Mais et associés

Mathilde Dargnat, Jacques Jayez

Received: March 2021 / Accepted: July 2021
Published online: December 2021

Mais et associés

Mathilde Dargnat
Université de Lorraine et ATILF (Université de Lorraine-CNRS, UMR 7118), Nancy
mathilde.dargnat@univ-lorraine.fr

Jacques Jayez
École Normale Supérieure de Lyon et LORIA (Université de Lorraine-CNRS-INRIA, UMR 7503), Nancy
jacques.jayez@ens-lyon.fr

Abstract

This paper focuses on the distribution of discourse marker sequences immediately on the left or right of *mais* (≈ *but*) in French, such as *oui non mais* (≈ *yes no but*), *mais quand même* (≈ *but still*) or *mais bon* (≈ *but well*). The goal is to determine what sequences are the most frequent and how they fit (or not) with the meaning of *mais*. Exploiting five spoken French corpora, we use two association measures to extract the most plausible candidate patterns. Following the literature on association measures, we explore the two complementary dimensions of frequency (MI^3 measure) and predictability (DeltaP measure). This procedure reveals that i) there are indeed discourse marker patterns around *bon,* ii) most clusters or associates smoothly combine with the basic 'argumentative' value of *mais* and iii) within this semantically coherent set, comparing and crossing the results of the two measures on the left and the right of *mais* helps to identify subsets of discourse markers with specific discourse functions.

Keywords: association measures, semantic integration, discourse markers, *mais*

Résumé

Dans cet article, nous étudions la distribution des séquences de marqueurs de discours dans l'environnement immédiat de *mais*, à droite ou à gauche, par exemple *oui non mais*, *mais quand même* ou *mais bon*. Notre objectif est d'identifier les séquences les plus fréquentes et de déterminer si elles sont compatibles avec le sens de *mais*. Nous utilisons cinq corpus de français parlé et deux mesures d'association pour extraire les motifs récurrents les plus plausibles. Nous appliquons une distinction, familière dans la littérature sur ces mesures, en explorant deux dimensions complémentaires : celle de la fréquence (mesure MI^3) et celle de la prédictibilité (mesure DeltaP). Cette méthode montre que : (i) il y a effectivement des motifs récurrents de marqueurs de discours autour de *mais*, (ii) la plupart de ces agrégats ou associés se combinent naturellement avec la valeur « argumentative » fondamentale de *mais* et (iii) à l'intérieur de cet ensemble sémantique homogène, la comparaison et le croisement des deux mesures sur la droite et sur la gauche de *mais* contribuent à l'identification de sous-ensembles exerçant des fonctions de discours spécifiques.

Mots-clefs : mesures d'association, intégration sémantique, marqueurs de discours, *mais*

1. Introduction au problème

Le but de cet article est d'étudier la combinaison de marqueurs de discours (MD). Le texte soulève la question de l'articulation entre la (non-)compositionnalité des MD que l'on peut dire « complexes » (Waltereit, 2007) et la fréquence d'appariement de leurs composants, en utilisant des mesures dites d'*association*. Cette recherche s'inscrit dans un ensemble de travaux autour des séquences de lexèmes ayant des propriétés statistiques particulières dans les corpus écrits ou oraux. Elle rejoint, dans le domaine des MD, les approches des *constructions* (Goldberg, 2006, 2019 ; Herbst *et al.*, 2014 ; Hilpert, 2014), des *collostructions* (Desagulier, 2015 ; Gries, 2019) et des *multi-mots* (voir Constant *et al.*, 2017 pour un état de l'art). Dans ces approches, l'accent est mis sur la régularité de séquences qui apparaissent intuitivement comme plus ou moins figées, voire lexicalisées et/ou non-compositionnelles. La fréquence d'association de lexèmes dans les données peut être un indice statistique de ce figement, mais n'en détermine pas la nature, notamment le caractère plus ou moins compositionnel (voir l'articulation des différents types de phrasèmes chez Mel'čuk, 2013). L'analyse quantitative de telles cooccurrences peut aussi donner des informations sur la force d'attraction de certaines valeurs sémantico-pragmatiques des cooccurrents, et de leur représentation dans des données et des contextes énonciatifs particuliers. Ainsi, certains des résultats d'association pourraient être interrogés sous l'angle des phraséologismes pragmatiques, relevant de routines ou stéréotypes discursifs (voir Métrich *et al.*, 2002 ; Burger *et al.*, 2007 ; Dziadkiewicz, 2007). Pour désigner les séquences de MD, nous utilisons le terme de *cooccurrence* de manière neutre, et le terme d'*associé* pour désigner les cooccurrences retenues par les mesures d'association.

Dans cinq corpus oraux (voir section 3.2), nous nous intéressons spécifiquement aux combinaisons ayant *mais* comme pivot et d'autres MD comme cooccurrents. L'expression *marqueurs de discours* désigne ici à la fois les connecteurs (*donc*, *pourtant*, *parce que,* etc.) et les particules énonciatives (*bon*, *ah*, *tu parles,* etc.) (voir section 2). Les cooccurrents pertinents sont définis à partir de deux aspects. Le premier aspect est intuitif : certains MD se combinent avec *mais* sans que leur valeur sémantique fasse écho à celle(s) de *mais*, c'est-à-dire utilise celle-ci pour réaliser un mouvement discursif intégré. Par exemple, un MD comme *après* peut avoir une valeur sémantique temporelle. En (1), le locuteur L1 explique les horaires suivis lors de son apprentissage du français. Le *après* de son deuxième tour de parole fait référence à la période postérieure à la limite des seize ans pour les apprenants.

(1) L1 – oui cent soixante minutes de onze ans à seize ans
L2 – cent soixante minutes
L1 – mais <u>après</u> nous avions euh six heures six (ESLO1)

En revanche, en (2), *après* renforce l'interprétation contrastive de *mais* (dépenses d'alimentation *vs* autres dépenses).

(2) L2 – ah oui je fais qu'une enveloppe pour euh mettons l'alimentation pour euh mais après tout ce qui est acheté en dehors
L1 – oui
L2 – euh de la nourriture c'est fait par un chèque comme ça je sais tout de suite où l'argent est passé (ESLO1)

Nous parlerons d'un emploi *indépendant* dans les cas comme (1), et d'un emploi *intégré* dans des cas comme (2).

Le deuxième aspect qui intervient dans la notion d'associé est plus « objectif » et repose sur des *mesures d'association*, c'est-à-dire des fonctions statistiques qui sont censées capturer la force et la direction de cooccurrence entre deux lexèmes (Brezina, 2018). De telles méthodes n'impliquent nullement l'existence d'hypothèses sémantiques préétablies, bien qu'elles puissent servir à les évaluer si elles existent. En général, il est impossible de prédire avec une certitude suffisante les environnements d'un lexème à partir de sa description sémantique (souvent pas assez précise ou complète, d'ailleurs). Cette difficulté à dériver les combinaisons est justement un des constats partagés par les diverses approches qui se fondent sur une intuition de *figement* (voir Hanks, 2013 pour une discussion de la relation entre sens et données d'observation).

