Mélanges CRAPEL n° 44/2

# CLAN : OUTIL DE TRANSCRIPTION ET DE TRAITEMENT DES DONNÉES MULTIMODALES

**Evgenia Bakaldina-Nicol**

Université Savoie Mont Blanc

**Mots-clés**

annotation – corpus – CLAN – linguistique – TAL

**Keywords**

annotation – corpus – CLAN – linguistics – NLP software

**Résumé**

Depuis quelques décennies les champs d'application des logiciels de traitement automatique de langue (TAL) s'étendent considérablement dans les domaines de la linguistique de corpus et la didactique des langues. Cet article met en lumière le logiciel CLAN. Premièrement, une présentation générale du logiciel est donnée, la question de la transcription des données et du codage d'erreurs à l'aide de CLAN est également abordée. Nous décrivons ensuite les problèmes auxquels le chercheur peut se confronter lors de la transcription des données dans CLAN. Les fonctionnalités saillantes du logiciel sont présentées, en particulier, les fonctions de calcul relatives à l'analyse morpho-syntaxique et lexicale de l'interlangue des apprenants. L'article se termine par un bilan des avantages et inconvénients de CLAN.

**Abstract**

Over the last few decades, the application range of natural language processing (NLP) software has expanded considerably in corpus linguistics and language teaching. This article focuses on CLAN. First, the software's general presentation is provided, along with data transcription and error coding routines using CLAN. We then describe the issues researchers may encounter when transcribing data using CLAN. The software's main features are equally addressed, in particular the computational functions for morpho-syntactic and lexical analysis of learners' interlanguage. The article concludes with a review of CLAN's advantages and disadvantages.

## Introduction

La présente contribution se situe dans un cadre plus large du domaine de Traitement Automatique des Langues (TAL). L'objet de cet article est de présenter un outil de transcription, de codage et d'analyse des données multimodales *CLAN*. Dans une première partie nous effectuons une présentation générale de *CLAN*, logiciel dont le champ d'application est très vaste au sein de la linguistique de corpus. L'article se poursuit avec une la question de la transcription des données et du codage d'évènements et d'erreurs au moyen de *CLAN* est également abordée. A la fin de cette première partie nous décrivons quelques difficultés auxquelles le chercheur peut se confronter lors de la transcription des données dans *CLAN*.

Dans une deuxième partie, nous examinons de plus près la question de traitement d'un corpus de linguistique à l'aide de *CLAN* dans un cadre de l'application des logiciels TAL dans la recherche. Notamment, nous mettons en lumière quelques fonctions saillantes de calcul relatives à l'analyse morpho-syntaxique et lexicale de l'interlangue des apprenants afin d'établir leur profil langagier. Parmi ces fonctions, les programmes *KWAL, EVAL, FLUCALC* et *MOR* sont explorés ; les analyses de la diversité et de la densité lexicales des discours sont présentées, entre autres. De nombreuses exemples tirés d'une étude de cas[1] corroborent cette présentation.

La contribution se termine avec la description des avantages et des inconvénients de *CLAN* en guise de conclusion.

## 1. Logiciels de transcription, d'analyse et de présentation des données multimodales

Quels outils de transcription et d'analyse choisir ? Chaque logiciel a des avantages fonctionnels, le choix doit être opéré en fonction de son utilité vis-à-vis du travail à effectuer (pour la liste des logiciels les plus connus en linguistique de corpus voir Bakaldina-Nicol, 2023).

En fonction des formats des fichiers (des métadonnées, des fichiers média et de transcriptions) qui peuvent varier, certains logiciels seront plus adaptés que d'autres pour but du codage et du traitement ultérieure d'un corpus. Par exemple, *EXMARaLDA, CLAN* et

1 Les exemples et les captures d'écran pour but de démonstration, sauf l'indication contraire, sont tirés de la recherche doctorale (l'étude de cas) menée à l'Université Savoie Mont-Blanc (2018-2023). Titre : « Les défis de l'enseignement d'une matière par intégration d'une langue étrangère (EMILE) en France : le rôle et le fonctionnement de la langue à l'intersection de deux disciplines en école secondaire. » 16 heures de cours EMILE ont été enregistrées lors de cette étude, ensuite transcrites et analysées. CLAN a été utilisé pour coder et analyser des données discursives orales récoltées.

*UAM Corpus Tool* sont prônés par des chercheurs pour des raisons suivantes : (1) leur disponibilité sans frais, téléchargeables en ligne et accompagnés d'un manuel détaillé et approfondi ; (2) leur compatibilité avec le système Windows ; (3) l'interopérabilité des formats qui permet un transfert facile des données si besoin est ; (4) la configuration possible en tenant compte des exigences d'autres logiciels[2].

## 2. Présentation du logiciel CLAN

***C**omputerized **L**anguage **AN**alysis* – est un outil développé par Leonid Spektor à l'Université Carnegie Mellon University[3]. Ce dispositif est conçu dans le but d'observer l'apprentissage du langage. *CLAN* est un logiciel de transcription et d'analyse des données multimodales, dans lequel les transcriptions sont effectuées en conformité avec le format *CHAT* (Ratner and Brundage, 2016, p. 2). *CHAT* et *CLAN* font partie du *CHILDES* (*CHild Language Data Exchange System*), système doté des outils pour but d'analyse des interactions discursives et qui sert comme une convention de codage des corpus reconnue un niveau mondial (MacWhinney, 2000, p.8). *CHAT* est un logiciel qui permet la transcription des fichiers sonores grâce aux règles préétablies connues sous le nom « *CHAT format »*.

Comme le précise un tutoriel de Saul Albert (Université de Loughborough), *CLAN* est utile pour les transcriptions de conversations « à tour de rôles » (*turn by turn transcription*)[4]. Le logiciel est aussi utilisé pour but d'exploration de l'interlangue des enfants. Plus précisément, *CLAN* permet d'effectuer les analyses ciblées en fonction de besoin grâce au calcul automatique des mesures, telles que : la longueur moyenne des énoncés ; le rapport entre le nombre des mots et leur type (*TTR*) ; les parties de discours et les morphèmes (en pourcentages) ; les morphèmes de Brown (pour étudier le développement langagier des enfants) ; la fluence et la disfluence dans les discours ; la rapidité et la densité des discours, etc. ; lemmatiser et analyser les énonces morphologique et syntaxique.

*CLAN* propose un champ très vaste des explorations des discours, allant des fonctions basiques comme *FLUCALC* (calcul du nombre de disfluences) ou *KWAL* (recherche d'un mot clé dans un contexte immédiat ou éloigné) jusqu'à des fonctions avancées comme *MLU run* (calcul de la proportion du nombre : morphèmes/énoncé), ou *COMBO* (séquences ciblées de recherche, ex. trouver l'infinitif dans tous les fichiers, etc.). Notre contribution se

2 Par exemple, la segmentation des transcriptions dans EXMARaLDA peut s'effectuer en tenant compte de la convention de transcription CHILDES afin de rendre les données compatibles avec le format CHAT opéré dans CLAN.

3 https://talkbank.org/manuals/CLAN.pdf , consulté le 18 mars 2020, logiciel téléchargé le 19 mars 2020 sur le site https://talkbank.org.

4 https://www.youtube.com/watch?v=pLmow45dyBQ consulté le 19/07/2020.

bornera à la présentation des fonctions de calculs, de lemmatisation et d'analyse morpho-syntaxique *KWAL, FLUCALC, EVAL, MOR* (voir la section 3.1.).

Le côté pratique de *CLAN* consiste en sa compatibilité avec d'autres logiciels de transcription, ce qui permet de profiter des avantages de chacun d'entre eux selon les besoins (ex. effectuer une analyse morphologique avec *CLAN,* coder les erreurs avec *Exmaralda*, etc.) sans la nécessité de transcrire à nouveau les conversations (ce qui est très chronophage).

