Codage et interprétation du langage spontané d'enfants de 1 à 3 ans

Aliyah Morgenstern, Christophe Parisse

Codage et interprétation du langage spontané d'enfants de 1 à 3 ans

Aliyah MORGENSTERN
ENS-LSH / ICAR

Christophe PARISSE
INSERM / MODYCO

Résumé : La transcription de corpus de langage oral est un art difficile car c'est une activité langagière. En particulier, elle inclut le processus d'interprétation des propos d'autrui que l'on trouve dans toute interaction langagière. Or ce qu'attend le scientifique est une description des données de langage qui s'affranchirait de cette interprétation, ce qui est impossible. On doit donc chercher à générer un processus d'interprétation simple, consensuel, qui puisse être compris par tout utilisateur d'un corpus. Pour cela, on utilise des normes de codage précises, claires, aussi peu ambigües que possible, ainsi qu'un alignement sur du matériel sonore ou audiovisuel. On peut aussi accompagner les transcriptions d'un contexte très riche, soit langagier (phonologique, syntaxique, sémantique, pragmatique), soit extra-langagier (situation, actions, description de scène). Ces difficultés techniques sont présentées dans le cas de corpus de jeunes enfants (projet Léonard) qui exemplifie les problèmes de transcription de langage oral. Les outils et formats utilisés sont ceux du projet CHILDES avec des évolutions spécifiques qui reflètent les difficultés que nous avons relevées dans notre travail de corpus et les solutions adoptées.

Abstract : Transcription of oral data is a difficult art for it is a complex form of language activity which includes an interpretation process of other people's verbal productions. Ideally, researchers would prefer descriptions free of any interpretation, but this is impossible. A solution would be to follow a clear and simple

consensual interpretation process, easily understood by all potential users of the data, using precise and clear coding standards, as unambiguous as possible, as well as linkage between the transcript and the multimedia recordings. This could be completed with an enriched context, both linguistic (phonology, syntax, semantic, pragmatic descriptions) and extra-linguistic (situation, action, descriptions of the surroundings…). In this paper, the technical difficulties of this type of coding are presented within the framework of the transcription process of young children's data. The tools and formats used are those of the CHILDES project with our own additions brought about by the choices we made to face the specific difficulties linked to transcription of child language data.

Le travail sur le langage de l'enfant et en particulier du jeune enfant conduit à se placer aux limites de l'interprétation, de la description linguistique, des techniques de recueil et de présentation de données. L'enfant produit soit des esquisses des formes adultes soit des formes qui lui sont propres, auxquelles même ses proches ont du mal à attribuer un sens. Hors contexte, toute interprétation est la plupart du temps impossible. Cette dimension interprétative nous conduit à reposer ou à préciser des questions théoriques. Quel sens attribuer aux énoncés de l'enfant ? Faut-il se placer du point de vue de l'enfant ou/et du point de vue de son interlocuteur ? Faut-il gloser, interpréter ?

Toutes ces questions sont liées à la nature du langage de l'enfant pour lequel on ne peut que travailler sur de l'oral spontané et non de l'écrit. Ceci oblige les chercheurs à se poser des questions importantes et bien sûr à faire de la linguistique de corpus, et de la linguistique de l'oral (Miller & Weinert, 1998; Parisse 2002, 2005 ; Morgenstern 2006).

Dans le cadre d'une recherche sur l'émergence et l'évolution des marqueurs grammaticaux (projet ANR jeunes chercheurs « Léonard »[1]), notre équipe a dû mettre en place un

[1] Voir site internet : http://anr-leonard.ens-lsh.fr/

recueil de corpus longitudinaux[2] d'enfants. Ce recueil de corpus sous-entend l'application de techniques et de conventions appropriées pour le langage d'enfants qui commencent à peine à parler. Lors de la réalisation pratique de ce recueil, un certain nombre de questions importantes quant à la nature du langage et à sa transcription se sont posées. Ces questions sont de deux ordres : théorique et technique.

D'un point de vue théorique, la pratique de la transcription de jeunes enfants fait ressortir de manière exacerbée la distance importante qu'il peut y avoir entre formes effectivement produites et transcriptions langagières. Un enfant d'un an environ produit en général des formes d'une ou deux syllabes, phonologiquement encore instables et dont l'interprétation est difficile et souvent subjective. Cette distance entre formes et transcriptions se réduit à mesure que l'enfant grandit mais ne disparaît jamais. On est donc face à des choix théoriques importants dans la mesure où ils induisent les résultats des recherches menées sur les transcriptions. De quelle nature doit être la transcription ? Phonétique, phonologique, lexicale, orthographique ? Jusqu'où aller dans les détails de la description ? Ces questions doivent être posées dans toute transcription de corpus d'enfant et résolues en fonction des buts de la recherche.

D'un point de vue technique, il faut se donner les moyens de décrire les sons, les gestes, le contexte, la situation, de manière suffisamment précise pour pouvoir partager les données et les analyses avec des personnes étrangères au recueil de données original. Pour respecter au mieux la situation de recueil, il est nécessaire d'inclure dans les corpus les vidéos enregistrées. Ces données vidéo doivent toujours être complétées par des descriptions textuelles qui permettent de mieux spécifier l'image originelle ou de présenter le contexte de recueil. Les transcriptions linguistiques doivent contenir au minimum des données phonologiques et orthographiques

[2] Il s'agit de suivis d'un même enfant vu par exemple une fois par mois pendant une période donnée. Dans notre cas, cette période allait de l'âge de un an à l'âge de trois ans.

complètes : les données phonologiques permettent de suivre pas à pas de manière quantitative et qualitative le développement du langage de l'enfant, les données orthographiques permettent d'accéder plus facilement à des informations externes au corpus (fréquences, catégorisation syntaxique ou sémantique, etc.). D'autres données (intonation, variations phonétiques, contexte pragmatique, etc.) peuvent être ajoutées en fonction des besoins de recherche spécifiques.

Notre travail sur les productions spontanées de l'enfant dans un contexte dialogique nous montre ainsi qu'il s'agit d'une situation exemplaire pour comprendre la nature interprétative des codages, mais aussi du langage. Toute transcription est réductrice, mais le chercheur se doit de laisser place à l'interprétation de l'autre et à trouver des compromis entre exhaustivité et souplesse de la transcription.

