MDMA. Identification et annotation des marqueurs discursifs « potentiels » en contexte

Catherine T. Bolly
CNRS & UMR 7023 SFL ; Université catholique de Louvain
catherine.bolly@uclouvain.be

Ludivine Crible
Université catholique de Louvain
ludivine.crible@uclouvain.be

Liesbeth Degand
Université catholique de Louvain
liesbeth.degand@uclouvain.be

Deniz Uygur-Distexhe
Université catholique de Louvain
deniz.uygur@uclouvain.be

Résumé

Partant du constat qu'il n'existe pas de catégorie fermée de Marqueurs Discursifs (MD) et que la définition de ces marqueurs varie fortement selon le cadre épistémologique adopté, l'objectif du projet MDMA (« Model for Discourse Marker Annotation ») est d'établir une méthode empirique d'identification et d'annotation des MD en français oral. La méthode vise tout d'abord à décrire les MD en faisceaux de variables et ensuite, d'un point de vue combinatoire, en patrons spécifiques. Notre démarche comprend trois étapes : (i) repérage manuel de tous les MD dits « potentiels » dans un corpus équilibré en français oral (5.000 mots ; Belgique et France) ; (ii) extraction automatique de toutes les formes qui correspondent aux MD potentiels identifiés précédemment (1.181 occurrences) ; (iii) analyse paramétrique d'un échantillon aléatoire de 200 MD potentiels en contexte (variables syntaxiques, formelles et sémantico-pragmatiques). L'hypothèse est que l'analyse statistique des contraintes distributionnelles imposées aux différents MD potentiels devrait révéler une certaine hiérarchisation entre variables annotées, concernant leur pertinence, leur fiabilité et leur généralisabilité (voire leur spécificité). Dans cet article, nous présenterons les principes d'annotation des MD, aborderons ensuite la problématique de l'accord inter-juges, pour finalement discuter de manière plus approfondie les résultats de l'analyse sur corpus.

Mots clés : marqueurs discursifs, annotation, corpus-based, analyses multivariées, français parlé

Abstract

Starting from the common observation that there is no recognized closed class of Discourse Markers (DMs) and that their definition may vary from one theoretical framework to another, the aim of the MDMA project ("Model for Discourse Marker Annotation") is to establish an empirical method for the identification and annotation of DMs in spoken French. Central to our proposal is that DMs may be described as clusters of features that, in specific patterns of combination, allow distinguishing between more or less prototypical uses of DMs in context. We proceeded in three steps : (i) manual identification of all so-called "potential" DMs in a balanced corpus of spoken French (5,000 words; Belgium and France) ; (ii) automatic extraction from the corpus of every token corresponding to the DMs candidates previously identified (1,181 tokens) ; and (iii) parameter analysis of a random sample of 200 potential DMs (syntactic, formal and semantic-pragmatic variables). The hypothesis is that the statistical analysis – based on the distributional constraints of the potential DMs at stake - should uncover a certain hierarchy between the different features under scrutiny, regarding their relevance, reliability, and generalizability (or even specificity). In the present paper, we first present the annotation procedure, then we discuss several aspects of inter-rater agreement, and finally discuss the results from the in-depth corpus-based and statistical analyses.

Keywords : discourse markers, annotation model, corpus-based, multivariate analysis, spoken French

1. Introduction

En analyse du discours, on reconnaît généralement l'absence de consensus sur les frontières, voire sur l'existence même (*cf.* Pons Bordería, 2008) d'une catégorie homogène qui rassemblerait de manière exhaustive et inclusive les éléments pragmatiques définis comme étant des « marqueurs discursifs » (MD) : « [i]t has become standard in any overview article or chapter on DMs to state that reaching agreement on what makes a DM is as good as impossible, be it alone on terminological matters » (Degand *et al.*, 2013 : 5). En effet, l'inclusion ou non d'une expression linguistique dans la catégorie des MD dépend largement de la définition adoptée (*cf.* Schourup, 1999 : 228), ce qui rend la portée et les résultats d'analyses peu comparables au sein de la communauté scientifique. Par ailleurs, les études de cas, qui prolifèrent depuis plusieurs décennies dans le domaine, témoignent d'une grande variété concernant (i) l'objet d'étude (le plus souvent un MD spécifique – *cf.* Dostie et Pusch, 2007), (ii) les données d'étude (variées en termes de langues et de registres) et (iii) les cadres théoriques adoptés (analyse conversationnelle, théorie de la pertinence, pragmaticalisation, théories de l'énonciation, etc.).

Dans le cadre de cette situation peu consensuelle et en réponse aux particularismes parfois contradictoires qui jalonnent le champ d'étude des MD, nous proposons de développer une méthode empirique inédite pour l'identification et l'annotation systématique – potentiellement automatisable – des MD en contexte. Cette méthode est le fruit du projet collaboratif MDMA (« Model for Discourse Marker

Annotation ») dont l'objectif principal est de couvrir toutes les étapes d'analyse des MD (de l'identification à l'annotation), tout en fournissant des critères sur corpus valides et réplicables qui permettent de mieux cerner les frontières de la catégorie des MD. Le présent article entend mettre l'accent sur la validation statistique du modèle et de la méthode adoptée, visant *in fine* une meilleure compréhension de ces éléments fréquents, hétérogènes et multifonctionnels que sont les MD.

Après un rapide point terminologique et définitoire sur la catégorie des MD et leurs fonctions en discours, l'article présente les hypothèses et les étapes de notre méthode, en détaillant la phase d'annotation et ses résultats. Ceux-ci sont présentés sous la forme de tableaux et de figures mettant en exergue la hiérarchisation des variables annotées. Cette hiérarchisation est le produit de méthodes statistiques multivariées qui ont pour but d'identifier des profils plus ou moins prototypiques de MD.

2. Marqueurs discursifs « potentiels »

Dans notre étude, les MD sont envisagés selon une perspective délibérément peu restrictive de la catégorie, laissant *de facto* à l'analyse sur corpus le soin de fournir ultérieurement des distinctions plus fines. Dans le cadre du projet MDMA, le terme « marqueurs discursifs » (*discourse markers – cf.* Schiffrin, 1987) couvre des éléments très divers, tels que les connecteurs *parce que* et *donc*, mais aussi des marqueurs non-relationnels ou interactifs comme *tu vois* et *ben*. Les MD correspondent donc ici à ce que d'autres appellent par ailleurs les « marqueurs pragmatiques » (*pragmatic markers – cf.* entre autres Brinton, 2008 ; Aijmer et Simon-Vandenbergen, 2011), catégorie qui inclut de nombreux types d'éléments à fonction pragmatique (interjections, particules modales, signaux de réponse) globalement réunis par leur fonction d'indice à l'interprétation du contexte, qui dépasse le décodage sémantique de base.

> A pragmatic marker is defined as a phonologically short item that is not syntactically connected to the rest of the clause (i.e., is parenthetical), and has little or no referential meaning but serves pragmatic or procedural purposes (Brinton, 2008 : 1).