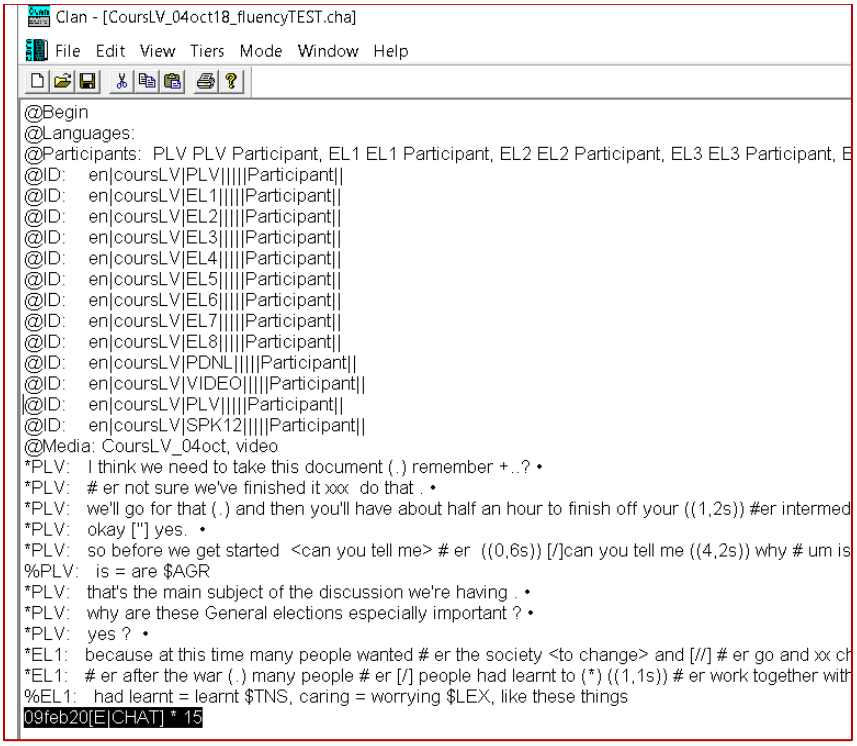



**Figure 1. Interface du logiciel CLAN**

## 2.1. Transcription et codage des événements et des erreurs dans CLAN

Le manuel *CHILDES* (MacWhinney, 2000, p.55-56) propose un protocole à suivre avant de se lancer dans un travail chronophage d'analyse de multitude de données. Ces préconisations sont destinées aux chercheurs n'ayant pas beaucoup d'expérience avec *CHAT* et *CLAN*. L'idée est d'économiser l'effort en suivant les étapes suivantes :

- entrer une petite quantité de données dans le fichier *CHAT* ;
- appliquer la fonction *CHECK* dans l'éditeur afin de vérifier la conformité de la syntaxe au format *CHAT* ;
- développer une série des codes à intégrer dans un panel de codage en lien avec les codes déjà existant de CLAN en fonction des analyses envisagées.

La transcription s'effectue dans la ligne principale attribuée à un participant (*tiers*) : *(*EL:* pour désigner un élève ; **PLV :* pour désigner un professeur de langue, etc.) Des éléments paralinguistiques (ex. rire, tousse ; tout sort de commentaire, comme « lit, chante ») peuvent être annotés soit dans la même ligne principale au moyen des chevrons et des parenthèses carrées *(*EL1: <The British Prime Minister> [=! lit]*), soit en les rapportant à la ligne secondaire en fonction du but de l'analyse (*%com :* pour des commentaires paralinguistiques*; %EL :* pour apporter des corrections d'erreurs) :

Ex. *%EL1: had learnt = learnt $TNS*

Mentionnons quelques particularités à prendre en compte lors de la transcription[5] et codage des données dans CLAN :

- les données méta-linguistiques doivent figurer dans l'ordre précis ;
- la ponctuation est obligatoire à la fin de chaque énoncé ;
- l'absence des majuscules à surveiller au début des phrases, sinon le logiciel *CLAN* classe ces mots comme les noms propres ;
- le début de chaque énoncé se précède par une tabulation après **:** et la fin d'énoncé est marquée avec ●
- l'étiquetage des pauses doit se présenter comme **(.)** ;
- les parenthèses rondes **(*)** ne sont pas acceptées par *CLAN* afin de désigner les erreurs, le logiciel propose de le remplacer par **0*** sinon il s'agit de modifier des parenthèses rondes en carrées **[*]** ;
- **xx** n'est pas acceptable pour marquer un mot inaudible, il faut toujours utiliser **xxx** quel que soit le nombre de mots concernés ;
- les mots incomplets doivent être restitués : ***develop(ping) countries***.

Par ailleurs, **["]** n'est pas reconnu par *CLAN* pour marquer les commentaires paralinguistiques. La solution est de marquer l'étiquetage des expressions métalinguistiques comme des « fillers » (&-) qui ne seront pas pris un compte pour des analyses lexicales mais seront comptés en tant que disfluences.

Il est possible de coder des erreurs de toute nature (syntaxiques, lexiques, morphologiques, phonologiques, etc.) à l'aide des codes proposés dans le manuel *CLAN*. La figure 2 montre les codes d'annotations des erreurs morphologiques suivis des exemples.

Si le codage est effectué avec un autre logiciel dont le format est compatible avec CLAN (comme *CHAT* dans *EXMARaLDA*), il est possible de transférer les transcriptions déjà

5 A noter que la transcription est aussi une opération de codage, car les choix opérés lors de la transcription sont définis par le chercheur en fonction des besoins des analyses ultérieures (ex. le degré d'exhaustivité ; la segmentation des énoncés ; la transcription en orthographe des mots dont le sens n'est pas clair, comme *choise,* ou des mots avec l'orthographe non-normée, comme *yea*).

codées dans *CLAN*. Cependant après le transfert, il convient d'appliquer la fonction *CHECK* du *CLAN* (Esc+L) afin de vérifier la conformité des données aux normes de la convention *CHILDES,* sans laquelle l'accès à la fonction de l'analyse se trouve bloqué par le logiciel.

| Code | Usage | Example |
|---|---|---|
| $PRE | error involving prefix | misforgiving = unforgiving |
| $SUF | error involving suffix | taked = taken |
| $NFX | error involving infix | |
| $NFL | error involving inflection | taked = taken |
| $DER | error involving derivation | misforgiving = unforgiving |
| $RED | error involving reduplication | sevenses = sevens |
| $AGA | error of agreement, agreer is wrong | el palma |
| $AGC | error of agreement, controller is wrong | la palmo |
| $AGB | error with both wrong | el palmo |
| $REG | regularization | eated = ate |
| $FUL | full regularization | throwed = threw |
| $PAR | partial regularization | threwed = threw |
| $HAR | vowel harmony error | ablakek = ablakok |

**Figure 2. Codes pour l'annotation des erreurs morphologiques (MacWhinney, 2000, p.96)**

### 2.2. De la difficulté de segmentation de flux de parole

La segmentation des énoncés lors de la transcription dans *CLAN* est d'une importance capitale, car elle va déterminer notamment le calcul de la longueur moyenne des énoncés (*mean length of utterance, MLU*). D'où le principal intérêt de la ponctuation (le point (.)) dans *CLAN.*

Il n'existe pas un seul critère de délimitation des énoncés ; il faut souvent associer critères syntaxiques (et parfois sémantiques), prosodiques et pausologiques. Ce sont généralement les conjonctions de coordination qui posent le plus de problèmes. Dans la grammaire traditionnelle, des propositions liées par *and*, *but* ou *so* constituent une seule phrase, mais à l'oral il est possible d'avoir tout un récit constitué de propositions liées par des conjonctions. Faut-il alors le considérer comme un seul énoncé ? Cela pourrait être une solution, mais fausserait sans doute les calculs.

Prenons un exemple du corpus EMILE (extrait de la production d'un élève).

*EL1 : *# 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 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> (*) 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 .*

L'extrait en question constitue certainement une unité discursive. Si l'on se fie à des dépendances syntaxiques, l'on peut distinguer les unités suivantes :

(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.*

Dans (3) le statut de ***and*** *it will be able* est plus ambigu, mais il semble être dépendant de *it's the first time that.* Globalement il s'agit probablement de trois énoncés, liés par des connecteurs.

Prenons un autre exemple (extrait d'un discours du professeur de langue).

