Méthodologie de corpus pour l'étude des interactions dans les écoles bilingues

Christophe PARISSE
MoDyCo - INSERM, CNRS, Université Paris Ouest Nanterre la Défense

Introduction

L'apprentissage et l'enseignement scolaire d'une langue, qu'elle soit première ou seconde, se déroulent dans un contexte langagier très riche : l'école et en particulier la salle de classe. On y trouve bien sûr les élèves, l'enseignant, mais aussi un environnement : la situation de la classe, le nombre d'élèves, et le matériel pédagogique à disposition des enfants.

Décrire toute cette situation scolaire peut se faire à travers la simple analyse directe du déroulement de la classe, ou à travers une retranscription et une analyse fine de ce qui est dit, qu'on peut obtenir par exemple à l'aide d'un enregistrement sonore. Néanmoins, quelle que soit l'indéniable qualité des nombreuses études qui ont été menées de cette manière, il est possible aujourd'hui d'aller encore plus loin, dans l'analyse comme dans la retransmission fidèle du travail fait à l'école bilingue.

Ceci peut se faire grâce à l'usage du film et en particulier de la vidéo. Aujourd'hui, et les technologies ne font que s'améliorer pour rendre les choses de plus en plus faciles, il est possible d'enregistrer à l'aide d'une caméra numérique, ou même d'un téléphone de bonne qualité, et d'obtenir instantanément ou presque un film disponible sur ordinateur pour le travail scientifique. Cet usage du film permet un travail nouveau qui restait difficile autrefois. Veut-on par exemple tenir compte des hésitations des élèves, de la manière dont ils s'expriment devant le tableau, de la manière dont l'enseignant s'adresse à eux, compter le nombre d'élèves qui essaient de répondre en même temps à une question, apprécier les gestes des enseignants et des élèves, tout ceci ne peut se faire sans un accès à l'image qu'offre le film. De plus l'usage du film améliore le travail sur le son : on peut être sûr de savoir qui parle, on peut s'aider dans la transcription des gestes co-verbaux ou de la forme d'articulation de la bouche. Dans de nombreux contextes, l'image permettra de plus facilement deviner ou contrôler ce qui est parfois difficile à entendre.

Ainsi, aujourd'hui, les corpus de langage ne sont pas uniquement constitués de textes écrits ou de transcription de ce qui est dit. Tous les corpus, en particulier les corpus oraux, mais aussi les corpus écrits, sont accompagnés de données d'une très grande variété qui sont nécessaire pour l'étude de ces

corpus. Par exemple, pour un corpus de langage oral, on pourra avoir une transcription phonologique, phonétique, un enregistrement sonore et/ou vidéo, une description de la situation, des éléments d'analyse linguistique, des données accompagnant le recueil et les travaux (photos, documents papiers, etc.).

L'utilisation de média électronique et de transcription langagière permet aussi une diffusion et une mise en commun des différentes sources scientifiques. En effet, le coût de la transcription complète d'un corpus est très important. On considère ainsi qu'il faut entre 30 et 60 heures de travail pour transcrire une heure de corpus (le nombre d'heures variant de façon importante selon la nature des informations codées et la densité sonore du corpus). Il faut ajouter les besoins nécessaires pour le recueil des données, qui peut être coûteux lorsqu'il nécessite un important travail de terrain. Pour un corpus comme PFC (Phonologie du Français Contemporain : Laks, Durand, et Lyche, 2009) on arrive ainsi à un euro du mot, ce qui pour des corpus de taille importante (50 000 mots pour un corpus standard, 1 million de mots pour un gros corpus), représente de très grosses sommes. Il faut donc absolument partager ces données pour gérer au mieux l'argent de la recherche.

Enfin, le partage de corpus permet aussi des travaux scientifiques impossibles à réaliser à l'aide d'un seul corpus. Par exemple, toutes les études d'ampleur importante portant sur la typologie des langues ne peuvent se faire sans un partage de données très variées (Evans et Levinson, 2009). Egalement, en dehors de ses avantages scientifiques, l'usage du film offre une mémoire de la culture qui est sauvegardée dans le support numérique. Il permet la diffusion à un large public, scientifique, enseignant, ou autre. On peut donc accéder, au-delà du travail scientifique, à une dimension pédagogique et culturelle.

