Un concordancier multi-niveaux et multimédia pour des corpus oraux

Giulia Barreca[1] George Christodoulides[2]
(1) Laboratoire MoDyCo, CNRS, Université Paris Ouest Nanterre La Défense
200, avenue de la République, FR-92001 Nanterre, France
(2) Centre Valibel, Institut Langue & Communication, Université de Louvain,
Place Blaise Pascal 1, 1348 Louvain-la-Neuve, Belgique
giulia.barreca@gmail.com, george@mycontent.gr

Résumé. Les concordanciers jouent depuis longtemps un rôle important dans l'analyse des corpus linguistiques, tout comme dans les domaines de la philologie, de la littérature, de la traduction et de l'enseignement des langues. Toutefois, il existe peu de concordanciers qui soient capables d'associer des annotations à plusieurs niveaux et synchronisées avec le signal sonore. L'essor des grands corpus de français parlé introduit une augmentation des exigences au niveau de la performance. Dans ce travail à caractère préliminaire, nous avons développé un prototype de concordancier multi-niveaux et multimédia, que nous avons testé sur le corpus de français parlé du projet Phonologie du Français Contemporain (PFC, 1,5 million de tokens de transcription alignée au niveau de l'énoncé). L'outil permet non seulement d'enrichir les résultats des concordances grâce aux données relevant de plusieurs couches d'annotation du corpus (annotation morphosyntaxique, lemme, codage de la *liaison*, codage du *schwa* etc.), mais aussi d'élargir les modalités d'accès au corpus.

Abstract. Concordances have always played an important role in the analysis of language corpora, for studies in humanities, literature, linguistics, translation and language teaching. However, very few of the available systems support multi-level queries against a richly-annotated, sound-aligned spoken corpus. The rapid growth in the development of spoken corpora, particularly for French, increases the need for scalable, high-performance solutions. We present the preliminary results of our project to develop a multi-level multimedia concordancer for spoken language corpora. We test our prototype on the PFC corpus of spoken French (1.5 million tokens, transcriptions aligned to the utterance level). Our tool allows researchers to query the corpus and produce concordances correlating several annotation levels (part-of-speech tags, lemmas, annotation of phonological phenomena such as the *liaison* and *schwa*, etc.) while allowing for multi-modal access to the associated sound recordings and other data.

Mots-clés : concordancier, annotation multi-niveaux, linguistique de corpus, didactique du FLE
Keywords: concordance tool, multi-level annotation, corpus linguistics, French language teaching

1 Introduction

L'utilisation d'outils informatiques tels que les concordanciers, qui permettent d'extraire des collocations de mots et leur contexte, est fréquente dans la communauté des chercheurs en sciences du langage. La taille de corpus actuellement disponibles est en augmentation constate, de même que les schèmes d'annotation, ce qui entraîne la création de nouveaux systèmes pour interroger ces données. Dans les pages qui suivent, après avoir décrit brièvement quelques-uns des travaux menés sur les concordanciers, nous présentons un prototype de concordancier multi-niveaux et multimédia que nous sommes en train de développer et tester sur les données du corpus de français parlé PFC (Phonologie du Français Contemporain, cf. Durand et al. 2002). Notre objectif est de nous rapprocher le plus possible des besoins des linguistes et des apprenants, qui souhaitent dépouiller de façon rapide et précise les contextes d'occurrences d'un lexème ou d'une séquence de lexèmes. Notre outil, contrairement à d'autres concordanciers, intègre en effet l'écoute des extraits d'enregistrements au travail de recherche des concordances. Il permet de lancer des requêtes croisant plusieurs couches (niveaux) d'annotation alignées temporellement avec le signal. Le moteur de recherche du concordancier construit de manière dynamique des requêtes SQL vers une base de données relationnelle, dans laquelle sont stockées les annotations. Dans le cas du corpus PFC, ces annotations associées à la transcription sont de nature morphosyntaxique (POS, lemme) et phonologique (codage des phénomènes de la liaison et du schwa). Nous montrons les applications possibles de l'outil dans les différents domaines de la linguistique et d'enseignement du FLE, son architecture et quelques-unes de ses fonctionnalités.