*ProfLV : *&-um ((0,8s)) just to make sure because of the other group ((5,4s)) it wasn't very clear what's the difference between the National (.) Health Service and the Social Security .*

Dans cet exemple, en ce qui concerne la deuxième partie du segment (*what's the difference between the National (.) Health Service and the Social Security)* il n'est pas très facile d'identifier ses limites. Tout d'abord, à la première écoute cette partie se trouve soudée avec ce qui précède (*it wasn't very clear*) c'est-à dire, aucune pause ne s'insère entre les deux.

Si on l'analyse comme un seul énoncé complexe, ce serait alors une extraposition : *It [=what the difference is…] wasn't very clear*, avec une inversion inhabituelle : *what the difference* à la place de *what the difference is.*

Comment définir si nous avons affaire à un seul énoncé complexe affirmatif ou à un ensemble d'énoncés séparés dont la seconde partie est la question ?

En effet, si nous re-écoutons ce segment, nous entendons une légère modulation de voix caractéristique des questions ouvertes. Qui plus est c'est l'indice syntaxique (l'inversion sujet-verbe *what's the difference)* qui nous conduit finalement à une réponse : cette deuxième partie est bel et bien une question ouverte et non pas une interrogative indirecte extraposée. Ainsi, en observant et en analysant l'ensemble des indices, nous nous penchons finalement vers la version où il est composé de plusieurs segments abandonnés (inachevés) et d'un segment qui est une question ouverte. Nous reconstituons finalement une série d'énoncés suivants :

(1) *&-um ((0,8s)) just to make sure because of the other group* ***+…*** (« *trailing off* » accompagné d'une pause très longue qui dépasse 5 secondes)

(2) ***+****it wasn't very clear* +//· (la professeure reprend elle même l'énoncé, l'abandonne, et sans pause, en commence un autre.)

(3) *what's the difference between the National (.) Health Service and the Social Security?* (énoncé interrogatif).

Ainsi les pauses (ou leur absence), peuvent être des indices précieux dans le découpage des segments (1) et (2) mais ne sont pas toujours des marqueurs fiables de ponctuation

dans tous les cas. Lorsque les pauses se retrouvent à l'intérieur de sous-segments elles ne signalent plus la fin logique mais peuvent relever de la recherche du mot par l'élève ou une pause qui se produit lorsque la professeure note au tableau le vocabulaire important. Par conséquent, il n'existe pas de recette universelle de délimitation des énoncés. Lors de la transcription les choix sont faits au cas par cas en s'appuyant sur les indices prosodiques et/ou morpho-syntaxiques.

## 3. Applications possibles de CLAN dans la recherche

*CLAN* permet une exploration quantitative de l'interlangue des élèves (à partir des productions écrites et orales) afin de mettre en lumière l'évolution et le profil langagier propres à l'acquisition de la L2. Notamment, il est possible d'explorer les traits morpho-syntaxiques de l'interlangue grâce aux fonctions de calculs, de lemmatisation et d'analyse morpho-syntaxique de CLAN. Également, les analyses des disfluences et de la pausologie (pauses remplies, abandons, répétitions) sont envisageables. Grâce au calcul du nombre d'erreurs nous pouvons non seulement tracer l'évolution des déviances sur l'année scolaire (le nombre des erreurs a-t-il diminué ?), mais aussi établir la comparaison du nombre d'erreurs effectués par les élèves dans les deux disciplines (en cours de langue et en histoire-géographie), ou encore entre les cours et les évaluations (les élèves se trompent-ils plus souvent pendant les cours ou lors des interrogations ?). Enfin, dans quelles catégories (lexicales, grammaticales, syntaxiques) les erreurs apparaissent-elles le plus souvent et dans quelles parties de discours ? Il est aussi possible de s'interroger quant à la place du vocabulaire dans l'apprentissage immersif EMILE. Quelles sont la diversité lexicale et la densité des discours des élèves ? L'étude nous renseigne vis-à-vis de l'interlangue écrite et orale des apprenants EMILE et contribue à la réflexion sur l'usage de L2 et la façon dont les élèves s'en servent.

### 3.1. CLAN : fonctions de calculs, de lemmatisation et d'analyse morpho-syntaxique (KWAL, FLUCALC, EVAL, MOR)

Les transcriptions du corpus EMILE codées à l'aide d'*EXMARaLDA* ont été transférées dans *CLAN* afin d'effectuer l'analyse morpho-syntaxique et lexique de l'interlangue des élèves.

Dans cette section nous présentons les fonctions de calculs, de lemmatisation et d'analyse morpho-syntaxique (*KWAL, FLUCALC, EVAL, MOR*) opérés dans *CLAN* et corroborés par des exemples tirés de l'analyse du corpus EMILE[6].

6 Analyses effectuées avec un échantillon du corpus (cours LV, durée 23 minutes, 04 octobre 2018) dans CLAN.

### 3.1.1. Fonction KWAL de CLAN

La fonction *KWAL* permet la recherche d'un mot clé et son affichage dans un contexte immédiat (=un énoncé dans lequel le mot recherché se retrouve). En plus, il est possible de paramétrer l'affichage d'un contexte éloigné en indiquant le nombre d'énoncés à afficher avant et après l'énoncé du contexte immédiate.

Exemple : rechercher le mot ***government*** dans le corpus EMILE avec 2 énoncés qui précèdent et 2 qui suivent l'énoncé du contexte immédiat (Fig.3).

```
> kwal +sgovernment -w2 +w2 CoursLV_04oct.cha
kwal +sgovernment -w2 +w2 CoursLV_04oct.cha
Thu Apr 09 09:55:05 2020
kwal (09-Feb-2020) is conducting analyses on:
 ALL speaker tiers
****************************************
From file <CoursLV_04oct.cha>
----------------------------------------
*** File "CoursLV_04oct.cha": line 77. Keyword: government
*PLV:   yes ?
*EL1:   because at this time many people wanted &-er the society <to change>
        [//] and &-er go and xxx change these fears .
*EL1:   &-er after the war (.) many people er [/] people had learnt to [*]
        (1.1) &-er work together without caring about [*] (.) social
        classes, or genders or all these things [*] so they wanted the
        government to (1.7) go deeper in their life and to [*] (1.0) have
        more [*] impact on (.) 0pron [*] health <or or or> [/] or work or like
        these things [*] 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 [*] .
*PLV:   so &-yes, you said it first who were &-er the two candidates to [///]
        in nineteen +//?
*PLV:   yes ?
----------------------------------------
*** File "CoursLV_04oct.cha": line 396. Keyword: government
*VIDEO:         yes that's just it .
*VIDEO:         we've got a [?] Parliament and that's for us to assign who goes there and to make sure
they do the job when they get there .
*VIDEO:         conservatives liberals labor right and minor parties have joined in
        supporting government sponsored programs for post war housing and
        health, for social security and education .
*VIDEO:         today even our children are being encouraged to take an active and critical interest in
planning the reconstruction of the towns and cities in which they live .
*VIDEO:         for it will take years perhaps decades to translate the blueprints of today into the
Britain of tomorrow .
----------------------------------------
*** File "CoursLV_04oct.cha": line 479. Keyword: government
*VIDEO:         would you like <us to make some> [?] coffee now ?
*VIDEO:         oh, gee !
*VIDEO:         the government have announced their plans to establish a
        National_Health_Service so that everyone may be provided with
        up to date medical care .
*PLV:   &-right .
*PLV:   this is xxx which is pretty much important .
```

**Figure 3. Fonction KWAL : Recherche du mot *government* dans un contexte éloigné dans CLAN**

### 3.1.2. Fonction FLUCALC de CLAN

*FLUCALC* est une fonction de *CLAN* qui permet de calculer le nombre de disfluences et leur type dans un discours. Dans d'autres mots, nous avons les statistiques concernant les pauses remplies, les répétitions, le lexique non terminé, les prolongations du son, etc.