Le recueil, la diffusion et l'utilisation des corpus

Il existe à travers le monde de nombreux centres historiques de recueil et de diffusion de corpus. L'un d'eux, dont nous nous sommes inspirés et donc nous avons utilisé les outils pour le projet Transferts, est le projet CHILDES. Il s'agit d'un projet américain, porté par l'université de Carnegie Mellon à Pittsburgh. CHILDES (Child Language Data Exchange System, voir http://childes.psy.cmu.edu) est un système d'échange et de description du langage de l'enfant mis sur pied en 1984 par Brian MacWhinney et Catherine Snow (Université Carnegie Mellon de Pittsburgh, USA). Ce projet vise à stocker, sauvegarder et diffuser des corpus à but scientifique portant sur l'acquisition du langage, en langue première comme en langue seconde, et

ceci dans toutes les langues (plus de 130 corpus correspondant à plus de 30 langues). Une des originalités de ce projet est que comme il porte en particulier sur l'étude des très jeunes enfants, l'usage de l'enregistrement audio d'abord puis de la vidéo s'est généralisé depuis de nombreuses années.

La difficulté de la transcription

L'étude du langage du jeune enfant permet particulièrement bien de se rendre compte de la subjectivité de toute situation langagière (Morgenstern & Parisse, 2007). En effet lorsque les enfants disent leurs premiers mots, il est souvent difficile de les comprendre hors de leur contexte de production. La situation permet de deviner ce que l'enfant cherche à dire, même si l'enfant ne prononce pas encore correctement ou complétement ce qu'il produit. Cette difficulté pour l'interlocuteur de l'enfant est la même que pour le linguiste qui étudie le langage de l'enfant. Le linguiste doit lui aussi utiliser le contexte ou tout autre connaissance pour décrypter ce que l'enfant dit.

Ces difficultés sont les mêmes pour la compréhension de plus grands enfants ainsi que d'adultes, même si les cas de grande difficulté de compréhension sont plus rares. Dans le contexte d'une classe, ceci est particulièrement vrai. Sans savoir exactement ce qui se passe dans une classe, à quel moment on en est d'une leçon, quel est le thème enseigné ou même de quel élève il s'agit, la compréhension et donc l'analyse qui suit peuvent être difficiles.

Comme il n'est pas toujours possible d'être présent le jour de l'enregistrement, ou comme on ne peut pas tout noter et tout remarquer lors de la collecte de données, l'utilisation de la vidéo est indispensable, et ce principe a été largement utilisé pour le projet Transferts. Le détail de la procédure de recueil et de codage est présenté ci-dessous.

Diffusion des corpus

La diffusion des corpus est rendue aisée grâce au stockage des données numériques. Non seulement il est possible de télécharger à distance des transcriptions ou des vidéos, mais aussi il est possible de visualiser en direct, à travers Internet, une transcription et la vidéo associée.

La diffusion des corpus impose une standardisation des formats et de la nature des transcriptions. Cette standardisation permet la réutilisation des corpus avec d'autres outils et par d'autres personnes. Elle permet aussi leur visualisation sur le site.

Enfin elle permet d'utiliser des outils sur une transcription, sur un ensemble de transcriptions, sur un corpus complet ou même sur l'ensemble des corpus.

Les exemples d'outils sont très nombreux. Les plus courants sont les suivants :

- Recherche de mots, de suites de mots, ou de motifs quelconque dans les corpus. Les versions avancées des outils de recherche permettent de limiter la recherche à certains types de locuteurs, certains champs linguistiques, etc.
- Création du lexique de toute une transcription, tout un corpus, pour un locuteur, pour un ensemble de locuteurs, pour une langue, etc. On pourra faire des recherches sur ce lexique, l'éditer, l'imprimer.
- Traitements du son par l'intermédiaire d'outils spécifiques comme Praat (2001, 2014), Winpitch (Martin, 2014). Ces outils permettent de réaliser de l'analyse très fine du signal sonore (spectrogrammes, formants, analyse du pitch, traitements du signal, etc.).
- Traitements textométriques de l'ensemble d'un corpus ou sous-corpus avec des outils comme TXM (Heiden, 2014 ; Heiden & al., 2010) ou Le Trameur (Fleury, 2014 ; Fleury & Zimina, 2014). Ces outils permettent des traitements syntaxiques, une analyse statistique de l'ensemble des corpus, de tous les usages de tous les mots. On peut par exemple mettre en valeur l'importance d'un usage par rapport à un autre, suivre l'évolution des usages au cours du temps. Ces logiciels disposent d'importantes possibilités de visualisation des données.

