Proposition de workshop dans le cadre du GDR
Langues et Langages à la Croisée des Disciplines

Titre : **Cooccurrences et marquage discursif**

Mots-clés : linguistique de corpus, phraséologie, marqueurs de discours, mesures d'association, (non-)compositionnalité

Organisatrices :

Mathilde Dargnat, Université de Lorraine & ATILF-CNRS, Nancy
mathilde.dargnat@univ-lorraine.fr

Agnès Tutin, Université de Grenoble-Alpes & LIDILEM, Grenoble
agnes.tutin@univ-grenoble-alpes.fr

Descriptif

Il existe une littérature abondante dans le domaine des expressions complexes (pour un état des lieux récent, voir Bathia et al. 2023) et de la phraséologie (pour un état des lieux récent, voir Mel'čuk 2023). Ces travaux visent entre autres à repérer les combinaisons stables et à évaluer leur mode éventuel de figement sémantique. Des questions du même type se posent pour des cooccurrences d'expressions qui ont prioritairement une fonction dite discursive, c'est-à-dire qui concernent l'organisation du discours ou la manifestation du locuteur dans l'énonciation, par exemple *ah* + *bon*, *non* + *mais* + *alors*, *donc* + *du coup*, *ah* + *ben* + *tu parles*, *mais* + *enfin* (voir Waltereit 2007, Dostie 2013, Crible 2018, Crible & Degand 2019, Cuenca & Crible 2019, Haselow 2019, Dargnat 2022). L'étude de ces cooccurrences soulève un certain nombre de questions, qui , bien que nécessitant des perspectives différentes, interrogent toutes le statut d'expressions discursives complexes. Ce workshop insistera en particulier sur les aspects suivants :

## 1. Annotation des composants des cooccurrences

Parmi ces composants, certains appartiennent à plusieurs catégories, par exemple *bon* comme adjectif, y compris dans des figements comme *à bon escient/droit*, comme nom ou comme adverbe voire interjection. Les étiqueteurs probabilistes ou à base de grands modèles de langue donnent des résultats très moyens et instables quant à leur étiquetage catégoriel (par exemple *Perceo*, aussi bien dans sa version probabiliste que dans sa version apprentissage profond). Il est possible d'utiliser des automates finis pour reconnaître la catégorie, mais, là encore, il y a des difficultés notamment pour les composants dont la fonction discursive se reconnaît à partir de dépendances à distance. Exemples :

(a) $[[\text{après}]_{PREP}$ [le train qui est arrivé en retard]$_{GN}]_{GP}$ [il y en avait un autre]$_{P}$
(b) $[\text{après}]_{ADV}$ [le train qui est arrivé en retard]$_{GN}$ [c'est pas la faute du conducteur]$_{P}$

D'autre part, la position initiale n'est pas identifiable dans les corpus oraux quand il n'y a pas eu de segmentation en unités prédicatives. Il est donc probablement nécessaire de combiner les méthodes automatiques et l'annotation manuelle ainsi que les informations phonétiques et prosodiques quand elles sont pertinentes (réduction, pauses, contours, durée, etc.). L'identification de la catégorie est essentielle pour le point 2.

## 2. Évaluation de la « force » d'attraction entre les composants

Les mesures d'association constituent la technique de référence pour évaluer la tendance d'unités à se regrouper. Ces mesures sont sensibles à deux dimensions principales (Brezina 2018), auxquelles il faut rajouter la directionnalité, c'est-à-dire le fait qu'un composant

permette de prédire la présence d'un cooccurrent à sa droite ou à sa gauche. Par exemple, est-ce que *ah* est un meilleur « prédicteur » de *bon* sur sa droite que d'un autre marqueur, et inversement ? Il existe une vingtaine de mesures et il est nécessaire d'en comparer les résultats et d'en tester l'efficacité et la stabilité sur différentes données. Par ailleurs, on peut évaluer les performances des grands modèles de langue (type BERT et ses dérivés) en ce qui concerne la prédiction avec masquage, autrement dit le fait que le modèle complète plus ou moins bien une cooccurrence dont un des éléments est masqué. Par exemple, le fait que le modèle propose *bon* à la place de <masque> dans *ah* + <masque> + *je ne savais pas* et dans un ensemble de phrases de structures similaires : *ah* + <masque> + phrase.