Le désavantage de cette fonction consiste dans l'obligation de choisir un seul participant à la fois dont le discours sera analysé. Ainsi, une analyse simultanée de plusieurs discours n'est pas possible. Par conséquent, si nous souhaitons analyser les disfluences de l'ensemble des élèves nous devons réunir tous les discours de tous les élèves sous le même identifiant (par exemple EL1). La requête est : *flucalc +t*CHI filename.cha.*

En revanche, la fonction *FLUCALC* permet d'analyser les disfluences dans plusieurs fichiers en même temps (il convient d'indiquer *.cha à la fin). Pour l'*output* des données le logiciel crée un fichier .xls dans le même dossier où se trouve un fichier source. Le tableur contient plusieurs relevés de données dont nous citons ici les valeurs les plus importantes (Ratner et Brundage, 2016, p.31) :

- Nombre total des énoncés dans l'échantillon (mor_Utts) ;
- Nombre total des mots ayant du sens (intended words, mor_Words) ;
- Nombre de syllabes (mor_syllables)
- Nombre de mots par minute : vitesse (words_min)
- Nombre minimal des syllabes (syllables_min)
- Nombre des prolongations des sons (#_Prolongation)
- Pourcentage des prolongations des sons (%_Prolongation)
- Nombre des mots abandonnés (#_Broken_word)
- Pourcentage des mots abandonnés ( %_Broken_word)
- Nombre des blockages devant un mot (#_Block)
- Pourcentage des blockages devant un mot ( %_Block)
- Nombre des répétitions d'une partie du mot (#_PWR)
- Pourcentage des répétitions d'une partie du mot ( %_PWR)
- Nombre d'unités de répétition (repetition units) ou réitérations : répétition excessive d'une partie du mot (RU)
- Pauses (#_Pauses)
- Pauses remplies (#_Filled_pauses)
- Répétitions des phrases (# Phrase repetitions)
- Disfluences typiques (# TD typical disfluencies)
- Nombre total des disfluences (# Total SLD+TD)

| File | mor_Utts | mor_Words | mor_ syllables | words_ min | syllables_ min | #_WWR | %_WWR | #_mono-WWR | %_mono -WWR | #_WWR-RU | %_WWR-RU | #_mono -WWR-RU | %_mono -WWR-RU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CoursDNL_17jan.cha | 46 | 401 | 624 | 85,623 | 133,238 | 10 | 1,603 | 9 | 1,442 | 10 | 1,603 | 9 | 1,442 |
| CoursLV_04oct.cha | 21 | 105 | 189 | 54,783 | 98,609 | 4 | 2,116 | 1 | 0,529 | 4 | 2,116 | 1 | 0,529 |
| CoursLV_11avr_H1P1.cha | 38 | 181 | 289 | 116,774 | 186,452 | 7 | 2,422 | 4 | 1,384 | 7 | 2,422 | 4 | 1,384 |

**Tableau 1. L'exemple de l'analyse des disfluences des productions orales des élèves en cours EMILE (fonction *FluCalc* du *CLAN*)**

### 3.1.3. Fonction EVAL de CLAN

*EVAL* (évaluation) est une des fonctions les plus importantes de *CLAN* qui procure l'ensemble des données statistiques d'un échantillon traité, dont le nombre total des énoncés, la longueur moyenne des énoncés en mots (*MLU*), le *MLU* calculée en morphèmes, le nombre de différents mots (*types*), le nombre total des mots (*tokens*), la diversité du vocabulaire (*VocD*), etc.

Le désavantage est le même que dans la fonction précédente : l'obligation de choisir seulement un tier (acteur) dans chaque fichier analysé.

Tableau 2 montre un exemple de traitement du discours de la professeure de LV avec la fonction *EVAL*.

| Duration (sec) | Total Utts | MLU Utts | MLU Words | MLU Morphemes | FREQ types | FREQ tokens | FREQ TTR | Words Min | Verbs Utt | %_Word Errors | Utt Error | Density | % Nouns | % Plurals | % Verbs | % Aux | % 3S | % 1S_3S | % PAST | % PASTP | % PRESP | % prep | % adj | % adv |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 620 | 191 | 186 | 7,989 | 9,054 | 355 | 1602 | 0,222 | 155,032 | 1,639 | 0,749 | 0 | 0,471 | 16,479 | 3,184 | 23,783 | 1,748 | 4,12 | 0,499 | 1,81 | 1,248 | 1,186 | 7,179 | 4,432 | 7,179 |

| % conj | % det | % pro | Noun verb | Open closed | #open-class | #closed-class | retracing | repetition |
|---|---|---|---|---|---|---|---|---|
| 1,998 | 7,553 | 16,542 | 0,815 | 0,808 | 696 | 861 | 8 | 12 |

**Tableau 2. Données statistiques du discours de la professeure de langue (fonction EVAL)**

Une autre fonctionnalité de *EVAL* est le calcul de type et de nombre de morphèmes dans un discours selon le système de Brown (exemples tirés de Ratner et Brundage, 2016, p.28-29) :

-**PRESP** participe présent -ing : *swimming.*
-**in** préposition in : *the cheese is in the bag*.
- **on** préposition on : *put it on*.
-**PL** pluriel régulier : *dogs*.
- **&PAST** passé des verbes irréguliers **:** *fell.*
**~poss** cas possessif : *John's*
- **cop** verbe copule non contractée : *Is Meg nice? Meg is.*
- **det:art** déterminant (article) : *the ball.*
-**PAST** passé des verbes réguliers : *jumped.*
**-3S** S à la 3 personne singulier des verbes au présent: *runs*.
**-&3S** S à la 3 personne singulier des verbes au présent: *does* ou *has.*
-**aux** verbe auxiliaire non contractée : *Is John running? Yes he is*.
**~cop** verbe copule contracté clitique : *Meg's tall*.
**~aux** verbe auxiliaire contracté clitique : *John's going*.

Ainsi ce programme permet de synthétiser de multiples données dans un seul fichier. Grâce à cette fonction nous sommes en mesure d'observer plusieurs phénomènes dans le corpus des élèves et des professeurs ce qui nous renseigne sur l'appropriation et l'utilisation de la langue par les apprenants en cours EMILE. Par exemple, il est possible de tracer l'utilisation des POS (parties de discours) en pourcentages, d'observer une richesse lexicale, voir la proportion (*ration*) d'utilisation des verbes par rapport aux substantifs.

### 3.1.4. Fonction MOR de CLAN

MOR (*morphology*) est une fonction très importante de *CLAN*[7]qui effectue un calcul automatique de la structure morpho-syntactique des transcriptions en format *CHAT*[8].

Le programme *MOR* ajoute une couche secondaire **%mor** (*dependent tier*) à la transcription et effectue une lemmatisation de tous les mots. En plus, le lancement de *MOR* entraîne l'activation d'autres programmes intégrés dans le *CLAN* : *POST, POSTMORTEM*, and *MEGRASP*. Ce dernier crée une couche secondaire **%gra** qui se charge d'une analyse plus approfondie des dépendances grammaticales entre les éléments de la couche **%mor**.

```
114   *PLV:   so &-yes, you said it first who were &-er the two candidates to [///]
115           in nineteen +//?
116   %mor:   adv|so cm|cm pro:per|you v|say&PAST pro:per|it adv|first
117           pro:rel|who cop|be&PAST det:art|the det:num|two n|candidate-PL
118           prep|in det:num|nineteen +//?
119   %gra:   1|4|JCT 2|1|LP 3|4|SUBJ 4|0|ROOT 5|4|OBJ 6|4|JCT 7|8|LINK 8|6|CMO
120           9|11|DET 10|11|QUANT 11|8|PRED 12|11|NJCT 13|12|POBJ 14|4|PUNCT
121   *PLV:   yes ? •
122   %mor:   co|yes ?
123   %gra:   1|0|INCROOT 2|1|PUNCT
124   *EL2:   &-er Churchill and Clement_Atlee . •
125   %mor:   n:prop|Churchill coord|and n:prop|Clement_Atlee .
126   %gra:   1|0|INCROOT 2|1|CONJ 3|2|COORD 4|1|PUNCT
127   *PLV:   what parties ? •
128   %mor:   det:int|what n|party-PL ?
129   %gra:   1|2|DET 2|0|INCROOT 3|2|PUNCT
130   *EL2:   &-er <Labour and the Conservative> [//] &-er the Conservative and
131           the Labour . •
132   %mor:   det:art|the n:prop|Conservative coord|and det:art|the n:prop|Labour.
133   %gra:   1|2|DET 2|0|INCROOT 3|2|CONJ 4|5|DET 5|3|COORD 6|2|PUNCT
134   *PLV:   yes &-alright . •
135   %mor:   co|yes .
136   %gra:   1|0|INCROOT 2|1|PUNCT
137   *PLV:   &-um (.) so we have two very different (.) men and two very
138           different platforms &-er at the turning point of the society
139           &-right . •
140   %mor:   co|so pro:sub|we v|have det:num|two adv|very adj|different n|man&PL
141           coord|and det:num|two adv|very adj|different n|platform-PL prep|at
142           det:art|the part|turn-PRESP n|point prep|of det:art|the n|society .
```