Application au projet Transferts d'apprentissage

Outils vidéo et audio

L'ensemble des recueils de données du projet Transferts a été réalisé avec des caméras numériques Canon permettant de stocker plusieurs heures de film sur un support mémoire SDHC. La principale difficulté technique est celle de l'autonomie limitée des caméras, liée aux limites de capacité des batteries. La réalisation d'un tournage de plus d'une heure nécessite plusieurs batteries ou un accès à l'électricité sur secteur. Dans ces conditions et en fonction du nombre limité de caméras (une par pays) que pouvait financer le projet, l'utilisation de moyens complémentaires appartenant à chaque pays a été parfois nécessaire. Autant que possible, les enregistrements vidéo ont été doublés en audio pour garantir la qualité du son. Dans la pratique, la qualité de micro des caméras modernes et l'aide donnée par le support de l'image ont permis de ne pas rendre l'exploitation de l'audio nécessaire pendant le projet.

Chaque enregistrement vidéo est stocké dans sa qualité originale (format MTS MPEG2 pour nos caméras) pour conservation sur ordinateur. Ensuite les enregistrements sont convertis (utilisation du logiciel Emicsoft) dans un format numérique compatible avec leur utilisation dans un logiciel de transcription. Les formats Quicktime MOV et MP4 H264/AAC sont deux des choix qui ont été utilisés. Les films peuvent être disponibles en deux formats, un grand format haute qualité pour la diffusion en public et un petit format basse qualité pour l'utilisation sur un ordinateur personnel pour la transcription.

Les films ont parfois fait l'objet de montage (utilisation du logiciel Emicsoft) pour rassembler des séances de classe qui peuvent avoir été dispersées sur plusieurs enregistrements selon les contraintes effectives des enseignants qui ont été filmés. Une description exhaustive des médias a été réalisée. Cette description porte sur les films montés qui sont utilisés pour l'étude mais conserve la référence aux films originaux. Les descriptions contiennent les champs qui permettent d'identifier les données (noms physiques des films) et les champs suivants : Pays, Date, Localité, Langue 1, Type, Année de scolarité, Langue d'enseignement, Durée, Matière/Domaine, Activité.

Outils de transcription

La transcription d'un enregistrement vidéo ou audio peut s'accompagner aujourd'hui d'un alignement parfait avec le média. Ceci veut dire qu'à tout énoncé de la transcription correspond un moment précis dans le média sonore ou visuel. Ce point est repéré par le logiciel et peut être joué instantanément à tout moment. Les transcriptions et les alignements sont aussi précis qu'on le désire. Il est possible de faire des retours en arrière, de réécouter, de découper. Les transcriptions peuvent être modifiées autant que l'on désire. Enfin il est possible de les transmettre et de les corriger aisément à plusieurs.

L'outil de transcription qui a été choisi est CLAN (Computerized Language Analysis : MacWhinney, 2014), celui du projet CHILDES (MacWhinney, 2000). Il présente, par rapport à d'autres outils, l'avantage de permettre une utilisation scientifique directe des transcriptions. Il représente aussi un standard connu et reconnu dans le domaine des corpus.

La transcription dans CLAN exige de respecter les conventions de transcription correspondant au format CHAT (Codes for the Human Analysis of Transcripts). L'avantage du respect de ces conventions, en dehors des possibilités de partage de données qui sont alors offertes, est de rendre les données compatibles avec l'utilisation des outils de CLAN (voir ci-dessous).

Principes de base

CLAN divise les données transcrites en trois niveaux. Chaque niveau correspond à un caractère de début de ligne différent dans un fichier de transcription. Chacun des caractères de tête de ligne est suivi d'un mot clé qui décrit la nature de la ligne transcrite. Enfin, après ce mot qui se termine par le signe « : » suit la transcription ou le codage des données transcrites.

- Lignes commençant par « @ » : données descriptives de l'ensemble d'un fichier ou d'une sous-partie du fichier.
- Lignes commençant par « * » : transcription orthographique. Chaque ligne correspond à un énoncé.
- Lignes commençant par « % » : codage complémentaire de la transcription orthographique.

