatiser les procédures de codage, d'analyse morpho syntaxique, et les calculs de fréquence.

Mais ce faisant les observables sont devenus plus complexes, et les généralisations plus ardues, du fait même de la pluralité des corpus accessibles. Nous proposons ici une approche des données basant les statistiques sur des analyses linguistiques minutieuses. Deux types d'illustrations en seront données. Nous présenterons et discuterons d'abord la création, dans le cadre du *projet Léonard*, d'un format de transcription enrichi permettant de condenser dans le « texte » le maximum d'informations tout en réduisant la part des choix, toujours subjectifs Nous montrerons ensuite l'impact des catégorisations utilisées ou produites par le chercheur, à travers la présentation d'une procédure de codage sémantique des premières prépositions.

Mots-clés : acquisition du langage, transcription, analyses quantitatives et qualitatives, codage.

[1] Child Language Data Exchange System : il s'agit d'un système d'échange de données en acquisition du langage. Cf http://childes.psy.cmu.edu/intro/

[2] http://anr-leonard.ens-lsh.fr

1. Introduction

De nos jours, les chercheurs qui travaillent sur des suivis longitudinaux d'enfants en milieu naturel peuvent filmer les enfants chez eux. Les enregistrements sont ensuite transcrits et informatisés. Des logiciels spécialisés offrent la possibilité d'aligner transcriptions et enregistrements. Mais avant de pouvoir s'appuyer sur toutes ces technologies, les premiers observateurs du langage enfantin avaient adopté une autre méthode de recueil des données : la méthode du journal.

Les journaux des scientifiques de la fin du dix-neuvième siècle et du début du vingtième siècle avaient l'extraordinaire avantage de permettre au chercheur, le plus souvent l'un des parents de l'enfant, de faire des descriptions quotidiennes. Ces premières observations contiennent à la fois des faits de langues notés au jour le jour, des questionnements, des hypothèses sur la naissance du langage chez l'enfant ancrée dans le contexte de son développement général, sur l'émergence des premiers mots, des premiers outils grammaticaux, des premières constructions, mais aussi des remises en question des catégories grammaticales adultes et des idées préconstruites sur le langage.

Cette méthode d'investigation des premiers chercheurs en acquisition du langage était déjà proche de ce que l'on nomme aujourd'hui « linguistique de corpus ». À ceci près que, depuis l'avènement des nouvelles technologies, le travail sur corpus se pratique sur des enregistrements vidéo alignés avec leurs transcriptions. Mais il s'agit également pour les chercheurs qui travaillent sur le langage en milieu familial de traiter des données « naturalistes » : le chercheur part des productions de l'enfant en dialogue et n'a pas à décider s'il les accepte - elles existent. Cependant, tout corpus est une construction, au sens où il est toujours déjà le produit des analyses du chercheur (Ochs, 1979). Ainsi les procédures de transcription puis de codage des données méritent une attention toute particulière, si l'on veut éviter de dénaturer les productions.

Même si les premiers travaux de psycholinguistique développementale ont pour la plupart ignoré les problèmes posés par la construction des données, les méthodes de compilation, de transcription et de présentation des données ont été très discutées depuis la fin des années soixante dix, en raison des nombreuses variables intervenant à chaque étape du processus. En particulier, faut-il opérer une sélection, et si oui selon quels critères ? À quelles conditions les informations ajoutées, et notamment celles qui ne concernent pas les productions verbales stricto sensu, sont-elles un enrichissement plutôt qu'une gêne ? Nous discuterons les réponses apportées à de telles questions en présentant la principale initiative de normalisation[3] et de partage des données dans la discipline, CHILDES, qui a permis de définir un format standard ; puis nous proposerons deux types d'enrichissement des transcriptions, avant d'aborder la question du codage. La question de la normalisation des données est particulièrement problématique, puisque nous souhaitons à la fois pouvoir partager les transcriptions et les méta-données avec la communauté et suivre ce principe fondateur de tout travail sur corpus en acquisition du langage, qui veut que l'on porte un regard non normatif sur les productions spontanées des enfants (Morgenstern, 2006).

[3] Au sens où une communauté de chercheurs adopte les mêmes pratiques, formats et conventions. Il ne s'agit évidemment pas d'imposer au langage de l'enfant les normes de la langue adulte.

