Journée d'études : Annoter l'oral

Organisée par Lotfi Abouda, Flora Badin, Florence Lefeuvre
CLESTHIA EA 7345 & LLL UMR 7270

19 novembre 2021
Maison de la Recherche, salle du Conseil, 4 rue des Irlandais, 75 005 Paris.

Argumentaire

Les annotations de corpus écrits axées sur des thématiques linguistiques sont nombreuses et performantes depuis des années (Pour une présentation et une évaluation de certains d'entre eux, cf. par exemple Neves et Seva 2021). En ce qui concerne l'oral spontané, le chemin a été plus long (cf. pour une vue d'ensemble, Bergounioux et *al.* 2017). Si les corpus oraux ont émergé depuis les années 60-70 (Enquêtes Sociolinguistiques à Orléans Eslo 1, Corpus de Montréal, Valibel), leur mise à disposition n'a été possible que depuis une quinzaine d'années (Eslo 2, CFPQ (Corpus de Français Parlé au Québec), CFPP / CFPB (Corpus de Français parlé parisien / Corpus de Français Parlé à Bruxelles), MPF (Multicultural Paris French), OFROM (Corpus Oral de Français de Suisse Romande), CLAPI (Corpus de LAngue Parlée en Interaction). A partir de ces bases de données orales, plusieurs corpus structurés ont vu le jour : corpus Rhapsodie (Lacheret et *al.* 2014), corpus Orféo (Benzitoun et *al.* 2016), corpus ESLO-MD (Abouda et Skrovec 2018). Dépendantes de la constitution de corpus oraux, les annotations de ces corpus (manuelles et automatiques) sont plus récentes. Des annotations en lemmes, catégories grammaticales et fonctions syntaxiques sont nées à partir de différents projets de recherche (Rhapsodie, Orféo). La segmentation en unités est au cœur de ces problématiques (cf Rossi-Gensane et *al.* 2019 et le projet SegCor). On peut citer également le projet LOCAS-F : un Corpus Oral Multigenres Annoté (Degand et *alii* 2014), qui propose d'annoter des corpus en fonction de l'unité discursive de base résultant de la corrélation entre unités prosodiques et unités syntaxiques. L'annotateur multi-niveaux DisMo quant à lui permet d'annoter des corpus oraux, il propose un étiquetage morphosyntaxique, une lemmatisation, une détection des unités poly-lexicales, une détection et annotation des phénomènes de disfluence et des marqueurs de discours, ainsi qu'un découpage en unités syntaxiques minimales (cf. Christodoulides et Barreca 2017). Des phénomènes ciblés ont pu être annotés, comme par exemple les « reformulations paraphrastiques » à partir d'un sous-corpus d'Eslo (Eshkol 2015). Des outils se développent pour décrire linguistiquement des corpus oraux, c'est le cas du logiciel TXM (Badin et *al.* 2021). Des logiciels sont déjà spécialisés dans le domaine de l'annotation comme le logiciel ELAN. Recenser ces outils et former les chercheurs à ceux-ci sont des missions du consortium CORLI. On peut citer également la plateforme Ortolang, réservoir de données et d'outils.
La journée d'études qui est proposée permettra de faire le point sur des annotations récentes en conviant linguistes et talistes à intervenir afin de croiser les approches. Il s'agira de comparer les types d'annotation selon l'objet linguistique et l'angle privilégié (syntaxique, sémantique, pragmatique (cf. par exemple Degand 2014, Abouda et Skrovec 2017, Lefeuvre 2021), de comparer les outils permettant d'annoter (TXM, le Trameur, ELAN, PRAAT, CLAN, …), de partager les pratiques pour procéder à l'enrichissement de ces corpus.

Bibliographie succincte

Abouda Lotfi et Skrovec Marie, 2018 : « Pour une micro-diachronie de l'oral : le corpus ESLO-MD », SHS Web of Conferences 46, 11004 (2018). https://doi.org/10.1051/shsconf/20184611004, CMLF 2018.

Abouda Lotfi et Skrovec Marie, 2017 : « Du rapport micro-diachronique futur simple / futur périphrastique en français moderne. Etude des variables temporelles et aspectuelles », *Corela* HS-21 (Eshkol-Taravella et Lefeuvre-Halftermeyer eds)

Badin Flora, Liégeois Loïc, Thiberge Gabriel, Parisse Christophe, 2021 : « Vers un outillage informatique optimisé pour corpus langagiers oraux en vue d'une exploitation textométrique : le cas des interrogatives partielles dans ESLO », *Corpus* [En ligne], 22 | 2021, mis en ligne le 28 janvier 2021, consulté le 13 mars 2021. URL : http://journals.openedition.org/corpus/5752.