2 Synthèse des études précédentes

Le concordancier a toujours été un outil de grand intérêt pour l'analyse des contextes d'emploi des lexèmes dans des corpus oraux ou écrits. Ses applications théoriques et pratiques sont nombreuses, et intéressent plusieurs disciplines : philologie, littérature (Caballero 1999, Magri-Mourgues 2006), syntaxe et sémantique (Gross 2000), traduction (Jacquet-Pfau 1994) et didactique des langues (Tognini-Bonelli 2001). Pourtant, comme le soulignent Pincemin et al. (2006), la plupart des concordanciers disponibles n'exploitent pas les informations linguistiques introduites par l'annotation des corpus (p. ex. catégories morphosyntaxiques, lemmatisation, etc.). Les résultats produits par ce genre de concordancier n'aboutissent qu'à de simples collections d'occurrences (p.ex. Lextutor et Lexiqum). Plusieurs concordanciers existent déjà pour les corpus écrits : parmi ceux-ci nous pouvons citer Condor (développé par le Laboratoire lorrain de recherche en informatique – LORIA), qui permet d'effectuer des recherches simples ou de cooccurrences utilisant le mot-forme ou le lemme, et d'indexer les formes présentes dans le corpus. La version publiquement disponible propose une collection des corpus écrits, pour la plupart littéraires. Le concordancier TXM (Heiden et al. 2010) permet de produire des concordances KWIC (*keyword in context*) à partir des propriétés des mots (ex. lemme, catégorie POS) ou de lancer de requêtes plus complexes grâce à une syntaxe propre à son moteur de recherche.

L'intervention des informations de nature linguistique, et donc l'affinement des résultats des concordances, demande un traitement préalable du corpus qui est souvent très couteux, comme c'est par exemple le cas du corpus French Treebank (Abeillé 2003). Pour cette raison, de nombreux concordanciers lemmatisés font appel à des dictionnaires. Le recours à ce type de ressources linguistiques permet notamment de mettre en relation les occurrences du texte avec les lemmes et les formes fléchies contenues dans les dictionnaires. Parmi les concordanciers qui utilisent ce type d'outils linguistiques, nous pouvons citer *Unitex* (Paumier 2002) et *Stella* (Bernard et al. 2002).

Toutefois, l'application des concordanciers conçus pour l'écrit à l'analyse des corpus oraux pose un problème de fond : ces outils ne permettent pas de tenir compte de l'alignement des annotations linguistiques avec le signal sonore. Les requêtes possibles sont ainsi limitées à ce qu'on peut représenter avec les conventions de transcription choisies. À notre connaissance, les concordanciers lemmatisés pour l'analyse de corpus de français parlé sont peu nombreux. Mis à part le concordancier de la base lexicale *Lexique 3* (New 2006), basé sur les transcriptions des sous-titrages de films, les autres concordanciers disponibles pour les corpus oraux[1] ne permettent d'obtenir que des relevés des occurrences et des données statistiques sur leurs collocations. Dans ce contexte, nous considérons qu'un concordancier multimédia spécifique pour l'analyse d'un corpus oral pourrait permettre aux utilisateurs de mieux prendre en compte l'articulation des différents niveaux d'annotation dans le temps et d'exploiter les résultats des requêtes tout en gardant le lien avec le signal sonore.

3 Objectif

Notre objectif est de développer un concordancier multi-niveaux et multimédia, exploitable pour des recherches linguistiques en français parlé. Pour ce faire nous avons utilisé le corpus PFC (Phonologie du Français Contemporain) (Durand et al. 2002) annoté en morphosyntaxe (POS et lemmes) à l'aide de l'étiqueteur *DisMo* (Christodoulides et al. 2014). Ce corpus, qui se compose de 36 points d'enquête pour un total de 394 locuteurs, constitue une ressource de grand intérêt pour tout linguiste intéressé au français parlé, grâce à la variété des origines géographiques des locuteurs enregistrés, des styles et de registres de parole, mais également des annotations des phénomènes phonologiques spécifiques au français parlé (liaison, schwa).