Cette procéduralité évoque le profil fonctionnel défini par Hansen (2006), pour qui les MD « provide instructions to the hearer on how to integrate their host utterance into a developing mental model of the discourse in such a way as to make that utterance appear optimally coherent » (2006 : 25). Sans entrer plus en détail dans le panel fonctionnel des différents membres de la catégorie, on peut donc *a minima* résumer le rôle des MD à leurs fonctions métalinguistique et intersubjective, qui couvrent des usages spécifiques très divers. La multifonctionnalité des MD contribue par ailleurs à rendre la tâche définitoire compliquée et potentiellement conflictuelle, car elle fait notamment se confronter des approches monosémique et polysémique (*cf*. la notion de *meaning potentials* – Aijmer, 2013) et tend à une fragmentation de la catégorie en sous-groupes fonctionnels (marqueurs relationnels *vs*. non-relationnels, par exemple).

Concernant leur comportement syntaxique, les MD peuvent provenir d'une variété de catégories grammaticales, principalement des conjonctions de coordination et de subordination (*et, donc, parce que, même si*) (ente autres, Bolly et Degand, 2009), des adverbes (*alors, enfin*) (entre autres, Barnes, 1995 ; Degand et Fagard, 2011) et des locutions prépositionnelles (*en fait, par contre*) (Lewis, 2006 ; Mortier et Degand, 2009 ; Defour *et al.*, 2010). Ils peuvent également appartenir à d'autres classes moins

prototypiques, comme c'est le cas des constructions parenthétiques verbales (*tu vois, je pense*) (Bolly, 2010, 2012a), des adjectifs (*bon*) (Waltereit, 2007) ou des particules, c'est-à-dire des interjections ou autres expressions sous-déterminées (*euh, ben*) (Hansen, 1995, 1998), ainsi que des acronymes (*lol*, *mdr*) (Uygur-Distexhe, 2012, 2014).

Corollairement à cette diversité grammaticale, le degré d'intégration syntaxique et la portée de ces éléments varient aussi, entre des connecteurs reliant deux unités distinctes et des éléments plus mobiles (parenthétiques ou non). Une conséquence de cette diversité est la coexistence, en synchronie, de formes propositionnelles pleines [1] et de leurs usages discursifs, grammaticalisés, où ils fonctionnent plutôt sur le plan de l'interaction et de la relation entre interlocuteurs [2].

[1] c'était chouette parce que j'avais jamais été et **tu vois** tous les instruments / (VALIBEL : corpus famES1r)

[2] mais il viendra te poser des questions sur euh genre / **tu vois** où tes parents sont nés euh / (VALIBEL : corpus blaNB1l)

Si la désambiguïsation du MD *tu vois* dans ces deux exemples est assez évidente, avec notamment un sens premier de perception visuelle très présent en [1] par contraste avec l'éloignement sémantique marqué en [2], les MD doivent le plus souvent être envisagés selon un continuum de discursivité (Bolly, 2012b, 2014) où le sens est plus ou moins pragmatique selon le contexte.

La difficulté qui émerge de ce double constat – concernant la diversité à la fois fonctionnelle et syntaxique des MD – est qu'une définition représentative de la catégorie est un projet pour le moins ambitieux et complexe, en raison des profils très contrastés de ses membres. De plus, il convient de prendre en compte des considérations pratiques sur l'application de cette définition dans un processus d'identification des MD sur corpus authentique, puisqu'aux exigences théoriques de validité et d'extension de la définition s'ajoutent des critères d'opérationnalisation et de réplicabilité.

Notre contribution s'inscrit dans la perspective de réconcilier ces deux extrêmes, entre définition extensive et annotation, en proposant une méthode de désambiguïsation de MD potentiels par le biais d'une grille opérationnelle de paramètres, testée et améliorée sur corpus. Ainsi, le projet MDMA présente la double particularité (i) d'incorporer, dans la sélection d'éléments à annoter, ces formes dont la fonction discursive n'est pas systématiquement activée en contexte, mais qui sont néanmoins des MD « en puissance » (pour une référence à ces formes sélectionnées par défaut comme étant des MD dits « potentiels », voir la suite de l'article) et (ii) d'adopter une approche onomasiologique et graduelle (voir aussi Uygur-Distexhe, 2012 ; Ciabarri, 2013 ; Bolly et Crible, 2015 ; Crible, soumis) qui repose sur la description des MD en termes de faisceaux de variables formelles, syntaxiques et sémantico-pragmatiques (pour une description détaillée de ces variables, voir 3.3).

L'intérêt de notre étude réside donc principalement dans l'originalité du protocole d'annotation développé, qui vise à examiner quels sont les paramètres – et leurs combinaisons – particulièrement pertinents pour discriminer les MD « potentiels » des MD « confirmés » sur la base de variables annotées en contexte. L'hypothèse est que l'analyse statistique des contraintes distributionnelles imposées aux différents MD devrait révéler une certaine hiérarchisation entre les variables annotées, en soulignant leur degré de pertinence, de fiabilité et de généralisabilité (voire de spécificité) dans la

démarche d'identification et de catégorisation des MD. L'objectif ultime de MDMA est donc de faire émerger, à partir de données de corpus, des critères prédictifs, fiables et aussi reproductibles que possible, dans la perspective d'une annotation (semi-) automatique. Il s'agit donc de montrer qu'un modèle d'annotation fondé sur corpus constitue une méthode qualitative valable pour asseoir la validité et la pertinence de traits discriminatoires dans l'attribution du statut de MD. La méthode et la démarche d'annotation, ainsi que les données utilisées, sont présentées dans la section suivante.

3. Données et méthode

3.1. Corpus de français oral (semi-)spontané

Pour ce projet d'annotation, nous avons réuni un échantillon de français oral à partir d'extraits de trois corpus existants : Corpage (Bolly *et al.*, 2012), VALIBEL (Dister *et al.*, 2009) et CLAPI (Balthasar et Bert, 2005). L'échantillonnage contient 5.000 mots, équilibrés entre les variétés du français de Belgique et de France. Les textes sélectionnés correspondent au registre (semi-)spontané et se composent d'entretiens en face à face et de conversations spontanées. L'objectif de cet équilibrage et le choix du registre reposent sur l'hypothèse selon laquelle les MD sont plus fréquents (Brinton, 1996) et plus variés à l'oral spontané (Georgakopoulou et Goutsos, 1998). L'impact du type de textes sur les relations de cohérence a été démontré par Louwerse et Mitchell (2003), indiquant une fréquence significativement plus élevée de relations de cohérence, et en particulier de particules discursives, dans les textes dialogaux (*vs.* monologaux), oraux (*vs.* écrits) et spontanés (*vs.* formels). Comme décrit plus bas (sous 3.2.2), les MD potentiels étudiés ont été extraits de manière exhaustive et automatique à partir des données de corpus ainsi collectées.

3.2. Annotation des marqueurs discursifs potentiels

La méthodologie générale employée pour cette étude peut se qualifier de *corpus-based* (Biber *et al.*, 1994), c'est-à-dire qu'elle constitue un va-et-vient constant entre cadre théorique et données de corpus, l'un renforçant l'autre. La clé de voûte principale de notre entreprise est l'élaboration d'un protocole d'annotation qui rende compte de l'hétérogénéité du comportement syntaxique, sémantico-pragmatique, prosodique et collocationnel des MD potentiels, tout en répondant aux exigences d'opérationnalisation nécessaires à toute démarche d'annotation. Notre démarche comprend trois étapes majeures: (i) repérage manuel de tous les MD potentiels par trois spécialistes dans un corpus équilibré en français oral (5.000 mots ; Belgique et France) ; (ii) extraction automatique de toutes les formes correspondant aux MD potentiels identifiés précédemment (1.181 occurrences pour 152 MD types) ; (iii) analyse paramétrique d'un échantillon aléatoire de 200 MD potentiels en contexte (variables syntaxiques, formelles et sémantico-pragmatiques).