Le calcul de ces mesures est utile mais génère du bruit par rapport à la distinction sémantique introduite ci-dessus. Les emplois de MD statistiquement associés peuvent être de « faux positifs sémantiques », c'est-à-dire relever d'emplois indépendants plutôt qu'intégrés. D'autre part, il peut y avoir des erreurs catégorielles pour les mots polyfonctionnels, par exemple si *bon* adjectif est inclus dans l'inventaire des cooccurrents. On peut alors parler de « faux positif catégoriels ». L'existence du bruit lié aux faux positifs s'explique par (i) le caractère local des mesures d'association (on exploite les environnements immédiats) ; (ii) la non prise en compte de la prosodie[1], et (iii), pour les résultats chiffrés, par un effet de cumul entre emplois indépendants et emplois intégrés, qui aboutit à gonfler artificiellement le nombre des cooccurrences.

L'article est organisé comme suit. La section 2 caractérise les unités étudiées. La section 3 présente les corpus retenus, le cadre général des mesures d'association et les choix qui ont été faits ainsi que les limites qu'ils imposent. La section 4 expose et discute les résultats obtenus. La section 5 conclut brièvement sur le problème de la (non-)compositionnalité.

[1] Sans données prosodiques, il est par exemple très difficile de distinguer *si* comme particule de réponse et *si* comme MD d'hypothèse ou de concession (*si habile qu'elle soit* …). Sur le rôle de la prosodie dans la discrimination des fonctions des marqueurs, voir par exemple Dargnat *et al.,* 2015 et Lee, 2021.

2. Les marqueurs de discours

La littérature sur les MD est vaste et difficile à embrasser et à unifier en raison notamment de la diversité des perspectives (études sémantiques, syntaxiques, interactionnelles, etc.) et des étiquettes (voir Dostie & Pusch, 2007 ; Dargnat, à paraître). Les MD constituent une catégorie fonctionnelle plutôt qu'une catégorie grammaticale (voir Hansen (1998) et Paillard (2011) pour des perspectives différentes). Ils désignent des lexèmes ou expressions simples ou complexes utilisé(e)s pour donner des instructions d'interprétation dans un contexte donné. C'est en cela qu'on leur attribue une signification *procédurale*. La portée des MD et le type d'instruction qu'ils véhiculent sont variables, et ce sont ces variations qui permettent de distinguer différents types de fonctionnement[2]. En suivant la réflexion de Dostie (2004), mais en modifiant sa terminologie, nous distinguons des MD connecteurs (MDC) et des MD particules énonciatives (MDP). Les MDC incluent des formes comme *alors, après (que), bien que, c'est-à-dire, donc, mais, pourtant*, etc. Ils permettent de tisser un réseau de relations de discours entre différents objets interprétatifs, en exprimant des relations de causalité, de justification, d'opposition, etc., telles qu'on les trouve formellement décrites par exemple dans la RST (Taboada & Mann, 2006) ou la SDRT (Asher & Lascarides, 2003). Les MDP comportent la classe traditionnelle des interjections, comme *ah, hein, zut*, etc., des expressions ou des constructions qui impliquent l'interlocuteur, comme *tu vois, n'est-ce pas, d'accord*, et d'autres expressions ou constructions plus diverses comme *alors, bon, disons, tiens, voilà*, etc. Ils inscrivent le locuteur dans le discours en train de se faire et fournissent une trace en temps réel de sa propre évolution attentionnelle, émotionnelle et intellectuelle. Les MDC contribuent à la cohérence du discours, les MDP contribuent à la manifestation du locuteur et à la gestion de l'interaction. Les MD n'apparaissent pas nécessairement seuls ; au contraire, ils sont souvent combinés en séquences (*ah bon, non mais quand même*, *mais enfin*, etc.) (voir Auchlin, 1981). Les contraintes de combinaison et les figements sont plus rarement étudiés que les propriétés des MD simples et il n'est pas toujours facile d'expliquer le processus de pragmaticalisation sous-jacent (voir par exemple Razgoulieva, 2002 ; Waltereit, 2007 ; Dargnat, 2021b).

3. Problèmes et techniques de l'association

Toute étude d'association présuppose une réponse à trois questions, non nécessairement distinctes : (i) Quel(s) type(s) d'associé veut-on étudier ?, (ii) Quel(s) type(s) de corpus veut-on utiliser et comment les explore-t-on ?, (iii) Quel(s) type(s) de mesure d'association doit-on choisir ?. Les trois sous-sections suivantes abordent ces trois questions dans le même ordre.

[2] Ces fonctionnements ne sont pas exclusifs, et il est courant de parler de *polyfonctionnalité* des MD.

3.1. Les associés de *mais*

Toutes les associations supposent qu'on se limite à des *fenêtres*. Les fenêtres sont définies en fonction de critères comme le type des unités de découpage du corpus (des mots, des syntagmes, des paragraphes, des tours de parole, etc.), leur longueur (nombres d'unités de découpage), leur direction et leur distance par rapport au *pivot*, ici *mais*. La caractérisation des fenêtres peut être complexe lorsqu'elle combine tous ces critères.

Un travail préliminaire sur *mais enfin* (Dargnat, 2021b) a confirmé l'impression que les lexèmes immédiatement à droite ou à gauche de *mais* sont indéfiniment variables (adjectif, adverbe, pronom, groupe nominal, verbe, etc.). Il n'y aurait donc pas grand intérêt à étudier ce type de cooccurrence en général. En revanche, *mais* se trouve souvent au contact de MD et c'est cette catégorie que le présent article cible. Des tests complémentaires indiquent que les MD qui cooccurrent avec *mais* sont généralement au nombre de un ou deux à droite et/ou à gauche, plus rarement trois. Il a donc été décidé de se limiter à un maximum de trois MD à gauche et/ou à droite du pivot *mais*. Cependant, ce choix ne détermine pas la distance de *mais* à ses cooccurrents. Par exemple, il est possible qu'un ou plusieurs lexèmes qui ne sont pas des MD séparent *mais* d'un MD à droite ou à gauche. En (3), la MDP d'hésitation *heu* et la MDP *hein* apparaissent autour du pivot, mais pas à son contact. Le problème est encore plus flagrant en (4) où l'on trouve des MDC (*donc, parce que*, *et*, *aussi*) et des MDP (*ben*, *euh*) dans un environnement large du pivot. Le pivot est en capitales et les cooccurrents sont soulignés.