**Figure 4. Les couches secondaires crées par MOR (%mor) et MEGRASP (%gra) dans CLAN**

Ainsi, si nous prenons un énoncé de la professeure LV :

**PLV: &-um (.) so we have two very different (.) men and two very different platforms &-er at the turning point of the society &-right .*

la fonction **%gra** va produire les résultats suivants :

%gra: 1|3|COM 2|3|SUBJ 3|0|ROOT 4|7|QUANT 5|6|JCT 6|7|MOD 7|3|OBJ 8|7|CONJ
9|12|QUANT 10|11|JCT 11|12|MOD 12|8|COORD 13|12|NJCT 14|16|DET 15|16|MOD 16|13|POBJ 17|16|NJCT 18|19|DET 19|17|POBJ 20|3|PUNCT

7 A tel point qu'un volet entier du manuel CLAN y est consacré.

8 Le manuel et les tutoriels vidéo sont disponibles sur https://talkbank.org

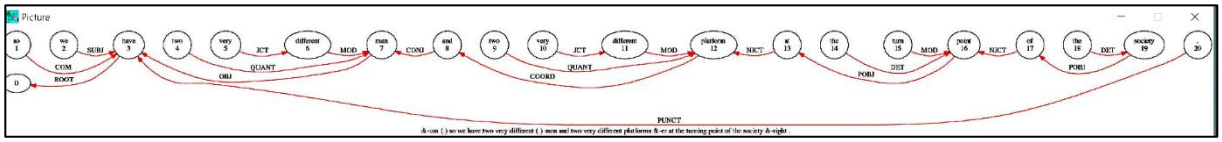



**Figure 5. Relations grammatico-syntaxiques dans l'énoncé du professeur de langue**

Ensuite, il faut cliquer sur une de ces lignes pour observer la représentation de ces relations grammaticales et syntaxiques entre les lexèmes.

Nous constatons ici que le premier mot *so* est relié au troisième mot (le verbe *have*) par le biais de sa fonction de communicateur ; en même temps, le mot à la septième position (*men*) est dépendant du même verbe *have* à travers sa fonction du complément d'objet direct, etc.

Plus précisément, nous pouvons consulter les informations telles que la quantité des morphèmes employées, le nombre des révisons et de répétitions, nombre total des erreurs, la densité des discours (nombre de verbes, d'adjectifs, d'adverbes, de prépositions et de marqueurs de conjonction divisés par le nombre total des mots), les données concernant des parties de discours et de morphèmes en particulier, la proportion du nombre des verbes par rapport aux substantifs, etc. Ces informations sont calculées par le logiciel à partir de la ligne *%mor line* à l'exception des mesures de nombre de mots et de nombre de types de mots dans les discours des élèves.

## 3.2. Analyses du profil de l'interlangue des élèves

### 3.2.1. La progression globale des élèves

Dans cette section nous suivons la progression globale des élèves dans les productions orales du mois d'octobre au mois de mai (dans tous les cours et toutes les matières confondus). L'évolution a-t-elle lieu dans les productions des élèves (le nombre d'erreurs a-t-il diminué) ? Pour répondre à cette question nous avons pris les données des colonnes « *Freq_tokens* » (nombre de mots) et « *Words_errors* » (nombre d'erreurs) de l'analyse *FLUCALC* de *CLAN*, puis les avons relativisés par 1000 mots.

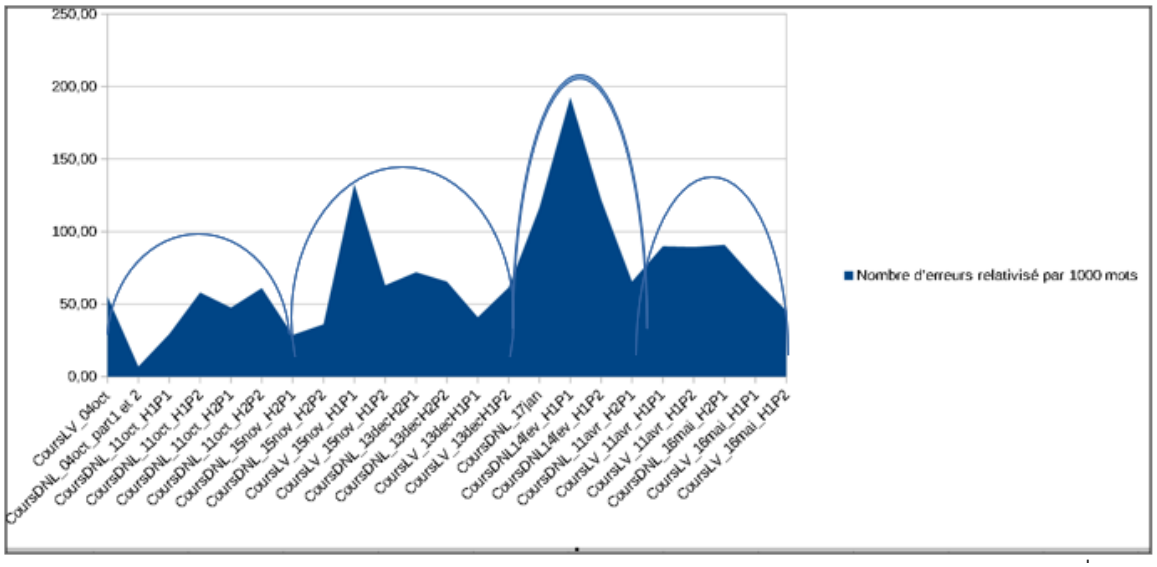



**Figure 6. Nombre d'erreurs dans le corpus oral des élèves (l'année scolaire 2018-19)[9]**

Dans le cas du premier grand pic (15 novembre), les élèves venaient de rentrer des vacances de Toussaint et la mise en route de la première semaine s'est révélée probablement difficile. Quant au cours de DNL du 14 février, il était consacré aux préparations des débats (sujet d'évaluation le jour même), l'exercice consistant à présenter un document inconnu à l'oral avec 15 minutes de préparation. Il s'agit donc d'une prise de parole assez longue dans la durée, semi-spontanée, dans des conditions de stress, ce qui peut expliquer peut-être le taux très élevé de déviances ce jour-là.

Un autre constat est que le nombre minimal d'erreurs augmente au fur et à mesure que l'année avance, sauf au mois de mai où le point minimal a diminué par rapport au mois d'avril.

L'évolution des erreurs s'effectue par cycles (illustrés par quatre arcs dans Fig.6). L'analyse suggère que chaque cycle correspondrait à une période scolaire. A l'intérieur de chaque cycle nous trouvons des hauts et des bas dans le nombre d'erreurs constatées, mais les quatre cycles sont organisés selon un schéma quasiment identique : une augmentation des erreurs au début → un ou deux pics vers le milieu du cycle puis une stabilisation → une diminution des erreurs vers la fin. L'augmentation du nombre d'erreurs au début de chaque cycle est vraisemblablement un signe typique des « redémarrages » après la rupture des vacances tandis que vers la fin du cycle le nombre d'erreurs diminue « naturellement » suite à l'évolution de l'interlangue, intégrant une meilleure maîtrise des nouveaux éléments introduits pendant le cycle et de leur contexte d'emploi.