2. Une méthode pour tous : normalisation et partage des données

Le premier problème que posent les transcriptions, c'est qu'elles portent chacune des conventions spécifiques, puisque chaque chercheur (ou groupe de recherche) développe ses propres méthodes de travail. La question de la fiabilité de telles données s'est posée dès que l'on a commencé à travailler sur des transcriptions tapées à la machine et à les partager (MacWhinney, 1995). D'autre part, l'échantillonnage (fréquence des enregistrements pour des suivis longitudinaux, par exemple) et la quantité de données permettant d'observer des tendances développementales, sont des paramètres dont la prise en compte est cruciale pour le chercheur en acquisition du langage, puisqu'il ne peut pas se fier à ses intuitions de locuteur natif pour juger des productions de l'enfant (Bloom, 1991). L'objectif premier du projet CHILDES, lancé en en 1984[4] sous la direction de Brian MacWhinney et de Catherine Snow, était justement la mise en partage d'une base de donnée informatisée de transcriptions, mais aucune normalisation n'était prévue. C'est seulement avec le nombre croissant des données ajoutées que le besoin d'une méthode standardisée, permettant la mutualisation des données s'est fait jour. Les trois composantes actuelles de CHILDES ont été développées pour répondre au mieux à ces exigences :

1 - Une banque de données.

2 - CHAT, un format de transcription et de codage qui permet d'informatiser le corpus.

3 - CLAN, une série de programmes d'ordinateur pour traiter et analyser les données.

Les méthodes de transcription ont été revues pour permettre l'adoption d'un format unique (CHAT), faisant des transcriptions un « texte » plus fiable, mais aussi réutilisable et partageable sous la forme d'une base de données internationale. D'autre part, les programmes de CLAN ont permis d'automatiser les procédures de codage, d'analyse morpho-syntaxique, et les calculs de fréquence d'occurrence.

Dans le même temps, les progrès technologiques ont fait des transcriptions un outil multimédia, lié non seulement au son mais aussi à l'image. L'enrichissement est considérable, puisque avec l'accès ainsi gagné à l'ensemble (ou presque) de l'interaction lors de la réalisation puis de l'usage de la transcription de nombreux problèmes peuvent être résolus (levée d'ambiguïtés, vérifications systématiques par exemple). De nouvelles questions de recherche ont aussi été ouvertes ou réouvertes, imposant leurs contraintes propres, à commencer par la nécessité d'enrichir les transcriptions : ainsi du récent regain d'intérêt pour le posturo-mimo-gestuel, dont nous discuterons à partir de l'exemple du pointage (3.1).

Ci-dessous, un exemple à partir du corpus Léonard (transcription Morgenstern 1994, format CHAT et alignement réalisés par Christophe Parisse).

[4] Child Language Data Exchange System : il s'agit d'un système d'échange de données en acquisition du langage. Cf. http://childes.psy.cmu.edu/intro/

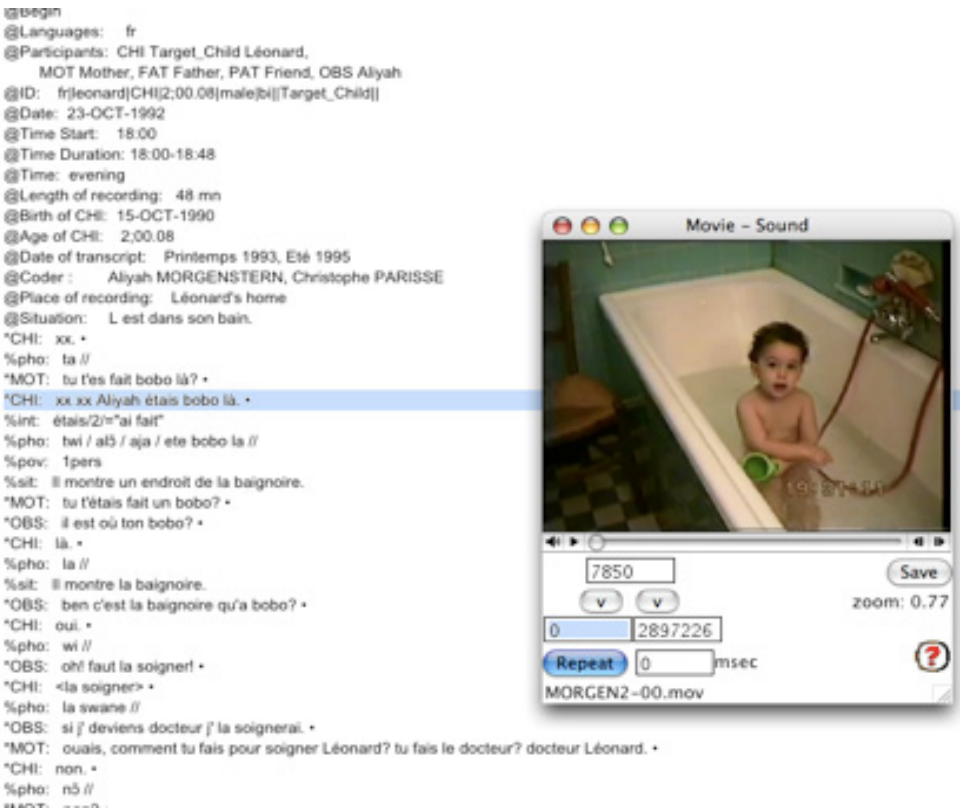

Figure 1 : Une page de transcription du corpus Léonard, alignée avec la vidéo.