(3) l'enseignement du français plus spé- spécialement + peut-être pour faire des des lettres pour euh écrire euh à des amis MAIS pas forcément pour euh faire des C.V. des lettres de motivation tout ça hein (CORPAIX)

(4) ben moi en cours je me plie à aux règles donc euh j'utilise ça parce que je tiens à avoir des notes raisonnables MAIS je trouve ça aussi très bien de pouvoir parler la langue et pouvoir la parler correctement (CORPAIX)

Les exemples (3) et (4) illustrent un phénomène général : lorsqu'un MD est relativement éloigné du pivot *mais*, les contributions de ce MD et du pivot sont le plus souvent indépendantes. Par exemple, dans (3), les hésitations marquées par *euh* sont indépendantes de la valeur de « correction » de *mais*, laquelle indique que le locuteur exclut certaines possibilités (les lettres professionnelles). En (4), les MDC *donc*, *parce que* et *et*, ainsi que les MDP *ben* et *euh* ont également une contribution spécifique. En revanche, *aussi* coopère avec *mais* dans un motif sémantico-rhétorique fréquent du type *non seulement ... mais aussi*. Dans une première phase, nous avons testé une recherche « tolérante » qui identifiait toutes les occurrences isolées, les couples d'occurrences ou les triplets d'occurrences de MD à gauche ou à droite de *mais*, chaque MD pouvant être séparé de *mais* ou du MD suivant par au maximum deux non-MD. Le résultat a été très décevant, car les MD non contigus à *mais* étaient en général sans rapport avec lui. Pour éviter ce bruit substantiel, nous avons supprimé les intervalles de non-MD sauf pour *aussi*. En résumé, nous avons recherché toutes les séquences d'au plus trois MD à gauche et à droite de *mais*.

3.2. Corpus et exploration

Les corpus utilisés sont des transcriptions d'interactions en face à face entre deux ou plusieurs locuteurs : CORPAIX (953188 mots), CRFP (380435 mots), DECLICS (160685 mots), une partie de ESLO1 (613965 mots) et FRA80 (185613 mots). Il s'agit donc de corpus oraux pour lesquels des notions traditionnellement utilisées en TAL (ponctuation, découpage syntagmatique ; voir Bender, 2013) ne sont pas nécessairement possibles à appliquer ni même pertinentes. Ces données ont été retraitées pour éliminer les indications diacritiques et les ramener à un format unique pour les tours de parole. Les analyseurs syntaxiques ne reconnaissent pas bien les catégories grammaticales sur ce type de corpus. Les étiqueteurs morphosyntaxiques entraînés sur l'écrit sont également insuffisants. La version des paramètres du français oral fournie par le projet PERCEO (Benzitoun *et al.,* 2012) pour l'étiqueteur multilingue *TreeTagger*[3] a de meilleurs résultats parce qu'elle intègre un certain nombre de lexèmes catégorisés comme interjections. Ainsi, elle reconnaît les emplois comme MDP de *bon* et *ben* et est capable de discriminer les emplois de *bon* comme adjectif et comme MDP. En revanche, elle ne discrimine pas les différents emplois de *bien* dans une phrase test comme « bien, j'ai commencé à faire du bien. C'est bien », et les étiquette tous comme adverbes.

Nous avons donc conservé l'approche initiée dans Dargnat (2021b) et fondée sur l'utilisation de l'étiqueteur lexical Unitex-GramLab[4]. Ce dernier ne cherche pas à « deviner » une catégorie à partir de probabilités, mais assigne à tout lexème toutes les catégories dont celui-ci peut relever d'après les dictionnaires utilisés. Il est donc en théorie très inférieur à un étiqueteur standard ou à un analyseur syntaxique, mais il est beaucoup plus efficace si l'on cherche des motifs à l'aide d'automates finis (des « graphes » dans sa terminologie). De plus, il comporte déjà une liste de MD complexes comme *quand même* ou *à la fois,* que ni PERCEO ni d'autres étiqueteurs ne reconnaissent correctement. Enfin et surtout, il permet de créer des dictionnaires spécialisés, et de compléter ou de simplifier ceux qui existent pour affiner les graphes de recherche de motifs. Dans un premier temps, nous avons modifié et étendu le dictionnaire de Roze *et al.* (2012), notamment en ajoutant les MDP. Nous avons conservé une idée importante du dictionnaire initial en considérant des prépositions causales, finales ou temporelles comme des MDC (*en raison de*, *pour*, *au moment de*, etc.). Nous avons ensuite intégré ce dictionnaire dans Unitex-GramLab et construit un dictionnaire hybride entre le dictionnaire du français fourni avec le logiciel et le dictionnaire des MD. Concrètement, ce dictionnaire identifie chaque MD comme MDC ou MDP et tous les lexèmes différents ou identiques mais relevant aussi d'une autre catégorie comme OTHER. Par exemple *bon* va figurer comme MDP et comme OTHER, car il est aussi adjectif.

La recherche des motifs qui nous intéressent est effectuée à l'aide d'un automate. Dans un premier temps, on récupère des triplets formés d'un contexte gauche, d'une combinaison de MD correspondant

[3] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[4] https://unitexgramlab.org/fr

à un motif et d'un contexte droit. À l'intérieur du motif nous distinguons le pivot (*mais*) et la séquence de MD à gauche / droite de *mais*, que nous appelons les *cooccurrents gauches* / *droits*. Ces séquences peuvent être nulles lorsque *mais* n'est précédé ou suivi d'aucun MD.

Les résultats sont retraités[5] à l'aide d'un script spécifique, pour éviter, entre autres, (i) de prendre en compte des emplois de *bon* et *bien* comme adjectif ou adverbe, des emplois de *attention* ou *remarque* comme nom (par exemple dans *porter une grande attention à* ou *bonne remarque*) et des emplois de *dis* ou *dites* comme verbe (après un pronom), (ii) de conserver *aussi* quand il est utilisé dans une expression comparative (on élimine *aussi* + adjectif / adverbe), (iii) d'avoir des doublons lorsqu'un même lexème est étiqueté comme MDC et MDP, (iv) d'avoir des combinaisons de MD au lieu d'un MD complexe (par exemple *quand même* pourrait être incorrectement découpé en *quand* + *même*). Par ailleurs, sept MD fréquents, mais dont il est difficile de déterminer s'ils sont plutôt MDP ou plutôt MDC[6], ont été réétiquetés par le script comme simplement MD, ce qui signifie qu'on conserve l'indétermination. Cela n'a pas d'influence sur les résultats quantitatifs.