Nous avons essayé de créer un outil qui permette établir une articulation entre ces différents niveaux d'analyse, et qui puisse ainsi offrir une observation linguistique plus précise des contextes d'occurrence d'un phénomène linguistique donné (phonologique, syntaxique, lexicale ou prosodique). L'ensemble de ces contextes regroupés peut permettre de révéler la présence de régularités liées à certaines combinaisons sémantiques, syntaxiques et discursives des mots. Plusieurs études ont déjà mis en évidence l'influence de la fréquence sur les représentations cognitives (Bybee & Hopper 2001), sur la morphologie et la syntaxe (Bybee & Thompson 1997), sur la phonologie (Bybee 2001), sur l'acquisition (Tomasello 2000) et évidemment sur les procès de grammaticalisation (Bybee & Scheibman 1999). En particulier cette dernière application pourrait être d'un intérêt particulier dans l'étude du procès de pragmaticalisation (Dostie 2004) des marqueurs discursifs du français parlé.

[1] Parmi lesquels ceux des corpus CFPP2000 (Corpus de Français Parlé Parisien), CLAPI (Corpus de Langue Parlée en Interaction) et OFROM (Corpus Oral de Français parlé en Suisse Romande).

L'utilisation d'un concordancier lemmatisé sur un corpus annoté pourrait aussi permettre d'extraire de façon automatique les contextes d'occurrences des marqueurs discursifs, et de repérer leurs positions prototypiques ainsi que leurs combinaisons possibles et effectives (p. ex. *« enfin bon »*), ceci afin de mieux saisir leurs fonctions pragmatiques en discours. De plus, l'utilisation d'un concordancier lemmatisé permettrait de développer une analyse multidimensionnelle des phénomènes phonologiques spécifiques au français parlé, phénomènes dont la réalisation est influencée par des facteurs relevant de l'articulation de plusieurs dimensions linguistiques (phonologique, syntaxique, lexicale, prosodique etc.). C'est le cas par exemple du phénomène de la liaison. Plus précisément une telle analyse permettrait de vérifier l'influence des facteurs lexicaux (fréquence des mots liaisonnés et de leur cooccurrence) sur la réalisation de la liaison, tout en donnant accès aux données nécessaires pour une analyse phonologique et syntaxique de la séquence liaisonnée. L'utilisation d'un concordancier lemmatisé pourrait aussi permettre de mieux analyser le rôle joué par la fréquence d'usage des mots et leur fréquence distributionnelle dans certaines constructions où ni la liaison, ni le figement, ne sont prévisibles. C'est le cas par exemple des séquences *temps en temps* vs *temps/à autres* (Laks 2005). Nous considérons de façon plus générale qu'un tel outil permettra de placer la notion d'usage (Bybee & Hopper 2001) au centre de la réflexion linguistique.

En ce qui concerne la didactique du FLE, l'application d'un concondancier lemmatisé du français parlé constituerait une ressource précieuse pour l'enseignement, favorisant un apprentissage basé sur des données authentiques et conçu comme une recherche (*data-driven learning*), dans la lignée des travaux réalisés dans le cadre du sous-projet PFC-EF (PFC Enseignement du Français, Detey et al. 2010). Une telle approche fournirait aux apprenants une observation directe de la langue, qui aiderait à rapprocher les productions des apprenants des usages réels des natifs. Dans les paragraphes suivants, nous illustrons plus en détails le fonctionnement du concordancier.

4 Architecture du système

Le concordancier fonctionne à partir d'une base de données relationnelle (SQLite, MySQL ou PostgreSQL), dans laquelle les annotations du corpus à plusieurs niveaux sont représentées sous le schéma du logiciel *Praaline* (Christodoulides 2014). Chaque niveau d'analyse correspond à une table, et chaque annotation de ce niveau à une colonne de la table. Des identifiants uniques des échantillons du corpus (`AnnotationID` et des locuteurs (`SpeakerID`) renvoient aux métadonnées. Les intervalles temporels sont représentés avec deux nombres entiers de 64 bits, en nanosecondes (`tMin`, `tMax`). Ce simple schéma permet de représenter des relations hiérarchiques entre deux ou plusieurs couches d'annotation, ainsi que les éventuels chevauchements sur la même couche. Il permet aussi de profiter du système d'indexation de la base de données pour améliorer la performance des requêtes, ou appliquer des restrictions sur les annotations (p. ex. un vocabulaire restreint au jeu d'étiquettes morphosyntaxiques est appliqué au champ `pos_min`). Un extrait du schéma est affiché ci-dessous (Figure 1).

