De l'analyse au partage des données, quel(s) format(s) choisir ? L'exemple d'un corpus d'interactions parents-enfant

Loïc Liégeois

Référence :

Liégeois, L. (2013). De l'analyse au partage des données, quel(s) format(s) choisir ? L'exemple d'un corpus d'interactions parents-enfant. In Damiani M., Dolar K., Florez-Pulido C., Loth R., Magnier J. & Pegaz A. (dir.) *Traitement de corpus (Actes de Coldoc 2012)*. Paris : Modyco, pp. 128-142.

De l'analyse au partage des données, quel(s) format(s) choisir ? L'exemple d'un corpus d'interactions parents-enfant

Loïc Liégeois[1]
(1) Clermont Université, Université Blaise Pascal, EA 999, Laboratoire de Recherche sur le Langage, BP 10448, F-63000 CLERMONT-FERRAND
loic.liegeois@univ-bp.fr

RESUME

Les enjeux inhérents à tout projet de constitution de corpus sont divers mais, parmi ceux-ci, le choix du format d'encodage des données est central. Cet article expose la chaîne de traitement utilisée dans le cadre du projet ALIPE dans le but de constituer un corpus d'interactions orales entre des parents et leur jeune enfant. Afin de constituer une ressource organisée, structurée, documentée, libre d'accès et au maximum interopérable, nous avons retenu deux formats d'encodage : le format CHAT et le format XML-TEI. Nous présentons dans cette étude les méthodes utilisées par l'équipe de recherche pour récolter les données, les annoter et les rassembler dans le but de constituer un corpus. Nous évoquerons également les avantages que l'utilisation du format XML peut apporter pour l'analyse des données ainsi que pour l'interopérabilité entre logiciels de traitement et d'analyse de corpus.

ABSTRACT

Any project dealing with corpus building will be faced with any array of different challenges. However, amongst these, the choice of the data encoding format will be central. This article describes the processing chain used during the ALIPE project whose aim is to build a corpus of verbal interactions between parents and their young children. In order to put together an organized, structured, documented, open-access resource with maximal interoperability, we selected two encoding formats: CHAT and XML-TEI. In this article, we introduce the methods used by the research team for data collection and annotation and describe how the data was assembled into a corpus. We also discuss the advantages of using the XML format with respect to data analysis as well as interoperability between corpus processing and analysis software.

MOTS-CLES : interactions parents-enfant, acquisition, variation phonologique, partage des données, interopérabilité
KEYWORDS : parents-child interactions, acquisition, phonological variation, data sharing, interoperability

1 Introduction

La constitution d'un corpus de données issues d'interactions en situation naturelle est une tâche qui se révèle difficile et coûteuse en temps. En effet, même si les évolutions technologiques ont fourni aux chercheurs un bon nombre d'outils d'aide à la transcription, l'annotation ou l'analyse des interactions, celles-ci ont également amené les chercheurs à se confronter à de nouvelles problématiques, telles que le choix du format de représentation du signal sonore ou de l'encodage des données. Ces décisions méthodologiques sont aujourd'hui au cœur des débats au sein de

différentes disciplines (Reffay, Betbeder, & Chanier, 2012) et font l'objet d'une littérature importante, notamment dans le champ disciplinaire de l'acquisition du langage (Behrens, 2008a). La mise en place d'infrastructures et de projets nationaux (Très Grand Equipement ADONIS, Très Grande Infrastructure de Recherche Corpus) et internationaux (Common Language Resources and Technology Infrastructure) témoigne également de l'envergure sociale et scientifique grandissante des corpus de données langagières.

Dans cet article nous montrerons comment, dans le cadre du projet ALIPE (Acquisition de la Liaison et Interactions Parents-Enfant), nous avons tenté de construire une méthodologie pertinente répondant à l'ensemble des critères nécessaires à la construction d'un corpus d'interactions orales à la base de nos recherches sur l'acquisition de la variation phonologique. Après avoir dressé un bref historique de l'utilisation des corpus dans le cadre de recherches sur l'acquisition du langage, notre étude s'articulera autour de trois axes : l'annotation, la mise en forme des données en corpus et enfin l'analyse des données.

2 Corpus et recherche en acquisition du langage

Dans le champ des recherches en acquisition du langage, la construction de corpus de données constitués à partir des productions de jeunes locuteurs en situation naturelle a depuis toujours occupé une part importante du travail du chercheur. Ingram (1989) prend comme point de départ des travaux sur corpus en acquisition la publication de Taine (1877). Dans cette étude, le chercheur analyse les productions orales de sa propre fille recueillies sous forme de notes rédigées dans un journal (*parental diary*). Cette méthode a été la plus employée à la fin du XIXème et au début du XXème siècle, permettant aux chercheurs de relever des indices de développement non seulement linguistique mais également cognitif en général. Taine (1877, p.256) rapporte par exemple :

From the 15th to the 17th month. Great progress. She has learnt, to walk and even to run, and is firm on her little legs. We see her gaining ideas every day and she understands many phrases, for instance: " bring the ball," " come on papa's knee," " go down," "come here," &c. She begins to distinguish the tone of displeasure from that of satisfaction, and leaves off doing what is forbidden her with a grave face and voice ; she often wants to be kissed, holding up her face and saying in a coaxing voice papa or mama - but she has learnt or invented very few new words.

Cette méthode de recueil de notes « à la volée », bien qu'elle soit limitée (Tomasello & Stahl, 2004) et parfois même jugée subjective et trop spécifique pour mener à bien des études quantitatives et générales sur le développement de l'enfant (Morgenstern & Parisse, 2007), a ouvert la voie aux études de corpus de données recueillies en situation naturelle. En effet, Nadelman (2004) note qu'entre 1890 et 1960, seulement 8% des travaux empiriques sur le développement des enfants et des adolescents étaient basés sur des situations d'observation en situation naturelle (par exemple lors de séances de jeux entre les parents et leur enfant). Par la suite, l'évolution des technologies de recueil de données (comme les magnétophones pour l'audio et les caméras pour la vidéo) a engendré un nombre croissant d'études menées à partir de corpus de productions enfantines. Pour Behrens (2008), les études menées par Brown (1973) représentent l'un des tournants majeurs pour la discipline, au niveau théorique et méthodologique. Au niveau méthodologique, l'utilisation de magnétophones permettant de capter le signal sonore original a poussé les chercheurs à se confronter à de nouvelles problématiques : comment matérialiser le signal sonore de façon pertinente et cohérente ? De quelle façon enrichir les données primaires en ajoutant les informations interprétatives qui seront à la base de la recherche ? Quel format de transcription et d'encodage des données choisir ? En réponse à ces préoccupations, MacWhinney et Snow (1985) et MacWhinney (2000) ont commencé à développer, dès 1983, la base de données CHILDES (CHIld Language Database Exchange System). Cette base de données a pour objectif d'héberger des corpus variés de productions enfantines. Ainsi, on peut retrouver par exemple dans cette base des productions d'enfants monolingues ou bilingues recueillies en situation naturelle ou en situation expérimentale. Les données originales sous format audio ou vidéo sont accompagnées