3.3. Cooccurrences et mesures d'association

3.3.1. Le choix des cooccurrences à étudier

La définition des cooccurrences conduit à plusieurs interrogations. Lorsque *mais* est environné de MD à gauche et à droite, faut-il considérer que les cooccurrents sont toute la séquence (MD à gauche + MD à droite), ou faut-il séparer les cooccurrents gauches et les cooccurrents droits ? Bien que la première approche corresponde plus strictement au terme de *cooccurrent*, elle méconnaît le fait que les cooccurrents gauches n'influencent pas les cooccurrents droits, ni bien entendu le contraire. Un test préalable de prédictibilité[7] des séquences de cooccurrents droits à partir des séquences de cooccurrents gauches n'a donné que des valeurs négligeables. Nous avons donc analysé les cooccurrents gauches et droits séparément.

Une deuxième question concerne la distance. Lorsqu'un cooccurrent comporte plusieurs mots, (simples ou complexes), on peut analyser le rapport de *mais* au cooccurrent dans son ensemble ou à chacun de ses composants. Par exemple, dans une expression comme *mais maintenant non*, on peut analyser le rapport entre *mais* et *maintenant non* à droite et/ou le rapport entre *mais* et *maintenant* et/ou le rapport entre *mais* et *non*. Nous avons retenu le rapport entre *mais* et toute la séquence droite (*maintenant non*) ainsi que le rapport entre *mais* et le cooccurrent contigu *maintenant*. Pour des raisons de temps, nous avons pour le moment ignoré le rapport entre *mais* et les cooccurrents non contigus (*non* dans l'exemple), qui fera l'objet d'une étude complémentaire. Ces choix s'appliquent aussi sur la

[5] L'écriture d'un automate intégré dans Unitex et évitant ce double traitement est en cours de test.

[6] Il s'agit de *alors*, *effectivement*, *enfin*, *mais, par exemple, remarque* et *soit*.

[7] Nous avons utilisé la mesure DPF décrite dans la section suivante.

gauche. Pour résumer, dans une structure [*mais* A B C], nous étudions les rapports *mais*-ABC, *mais*-AB et *mais*-A ; dans une structure [A B C *mais*], nous étudions les rapports ABC-*mais,* BC-*mais* et C-*mais*.

Lorsqu'on étudie les rapports entre *mais* et les cooccurrents contigus, faut-il séparer les cas où le cooccurrent est tout seul (structure [*mais* A] ou [A *mais*]) et les cas où il figure dans une séquence (structure [*mais* A B (C)] ou [(A) B C *mais*]) ? Nous avons considéré qu'il valait mieux réunir tous les cas dans la mesure où nous nous intéressons au premier chef aux relations avec *mais*. Dans la structure [*mais* A B (C)], il est peu probable que A soit influencé par B (C). Dans la structure [(A) B C *mais*], il est peu probable que le rapport entre C et *mais* soit influencé par (A) B. Cependant, il est possible que la présence de C soit influencée par (A) B et le choix que nous avons fait demanderait à être réévalué en détail, ce que nous n'avons pas fait dans cet article.

Nous avons appliqué la même analyse aux cooccurrents de longueur 2 en prenant en compte, outre ces cooccurrents, les fins (à droite) ou les débuts (à gauche) de séquences de longueur 3.

3.3.2. Les mesures d'association

Brezina (2018) décrit les mesures d'association en utilisant deux dimensions : la fréquence et l'exclusivité. Les mesures sensibles à la fréquence reposent sur la fréquence *observée*, correspondant au nombre effectif de pivots (*mais* dans notre cas), de cooccurrents ou de combinaisons impliquant un cooccurrent et le pivot et sur la fréquence *attendue*, qui se détermine à partir de la présence ou de l'absence du pivot et/ou du cooccurrent. L'exclusivité estime la tendance pour le pivot et le cooccurrent à figurer davantage ensemble que séparément. Seize mesures d'association ont été testées dans Dargnat (2021b). N'ont été retenues ici que les deux plus efficaces : la MI^3 (MI pour *Mutual Information*, 3 pour le degré du numérateur dans la fonction) et la *Delta-P-forward/backward* (DPF ou DPB). En simplifiant, la MI^3 est surtout sensible à la *fréquence* d'association alors que la DPF mesure la *prédictibilité* d'un lexème (ou groupe de lexèmes) à partir d'un autre lexème (ou groupe de lexèmes), par exemple *mais* à partir de *ah* ou *quand même* à partir de *mais*. Nous ne développons pas les caractéristiques techniques de ces mesures (voir Jenkins & Ward, 1965 ; Allan, 1980 ; Church & Hanks, 1990 ; Evert & Krenn, 2001 ; Ellis, 2006, et Schneider, 2020 pour des discussions précises).

4. Résultats

Par commodité, nous diviserons la présentation et la discussion des résultats en quatre rubriques : les chiffres globaux pour les différentes tailles de fenêtre à gauche et à droite, les chiffres pour MI^3 et ceux pour DPF et DPB.

4.1. Répartition des fenêtres

Les cinq corpus retenus permettent de produire 19 297 occurrences de *mais* tout seul ou avec un environnement de MD sur un total de 2747166 mots (une proportion de 7‰). La proportion de *mais* isolés est proche de 50 % dans tous les corpus. Il est difficile d'interpréter cette stabilité. Elle pourrait par exemple être liée à une tendance générale dans la répartition de *mais* à l'oral et/ou relative au type de corpus oral considéré (entretiens dirigés, échanges libres). Pour chaque corpus, il existe seize distributions possibles de cooccurrents à gauche et à droite : 0 à gauche et à droite (*mais* isolés), 0 à gauche et 1 à droite, etc. Lorsqu'on ordonne les résultats, on constate que les résultats 3-3 (3 cooccurrents à gauche et 3 à droite) sont rares (nombre de cas pour chaque corpus des tableaux = 5,1,0,3,0). Cette possibilité n'est donc pratiquement pas représentée dans les corpus. En dehors des 0-0 (*mais* isolés) qui, sans surprise, arrivent toujours en tête, ce sont les 0-1 et les 1-0 qui sont les mieux représentés avec une nette domination des premiers. On trouve plus d'occurrences d'associés à droite qu'à gauche de *mais*. Sur l'ensemble des corpus, les associations de type (0-1) sont 2.18 fois plus nombreuses que les associations de type (1-0), le rapport étant de 2.4 pour le type 0-2 par rapport à 2-0. Les comparaisons sont plus équilibrées pour les configurations mixtes, avec un associé ou deux à gauche ou à droite, par exemple on note, toujours pour l'ensemble des corpus, un rapport de 1.16 pour 1-2 comparé à 2-1.