Il ne s'agit pas ici de répondre à toutes ces questions. Nous voudrions montrer comment, tout en étant conscients des problèmes de fond que nous ne pouvons pas résoudre et des limites inhérentes à toute transcription, nous avons pu choisir une méthode de codage et un format. Ceux-ci nous semblent suffisamment souples pour ne pas mettre trop de présupposés dans nos transcriptions et suffisamment standards pour utiliser presque exclusivement les conventions de transcription d'un système existant : CHAT de CHILDES (MacWhinney, 2000).

Nous ferons tout d'abord un bref historique du travail sur corpus longitudinal spontané en acquisition du langage. Nous préciserons ensuite les spécificités de notre projet de recherche et les principes que nous avons voulu mettre en œuvre dans nos transcriptions : utilisation de la phonologie, formes orthographiques redressées avec notification du degré d'interprétation, description des gestes et des contextes, lien avec la vidéo (son et image). Nous décrirons à l'aide d'exemples précis et en présentant des extraits de nos transcriptions comment nous avons utilisé le système CHAT pour inclure tous ces éléments, les problématiques que nous avons eu à résoudre et quels éléments nous avons choisi d'ajouter.

1. Historique du travail sur corpus spontané longitudinal en acquisition

A partir du milieu du dix-neuvième siècle, des théories sur le langage de l'enfant commencent à être élaborées et les scientifiques tiennent des journaux concernant les particularités du langage et du développement de leurs propres enfants. Ils notent quand un phénomène apparaît, se répète, revient ou disparaît. L'intérêt de ces journaux réside dans le fait que le parent connaît bien son enfant qu'il suit quotidiennement et peut justement remarquer les "faits particuliers et intéressants" comme le formule Grammont (1902 : 61).

Ce mode de recueil s'inspire des prises de notes faites par des naturalistes comme Darwin (1945) qui a également été l'un des premiers à écrire un article sur le langage de son propre fils (1877). Il s'agit d'analyses qualitatives. Or, l'histoire de la recherche en acquisition du langage montre que quand on passe d'une approche générale avec des enjeux scientifiques très larges à la fois d'ordre philosophique, linguistique, psychologique, naturaliste, à des problématiques précises qui font appel à des données détaillées, d'autres formes de recueils de données que les journaux deviennent nécessaires afin de pouvoir affiner les observations au niveau quantitatif et qualitatif. Si on prend l'exemple d'une recherche sur les formes d'auto-désignations sujets :

- La quantification des phénomènes permet de connaître le pourcentage de chaque forme selon l'âge de l'enfant, quand on observe comme Cooley dans son journal sur sa fille (1908) que l'enfant emploie plusieurs formes d'auto-désignations (prénom, moi, tu, il/elle, je, absence de forme).

- Un travail qualitatif avec des analyses de formes en contexte dans le cas des auto-désignations donnera lieu à une catégorisation des formes utilisées en fonction du contexte.

Les chercheurs ont donc commencé au milieu du vingtième siècle à travailler à partir de corpus longitudinaux. Les données sont réunies de façon intermittente, une demi-heure tous les quinze jours, ou une heure tous les mois, à des moments fixes et pour une durée déterminée afin de collecter un échantillon représentatif, mais non exhaustif, du langage de

l'enfant et de son développement. Le travail en acquisition du langage a changé radicalement à partir de l'apparition du magnétophone. Grâce aux progrès techniques depuis les années soixante jusqu'à nos jours, les enfants sont enregistrés puis filmés. On a découvert les problèmes de transcription et d'interprétation. La nature du corpus oral et du corpus écrit a commencé à diverger. L'arrivée de la vidéo a permis de mieux prendre en compte le contexte, la situation, les gestes, les mimiques. Les enfants étudiés ne sont plus forcément ceux des chercheurs. Moreau et Richelle (1981) parlent de cette période comme étant le début de la psycholinguistique développementale.

Le processus de collecte, de transcription prend un temps considérable et exige un travail fastidieux. Une session d'enregistrement vidéo d'une heure peut demander une cinquantaine d'heures de travail. Hélas, les pratiques du domaine de la recherche ont révélé un problème général et très sérieux : chaque chercheur a sa propre méthode pour encoder les données. Beaucoup ont un système personnel pour représenter différents aspects des informations collectées, et leur matériel devient inutilisable pour les autres chercheurs.

Une méthode standardisée est vite devenue nécessaire. Au début des années quatre-vingt, deux spécialistes reconnus du langage de l'enfant, Brian MacWhinney et Catherine Snow (1985) ont proposé de mettre en place une base de données informatique sur la production langagière de l'enfant, qui serait disponible pour l'ensemble de la communauté universitaire. Leurs idées ont abouti à la création du CHILDES : Child Language Data Exchange System. Il s'agit d'un lieu de partage de corpus d'enfants sous forme électronique standardisée. Il y a aujourd'hui plus de trente langues différentes auxquelles on peut aisément accéder par Internet. Des centaines de chercheurs ont déjà consulté cette base de données et ont publié des articles où ils citent CHILDES comme leur source principale de recherche.

Il convient toutefois de ne pas surestimer la valeur d'une base de données comme CHILDES. Avec un seul clic de souris, il est tentant de l'utiliser comme unique source et de se

dispenser de constituer sa propre collecte de données : on pourrait finir par ne travailler que sur des transcriptions écrites sans jamais entendre le discours des sujets vivants. Or, se confronter directement au langage d'un enfant grâce à l'observation ou à l'expérimentation est absolument primordial pour tous ceux qui étudient l'acquisition du langage. Travailler uniquement sur des enfants virtuels peut amener à oublier le caractère très interprétatif des phénomènes langagiers et les limites de tout type de transcription. Il est important à la fois d'utiliser mais aussi de participer, de contribuer à l'échange des données au niveau international, et de rester constamment en contact étroit avec les enfants.