Exemple d'extrait d'une transcription (leçon en langue première – songhay – cours de lecture, classe de deuxième année)

@Begin

@Languages: son. fra

@Participants: ELV Child, MTR Teacher, ELV-MTR Group

@ID: son. fra|video-première-sortie-gao-bon-format|ELV|||||Child|||

@ID: son. fra|video-premiere-sortie-gao-bon-format|MTR|||||Teacher|||

@ID: son. fra|video-premiere-sortie-gao-bon-format|ELV-MTR|||||Group|||

@Media: Mson-A2-lec-L1-171011. video

@G: Debut

…

*ELV: ka senni hantumanta yan caw kan ga waani waani

%gls: lecture de syllabes hétérogènes

*MTR: may no ma kaa ka caw ya ne ?

%gls: qui va lire?

*ELV: agay agay

%gls: moi moi

*MTR: he@i wa kay ay ga sintin war se wa haŋa jer!

%gls: hé! arrêtez je vais commencer la lecture

*MTR: a a a

*ELV: a a a

*MTR: ne na na.

*ELV: ne na na.

*MTR: e ne nee.

*ELV: e ne nee

*MTR: o no na.

*ELV: o no na

…

@End

On voit dans cet exemple des lignes commençant par « @ » qui désignent l'ensemble du fichier. Par exemple, la première ligne @ID décrit les caractéristiques de l'interlocuteur correspondant à un élève. La ligne @Media donne le nom du fichier vidéo associé à la transcritption.

Les lignes commençant par des « * » désignent les interlocuteurs de cette session : le maître « *MTR » et un élève « *ELV ». Les transcriptions sont dans la convention orthographique standard du songhay.

Enfin les lignes commençant par des « %gls » sont des éléments complémentaires des lignes « * » qui les précèdent. Elles contiennent ici la glose en français.

Exemple d'extrait d'une transcription (leçon en langue seconde – Fulfude langue première – cours d'arithmétique, classe de cinquième année)

*MTR: +, soulève jusqu'ici (.) viens par là [/] viens par là .

*MTR: [- ful] ga .

%gls: ici.

%ac: inter|phra|crea|MTR

*MTR: +, voila .

*ELV: deux mille +//.

%act: l'élève vient au tableau et écrit en chiffres deux mille.

*MTR: deux mille francs (.) mais tu as fait comment (.) tu as dit mille plus mille non .

*MTR: +, donc ça fait deux mille c'est bien .

*MTR: ok (.) ceux qui ont trouvé (.) montrez .

*ELV: 0 .

%xpnt: les élèves soulèvent leurs brouillons et présentent au maître leurs travaux.

*MTR: +, ok (.) montrez tout simplement .

*MTR: +, un deux trois quatre +...

%com: MTR compte le nombre de ceux qui ont trouvé.

On voit dans ce nouvel exemple une variété plus importante de codages secondaires. On retrouve l'exemple de la glose (%gls) mais on trouve aussi un codage des alternances codiques (%ac). Cet élément est spécifique du projet Transferts (voir ci-dessous). On voit un exemple de « %act » qui correspond à une description de l'action en cours, un exemple de « %xpnt » qui permet de décrire les pointages effectués, et un exemple de « %com » qui permet d'ajouter des commentaires libres.

Les codes de transcription de CHAT permettent un codage complet de toute une séries de détails linguistiques. Le codage est décrit de manière exhaustive dans la documentation de CHILDES. Un complément de description des conventions spécifiques du projet Transferts est disponible à l'adresse Internet http://modyco.inist.fr/transferts/pdf/conventions-transferts.pdf
Les principaux éléments de codage sont les énoncés, la manière dont ils peuvent être interrompus ou commencés. On peut coder les citations, les indications de mots incompréhensibles ou que l'on ne peut pas transcrire, les éléments phonétiques, les intonations, les mots composés, les élisions, les reprises, les pauses, etc.

Ajustements spécifiques du projet Transferts
Dans le cadre du projet Transferts, certains des éléments qui figurent dans les possibilités du système CHILDES ont été ajustés pour les besoins du projet. Nous avons en particulier utilisé les possibilités de codage multilingue. Il est en effet possible d'indiquer de manière précise dans quelle langue chaque mot est produit.

Le codage de la langue commence par la description des langues « un » et « deux » qui sont utilisées dans le document. Par exemple, on a :

@Languages: fra, ful

qui indique le français (fra) en langue « un » et le fulfulde (ful) en langue « deux ». Attention, il ne s'agit pas du français langue première (c'est-à-dire langue nationale) mais du français langue principale de l'enregistrement. Ici c'est tout à fait normal car la session est une session de géométrie de cinquième année au Burkina-Faso, année durant laquelle la proportion de

français dans l'enseignement est nettement plus importante que la proportion de fulfulde (langue nationale des enfants).