4.2. La mesure MI^3

Pour MI^3, nous avons fixé des seuils en dessous desquels les résultats sont peu pertinents. Pour une taille de fenêtre de 1 ou 2, le résultat de la mesure doit être au moins de 7, avec, dans le corpus, au moins 10 occurrences des associés et 10 cooccurrences de type [associé(s) + pivot *mais*] ou [pivot *mais* + associé(s)]. Pour une taille de fenêtre 3, il faut au moins 8 occurrences des associés et 8 cooccurrences de type [associés + pivot *mais*] ou [pivot *mais* + associés].

4.2.1. MD à gauche

Comme c'est souvent le cas quand on veut comparer les résultats entre corpus, il est plus parlant de donner la préférence au rang sur le score brut. La répartition des valeurs de MI^3 pour les différents corpus est hétérogène tant pour les médianes que pour la distribution des valeurs. Voici les résultats par corpus. L'ordre des lexèmes correspond à l'ordre décroissant de la MI^3.

	1 associé	2 associés	3 associés
CORPAIX	*non, oui, ouais, hein, euh, quoi, ah, oh, voilà, là, bon, je veux dire, eh, hum, voyez, et*	*oui non, ah non, non non, mais non, ah oui, ouais non, ben oui, mais oui, oui oui, ah ouais, euh oui*	
CRFP	*non, oui, quoi, hein, euh, ouais, voilà, là, ah, bon, hum*	*non non, oui oui*	*non non non*
DECLICS	*non, oui, ouais, hein, euh, voilà, ah, si vous voulez, là*	*oui non, non non*	
ESLO1	*oui, non, hein, euh, ah, oh, si vous voulez, là, bon, quoi*	*non non, mais oui, ben oui, mais non, oui non, oui oui, ben oui, ah oui, ah non, ben non, euh oui, oh oui*	
FRA80	*non, oui, hein, ah, bon*	*ben oui*	

Tableau 1. Résultats de la MI^3 à gauche du pivot *mais*.

On peut dégager plusieurs traits de ces résultats. (i) Les MDC sont quasiment absents ; *mais* se combine donc sur sa gauche essentiellement avec des MDP. (ii) Il n'y a pratiquement pas de séquences de trois cooccurrents suffisamment stables avant *mais* pour être repérées par la MI^3. (iii) Les *oui* et les *non* isolés, répétés ou combinés dominent largement. Par rapport au nombre important de MDP existant en français, la variété récupérée par le filtre de la MI^3 est assez faible.

À ce stade, on peut déjà, avec prudence, envisager quelques pistes interprétatives. La présence massive de *oui* ou *non* et leurs dérivés est compatible avec le statut argumentatif de *mais*, bien dégagé par Bruxelles *et al.* (1976) et Anscombre et Ducrot (1983). L'idée fondamentale est que *mais* sert à introduire un constituant de discours qui s'oppose aux conclusions potentielles qu'on pourrait tirer d'un autre constituant. Un couple *oui mais* peut ainsi servir à accepter une proposition dont on va affaiblir la portée argumentative implicite en introduisant une autre proposition que *mais* marque comme opposée à la première ou, tout au moins, comme favorisant d'autres conclusions que la première (mouvement concessif). Cette conception de l'argumentation est élargie dans Dargnat (2021c), notamment à propos des séquences *non mais*, pour tenir compte de la grande flexibilité des enchaînements conversationnels. *Oui mais* peut également servir à prendre en compte une intervention de l'interlocuteur (*feedback*) et à enchaîner sur une argumentation. Les séquences avec *ah* et *oh* associent des MDP qui marquent une modification de l'état émotionnel ou attentionnel du locuteur et un mouvement argumentatif. C'est une séquence fréquente dans les corpus oraux en général, qui signale par exemple qu'un locuteur ne s'attendait pas à une question ou à un thème de discours ou qu'il y accorde de l'importance. Le rôle de *voilà* et de *bon* est plus complexe. Ils ont en commun de pouvoir marquer le terme d'un processus, soit directement (*bon*), soit par le biais d'un regroupement d'entités et/ou de procès (Col *et al.*, 2015). Avant *mais*, ces deux marqueurs indiquent souvent une

étape, soit dans la gestion d'un échange, soit dans le propre discours du locuteur. La présence de *mais* indique la reprise ou le début d'un mouvement argumentatif. *Quoi* indique qu'une assertion est « optimale » pour le locuteur dans un contexte donné. Cela peut signifier que le locuteur estime cette assertion comme appropriée pour résumer sa pensée ou que, du moins, il n'a pas la volonté ou la possibilité de formuler ce qu'il a à dire autrement (Dargnat & Jayez, 2020). Dans cette perspective, *quoi* marque également une étape dans la trajectoire discursive/mentale du locuteur et *mais* est susceptible, entre autres, de revenir argumentativement sur le(s) constituant(s) de cette étape (mouvement concessif, surtout avec *bon*) ou sur une phase antérieure de l'échange. Nous ne commenterons pas les marques d'hésitation ou de temporisation (*heu*, *hum*), qui sont omniprésentes dans les corpus d'oral spontané.

4.2.2. MD à droite

Comme pour les MD à gauche, la répartition entre les différents corpus est hétérogène.

	1 associé	2 associés
CORPAIX	*euh, enfin, bon, sinon, quand, je veux dire, disons, alors, mais, autrement, là, par contre, si, pour, en même temps, non, à, en tout cas, en fait, maintenant, par exemple, après, sans, tu sais, oui, en, quand même, aussi, avant*	*enfin euh, bon euh, euh euh*
CRFP	*bon, euh, enfin, en même temps, là, après, mais, quand, si, non, alors, aussi, en fait, pour, à*	*enfin bon, bon euh, euh bon*
DECLICS	*bon, euh, après, quand, pour, là, si*	
ESLO1	*enfin, alors, euh, si, non, pour, par contre, autrement, quand, maintenant, disons, là, à, vous savez, sans, voyez, oui, en, quand même, mais, avant, bon, même*	*enfin euh*, *enfin*, *pour*, *alors là*, *alors euh*
FRA80	*enfin*, *euh*, *alors*, *autrement*, *quand*, *pour*, *bon*, *là*, *si*	

Tableau 2. Résultats de la MI^3 à droite du pivot *mais*.