Plus d'une centaine de groupes de chercheurs du monde entier travaillent avec ces outils. Ils transcrivent des entretiens sous le format CHAT ce qui enrichit la banque de données, et les analysent grâce aux différents programmes de CLAN. Les propositions et analyses que nous discutons ici (points 3 et 4) sont autant d'illustrations des apports du système, qui permet de mener des études à la fois minutieuses et à grande échelle, mais aussi de ses limites.

Il faut ajouter à cet aperçu que bien peu des outils décrits sont performants pour les besoins de la recherche sur le développement phonologique. PHON[5] est un nouveau logiciel conçu par Yvan Rose (Rose et al. 2006), Professeur à Memorial University (Canada), pour combler ce manque. Il est doté de fonctionnalités permettant d'aligner des données multimédia avec une transcription phonétique et graphémique, de segmenter, de faire des transcriptions multiples en aveugle, et possède des fonctions d'analyse très souples, accessibles à travers une interface graphique conviviale. PHON fonctionne sur plusieurs systèmes informatiques (Windows, Macintosh, Linux), est compatible avec le format CHILDES (Talkbank XML) et permet le codage avec les fontes Unicode, ce qui facilite l'échange de données entre chercheurs.

PHON est mis à la disposition de la communauté sous forme de logiciel « open-source ». Il répond aux besoins de l'étude du développement phonologique en langue maternelle (y compris le babillage), de l'acquisition langue seconde, et des troubles du langage. Il fournit ainsi un standard unifié pour obtenir des représentations de données phonologiques dans le cadre de la continuation des projets d'échange de données CHILDES.