3.2.1 Identification des marqueurs discursifs potentiels

Dans une première étape, trois analystes experts ont identifié de manière indépendante et manuelle toutes les occurrences des formes qu'ils considéraient comme étant susceptibles d'être un MD au sein d'un corpus équilibré de 5.000 mots en français oral (semi-)spontané. À ce stade, comme mentionné précédemment, l'acception des MD a délibérément été la plus large possible, suivant la définition indexicale et procédurale des *marqueurs pragmatiques* (voir plus haut sous 2.), au point d'inclure certaines expressions présentant un sens pragmatique minimal, telles

que les adverbes modaux [3], les énoncés parenthétiques [4] ou les particules d'hésitation [5].

[3] et bon j'ai choisi ça **un peu** euh / comme ça parce que comme dernier choix en fait (VALIBEL : corpus ilrCP1r)

[4] au fond ce n'était pas le / le / le grand bonheur **si tu veux** / c'est c'est un peu / un peu raisonné (Corpage : corpus ageNM1)

[5] puis vous êtes **euh** / peut-être un peu conseiller **euh** / comme vous me disiez **euh** / la fois dernière // comme par exemple **euh** / avec les ruches / vous donnez des conseils (Corpage : corpus ageFB1)

Ce principe d'inclusivité se révèle indispensable, comme prérequis à l'analyse, pour ne pas exclure *a fortiori* de l'étude des éléments linguistiques situés à l'intersection de plusieurs niveaux de l'analyse linguistique (syntaxique, sémantique et pragmatique). C'est le cas, par exemple, de certains emplois parenthétiques (*tu vois*, *je pense* – Bolly, 2012b, 2014) et d'emplois pronominaux (*(c'est) ça* – Dostie, 2014), qui peuvent se caractériser à la fois par une certaine rétention syntaxique et une fonction proprement discursive. Cette sélection intuitive a produit un total de 152 types, tous analystes confondus, certains attendus (*alors que, ben, parce que*) et d'autres moins consensuels (*à mon avis, bien sûr, bref*). Parmi les expressions les moins prototypiques de la classe des MD, on trouve notamment des membres d'autres catégories comme des adverbes phrastiques (*à la limite*), des expressions figées (*au fur et à mesure*), des interjections (*ah*) ou encore des locutions verbales (*c'est bon*).

3.2.2 Extraction automatique des marqueurs discursifs potentiels

Dans un deuxième temps, toutes les occurrences (*tokens*) de ces 152 types (incluant leurs diverses graphies) ont été extraites automatiquement du corpus par le biais d'un concordancier. Il en a résulté un total de 1.181 occurrences de formes linguistiques susceptibles de remplir une fonction pragmatique en contexte, autrement dit, d'être un MD. Rappelons que ces occurrences sont des MD dits « potentiels », incluant donc des usages non discursifs [6] de formes susceptibles d'être employées par ailleurs comme MD [7], ainsi que des expressions dont le statut de MD est controversé selon la définition des MD, plus ou moins étendue, que l'on adopte [8].

[6] le loup s'empresse/ chemin pour arriver chez la mère-grand avant le petit chaperon rouge mais là faut mettre pas pris le **bon** chemin /euh/ (CLAPI : corpus Chaperon Rouge, « Jean-Pierre et Magali »)

[7] ben c'est-à-dire que **bon** comme **bon** quand je suis sorti de troisième année j'avais j'ai **bon** j'ai doublé ma troisième année (VALIBEL : corpus ilrCP1r)

[8] je travaille sur euh sur ordinateur hein sur Macintosh **et cetera** / oui hein ces derniers temps (VALIBEL : corpus ilrCP1r)

Ces trois groupes d'éléments forment un ensemble d'entités linguistiques qui, malgré leur hétérogénéité, devront être soumises à la même grille d'annotation. Par conséquent, les paramètres en jeu dans le codage ont été élaborés de manière

suffisamment flexible pour pouvoir s'appliquer à des expressions linguistiques qui divergent tant au niveau de leur forme que de leur fonction en contexte.

3.2.3 Annotation paramétrique : du marqueur « potentiel » au marqueur « confirmé »

L'étape suivante a consisté en l'élaboration du protocole d'annotation, sur la base de paramètres existant dans la littérature et reconnus comme faisant l'objet d'un certain consensus dans le domaine. Pour ce faire, nous avons donné la priorité à des variables formelles, syntaxiques et sémantico-pragmatiques, tout en couvrant le maximum de traits potentiellement impliqués dans la définition relativement consensuelle des MD. À notre connaissance, un tel protocole n'existe pas, et nous avons donc dû procéder à une sélection de critères mentionnés dans des définitions de MD en vigueur (entre autres Schiffrin, 1987 ; Vincent, 1993 ; Brinton, 1996, 2008 ; Fraser, 1999 ; Schourup, 1999 ; Gonzàlez, 2005 ; Dostie, 2007). Par la suite, nous avons adapté et opérationnalisé ces critères définitoires pour permettre une application optimale et réplicable sur corpus authentique. Ces variables couvrent des aspects syntaxiques, sémantico-pragmatiques, prosodiques et collocationnels. Parmi celles-ci, seules certaines variables se sont révélées être pertinentes pour la présente étude (pour les détails, voir le point suivant).

Ce travail théorique de paramétrisation de critères définitoires est intrinsèque aux différentes phases d'entraînement du protocole d'annotation, phases pendant lesquelles les analystes ont appliqué les versions successives du protocole au même jeu de données, puis ont discuté les résultats divergents. La résolution des désaccords a non seulement permis aux quatre chercheurs de coordonner leurs interprétations, mais également de renforcer la robustesse et la réplicabilité des critères prévus par le protocole, en améliorant constamment la précision des définitions, la pertinence des exemples et la représentativité des différentes valeurs possibles (voir Zufferey *et al.*, 2012 pour une démarche similaire). Ces phases d'annotation ont été réalisées sur un échantillon aléatoire de 200 MD potentiels, présentés avec leur contexte (150 caractères à gauche et à droite) dans un concordancier. Dans de rares cas, le contexte ne s'est pas révélé suffisamment large pour pouvoir décrire de manière non-ambiguë le marqueur visé. Dans ces cas, la valeur « non interprétable » (voir Tableau 1, plus bas) est attribuée au marqueur pour la variable paramétrique concernée.

En résumé, cette méthodologie peut s'envisager comme un échange constant et constructif entre théorie et données, qui résulte en un protocole d'annotation forgé par sa propre applicabilité à des corpus authentiques. Il va sans dire que la plupart des critères ont été reformulés, voire supprimés, d'autres ajoutés, au cours des multiples phases d'annotation, toujours dans l'objectif de maximiser l'efficacité du processus de codage. L'originalité de cette approche est de couvrir à la fois les MD et leurs équivalents non discursifs (en plus d'éléments limitrophes de la catégorie), en identifiant de manière fine et à partir de données observables en contexte, non seulement les critères d'appartenance de certaines expressions linguistiques à la catégorie des MD, mais aussi les critères qui les en excluent (pour une présentation des différents moyens de mesurer, qualitativement et quantitativement, la validité globale de cette méthode, voir la section sur les résultats).