## 3. Contribution sémantique des composants en fonction de leur intégration dans le contexte et en fonction des contours prosodiques

Il y a a priori deux questions fondamentales.
Premièrement, est-il possible de faire l'hypothèse que la contribution sémantique d'une cooccurrence est « additive », autrement dit que les différents composants contribuent séparément à l'interprétation de leur énoncé-hôte ? Cela paraît être le cas pour *mais enfin*. Faut-il, au moins dans certains cas, envisager que la cooccurrence a une contribution qui lui est propre, parce qu'elle ne retient que certains traits sémantiques de ses composants, ou qu'elle a un sens global indécomposable, comme en première analyse pour *ah bon* ?
Deuxièmement, quel est le rôle de la prosodie ? Dans le cas de marques isolées (par ex. *ah*, *tu sais, du coup, tu plaisantes*), la prosodie permet souvent de saisir les différentes valeurs (surprise, mécontentement, ironie, etc.). Dans le cas des cooccurrences (par ex. *allez* + *bon*, *non* + *mais* + *oh*, *tiens* + *donc*), les contours prosodiques observés sont-ils une simple juxtaposition de contours associés à chacun des composants ? Un des contours appropriés pour un des composants est-il privilégié et étendu à la cooccurrence ? Dans ce dernier cas, quelle est l'interaction entre la prosodie et la sémantique, un des composants a-t-il plus de poids que l'autre ?

## 4. Prise en compte de la variation

On peut distinguer plusieurs variables susceptibles d'affecter la production de ces cooccurrences, dont le genre discursif (par ex. conversation spontanée, conférence, débat, entretien dirigé, exposé scolaire, textes de fiction, etc.), la situation d'énonciation (notamment les positions plus ou moins hiérarchiques des interlocuteurs), les paramètres individuels (âge, situation sociale, genre, etc.), l'époque de constitution des corpus. La prise en compte de la dimension diachronique, courte ou longue, est également importante pour la mise en évidence de l'émergence des cooccurrences et de leur figement éventuel. Par exemple, l'apparition de *du coup* comme connecteur de conséquence et sa combinaison avec *donc* et *alors*, la constitution des séries verbe + *donc* (*tiens donc*, *coudonc* en français québécois, *(non mais) dis donc*, *va donc*), qui peuvent pour certains prétendre à la pragmaticalisation (voir entre autres Dostie 2004, Waltereit 2007, Heine et al. 2021).

## Références

Bhatia A., Evang K., Garcia M., Giouli V., Han L., Taslimipoor S. 2023. *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*. Association for Computational Linguistics, Dubrovnik, Croatia.

Brezina V. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge : Cambridge UP.

Crible L. 2018. *Discourse Markers and (Dis)fluency Forms and functions across languages and registers*. Amsterdam : John Benjamins.

Crible L. et Degand L. 2019. « Domains and Functions: A two-dimensional account of discourse markers ». *Discours* 24, 35 p. (en ligne)

Cuenca M.-J. et Crible L. 2019. « Co-occurrence of discourse markers in English: From juxtaposition to composition ». *Journal of Pragmatics* 140, 171-184.
Dargnat M. 2022. « *Mais enfin*: construction et association ». *Langages* 225, 49-63.
Dostie G. 2004. *Pragmaticalisation et marqueurs discursifs. Analyse sémantique et traitement lexicographique*. Liège : De Boeck/Duculot.
Dostie G. 2013. « Les associations de marqueurs discursifs. De la cooccurrence libre à la collocation ». *Linguistik* 62(5), 15-45. (en ligne)
Haselow A. 2019. « Discourse Marker Sequences: Insights into the Serial Order of Communicative Tasks in Real-Time Turn Production ». *Journal of Pragmatics* 146, 1-18.
Heine B., Kaltenböck G., Kuteva T. & Long H. 2021. *The Rise of Discourse Markers*. Oxford : Oxford UP.
Mel'čuk I. 2023. *General Phraseology, Theory and Practice*. Amsterdam : John Benjamins.
*Perceo.* Version probabiliste https://www.ortolang.fr/market/corpora/perceo ; version apprentissage profond : https://huggingface.co/waboucay/french-camembert-postag-model-finetuned-perceo
Waltereit R. 2007. « A propos de la genèse diachronique des combinaisons de marqueurs. L'exemple de *bon ben* et *enfin bref* ». *Langue française* 154, 94-109.

Ce workshop sera soutenu par deux projets ANR :

CODIM
Compositionality and Discourse Markers

ANR-22-CE38-0002

Coord. Mathilde Dargnat, ATILF
Autres laboratoires partenaires : LLF, LORIA

PRÉFAB
Construction des phrases PRÉFABriquées dans les interactions langagières
ANR-22-CE54-0013-02

Coord. Agnès Tutin, LIDILEM
Autres laboratoires partenaires : ATILF, BCL, ICAR,