de la transcription encodée dans un format spécifique, le format CHAT (Codes for the Human Analysis of Transcripts), permettant l'analyse de corpus via le programme CLAN (Computerized Language ANalysis). Cet ensemble d'outils permettant de transcrire, coder/annoter et analyser les données orales est devenu standard dans le domaine. Les aides technologiques de la sorte ont bouleversé le travail du chercheur au niveau méthodologique (Parisse & Morgenstern, 2010a) : une fois que l'outil que l'on souhaite utiliser est maîtrisé, les tâches de transcription et d'annotation deviennent un peu moins fastidieuses et surtout plus rigoureuses. Ainsi, les recueils de corpus longitudinaux se sont multipliés. Pour ce type de corpus, il s'agit d'enregistrer et/ou filmer un même enfant à intervalle régulier pendant une période importante de son développement langagier dans le but « de collecter un échantillon représentatif, mais non exhaustif, du langage de l'enfant et de son développement » (Morgenstern & Parisse, 2007 p.58).

Ces bouleversements méthodologiques ont engendré une multiplication et une diversification des données. Auparavant cantonnés à l'observation de leur propre enfant (ou, dans le meilleur des cas, de l'enfant d'un proche), les chercheurs se sont trouvés face à des données beaucoup plus hétérogènes à plusieurs niveaux. Aujourd'hui, les jeunes locuteurs étudiés sont issus de milieux sociaux divers et contrastés, ce qui a ouvert la voix à des études sociolinguistiques de l'acquisition de certains phénomènes langagiers (Chabanal, 2003). La diversification et la multiplication des sujets observés permettent également une prise en compte de la variation inter-locuteur et intra-locuteur, facilitant ainsi la mise en relief des phénomènes variant et invariant du processus d'acquisition du langage.

En même temps que l'accessibilité à des données variées s'est développée, la communauté a pu observer une multiplication des outils de traitement et d'analyse des corpus[1]. Ainsi, le chercheur souhaitant constituer un corpus de données orales ou multimodales mêlant images et sons se retrouve face à une problématique centrale : quel format d'encodage des données choisir ? Nous retiendrons ici trois critères principaux pouvant guider son choix :

1. L'expressivité du format : le format choisi doit permettre la transcription et l'annotation des données brutes (images et/ou sons) en rapport avec les phénomènes que le chercheur souhaite étudier sans remettre en cause l'expression de phénomènes déjà étudiés.
2. Le caractère standard et extensible du format : le format choisi doit être standard et extensible dans le but de faciliter l'échange et le partage des données au sein de la communauté de chercheurs. L'extensibilité du format permet d'y incorporer la description de nouveaux phénomènes non pris en compte jusqu'à maintenant.
3. L'interopérabilité du format : le format choisi doit faciliter l'interopérabilité entre les logiciels de traitement et d'analyse des corpus.

Si le choix du format ne s'effectue que par rapport au besoin de l'étude d'un ou de plusieurs phénomènes particuliers, tel que perçu en phase de démarrage du projet de recherche sans prendre en compte l'empan temporel important des projets de recherche sur corpus, alors le chercheur aura tendance à choisir une mise en forme dans des formats souvent propriétaires. Ce choix ne permet que rarement de répondre aux critères 2 et 3 précédemment cités qui sont pourtant primordiaux. En effet, favoriser la mise à disposition des corpus/données de recherche reflète un triple enjeu. Le partage des données permet en premier lieu aux autres chercheurs de la communauté de mener des études à partir de ces données, mais cela leur donne également l'occasion de faire des retours sur les analyses menées et d'enrichir les corpus en ajoutant des couches d'annotation. L'utilisation de formats standard permet également au chercheur de valoriser son travail par le référencement de son corpus dans des répertoires internationaux (par exemple OLAC[2]).

[1] On notera CLAN, ELAN, EXMARALDA, Praat, TRANSCRIBER pour ne citer qu'eux.

[2] Open Language Archives Community, www.language-archives.org

3 Le projet ALIPE

Le projet ALIPE est un projet structurant du Laboratoire de Recherche sur le Langage (LRL) qui vise à étudier l'acquisition de la variation phonologique et plus particulièrement les phénomènes de liaison et d'élision. En prenant comme cadre théorique le modèle basé sur l'usage (Kemmer & Barlow, 2000) et son application à l'acquisition du langage (Tomasello, 2003) les objectifs du projet ALIPE s'articulent autour de deux axes de recherche :

- La description et la caractérisation des particularités du discours adressé à l'enfant (DAE) au niveau de la variation phonologique.
- La mise en relation des productions enfantines et des productions parentales dans le but de mesurer l'impact du discours parental sur la vitesse et la qualité d'acquisition de la variation phonologique chez le jeune enfant pré-lecteur.

Alors qu'il est convenu que, comparé au discours adressé à l'adulte (DAA), le DAE comporte des énoncés plus courts (Phillips, 1973), syntaxiquement plus simples (Rondal, 1980) et produits avec « une hauteur tonale élevée et une intonation exagérée » (Jisa & Richaud, 1994, p.22), la littérature fait peu mention des particularités phonologiques du DAE. L'objectif du projet ALIPE est donc de combler ce vide en comparant les productions d'adultes en fonction de l'adresse de leur discours (à l'enfant ou à l'adulte)[3]. À partir de ces données, nous souhaitons également mesurer l'impact des caractéristiques du DAE sur le développement linguistique du jeune locuteur, et plus spécifiquement sur son acquisition des variables phonologiques.