2. Spécificités de notre projet de recherche

2.1. Objectifs

Il est essentiel selon nous que les chercheurs fassent eux-mêmes l'observation directe de jeunes enfants et nous avons voulu trouver un compromis : construire une base de données à petite échelle permettant à tous les chercheurs de l'équipe de travailler sur 10 à 15 enfants (participation aux transcriptions, analyse des vidéos, réunions fréquentes du groupe de recherche et discussions sur le développement de chacun des enfants). En parallèle avec les transcriptions complètes, nous mettons en place une base de données de descripteurs, comprenant un résumé de la séquence, les principaux traits saillants au niveau moteur, cognitif, relationnel et linguistique que l'on peut remarquer chez l'enfant à chaque séance, un découpage de la séance en « saynètes » permettant ainsi par la suite de mettre en lien un phénomène avec son contexte (repas, lecture d'album, bain, jeu avec la mère ou le père…) et d'observer finement les phénomènes de ritualisation souvent observés dans les dyades parents-enfants.

Notre objectif scientifique consiste à nous donner les moyens techniques de repérer l'émergence et l'évolution des formes des premiers morphèmes libres grammaticaux chez l'enfant de 1 à 3 ans. Les données doivent être transcrites de telle façon que l'analyseur morpho-syntaxique puisse

fonctionner avec une pertinence optimale. Nous développons un programme qui permettra de repérer l'apparition des marqueurs, de classer les occurrences et d'observer l'évolution des formes avec une attention particulière à la phonologie. Nous pourrons ainsi constituer un panorama des acquisitions de l'enfant et les replacer par rapport à son développement moteur, cognitif et relationnel.

2.2. Principes fondamentaux des transcriptions du corpus Léonard

La réflexion théorique nous conduit à définir des principes qui orientent nos choix techniques de façon précise. Ces principes représentent les éléments qui nous semblent, après réflexion théorique, incontournables dans la transcription de corpus de jeunes enfants. Les choix techniques qui seront faits devront nous permettre de respecter ces principes.

- Donner les moyens au chercheur de se faire sa propre interprétation du langage produit par l'enfant (d'où l'utilisation de l'alignement avec les vidéos, de transcriptions phonétiques, et de descriptions de la situation, des actions, des gestes…).

- Fournir des transcriptions orthographiques et phonologiques précises pour permettre des études statistiques à travers toutes les séances d'un même enfant et entre les différents enfants suivis.

- Utiliser un codage courant dans la communauté et facile à diffuser et à manipuler.

2.3. Sélection d'un logiciel et d'un format de transcription

Après un tour d'horizon, nous avons choisi le logiciel CLAN et le format CHAT. En effet, la transcription en phonétique, l'explicitation du non-verbal, le comptage d'occurrences et l'analyse morpho-syntaxique sont possibles. L'alignement avec la vidéo est assez simple à effectuer et l'on peut facilement vérifier les transcriptions en faisant jouer l'enregistrement. L'alignement se fait sur la ligne dite principale sauf en cas de recouvrement d'énoncés ou d'énoncés très courts. Dans ce cas, un alignement peut correspondre à plusieurs énoncés ou tours de parole. Par ailleurs, le logiciel est

gratuit et la formation des étudiants peut se faire grâce à plusieurs membres du groupe qui connaissent bien le logiciel. Enfin, une large communauté internationale utilise ce logiciel et ce format, ce qui permet de partager très facilement les données.

2.4. Méthode de travail

Dans la pratique, la mise en place d'un corpus se fait de manière incrémentale. On commence par la transcription, ce qui permet d'en comprendre les limites et impose de développer des interprétations (ajoutées au corpus via des lignes d'annotations) et d'augmenter les éléments de contexte par des lignes de description. Une fois ces éléments mis en place, on revient au codage, soit pour le réduire si certains éléments ne sont plus nécessaires, soit pour l'enrichir s'il se révèle impossible d'inclure des informations dans le contexte ou l'interprétation. Cette boucle se répète alors grâce à la confrontation avec d'autres utilisateurs ou codeurs jusqu'à trouver un accord mutuel.

Afin d'éviter d'alourdir le processus d'accord inter-juges, nous avons décidé que toutes les transcriptions seraient simplement vérifiées par un chercheur qui ne connaît pas l'enfant plutôt que réalisées en double à l'aveugle avec confrontation entre les transcriptions. Par ailleurs, les corpus sont « exploités » par les différents chercheurs de l'équipe qui travaillent sur des thématiques différentes. Ces derniers peuvent après discussion apporter des modifications et enrichir le corpus en permanence. Il s'agit donc d'un corpus « dynamique » et non « statique ».

Tout au long du travail de transcription, un guide de codage établi dès le démarrage du projet est continuellement enrichi par les membres du groupe : toute convention nouvelle adoptée est insérée dans le guide.

3. Transcription, interprétation et utilisation des données

La transcription de corpus de langage oral n'est pas une activité scientifique banale : il faut utiliser l'objet même que l'on étudie comme outil d'analyse et de mesure. Cet objet

d'étude est le processus d'interprétation de langage : il consiste, à partir d'un message langagier, à comprendre ce qui a été dit. Dans un échange langagier ordinaire, le résultat de cette compréhension n'a pas nécessairement un format langagier. Il s'agit au minimum d'une modification de l'état cognitif, entraînant une réponse, langagière ou non, qui s'inscrit dans le processus dialogique au cours d'un échange langagier. Ce processus d'interprétation est largement dépendant des locuteurs et des circonstances, et son résultat est éminemment variable. On voit ceci de manière exacerbée dans les échanges avec de jeunes enfants car ceux-ci sont souvent difficiles à comprendre même pour les adultes qui les entourent. La « distance » entre les productions vocales et le résultat de l'interprétation peut être très grande en ce sens que le résultat peut être très largement « fabriqué » par l'interlocuteur adulte de l'enfant (voir en particulier les exemples 1 et 8).

