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) LRL, 4, rue Ledru, 63057 Clermont-Ferrand Cedex1

loic.liegeois@univ-bpclermont-fr

Contexte

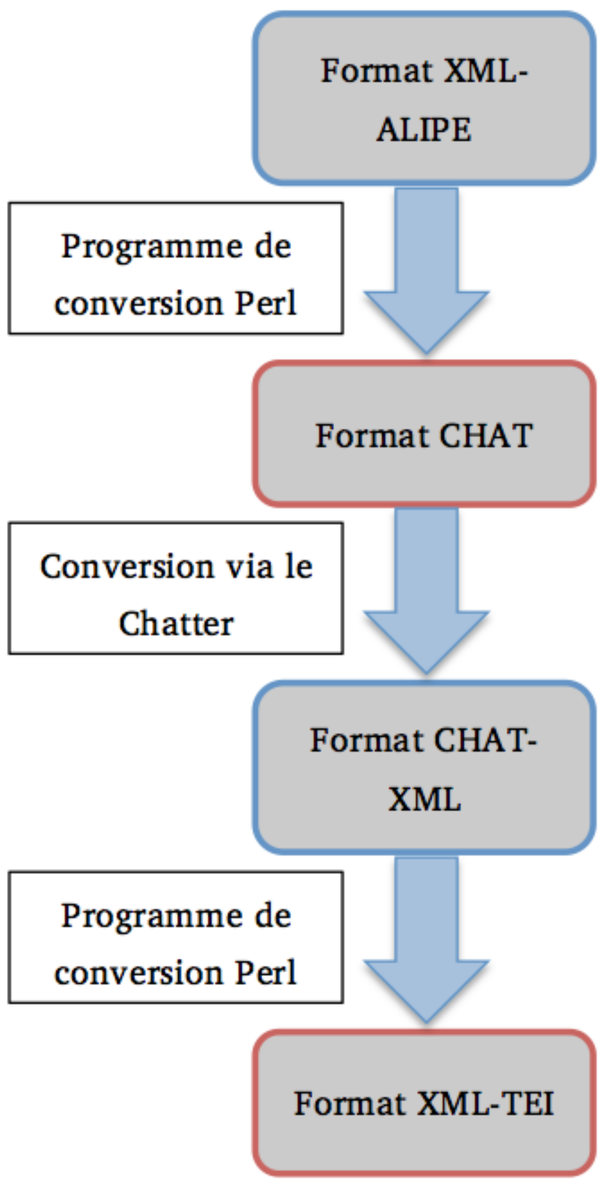
- Mise en place d'infrastructures et de projets nationaux (Très Grand Equipement ADONIS, Très Grande Infrastructure de Recherche Corpus) et internationaux (Common Language Resources and Technology Infrastructure) qui témoigne de l'envergure sociale et scientifique grandissante des corpus de données langagières.
- 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 (Behrens, 2008).
- L'objet corpus, un paradigme comportant quatre points (Chanier et Ciekanski, 2010) :
  - Recueil de documents avec prise en compte de la couverture et de la taille des données
  - Organisation des données dans le but de rendre le corpus utilisable par d'autres (interopérabilité)
  - Description des contextes (métadonnées)
  - Dépôt en vue de l'échange et du partage du corpus

Problématique

- Parmi les problématiques liées à tout projet de constitution d'un corpus de données langagières, celle concernant le choix du format de structuration est centrale.
- Quel(s) format(s) choisir ? Pour répondre à cette question, nous avons retenu trois critères de sélection dans le cadre du projet ALIPE (Acquisition de la Liaison et Interactions Parents-Enfant :
  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 analyser, 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.

Méthodologie de traitement des corpus du projet ALIPE

Figure 1 : Chaîne de traitement des données du projet ALIPE



Etape 1 : Transcription et annotation des données au format XML-ALIPE. Définition d'une structure d'annotation au moyen de balises XML. Annotation spécifique de l'adresse du discours, de la liaison et de l'élision.

Etape 2 : Conversion du format XML-ALIPE vers le format CLAN. Réalisation de l'alignement de la transcription avec le document sonore. Possibilité de mener des analyses spécifiques (calcul de diversité lexicale et calcul de longueur moyenne d'énoncé par exemple) via le logiciel CLAN.

Etape 3 : Conversion automatique du format CHAT vers le format CHAT-XML via le Chatter (outil disponible sur http://talkbank.org/software/chatter.html).

Etape 4 : Conversion du format CHAT-XML vers le format XML-TEI. Possibilité d'extraire des informations via des programmes en langage Perl et XSLT (cf. **Figure 2**).

Tableau 1 : Durées des enregistrements récoltés et transcrits

Enfant	Age	Durée totale des enregistrements	Durée des enregistrements transcrits
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	6h44	5h

Figure 2 : Extraction d'informations annotées dans le corpus au format XML-TEI : l'exemple de l'élision du schwa

```
<u who="#FAT-Prune" xml:id="u322-ali-prune-071120-1"> <anchor
synch="u322-ali-prune-071120-1-start"/> <w>ben</w> <w>alors</w>
<w>fais</w> <w>le</w> <w>tour</w> <anchor synch="u322-ali-prune-071120-
1-end"/> <fs type="addressee"><f name="target" fVal="CHI"/></fs> </u>
```

```
<u who="#FAT-Prune" xml:id="u1260-ali-prune-071120-1"> <anchor
synch="u1260-ali-prune-071120-1-start"/> <w>pas</w> <w>tout</w>
<w><choice><orig>l'</orig><reg>le</reg></choice></w> <w>temps</w>
<anchor synch="u1260-ali-prune-071120-1-end"/> <fs type="addressee"><f
name="target" fVal="CHI"/></fs> </u>
```

Programme d'extraction Perl

```
LOCUTEUR ; ADRESSE ; CTXG ; MONOSYL ; CTXD ; ELISION
FAT-Prune ; CHI ; fais ; le ; tour ; Non
FAT-Prune ; CHI ; tout ; le ; temps ; Oui
```

Conclusions

❖ Atouts de la structure XML :
- Permet 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).
- Permet de 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.
- Permet la création d'une structure d'annotation.
- Permet une extraction rapide des phénomènes annotés.
- Permet une conversion vers d'autres formats de structuration (dans notre cas, CHAT et XML-TEI).

❖ Atouts du format CHAT :
- Permet de mener des analyses via le logiciel spécialisé CLAN.
- Permet de déposer les corpus dans la banque CHILDES.
- Standard du domaine.

❖ Atouts du format XML-TEI :
- Permet d'encoder dans un même fichier données et métadonnées détaillées.
- Format standard, amené à devenir un format « pivot » entre les différents formats étant donné son extensibilité et sa capacité à encoder, pour un même énoncé, les particularités de codage des autres formats (Parisse et Morgenstern, 2010 ; Schmidt, 2011)

Bibliographie

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

CHABANAL, D., LIÉGEOIS, L. ET CHANIER, T. (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]

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

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

PARISSE, C., et MORGENSTERN, A. (2010). 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.

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

TEI CONSORTIUM, eds. *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).