Comme précédemment, les combinaisons ternaires sont nulles ou négligeables avec le filtre de la MI^3 et des seuils appliqués. On ne trouve pratiquement pas de séquence de trois cooccurrents. Certains associés, bien que relevés ci-dessus sont de « faux positifs sémantiques », en ce sens qu'ils n'ont pas un lien étroit avec la signification de *mais* (voir section 1). Il s'agit de *à, avant, en, là, maintenant, pour, quand, sans* et *si*. *Après* et *en même temps* sont plus ambivalents entre fonction temporelle et fonction concessive, comme dans (6) :

(5) c'est important l'orthographe important mais en même temps c'est pas c'est insignifiant quoi (CORPAIX)

Parmi les autres MD, on trouve des approximateurs (*disons, je veux dire*), le reformulatif *en fait* (Rossari *et al.,* 2018), le contrastif *par contre*, le concessif *quand même*, des appels à l'interlocuteur[8] (*vous savez*, *voyez*), des MD qui suggèrent un changement de thème de discours ou un contraste (*autrement*, *sinon*), le MD de repérage énonciatif *alors* (voir Franckel, 1987 ; Jayez, 1988 et Hansen, 1998 pour des analyses), l'additif *aussi*, le marqueur de fin de séquence *bon* (Jayez, 2004) et le MD *enfin* analysé dans Razgoulieva (2002) et Dargnat (2021b). Malgré leurs différences, un certain nombre de ces MD ont en commun d'évoquer un contraste et/ou une concession ou une addition dans le schéma de type *non seulement ... mais aussi* et ont donc une affinité claire avec *mais*. Pour d'autres (*enfin* et *bon*), l'association avec *mais* permet de clore une continuation discursive potentielle, paraphrasable par « je pourrais continuer de parler mais je m'arrête là ».

4.3. Les mesures DPF et DPB

Ces mesures donnent des valeurs entre 1 et -1. Nous nous sommes fixé un seuil de 0,02 en utilisant les mêmes seuils d'effectifs que pour la MI^3, à savoir 10 et 8. La DP(F/B) mesure la prédictibilité d'une forme à partir d'une autre, ici le fait que *mais* prédise ses associés ou l'inverse. Concrètement, plus le résultat de la mesure est bas, moins la prédictibilité est bonne.

4.3.1. MD à gauche

Les résultats globaux suggèrent que les associés gauches sont des prédicteurs de *mais* plus nombreux et plus efficaces (valeurs de la DPF) que *mais* vis-à-vis de ses associés gauches (valeurs de la DPB). Dans le Tableau 3, où toutes les tailles de fenêtre sont regroupées, l'astérisque note les expressions dont le score est d'au moins 0.1, qui est ici un score de bonne prédiction. L'asymétrie DPF-DPB suggère que certaines expressions ont une affinité relativement spécifique avec les types de relations de discours que *mais* véhicule (en particulier le contraste, le concession et l'opposition) et que cette affinité est suffisamment prononcée pour laisser une trace quantitative, détectée par la DPF. Cela n'implique pas que ces expressions soient réservées à ces relations de discours, puisqu'on peut les trouver avec des relations causales, explicatives, etc.

[8] Ce terme générique cache des différences complexes liées aux présupposés sur l'état de croyances et l'état attentionnel de l'interlocuteur.

	DPF	DPB
CORPAIX	*non*, *oui*, *ouais*, *oh*, *hein*, *voilà*, *oui non**, *ouais non**, *mais non**, *ah non**, *ben oui*, *non non*, *ah oui*, *oui oui*	*oui*, *non*
CRFP	*non*, *oui*, *voilà*, *quoi*, *ouais*, *hein*, *non non**, *oui oui*	*oui*, *non*
DECLICS	*non**, *oui*, *ouais*, *hein*, *non non*	*oui*, *non*
ESLO1	*non**, *oui*, *si vous voulez*, *hein*, *mais oui**, *oui non**, *mais non**, *ben oui**, *bah oui**[9], *non non**, *ah non*, *oui oui*	*oui*, *non*
FRA80	*non*, *hein*, *oui*	*oui*, *non*

Tableau 3. Résultats de la DPF et de la DPB à gauche du pivot *mais*.

Ces résultats montrent que, parmi les prédicteurs de *mais*, ce sont ici encore les *oui*, *non* et leurs composés qui dominent. Comme nous l'avons noté, cette observation est en accord avec la nature argumentative de *mais*. En revanche, cela n'explique pas que certaines combinaisons (*mais oui*, *non non*, etc.) soient de meilleurs prédicteurs que les MD simples. Il y au moins quatre éléments à mentionner ici. (i) Les combinaisons avec astérisque utilisent *oui* et *non*, ce qui indique qu'elles participent à des mouvements de réfutation / rejet ou concession. (ii) La présence d'une répétition plutôt qu'un MD simple peut correspondre à une plus forte expressivité (*non non* ou *mais non mais non* dans ESLO1). (iii) Les combinaisons permettent de communiquer des mouvements complexes de manière condensée. Dans l'échange (6), L1 explique que le succès d'une vente commerciale dépend du contact qui se noue avec la clientèle.

(6) L1 – ah oui ça c'est sûr qu'il y a des il y a des gens avec lesquels tu ne passes absolument pas euh il y a des il y a des clients enfin il y a des il y a des il y a des clients qui sont euh
L2 – ça dépend le type de clientèle l'âge ça y joue aussi ça y joue aussi
L1 – oui non mais c'est vrai bon et puis bon c'est p-c'est p- c'est plus expliqué (CORPAIX)

À la remarque de L2 sur le rôle de l'âge, il répond par un feedback ou une approbation (*oui*) puis utilise un type de *non mais* qui introduit habituellement l'expression d'un accord à travers l'évocation d'une croyance attribuée à l'interlocuteur (L2), selon laquelle ce qu'il affirme ou sous-entend pourrait ne pas être accepté par le locuteur (L1)[10]. Le *non* exclut que L1 refuse le rôle de l'âge et le *mais* introduit l'argument qui appuie cette exclusion (« non je ne refuse pas que l'âge joue un rôle, mais au contraire je suis d'accord »). Si cette interprétation est correcte, le *oui* se combine donc sémantiquement avec un *non mais* dont la valeur peut apparaître soit comme opaque soit comme dérivable des valeurs de *non* et de *mais*. (iv) Les séquences de MD n'entrainent pas forcément un

[9] *Bah* est très vraisemblablement une variante de transcription de *ben*. La distinction n'affecte pas la mesure.

[10] Exemple typique : « Non mais je suis d'accord sur ce point-là, la barre n'est pas la même » (https://twitter.com/srhxsdy/status/1250530819204091906)

traitement cognitif plus lourd. En fait, les études psycholinguistiques des figements vont dans le sens d'une facilitation du traitement vraisemblablement parce qu'une signification complexe est encapsulée dans une expression toute prête (voir Tremblay, 2009 ; Gries, 2010). Ces quatre aspects pourraient contrbuer à expliquer pourquoi des combinaisons de marqueurs collaborent facilement avec *mais.*

4.3.2. Marqueurs à droite

La situation est analogue à celle des marqueurs à gauche : les associés droits sont des prédicteurs de *mais* plus nombreux et plus efficaces (valeurs de la DPB) que *mais* vis-à-vis de ses associés droits (valeurs de la DPF). Cependant, les associés repérés à droite diffèrent de ceux récoltés à gauche, probablement pour deux raisons : (i) la place de *mais* à l'initiale qui exclut les MDC mais pas les MDP à gauche ; (ii) la dynamique du discours : avec les associés à droite, les relations de discours propres à *mais* sont déjà introduites et ouvrent des possibilités d'enchaînements différentes de celles qui existent lorsque les mêmes relations de discours ne sont pas encore introduites (associés à gauche).