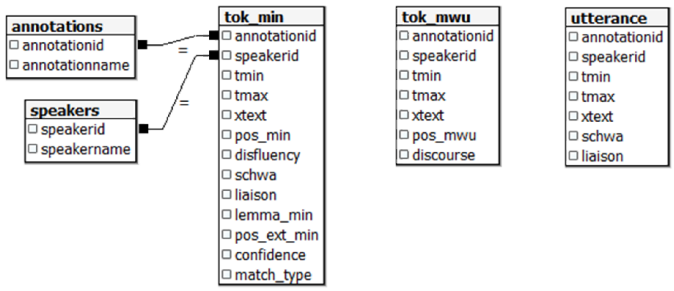

FIGURE 1 Extrait du schéma de la base de données. Les tables qui représentent les couches d'annotation (`tok_min`, `tok_mwu`, `utterance`) ont comme clé primaire la triplette (`annotationID`, `speakerID`, `tMin`).

Plusieurs méthodes ont été proposées pour le stockage des annotations d'un corpus, notamment sous la forme de fichiers XML, de bases de données NoSQL, ou même de structures de données propres à des logiciels d'annotation. Notre choix d'une base de données relationnelle est motivé par le fait que les annotations textuelles du corpus sont essentiellement structurées. Les indices temporels permettent au logiciel d'exploitation de corréler ces annotations avec le signal sonore (qui est stocké au format WAV), et avec d'autres annotations stockées au format HDF5. Ce dernier est le format de préférence pour les annotations qui impliquent un grand nombre de vecteurs à plusieurs dimensions, avec une fréquence d'échantillonnage élevée (p. ex. une série de mesures acoustiques et prosodiques).

Le logiciel que nous avons développé permet de construire de manière dynamique les requêtes SQL adéquates, selon les choix de l'utilisateur. Par exemple, la requête suivante renvoie les annotations associés à tous les 4-grammes « c'est-à-dire » (temps d'exécution : 649 ms sous SQLite 3.8.1, CPU i7 2.4 GHz, mémoire 24 Go) :

```
SELECT * FROM tok_min AS T1, tok_min AS T2, tok_min AS T3, tok_min AS T4
WHERE   T1.AnnotationID = T2.AnnotationID AND T1.SpeakerID = T2.SpeakerID AND T1.tMax = T2.tMin
 AND    T2.AnnotationID = T3.AnnotationID AND T2.SpeakerID = T3.SpeakerID AND T2.tMax = T3.tMin
 AND    T3.AnnotationID = T4.AnnotationID AND T3.SpeakerID = T4.SpeakerID AND T3.tMax = T4.tMin
 AND    T1.xText = 'c''' AND T2.xText = 'est' AND T3.xText = 'à' AND T4.xText = 'dire'
```

Une interface graphique permet à l'utilisateur de définir sa requête dans le corpus, qui est traduite à des requêtes de la base de données. Le système du moteur de recherche et du concordancier est indépendant du corpus : la structure de l'annotation est définie préalablement, et le schéma de la base de données est dynamique (modifié par le logiciel quand une couche d'annotation est ajoutée ou supprimée). L'outil utilise cette définition afin de créer les jonctions entre les tables qui représentent les différentes couches d'annotation.

5 Illustration: recherche de concordances lemmatisées et multimédia

Dans les paragraphes suivants nous allons présenter une démonstration des requêtes qui peuvent être réalisées à partir du prototype de notre concordancier. Nous allons montrer des extraits des tableaux des occurrences résultant de trois différents types de recherche: (a) par lexème; (b) par lemme; (c) par chaîne de caractères. Enfin nous illustrerons les résultats de la recherche de la cooccurrence de la séquence *je suis* et un exemple de l'affichage des résultats des concordances multimédia correspondantes.