[5] http://childes.psy.cmu.edu/phon/

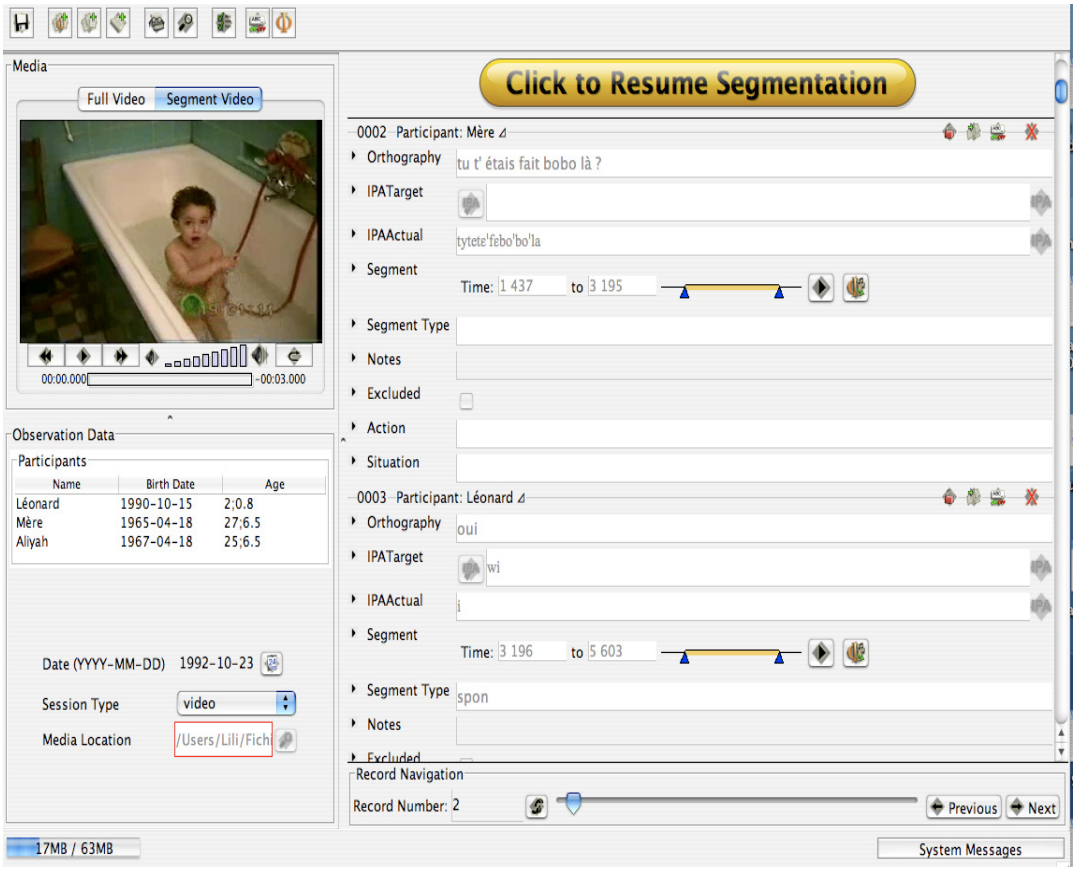

Figure 2 : Segmentation et transcription (d'un énoncé) avec PHON.

Nous avons vu que la conjonction d'une réflexion méthodologique et des progrès technologiques récents est à l'origine du développement d'outils communs utilisés par toute une communauté de chercheurs, ainsi que la mise en partage des données recueillies. Mais ce faisant les observables sont devenus plus complexes, et les généralisations plus ardues, du fait même de la pluralité et de la richesse des corpus accessibles. Pour répondre à ces questions, proposons ici une approche des données basant les statistiques sur des analyses linguistiques minutieuses : nous présentons d'abord deux types d'enrichissement des transcriptions, puis un système de codage.

3. L'enrichissement des transcriptions

Si transcrire, c'est trahir, en ce sens que le chercheur impose toujours déjà une interprétation aux données brutes qu'il traduit, l'on pourrait voir dans une transcription minimale le moyen de limiter la part de ces impondérables. Pourtant, la sélection ainsi opérée est un choix qui engage bien plus encore que la tentative, quoique toujours imparfaite, de tout transcrire. En effet, sur quels critères privilégiera-t-on un niveau (le verbal, par exemple) au détriment d'un autre ? Il est évident que la sélection n'est jamais neutre. C'est pour cette raison, ainsi que pour répondre aux besoins d'un groupe de chercheurs aux intérêts et compétences divers, que nous avons créé dans le cadre du *projet Léonard*[6], un format de transcription enrichi permettant de condenser dans le « texte » le maximum d'informations et d'encadrer les

[6] http://anr-leonard.ens-lsh.fr

risques liés à la subjectivité de chaque chercheur (Snyder, 2007), tout en lui assurant la primauté.

3.1. Le non verbal

Peut-on travailler sur le mimo-gestuel comme on travaille sur le langage, et examiner conjointement le développement de l'un et de l'autre ? Pour Teubert (2005), la linguistique de corpus ne peut pas s'attacher au non verbal, puisqu'on ne peut en donner d'interprétation sûre et fiable, de paraphrase semblable à celle qu'un linguiste est à même de proposer pour le langage écrit comme oral. Cette affirmation polémique a le mérite de montrer la difficulté de travailler sur deux media différents, et la nécessité de mener une réflexion préalable sur le bon usage d'outils avant tout conçus pour l'examen du langage oral. Lorsqu'on regarde une transcription au format CHAT (cf. supra, figure 1) les lignes dépendantes généralement utilisées à cette fin (qui apparaissent au-dessous des lignes de transcription orthographique et phonétique) montrent bien que transcrire le mimo-gestuel, comme les actions et les situations, revient à décrire et non pas à interpréter ou à spéculer sur des intentions des locuteurs en présence. À cela s'ajoute une particularité forte du type de discours et d'interaction étudiés en acquisition du langage : il est fréquent que les productions langagières et/ou gestuelles ne soient pas univoques, ainsi des onomatopées comme « poum », « hop », etc., à propos desquelles il importe de savoir si elles ont été prononcées en tapant dans une balle ou en sautant, par exemple.

On veut donc, grâce à la transcription systématique des événements et gestes accompagnant ce qui est dit, pouvoir associer la production verbale au contexte, et tenir compte de cette association dans nos analyses quantitatives et qualitatives. C'est ainsi que nous avons été amenés à développer un système d'annotation spécifique permettant de répertorier les premiers pointages de l'enfant en contexte.

Pour le chercheur en acquisition du langage, le pointage est un moment important puisqu'il correspond à l'entrée dans la symbolisation, bien avant les premiers mots. On distingue (depuis Bates E., Camaioni L. & Volterra V., 1975) deux types de pointage : le pointage proto impératif est l'expression d'une demande, alors que le pointage proto déclaratif établit un partage d'attention sur l'objet ou l'événement désigné. Notre système d'annotation (sur une nouvelle ligne dépendante, intitulée % point) était destiné à déterminer s'il y avait des formes distinctives pour l'un et l'autre. Or, il nous a permis de combiner deux niveaux d'analyse, puisqu'on observe qu'une grande majorité des pointages sont accompagnés de vocalisations. Et c'est quand on en analyse la prosodie que les différences apparaissent nettement : le pointage proto déclaratif se caractérise par une courbe intonative descendante alors que le proto impératif a une intonation montante[7].