Enfant	Age	Durée totale des enregistrements	Durée des enregistrements transcrits et annotés
Salomé	2;4 ans	8h06	5h
	3;0 ans	6h42	5h
Baptiste	3;0 ans	5h30	5h30
	3;7 ans	4h23	4h23
Prune	3;4 ans	8h34	5h
	4;0 ans	2h02	2h02
	5;4 ans	4h21	4h21

Table 1 : Durées des enregistrements récoltés et transcrits

Dans cet objectif, les recherches menées dans le cadre du projet prennent appui sur des études de corpus constitués de productions recueillies en situation naturelle d'interaction (Corpus-ALIPE). Plus spécifiquement, il s'agit de corpus relativement denses (une heure par jour pendant une semaine) recueillis en deux temps distants de plusieurs mois (T1 et T2). Ce type de corpus comporte plusieurs avantages. En effet, la densité des enregistrements permet d'obtenir un inventaire tout à fait correct des formes et des constructions que l'enfant est capable de

[3] On notera que Andreassen (2011) a en effet comparé le taux d'élision du schwa en DAE, recueilli dans des corpus d'interaction parents-enfant, avec des données de DAA extraite du corpus du projet PFC (Durand et Lyche, 2009).

comprendre et de produire. De plus, le recueil d'autant d'heures d'interaction à un point précis du développement linguistique de l'enfant rend possible l'étude de phénomènes linguistiques relativement rares (Tomasello & Stahl, 2004). Au sein du projet, ce point nous est apparu crucial au regard de la fréquence d'apparition des erreurs enfantines en contextes de liaison catégorique par exemple. Ainsi, en disposant de deux temps de récolte d'enregistrements relativement denses par enfant, nos données sont compatibles avec nos problématiques de recherche et permettent des mesures adéquates de l'évolution des productions de l'enfant et des caractéristiques du DAE entre T1 et T2.

Pour récolter ces corpus, nous avons confié aux parents un enregistreur numérique équipé d'un microphone omnidirectionnel intégré. De cette façon, nous avons minimisé les biais qui auraient pu être engendrés par l'intrusion d'un observateur inconnu de l'enfant à son domicile (Tomasello & Stahl, 2004). La seule consigne donnée aux parents était d'enregistrer leur enfant dans des situations propices aux interactions telles que le bain, le repas ou des séances de jeu et de lecture. À raison d'environ une heure de recueil de données par jour pendant une semaine, nous avions donc à notre disposition un peu moins de 53 heures d'enregistrement[4]. Cependant, les durées des enregistrements variant en fonction de la famille les ayant recueillis, nous avons décidé d'harmoniser les données en ne sélectionnant que 10h de piste audio par enfant pour nos études (cf. Table 1).

4 Annotation des données

La transcription et l'annotation de données orales est une tâche primordiale à laquelle tout chercheur souhaitant constituer un corpus d'interactions naturelles est confrontée. Cette activité se révèle particulièrement coûteuse en temps et/ou en argent. Pour transcrire et annoter une heure d'enregistrement audio, il faut compter jusqu'à vingt heures de travail pour un transcripteur confirmé (Behrens, 2008b ; Parisse & Morgenstern, 2010). Ce temps moyen peut cependant varier en fonction de plusieurs paramètres. Par exemple le chercheur peut, selon les objectifs de sa recherche, utiliser une méthode de transcription plus ou moins détaillée (Delais-Roussarie, 2004). Ainsi, on pourra privilégier la transcription orthographique pour des corpus de grandes tailles sur lesquels on souhaite mener des analyses lexicales sans trop se soucier de la façon dont les formes lexicales ont été prononcées. À l'inverse, le chercheur s'intéressant à des phénomènes phonétiques et/ou acoustiques précis tels que l'accentuation ou le dévoisement optera davantage pour la méthode de transcription acoustique-phonétique. Le nombre de phénomènes linguistiques et d'informations paralinguistiques que le chercheur doit annoter ainsi que leur nature (geste de pointage ou déplacement par exemple) influe également sur le temps que l'annotateur va passer à coder ces informations.

Dans le cadre du projet ALIPE, bien que nous nous intéressions à des phénomènes phonologiques, nous avons choisi de transcrire orthographiquement les productions orales des locuteurs. Ainsi, les analyses lexicales se trouvent facilitées, chaque variante phonologique d'une même forme sous-jacente étant transcrite de la même manière (par exemple, les variantes /mɛtsɛ̃/ et /medəsɛ̃/ sont toutes les deux transcrites « médecin »). Cependant, certains phénomènes phonologiques ont été annotés :

- La liaison : la liaison consiste en la réalisation d'une consonne entre un premier mot (Mot1) et un deuxième à initiale vocalique (Mot2), alors que cette consonne n'est pas réalisée lorsque le Mot1 est produit en isolation (par exemple [dezami] pour « des amis » mais [de] pour « des » et [ami] pour « amis »).
- L'élision variable du schwa : l'élision variable du schwa consiste en l'effacement du schwa dans un contexte où il aurait pu être maintenu (par exemple [ʒprãltrɛ̃] pour

[4] Les données de Salomé et Prune ont été récoltées dans le cadre du projet ANR Phonlex « De la phonologie aux formes lexicales : liaison et cognition en français contemporain ».

[ʒ ə p r ɑ̃ l ə t r ɛ̃] → « je prends le train »).

Principales informations à annoter	**Portée de l'annotation : énoncé entier > partie de l'énoncé > forme lexicale > phonème**	**Type d'annotation : situationnelle, paralinguistique, extralinguistique ou linguistique**
Locuteur	Énoncé	Situationnelle
Adresse du discours		
Chevauchement	Énoncé entier ou partie de l'énoncé	
Type d'énoncé (interrogatif ou exclamatif par exemple)	Énoncé entier	Paralinguistique
Mode de production (énoncé produit en riant, en criant ou en pleurant par exemple)	Énoncé entier ou partie de l'énoncé	
Événement extralinguistique (bruit parasite couvrant la voix du locuteur par exemple)	Énoncé entier ou partie de l'énoncé	Extralinguistique
Liaison	Partie de l'énoncé	Linguistique
Formes spécifiques (forme d'une autre langue ou onomatopée par exemple)	Forme lexicale	
Élision	Phonème	

Table 2 : Principales informations annotées dans le corpus ALIPE

Outre les phénomènes phonologiques étudiés dans le cadre du projet ALIPE, d'autres informations, de diverses natures, nécessitaient une annotation manuelle de la part de l'équipe de recherche (cf. Table 2). La nature et la portée différentes de chacune d'entre elles nécessitaient un format de transcription et d'annotation capable de :

- Représenter correctement la portée de l'annotation en permettant une annotation à un point précis de l'énoncé (pour la liaison par exemple) comme une annotation portant sur une partie de l'énoncé (pour un chevauchement de la parole par exemple).
- Représenter correctement les « enchâssements» d'annotations. Par exemple, un énoncé

entier produit en criant peut contenir une partie d'énoncé se chevauchant avec l'énoncé d'un autre locuteur.