Dans les exemples ci-dessous nous pouvons observer les résultats de la recherche de *suis* en tant que lexème et en tant que chaîne de caractères et en tant que lemme du verbe *être*. Les occurrences calculées sont présentées de façon synthétique par des tableaux avec, en colonne, de gauche à droite : le lexème, l'étiquette POS, le lemme, le nombre d'occurrences, la fréquence du lexème et la fréquence du lemme. De plus, chaque tableau est précédé par l'indication du nombre total des occurrences toute catégories confondues.

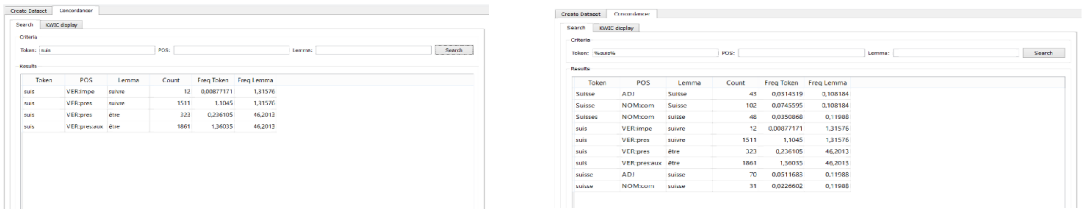

FIGURE 2 Accès à la concordance par tableau des résultats synthétiques de la requête du lexème *suis* (gauche) et de la chaîne de caractères %suis% (droite)

Il en va de même pour la recherche des cooccurrences par lexème et par lemme. Dans l'exemple ci-dessous, nous observons pour la séquence *je suis* deux différentes typologies de cooccurrences selon le lemme correspondant (*être* vs *suivre*).

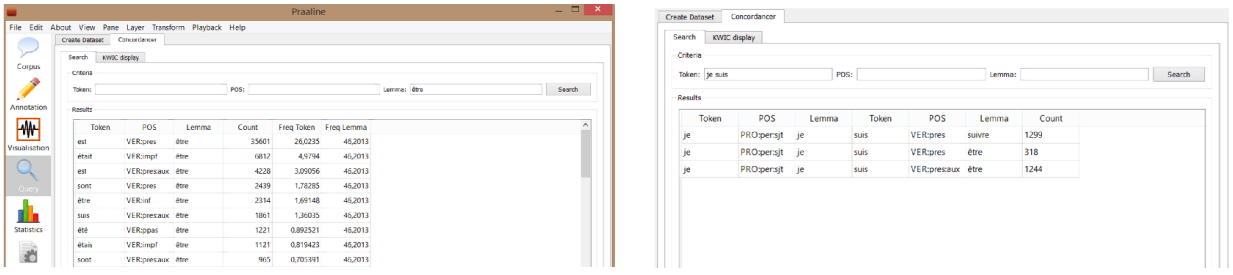

FIGURE 3 : Accès à la concordance par tableau des résultats synthétiques (1) du lemme *être* (gauche), (2) de la chaîne de caractères *je suis* (droite).

Ensuite l'ensemble de ces tableaux donne accès, par des liens hypertextes sur chaque ligne, aux extraits des concordances correspondants. Par exemple, si l'on choisit la deuxième typologie de la cooccurrence *je suis* (lemme *être*, Fig. 3, droite), nous pouvons accéder, entre autres, à la concordance suivante :