	DPF	DPB
CORPAIX	*euh, enfin*	*autrement**, *en tout cas**, *en même temps**, *sinon*, *par contre*, *disons*, *enfin*, *je veux dire*, *bon*, *tu sais*, *enfin euh**, *bon euh*
CRFP	*bon, euh*	*en même temps**, *enfin*, *bon*, *enfin bon*, *bon euh*, *euh bon*
DECLICS	*euh, bon*	*bon*, *après*
ESLO1	*enfin*	*enfin**, *par contre**, *autrement**, *disons*, *enfin quand même**, *enfin euh**, *enfin pour**, *alors là*, *alors euh*
FRA80	*enfin*	*enfin**, *autrement**

Tableau 4. Résultats de la DPF et de la DPB à droite du pivot *mais*.

4.4 Bilan

La fonction générale de *mais* est de mettre en cause un objet interprétatif au sens large. Les scénarios de détail sont variés et dépendent du type d'objet interprétatif ciblé par *mais*. Par exemple, *mais* peut remettre en cause un acte de langage, ses conditions de légitimité, un sous-entendu ou un comportement (voir Dargnat, 2021c). Nous avons regroupé ces scénarios sous l'étiquette « argumentatif » en suivant et étendant l'approche fondamentale d'Anscombre et Ducrot. Si l'on revient sur les MD qui s'associent à *mais*, on voit qu'on peut les classer en 4 groupes.

– Groupe 1 : ceux qui ont un rôle direct dans le fonctionnement argumentatif : *oui*, *non*, *ouais* et leurs combinaisons (*mais non*, *ben oui*, etc.), *par contre*, *quand même, en fait*, *aussi*, *maintenant*, *en même temps* et *après* dans leurs interprétations non temporelles.

– Groupe 2 : les MD qui interagissent avec le fonctionnement argumentatif en abandonnant son élaboration (*bon*) ou en évaluant l'acte *mais* sous un certain angle (*enfin*, *en tout cas*).

– Groupe 3 : les MD qui se greffent sur le mouvement argumentatif sans y participer directement, comme des marques d'approximation (*disons*, *je veux dire*, *si vous voulez*), de repérage énonciatif (*alors*), de « bilan » (*voilà*), d'hypothèse (*si*), de repérage temporel (*quand*), d'appel à l'interlocuteur (*voyez*, *tu sais*, *vous savez*, *hein*), d'évaluation de l'acte (*quoi*, *par exemple*), d'hésitation (*euh*, *hum*), de variation du niveau émotionnel ou attentionnel (*ah*, *oh*). Ce sont ces MD que nous avons désignés comme des « faux positifs sémantiques » parce qu'ils sont détectés par les mesures d'association mais n'entretiennent pas une relation directe avec le fonctionnement de *mais*.

– Groupe 4 : les lexèmes dont le statut sémantique exact est douteux (*là*, *pour*, *à*, *sans*, *en*), et que nous avons laissés de côté.

La mesure MI^3 capture beaucoup plus d'associés des types 3 et 4 (les associés « indépendants » dans notre terminologie), que les mesures DPF et DPB qui détectent des associés « intégrés », dont les interprétations collaborent avec celles de *mais*. Ces deux dernières mesures apparaissent donc comme plus sélectives. La principale explication de cette différence est que la DPF et la DPB sont sensibles à la *dépendance*, alors que la MI^3 est sensible à la fréquence. Étant donné que les MD instaurent des relations de discours ou des images du locuteur, il n'est pas illogique que DPF et DPB reflètent les affinités entre ces relations et ces images. Étant donné par ailleurs que cette recherche gravite autour d'un seul MD, la cohérence des résultats de dépendance sera fonction du profil de ce MD, qui peut être plus ou moins homogène. Dans le cas de *mais*, comme on l'a vu, les associés gravitent majoritairement autour de la fonction de mise en cause argumentative.

5. Conclusion

L'analyse quantitative proposée montre que, dans le domaine des MD, *mais* favorise certaines associations, plus restreintes dans le cas des mesures de prédiction (DPF et DPB), parce que centrées sur la fonction de *mais* en vertu de l'importance de la dépendance pour de telles mesures.

Les résultats concernant les associés intégrés, bien repérés par la DPF et la DPB, soulèvent la question de leur figement. Il est possible que parmi toutes les cooccurrences relevées, certaines fassent à présent partie du « prêt à parler » des locuteurs du français. Les meilleurs candidates seraient celles qui sont détectées à la fois par la MI^3 (fréquence de l'association) et par la DPF et/ou la DPB (force de la dépendance). On mentionnera en particulier *bon*, *enfin* et *alors* à droite de *mais*. L'utilisation conjointe d'une mesure de fréquence et d'une mesure de prédictibilité permet de prendre en compte en même temps le facteur de répétition et la proximité de fonction discursive.

Ces figements n'impliquent pas une perte de compositionnalité. Il est en fait très difficile de statuer sur ce point, car, dès qu'on quitte certains domaines bien connus (modification par un adjectif, valence verbale), l'étude de la (non-)combinaison des significations donne lieu à des analyses très diverses et

pas forcément comparables, comme en témoigne la vaste littérature sur les MD ou, dans d'autres domaines, sur la détermination, l'évidentialité, la modalité ou la scalarité. Dans ces cas complexes, la (non-)compositionnalité est relative à une représentation et ne semble pas se laisser appréhender dans l'absolu. Une prolongation du présent travail consistera précisément à élaborer des paramètres de représentation du fonctionnement discursif des MD afin de mieux aborder la question de la (non-)compositionnalité des associations de MD.

Bibliographie

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147-149.

Anscombre, J. C., & Ducrot, O. (1983). *L'argumentation dans la langue*. Bruxelles : Mardaga.

Asher, N., & Lascarides, A. (2003). *Logic of Conversation*. Cambridge: Cambridge University Press.

Auchlin, A. (1981). *Mais*, *heu*, *pis bon*, *ben alors voilà*, *quoi !* Marqueurs de structuration de la conversation et complétude. *Cahiers de linguistique française, 2*, 141-159.

Bender, E. M. (2013). *Linguistic Fundamentals for Natural Language Processing. 100 Essentials from Morphology and Syntax*. Morgan & Claypool.

Benzitoun, C., Fort, K., & Sagot, B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe. In *Actes de JEP-TALN-RECITAL 2012* (Grenoble)*, 2*, 99-112.

