EXMARALDA : outil de traitement des données discursives orales

Evgenia Nicol-Bakaldina

Mélanges CRAPEL n° 44/1

Evgenia Nicol-Bakaldina

EXMARALDA : OUTIL DE TRAITEMENT DES DONNEES DISCURSIVES ORALES

Mots-clés

annotation – corpus – EXMARaLDA – linguistique – TAL

Keywords

annotation – corpus – EXMARaLDA – linguistics – software

Résumé

Ces dernières années les logiciels de traitement automatique de langue (TAL) sont de plus en plus impliqués dans la linguistique de corpus et la didactique des langues. Cet article met en lumière les logiciels TAL, en présentant en particulier EXMARaLDA. Les fonctionnalités principales du logiciel sont décrites, comme la transcription, l'annotation et le codage des données orales discursives. Les applications possibles sont présentées dans les domaines de la recherche en linguistique de corpus et de la didactique des langues, suite à une recherche doctorale sur l'EMILE (L'enseignement d'une matière par l'intermédiaire d'une langue étrangère) effectuée à l'Université de Savoie Mont Blanc. L'article se termine par un bilan des avantages et inconvénients.

Abstract

In recent years, language processing software tools have become more and more implicated in corpus linguistics research and language didactics (CALL) fields. In this paper, some essential information about multimodal annotation tools is summed up with a focus on EXMARaLDA. The tool's main functions are described such as transcription, annotation and coding of oral discursive data. Examples of the software use for corpus annotation and analysis purposes are provided drawing on the Content and Language Integrated Learning (CLIL) case study research conducted at the Savoy Mont Blanc University. EXMARaLDA's strengths and weaknesses are outlined in the conclusion.

Introduction

Cet article met en lumière les logiciels TAL, en présentant en particulier *EXMARaLDA*. Nous décrirons ses fonctionnalités principales, préciserons les applications possibles dans les domaines de la recherche et de la didactique. L'article se termine par un bilan des avantages et inconvénients en termes de conclusion.

Le Traitement Automatique des Langues (TAL) est un domaine scientifique pluridisciplinaire qui se situe au croisement de la linguistique et de l'informatique souvent associé à l'intelligence artificielle[1].

Il existe un lien étroit, voire une interaction interdépendante, entre les TAL et la linguistique de corpus, les logiciels de traitement de données étant de plus en plus impliqués dans un champ de recherche en linguistique et didactique des langues. D'un côté, les linguistes de corpus recourent aux logiciels TAL à des fins de transcription, de constitution et de traitement d'un corpus à partir de données volumineuses et hétérogènes rendues de ce fait exploitables. D'un autre côté, un corpus constitué sert de support d'observation et d'extraction des données linguistiques à l'aide des outils de TAL afin de concevoir de nouveaux outils comme des dictionnaires, des listes lexicales, des bases hypertextuelles, des ontologies, etc. C'est ainsi que grâce à la linguistique de corpus les logiciels TAL évoluent et leur efficacité peut être testée (Tutin et al., s.d.).

1. Présentation des logiciels de traitement automatique des langues

Dans la linguistique de corpus, il existe une pléthore de logiciels d'analyse/exploitation de données, tant discursives que textuelles (perspective textométrique), tout en sachant que les données initiales discursives, pour être analysées, doivent être transcrites, donc se présenter sous la forme d'un texte (corpus). Les programmes TAL ont des fonctionnalités nombreuses : elles permettent la reconnaissance ou la synthèse vocale, des navigations dans des contextes élargis, l'établissement d'index sélectifs ou systématiques, le codage (étiquetage ou *tagging*[2]), l'extraction des fréquences, des concordances ou de cooccurrences, tout en favorisant la recherche lexicale, analyses morpho-syntaxiques ou sémantiques, des interprétations statistiques. De plus, l'intelligence artificielle s'invite désormais dans le domaine de la didactique, entre autres, avec un but d'apprentissage des langues qui a considérablement élargi le potentiel d'exploitation du TAL.

Dressons une liste non exhaustive des logiciels les plus connus en linguistique de corpus :

---

[1] https://www.universalis.fr/encyclopedie/traitement-automatique-des-langues/

[2] Une étiquette assignée à chaque mot du corpus indique une partie du discours et d'autres catégories grammaticales. "A POS tag (or part-of-speech tag) is a special label assigned to each token (word) in a text corpus to indicate the part of speech and often also other grammatical categories such as tense, number (plural/singular), case etc. POS tags are used in corpus searches and in text analysis tools and algorithms." Source: https://www.sketchengine.eu/blog/pos-tags/

- *AntConc (Anthony Concordancier)* développé par Anthony[3] (Université de Waseda, Japon), la dernière version 4.2.0 est sortie en 2022. Le logiciel a fait preuve de son utilité dans le traitement des données des corpus, ainsi les sites des universités françaises en proposent des guides assez détaillés[4]. Doté d'une interface intuitive, il permet d'effectuer des requêtes simples et élaborées[5], d'observer la répartition (plot) d'un mot simultanément dans plusieurs corpus, etc. Notons aussi *AntWordProfiler*[6]*,* la version plus moderne du même logiciel dotée de plusieurs fonctions supplémentaires, comme l'analyse de n'importe quelle langue à condition qu'il y ait une liste de mots constituée. *ProtAnt*[7], outil d'analyse de la prototypicalité des textes, a été développé en 2017 par Anthony et Baker. Encore d'autres versions du même logiciel, plus récentes, sont : *FireAnt* (doté d'un système de visualisation intégré)*, AntGram* (focalise sur la recherche des N-grams).

- *CLAN* (Computerized Language *Analysis*) est un logiciel phare d'analyse de données textuelles, qui permet également d'effectuer les transcriptions des fichiers sonores grâce aux règles préétablies en conformité avec le format *CHAT*[8]. *CHAT* et *CLAN* font partie du *CHILDES* (*CHild Language Data Exchange System*), système doté d'outils ayant pour but l'analyse des interactions discursives et qui sert en tant que convention de codage des corpus reconnue au niveau mondial[9] .

- *Hyperbase*[10]. Ce dispositif met en relation plusieurs textes d'un corpus (ou plusieurs corpus), en faisant ressortir des spécificités lexicales (grâce aux indices de spécificité semi-automatiques), l'évolution du vocabulaire (ex : la distance ou connexion entre mots/groupes nominaux, la richesse lexicale), et permet également d'effectuer des analyses factorielles ou « arborées » des données, entre autres. Qui plus est, il est possible d'établir des relations entres les données textuelles grâce à la coloration thématique (voir l'annexe 1).