3.3. Protocole d'annotation

Nous présentons ici les différents paramètres prévus par le protocole d'annotation dans sa version finale, ainsi que leurs critères d'application. Cette section reprend donc succinctement les huit variables en précisant leur définition, leur intérêt dans un modèle d'annotation de MD, et en discutant des aspects les plus sujets aux désaccords inter-juges.

Les paramètres qui ont été opérationnalisés pour l'analyse des MD potentiels peuvent être regroupés en trois catégories linguistiques majeures : syntaxique, sémantico-pragmatique et contextuelle. La catégorie syntaxique comprend trois variables formelles qui convergent pour décrire le comportement syntagmatique des MD potentiels : (i) l'identification de la catégorie syntaxique à laquelle appartient l'expression, prise dans son acception normée (par exemple, le statut d'adverbial a été attribué aux occurrences de *là* compte non tenu de leur possible emploi comme déictique textuel en contexte) et dans sa globalité (compte non tenu de la classe grammaticale des éléments qui composent, par exemple, les unités complexes ou phraséologiques) (*cf. self-category* dans le Penn Treebank, Marcus *et al.*, 1994) ; (ii) codage de la position du marqueur au sein de l'énoncé, par rapport à l'élément prédicatif ou recteur (le plus souvent un noyau verbal) de son énoncé hôte (Blanche-Benveniste *et al.,* 1990 ; Lindström, 2001 ; Blanche-Benveniste, 2003 ; Lindström et Karlsson, 2005) ; (iii) annotation du caractère mobile (ou non) du marqueur au sein de l'énoncé hôte.

Il convient ici de préciser quels sont les critères ayant permis de distinguer les différentes positions syntaxiques. Après avoir identifié le noyau prédicateur au sein de l'énoncé hôte, les éléments syntaxiquement dépendants de ce noyau et situés à l'intérieur de l'unité valentielle ont été identifiés comme étant en position médiane dans l'énoncé [9] ('Middle field'). Les expressions intégrées syntaxiquement à l'unité rectionnelle, mais situées en marge de la valence, ont été codées en position initiale interne [10] ('Initial field' : à gauche de l'unité valentielle) ou en position finale interne [11] ('End field' : à droite de l'unité valentielle). Enfin, les expressions non régies, situées en périphérie de la syntaxe de l'énoncé, sont considérées comme étant soit en position initiale externe [12] ('Pre-front field' : en périphérie gauche de l'unité rectionnelle) ou finale externe [13] ('Post field' : en périphérie droite de l'unité rectionnelle).

[9] mm et il y a **déjà** des documents que tu as créés dont tu es très très content (VALIBEL : corpus ilrCP1r)

[10] mais on se dit / tiens / **ça** c'est ton âge (Corpage : corpus ageFB1)

[11] il faut admettre que / bon voilà il y a des choses qu'on ne fait plus et c'est / ce n'est pas plus mal **aussi** (Corpage : corpus ageNM1)

[12] **mm** c'est ça oui oui (VALIBEL : corpus ilcDA1r)

[13] oui hein ces derniers temps ça prend beaucoup de d'expansion **quoi** (VALIBEL : corpus ilrCP1r)

À noter que, par défaut, les conjonctions et prépositions sont codées en position initiale interne ('Initial field'); toutefois, quand elles introduisent un élément non régi

(souligné dans l'exemple [14]), elles sont codées en position externe non régie ('Pre-front field' ou 'Post-field').

[14] donc / on va en venir à l'entretien / à proprement dit // **et** c'est plus / un entretien cette fois-ci / basé sur comment vous vous sentez actuellement euh / dans votre âge (Corpage : corpus ageFB1)

Quant aux particules, elles sont uniquement considérées au niveau de leur position de surface dans l'énoncé (*cf.* la particule *euh* considérée comme étant en position 'Pre-front' dans l'exemple [15] et en 'Middle field' dans l'exemple [16]), au vu de la difficulté de déterminer leur degré d'intégration syntaxique via l'application de tests de transformation traditionnels (focalisation, pronominalisation, transformation interrogative, etc.).

[15] alors / je vais commencer par ma première question // **euh** // ben / d'après votre expérience / qu'est-ce qui change au fur et à mesure lorsque vous avancez dans l'âge (Corpage : corpus ageFB1)

[16] bon ben je ce qui est assez intéressant maintenant c'est qu'on peut **euh** apporter **euh** / sa façon de voir les choses **euh** sa sa **euh** (VALIBEL : corpus ilrCP1r)

Ces trois paramètres syntaxiques (catégorie, position et mobilité), en tant qu'indices de surface se situant au niveau formel du langage, présentent l'avantage de ne faire intervenir qu'un minimum d'interprétation subjective de la part de l'analyste. Toutefois, force est de constater que le codage de la position peut faire l'objet de quelques désaccords, car il peut varier en fonction de l'interprétation sémantique du contexte. Ainsi, dans l'exemple [17], le MD potentiel *faut dire que* peut être envisagé soit comme étant intégré à la structure verbale de l'énoncé, dans son sens propositionnel plein, soit comme périphérique, en marge de la syntaxe de dépendance, dans un usage discursif plus marqué. Cette double interprétation a conduit à des annotations divergentes d'un codeur à l'autre en ce qui concerne la position du marqueur. Des divergences d'annotation entre codeurs peuvent également résulter d'une interprétation différente des frontières de l'unité hôte. En [18], le MD potentiel *là* peut être considéré comme clôturant le noyau verbal porté par *passer* ou comme introduisant celui porté par *comprendre*.

[17] non mais **faut dire qu'**en fait il a réfléchi que il pouvait non ? (CLAPI : corpus Chaperon Rouge, « Jean-Pierre et Magali »)

[18] mais après en expliquant ce qui se passe **là** on comprend tout de suite c'est c'est /logique (CLAPI : corpus Chaperon Rouge, « Jean-Pierre et Magali »)

Notons que l'absence de recours au signal sonore repose ici sur un principe méthodologique de dissociation des niveaux de l'analyse linguistique. Il ne fait cependant aucun doute que certaines divergences d'interprétation pourraient être résolues par l'écoute des extraits. Une étude portant sur le rôle du contexte sonore et visuel dans l'interprétation des marqueurs discursifs est actuellement en cours dans le cadre du projet MDMA (Bolly et Crible, 2015).

La deuxième catégorie regroupe deux variables sémantico-pragmatiques : (i) le degré de procéduralité et (ii) le maintien du sens codé (*coded meaning*) du MD

potentiel en contexte. Ces deux paramètres donnent respectivement une indication sur l'éloignement du marqueur potentiel par rapport (i) à son sens conceptuel (s'il en a un) et/ou (ii) son sens conventionnel, prototypique ou premier. Le degré de prototypicalité des MD potentiels a été déterminé en s'appuyant sur une ressource lexicographique spécialisée (à savoir, le TLFi), afin d'éviter toute considération de saillance cognitive ou de prototypicalité fréquentielle (*cf.* Gilquin, 2008) difficilement opérationnalisable dans le cadre de cette étude, tout en garantissant la réplicabilité de l'analyse par le choix d'un outil librement accessible et reconnu par la communauté scientifique. À noter que lorsqu'une acception était absente du TLFi, nous avons eu recours à la littérature scientifique pour déterminer le sens prototypique du marqueur. Par exemple, nous avons formulé le sens de *eh ben* comme 'introduisant une situation contextuelle où la réaction introduite est inattendue', en nous basant sur les études menées par Gülich (1970) et Hansen (1995) sur le MD *ben*. La procéduralité se réfère, quant à elle, à la présence d'une valeur d'instruction guidant l'interprétation, présente ou non dans le MD en contexte, et s'oppose ainsi aux expressions purement conceptuelles à fonction référentielle (*cf.* Bolly et Degand, 2013). Cette catégorie d'annotations sémantiques, bien qu'intrinsèquement plus interprétatives que la précédente, porte à relativement peu de désaccords, excepté de rares cas d'emplois phraséologiques, comme dans l'exemple [19] où il est difficile d'affirmer que l'adverbe *bien* dans *j'aimerais bien* est sémantiquement similaire au sens premier de 'conformité à une norme de qualité'.