- Permettre la création d'une structure d'annotation. Par exemple, annoter une liaison entre un Mot1 et un Mot2 consiste à annoter plusieurs informations : le contexte syntaxique (entre déterminant et nom ou entre adjectif et nom par exemple), la consonne attendue si la liaison est réalisée, la consonne effectivement produite (ou l'absence de réalisation de la liaison) et le caractère variable ou catégorique de la liaison.
- Permettre une extraction rapide des phénomènes étudiés.

Dans cet objectif, nous avons décidé de transcrire et d'annoter manuellement nos données audio en utilisant le langage de balisage XML (eXtensible Markup Language) et l'éditeur XML Oxygen (SyncRO Soft SRL, 2012). En effet, ce format nous a semblé être le plus performant pour répondre aux exigences des annotations spécifiées plus haut, et ce pour plusieurs raisons. Premièrement, le balisage XML offre deux possibilités particulièrement intéressantes pour l'annotation de corpus : les éléments hiérarchisés et les éléments vides. En langage XML, chaque élément bien formé se compose d'une balise ouvrante, d'un contenu (optionnel) et d'une balise fermante. De plus, un élément peut lui-même contenir un autre élément, entraînant ainsi une relation « père-fils » entre le premier et le second. Ainsi, dans la Figure 1, l'élément « L1 » contient l'élément « crie », qui contient lui-même l'élément « ens1 », le tout formant une structure arborescente basée sur le rapport « père-fils ». Cette caractéristique nous est apparue très intéressante à exploiter. En effet, cette hiérarchisation des éléments nous a permis d'annoter le fait qu'un énoncé produit par le locuteur « L1 » était en partie produit en criant (« prends ton ours »).

Exemple extrait du corpus ALIPE	Signification des balises
<L1> <AE/> <crie> prends <ens1> ton <Ann1/> ours </ens1> </crie> pour l(e) trajet ! </L1>	<L1> </L1> : Ces balises encadrent l'énoncé pour spécifier quel locuteur le produit. <AE/> : Cette balise, en début d'énoncé, indique l'adresse du discours. <crie> </crie> : Ces balises encadrent une partie d'énoncé produit en criant. <ens1> </ens1> : Ces balises encadrent la partie de l'énoncé qui se chevauche avec la production d'un autre locuteur. <Ann1/> : La balise de liaison est placée entre le Mot1 et le Mot2.

Figure 1 : Exemple d'énoncé transcrit et annoté au format XML

Les éléments vides, quant à eux, rendent possible une annotation qui n'englobe pas un énoncé ou une partie d'énoncé mais qui porte sur un point précis de la chaîne parlée. Par exemple la liaison, qui apparaît entre un Mot1 et un Mot2 peut être annotée à l'aide d'un élément vide à l'endroit précis de sa réalisation, comme dans la Figure 1 pour la liaison entre « ton » et « ours ». Dans ce cas précis, la balise peut être considérée comme un codage portant les informations nécessaires à l'étude d'un phénomène particulier. Ainsi, dans le corpus ALIPE, l'annotation de la liaison comporte quatre informations : le contexte syntaxique (par exemple « A », entre déterminant et

nom), la consonne attendue, la consonne réalisée et le caractère variable ou catégorique de la liaison (liaison variable : « 0 » ; liaison catégorique : « 1 »).

Le choix d'utiliser le langage XML s'est donc révélé judicieux pour la transcription et l'annotation de nos données : outre la possibilité de créer des jeux de balises d'annotation permettant de coder des informations générales et spécifiques à notre projet de recherche, le langage XML s'est révélé particulièrement utile dans le but de transformer nos données encodées dans notre propre format (le format XML-ALIPE) en corpus.

5 Mise en forme des données en corpus

La mise en forme des données en corpus est une étape importante pour tout projet de recherche s'appuyant sur des études de données recueillies en situation naturelle. Comme nous l'avons spécifié plus haut, la création d'un corpus libre et accessible à la communauté de chercheurs dans des formats standard se révèle particulièrement importante à plusieurs niveaux. Dans le cadre du projet ALIPE, nous avons retenu deux formats standard pour la mise en forme de nos données en corpus : le format CHAT et le format XML-TEI (Text Encoding Initiative). En effet, ces deux formats nous apparaissent complémentaires.

Format XML-ALIPE

```
<L1>
<AA/> +, je sais pas si c'est <Htt0/> une bonne chose
mais
<rit> de toutes façons </rit>
c'est fait .
</L1>
```

Format CHAT

```
*MOT: +, je sais pas si c'est [^ Syntctx=H expecCons=t
realCons=t obliOpt=0] une bonne chose mais <de toutes façons>
[=! rit] c'est fait . •1465906_1470800•
%add: à FAT
```

Format XML-TEI

```
<u who="#MOT-Prune" xml:id="u637-ali-prune-071121-1">
<anchor synch="u637-ali-prune-071121-1-start"/>
<w>je</w> <w>sais</w> <w>pas</w> <w>si</w> <w>c'est</w>
<fs type="liaison"> <f name="Word1" fVal="c'est"/> <f
name="Word2" fVal="une"/> <f name="SynctacticContext"
fVal="H"/> <f name="ExpectedConsonnant" fVal="t"/> <f
name="ProducedConsonnant" fVal="t"/> <f
name="ObligatoryOptional" fVal="0"/> </fs>
<w>une</w> <w>bonne</w> <w>chose</w> <w>mais</w>
<shift new="laughing"/>
<w>de</w> <w>toutes</w> <w>façons</w> <shift/>
<w>c'est</w> <w>fait</w>
<anchor synch="u637-ali-prune-071121-1-end"/>
<fs type="addressee"><f name="target" fVal="FAT"/></fs>
</u>
```

Figure 2 : Exemple d'un énoncé encodé dans les différents formats utilisés pour la mise en forme des données du projet ALIPE en corpus