---

[3] https://laurenceanthony.net/

[4] Université de Caen : https://ecampus.unicaen.fr/pluginfile.php/345933/mod_resource/content/6/co/1_-_7_-_Grain_video.html, Université de Lyon : http://cid.ens-lyon.fr/ac_article.asp?fic=antconc.asp , traduit en français par Stefania Solofrizzo, Guide by Warren Tang (Hiroshima University, Japan).

[5] Par exemple, grâce à la fonction « wordcard/joker » « * » il est possible de chercher des formes de verbes.

[6] Laurence Anthony's AntWordProfiler

[7] https://www.laurenceanthony.net/software/protant/resources/anthony_and_baker_2015_FR.pdf

[8] Ratner and Brundage (2016, p. 2)

[9] MacWhinney (2000)

[10] Le logiciel est développé au laboratoire BCL « Bases, Corpus, Langage » à Nice, sous la double tutelle du C.N.R.S. (sections 34 et 26) et de l'Université Nice Sophia Antipolis, http://bcl.cnrs.fr/, http://ancilla.unice.fr/.

- D'autres logiciels de traitement des données multimodales (écrites et/ou orales) conçus en vue d'analyses statistiques, de traitements automatiques et d'interprétations des corpus : *ELAN*[11], *SketchEngine*[12], *TRAMEUR*[13]; *Range*[14] (connu auparavant sous les appellations *VORDS*, *FVORDS* et *VocabProfile)* ; *TXM*[15] *; CQPweb Lancaster*[16] *; Lexico 3*[17], *UAM Corpus Tool*[18], etc.

- *EXMARaLDA*[19], outil de traitement des données orales discursives (transcription, organisation, analyse), est présenté dans la section suivante.

2. Présentation du logiciel *EXMARaLDA*

Le logiciel *EXMARaLDA* (***Ex****tensible* ***Mar****kup* ***L****anguage for* ***D****iscourse* ***A****nnotation*) est conçu par un chercheur allemand Thomas Schmidt en 2001. Le programme a pour vocation principale l'annotation de la parole comme son nom le suggère. Très en vogue de par ses nombreuses fonctionnalités à visée d'analyses, il ne cesse d'être adopté par un nombre toujours plus important de chercheurs en linguistique de corpus.

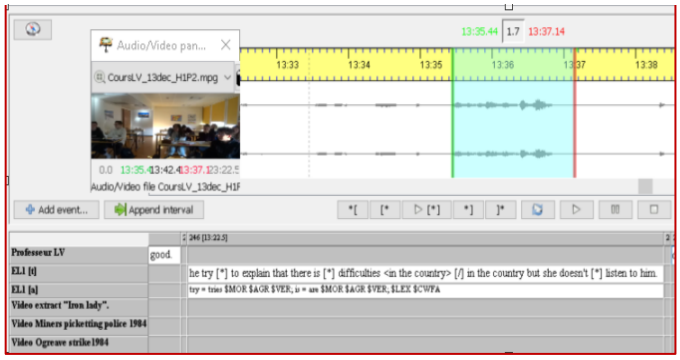


Figure 1. Interface générale du logiciel *EXMARaLDA*[20]

[11] https://archive.mpi.nl/tla/elan
[12] https://www.sketchengine.eu/
[13] http://www.tal.univ-paris3.fr/trameur/
[14] la version simplifiée développée par Tom Webb est consultable sur le site Compleat Lexical Tutor, https://www.lextutor.ca/
[15] http://textometrie.ens-lyon.fr/?lang=en
[16] https://CQPweb.lancs.ac.uk/
[17] http://www.tal.univ-paris3.fr/lexico/Lexico3doc0.pdf
[18] http://www.corpustool.com/
[19] Www.exmaralda.org
[20] 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. *EXMARaLDA* a été utilisé pour transcrire, coder et analyser des données discursives orales récoltées.

Tout d'abord, le logiciel peut être utilisé à des fins de transcriptions des données audio-visuelles. La délimitation précise du début et de la fin de l'énoncé est possible grâce à un oscillogramme : le déroulement de la bande son s'effectue en parallèle avec la transcription. Le logiciel permet également de créer de multiples couches (tiers) en lien avec les objectifs d'exploration (ex : la couche de transcription, la couche d'annotation, ou encore la couche de commentaires). Il est possible ainsi d'annoter et d'analyser des phénomènes discursifs divers : linguistiques (lexicaux, grammaticaux, prosodiques), ou paralinguistiques (rires, pauses, gestes), voire d'explorer des caractéristiques suprasegmentales (intonation, modulations de la voix, etc.). L'annotation des phénomènes recherchés est facilitée par le fichier .xml intégré au panel d'annotation configurable (annexe 2). Le logiciel utilise un standard informatique Unicode. Enfin, les transcriptions peuvent être fusionnées et exportées dans différents formats.

Le logiciel *EXMARaLDA* est un système composé d'un ensemble d'outils (programmes) : *PartiturEditor*, *COMA* et *EXAKT*. Le premier est un éditeur *Partitur-Editor* permettant de transcrire et d'annoter (coder) les données à partir des fichiers audio-visuels avec une possibilité d'éditer les résultats (segmenter les annotations, choisir seulement les extraits des transcriptions voulus et les mettre en valeur, exporter les données dans plusieurs formats : *html, pdf, rtf,* etc.). Une des fonctions principales de Partitur-Editor est la segmentation. Celle-ci est nécessaire afin d'analyser par la suite les données à l'aide de *EXAKT* (voir la description ci-dessous). 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*.

Les deux autres outils sont *COMA* et *EXAKT*. *COMA* (*Corpus Manager*) est nécessaire afin de rassembler les données transcrites et de compiler un corpus à l'aide de métadonnées assignées aux transcriptions et aux enregistrements. Il peut s'avérer également utile pour renseigner des métadonnées détaillées concernant les participants et les interactions à analyser. *EXAKT* est un outil intégré dans *EXMARaLDA* en vue des analyses des transcriptions telles que : la recherche d'un mot (ou d'un symbole) dans un contexte, en utilisant la fonction *KWIC*[21] ou *RegEx* (les expressions régulières), ensuite il est possible d'exporter les résultats de recherche.

2.1. Transcription et codage des événements

Afin de pouvoir se servir du corpus et notamment travailler avec l'interlangue des élèves (analyser les erreurs, etc.) dans *EXMARaLDA*, il est nécessaire de transcrire les paroles et de coder les éléments (événements) que l'on souhaite analyser. La transcription et le codage s'effectuent à l'aide du Partitur-Editor du système *EXMARaLDA*, tandis qu'une analyse quantitative et qualitative

[21] Key Word In Context

des données est opérée au moyen de l'outil *EXAKT*.