Brezina, V. (2018). *Statistics in Corpus Linguistics. A Practical Guide.* Cambridge: Cambridge University Press.

Bruxelles, S., Ducrot, O., Fouquier, E., Gouazé, J., dos Reis Nunes, G., & Rémis, A. (1976). Mais occupe-toi d'Amélie. *Actes de la Recherche en Sciences Sociales*, *2-6*, 47-62.

Burger, H., Dobrovolskij, D., Kühn, P., & Norrick, N.R. (2007). *Phraseologie/Phraseology, An International Handbook of Contemporary Research.* Berlin/New York: Mouton de Gruyter.

Church, K., & Hanks, P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, *16*(1), 22-29.

Col, G., Danino, C., & Rault, J. (2015). Éléments de cartographie des emplois de *voilà* en vue d'une analyse instructionnelle. *Revue de Sémantique et Pragmatique, 37,* 37-59.

Dargnat, M. (2021a). Interjections et particules de discours. In A. Abeillé & D. Godard (dir.), *Grande Grammaire du Français*, tome 2 (pp. 2015-2025). Arles /Paris : Actes Sud / Imprimerie Nationale.

Dargnat, M. (2021b). *Mais enfin :* construction et association. *Langages* (sous presse).

Dargnat, M. (2021c). Explication et argumentation. In M. Blasco & E. Auriac-Slusarczyk (dir.), *Parler à l'hôpital : écouter ce qui est dit, décrypter ce qui se dit.* Münster : NODUS Publikationen. (sous presse).

Dargnat, M. (à paraitre) Les particules énonciatives. *Encyclopédie Grammaticale du Français*. http://encyclogram.fr

Dargnat, M., Bartkova, K., & Jouvet, D. (2015). Discourse particles in French: prosodic parameters extraction and analysis. In A.-H. Dediu, C. Martín-Vide & K. Vicsi (Eds), *Statistical Language and Speech Processing* (pp. 39-49). Cham: Springer.

Dargnat, M., & Jayez, J. (2020). Presupposition projection and main content. In A. Abeillé & O. Bonami (Eds), *Constraint-based Syntax and Semantics. Papers in Honor of Danièle Godard* (pp. 99-121). Stanford: CSLI Publications.

Desagulier, G. (2015). Le statut de la fréquence dans les grammaires de constructions : *simple comme bonjour. Langages, 197*, 99-128.

Dostie, G., & Pusch C. (2007). Les marqueurs discursifs. Sens et variation. *Langue française*, *154*(2), 3-12.

Ducrot, O., *et al.* (1980), *Les mots du discours.* Paris : Éditions de Minuit.

Dziadkiewicz, A. (2007). La traduction automatique de phraséologismes pragmatiques : quelles représentations à travers la diversité formelle et culturelle ? *Corela, 5*(2). doi : 10.4000/corela.383

Ellis, N. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1-24.

Evert, S., & Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the ACL*, 188-195.

Franckel, J.-J. (1987). *Alors – Alors que*. *BULAG*, *13*, 17-49.

Goldberg, A. (2006). *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Goldberg, A. (2019). *Explain Me This. Creativity, Competition, and the Partial Productivity of Constructions*. Princeton / Oxford: Princeton University Press.

Gries, S. Th. (2010). Corpus linguistics and theoretical linguistics: a love-hate relationship? Not necessarily… *International Journal of Corpus Linguistics*, *15*(3), 327-343.

Gries, S. Th. (2019). 15 years of collostructions: some long overdue additions/corrections (to/of actually all sorts of corpus-linguistics measures). *International Journal of Corpus Linguistics, 24*(3), 385-412.

Hanks, P. (2013). *Lexical Analysis. Norms and Exploitations*. Cambridge (MA)/Londres: MIT Press

Hansen, M.-B. (1998). *The Function of Discourse Particles. A Study with Special Reference to Spoken Standard French*. Amsterdam: Benjamins.

Hilpert, M. (2014). *Construction Grammar and its Application to English*. Edinburgh: Edinburgh University Press.

Jayez, J. (1988). *Alors* : description et paramètres. *Cahiers de Linguistique Française*, *9*, 133-175.

Jayez, J. (2004). *Bon. Le mot de la fin.* Manuscrit non publié. http://perso.ens-lyon.fr/jacques.jayez/doc/bon.pdf

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*(1), 1-17.

Lee, L. (2021). *Fonctions pragmatiques et prosodie des marqueurs discursifs en français et en anglais*. Thèse de l'Université de Lorraine.

Mel'čuk, I. (2013). Tout ce que nous voulions savoir sur les phrasèmes, mais… *Cahiers de lexicologie, 102*, 129-149.

Métrich, R., Faucher, E., & Courdier, G. (2002). *Invariables difficiles, Dictionnaire allemand-français des particules, connecteurs, interjections et autres mots de la communication*. Nancy : Groupe de lexicographie franco-allemande et ATILF.

Razgoulieva, A. (2002). Combinaison des connecteurs *mais enfin*. *Cahiers de Linguistique Française, 24*, 143-68

Rossari, C., Ricci, C., & Wandel, D. (2018). Introducteurs de cadres et connecteurs de reformulation : étude contrastive sur corpus. *Langages*, *212*, 51-67.

Roze, C., Danlos, L., & Muller, P. (2012). LEXCONN: A French Lexicon of Discourse Connectives. *Discours*, *10*. http://journals.openedition.org/discours/8645

Schneider, U. (2020). Δ*P* as a measure of collocation strength. *International Journal of Corpus Linguistics*, *16*(2), 249-274.

Taboada, M., & Mann, W. C. (2006). Rhetorical Structure Theory: Looking Back and Moving Ahead. *Discourse Studies, 8*(3), 423-459.

Tremblay, A. (2009). *Processing advantages of lexical bundles: Evidence from self-paced reading, word and sentence recall, and free recall with event-related brain potential recordings*. Ph.D. Dissertation, University of Alberta.

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: investigating the interaction of words and constructions. *International Journal of Corpus Linguistics, 8*(2), 209-243.

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.

Corpus

CORPAIX : *CORpus d'AIX-en-Provence*, équipe DELIC, Aix-en-Provence.

CRFP : *Corpus de Référence du Français Parlé*, équipe DELIC, Aix-en-Provence.

DECLICS : *Dispositif d'Etudes CLIniques sur les Corpus de Santé*. http://lrl.uca.fr/rubrique97.html

ESLO1 : *Enquêtes SocioLinguistiques à Orléans*, http://eslo.huma-num.fr/index.php/pagecorpus/pageaccescorpus

FRA80 : *Cahiers du Français des Années Quatre-Vingts*, hors-série 1 (1989), CREDIF, ENS de Saint-Cloud.