On le voit, ce sont plutôt les vocalisations qui permettent de faire le départ entre l'une et l'autre forme, mais il fallait d'abord repérer les pointages en contexte pour s'en apercevoir. L'enrichissement proposé ici (ajout du non verbal) permet donc de revenir aux productions vocales pour en donner une analyse plus fine, car multimodale. Le format CHAT, qui n'impose aucune limite sur le nombre des lignes dépendantes, et permet de réaliser des requêtes portant sur une ou plusieurs d'entre-elles, se prête facilement à ce type d'enrichissement.

[7] Analyses conduites par Marie Leroy.

3.2. Le développement grammatical

Suite aux travaux pionniers de Venezziano & Sinclair (2000), et de Peters (2001), les psycholinguistes ont prêté une plus grande attention aux fillers, ou éléments préfixés additionnels : ces éléments monosyllabiques, souvent vocaliques ou nasalisés, que beaucoup d'enfants ajoutent à leurs premières productions verbales. Ils sont essentiels en ceci qu'ils nous renseignent sur le traitement que fait l'enfant de certains marqueurs linguistiques présents dans le discours qui lui est adressé. En effet, les fillers remplacent aussi bien les déterminants que les auxiliaires, clitiques et prépositions, par exemple. Il nous semble donc tout aussi essentiel d'intégrer dans les transcriptions cette variante développementale, et de ne pas en donner une interprétation univoque sur la ligne principale (transcription orthographique).

Il existe un encodage optionnel des fillers dans le format CHAT, que nous avons repris et enrichi : il consiste à ajouter @fs aux fillers, qui sont toutefois orthographiés comme la forme cible adulte. Or celle-ci n'est par définition pas toujours en correspondance biunivoque avec une forme adulte. Nous avons donc conçu un enrichissement issu d'une réflexion sur les différentes formes, qui nous a conduit à distinguer les trois étapes suivantes :

@fsa : aucune structure syntaxique, vrai filler, pas proche de la forme adulte
avant l'arobase on code la forme phonétique
si l'enfant dit [ə gato] ça va donner *CHI : ə @fsa gateau

@fsb : il y a une voyelle commune entre la cible et la production effective. Il faut que la cible soit claire
si l'enfant a dit [adam] = *CHI : a@fsb dame

@fsc : la forme est réalisée, elle correspond à la cible, mais l'enfant n'est pas dans un stade de développement où on lui attribue des marques syntaxiques fiables. Cette fois on met la forme cible en orthographe
l'enfant a dit [ladam] = *CHI : la@fsc dame

Cet enrichissement incorporé dans les transcriptions permet de rester au plus près des productions de l'enfant. Ce faisant, nous intégrons des décisions théoriques nouvelles, qui montrent bien que la transcription n'est jamais neutre. Dans le cas présent, la décision revient à distinguer une production verbale qui, au stade « c », peut sembler indiscernable de la cible, mais que l'on appelle filler parce qu'on considère que l'enfant ne construit pas le déterminant. On dispose pour ce faire de critères comme l'absence de substitutions (d'un déterminant à l'autre par exemple) : un usage qui n'est pas paradigmatique ne saurait être rangé dans une catégorie grammaticale qui, à l'évidence, n'existe pas encore dans le système de l'enfant.

Un tel enrichissement témoigne donc d'une attention particulière prêtée à ces voyelles comme étant le fruit d'un travail grammatical, et des esquisses de la production adulte. Toutefois, l'absence d'encodage de ces formes dans la transcription revient à effacer toute une dimension du développement grammatical en « régularisant » les productions de l'enfant.

3.3. De l'enrichissement au codage

On le voit à travers les deux formes d'enrichissement que l'on vient de présenter : nous sommes déjà proches des procédures de codage des données. Dans le format CHAT, ce sont les lignes dépendantes qui permettent le codage proprement dit, mais on peut aussi l'intégrer aux lignes principales en fonction des besoins.

À la différence de la transcription simple, de tels enrichissements permettent déjà de repérer et d'analyser un ensemble d'énoncés. Ils permettent surtout de faire ce qu'aucune analyse

automatiques ne peut réaliser à partir d'une transcription brute : repérer l'absence de marqueur. Cela peut se faire à partir du non verbal : ainsi, avec le pointage proto déclaratif on peut repérer les premières dénominations, accompagnées ou non de productions verbales. Ou encore à partir du filler, que l'on peut penser comme une amorce de préposition, par exemple. Ainsi dans l'énoncé suivant :

(1) *CHI : a@fsa doudou a@fsa Madeleine.