Plan du travail avec *Partitur-Editor* **:**

- la création d'une piste (couche) d'annotations en plus de celle de transcription ;
- la segmentation et la transcription des énoncés (manuelle ou automatique) ;
- l'étude d'un système de codage (format *CHAT* conformément à la convention *CHILDES*) ;
- l'écoute des enregistrements et le codage des événements (erreurs) dans la couche d'annotations[22]. A noter que certains éléments prosodiques – l'hésitation, les pauses, l'interruption de l'énoncé – peuvent être codés directement dans la piste (couche) de transcription ;
- la possibilité de réunir les transcriptions dans un seul fichier (fonctionnalité : fusionner des transcriptions) afin de chercher et d'analyser les événements dans l'ensemble des transcriptions ;
- l'édition des fichiers. La transcription peut être sauvegardée dans le format de présentation souhaité, soit par le biais de browser (.html), dans un document *Word*, en RTF ou en PDF. **Fichier > Édition**[23]**.** Ainsi, les transcriptions sont exportées en format .html (Fig.2).

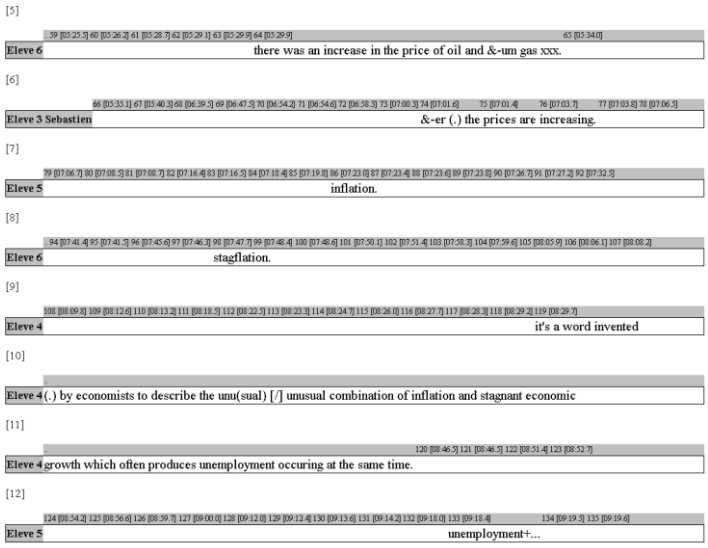


Figure 2. Transcription des paroles des élèves exportée dans le format .html

Donnons maintenant un exemple d'utilisation de *Partitur-Editor*. Dans notre étude de cas nous avons créé deux types de couches : la transcription et l'annotation.

[22] Le panneau d'annotation fonctionne à partir d'un fichier configurable au format .xml qui permet d'insérer n'importe quel ensemble d'annotations. Pour des fins d'annotation des erreurs nous avons utilisé le fichier xml élaboré au sein du laboratoire LLSETI afin d'insérer les codes d'erreurs de CHILDES après l'avoir incorporé dans le logiciel EXMARaLDA.

[23] La fonctionnalité « Range » permet de définir les modalités de la présentation de l'output, à savoir, si toutes les couches ont été prises en compte, si l'intégralité de la transcription a été présentée ou seulement une partie, etc.

- Des couches « transcription » (une pour chaque type de locuteur : professeurs, élèves, natifs anglophones) contiennent les transcriptions proprement dites. Conformément à la convention *CHILDES* un astérisque figure à la fin du chaque mot contenant une erreur[24].
- Une couche « annotation » contient l'annotation des erreurs identifiées dans la couche des transcriptions (nous avons annoté uniquement les erreurs des élèves, conformément à l'objectif de notre recherche qui est l'exploration de l'interlangue des élèves).

2.2. Codes d'annotations des événements prosodiques

L'annotation des erreurs se fait sous forme de codage selon la convention *CHILDES*. Les codes d'annotations des disfluences principales[25] (MacWhinney, 2000, p. 66-76) sont présentées en annexe 3. Nous allons expliciter et illustrer ces codes avec des exemples tirés de notre étude de cas.

Symbole *[:* . Dans les cas où le mot n'est pas utilisé dans sa forme standardisée[26], il peut être remplacé par une forme normalisée de la langue cible (*target language form*) précédé par un symbole *[:* suivi par un espace. Cette correction se fait directement dans la couche de transcription. Si l'on veut que le logiciel marque une variation orthographique comme une erreur **et** en même temps proposer le remplacement, il faut ajouter **[*].** Cependant, afin de ne pas trop « charger » la couche de transcription il est recommandé d'assigner les corrections à la couche d'annotation.

Remplacement (assimilations).

*so they **gonna [: going to]** just agree with the (*) other countries of the [*] Europe*

Shortening. Une lettre ou une partie du mot manquant (le mot considéré comme incomplet par le logiciel) s'ajoute grâce aux parenthèses, ensuite il est possible de demander au logiciel de prendre en compte ou pas les parenthèses lors du calcul des données.

---

[24] Rappelons que l'erreur n'est pas la même chose que « mistake » (James, 1998), cette dernière se distingue par la capacité du locuteur de s'auto-corriger, sans ou avec l'aide de l'extérieur. Ainsi, toute tentative de reprendre ou de reformuler la parole ou encore les hésitations dans le discours ne peuvent être considérés comme erreur. Ces événements sont cependant codés en tant que disfluences au sein de cette même couche « transcription » pour des raisons statistiques.

[25] Productions autres que normalisées.

[26] Selon Baggioni (1994), l'étape de normalisation précède l'étape de standardisation, bien que les deux soient étroitement liées. Les définitions données par l'auteur (1994, pp.84-85) sont : « L'aspect normalisation concerne d'une part les écrits et les actions pour la promotion de la variété vernaculaire, d'autre part tout ce qui contribue à orienter le choix de la norme vers une variété précise (variété littéraire toscane vs autres dialectes de la péninsule, koinè saxonne vs dialectes alémaniques, etc.) ou un choix de registre (variété plus ou moins littéraire vs variété plus proche des dialectes parlés, type de locuteurs-scripteurs de référence, etc.). L'aspect " standardisation " touche tout ce qui a trait au travail de description-fixation de la langue, ce qui se manifeste aussi bien par l'élaboration d'outils métalinguistiques (grammaires, dictionnaires, rhétoriques, manuels d'enseignement) que par des œuvres littéraires de référence servant de corpus pour l 'élaboration-justification des « règles » sans lesquelles on ne peut parler de " langue standard " ».

they start(ed) panicking, per capi(ta), chair(man), victori (ous), ideologic(al).

En revanche, l'item lexical *ya* est considéré par le logiciel comme différent de *you*, bien que leur équivalence sémantique soit maintenue au moyen d'une liste formalisée des variations dialectales orthographiques (dialectal spelling variations).