[19] j'aimerais **bien** travailler dans ce genre de matériel (VALIBEL : corpus ilrCP1r)

Ces deux paramètres contribuent à déterminer l'ancrage sémantico-pragmatique des MD potentiels et leur actualisation en contexte.

La troisième catégorie concerne les paramètres dits « contextuels » qui visent à décrire : (i) la présence ou non d'une pause contiguë, directement à gauche et/ou à droite du MD potentiel, repérée grâce aux conventions de transcription adoptées dans les corpus respectifs (la barre oblique simple ou double pour les pauses perçues dans Valibel et Corpage ; les parenthèses avec un point pour les micro-pauses et avec un chiffre pour les pauses chronométrées dans CLAPI) ; (ii) la position du MD potentiel au sein du tour de parole, d'après l'hypothèse selon laquelle les MD peuvent être mobilisés dans des stratégies interactives de prise ou de cession de parole (*cf.* Fischer, 2000). Ces variables, moins essentielles à la définition de la catégorie des MD, fournissent néanmoins des caractéristiques formelles fiables du comportement contextuel des MD dans la chaîne parlée, laissant peu de place aux désaccords inter-juges.

Enfin, la dernière variable, non linguistique, consiste en un codage graduel et chiffré (de 1 à 3) qui rend compte du degré de certitude des annotateurs quant au statut de MD de l'item étudié. En d'autres termes, il s'agit d'indiquer, pour chacune des 200 occurrences en contexte, dans quelle mesure le MD potentiel concerné peut être catégorisé comme étant un MD ou non. Cette variable de l'annotation permet de garder une trace de l'aspect intuitif de la sélection effectuée en première étape de la méthode (pour rappel, voir la section précédente). Elle fournit par ailleurs un point de comparaison, à dimension qualitative, pour l'interprétation des résultats statistiques qui ont émergé de l'analyse paramétrique, plus objective. Contrairement à tous les autres paramètres, pour lesquels les désaccords sont résolus par discussion ou par une règle de majorité, les scores donnés par les différents analystes pour le degré de

certitude sont résumés sous forme d'une moyenne arithmétique pour chaque MD potentiel pris dans son contexte d'énonciation, constituant la valeur de référence dans l'analyse des résultats.

Tous ces paramètres sont pris en considération indépendamment les uns des autres lors de la phase d'annotation. Bien que la plupart d'entre eux restent difficiles à systématiser, la part de subjectivité de l'analyste a été réduite au minimum possible, et ce grâce à la précision et à l'opérationnalité des différents critères, ainsi qu'au choix de variables les plus objectivables possibles (pour une évaluation quantitative de la validité générale de ce protocole d'annotation, voir la section 4.1 des résultats). Les paramètres ainsi que les valeurs qui peuvent leur être attribuées sont repris dans le Tableau 1.

	PARAMETRES	VALEURS
SYNTAXE	1. Catégorie syntaxique	Syntagme verbal, Syntagme nominal, Pronom, Syntagme adjectival, Syntagme adverbial, Conjonction/locution conjonctive, Préposition/locution prépositionnelle, Particule ou interjection, Autre
	2. Position syntaxique	*Pre-front field* : initiale, non-régie, à gauche
		Initial field : initiale, régie, à gauche
		Middle field : élément prédicateur ou intégré à la valence
		End field : finale, régie, à droite
		Post field : finale, non-régie, à droite
		Non interprétable
		Non applicable
	3. Mobilité syntagmatique	Non déplaçable
		Mobile
		Non interprétable
		Non applicable
SEMANTICO-PRAGMATIQUE	4. Activation du sens codé en contexte	Oui (sens codé seul)
		Non (éloignement du sens codé)
	5. Procéduralité	Sens conceptuel
		Sens conceptuel-procédural
		Sens procédural
		Non interprétable
PROSODIE	6. Présence d'une pause contiguë	Pause à gauche
		Pause à droite
		Pause à gauche et à droite
		Pas de pause contiguë
		Non interprétable
ENONCIA-TION	7. Position dans le tour de parole	Initial
		Final
		Autonome
		Autre
		Non interprétable
NON-LING.	8. Degré de certitude	1 : non, (quasi-)certitude que ce n'est pas un MD
		2 : statut incertain, hésitation entre un statut de MD ou non
		3 : oui, (quasi-)certitude que c'est un MD

Tableau 1. Annotation paramétrique : variables et valeurs attribuées aux MD potentiels

À noter que d'autres paramètres, susceptibles de jouer un rôle dans la définition des MD, ont été pris en considération dans le projet MDMA et ont également fait l'objet d'une annotation systématique. Si nous ne tenons pas compte de ces paramètres ici, c'est que les annotations n'étaient pas suffisamment quantifiables ou réplicables pour garantir une analyse statistique valide. Ces variables non retenues dans le cadre de la présente étude sont : (i) l'identification et la délimitation de la portée sémantique du MD potentiel ; (ii) l'identification du type de relation instaurée entre segments mis en présence (critère applicable exclusivement aux MD de type « connecteurs », en termes de type de contenu et de taille des segments concernés) ; (iii) l'optionalité syntaxique du MD potentiel ; (iv) la collocabilité ou cooccurrence avec d'autres MD prototypiques (critère non retenu pour l'analyse statistique, vu la circularité du raisonnement qui présuppose que les MD « confirmés » aient été préalablement identifiés comme tels dans l'environnement linguistique du MD « potentiel »).

4. Résultats : vers des critères définitoires fondés sur corpus

4.1. Réplicabilité du protocole d'annotation

Dans le domaine de la linguistique de corpus, ainsi qu'en linguistique computationnelle, les mesures quantitatives d'accord inter-juges, notamment l'indice Kappa de Fleiss (1971), sont régulièrement utilisées comme critère principal pour l'évaluation d'annotations. Toutefois, le choix de l'un ou l'autre score – α de Krippendorff (1980), π de Scott (1955), κ de Fleiss (1971) ou de Light (1971) – dépend de plusieurs facteurs (avec ou sans factorisation de la chance, nombre d'annotateurs, équivalence ou non des catégories annotées) et implique certaines précautions d'interprétation. Au vu de la complexité des facteurs et des limites de certains scores, nous proposons ici d'utiliser une combinaison de mesures quantitatives (d'après les recommandations de Spooren et Degand, 2010) et qualitatives pour évaluer la réplicabilité du protocole.

