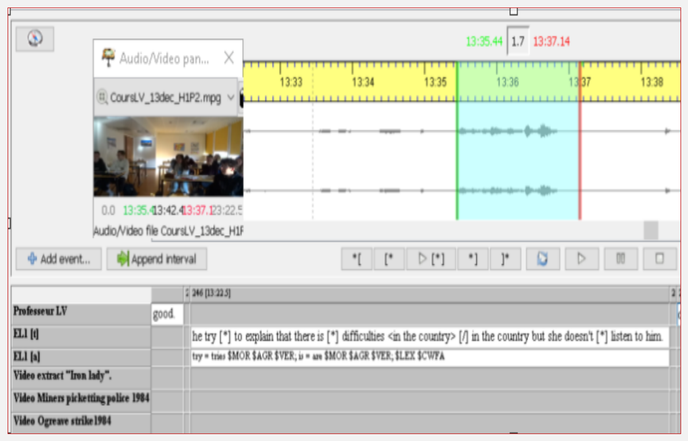
Corpus multimodal des apprenants en EMILE : constitution, traitements, outils

Evgenia Nicol-Bakaldina

Constituer et analyser un corpus multimodal EMILE : quels enjeux méthodologiques ?



Transcription et codage dans EXMARaLDA

Il est important de trouver un compromis raisonnable entre des exigences spécifiques des logiciels TAL et des conventions communément employées (ex.CHILDES), tout en prenant en compte les avantages et les limites des outils de transcription et d'analyse des données.

CONTEXTE DE RECHERCHE : ETUDE DE CAS

La contribution se situe dans le cadre plus large d'un projet d'observation d'une classe EMILE sur un an dans un contexte de lycée français. 16 heures de cours (histoire-géographie en L2 : anglais) ont été enregistrés ; 280 documents (supports des professeurs et productions des élèves) ont été récoltés, puis transcrites (avec EXMARaLDA). Le corpus multimodal (écrit+oral) de 119.000 mots a été constitué et analysé. Le travail de recherche étudie dans quelle mesure un statut particulier de la langue en cours EMILE (à la fois un objet et un outil d'apprentissage) modifie les conditions et les résultats de l'apprentissage.

METHODOLOGIE

Le fonctionnement de l'interlangue des élèves, ainsi que l'interaction entre *l'input* et *l'output* sont étudiés par le biais de l'observation des cours EMILE. Les composantes linguistiques du corpus sont analysées avec : Exmaralda, CLAN, Hyperbase, SketchEngine, UAM Corpus Tool, AntConc.

CONCLUSION

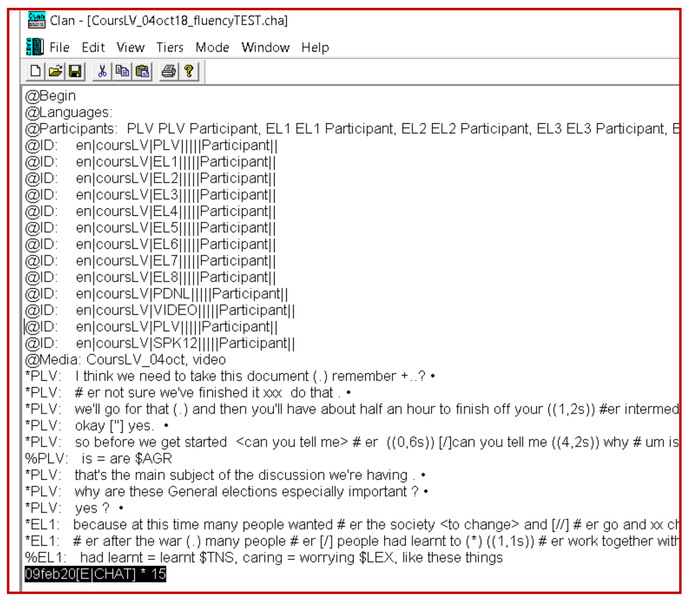
La compilation du corpus doit obéir aux règles suivantes :

- être **représentative** du langage des apprenants (= représenter des phénomènes linguistiques ; Rastier, 2004) ;
- **correspondre aux normes** langagières et aux exigences des conventions communes (CHILDES) ;
- être **systématique et régulière** (codage) afin de permettre des prolongements possibles d'exploration (Granger, 2015).