Transcription incertaine [?]. Il existe des situations où le bruit de fond ou une qualité moyenne de l'enregistrement empêchent une bonne compréhension. Si malgré tout on croit reconnaître un mot, sans en être sûr, un procédé « best guess » s'applique : on adopte la transcription qui semble la plus plausible parmi toutes les variantes :
*It's about &-er the social service, about the ratio **[?]** of taxes.*

Répétitions [/].

- un mot est répété sans correction : *after the war (.) many **people [/]** # er **people ;***
- un groupe de mots est répété sans correction : *so before we get started **<can you tell me>** # er ((0,6s)) **[/]can you tell me ;***
- s'il y a des pauses et des marqueurs d'hésitations entre l'énoncé initial et la répétition il convient de les mettre après le symbole de répétition **[/]** : *to answer to **<what they> [//] # er what they** wanted ;*
- quand un mot ou un groupe de mots est répété sans événements, toutes les répétitions, sauf la dernière, doivent être placées dans le même groupe : ***<or or or> [/] or work.***

Répétitions avec des changements [//].

- un mot répété : *the two candidates **to [//] in** nineteen*
- un groupe de mots répété avec un changement :

<Labour and the Conservative> [//]** # er **the Conservative and the Labour .

Reformulations [///]. Parfois les changements impliquent une reformulation complète du message en abandonnant la partie entamée.
<so do you> [///] before that there were already (.) changes in the society.

Recouvrement. Il existe plusieurs façons d'indiquer un chevauchement dans un énoncé interrompu. La plus simple serait d'utiliser **+<** (lazy marking) au début de la phrase qui a chevauché la précédente, mais les mots exacts superposés ne seront pas marqués.

Elève : *#er social security includes the (.) <all types of> [?] insurance for # er employment xxx (.) children, maternity and the NHS.*
ProfLV: *+< Exactly.*

Une autre façon de marquer les chevauchements est d'insérer les paroles du second locuteur directement à l'endroit de transcription où le premier locuteur a été interrompu.
Elève : *the belief that purely [*] economic [*] integration F_T_A &* ****PDNL:yes*** *would be superior to deeper, politico economic alternatives.*

Auto-interruptions. Un locuteur arrête soudainement un énoncé à sa propre initiative (celui-ci reste ainsi inachevée, abandonnée), sans pause et en commence un autre : ***+//.*** (+//? pour indiquer une question abandonnée).
Elève : ***because the character here is Atlee +//.***
Elève : ***no, not Atlee +//.***
Elève : ***Aneurin Bevan.***

Enoncés non terminés[27]. Un locuteur en incite un autre à terminer un énoncé ou un locuteur abandonne une phrase et un autre locuteur la termine. Ceci est souvent accompagné de pauses. Ce genre d'interruption peut être volontaire ou pas. Pour le premier locuteur qui ainsi abandonne une phrase l'on indique **+… ;** pour celui qui reprend l'énoncé les symboles **++** sont utilisés.
Prof LV ***: so ((0,7s)) freedom from +..?*** (le professeur incite les élèves à répondre en laissant l'énoncé inachevé, « trailing off » type interrogatif)
Elève ***: ++ want.*** (L'élève réagit à la sollicitation et termine l'énoncé).
Prof LV ***: +, want.*** (Le professeur reprend cette même phrase abandonnée au début et la termine en répétant ce qui était dit par l'élève. Ainsi un double *input* pour les élèves est assuré).[28]

Citation. La lecture d'un texte par des élèves à voix haute est marquée par **+"** :
+" *late entry into the EEC, an outside member.*

Les segments inintelligibles des énoncés sont transcrits comme **xxx :**
 # er not sure we've finished it ***xxx*** *do that* ***.***

[27] *invited interruptions*, aussi appelées *trailing off*
[28] L'exemple observé est un schéma classique de l'échange en classe, en effet, correspondant à la structure IRF : *initiation, response, feedback* selon Sinclair et Coulthard (1975).

www. L'information que l'on ne souhaite pas transcrire (non-pertinente)**.** Le code est utilisé dans la ligne principale (couche des transcriptions).
*they were not (3.3) <doing guerre> [*] **www** +...*

3. Applications d'*EXMARaLDA* dans la recherche et dans l'apprentissage des langues

EXMARaLDA peut avoir plusieurs applications possibles, dont les principales sont : l'analyse des données discursives orales dans la recherche en linguistique appliquée ; l'apprentissage des langues et de multilinguisme ; la phonologie et la phonétique ; le domaine de la sociolinguistique et de la dialectologie, etc.
Nous nous bornerons dans le présent article à proposer quelques applications du logiciel *EXMARaLDA* en linguistique de corpus appliquée et en apprentissage des langues.
Tognini-Bonelli (2001) distingue deux approches de recherche en linguistique de corpus : « *corpus-based* » et « *corpus-driven* ». Dans la première approche un chercheur définit *a priori* des structures et des formes linguistiques dont les variations et l'utilisation seront testées/analysées grâce au corpus linguistique (corpus en tant qu'objet de recherche). La seconde approche « *corpus-driven* » se veut plus inductive et permet de faire émerger les caractéristiques/phénomènes linguistiques à partir du corpus choisi (le corpus devient un sujet, une source de données).
Ainsi, les productions orales des apprenants sont susceptibles de faire émerger des informations utiles afin de comprendre l'utilisation de la langue par les apprenants. Ces données peuvent renseigner au sujet de la qualité des productions des élèves, des mots et des expressions les plus fréquents, mais aussi permettent d'observer le taux de mots erronés et leur fréquence, les types d'erreurs et leur évolution dans les élocutions des apprenants.

3.1. Établir un profil linguistique : recherche des mots par type d'erreur

Le logiciel *EXMARaLDA* est doté d'un outil *EXAKT* (***EX****MARaLDA* ***A****nalyse und* ***K****onkordanz****t****ool*) permettant d'effectuer des recherches dans les données transcrites et annotées, et de les analyser quantitativement et qualitativement.

- Recherche morphologique. Trouver un mot ajouté. Afin de trouver les erreurs morphologiques dans lesquelles les ajouts ont été effectués (par exemple, un), il faut aller dans « File → Exakt search » et choisir dans le tableau la couche où il faut chercher les erreurs. La requête dans RegEx permet de cibler la recherche. Pour voir le contexte dans

lequel l'événement cherché apparait il faut cliquer sur l'événement souhaité (la sélection est marquée en couleur bleue) ; le contexte de l'emploi apparait en bas du tableau (Fig.3).