Que l'on pourra repérer et distinguer à la fois de (2) et de (3) :

(2) *CHI : doudou Madeleine.

(3) *CHI : le doudou de Madeleine / pour Madeleine.

Pour généraliser ce type d'analyses, nous sommes en train de réfléchir à la création d'une ligne dépendante permettant de repérer l'absence de forme, là où elle serait attendue dans la langue cible. Ici encore, c'est un enrichissement des transcriptions qui constitue aussi une première forme de codage.

4. Codage et analyses

Nous examinons pour finir une procédure de codage élaborée pour les besoins de nos études récentes sur les premières prépositions, en vue de montrer l'impact des catégorisations utilisées ou produites par le chercheur.

Nous avons codé le sémantisme de chaque préposition en contexte, et réalisé une comparaison inter-langues (français/anglais) afin de réexaminer la pertinence de l'hypothèse localiste pour l'ontogenèse. Comme nous allons le voir, il s'agit d'une procédure que l'on ne peut automatiser qu'en partie, et qui par conséquent met en jeu des analyses raisonnées en amont, mais aussi tout au long du processus.

4.1. L'analyse en amont

Les programmes de CLAN sont capables, à partir d'une transcription même minimale (c'est-à-dire ne comprenant que des lignes orthographiques), de repérer toutes les prépositions produites par l'enfant et de les lister par types avec leurs fréquences d'occurrences. Même si de telles analyses sont précieuses comme point de départ d'une étude comme la nôtre, on ne peut apercevoir dans les résultats obtenus qu'un panorama peu significatif et difficilement interprétable.

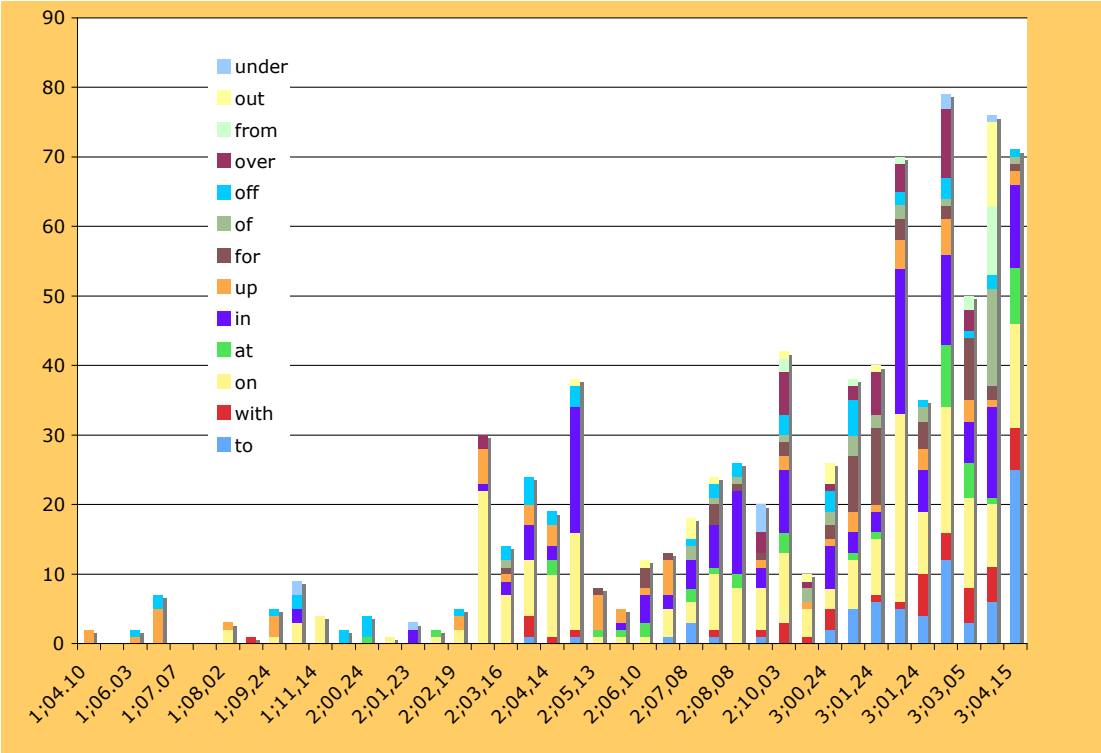

Figure 3 : Panorama des premières prépositions (types et occurrences) produites par un enfant anglophone entre un an et quatre mois et trois ans et quatre mois.