Inversement on trouvera dans une leçon de grammaire de deuxième année au Mali :

@Languages: son, fra

qui indique le songhay (son) en langue « un » et le français (fra) en langue « deux ».

Dans le codage de CLAN, on considéra alors que tous les énoncés qui sont codés sans indication supplémentaire sont des énoncés dans la langue « un ». Si un mot de l'énoncé n'est pas en langue « un » mais en langue « deux », on indiquera le mot en question par le suffixe « @s ».

@Languages: fra, ful

*MTR: vingt+mille+francs foti@s ?

%gls: vingt mille francs égale combien?

Ici dans cet exemple, foti n'est pas en français mais en fulfulde et donc est indiqué par « @s » : foti@s.

Pour les énoncés qui ne seraient pas en langue « un », on pourra indiquer quelle est la langue de l'énoncé pour ne pas utiliser @s sur tous les mots. Par exemple :

@Languages: fra, ful

*ELV: moi [/] moi monsieur [=! par quelques élèves] .

*ELV: carré .

*MTR: [- ful] njaajeefi carré@s wiete ?

%gls: est-ce que les largeurs se dit carré?

%ac: inter|phra|suit|reqt|MTR

%com: MTR contredit la répoonse de l'élève.

*ELV: moi [/] moi monsieur [=! par quelques élèves] .

L'énoncé du maître (*MTR) est presque complétement en fulfulde et c'est indiqué par « [- ful] » en début d'énoncé. Enfin, un des mots de cet énoncé (carré) n'est pas en fulfulde et donc cette exception s'indique avec un « @s ». Ce codage peut paraître complexe mais il comprend plusieurs avantages. Dans les cas d'ambiguïté (et notamment pour les mots d'emprunt) on indiquera de manière explicite la langue. De ce fait, le codage pourra être

évalué et prêter à discussion. Par ailleurs, il est possible lors de l'utilisation des outils (voir ci-dessous) de préciser la langue. On pourra ainsi constituer un lexique complet d'une langue particulière, ou faire des recherches sur tous les mots d'une langue précise.

Conventions de transcription spécifiques du projet Transferts

Pour le projet Transferts, comme pour tout autre projet de linguistique, il est possible d'utiliser le logiciel de transcription pour faciliter l'analyse des corpus. Ceci peut se faire en exploitant les possibilités des codages des lignes secondaires et des outils d'interrogation de corpus. Ainsi dans l'exemple ci-dessus, on trouve un exemple de codage spécifique qui concerne l'alternance codique (cf. contribution d'Inoussa Guiré dans ce volume).

%ac: inter|phra|suit|MTR

1. « inter » correspond à interphrastique, selon la typologie de Poplack (1980). Dans cette position on peut aussi coder « intra » (intraphrastique) et « extra » (extraphrastique)
2. « phra » correspond à la nature de l'élément inséré : « phra » pour phrase, « synt » pour syntagme, « lexi » pour élément lexical
3. « suit » correspond à la formulation de l'énoncé. De nombreuses situations sont possibles ici, comme « suit » pour appartient à une succession d'alternances codiques, « crea » pour formulation à l'initiative du locuteur, « imit » pour reprise par imitation d'un interlocuteur, « modi » reprise du tour précédent avec modification de la forme produite, etc.
4. « MTR » indique la personne qui produit l'alternance codique : « MTR » pour maître, « ELV » pour élève.

L'utilisation de codes variés dans chacune des quatre positions permettra de faire la liste des éléments codés ou de faire des requêtes sur une partie ou sur l'ensemble du corpus.

Dans un autre exemple ci-dessous, on utilisera les lignes %m pour coder les phénomènes métalangagiers (cf. contribution de Zakaria Nounta dans ce volume).

*MTR: mayno ma hin ka har ir se koyraboro cennira ?

%gls: qui peut le traduire en songhay?

%m: MEQ MGI MTC

*ELV: madame agay madame agay madame agay!