Démonstration 1. *EXAKT* permet d'établir un profil linguistique de l'interlangue des élèves à partir des éléments morphologiques, syntaxiques ou lexicaux. Dans notre cas, nous voudrions chercher toutes les erreurs (déterminants ajoutés avant le nom, codés par conséquent comme $MOR $ADD). Ces codes se trouvent dans la couche des annotations. Nous faisons la requête dans *RegEx* : « ADD » afin d'obtenir les résultats (Fig.3). Au sein de cette transcription nous trouvons 8 incidences d'ajouts morphologiques, notamment le déterminant *THE*. Nous pouvons par la suite isoler *THE* à l'aide de filtrage manuel ou automatique.

Démonstration 2. Afin de relever les erreurs morphologiques liées au manque d'un élément dans l'énoncé, nous devons nous référer à la couche « Transcriptions », puisque le codage 0 (zéro) s'effectue directement dans la couche principale[29]. Autrement, la recherche est possible, comme dans l'exemple précédent, dans la couche des annotations au moyen de la requête : *RegEx* : « *LOS* ».

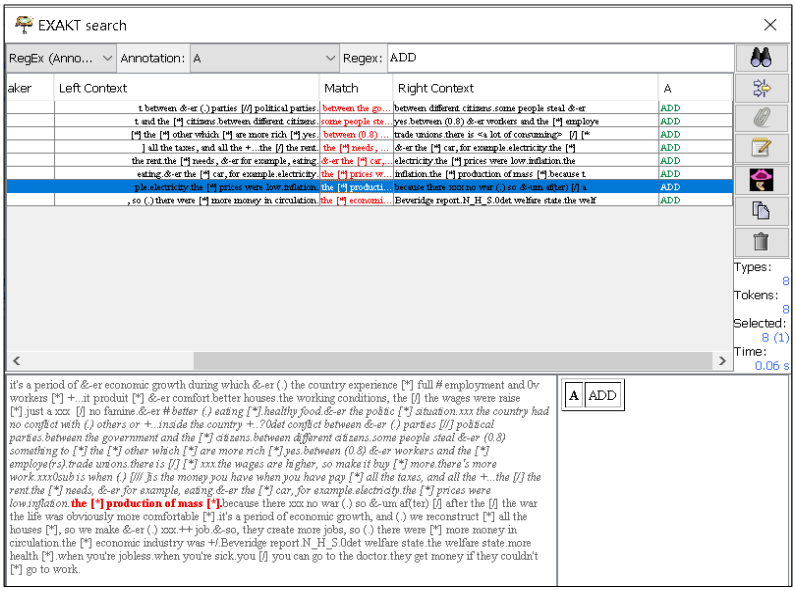


Figure 3. Recherche des erreurs morphologiques dans *EXAKT*

- Recherche lexicale. A l'aide du concordancier *KWIC* intégré dans *EXAKT* il est possible de faire des requêtes précises pour trouver le lexique dans les transcriptions.

[29] Il est à noter toutefois que le 0 (zéro) est également utilisé pour le codage de la longueur des pauses (ex. 0.7). Donc, il existe deux solutions : soit il convient de préciser quel type de manque est recherché (ex. **0det** pour la recherche d'un déterminant manquant), soit il faut signaler un manque dans la couche des annotations (lors de l'étape du codage des erreurs) par le code : $MOR $LOS en conformité avec la convention de transcription CHAT.

Ex. [Tt]h(is|at|ose| permet de trouver *this, that, those* et *these* (majuscules et miniscules)[30].

Ex. \bin[a-z]+abl[ey]\b correspond à la recherche des mots qui commencent par *in* et terminent par *able* ou *ably*, comme dans *indisputable, indescribably, ineffable*, etc.

Ex. (\[A-Za-z]+\b){3,3}\ ? fait ressortir toutes les séquences de trois mots suivis par un point d'interrogation, autrement dit, il s'agit des trois derniers mots des questions.

Parfois pour les besoins des analyses quantitatives il est nécessaire de comparer les chiffres relevant de plusieurs attributs des métadonnées. Afin de regrouper ces résultats, *EXAKT* permet d'appliquer les feuilles de style XSL à un résultat souhaité (Fig. 4). Les résultats sont groupés tout d'abord par le locuteur, après par l'âge du locuteur et finalement par le type de mot.

Age	Count	Types
6;4	3	weem (1) wenn (1) wer (1)
6;5	10	Wer (2) welche (4) wenn (2) wer (2)
6;9	1	Wer (1)
6;10	4	Wer (2) wenn (1) wer (1)

Figure 4. Quantification d'un résultat de recherche groupée[31]

Démonstration de recherche de lexique. La figure 5 montre la recherche des erreurs type « Lexique » (requête : *RegEx* : « *LEX* ») avec la représentation du mot pivot dans un contexte.

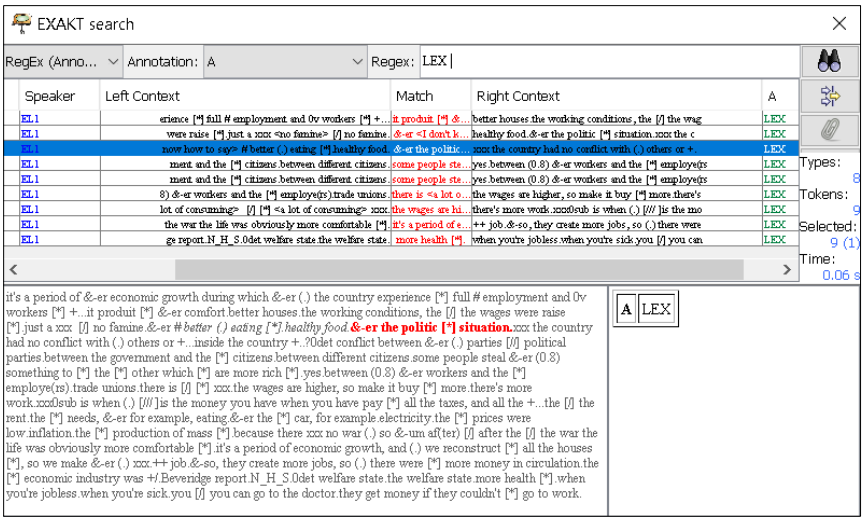


Figure 5. Recherche des erreurs lexicales dans *EXAKT*

Les résultats de la recherche se présentent sous la forme de l'expression recherchée qui apparait dans son contexte immédiat juste avant et après, prononcée par le même locuteur. Il est possible

[30] Exemples de *Schmidt* et *Wörner*, 2011.
[31] Figure de *Schmidt* et *Wörner*, 2011.

de filtrer le contexte à droite ou à gauche en cliquant sur la colonne correspondante (*right context* ou *left context* respectivement).