Si les pourcentages seuls ne suffisent pas à mesurer l'accord inter-juges, ils complètent néanmoins les résultats des protocoles « corrigés », c'est-à-dire qui prennent en compte un facteur chance. Nous avons donc calculé différentes mesures pour chaque variable, à l'aide de l'outil conçu par Geertzen (2012). Dans l'ensemble, les scores Kappa sont intermédiaires (moyenne de 0,555 sur les huit variables), tandis que les accords observés (simples pourcentages) sont plutôt bons (72,45% en moyenne). Il n'est pas surprenant d'obtenir de tels scores si on prend en considération la complexité de l'objet d'étude et le panel de variables concernées, certaines impliquant, comme on l'a vu, une part relative d'interprétation. C'est la catégorie syntaxique qui apparaît – sans surprise – comme la plus réplicable (0,766 quel que soit l'indice utilisé, 81% d'accord), tandis que le degré de certitude montre le plus de désaccord (α = 0,361, 59,3%), ce qui reflète les divergences entre annotateurs quant aux représentations qu'ils ont *a priori* de la catégorie de MD.

Pour ce qui concerne plus particulièrement les objectifs de la présente étude, on constate que la position syntaxique affiche un score Kappa (par paires) raisonnable (κ = 0,637, 71,7%), d'autant plus qu'il est affecté négativement par des catégories rares, comme les valeurs « non interprétable » et « non applicable ». Bien qu'il reste sous le seuil généralement admis de 0,8, en considérant la part d'interprétation sémantique impliquée dans le processus d'annotation de cette variable, un tel résultat

est encourageant et permet de consolider les résultats discutés plus bas concernant ce paramètre syntaxique. Quant aux autres variables, les différences de scores s'expliquent par différentes raisons : subjectivité de l'annotateur, présence de catégories rares, difficultés liées au format de transcription utilisé, etc. En résumé, la réplicabilité du protocole d'annotation n'est pas parfaite et pourrait sans doute bénéficier de quelques autres phases d'entraînement, mais reste certainement correcte par rapport à la complexité du phénomène étudié.

Pour compléter cette évaluation quantitative, nous tenons à ajouter que les précautions méthodologiques discutées plus haut (paramétrisation des définitions, amélioration des critères par phases successives d'annotation, discussion des désaccords) fournissent à elles seules une forme de validation qualitative du protocole final. L'objectif n'est en effet pas d'aboutir à un protocole qui serait réplicable sans entraînement préalable ni prérequis théorique dans le domaine. Il s'agit plutôt d'élaborer un modèle cohérent qui rende compte de la diversité des comportements et de la complexité du phénomène étudié.

Dans la suite de l'article, ce sont les annotations finales, c'est-à-dire les valeurs attribuées lors de la dernière phase d'entraînement du protocole (obtenues soit par accord parfait, par règle de majorité ou par discussion) qui feront l'objet des analyses. On montrera que certains résultats contribuent à leur tour à asseoir la validité du protocole proposé dans son ensemble.

4.2. Degré de prototypicalité des marqueurs potentiels : analyse distributionnelle

L'analyse des annotations sur l'échantillon aléatoire de 200 MD potentiels fournit des résultats intéressants quant à la pertinence des paramètres prévus par le protocole. La distribution de certaines valeurs au sein de l'échantillon, tous marqueurs confondus, est d'abord révélée par des statistiques descriptives. On constate notamment une étonnante répartition des degrés de certitude pour les MD potentiels étudiés. En effet, la proportion de MD affichant un degré élevé de certitude (2,75 ou 3) est minoritaire (environ 1 cas sur 5), par rapport à une proportion plus élevée de MD potentiels qui ne sont *a priori* pas reconnus comme MD en contexte (environ 1 cas sur 3 ayant un score de 1 ou 1,25). Ce sont en fait les cas dont le statut est incertain ou non consensuel, qui sont les plus nombreux (environ 1 cas sur 2 ayant un score moyen 1,5 à 2,5) (*cf.* Tableau 2).

Degré de certitude	**MD potentiels**
1 – 1,25	34 %
1,5 – 2,5	46,5 %
2,75 – 3	19,5 %

Tableau 2. Distribution des marqueurs discursifs potentiels en fonction de leur degré de certitude[1]

Autrement dit, peu de MD potentiels ont été reconnus de manière consensuelle comme étant *a priori* des MD par les annotateurs, lors de la phase d'identification préliminaire à l'analyse paramétrique. L'échantillonnage aléatoire des 200 occurrences semble au contraire rassembler une majorité de cas limite correspondant soit à des expressions qui ne sont pas considérées comme des MD, mais qui ont une

[1] Plus le score est élevé, plus le MD potentiel est considéré *a priori* comme un MD.

fonction pragmatique dans le contexte étudié (*oui* dans le Tableau 3), soit à des expressions qui font l'objet de désaccords entre annotateurs quant à leur statut de MD (*là-dessus* dans le Tableau 3). Un aperçu des différents cas limite ou écartés, reprenant les degrés de certitude des différents annotateurs, est repris dans le Tableau 3.

Degrés de certitude				**MD potentiel en contexte**
3	3	2	1 (moyenne 2,25)	ben oui enfin bon **disons que** je (bruits de voix) / je ferai un peu ce qui ce qui se présentera euh ce que je trouverai comme boulot quoi (VALIBEL : corpus ilrCP1r)
2	3	1	1 (moyenne 1,75)	STE hm hm d'accord= DIR okay/ STE [d'accord] DIR [voilà] (0.2) et pour votre loyer/ STE hm (0.3) **oui** je pensais que: on fait avec ça pa'ce que maint'nant j'ai pas les-d'argent liquide (CLAPI : corpus Bielefeld_Caution1)
1	1	1	1 (moyenne 1)	CG c'était pas une priorité pour vous c'est pas une priorité de revenir **là-dessus** CEC voilà c'est-à-dire que faut le faire faut pas le laisser de côté (CLAPI : corpus Adi étudiants)
2	1	1	1 (moyenne 1,25)	bon ben je ce qui est assez intéressant maintenant c'est qu'on peut euh apporter euh / sa façon de voir les choses euh sa sa euh **il y a** une certaine créativité donc dans la façon d'organiser de de créer euh (VALIBEL : corpus ilrCP1r)

Tableau 3 : Illustration de l'évaluation du statut du MD potentiel (avant l'analyse paramétrique)

En outre, il ressort du graphique ci-dessous (Figure 1) que les catégories syntaxiques les plus représentées dans l'échantillon sont des syntagmes adverbiaux (ADV-P ; *alors, voilà, puis, là, oui, non*), des conjonctions (CONJ ; *donc, et, mais, parce que, si*) et des particules (PART ; *ben, hein, euh, mm*).