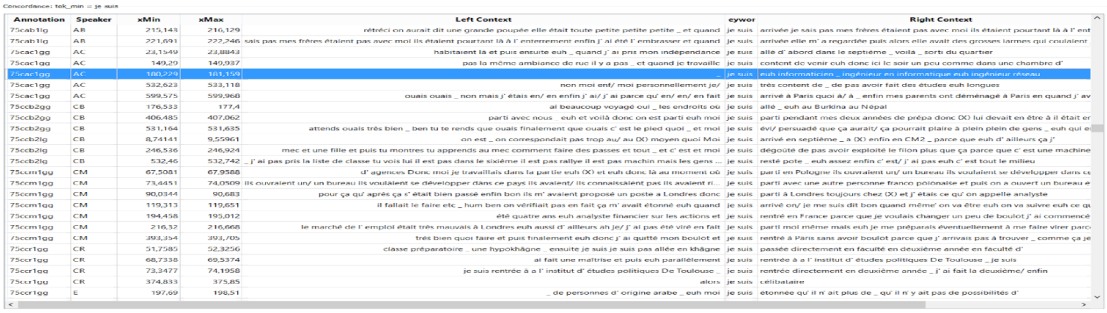

FIGURE 4 : Concordance « *AC : je suis11z euh informaticien. E : Hum hum. AC : Ingénieur en informatique euh ingénieur réseau voilà* » (Corpus PFC_locuteur75cac_discussion guidée).

Enfin un lien multimédia sur chaque pivot (ex. *je suis*) permet de visualiser (à l'aide du logiciel *Praat*, Boersma & Weenink 2014) l'intervalle de la transcription de l'enregistrement contenant les occurrences ou les cooccurrences des mots ciblés, d'accéder aux différentes couches d'annotation et de codage du corpus PFC. Dans l'exemple ci-dessous, plusieurs couches sont associées à la concordance de la séquence *je suis*: (a) transcription de l'intervalle contenant la concordance ciblée ; (b) codage de la *liaison*; (c) codage du *schwa*; (d) étiquettes d'annotation morphosyntaxique.

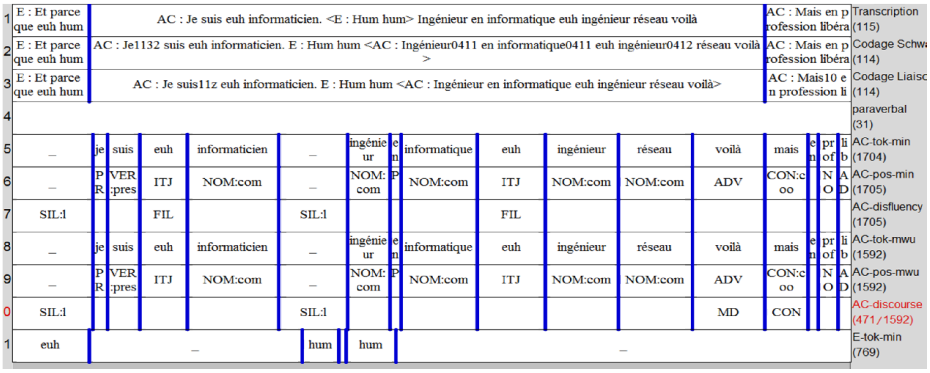

FIGURE 5 : Accès à l'annotation multi-niveaux (au format *TextGrid*) à partir de la concordance.

6 Conclusion et perspectives

Dans cette étude à caractère préliminaire, nous avons présenté les caractéristiques principales d'un prototype de concordancier lemmatisé et multimédia, basé sur le corpus de français oral PFC. Nous cherché à montrer l'intérêt de la création d'un tel concordancier lemmatisé pour le français oral, où les informations linguistiques sont exploitées afin d'améliorer le rappel et la précision des concordances, et de répondre aux besoins de dépouillage précis dans le cadre de la linguistique de corpus. Nous avons utilisé les fonctions traditionnelles des autres concordanciers, tout en proposant un enrichissement des résultats des requêtes grâce à l'apport des données relevant de l'annotation morphosyntaxique du corpus, et du codage de phénomènes phonologiques tels que la liaison et le schwa. De plus, la possibilité d'accéder aux écoutes des intervalles des enregistrements contenant les concordances permet d'élargir les modalités d'accès à l'analyse des contextes d'occurrences des mots ciblés. Dans le cadre de notre travail, nous envisageons d'améliorer l'affichage des annotations supplémentaires dans les concordances détaillées, permettre la sauvegarde des requêtes plus compliquées, d'ajouter une passerelle entre le système de requêtes (concordancier) et le logiciel d'analyse statistique *R*, et éventuellement mettre à les outils développés à la disposition de la communauté (en source libre). L'ensemble de fonctionnalités offertes fait du concordancier un outil fondamental pour l'analyse des corpus linguistiques pour toute discipline des sciences du langage ainsi que pour les domaines de la traduction et de la didactique du FLE.