3.2. Exemples des interactions entre *l'input* et *l'output*

Grâce au module *EXAKT* il est possible non seulement d'établir le profil linguistique de l'interlangue des élèves, mais aussi d'effectuer le repérage des traits caractéristiques des relations entre ce qui est donné (enseigné) – *l'input* – et ce qui en ressort comme un résultat (appris) – *l'output.*
Une fonction de recherche des phénomènes particuliers (*search in events*) d'*EXMARaLDA* permet de découvrir une technique d'interaction du professeur de langue avec les élèves (Fig. 6).

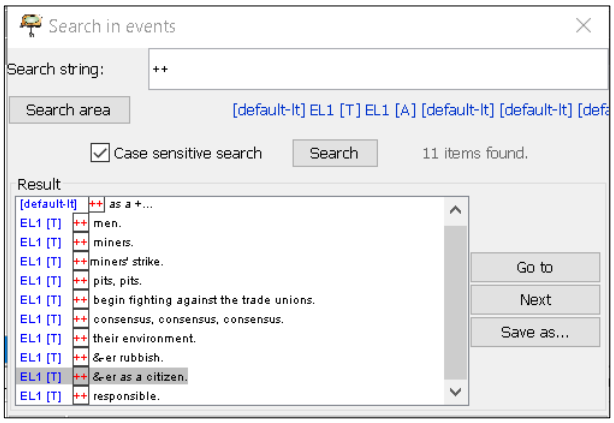


Figure 6. Recherche de téchnique *trailing off* dans *EXMARaLDA*

Nous avons ainsi repéré les différents types d'interaction entre *l'input* et *l'output* (Fig.7, 8, 9) :

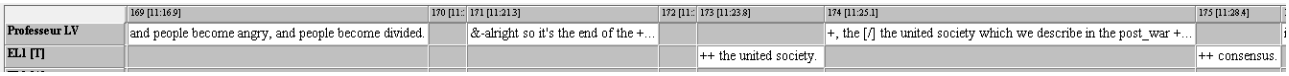


Figure 7. Interaction composée de quatre éléments (Professeur → Élève → Professeur → Élève)

Figure 8. Interaction composée de cinq éléments (Professeur → Élève → Professeur → Élève→ Professeur)

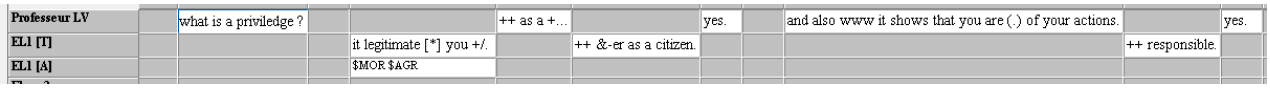


Figure 9. Interaction scindée (Professeur → Elève. Professeur → Elève. Professeur → Elève)

Dans l'extrait du cours analysé presqu'un tiers des énoncés produits par les apprenants (11 sur 35, Fig.6) est codé avec **++** [32] ce qui sous-entend un modèle particulier de l'interaction en cours. Cette technique d'interaction permet d'observer comment le professeur travaille l'aspect linguistique (apprentissage du vocabulaire) avec les élèves. En effet, le professeur de langue commence une réponse et l'abandonne à sa propre initiative proposant ainsi aux élèves une possibilité de la terminer avec un mot (ou une expression) attendu plutôt que de recourir à un schéma classique « question – réponse ». A chaque fois le professeur valide les réponses des élèves (voir Fig. 7, 8, 9).

Malgré plusieurs découpages, la phrase entière se devine facilement puisque construite autour de la même thématique : *A privilege legitimates you as a citizen and also it shows that you are responsible of (= for) your actions.* Cette phrase a un parcours particulier dans la mesure où sa construction passe par des techniques variées, dont l'abandon et la relance, en interaction étroite avec l'élève. L'interaction est guidée afin de faire émerger un terme attendu « citizen ». Le professeur se montre garant de l'énoncé complet et entier, puisqu'il en « soude » des segments constitutifs par le biais des mots de liaisons et de connecteurs (*as, and, also*). De cette façon, le professeur de langue s'assure de la maitrise par l'élève à la fois des compétences linguistiques et des connaissances civilisationnelles.

Dans sa tentative de guider les élèves vers la réponse précise *(responsible),* l'enseignant vise un double objectif : créer un contexte morpho-syntaxique reconnaissable par des apprenants (*you are [?] of your actions),* faisant appel non seulement à la connaissance des expressions phraséologiques (*be responsible for*) mais aussi au contexte de son évocation (extrait du discours de Margaret Thatcher tiré du film « Iron Lady » visionné juste avant).

Par ailleurs, l'interaction entre l'input et l'output se distingue grâce à l'emploi d'une technique particulière « phrase à trous » : l'enseignant choisit de marquer une pause dans un endroit où le terme précis est attendu, puis termine la phrase (*and also it shows that you are (.) of your actions)*. Il appartient donc aux élèves de restituer un mot absent à partir du contexte. Ceci est une technique classique d'apprentissage dans les supports écrits, or nous sommes ici dans une interaction orale.

La recherche des phénomènes dans *EXAKT* a permis de faire le constat suivant. Grâce à ces formes d'interaction variées *l'input* s'en trouve pour le moins doublé : le premier *input* pour la classe s'effectue lorsqu'un des élèves produit la réponse souhaitée, et le second quand le professeur reprend cette même expression. Le second *input* fait par le professeur valide l'existence du premier et ainsi ouvre le chemin pour continuer l'énoncé ; ensuite le schéma se répète avec un autre mot. Si les mêmes mots/expressions sont utilisés encore n-fois durant le cours, *l'input* par conséquent

[32] Code qui désigne une phrase terminée par quelqu'un d'autre que celui qui l'avait commencée.

se multiple par n-fois[33]. Par conséquent, l'acquisition de la terminologie et des expressions phraséologiques propres aux cours EMILE est favorisée à la fois par la répétition fréquente des mots (grâce aux multiples *inputs*) et par la participation active des apprenants dans le processus de l'apprentissage.

3.3. Applications possibles d'*EXMARaLDA* dans l'apprentissage des langues

Nous proposons dans le Tableau 1 quelques pistes non-exhaustives d'exploitation de l'outil *EXAKT* d'*EXMARaLDA* en vue de l'apprentissage des langues.