9 Données relativisées par 1000 mots, les arcs correspondent aux périodes scolaires.

Comparaison de la progression annuelle par matière (LV/DNL)

Nous avons vu que les « vagues » d'erreurs sont probablement rythmés par l'alternance cyclique des périodes scolaires et des vacances. Mais que se passe-t-il dans les deux disciplines vues séparément ? L'interlangue des élèves évolue-t-elle différemment en fonction de la matière ce qui aurait éventuellement un impact sur le nombre d'erreurs ?
A partir des données statistiques de *FLUCALC* nous avons obtenu un graphique (Fig.6)[10]. Nous constatons que la répartition des erreurs n'est pas identique dans les deux matières concernées. En DNL la courbe des erreurs a une tendance à grimper au fur et à mesure que les mois passent (et ce, sans prendre en compte le mois de février, en non-comparable entre les disciplines en raison de l'absence des observations en LV). En revanche, l'évolution des erreurs en cours de langue adopte un schéma façon « montagnes ». L'interlangue des élèves par conséquent évolue différemment selon la matière enseignée, avec une stabilité des erreurs (par rapport au début de l'année) en LV mais l'augmentation de celles-ci en DNL en fin d'année scolaire.

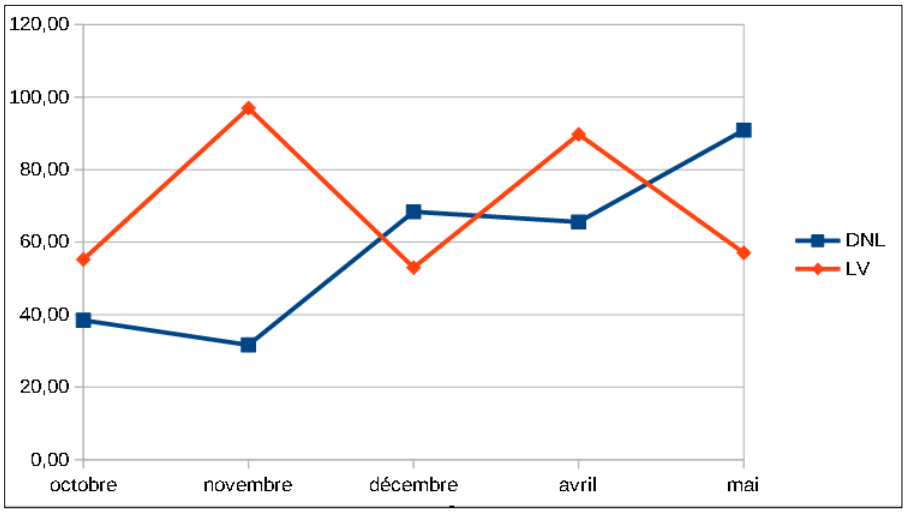



**Figure 7. Nombre d'erreurs sur l'année scolaire en LV et en DNL[11]**

### 3.2.2. Diversité lexicale

**La diversité lexicale** est une mesure utile pour observer l'évolution de l'interlangue des apprenants. Le calcul de diversité effectué par *CLAN* est basé sur *VocD* (Malvern et Richards, 1997) Sur de petits corpus il est plus précis qu'un simple *type-token ratio*, qui peut fluctuer selon la taille du corpus examiné (MacWhinney, 2000, p.116).

10 A noter que nous avons pris en compte uniquement les périodes scolaires pendant lesquelles les deux professeurs étaient présents afin de rendre les données comparables de façon égale. Ainsi les mois de février, janvier et mars sont exclus de l'analyse.

11 Données relativisées par 1.000 mots.

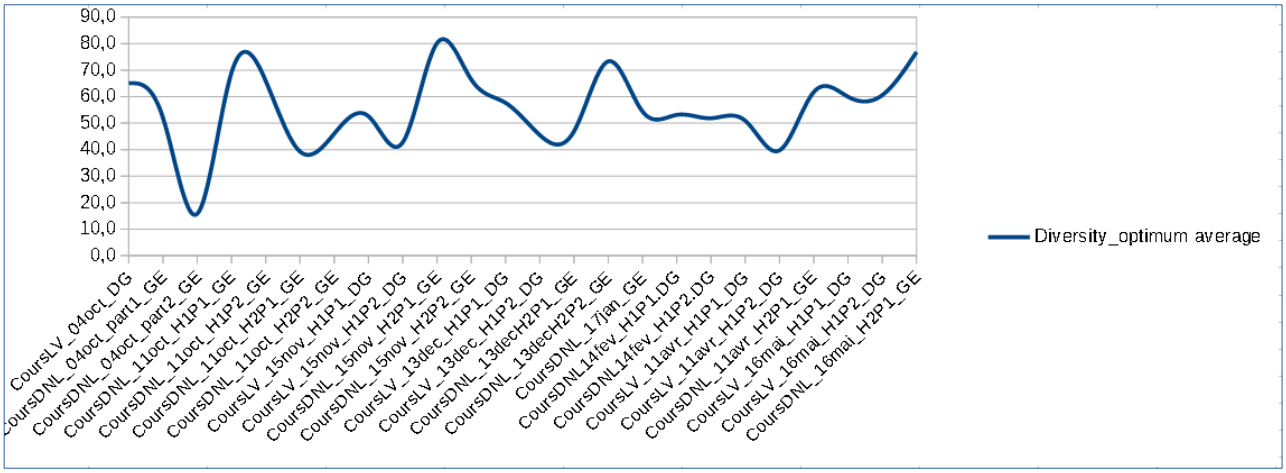



**Figure 8. Diversité lexicale dans les productions orales des élèves en EMILE durant l'année scolaire**

Nous avons retenu pour l'analyse *VocD* seulement les cinq mois de l'année durant lesquels les deux professeures ont assuré les cours ensemble. Ces périodes couvrent le 1er et le 3ème trimestre scolaires.

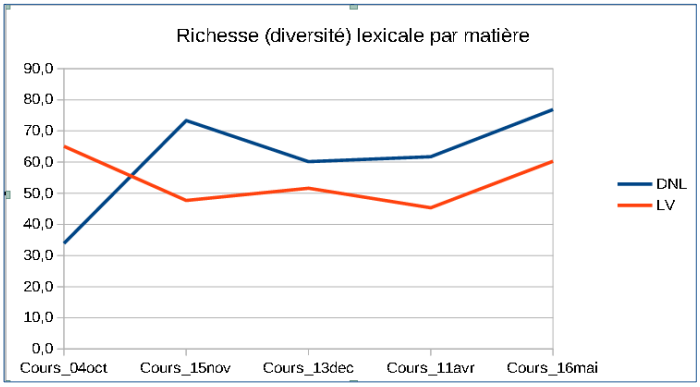



**Figure 9. Diversité lexicale par discipline enseignée**

| | Diversité (moyenne) par discipline | |
|---|---|---|
| | DNL | LV |
| Cours_04oct | 33,9 | 65,1 |
| Cours_15nov | 73,4 | 47,7 |
| Cours_13dec | 60,155 | 51,6 |
| Cours_11avr | 61,7 | 45,3 |
| Cours_16mai | 76,9 | 60,3 |
| Moyenne | 61,208 | 54,003 |

**Tableau 3. Diversité lexicale par discipline enseignée**

Figure 9 montre que la charge lexicale est relativement importante dans les deux disciplines (la diversité peut venir de l'utilisation d'un plus grand nombre de mots déjà connus ou de l'introduction de mots nouveaux). En même temps, le parcours de la courbe n'est pas tout fait le même dans les deux disciplines pendant les mois observés. Si en DNL le début de

l'année se voit chargé en vocabulaire (montée brusque de la courbe), en revanche, la diversité lexicale en LV est à la baisse à cette même période. Ensuite, entre novembre et avril (sans prendre en compte les mois d'hiver) les deux courbes se stabilisent et évoluent en parallèle (avec une charge légèrement supérieure en DNL). Enfin, vers la fin d'année scolaire les deux courbes ont une tendance à monter, tout en gardant l'écart entre les deux disciplines concernées.