Références

BERNARD, P., LECOMTE, J., DENDIEN, J., PIERREL, J-M. (2002). Computerized linguistic resources of the research laboratory ATILF for lexical and textualanalysis: Frantext, TLFi, and the software Stella Actes de Language Resources and Evaluation (LREC 2002), 1090-1096, Las Palmas, Espagne.

BOERSMA, P., WEENINK, D. (2014). *Praat: doing phonetics by computer*, ver. 5.3.77, www.praat.org

BOKAN, N. éditeur (2000). *Contemporary Mathematics. Proceedings of the Symposium*, Belgrade.

BYBEE, J. (2001). Frequency effects on French liaison. In (Bybee, Hopper, 2001), 337-359.

BYBEE, J., SCHEIBMAN, J. (1999). The effect of usage on degrees of constituency: the reduction of *don't* in English. *Linguistics* 37(4): 575-596.

BYBEE, J., HOPPER, P., éditeurs (2001). *Frequency and the Emergence of Linguistic Structure*. Amsterdam : John Benjamins.

BYBEE, J., THOMPSON, S. (1997). Three frequency effects in syntax. *Pragmatics and Grammatical Structure* 23: 378-388.

CABALLERO, M. R. (1999). Using a Concordancer in Literary Studies. *TheEuropean English Messenger* 8(2): 59-62. http://www.edict.com.hk/Concordance/

CHRISTODOULIDES, G. (2014). Praaline: Integrating tools for speech corpus research. Actes de *IX Language Resources and Evaluation Conference (LREC 2014)*, 26-31 mai, Reykjiavic, Islande.

CHRISTODOULIDES, G., AVANZI, M., GOLDMAN, J.P. (2014). DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. Actes de *IX Language Resources and Evaluation Conference (LREC 2014)*, 26-31 mai, Reykjavik, Islande.

DETEY, S., DURAND, J., LAKS, B., LYCHE, C. (2010). *Les variétés du français parlé dans l'espace francophone: ressources pour l'enseignement.* Paris: Ophrys.

DOSTIE, G. (2004). *Pragmaticalisation et marqueurs discursifs. Analyse sémantique et traitement lexicographique*. Bruxelles: Duculot

DURAND, J., LAKS, B., LYCHE, C. (2002). La phonologie du français contemporain: usages, variétés et structure. In (Pusch et Raible, 2002), 93-106.

GROSS, M. (2000). A bootstrap method for constructing local grammars. In (Bokan, 2000), 229-250.

HEIDEN, S., MAGUÉ, J-P., PINCEMIN, B. (2010). *TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement*. In. I. C. Sergio Bolasco (Ed.), Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010) (Vol. 2, p. 1021-1032). Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy.

JACQUET-PFAU, C. (1994). L'intérêt des logiciels de concordances pour la traduction. *Langages* 116: 82-86.

LAKS B. (2005). Phonologie et construction syntaxique: la liaison, un test de figement et de cohésion. *Linx* 53, 155-171.

NEW, B. (2006). Lexique 3 : Une nouvelle base de données lexicales**.** *Actes de la Conférence Traitement Automatique des Langues Naturelles (TALN 2006),* avril 2006, Louvain, Belgique.

PAUMIER, S. (2003). A Time-Efficient Token Representation for Parsers. Actes de EACL Workshop on Finite-State Methods in Natural LanguageProcessing, 83–90, Budapest, Hongrie.

PINCEMIN, B., ISSAC, F., CHANOVE, M., MATHIEU-COLAS, M. (2006). Concordanciers: Thème et variations, *Lexicometrica*, numéro spécial, 769-780. Actes des Journées d'analyse statistiques des données textuelles (JADT).

PUSCH, C., RAIBLE, W., éditeurs (2002). Romanistische Korpuslinguistik- Korpora und gesprochene Sprache/Romance Corpus Linguistics - Corpora and Spoken Language. Gunter Narr Verlag.

TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work*. Amsterdam/Philadelphia : John Benjamins.

TOMASELLO, M. (2000). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics* 11 : 61-82.