Dans une transcription de corpus, on doit de la même façon transformer ce qui a été dit dans un format de référence pour le travail linguistique proprement dit. Le plus souvent, cette représentation vise à refléter aussi parfaitement que possible la forme originale sans invoquer de processus de compréhension (par exemple, en utilisant une transcription phonétique). En effet, il ne s'agit pas de transcrire ce qui est entendu par l'interlocuteur, mais ce qui est dit par le locuteur. Même lorsqu'on utilise une transcription graphémique, on cherche à coller à la forme originale (souvent en modifiant la forme graphémique « officielle »). Pourtant, qu'il s'agisse d'une représentation orthographique ou phonétique, il existe toujours une distance entre la représentation et l'activité originale. Cette distance est variable en fonction des circonstances, de l'interprète, et du format choisi. Ainsi, les pourcentages d'accord interjuges lors d'une transcription phonétique pure ne dépassent pas dans le meilleur des cas 70% chez des transcripteurs expérimentés. Paradoxalement, les accords interjuges sont plus élevés dans des transcriptions orthographiques, ce qui montre bien que le processus d'interprétation et de compréhension est indispensable pour arriver à obtenir un format langagier fiable. La transcription de

langage oral ne peut donc s'abstraire d'une phase d'interprétation qui permet de produire une forme linguistique écrite représentant au mieux mais sans universalité la forme orale produite originalement.

Dans une situation normale de langage, les erreurs d'interprétation engendrent en général une incompréhension entre l'enfant et l'adulte qui, si elle pose problème, se résout de manière dialogique. Ceci est bien sûr impossible dans une transcription (même s'il peut exister un certain dialogue entre créateurs et utilisateurs de corpus ou entre plusieurs transcripteurs d'un même corpus), et le chercheur doit donc effectuer ses propres vérifications. Le principe que nous adoptons est que, plutôt que de nier l'aspect interprétatif de la transcription, il faut l'accepter et fournir les éléments qui l'accompagnent et permettent de le comprendre ou de le remettre en cause. Pour cela, il faut prendre en compte le contexte dialogique et la situation et utiliser des normes de codage claires et aussi complètes que possible. Il est important de multiplier les codages quand cela est possible (phonologique, pragmatique, sémantique, syntaxique) ainsi que d'intégrer de nombreux éléments fournis par l'audio et la vidéo, (gestes, mimiques, prosodie, description des éléments non visibles parce que hors champ).

Ces multiples éléments permettent aussi de favoriser la deuxième étape de l'utilisation des corpus de langage : leur utilisation par des personnes n'ayant pas participé au recueil de données original. Cette deuxième étape est elle-même un processus langagier à part entière et peut conduire à des interprétations éloignées, à tort ou à raison, de l'idée originale du premier transcripteur. Les capacités de logiciels d'interrogation et de visualisation de corpus ont ici un rôle important à jouer pour faciliter le travail d'utilisation des corpus.

3.1. Problèmes de transcription

Les normes de transcription contiennent non seulement des éléments de description à respecter, mais surtout un cadre

méthodologique qui permet d'aider à sélectionner et analyser les éléments que l'on veut y faire figurer. Le texte écrit, souvent utilisé comme support de représentation des productions orales, est un exemple typique des limites de tout codage. Il manque beaucoup d'éléments du discours oral (intonation, prosodie, hésitations, reprises, circonstances de production) et il contient par ailleurs des éléments absents de la langue orale : éléments non prononcés et variantes morphologiques (*vert*/*verre*/*vair*/*ver*/*vers*), marques orthographiques de l'histoire de la langue (le *th* de *thèse*, le *ch* de *chaos*, l'accent circonflexe de *hôtel*), ponctuation… Tout codage de langage oral même quand il ne repose pas sur une base orthographique possède les mêmes limites : des sous-spécifications dues à la normalisation des codages qui pose problème face à des situations nouvelles et des sur-spécifications variables d'un transcripteur à l'autre au cours du processus d'interprétation des données à transcrire. Soulignons que les sur-spécifications sont souvent nécessaires dans les corpus de jeunes enfants pour que le codage soit utilisable hors du contexte original.

> Exemple1 : quand l'enfant dit /ko/, si l'on ne spécifie pas qu'il signifie en contexte de « encore », il est difficile de le deviner car il n'y a pas de référent explicite visible dans la vidéo et la forme phonologique peut pour ceux qui ne connaissent pas le langage de l'enfant sembler trop éloignée de la cible pour retrouver l'équivalent chez l'adulte. On choisira donc de transcrire sous format CHAT :
> *CHI: encore
> %pho: kɔ
> %sit: l' enfant veut manger un autre biscuit.
> *CHI = la ligne principale sur laquelle vont se faire les analyses quantitatives (fréquence des mots par exemple).
> %pho = ligne secondaire sur laquelle on transcrit en phonétique, de façon plus ou moins fine la production de l'enfant.
> %sit = linge secondaire dans laquelle on code des éléments de la situation.

Le travail sur le langage de l'enfant place souvent le transcripteur dans une situation qui est plus rare ou moins extrême chez l'adulte. Les productions de l'enfant sont

« fragmentaires » ou extrêmement différentes de celles qui sont attendues dans la norme de la langue (on trouvera le même type de problème chez l'adulte ayant des troubles de langage ou dans des situations « réelles » en milieu bruité par exemple). Cette situation entraîne une interprétation des productions de l'enfant pour « savoir ce qui est dit » qui relève parfois de la devinette – situation souvent réelle dans un dialogue entre un enfant et un adulte, même avec un proche parent. Cette interprétation peut porter sur les phonèmes, les éléments lexicaux, la structuration grammaticale comme le sens des propos à transcrire. Souvent, chez un jeune enfant, un mot sera prononcé incomplètement, par exemple /aty/ pour voiture. Parfois un ou plusieurs sons sont difficiles à identifier comme le /a/ dans /aty/ qui peut ressembler à un /ə/. S'agit-il alors vraiment d'un mot, par exemple « voiture », ou du lexème représenté par /ty/ et d'un embryon de détermination représenté par /a/ ce qui permettrait de gloser l'ensemble par « la voiture » ? De nombreuses interprétations sont souvent possibles en fonction de la qualité des phonèmes effectivement identifiés. Par exemple, /ə/ ne suggère pas (a priori) l'usage d'un article alors que /a/ le suggère de par sa plus grande proximité phonique avec l'article défini féminin. La difficulté est que l'on est plus ou moins obligé de normaliser les données transcrites pour pouvoir suivre le développement longitudinal d'un enfant, ou le comparer avec d'autres. Par exemple, on a besoin de pointer toutes les occurrences d'utilisation de « voiture » ou de l'article défini pour étudier le développement du vocabulaire, de la qualité des réalisations phonologiques, de la morphosyntaxe du déterminant, etc. Cette méthode qui va de la langue de l'enfant vers celle de l'adulte consiste à « redresser les énoncés ». Il faut absolument être conscient de ses limites et de ses conséquences afin d'éviter de faire des erreurs ou des contresens dans les analyses.