Constitution et codage du corpus : défis méthodologiques.

	Transcription des données EXMARaLDA	Codage pendant la transcription EXMARaLDA	Nettoyage et standardisation WORD	Codage post- transcription pour le traitement dans AntConc, CLAN
Constitution du corpus	• Enregistrements, documents, productions écrites des élèves. • Phénomènes méta- et supralinguistiques (toux, gestes, rires) ne sont pas pris en compte. • Transcription du contenu textuel uniquement. • Création de deux couches par locuteur : transcription et annotation.	Productions des élèves : • Corpus oral : disfluences principales. • Corpus écrit : orthographe. • Corpus écrit et oral : lexique, morphologie, syntaxe.	Semi-automatique : • Conversion des chiffres en lettres : ***15 = fifteen*** • Abréviations et orthographe acceptées internationalement : ***PM = Prime Minister*** • Variations de langue ignorées ***color = colour***	•Erreurs doublées de corrections ***not to <invaded*> invade*** •Mots non-terminés restitués: ***<contribu*> contribution*** •Données non-pertinentes cachées -mots en L1 ***<guerre>***, -données métalinguistiques: ***<I don't know how to say>***
Défis	• Processus très chronophage, vérification manuelle s'impose. • Que transcrire ? • Quelles ponctuation pour transcrire des données orales (une virgule, un point, un point-virgule) ? • Qualité des enregistrements et du son, segmentation des énoncés ont un impact sur le calcul des erreurs. ***it's the Cold war right.*** OU: ***it's the Cold war. right?***	• Codage systématique, homogène, claire. • Prise en compte des limites du logiciel. • Codage dans les deux couches influe sur le calcul du nombre des erreurs : ***the politic* situation*** 1) ***politic*0*** (mot non-terminé) 2) $MOR $SUF $LOS	• Vérification manuelle s'impose pendant la standardisation automatique : **1950s** → ***nineteen fiftyS**** au lieu de **1950s** → ***nineteen fiftIES*** • Eviter le « bruit » : Liens, sites Internet, marqueurs de structuration (puces, numérotation).	• Comment coder les mots en grammaire non normée ? ***Gonna, yea***, etc. • Quelle déviance considerer comme « erreur » ? • Comment coder en respectant les normes des conventions (CHILDES) ? • Quelle catégorie à attribuer lors du codage ? • Double entrée (erreurs + corrections) => incidence sur la taille du corpus

CODAGE DANS CLAN



1. Les métadonnées doivent figurer dans un ordre précis.
2. La ponctuation est obligatoire à la fin de chaque énoncé, mais n'est pas possible à l'intérieur de celui-ci.
3. Les majuscules ne doivent pas être utilisées, sinon le logiciel *CLAN* classe ces mots comme des noms propres.

Problème : *It's one of the five giants, remember ?*
Solution : *it's one of the five giants. remember ?*

Segmentation d'un énoncé ayant des fonctions syntaxiques multiples

(1) *# er after the war (.) many people # er [/] people had learnt to (*) ((1,1s)) # er work together without caring about (*) (.) social classes or genders or all these things*
(2) ***so*** *they wanted the government to ((1,7s)) go deeper in their life* ***and*** *to ((1,0s)) have more impact on (.) health <or or or> [/] or work, or <like these things> (*)*
(3) ***and so*** *# er it's important because it's the first time that the Labour Party is (*) elected* ***and*** *it will be able to (.) [/] to answer to <what they> [//] # er what they wanted.*

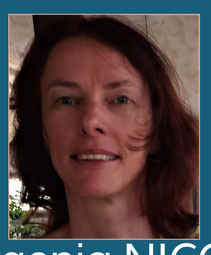
THESE DE DOCTORAT

L'enseignement d'une matière par intégration d'une langue étrangère (E.M.I.L.E) en France : le rôle et l'utilisation de la langue à l'intersection entre deux disciplines dans l'enseignement secondaire.

Autrice : Evgenia Nicol-Bakaldina

Laboratoire LLSETI : Langages, Littératures, Sociétés. Etudes Transfrontalières et Internationales, l'École Doctorale : Cultures Sociétés Territoires

UNIVERSITÉ SAVOIE MONT BLANC



Evgenia NICOL-BAKALDINA
eve.nicol83@gmail.com