Fonctionnalité d'EXAKT	Exploitation possible dans la didactique des langues
Requêtes (*RegEx*) pour chercher les événements codés ou transcrits dans un corpus annoté de productions des élèves.	Les apprenants relèvent des erreurs (toutes ou par catégorie), proposent des corrections ; PRL peut être abordé par le professeur en lien avec la catégorie travaillée (phonétique, grammaticale, syntaxique, lexicale, stylistique, etc.).
Requêtes (*RegEx*) pour chercher les phénomènes dans un corpus de natifs anglophones. *KWIC* : contexte d'emploi des phénomènes recherchés.	Les apprenants repèrent les phénomènes (mots, déterminants, expressions phraséologiques, etc.), observent leur utilisation dans un contexte, en déduisent la signification et la/les règle(s) de l'emploi. Mise en commun et création de « pense-bêtes », *flashcards* avec des exemples tirés du corpus.
KWIC : contexte d'emploi des phénomènes recherchés. Filtrage et tri (automatique ou manuel) des résultats de recherche.	Créer des « textes à trous » en se basant sur le repérage des mots clés (*KWIC*). Chercher les mots clés, concevoir les fiches d'évaluation du vocabulaire à partir d'un contexte d'emploi (traduire les mots, compléter les amorces des énoncés, trouver les synonymes, etc.) .
Visionnage/ écoute de l'enregistrement qui correspond à un segment de la transcription. Corrélation des résultats de recherche avec des métadonnées.	Entrainement du lien graphie-phonie. Les élèves écoutent un/des segment(s) sans regarder la transcription, puis écrivent les mots qu'ils ont entendus ; vérification avec la transcription. Entrainement à CO : après avoir écouté un segment, les élèves font une carte mentale avec les mots clés, ensuite restituent le contenu/le contexte/ les repérages essentiels ; puis vérification avec les métadonnées (locuteur, date, etc.). Prononciation : les élèves répètent le mot/son travaillé après le locuteur natif, trouvent dans le corpus d'autres mots qui contiennent le même son. Travail sur les accents variés grâce à la sélection des locuteurs souhaités.
Sauvegarde des résultats des recherches. Exportation dans d'autres logiciels.	Sauvegarder les résultats (mots clés), les exporter au format Excel et imprimer ; les élèves s'en servent comme un support pour mémoriser le vocabulaire.

Tableau 1. Utilisation d'*EXMARaLDA* dans l'apprentissage des langues

[33] Ces exemples cités sont des variantes du schéma classique IRF de Sinclair et Coulthard (1975).

4. Avantages et limites

EXMARaLDA dispose d'avantages certains :

- logiciel gratuit, téléchargeable en ligne et accompagné d'un manuel complet ;
- compatibilité avec le système d'exploitation Windows ;
- interopérabilité des formats qui permet un transfert facile des données vers d'autres logiciels si besoin est (voir la présentation d'*EXMARaLDA*, section 2) ;
- segmentation automatique des énoncés qui rend la transcription moins chronophage (voir la description de *Partitur-Editor*, section 2.1.) ;
- configuration possible en tenant compte des exigences d'autres logiciels[34] ;
- interface intuitive, facile à maîtriser même pour les utilisateurs non aguerris (voir Fig.2) ;

Cependant, il est à noter quelques difficultés :

- 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*) ;
- il n'est pas possible d'effectuer des analyses morphologiques approfondies[35].

Par ailleurs, des codes attribués[36] (ex. catégorisation) peuvent avoir une incidence sur le calcul des erreurs. Ainsi il faut se garder de l'envie d'utiliser tous les codes possibles en même temps pour la même erreur. De même, il est préférable de coder les erreurs dans la ligne secondaire (couche des annotations) afin de standardiser les types d'erreurs et ainsi d'en faciliter la recherche dans *EXAKT*.

Conclusion

Dans notre article nous avons présenté le système de transcription, de traitement et d'analyse de données multimodales *EXMARaLDA*. Les fonctions principales ont été décrites et accompagnées par des démonstrations. Nous avons également proposé des applications possibles dans les domaines de la linguistique de corpus ainsi que dans la didactique des langues. Nous espérons par notre travail répondre de façon efficace aux exigences des linguistes, chercheurs et praticiens sensés opérer des corpus volumineux contenant des données empiriques.

---

[34] 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*.

[35] Pour ce faire il possible d'utiliser, par exemple, le logiciel tiers *CLAN*.

[36] Par exemple, le code 0 (zéro) dans la couche des transcriptions et le code $MOR $LOS dans la couche des annotations signifient le même type d'erreur (manque) dans le même mot et par conséquent seront calculés par *CLAN* comme deux fautes morphologiques, bien qu'il s'agisse de la même erreur.

BIBLIOGRAPHIE

Baggioni, D. (1995). Normalisation/standardisation des langues nationales dans l'espace européen. *Archives et documents de la Société d'histoire et d'épistémologie des sciences du langage, 11*, 73-86.

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.

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

James, C. (1998). *Errors in language learning and use*. *Exploring error analysis*. Longman.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. Transcription format and programs* (vol. 1). Lawrence Erlbaum Associates.

Rohlfing, K., Loehr, D., Duncan, S., Brown, A., Franklin, A., Kimbara, I., Milde, J.-T., Parrill, F., Rose, T., Schmidt, T., Sloetjes, H., Thies, A. et Wellinghoff, S. (2006). Comparison of multimodal annotation tools. *Gersprachsforschung, 7*, 99-123. http://www.gespraechsforschung-ozs.de/heft2006/tb-rohlfing.pdf

Schmidt, T., Al-Jaf, T., Ferger, A., Frontzek, C., Kaminska, K. et Sambale, H. (2016). *Exmaralda. Partitur-Editor 1.6.* https://www.exmaralda.org/pdf/Partitur-Editor_Manual.pdf

Schmidt, T. (2010). *Exmaralda. EXAKT 1.* http://www.exmaralda.org/files/EXAKT_Manual.pdf

Schmidt, T. et Wörner, K. (2011). Creating, analysing and sharing spoken language corpora for pragmatic research. *Pragmatics, 19*(4).

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.