3.2. Point de vue adulte vs. point de vue enfant

L'opération de normalisation des productions de l'enfant n'est pas neutre d'un point de vue théorique car elle suppose qu'il y a un « modèle de référence », en général celui

de la langue adulte. En se plaçant *du point de vue de l'adulte*, on accepte nécessairement une interprétation forte qui entraîne une grande distance entre le langage effectivement produit et la transcription. C'est en soi une situation naturelle qui fait partie intégrante du processus d'acquisition en dialogue. Sans interprétation, sans va-et-vient entre compréhension et incompréhension entre l'enfant et l'adulte, le langage de l'enfant n'évoluerait pas de la même façon. Toutefois ce qui est naturel dans l'interaction enfant/parent pose des problèmes lors du travail sur corpus. Il faut être attentif aux cas de trop forte interprétation, appellés surinterprétations. Il s'agit des cas où la distance entre la production effective de l'enfant et la compréhension de l'adulte est tellement grande qu'elle est discutable, contestable et éventuellement source d'erreur. On peut en limiter la portée en s'autorisant à « noter » le degré de certitude des transcriptions, ce qui permet de faire savoir aux utilisateurs de corpus que certains points sont très fortement interprétés sans s'interdire de réaliser une transcription riche. La représentation adultomorphe des productions de l'enfant pose aussi le problème qu'elle n'est pas nécessairement une bonne représentation des connaissances effectives de l'enfant étudié et peut conduire à des erreurs scientifiques si elle n'est pas utilisée avec grande précaution.

L'autre attitude est celle qui consiste à se placer *du point de vue de l'enfant*, c'est-à-dire de ne pas utiliser dans les transcriptions de matériel langagier non maîtrisé par l'enfant. Par exemple, on ne transcrira pas d'article tant que l'enfant n'a pas un système de détermination nominale clairement développé. Il est alors nécessaire de disposer de notations alternatives ou d'utiliser simplement un codage phonologique.

C'est par exemple le principe que nous avons utilisé pour coder ce que l'on appelle les fillers. Il s'agit d'éléments qui sont produits par l'enfant là où on s'attendrait à trouver une marque morphosyntaxique (déterminants, pronoms préverbaux, auxiliaires, etc.) chez un enfant qui n'utilise pas encore de manière systématique ces marques. Le développement des fillers passe par plusieurs périodes. La première période est celle où les formes produites ne permettent pas de déterminer la

cible adulte. Ainsi dans l'exemple ci-dessous, la forme /ə/ est produite là où on attendrait un déterminant, mais on ne peut pas savoir lequel. La deuxième période est celle où les formes phonétiques correspondent en partie à la cible adulte, ce qui permet par exemple de faire la différence entre les fillers utilisés devant les noms ou devant les verbes. Dans l'exemple ci-dessous, l'enfant produit /a/ qui correspond à la voyelle du déterminant « la ». Enfin, la troisième période est celle où les formes produites sont identiques aux formes cibles adultes, mais où l'enfant ne présente pas par ailleurs des productions suffisamment organisées pour lui attribuer la maîtrise des déterminants.

Exemple 2 : Codage des fillers

*CHI: donne ə @fil1 tarte.
%pho: dɔ n ə tat
*CHI: donne a@fil2 tarte.
%pho: dɔ n a tat
*CHI: donne la@fil3 tarte.
%pho: dɔ n la tat

3.3. Multiplicité des points de vue

Nous rencontrons un second problème dans la transcription et l'analyse des interactions adulte-enfant : il ne peut pas y avoir une seule approche, mais au contraire de nombreuses façons très différentes de concevoir le travail de recherche avec les mêmes outils. Chaque problème particulier doit parfois être traité de manière spécifique. Par exemple un codage phonétique très fin est complètement différent d'un codage phonologique alors que la plupart des systèmes de codage de corpus utilisent les mêmes conventions de transcription pour les deux formes de codage.

Exemple 3 : Codage phonétique et codage phonologique

*CHI: elle est très jolie.
%pho: ɛ : ë tʷ ɛ ʒ ɔ li: [= phonétique]
%pho: ɛ e te ʒ ɔ li [= phonologie]

Ainsi, dans l'exemple 3, la transcription phonétique de « elle est très jolie » utilise des diacritiques pour spécifier le détail des productions de l'enfant, tandis que la transcription phonologique n'utilise pas ces diacritiques car ils ne servent pas à marquer une différence de sens en français de France. Par contre, cela ne serait pas le cas en français de Belgique où l'allongement final du /i/ sert à marquer la différence entre féminin et masculin. Dans ce cas, le diacritique d'allongement figurerait sur la ligne phonologique. CHAT ne permet pas de marquer la différence entre phonétique et phonologie. Dans le cas du corpus Léonard, seule une transcription phonologique a été utilisée. Il n'a donc pas été nécessaire de développer un formalisme spécifique[3].

La plupart du temps, on n'envisage qu'une transcription de ce type, ce qui ne satisfera pas les spécialistes de ces domaines. On est donc amené dans la pratique à amender et compléter les systèmes de notation existants. Ainsi, au cours du développement pratique d'un corpus, se met en place une interaction complexe entre trois notions fondamentales dont le jeu définit la nature des transcriptions réalisées : le codage, l'interprétation et le contexte.

3.4. Codage et interprétation

Le codage est l'ensemble des codes utilisés dans les transcriptions, c'est-à-dire tous les éléments qui sont le résultat du processus d'interprétation. Il décrit les données de langage et leurs modes de production. Par exemple un codage complet pourra contenir une trace phonétique, phonologique, lexicale, syntaxique, sémantique et pragmatique du discours tenu (voir exemples 6-7 ci-après).