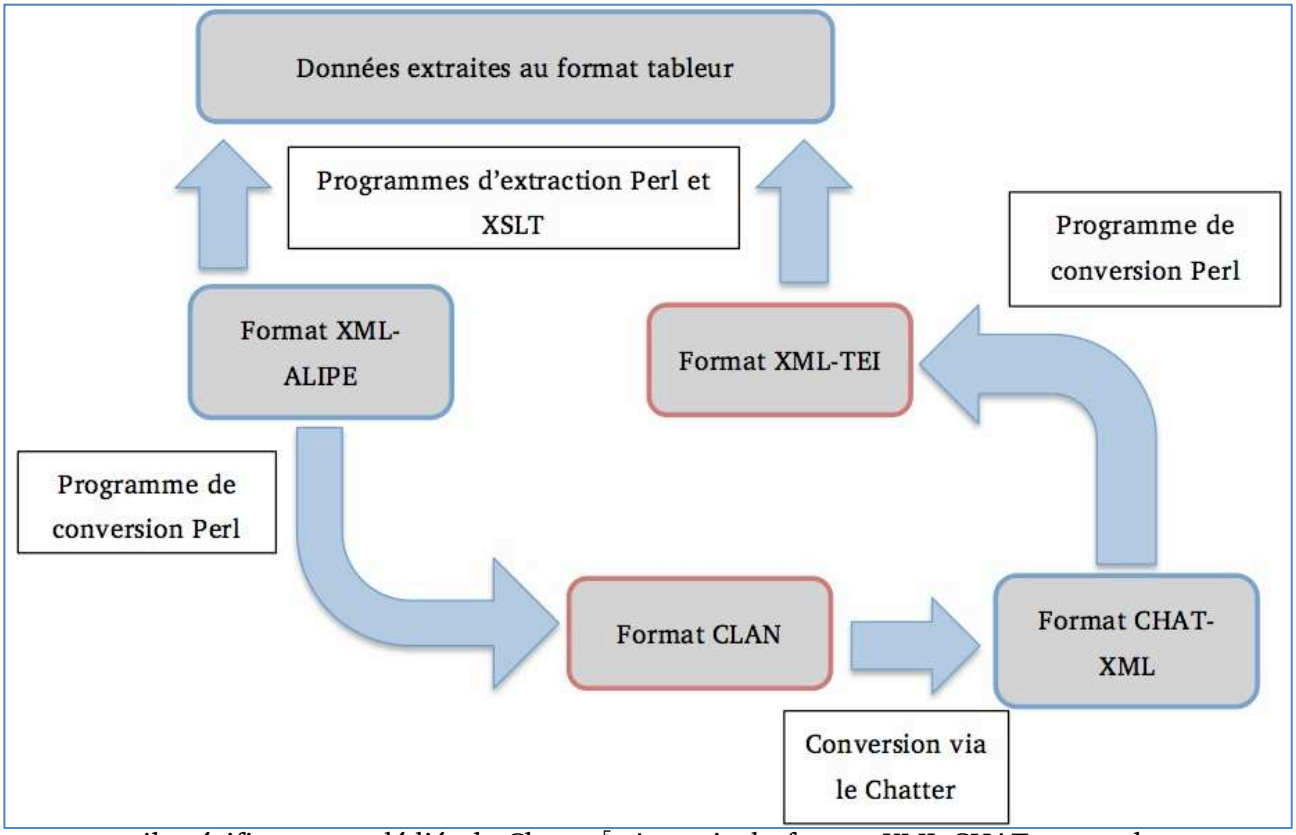
Le format CHAT s'est imposé depuis plusieurs années comme le format standard utilisé pour mettre en forme des corpus à la base d'études en acquisition du langage. En effet, le logiciel CLAN associé à ce format de transcription permet au chercheur de mener un nombre important d'analyses sur une grande quantité de données. En outre, l'interopérabilité du format CHAT nous a semblé particulièrement intéressante : à partir de fonctions d'import/export ou de programmes de conversion, un fichier au format CHAT peut être lu par un autre logiciel de traitement de corpus comme Praat par exemple. Ainsi, l'équipe de recherche se trouve dans la capacité de mener des analyses de différents niveaux (syntaxique, lexical, prosodique) dans différents logiciels et ce à partir des mêmes données, sans travail de mise en forme supplémentaire. Le format XML-TEI, quant à lui, se révèle particulièrement intéressant pour son expressivité et son extensibilité. Au niveau de l'expressivité, le format XML-TEI est un format ouvert, permettant l'ajout de balises spécifiques au projet de recherche. Ainsi, nous avons pu annoter diverses informations autour de la variation phonologique en créant des jeux de balises compatibles avec la grammaire de la TEI. En effet, la grammaire de la TEI propose un système d'éléments hiérarchisés conçus dans le but de permettre toute annotation linguistique. Au préalable, il suffit de décrire la portée et la nature de l'annotation ainsi que la structure de traits utilisés comme argument des éléments dans les métadonnées du corpus. Cette description agira comme une DTD (Document Type Declaration) et rendra compatible le codage créé par le chercheur avec la grammaire de la TEI. C'est en suivant ce procédé qu'à terme un codage mis en place dans un objectif particulier « pourra aboutir à l'intégration dudit codage dans la TEI » (Luzzati, 2009, p.101). En suivant ce protocole, nous avons défini, dans le cadre du projet ALIPE, plusieurs structures d'annotation dont une, par exemple, pour coder l'adresse du discours parental (cf.

Figure 2). Pour le reste de nos annotations, nous avons utilisé les balises disponibles dans la grammaire afin de renseigner le locuteur, l'alignement avec le document sonore et les incertitudes dans la transcription par exemple. Au niveau de l'interopérabilité, le format XML-TEI est amené à devenir le format pivot entre les différents logiciels de traitement de corpus, étant donné son extensibilité et sa capacité à encoder, pour un même énoncé, les particularités de codage des autres formats (Parisse & Morgenstern, 2010b ; Schmidt, 2011).

Afin de transformer nos données transcrites et annotées au format XML-ALIPE en corpus encodés au format CHAT et XML-TEI, nous avons utilisé le langage de programmation Perl (Wall, Christiansen & Orwant, 2001). Perl est un langage de programmation particulièrement adapté aux documents textuels. En effet, ce langage opère sur des chaînes de caractères en différenciant texte et données numériques. Les programmes que nous avons rédigés en langage Perl sont basés sur des expressions régulières et des opérateurs d'expressions régulières permettant par exemple des séries d'opérations de transformation. L'ensemble de la chaîne de traitement suivie par nos données, détaillée ci-dessous, est représenté par la Figure 3.