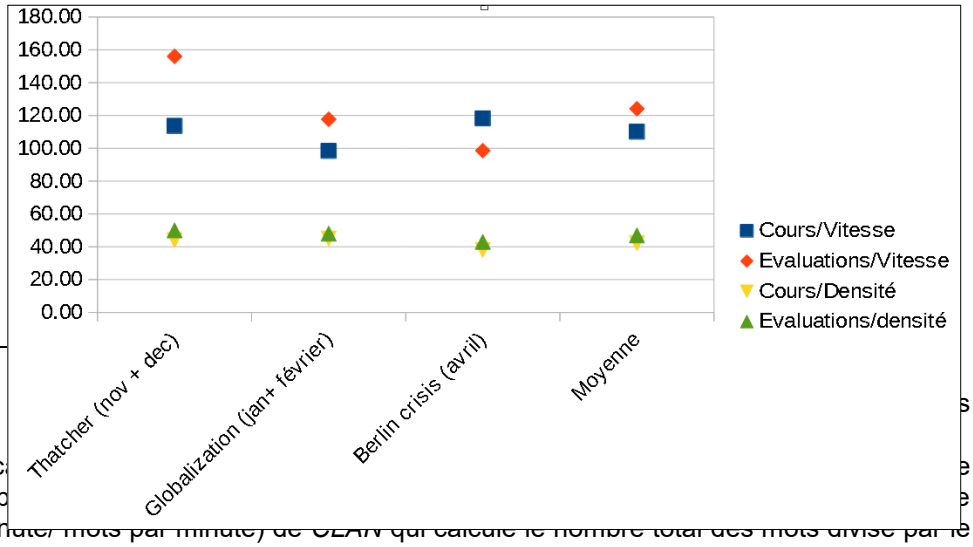
Nous constatons, entre les mois d'octobre et novembre, un effet compensatoire entre les disciplines : quand dans une des matières les élèves ont un influx important de vocabulaire, l'autre matière connaît une baisse en diversité, et inversement (voir le parcours des courbes). Enfin, l'introduction du vocabulaire se fait de façon plus ou moins homogène, l'amplitude des fluctuations étant presque identique dans les deux disciplines. Ceci semble suggérer que le vocabulaire utilisé par les apprenants en histoire-géo n'est pas beaucoup plus varié par rapport aux cours de langue[12].

### 3.2.3. Débit et densité de la parole

Débit et densité dans les cours et les évaluations. [13] Nous nous sommes appuyés sur les données des trois séquences enseignées en classe EMILE dont les moyennes ont été calculées par la suite.

| Séquence | Cours/Vitesse | Evaluations/ Vitesse | Cours/ Densité | Evaluations/ densité |
|---|---|---|---|---|
| Thatcher (nov + dec) | 113,6 | 156,0 | 44,0 | 50,0 |
| Globalization (jan+ février) | 98,4 | 117,7 | 45,0 | 48,0 |
| Berlin crisis (avril) | 118,2 | 98,6 | 38,0 | 43,0 |
| Moyenne | 110,1 | 124,1 | 42,3 | 47,0 |

**Tableau 4. Débit (WPM) et densité des productions des élèves en cours et dans les évaluations**




12Cette hypothèse s
observations.

13 La densité lexic e
conjonctions divisé p e
WPM (words per minute/ mots par minute) de CLAN qui calcule le nombre total des mots divisé par le temps de parole d'un locuteur ((McWhinney, 2000, p.127).

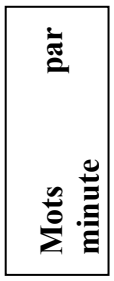



**Figure 10. Débit (WPM) et densité des productions des élèves en cours et dans les évaluations**

La première chose à remarquer est le débit plus important pendant les évaluations qu'en cours dans deux séquences sur trois. Le débit plus faible de la séquence *Berlin crisis* peut avoir deux explications :

- Premièrement, l'évaluation en question (*Berlin crisis*) est basée sur l'analyse de la production d'un seul élève. Il serait intéressant d'analyser les productions d'autres élèves et voir si les résultats vont modifier les données par la suite.
- Deuxièmement, l'élève prononce son texte (aussi préparé et enregistré en dehors de classe) très attentivement tout en concentrant son attention sur la prononciation, l'intonation et l'articulation des mots dans l'effort d'être compris.

Dans la séquence *Thatcher* la parole des élèves est beaucoup plus rapide mais aussi plus dense que dans n'importe quelle autre production orale, dans un discours adressé au premier ministre britannique Margaret Thatcher, préparé et enregistré en dehors de classe. De façon générale, nous remarquons que la densité de la parole dans les évaluations (toutes productions confondues, Fig.10) est toujours un peu plus élevée que pendant les cours.

Les deux paramètres étudiés sont en hausse dans les évaluations, les productions étant à la fois un peu plus rapides et plus denses, caractérisées par un nombre plus important de mots lexicaux que pendant les cours.

Comment expliquer ce fait ? Même s'il est difficile de parler de l'évolution proprement dit de l'interlangue des apprenants suite aux trois séquences observées, il est évident que les conditions dans lesquelles l'interlangue est déployée changent. Vraisemblablement, lors de l'évaluation, certaines opérations (d'encodage grammatical et d'activation lexicale notamment) ont été partiellement préparées à l'avance, ce qui réduit le temps de mise en œuvre de ces opérations et les risques de ratés ou de blocage.

Débit et densité des productions des élèves et du discours des professeurs.

La question de la compréhension de l'oral en cours d'anglais a toujours été un point délicat du point de vue des apprenants. Qu'en est-il en cours EMILE ? Cette section examine une question d'interdépendance entre la vitesse et la densité avec laquelle *l'input* oral est livré, ce qui serait l'indicateur de l'efficacité de *l'intake* des élèves en cours d'anglais de section

européenne. C'est une problématique à laquelle très peu de chercheurs se sont intéressés jusqu'à présent.

Est-ce que les professeures de LV et de DNL parlent avec une vitesse juste pour que les apprenants en relèvent le maximum d'informations utiles ? Leur densité lexicale présente-elle un obstacle pour une compréhension efficace des élèves ?

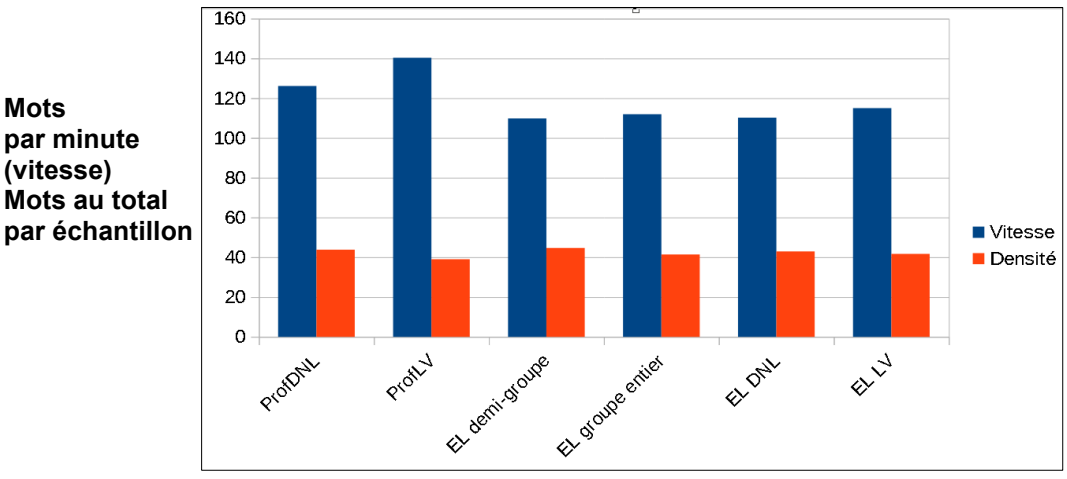
Nous avons comparé la rapidité et la densité du discours des professeures ; les mêmes paramètres des élèves en fonction de la matière (LV et DNL), et en fonction des modalités d'enseignement (en demi-groupes ou en classe entière). Pour effectuer cette analyse, nous avons relevé les données d'un échantillon du discours (durée environ 3 minutes) des professeures et des élèves pendant les périodes différentes de l'année scolaire (au début au milieu et la fin), en demi-groupes et en classe entière ; ensuite les données ont été analysées avec le logiciel *CLAN* (fonction *EVAL*).

| | **Prof DNL** | **Prof LV** | **EL demi-groupe** | **EL groupe entier** | **EL DNL** | **EL LV** |
|---|---|---|---|---|---|---|
| **Débit (WPM)** | 126,1 | 140,3 | 109,8 | 111,9 | 110,2 | 115 |
| **Densité** | 43,8 | 39 | 44,7 | 41,4 | 42,9 | 41,7 |

**Tableau 5. Débit et densité dans le discours des professeures et des productions des élèves en fonction de la matière et de la modalité d'enseignement**

Les résultats montrent que le débit et la densité lexicale des professeures et des élèves sont comparables quelle que soit la matière et le mode d'enseignement. La professeure de LV parle à une vitesse légèrement supérieure à celle de son homologue en DNL. En revanche, la densité de la parole de l'enseignante de la langue est moins importante que celle de la professeure d'histoire-géographie.