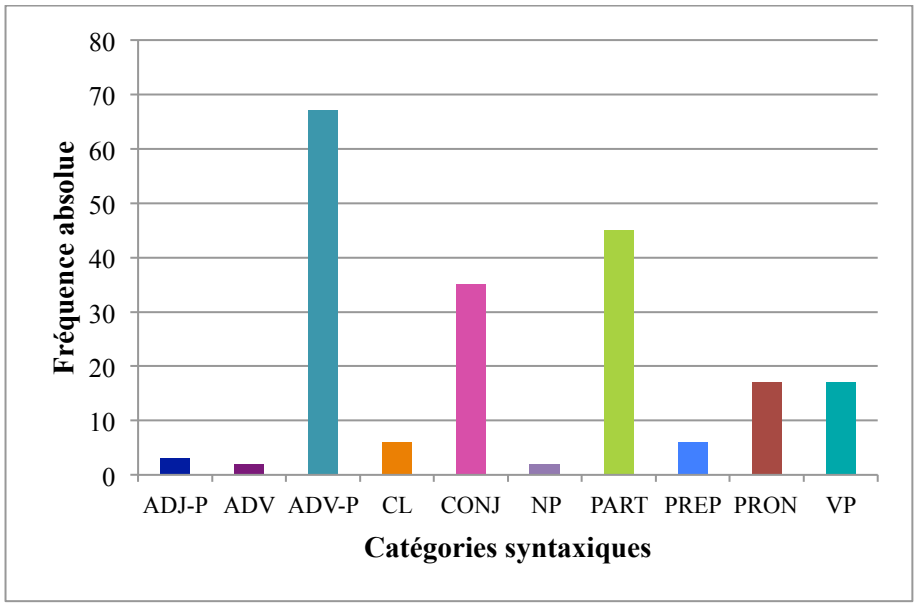

Figure 1. Distribution de la catégorie syntaxique des marqueurs discursifs potentiels

Leurs fréquences absolues élevées, largement supérieures à celles des autres valeurs et statistiquement significatives (χ^2 = 220,3, dl = 9, p < 0,01), témoignent du statut privilégié qu'entretiennent ces classes grammaticales avec la catégorie des MD : au-delà de leur fréquence élevée, il semblerait aussi que ces trois catégories grammaticales tendent à être identifiées en tant que MD. En effet, parmi les MD potentiels à haut degré de certitude, 87% (34/39) appartiennent à l'une de ces trois catégories. La proportion est similaire pour les cas incertains (degré de certitude intermédiaire), soit 88% (82/93), alors qu'elle est plus faible (48%, soit 33/68) pour les MD potentiels à faible degré de certitude. On peut en conclure qu'adverbes, conjonctions et particules sont si typiques de la catégorie de MD qu'ils élargissent l'identification à des expressions dont le statut discursif reste en suspens.

Une observation similaire concerne le degré de procéduralité des MD potentiels. On observe la prédominance d'expressions procédurales (PR) dans 56,5% des cas, par rapport aux valeurs conceptuelles (C) (13,5%) et conceptuelles-procédurales (C-PR) (28,5%)[2] (pour l'aspect théorique des notions « procédural » et « conceptuel », *cf.* Wilson, 2011 ; pour une application de ces notions aux MD, *cf.* Bolly et Degand, 2013). Il semblerait ainsi que le sens procédural soit un facteur décisif dans l'identification des expressions linguistiques susceptibles de remplir une fonction de MD, dès la phase préalable à l'échantillonnage des données (*cf.* 3.2.1). Cela induit donc une certaine circularité dans l'analyse, puisque la variable *procéduralité*, bien plus qu'un simple critère correspondant à un composant ou un niveau de l'analyse linguistique identifiable (syntaxique, sémantique, prosodique, etc.), rend compte de la fonction proprement dite du marqueur dans son contexte d'énonciation. Autrement dit, si nous considérons qu'un MD se définit avant tout par sa fonction pragmatique, indexicale et instructionnelle (*cf.* la définition de Hansen, 2006, sous 2.), nous le définissons implicitement comme étant par essence procédural. Autrement dit encore, attribuer à un MD potentiel un sens procédural, c'est déjà lui reconnaître son statut de MD. Par conséquent, afin de garder une pondération équilibrée entre paramètres d'analyse sans introduire de biais au niveau de la procéduralité, nous exclurons ce paramètre de certaines analyses. En résumé, bien que le trait procédural soit décisif pour l'identification des MD, il n'est que peu pertinent dans un modèle prédictif comme le nôtre, puisqu'il ne fait que refléter les choix théoriques préalables à l'annotation en désignant une propriété intrinsèque à la fonction pragmatique des MD.

D'après le degré de certitude attribué aux occurrences des entités identifiées comme étant potentiellement des MD, un portrait prototypique des MD peut déjà être dessiné. Ainsi, les items ayant obtenu un degré de certitude élevé quant à leur statut de MD sont le plus souvent des conjonctions, des particules ou des syntagmes adverbiaux, en position initiale non-régie, qui ne réalisent pas leur sens codé et qui ont un sens procédural. À l'inverse de l'exemple [20], qui remplit toutes ces conditions, les items pour lesquels un relatif consensus a été obtenu en faveur de leur exclusion de la catégorie des MD (*i.e.* ayant un degré de certitude faible) correspondent le plus souvent à des pronoms ou des syntagmes verbaux, en position médiale ou finale régie, affichant leur sémantisme de base et un sens conceptuel-procédural, comme dans l'exemple [21].

[20] et alors c'est très compliqué de pouvoir régler ces problèmes **donc** c'est même je dirais c'est presque plus du français (VALIBEL : corpus ilcDA1r)

[2] Les 1,5% des cas restant sont indéterminés en termes de procéduralité.

[21] prendre tout le temps pour ça alors que **je crois qu'**en France actuellement on a d'autres problèmes (CLAPI : corpus Adi étudiants)

Il convient toutefois de préciser que ces remarques ne sont valables que pour l'échantillon concerné, puisque toute généralisation nécessiterait un traitement quantitatif plus approfondi.

4.3. Profils multifactoriels des marqueurs potentiels

Afin de valider statistiquement les traits prototypiques mis en exergue lors de notre première analyse, nous avons eu recours à des méthodes statistiques multifactorielles. Ces méthodes permettent de faire émerger des profils de MD potentiels en fonction du degré de certitude. Au-delà des traits prototypiques mis en avant par l'examen des fréquences, l'approche multifactorielle, qui consiste en une analyse des correspondances multiples (ACM) dans notre cas, offre une visualisation graphique du degré d'association entre les différentes variables. Autrement dit, une ACM s'interprète en termes de probabilité de cooccurrence entre variables, d'après leur proximité sur le graphique.

Nous avons appliqué une ACM séparément aux trois groupes de marqueurs potentiels, à savoir les groupes formés par les trois niveaux de certitude, faible (1 – 1,25), intermédiaire (1,5 – 2,5) et élevé (2,75 – 3), comme nous l'avons fait pour l'analyse distributionnelle. Cette ACM révèle des niveaux d'association entre variables qui diffèrent selon le profil plus ou moins discursif accordé (indépendamment des autres traits susceptibles de les caractériser) aux MD potentiels en contexte.

Pour le groupe des marqueurs reconnus comme ayant *a priori* un haut potentiel discursif (*i.e.* montrant un degré de certitude élevé) (Figure 2), hormis quelques cas résiduels qui correspondent pour la plupart à des valeurs sous-déterminées, on observe très clairement que la grande majorité des valeurs prévues par le protocole d'annotation et assignées à ces items sont très proches les unes des autres, au croisement des deux dimensions (ou axes). Ce graphique montre donc une forte association entre toutes les différentes caractéristiques possibles pour les marqueurs de ce groupe. On peut en conclure une certaine homogénéité de profils pour les MD confirmés, qui suggère un nombre très limité de types différents ou, à tout le moins, des types bien définis se déclinant en variables covariant ensemble.