Dans un premier temps, nos données transcrites et annotées au format XML-ALIPE sont converties au format CHAT via une série de programmes rédigés en langage Perl. Le passage par le format CHAT nous a semblé primordial, et ce pour trois raisons majeures. Premièrement, l'utilisation de ce format nous permet de mener diverses analyses (calcul de la longueur moyenne d'énoncé ou de la diversité lexicale par exemple) à l'aide de l'outil dédié CLAN. De plus, à l'aide de cet outil, nous avons pu réaliser l'alignement de la transcription avec la source de donnée sonore d'une façon simple et efficace. Enfin, cette structuration de nos données nous permettra d'enrichir la base CHILDES avec le dépôt de nos corpus alignés et annotés.

Les transcriptions alignées et annotées au format CHAT sont ensuite converties dans un format XML propre au format CHAT (le format XML-CHAT). Cette conversion est gérée automatiquement



par un outil spécifiquement dédié : le Chatter[5]. A partir du format XML-CHAT, nous obtenons nos corpus encodés au format XML-TEI à l'aide, à nouveau, de programmes de transformation rédigés en langage Perl.

Figure 3 : Chaîne de traitement des données du projet ALIPE

En sortie de notre chaîne de traitement des données, nous avons donc à notre disposition quatre versions de notre corpus. Chacune de ces versions comporte ses particularités. Les corpus aux formats CHAT et XML-TEI sont les versions du corpus qui sont ou qui seront déposées sur des bases en accès libre[6], respectivement sur la plateforme CHILDES et sur le site de diffusion du Laboratoire de Recherche sur le Langage[7]. Le format XML-TEI est un format central encodant l'ensemble des métadonnées du corpus. En effet, alors que dans la plupart des formats propriétaires, le renseignement des métadonnées est succinct et indépendant du fichier de données, le format TEI présente l'avantage de regrouper dans un même élément données et métadonnées. De plus, les balises disponibles dans la grammaire sont nombreuses et nous ont permis de développer plusieurs aspects. Ainsi, outre les informations classiques sur les rôles des chercheurs concernés par le projet ou l'identification des sources de données, nous avons pu encoder diverses informations sociolinguistiques sur les locuteurs : statut socioéconomique, domiciles successifs, âge… Dans les métadonnées, les conditions de récolte des corpus sont également explicitées, accompagnées d'un descriptif du projet de recherche. Ces informations se révèlent essentielles une fois les corpus mis à disposition de la communauté de chercheurs. En effet, elles permettent à la personne qui souhaite utiliser les données de saisir comment et dans

[5] Il s'agit d'un logiciel libre permettant de valider les transcriptions au format CHAT et de les transformer dans le format XML-CHAT. Ce logiciel est disponible sur le site de la TalkBank : http://talkbank.org/software/chatter.html

[6] Au moment de l'écriture, l'équipe de recherche est en train de procéder au dépôt des corpus.

[7] http://lrl-diffusion.univ-bpclermont.fr/

quel objectif celles-ci ont été constituées. Les corpus aux formats XML-ALIPE, non standard, XML-TEI et CHAT sont utilisés par l'équipe de recherche pour les analyses (via le logiciel CLAN) et l'extraction de données qui seront à la base de la recherche.

6 Analyse des corpus

Une fois l'ensemble des traitements sur les données effectué (cf. Figure 3), les chercheurs du projet ALIPE ont à leur disposition trois formats de corpus pour mener à bien leurs analyses : les formats XML-TEI, XML-ALIPE et CHAT. À partir de ces corpus, deux types d'analyses différents vont pouvoir être menés : des analyses sur les corpus, via le logiciel CLAN et des analyses de données extraites des corpus encodés au format XML. Le logiciel CLAN propose une multitude de fonctionnalités allant du simple calcul de fréquence de formes à des calculs de diversité lexicale comme le TTR (Type Token Ratio) ou le VOCD (VOCabulary Diversity, McKee, Malvern, & Richards, 2000). Nous ne détaillerons pas ici la méthodologie d'analyse de corpus via CLAN, celle-ci étant particulièrement bien développée dans le manuel du logiciel (Macwhinney, 2000).

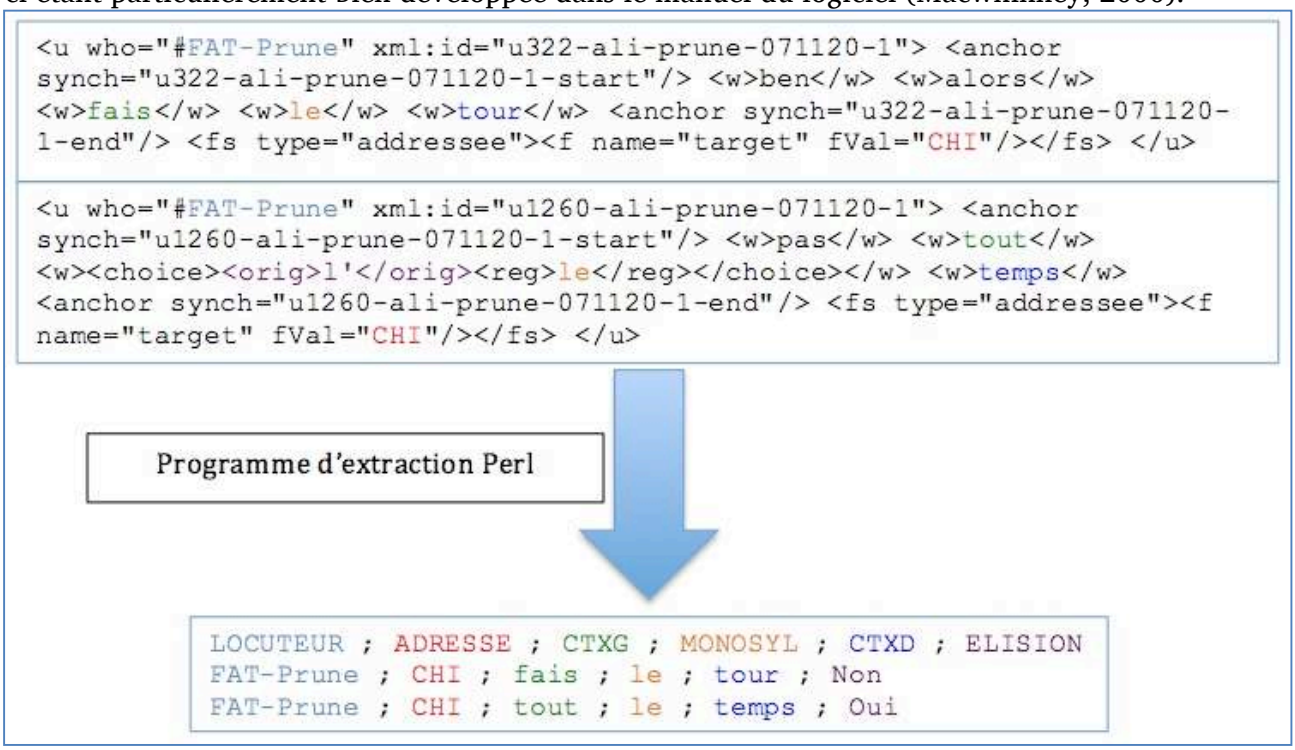


Figure 4 : Illustration du processus d'extraction d'informations du corpus encodé au format XML-TEI