Par définition, le codage est le résultat d'un choix qui correspond à l'analyse des transcripteurs. Il convient d'être conscient de leur travail interprétatif mais aussi et de manière

[3] Dans le cas de besoin précis en phonétique, nous avons préféré utiliser un autre logiciel plus spécifique et plus puissant, PHON développé par Yvan Rose et collaborateurs à Memorial University of Newfoundland, Canada (http://childes.psy.cmu.edu/phon/).

peut-être plus importante à l'interprétation que feront les utilisateurs des données. En effet, les analyses des utilisateurs seront toujours colorées par les choix originaux de codage et de convention. Pour être utilisé par de nombreuses personnes, un corpus doit laisser une part importante de liberté dans ce processus d'interprétation. Sinon les utilisateurs de corpus peuvent avoir l'impression qu'on leur force la main en imposant des choix et des points de vue théoriques, ce qui non seulement serait préjudiciable à la recherche, mais également à l'utilisation du corpus par autrui.

Exemple 4 :
*CHI: par terre met à terre # Esther met de l' eau

qui correspond en fait à

%pho: pa dɛ ma e tɛ: # etɛ ema du yɔ:

Ainsi, dans l'exemple 4, l'enfant a 22 mois et n'a pas encore du tout développé de manière systématique son système de détermination et ses catégorisations lexicales. Il n'est pas capable de faire la différence subtile entre « par terre » et « à terre ». Il serait alors plus juste d'utiliser le codage « par+terre » (avec comme transcription phonologique /padɛ/) ou « à+terre » (avec comme transcription phonologique /etɛ/) qui spécifie que les formes différentes qui sont produites sont traitées de manière globale comme des éléments figés. Cette décision peut être laissée au transcripteur.

3.5. Codage et contexte

Le contexte fait référence à l'ensemble des éléments qui aident à comprendre la situation de langage et les interprétations que l'on peut en faire, mais qui ne sont pas soumis à une interprétation linguistique. Typiquement, il s'agit du son, de la vidéo, mais aussi de la description écrite de la situation, dans la mesure où il ne s'agit pas d'une interprétation des productions orales. Le contexte peut être sujet à interprétation, à condition qu'elle ne soit pas « spécifiquement » langagière. Il a pour but de justifier les choix de codages et

d'interprétation et de permettre aux utilisateurs des corpus de se faire leur propre appréciation des données. Pour cela, nous utilisons deux techniques complémentaires : notations (notations non-verbales, notation du contexte, etc.) et fourniture en liaison avec le corpus des données originales sonores ou vidéo.

La définition du codage non-verbal varie d'un auteur à l'autre. Elle peut être assez spécifique ou très générale selon la manière dont on définit le non-verbal. Ici, on prend ce terme dans son sens le plus large. Il s'agit donc de coder tous les éléments de contexte qui ne figurent pas dans les productions verbales proprement dites, en particulier les gestes, les regards, les mimiques, l'intonation, les actions, les situations, les antécédents, ou toute autre information utile pour l'interprétation et l'analyse des données. Le non-verbal (qui inclut donc le co-verbal) sert à vérifier si l'enfant utilise les termes de façon appropriée et si ses productions verbales sont insérées de manière « pertinente » dans le dialogue avec ses interlocuteurs. Si l'enfant dit /koko/ en montrant la caméra, il peut être utile de savoir que c'est le terme qu'il utilise pour référer aux yeux. On pourra éventuellement en déduire qu'il met en œuvre un processus d'analogie et fait une surextension. Le codage de ce type d'information permet d'aller plus loin que les données fournies avec le son ou la vidéo. Il permet en particulier de donner des interprétations (non-verbales) des situations. On peut également faire référence à des éléments qui ne sont pas visibles sur la vidéo.

Exemple 5 :
%sit: on entend la sonnerie de la porte.
*CHI: oh c'est qui ça ?
%pho: o se ki sa
*MOT: oh qui voilà ?
%sit: P entre. L court l'accueillir et l'accompagne au salon tout excité.

En plus de la notation du contexte, comme l'on fournit avec tout corpus les sources originales audio ou vidéo alignées sur les productions de langage, on pourrait par ailleurs imaginer de multiplier les enregistrements vidéo d'une même situation

afin d'augmenter le nombre de points de vue (Mondada (2006) a d'ailleurs mis en pratique cette méthode).

Si la vidéo ne peut pas remplacer le codage du non verbal faute de temps pour vérifier à chaque fois la pertinence de la transcription, elle est nécessaire. Un bon travail implique d'avoir visionné au moins une fois (et même plusieurs fois) la vidéo avant et en même temps que l'analyse des données transcrites. Il y a un va-et-vient nécessaire entre la transcription et l'enregistrement vidéo.

Dans le codage CHAT, il est possible d'intégrer de multiples champs d'information. Il y a toutefois une ligne « principale » (codée avec le symbole *) qui est considérée comme plus fondamentale qu'une autre, ne serait-ce que par le simple fait qu'elle est obligatoire et sert à définir la découpe des énoncés. Le concept de ligne principale présente un danger car elle est alors généralement considérée comme LA référence de la transcription et trop souvent utilisée comme telle dans les travaux qui analysent le corpus sans avoir participé à sa création. Il serait souhaitable que les conditions de création de cette ligne principale soient clairement définies, et que son utilisation (ou sa non utilisation) ne conditionne pas les autres éléments de codage.

Les transcripteurs peuvent utiliser autant de lignes « secondaires » (codées avec le symbole %) qu'ils le souhaitent.

Exemple 6 :

@Situation: CHI et MOT sont assis à table.
*MOT: j' ai fait rouge.
%act: MOT dessine.
*MOT: c' est fini.
%act: MOT retire sa main de la feuille. CHI fait violemment non de la tête.
*CHI: yy donne.
%pho: ma don
%act: CHI fait le geste de saisir le stylo des mains de sa mère.
%reg: CHI vers MOT
%pov: 1 personne
%spa: injonction

Explication des codages : %act=action des interlocuteurs ; %reg=regard ; %spa=acte de parole ; %pov=point de vue (1ère, 2e ou 3e personne)

Le manuel de CLAN présentant le codage CHAT offre déjà de nombreuses possibilités de lignes « secondaires » et tout transcripteur ou utilisateur du corpus peut ajouter les lignes dont il a besoin, à partir du moment où il les définit dans un fichier descriptif.