%gls: moi madame moi madame moi madame

%m: GEAS GGIS GTRS

Dans les deux exemples ci-dessus, le codage des transcriptions est réalisé en contexte. Il est possible de contrôler l'image et le son et garantir que le

codage fait sera précis et exact. Une fois codés de cette manière, tous les éléments pourront être récupérés en une seule opération (voir ci-dessous) et intégrés dans un traitement de texte ou un tableur.

Outils de traitement des corpus

Un des buts de CLAN, en dehors de la transcription des textes, est de permettre l'exploitation scientifique des corpus. Pour cela, CLAN dispose d'un ensemble de commandes permettant l'édition des corpus, comme le remplacement d'un mot par un autre, la recherche de tous les mots inconnus (ce qui permet de repérer les hapax et les fautes de frappe), la suppression de certaines lignes secondaires, le codage rapide de certaines lignes en fonction d'un vocabulaire précis.

En dehors des commandes d'édition, CLAN dispose de deux commandes permettant une analyse des corpus. La première commande, FREQ, permet de récupérer automatiquement le lexique intégral d'un corpus avec le nombre d'apparitions de chaque mot (d'où le nom de FREQ pour « fréquence »). On peut limiter cette recherche de lexique à un certain locuteur, à certains fichiers, à certains motifs. On peut aussi n'extraire les listes de fréquences que pour certains codages secondaires. Ainsi, en utilisant les codages complémentaires ci-dessus, on peut extraire tous les codages métalinguistiques pour le maître, pour l'élève, etc. On peut limiter cette recherche aux énoncés d'une certaine longueur, d'une certaine langue, etc. Les résultats de cette commande peuvent être intégrés dans un traitement de texte ou dans un tableur.

Extrait d'une sortie de la commande FREQ pour une transcription de calcul en première année pour la langue nationale fulfulde :

```
 2 sukaabe@s:ful
 5 ta@s:ful
 3 tableau@s:fra
 6 tablo@s:ful
 1 tan@s:ful
45 tati@s:ful
 1 timmina@s:ful
 1 timminta@s:ful
 1 timmore@s:ful
 3 tinna@s:ful
 5 to@s:ful
```

Exemple de sortie graphique du logiciel CLAN

La deuxième commande, COMBO, permet de rechercher des motifs sur tout un ensemble de fichiers. Cette commande permet d'afficher les lignes résultats ainsi que plusieurs lignes autour si nécessaire. Il est possible de rechercher des mots complets, des parties de mots, des suites de plusieurs mots, etc. Par exemple, pour le même enregistrement que pour l'exemple précédent, on cherche le mot « tableau ». Le résultat se présente ainsi :

```
From file <Bful-A1-calc-L1-211111.cha>
----------------------------------------
*** File "Bful-A1-calc-L1-211111.cha": line 1146.
*MTR:  [- fra] au (1)tableau .
----------------------------------------
*** File "Bful-A1-calc-L1-211111.cha": line 1172.
*MTR:  [- fra] Mousa [//] Hamma Mousa (.) au (1)tableau <jannginoowo@s> [<] .
----------------------------------------
*** File "Bful-A1-calc-L1-211111.cha": line 1892.
*MTR:  +, njehe njoodowe (.) waarɗo ga fu m(i) wi'ay mo o wara (1)tableau@s
       +...
  Strings matched 3 times
```

Exemple de sortie graphique du logiciel CLAN

On peut voir qu'effectivement le mot tableau apparaît trois fois, à chaque fois prononcé par le maître, et une fois dans une alternance codique.
Les deux commandes FREQ et COMBO permettent de récupérer rapidement les résultats nécessaires à un travail scientifique de qualité.

Exemple d'utilisation

Dans les deux exemples de codage présentés ci-dessus, l'alternance codique et le métalangage, il est nécessaire d'étudier en détail tous les codages réalisés puis de démontrer, autant que possible, l'existence de différences entre telle et telle population, ou telle et telle situation.
Par exemple, on va vouloir savoir si l'on trouve plus d'alternances codiques chez les enseignants que chez les élèves, et comment leur utilisation change avec la situation (par exemple selon la discipline enseignée) ou selon l'année scolaire (les élèves débutants produisent peut-être plus d'alternances que les élèves confirmés).