Les corpus encodés au format XML nous offrent d'autres possibilités d'analyse, indispensables pour mener à bien nos recherches sur l'acquisition de la variation phonologique. Afin de faciliter ces analyses, l'un des objectifs de l'équipe de recherche était d'avoir à disposition des données sous format de tableur permettant ainsi l'analyse au moyen de logiciels de traitements statistiques tel que R. Dans cet objectif, nous avons développé une série de programmes rédigés en langage Perl et XSLT permettant d'extraire les informations nécessaires à nos études. Dans le cas, par exemple, d'une comparaison du comportement du schwa dans les monosyllabiques[8] en DAA et en DAE, plusieurs informations doivent être extraites du document XML, et ce pour chaque contexte

[8] En français, les monosyllabiques sujets à l'élision variable du schwa sont représentés par la classe fermée des clitiques *ce*, *de*, *je*, *le*, *me*, *ne*, *se*, *te* et *que*.

d'élision ou de maintien variable du schwa :

- Le locuteur
- L'adresse du discours
- Le monosyllabique concerné
- Les contextes gauche et droit
- L'élision ou le maintien du schwa

Pour extraire ces informations, le programme rédigé en langage Perl, par exemple, parcourt linéairement la suite de caractère du corpus. À chaque fois que celui-ci rencontre une occurrence d'un des clitiques étudiés, les informations nécessaires sont sélectionnées puis extraites dans un fichier texte de sortie (cf. Figure 4).

A partir du fichier texte de sortie, il est assez simple d'intégrer ces données à un tableur classique ou à un logiciel d'analyse statistique plus spécifique. Par exemple, nous avons voulu savoir si, au niveau de l'élision variable du schwa dans les monosyllabiques, les parents de Prune modulaient leur langage en fonction de l'adresse de leur discours. En effet, les premiers travaux du projet ALIPE ayant fait apparaître une tendance au maintien du schwa variable (Liégeois, Saddour & Chabanal, 2012) et à la réalisation de la liaison variable (Liégeois, Chabanal & Chanier, 2011), nous avons souhaité observer si les données issues des productions des parents de Prune corroboraient ces résultats. Le tableau ci-dessous résume les résultats obtenus pour les deux périodes de récolte des données (T1 et T2) :

Temps de récolte	**Taux d'élision en discours adressé à l'enfant**	**Taux d'élision en discours adressé à l'adulte**
T1	31,1% (286/918)	67,9% (178/262)
T2	34,8% (337/969)	87,4% (581/665)

Table 3 : Taux d'élision relevés dans les productions des parents de Prune en fonction de l'adresse du discours

Les résultats obtenus à partir des données des parents de Prune extraites de nos corpus corroborent ceux obtenus précédemment. En effet, les parents de Prune semblent moduler leur langage en fonction de l'adresse du discours. Au T1, alors que le taux d'élision variable dans les monosyllabiques en discours adressé à l'enfant est relativement faible (31,1%), celui-ci est plus de deux fois plus élevé en DAA (67,9%). Cette nette différence au T1 se révèle significative au regard du test du Chi2 de conformité ($Chi2 = 114{,}053$; $p < 0{,}0001$) tout comme au T2 ($Chi2 = 440{,}9166$; $p < 0{,}0001$).

7 Conclusion

La constitution d'un corpus de données langagières spontanées est un travail coûteux en temps (et donc en argent) mais aujourd'hui quasiment indispensable lorsque l'on souhaite étudier l'acquisition du langage. Au niveau national et international, les projets de bases de corpus (comme CHILDES par exemple) et de répertoires indexant des corpus (comme CLARIN par exemple), témoignent de l'engouement actuel autour du travail sur corpus, indifféremment de la discipline ou de la sous-discipline. Comme nous l'avons montré, l'évolution des technologies, entre autres, a engendré une redéfinition de l'objet corpus. Celui-ci peut aujourd'hui être défini sous la forme d'un paradigme comportant les quatre points suivants (Chanier & Ciekanski, 2010) :

- Le recueil systématique des documents liés à l'objet d'étude, en prenant en compte la

couverture et la taille des données recueillies.
- L'organisation et l'instrumentalisation en vue de traitements, qui consiste à rendre le corpus utilisable par d'autres équipes de recherche et analysable par d'autres outils que ceux initialement considérés lors de l'élaboration du projet de recherche.
- La description du contexte, qui regroupe par exemple les informations sur la situation d'énonciation ainsi que les méthodes de recueil.
- Les dispositions en vue de l'échange et du partage du corpus. Celles-ci doivent être prises dans l'optique d'un dépôt en accès libre du corpus.

La méthodologie mise en place pour la constitution du corpus ALIPE, par son encodage dans des formats standard, nous semble répondre à ces enjeux. Comme nous l'avons souligné, l'utilisation du langage de balisage XML pour la transcription et l'annotation de nos données se révèle particulièrement utile à deux niveaux. Premièrement, l'extraction d'informations en vue des analyses se trouve simplifiée grâce à l'utilisation de langages de programmation comme Perl ou XSLT. De plus, le format XML facilite grandement la dérivation du corpus vers différents formats, permettant ainsi une meilleure interopérabilité entre les différents logiciels de traitement des corpus. Ces dérivations nous ont permis d'obtenir deux versions du corpus ALIPE au format CHAT et XML-TEI. Ces formats choisis pour mettre en forme les données en corpus sont libres, expressifs et extensibles. En outre, le format XML-TEI permet de regrouper dans un même objet corpus les données et les métadonnées. Ces dernières sont primordiales et font référence aux deux derniers points du paradigme corpus. En effet, les métadonnées comportent l'ensemble des informations permettant de situer le corpus et de le définir comme un objet scientifique. On retrouve ainsi la description de la méthode, du contexte et de l'objectif du recueil des données, les structures d'annotation utilisées mais également les informations légales régissant la libre circulation du corpus dans la communauté de chercheurs.