Pour répondre à la question qui nous occupait, nous avions besoin de catégoriser ces usages en fonction de leur sémantisme spatial ou non. Décider que tous les usages de la préposition anglaise « to » par exemple, étaient spatiaux, sans examen des occurrences en contexte n'aurait pu produire que des résultats peu fiables puisque la vérification de notre hypothèse aurait reposé uniquement sur la fréquence relative de certains types de prépositions jugées intrinsèquement spatiales. La plus grande variété de prépositions spatiales dans la langue anglaise aurait probablement suffi à faire apparaître une primauté du spatial en anglais, et non en français. De plus, il est évident que de nombreux usages canoniques de la préposition « to » sont bien éloignés du sémantisme spatial originaire. Il fallait donc faire reposer nos comptages sur une distinction issue d'analyses linguistiques.

4.2. Le codage

Nous avons codé chaque occurrence sur la base d'une analyse des traits lexicaux (sémantisme spatial) et / ou fonctionnels (assignation d'un cas, appartenance à la structure argumentale du verbe) activés. Le recours à une telle distinction, certes simplificatrice, est destiné à faciliter le repérage d'usages spatiaux, et la stabilité des décisions prises d'un chercheur à l'autre (*inter coder reliability*) montre qu'elle est opératoire.

Il n'existe actuellement pas de programme permettant de faire prendre à un ordinateur des décisions aussi complexes, parce qu'elles résultent non seulement du bon usage d'une catégorisation donnée, mais aussi et surtout d'analyses en contexte. Il faut donc classer chaque occurrence dans un tableau et procéder à un comptage manuel.

Les résultats sont significatifs, comme le montre la figure ci-dessous.

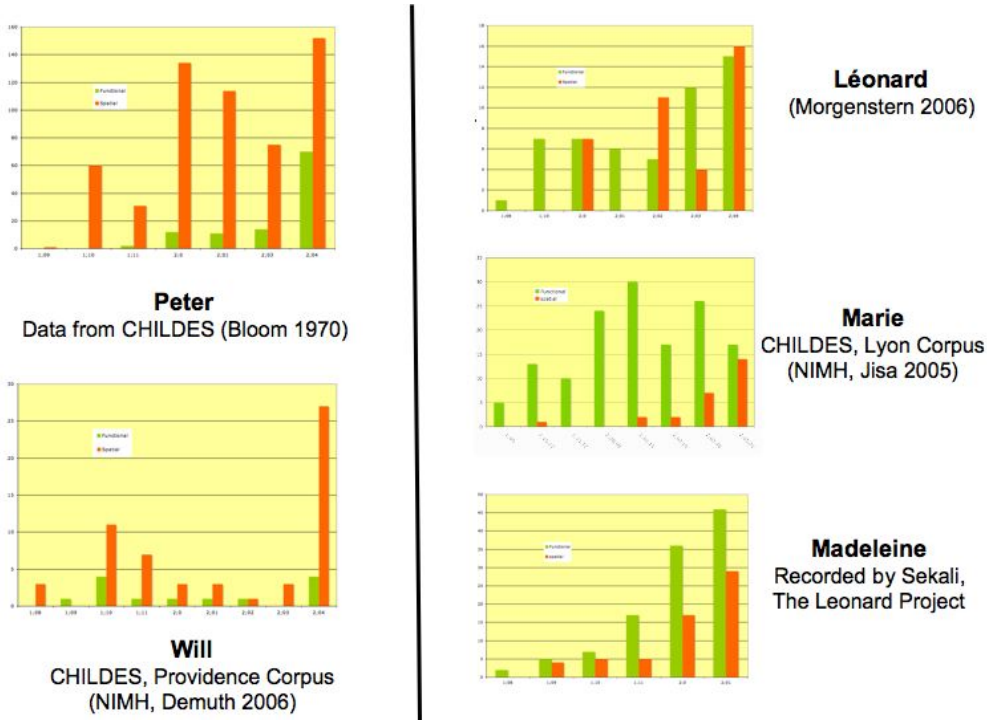

Figure 4 : Usages spatiaux (orange) et fonctionnels (vert) des prépositions chez deux anglophones et trois francophones, jusqu'à trois ans.

La vérification de l'hypothèse localiste chez des enfants anglophones (confirmée ici) est contredite par les résultats obtenus pour les enfants francophones : elle reflèterait donc une différence de structuration des langues, et non une primauté du spatial dans l'ontogenèse. Si les différences structurelles peuvent être repérées assez aisément, c'est cette dernière conclusion sur la primauté du spatial que l'on n'aurait pu atteindre à partir d'une catégorisation des types de préposition (spatiale ou non). En effet il est toujours possible que l'enfant utilise une préposition comme « à », généralement analysée comme « incolore » ou fonctionnelle dans un énoncé référant à une localisation spatiale.