Exemple 7 : liste non exhaustive de champs secondaires possibles

Lignes standards

*CHI: transcription orthographique « principale » et annotations diverses (rires, bruits, intervention non-verbale, etc.)
%pho: phonétique
%sit: situation pendant la production de l' énoncé
%act: actions des interlocuteurs pendant la prise de parole
%spa: actes de langage
%mod: modèle phonologique tiré de la langue adulte
%com: commentaire du transcripteur
%mor: morphosyntaxe
%syn: structure syntaxique
%imi: imitation
%fac: mimiques faciales
%adr: à qui le locuteur s' adresse
%sem: rôle sémantique
%gpx: gestes

Lignes que nous avons ajoutées dans le projet Léonard
%point: pointage
%int: degré d' interprétation de la part du transcripteur

3.6. Liens entre codage, interprétation, et contexte

Comme nous commençons à l'entrevoir, ces notions sont interdépendantes ; jouer sur l'une d'elles modifie les deux autres et vice-versa. Par exemple, en maximisant le codage, on vise à obtenir une description qui soit totale et ne présente qu'une seule interprétation possible. Si l'on veut laisser un large champ interprétatif, il faut réduire le codage de manière à ne

pas être trop fortement influencé par les codifications existantes et pour pouvoir aller à leur encontre. De la même manière, l'existence de beaucoup d'éléments de contexte modifie le codage et l'interprétation. Une description de contexte permet d'alléger le codage (réduire le nombre d'éléments fortement interprétés) et permet aux utilisateurs de corpus d'arriver à une vision différente de celle des transcripteurs originaux. Inversement, un codage fin masque le contexte et maximise l'importance de l'interprétation du transcripteur.

Pour clarifier le codage, il faut préciser son but, la visée du projet, et l'utilisation qui est envisagée des corpus. En effet, un codage ayant pour but une étude phonétique, phonologique, syntaxique, sémantique, ou pragmatique n'aura pas du tout la même structure et ne sera pas basé sur les mêmes présupposés.

Il est parfois nécessaire de faire figurer côte à côte des éléments de différentes natures, ou éventuellement contradictoires, surtout lorsqu'une transcription a été enrichie par les versions proposées par différents transcripteurs. C'est ce que nous faisons avec l'utilisation du code %int qui donne le moyen de marquer explicitement différentes interprétations et le degré de certitude du transcripteur. Les deux transcripteurs diffèrent dans l'exemple suivant sur l'interprétation de « koko ».

Exemple 8 :

@Situation: dans le bain. MOT est accroupie près de la baignoire.
*CHI: oh !
%pho: o:
*MOT: oh il est bon ce bain .
%sit: CHI se lève et montre la caméra du doigt.
*MOT: il a une bonne température.
*CHI: coco .
%pho: kɔkɔ
%int1: coco/2/=caméra
%int2: coco/3/=œil
Ici, sur une échelle de certitude qui va de 1 à 3 (par degré de certitude décroissant), le transcripteur 1 a accordé un 2 à sa proposition, le transcripteur 2 a accordé un 3 à son interprétation.

L'utilisateur du corpus peut quant à lui regarder l'enregistrement vidéo et soit choisir l'une des interprétations, soit en avoir une troisième. La transcription ne l'enferme pas et lui laisse une grande souplesse.

Deux autres éléments viennent aussi modifier fondamentalement la nature des codes utilisés : le degré de complexité du codage et l'ouverture vers d'autres interprétations ou notations. La grande complexité d'un code permet de le rendre plus précis mais nuit à son utilisation qui devient trop lourde et à sa clarté. Par exemple, un codage phonétique complet nécessite un entraînement particulier pour être facilement lu, ce qui n'est pas le cas d'une transcription orthographique. Des indicateurs pragmatiques, interactionnels ou situationnels viendront également en surcharge s'ils sont directement inclus dans la transcription même. Ces problèmes de complexité peuvent être minimisés par l'utilisation de plusieurs niveaux de codages qu'on est alors libre ou pas d'activer. Un autre problème lié à la complexité est l'accord inter-juges. Pour un élément de codage (par exemple l'intonation d'un énoncé), plus le choix mis à la disposition du codeur est grand, plus il sera difficile d'obtenir un accord inter-juge. Aussi certains éléments seront mal utilisés ou peu utilisés. Inversement, un codage trop simple sera réducteur. Il s'agit donc de trouver pour ce type de problème un niveau de complexité adéquat pour la tâche envisagée et éventuellement modulable en fonction des compétences des utilisateurs des corpus.

L'ouverture des codages vers des interprétations alternatives doit être envisagée. Elle permet souvent de résoudre des cas difficiles, de confronter des points de vue multiples et de rappeler aux utilisateurs de corpus qu'aucun choix de transcription n'est inscrit dans le marbre. Toutefois, pour des raisons de complexité et de temps de codage, il faut être conscient qu'il est impossible de proposer des alternatives pour tout élément codé. La possibilité d'un marquage d'incertitude dans un codage nous a donc semblé nécessaire dans notre méthode de travail.

3.7. Utilisation des corpus

On voit que le codage peut comporter des éléments d'incertitude et prêter à discussion. Il faut être conscient que toute transcription et système de codage est discutable et doit être utilisé en connaissance de cause, c'est-à-dire en faisant attention aux règles qui ont servi au codage. Dans notre corpus, nous avons choisi de nous conformer aux principes de base du système CHILDES, du format CHAT et du logiciel CLAN. Ceci impose l'utilisation d'une ligne principale. Comme conventionnellement il s'agit de la ligne orthographique, c'est sur elle que portent la plupart des analyses lexicales et morphosyntaxiques. Certains parmi nous ont d'ailleurs beaucoup de mal à accepter la survalorisation de cette ligne principale et les erreurs qu'elle peut induire chez des personnes utilisant notre corpus sans avoir participé à sa création. Rappelons que cette ligne principale doit être considérée comme une glose de la production verbale de l'enfant et qu'il faut absolument tenir compte de tous les compléments d'information contenus dans les lignes secondaires lorsqu'on utilise les corpus pour la recherche. Nous avons cherché à limiter le focus sur la glose en apportant une richesse d'information complémentaire (codage phonétique en sus du codage orthographique, information sur un lexique spécifique à la famille enregistrée comme « bouba bouba » qui signifie nourriture pour la famille du petit Léonard).