Si mettre en forme des données recueillies en situation naturelle d'interaction dans des formats standard tels que les formats CHAT et XML-TEI est long et parfois fastidieux, cette perte de temps initiale peut être rapidement compensée, et ce à plusieurs niveaux. Premièrement, la possibilité de pouvoir utiliser à posteriori un outil de traitement de corpus spécifique par simple dérivation automatique du format initial du corpus peut se révéler utile. Cette tâche de dérivation, qui peut être automatique entre certains programmes (comme entre CLAN et Praat), se trouve grandement facilitée par l'utilisation du langage XML. L'utilisation d'un format standard peut également permettre une réutilisation ou un enrichissement du corpus de la part d'une autre équipe de recherche. Dans cette optique, le format XML-TEI apparaît le plus à même pour représenter, à l'intérieur d'un même objet, une grande diversité de couches d'annotation reliées chacune à la définition du système d'annotation dans les métadonnées. Enfin, l'encodage des corpus dans un format standard rend possible leur dépôt dans une base de corpus ouverte, facilitant l'échange des données au sein de la communauté et la reconnaissance du travail de l'équipe ayant constitué la ressource.

Remerciements

Merci à Inès Saddour pour sa participation à la transcription et à l'annotation des données.

Références

BARLOW, M. et KEMMER, S., (dir.) (2000). *Usage Based Models of Language*. Stanford California: CSLI Publications.

BEHRENS, H. (dir.) (2008a). *Corpora in Language Acquisition Research : History, methods, perspectives*. Amsterdam: John Benjamins Publishing Company.

BEHRENS, H. (2008b). Corpora in language acquisition research. History, methods, perspectives. In (Behrens, 2008a), pp. XI–XXX.

BOERSMA, P. et WEENINK, D. (2009). *Praat: doing phonetics by computer* (Version 5.3.23) [Computer program]. http://www.praat.org/

BROWN, R. (1973). *A First Language: The Early Stages*. Cambridge: Harvard University Press.

CHABANAL, D. (2003). *Un aspect de l'acquisition du français oral : la variation sociophonétique chez l'enfant francophone*. Université Paul-Valéry - Montpellier 3.

CHANIER, T. et CIEKANSKI, M. (2010). Utilité du partage des corpus pour l'analyse des interactions en ligne en situation d'apprentissage : un exemple d'approche méthodologique autour d'une base de corpus d'apprentissage. *Alsic*, 13, Para. 2. doi:10.4000/alsic.1666

CHANIER, T., LIEGEOIS, L., CHABANAL, D. et LOTIN, P. (2012). *Projet Acquisition de la Liaison et Interactions Parents-Enfant*. Laboratoire de Recherche sur le Langage. Clermont Université. [http://lrl-diffusion.univ-bpclermont.fr/alipe]

DELAIS-ROUSSARIE, E. (2004). Constitution et annotation de corpus : Méthode et Recommandations. In Delais-Roussarie, E. et Durand, J. (dir.), *Corpus et Variation en Phonologie : Méthodes et Analyses*. Toulouse : Presse Universitaire du Mirail. pp. 89-126.

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

JISA, H. et RICHAUD, F. (1994). Quelques sources de variation chez les enfants. *Acquisition et Interaction en Langue Étrangère*, 4, pp. 5–51.

KEMMER, S. et BARLOW, M. (2000). Introduction: A usage-based conception of language. In (Barlow, M. & Kemmer, S., 2000). pp. VII–XXVIII.

LIEGEOIS, L., CHABANAL, D., et CHANIER, T. (2011). La liaison en discours adressé à l'enfant, spécificités et impacts sur l'acquisition. Communication au *Colloque du Réseau Français de Phonologie*, Tours (1-3 juillet 2011).

LIEGEOIS, L., SADDOUR, I. et CHABANAL, D. (2012). L'élision du schwa dans les interactions parents-enfant : étude de corpus. In Besacier, L., Lecouteux, B. et Sérasset, G. (dir.), *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*. Grenoble: ATALA & AFCP. pp. 313–320.

LUZZATI, D. (2009). Corpus d'hier et d'aujourd'hui : progrès quantitatifs ou progrès qualitatifs ? *Cahier de linguistique*, 32(2). pp. 97–112.

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

MACWHINNEY, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah: Lawrence Erlbaum Associates.

MCKEE, G., MALVERN, D. et RICHARDS, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing*, 15(3). pp. 323–338. doi:10.1093/llc/15.3.323

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

NADELMAN, L. (2004). *Research manual in child development*. Mahwah: Lawrence Erlbaum Associates.

PARISSE, C. et MORGENSTERN, A. (2010a). Transcrire et analyser les corpus d'interactions adulte-enfant. In J. Bernicot, A. Bert-Erboul, M. Musiol, & E. Veneziano (dir.), *Interactions verbales et acquisition du langage*. Paris : L'Harmattan. pp. 201–222.

PARISSE, C. et MORGENSTERN, A. (2010b). A multi-software integration platform and support for multimedia transcripts of language. *LREC 2010 : Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*. La Valette.

PHILLIPS, J. R. (1973). Syntax and vocabulary of mothers' speech to young children: Age and sex

comparisons. *Child Development*, 44(1). pp. 182–185.

REFFAY, C., BETBEDER, M. L. et CHANIER, T. (2012). Multimodal learning and teaching corpora exchange: lessons learned in five years by the Mulce project. *International Journal of Technology Enhanced Learning*, 4(1/2). pp. 11–30.

RONDAL, J. A. (1980). Father's and mothers' speech in early language development. *Journal of Child Language*, 7(2). pp. 353–369.

SCHMIDT, T. (2011). A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*, 1(1). doi:10.4000/jtei.142

SyncRO Soft SRL (2012). Oxygen XML Editor version 14.0. Craiova: Syncro Softsrl. [http://www.oxygenxml.com/]

TAINE, M. (1877). M. Taine on the Acquisition of Language by Children. *Mind*, 2(6). pp. 252–259.

TEI CONSORTIUM (dir). (2012). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.1.0. Last modified 17th June 2012. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (29/08/2012).

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

TOMASELLO, M. et STAHL, D. (2004). Sampling childrens spontaneous speech: how much is enough ? *Journal of Child Language*, 31(1). pp. 101–121. doi:10.1017/S0305000903005944

WALL, L., CHRISTIANSEN, T. et ORWANT, J. (2001). *Programmation en Perl* (3ème éd.). O'Reilly.