4.3. Analyses quantitatives et qualitatives

Si l'examen attentif des usages en contexte est essentiel, c'est aussi parce qu'il importe de distinguer, autant que possible, ce qui est réellement construit par l'enfant et ce qui est le produit des catégorisations projetées par le chercheur. En effet, si l'on accepte les hypothèses constructionnistes (voir par ex. Tomasello, 2003), l'enfant ne dispose pas d'emblée de notions grammaticales, et doit catégoriser l'input en exploitant l'information livrée par le contexte. Prêter une attention particulière au contexte de l'interaction, c'est donc aussi rester au plus près de la démarche de l'enfant. L'idéal serait que les catégorisations du chercheur concordent avec les « catégories émergentes » que l'on observe chez le tout jeune enfant (Clark, 2001) sans être pour autant des catégories ad hoc. La combinaison d'une analyse linguistique et d'un examen quantitatif minutieux doit permettre sinon d'atteindre, du moins d'approcher cet idéal.

Nous proposons que ce va-et-vient soit opéré non seulement en amont des comptages statistiques, mais aussi qu'une fois les résultats quantitatifs obtenus, on procède à un réexamen des productions repérées afin de comprendre ce qui se joue dans l'interaction. C'est la démarche que nous avons adoptée dans nos travaux récents (Kochan, Morgenstern, Rossi & Sekali, 2007), où l'analyse de micro-séquences nous a permis de mettre en évidence la fonction pragmatique des premières prépositions chez les enfants francophones.

Léonard (1;08) à table au cours du dîner.

Père : Les belles saucisses !
Léonard : Donne.
Mère : J(e) te l'ai donnée, ouais.
Léonard : **Pour** papa,

Figure 5 : Une micro séquence

On voit ici que face à l'incompréhension des parents, la préposition « pour » permet à l'enfant de lever l'ambiguïté en explicitant l'argument du verbe « donner ». En répliquant ce type d'analyses sur un ensemble de micro séquences, nous avons fait l'hypothèse d'une organisation des premières prépositions, catégorie grammaticale émergente, en paradigme pragmatique, autour des fonctions de désambiguïsation, d'argumentation ou de positionnement interpersonnel.

5. Conclusion

En définitive, nous espérons avoir fait apparaître que si la normalisation est un préalable essentiel à tout travail sur corpus en acquisition du langage, du fait de la nécessité de travailler sur un échantillonnage aussi riche et divers que possible, l'importance des analyses de détail ne doit pas être perdue de vue pour autant. Les illustrations proposées ici sont autant de manières de combiner de telles analyses avec un usage raisonné des outils. Nous plaidons donc pour une démarche plurielle, alliant analyse linguistique et examen quantitatif détaillé à chaque étape du traitement des données, et nous montrons que les analyses qualitatives commencent dès la transcription du corpus. Cette fidélité aux données est d'autant plus importante que nous avons désormais la chance de pouvoir conserver les données sous forme de transcriptions liées aux enregistrements vidéo, que tout chercheur peut consulter en permanence. Cela permet un va-et-vient constant entre l'observation des sujets et l'analyse des données, qui seul peut nous aider à appréhender l'entrée de l'enfant dans la langue.

Références

Bloom L. (1991). *Language development from two to three*. Cambridge University Press.

Clark E. (2001). Emergent categories in first language acquisition. In Bowerman M. and Levinson S. C. editors, *Language acquisition and conceptual development.* Cambridge University Press, pp.379-405.

Kochan A., Morgenstern A., Rossi C. & Sekali M. (2007). Children's early prepositions in English and French : a pragmatic device. In *Proceedings of the LingO Conference.* Oxford.

MacWhinney B. (1995). *The CHILDES project : Tools for analyzing talk.* Erlbaum.

Ochs E. (1979). Transcription as Theory. In Ochs E. and Schieffelin B. editors, *Developmental Pragmatics.* Academic Press.

Peters A. (2001). Filler Syllables : what is their status in emerging grammars. *Journal of Child Language,* 28: 229-242.

Snyder W. (2007). *Child language; the parametric approach*. Oxford University Press.

Teubert W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*, Vol(1).

Tomasello M. (2003). *Constructing a Language : A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Veneziano E. & Sinclair H. (2000). The Changing Status of "Filler Syllables" on the way to Grammatical Morphemes. *Journal of Child Language* 27(3): 461-500.