Ces problèmes pourraient aussi être levés avec l'utilisation de logiciels d'interrogation et de visualisation de corpus plus puissants que ceux qui existent actuellement. Il n'y a aucune raison théorique pour qu'il existe une ligne de codage « principale » car cela sous-entend qu'il y aurait un « meilleur » codage que les autres. En réalité ce concept de ligne principale n'est pas techniquement nécessaire et l'on pourrait envisager un logiciel grâce auquel toutes les informations liées à une partie du corpus seraient codées au même niveau. L'utilisateur du corpus serait alors libre de décider quelle est ou quelles sont la ou les lignes principales qui correspondent à son analyse. On pourrait ainsi envisager d'utiliser comme référence principale le codage phonologique, le contexte pragmatique, la situation du

discours, aussi bien que la glose orthographique, sans que la structure du système guide le travail d'appropriation d'un corpus que fera chacun.

Conclusion

Les problèmes de transcription de langage oral ne peuvent être résolus que dans la pratique de la transcription et de manière spécifique à un type de travail. Toutefois, il est essentiel d'ouvrir autant que possible ses propres techniques à une communauté large et de maximiser les informations de contexte disponibles pour rendre les transcriptions accessibles aux personnes qui ne connaissent pas la situation originale de recueil de données. C'est ce que nous avons cherché à faire dans notre projet qui regroupe un ensemble de chercheurs travaillant sur des thèmes différents à partir des mêmes corpus. Ainsi, au cours de la création de notre base de données, nous apporterons des solutions de compromis qui pourront satisfaire une grande partie des utilisateurs de données de langage oral, en conservant en tête à tout moment une idée fondamentale : tout système de codage et de transcription doit être évolutif et ouvert.

Le corpus en lui-même est un langage, les limites inhérentes que nous découvrons en transcrivant les productions de l'enfant reflètent les limites de l'interprétation en situation d'interlocution. Par ailleurs, il est important d'aborder en équipe et de résoudre ensemble les problèmes posés par les transcriptions du langage de l'enfant car les réponses apportées, les choix adoptés sont le reflet des approches théoriques utilisées dans les analyses des données (Ochs 1979).

Il faut toujours garder à l'esprit que les choix techniques et scientifiques pris lors de la création de transcriptions sont limités à un ensemble de corpus et à une communauté d'utilisateurs. Le but est bien sûr que cet ensemble soit le plus large possible, tout en sachant que plus la communauté sera large et plus il sera difficile d'arriver à des choix satisfaisants pour chacun.

Ces limites inhérentes au problème posé doivent aussi être repoussées par l'utilisateur même des corpus mis à la

disposition de la communauté scientifique. Il est important que tout utilisateur s'approprie le corpus d'un autre chercheur avec un esprit ouvert, c'est-à-dire en restant conscient du fait que les choix d'autrui ne peuvent recouvrir exactement les siens propres. Mais inversement, il faut aussi que cet utilisateur soit critique pour comprendre les limites inhérentes à tout corpus et en particulier au corpus sur lequel il va travailler. Pour cela, il lui faut connaître clairement les conditions de production et de transcription du corpus, en prenant connaissance des choix, des techniques et des codages utilisés, à condition bien sûr que ceux-ci soient diffusés en même temps que le corpus lui-même. Transcrire et partager des corpus est parfois difficile, mais pour être possible cet effort doit être mutuel, de la part des transcripteurs comme des utilisateurs. Il sera récompensé par la démultiplication des possibilités de recherche qu'offre la mutualisation des données et des opportunités d'analyser un même corpus avec des approches variées et complémentaires.

Références bibliographiques

Cooley, C. (1908). A study of the early use of self-words by a child, *Psychological review*, XV, 6, p.339-57.

Darwin, C. (1945). The voyage of the Beagle, Dover Publications, 2002 (publié sous le titre Journal of Researches into the Natural history and Geology of the Countries Visited During the Voyage of H.M.S. "Beagle" Round the World, under the Command of Capt. Fitz Roy, R.N. John Murray, London 1845).

Darwin, C. (1877). A Biographical Sketch of an Infant, *Mind, vol. 2 :* 285-294.

Grammont, M. (1902). « Observations sur le langage des enfants », *Mélanges linguistiques offerts à M. Antoine Meillet*. Paris, Klincksieck.

MacWhinney, B., & Snow, C. E. (1985). The child language data exchange system. *Journal of Child Language*, 12, 271-296.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk.* Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.McWhinney

Miller, J. & Weinert, R. (1998). *Spontaneous Spoken Language*. Oxford: Clarendon Press.

Mondada, L., (2006), Video Recording as the Preservation of Fundamental Features for Analysis, in Knoblauch, H., Raab, J., H.-G. Soeffner, Schnettler, B. (eds.).

Moreau, M-L. & Richelle, M. (1981). *L'acquisition du langage.* Bruxelles: P. Mardaga.

Morgenstern A. (2006). *Un JE en construction. Genèse de l'auto-désignation chez le jeune enfant*. Bibliothèque de faits de langues. Paris : Ophrys. 177 pages.

Ochs E. (1979). 1979. Transcription as theory. *Developmental pragmatics*, ed. by E. Ochs & B. Schieffelin. New York : Academic Press, pp. 43-72.

Parisse, C. (2002). Oral language, written language and language awareness. *Journal of Child Language*, 478-81.

Parisse, C. (2005). New perspectives on language development and innateness of grammatical knowledge, *Language Sciences*, 27, 4, 383-401.