Benzitoun Christophe, Jeanne-Marie Debaisieux et Henri-José Deulofeu, 2016 : « Le projet ORFÉO : un corpus d'étude pour le français contemporain », *Corpus* 15, Corpus de français parlé et français parlé des corpus.

Bergounioux Gabriel, Jacobson Michel, Pietrandrea Paola, 2017 : *L'annotation des corpus oraux*, halshs-03082419

Branca-Rosoff Sonia, Fleury Serge, Lefeuvre Florence et Pires Mat (2012) : « *Discours sur la ville. Présentation du Corpus de français parlé parisien des années 2000 (CFPP2000)* », cfpp2000.univ-paris3.fr/Corpus.html.

Christodoulides George et Barreca Giulia, 2017 : « Expériences sur l'analyse morphosyntaxique des corpus oraux avec l'annotateur multi-niveaux DisMo », *Corela* HS-21 (Eshkol-Taravella et Lefeuvre-Halftermeyer eds)

Degand, Liesbeth, 2014 : "'So Very Fast Very Fast Then' Discourse Markers at Left and Right Periphery in Spoken French." In *Discourse Functions at the Left and Right Periphery: Crosslinguistic Investigations of Language Use and Language Change*, Kate Beeching and Ulrich Detges (ed), *Studies in Pragmatics*, Volume 12. Leiden ; Boston : Brill, p. 151–78.

Degand, Liesbeth, Laurence Martin, et Anne-Catherine Simon, 2014 : « LOCAS-F : Un Corpus Oral Multigenres Annoté », CMLF 2014 - 4e Congrès Mondial de Linguistique Française, 2613–26. Berlin.

Eshkol-Taravella Iris, 2015 : *La définition des annotations linguistiques selon les corpus : de l'écrit journalistique à l'oral* (mémoire d'HDR), Linguistique, Université d'Orléans, tel-01250650

Fleury Serge, 2014-2015 : « Le Trameur. Base textométrique de textes alignés », (http://www.tal.univ-paris3.fr/sfleury/3.html)

Lacheret Anne, Kahane Sylvain, Pietrandrea Paola, 2014 (ed.): *Rhapsodie: a Prosodic and Syntactic Treebank for Spoken French*, col Studies in Corpus Linguistics, Amsterdam, Benjamins.

Lefeuvre Florence, 2021 : « Analyse outillée du marqueur discursif *bien sûr* », *L'Information grammaticale*, n°170, p. 32-42.

Rossi-Gensane Nathalie, Ursi Biagio, Eshkol-Taravella Iris, Skrovec Marie, 2019 : « La syntaxe en empirie et en théorie. La proposition de segmentation multiniveau du projet SegCor pour le français parlé », *Types d'unités et procédures de* segmentation, Béguelin, Corminboeuf, Lefeuvre (eds), Lambert Lucas, p. 203-220.

Programme provisoire

19 novembre 2021
Maison de la Recherche, 4 rue des Irlandais, 75 005 Paris.

9h-9h30 : accueil des participants

9h30-10h10 : Lotfi Abouda & Flora Badin (LLL-UMR7270 – CNRS / Université d'Orléans)
« Projet Ravioli : annoter les injonctives à l'oral en interaction »

10h10-10h50 : Florence Lefeuvre (CLESTHIA-EA 7345, Paris 3)
« Annotation topologique de quelques marqueurs discursifs »

Pause-café (10h50-11h05)

11h05-11h45 : Jean-Yves Antoine (LIFAT - EA 6300 – Université de Tours) :
« Prédire ou expliquer la langue ? Lorsque les spécificités de l'oral et les contraintes applicatives du TAL nous imposent la modestie en matière d'annotation »

11h45-12h25 : Christophe Parisse (coordinateur du consortium Corli) :
« L'Annotation de l'oral au sein du consortium CORLI »

Déjeuner : 12h30-14h

14h15-14h55 : Christophe Benzitoun (ATILF - UMR 7118 - CNRS / Université de Lorraine) :
« Annotation automatique en POS de productions de jeunes enfants »

14h55-15h35 : Liesbeth Degand (UC Louvain) & Anne Catherine Simon (UC Louvain) :
« Variation des structures syntaxiques dans un corpus annoté de français parlé »

Pause-café (15h35-15h50)

15h50-16h30 : Serge Heiden (ENS-Lyon) :
« Annotation linguistique automatique et assistée pour l'exploitation textométrique de transcriptions d'enregistrements multimédia avec TXM »