Pour opposer enseignants et élèves, il suffit d'utiliser une fois la commande FREQ pour toutes les alternances qui ont le codage « MTR » et une fois la même commande mais pour les alternances qui ont le codage « ELT ». On obtiendra deux listes d'occurrences, que l'on pourra compter (le logiciel le fait automatiquement) ou qu'on l'on pourra présenter sur deux colonnes de tableau dans un fichier Microsoft® Excel.
Si l'on veut observer l'évolution des alternances codiques en fonction des années scolaires, on pourra lancer la commande FREQ séparément pour les transcriptions des séquences de première année, puis de deuxième année, et ainsi de suite. Ou on pourra tirer parti du fait que la commande FREQ peut produire des statistiques, soit globales sur l'ensemble des fichiers que l'on lui donne en paramètres, soit fichier par fichier directement. Cette forme fichier par fichier peut être, encore une fois, directement insérée dans un fichier Microsoft® Excel.
On pourra aussi croiser les recherches. Par exemple on pourra chercher tous les cas où on a un codage d'alternance interphrastique (code « inter ») dans lesquels il y a une création (code « crea ») ou une modification (code « modi »). Pour cela il faut chercher tous les codes « inter » suivis du code « crea » (sans que les codes soient contigus), ou suivis du code « modi ». Cette recherche de deux (ou plus) éléments non contigus fait partie des possibilités de CLAN. Les possibilités dans ce domaine sont presque infinies et peuvent couvrir de nombreuses questions de recherche.

Conclusion

Dans les projets récents de linguistique, il est devenu indispensable de disposer de corpus pour pouvoir présenter de manière explicite les faits linguistiques et démontrer les résultats que l'on défend. Pour cela, il faut recueillir avec soin les données, les transcrire, les analyser, mais aussi les stocker de manière partageable et pérenne. Sinon, comment un résultat pourrait être vérifié par la suite, ou même amélioré au cas où.
Egalement, la conservation des corpus permet leur réutilisation pour des recherches futures. L'investissement du projet Transferts est en effet trop précieux pour être dispersé sans une bonne conservation des données.
Pour cela, en dehors du stockage sur notre site, de l'utilisation d'outils permettant le partage et l'étude qualitative précise, nous envisageons de conserver les données dans les instituts des différents pays impliqués, comme par exemple le service Ortolang pour la partie française, ainsi que dans les sites des institutions qui ont financé le projet (AUF, OIF).
Toutes ces conditions, qualité du recueil de données, qualité des outils et des études, qualité de la diffusion des corpus, pourront être réunies dans le projet Transferts.

Références bibliographiques

Boersma Paul (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

Boersma Paul & Weenink David (2014). Praat: doing phonetics by computer [Computer program]. Version 5.3.84, téléchargée le 26 août 2014 depuis http://www.praat.org/

Evans, Nicholas, & Levinson, Stephen C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(05), 429–448. doi:10.1017/S0140525X0999094X

Serge Fleury and Maria Zimina (29014) . Trameur: A Framework for Annotated Text Corpora Exploration, *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*: System Demonstrations, August 2014, Dublin, Ireland, pages 57-61.

Fleury, Serge (2014). Le Trameur : [Computer program]. Version 11.78, téléchargée le 10 septembre 2014 depuis http://www.tal.univ-paris3.fr/trameur/

Heiden, Serge, Magué, Jean-Philippe, Pincemin, Bénédicte (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In I. C. Sergio Bolasco (Ed.), *Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010* (Vol. 2, p. 1021-1032). Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy.

Heiden, Serge (2014). TXM : plateforme modulaire et open-source [Computer program]. Version 0.7.6, téléchargée le 21 juillet 2014 depuis http://textometrie.ens-lyon.fr/

Laks, Bernard, Durand, Jacques, & Lyche, Chantal. (2009). Le projet PFC (Phonologie du Français Contemporain) : une source de données primaires structurées. In *Phonologie, variation et accents du français*. (pp. 19–6). Paris : Hermès.

MacWhinney, Brian (2000). *The CHILDES project : Tools for analyzing talk* (3rd ed.). Hillsdale, N.J: Lawrence Erlbaum.

MacWhinney, Brian (2014). Clan: Computerized Language ANalysis [Computer program]. Version 27-Aug-2014, téléchargée le 27 août 2014 depuis http://childes.psy.cmu.edu/

Martin, Philippe (2014). Winpitch [Computer program]. Version 1.00 May 8 2013, téléchargée le 8 mai 2013 depuis http://www. winpitch.com/

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

Poplack Shana (1980). Sometimes I'll start a sentence in Spanish y termino en español: towards a typology of code-switching, *Linguistics*, 18, pp. 581-618.