L'analyse suggère que les facteurs objectifs, comme le débit et la densité lexicale des professeures, est semblable de ceux des élèves, qui se trouvent « en phase » avec leurs professeures. Dans le chapitre neuf nous analysons les facteurs subjectifs, à savoir, la perception du débit de la parole des professeures par les élèves.



**(densité[14])**

**Figure 11. Débit et densité dans le discours des professeurs et des productions des élèves en fonction de la matière et de modalité d'enseignement**

En conclusion, les fonctions *MOR, EVAL, FLUCALC* et *KWAL* du logiciel *CLAN* (décrites plus haut dans notre travail) permettent d'effectuer des analyses quantitatives dont les résultats sont ensuite interprétés qualitativement afin de visualiser l'ensemble des caractéristiques linguistiques dans les paroles des apprenants en cours EMILE. Ces données ont comme valeur non seulement de dessiner le profil morphologique et lexique de l'interlangue des élèves mais aussi de voir son évolution dans le temps ce qui donne comme résultat une représentation de son image globale.

## 4. Avantages et inconvénients de CLAN

Citons quelques avantages principaux de *CHAT-CLAN*[15]. Le logiciel est gratuit, téléchargeable en ligne et accompagné d'un manuel complet ; il est compatible avec le système d'exploitation Windows.

Grâce au logiciel *CLAN* il est possible d'étudier les interactions conversationnelles ou repérer les dysfonctions langagières. Le logiciel propose des fonctionnalités puissantes. Tout d'abord, il est possible de lier la transcription à un fichier vidéo ou audio pour une meilleure lisibilité et l'interprétation des données. Les transcriptions sont uniformes et donc exploitables dû aux règles précises de « syntaxe » de *CLAN* : l'absence des majuscules en début d'énoncé, pas de ponctuation possible à l'intérieur du segment, ponctuation obligatoire à la fin d'énoncé, suivie par une balise ●.

En second lieu, afin d'effectuer l'analyse des conversations CLAN dispose d'un système d'étiquetage et d'un panel de codage configurable, ce qui rend le processus de transcription plus rapide. Par exemple, lors du codage du corpus EMILE dans le but d'analyser des erreurs des apprenants nous avons étiqueté des déviances, comme : $LOS (manque d'un élément) ; $ADD (un élément ajouté), $VER (erreur dans la catégorie des verbes) ; $MOR (erreur morphologique), etc.

Ensuite, *CLAN* permet une prise en compte des données pour l'analyse automatique acoustique ou morphosyntaxique. Il est aussi possible d'effectuer des analyses approfondies

14 Densité (*propositional idea density*) : le nombre de verbes, d'adjectifs, d'adverbes, de prépositions et de conjonctions divisé par le nombre total des mots**.** (MacWhinney, 2000,p.137).

15 https://talkbank.org/manuals/Clin-CLAN.pdf

morpho-syntaxiques de l'interlangue des apprenants grâce à des multiples programmes intégrés dans le logiciel : *FLUCALC* (calcul du nombre de disfluences), *KWAL* (recherche d'un mot clé dans un contexte immédiat ou éloigné), *MLU run* (calcul de la proportion du nombre : morphèmes/énoncé), ou *COMBO* (séquences ciblées de recherche, ex. trouver l'infinitif dans tous les fichiers, etc.)

Quelques inconvénients sont à noter, cependant.

La transcription et l'annotation des événements précis du corpus nécessite une étude préalable des codes de transcription en accord avec les conventions de transcription en vigueur (ex.CHILDES). Les règles syntaxiques imposées par le logiciel CLAN peuvent présenter quelques inconvenients lors de la transcription. Par exemple, l'énoncé : *It's one of the five giants, remember ?* sera catégorisé par CLAN comme une question puisqu'il y a un point d'interrogation à la fin (la ponctuation à l'intérieur du segment n'étant pas possible) ; il convient donc de le scinder en deux énoncés distincts. Par ailleurs, métadonnées doivent être placées en ordre précis précédées par @ et placées à la ligne (@Begin @Languages @ Participants @ID @Media, voir Fig.1). Qui plus est, le codage des évènements ou des erreurs doit être régulier, homogène et systématique afin de permettre des exploitations futures possibles par d'autres chercheurs (voir Tableau 6 pour quelques exemples du codage des erreurs dans les productions orales des élèves).

| Catégorie des erreurs et codes attribués | Exemples des erreurs du corpus oral des élèves |
|---|---|
| - accord (AGR) | he <try*> tri**es** to explain that there <is*> **are** difficulties in the country |
| - aspect (ASP) | -many people <had learnt>* **learnt** to work together;<br>-it cost less to the European [country] because it <is exporting>*** exports** |
| - temps (TNS) | -they could <sold*> **sell** their goods ;<br>-they <start*> start**ed** panicking ;<br>- the British could send or <received*> **receive** |
| - préposition devant le nom (PREP) | -exports **with*** (to) the other countries;<br>-they're looking ***** (at) each other; |

**Tableau 6. Exemples du codage des erreurs les plus typiques**

## Conclusion

Dans la contribution actuelle nous avons présenté le logiciel de transcription et de traitement des données CLAN, ainsi que décrit ses fonctions principales. Nous avons également montré l'application du logiciel dans la recherche (l'étude de cas) pour but des analyses qualitatives et quantitatives du corpus EMILE. Notamment, le profil langagier morpho-syntaxique et lexical des élèves a été observé à travers des mesures de CLAN, comme : la diversité et la densité lexicale, des disfluences dans les discours, le calcul du nombre d'erreurs dans le corpus des élèves, entre autres. Ces données permettent de dégager les

caractéristiques et les traits spécifiques de l'interlangue des élèves, ainsi que mieux appréhender les enjeux de l'acquisition de la L2 en classe immersive EMILE dans une école française.

## BIBLIOGRAPHIE

Bakaldina-Nicol, E. (2023). *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.* [Thèse de doctorat ; Université Savoie Mont Blanc ; LLSETI].

Bakaldina-Nicol, E. (2023). Exmaralda: outil de traitement des données discursives orales. Mélanges Crapel (44/1), p. 330-348. https://www.atilf.fr/wp-content/uploads/publications/MelangesCrapel/Melanges_44_1_17_Bakaldina-Nicol_2023.pdf

Corder, S. P. (1981). *Error analysis and interlanguage.* Oxford University Press.

Granger, S., Gilquin, G. et Meunier, F. (dir.) (2015). *The Cambridge handbook of learner corpus research* (1ère éd.). Cambridge University Press.

MacWhinney, B. (2000). The CHILDES Project: Tools for Analyzing Talk. 3rd Edition. Lawrence Erlbaum Associates.

Ratner, B.N. et Brundage, S. (2016). *A clinician's complete guide to CLAN and PRAAT.* URL: https://vandammark.com/WSU/BernsteinRatnerBrundage_2016_ClinClan.pdf

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, *10*(3), 219-31.

Sinclair, J. et Coulthard, M. (1975). *Towards an analysis of discourse*. Oxford University Press.

Sinclair, J. (dir.) (1991). *Corpus, concordance, collocation*. Oxford University Press.

Stubbs, M. (1986). Lexical density: A computational technique and somefindings. Dans M. Coulthard. (dir.), *Talking about text* (p.27-48). English Language Research.

Tognini-Bonelli, E. (2001). Corpus Linguistics at work*. Studies in corpus linguistics*. (6). John Benjamins Publishing.

Tutin, A., Jaques, M-P., Kraif, O. et Hartwell, L. (s.d.). *Introduction à la linguistique de corpus*. FUN MOOC. URL : https://www.fun-mooc.fr/fr/cours/introduction-a-la-linguistique-de-corpus/.