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

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

SITOGRAPHIE

Laurence Anthony's Homepage. https://laurenceanthony.net

CID.ens-lyon.fr. http://cid.ens-lyon.fr/http://bcl.cnrs.fr/

CQPweb. https://cqpweb.lancs.ac.uk/

ECAMPUS UNICAEN. https://ecampus.unicaen.fr

ELAN | The Language Archive. https://archive.mpi.nl/tla/elan

Exploration de corpus : outils et pratiques. http://explorationdecorpus.corpusecrits.huma-num.fr/

Sketch Engine. https://www.sketchengine.eu/

EXMARaLDA. https://www.exmaralda.org/

Secteur TAL Informatique | Universite Paris 3 Sorbonne nouvelle. http://www.tal.univ-paris3.fr/outils-cla2t.html

CLESTHIA EA 7345 – Université Sorbonne Nouvelle Paris 3 (s.d.). *Le Trameur aka Le Métier Textométrique.* Le trameur. http://www.tal.univ-paris3.fr/trameur/

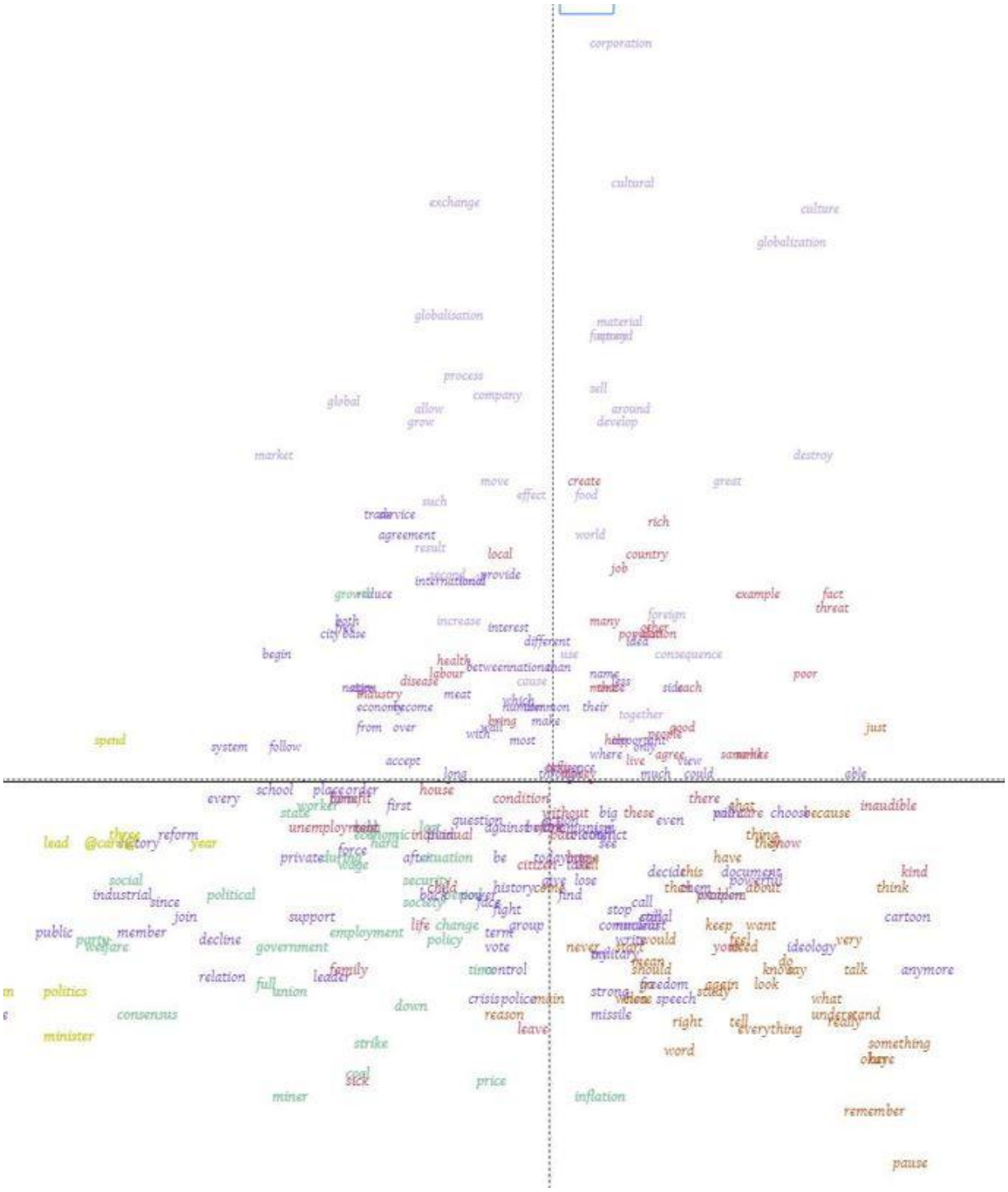
Lextutor. https://www.lextutor.ca/

UAM CorpusTool Homepage. http://www.corpustool.com/

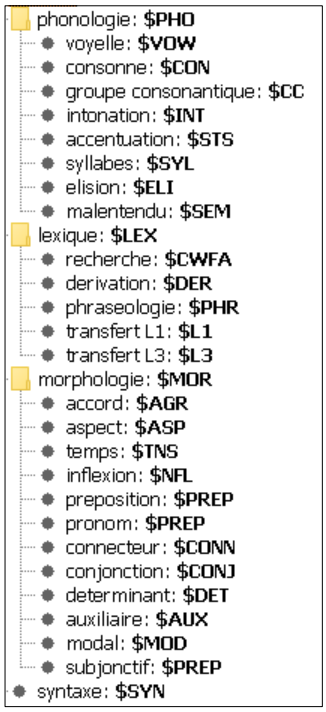
Bases, Corpus, Langues UMR 7320 – Université Nice Sophia Antipolis. Hyperbase 10. http://ancilla.unice.fr/

Annexes

Annexe 1. Coloration thématique de l'Hyperbase



Annexe 2. Panel d'annotation des erreurs intégré dans l'EXMARaLDA



Annexe 3. Codage des disfluences principales dans EXMARaLDA

Stuttering behavior	**Code**	**Example**	**Notes**
Prolongation	:	s:paghetti	Place after prolonged segment
Broken word	^	spa^ghetti	New code
Block	Unicode2260 ("not equal to" sign); *shortcut:* hold F2 and =	≠butter	This example illustrates a block before word onset
Repeated segments	21AB (curly left arrow); *shortcut:* hold F2 and /	↫r-r-r↫rabbit OR like↫ike↫	The curly left arrow brackets the repetition but leaves a recognizable target for mor; iterations inside of the sequence are marked with hyphens
phonological fragment	&+	&+sn dog	Changes from "snake" to "dog"
other non-word strings	&	&gara	Word play etc.
Typical Disfluencies			
Whole word repetition	follow word with [/]	butter [/] butter	Repeated word counts once
Multiple whole word repetition	indicate number of repetitions in brackets	butter [x 7] x space N	Indicates that the word 'butter' was repeated seven times
Phrase repetitions	<> [/]	<that is a> [/] that is a dog.	Repeated phrase counts once
Phrase revisions	<> [//]	<what did you> [//] how can you see it ?	Revised phrase counts once
pause	(.) or (..) or (…)	(.)	Counts the number of short, medium, long pauses
pause duration	(2.4)	(2.4)	Adds up the time values, if marked
Filled pauses	&-	&-um &-you_know	Note: multiword fillers should be connected with an underscore to avoid wrong word count