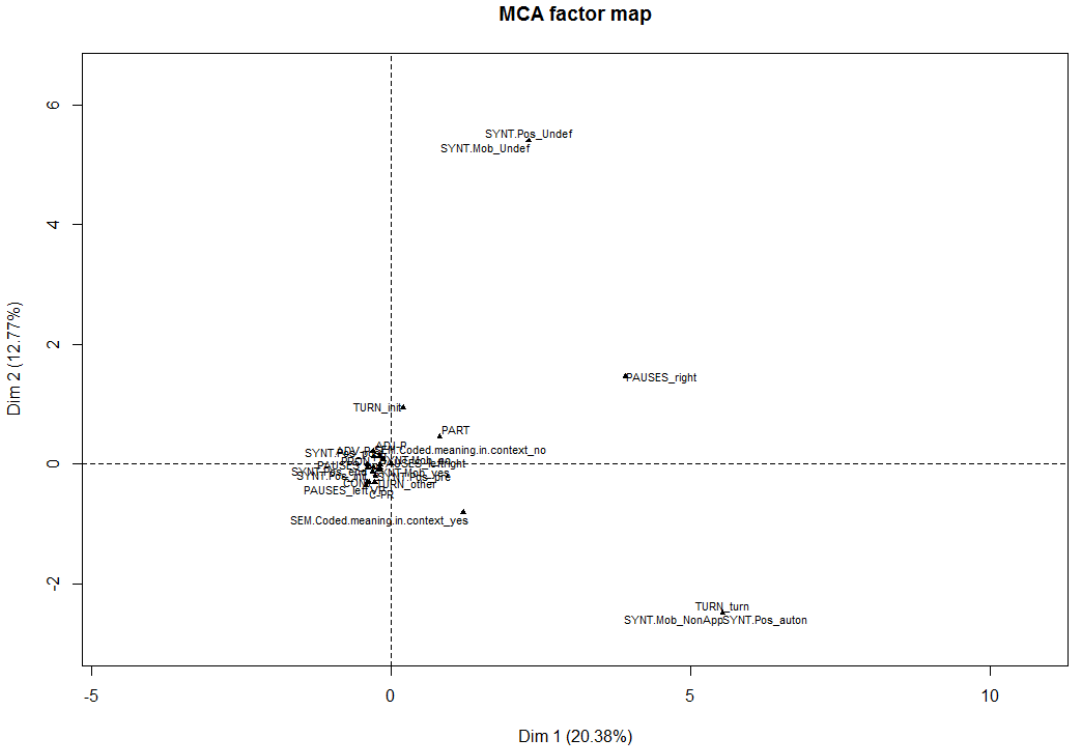

Figure 2. Analyse des correspondances multiples des MD à degré de certitude élevé

En revanche, contrairement à un amas concentré de points convergeant vers le centre des dimensions, le graphique représentant le groupe intermédiaire des MD potentiels (Figure 3) est quant à lui éclaté. S'il est possible de repérer des groupes de traits associés, il est difficile de les interpréter malgré la connaissance des données. De plus, le total de la variation pris en compte par les deux dimensions de ce graphique est inférieur (18,71%) à celui du graphique précédent (34,15%), ce qui indique que la variation interne aux données du groupe intermédiaire ne peut être expliquée ni visualisée en seulement deux dimensions. L'hétérogénéité de ce groupe est telle qu'il est impossible, ou non pertinent, d'identifier un profil clair de MD intermédiaire. On est plutôt face à une diversité d'expressions linguistiques qui frôlent les limites de la catégorie de MD.

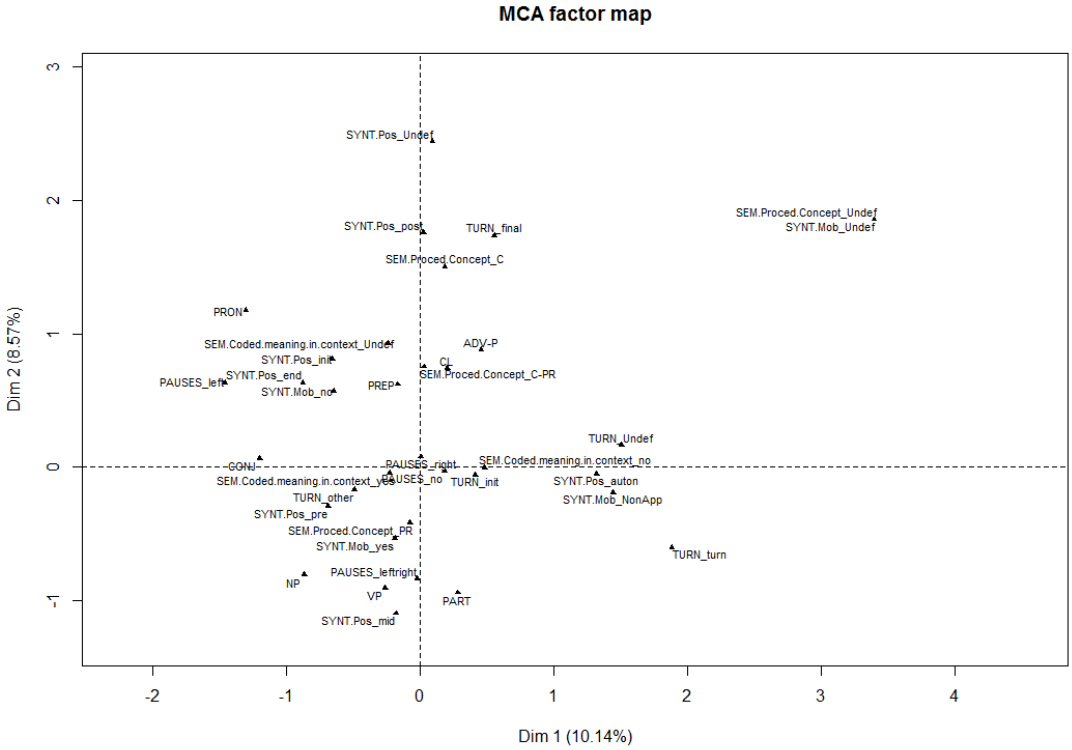

Figure 3. Analyse des correspondances multiples des MD à degré de certitude intermédiaire

Cette première approche statistique et multifactorielle des MD potentiels composant notre échantillon permet déjà d'affirmer que les variables prévues par notre protocole d'annotation sont suffisamment pertinentes pour faire émerger des profils. Ainsi, l'étape intuitive de la sélection d'une part, et du codage en degré de certitude d'autre part, est en quelque sorte validée statistiquement en ce que les différents scores, et les valeurs qui y sont associées, reflètent des questions théoriques de définition de la catégorie. En effet, si les membres prototypiques de la catégorie des MD sont suffisamment identifiables et partagés entre annotateurs, il apparaît que les cas limite sont beaucoup plus hétérogènes. En résumé, outre l'émergence de profils statistiques des MD potentiels, les analyses de cette section offrent un retour empirique sur le travail théorique préliminaire au projet MDMA, c'est-à-dire qu'elles font émerger, à partir des données annotées elles-mêmes, les outils et critères (syntaxiques, sémantico-pragmatiques, prosodiques et contextuels) nécessaires à la définition d'une catégorie pragmatique.

4.4. Hiérarchisation statistique des paramètres

Afin de faire émerger, grâce à diverses méthodes de visualisation et de hiérarchisation, les (combinaisons de) variables qui apparaissent comme étant les plus fiables pour identifier un MD, des analyses statistiques ont été appliquées aux données. Les tests statistiques ont fourni des informations diverses mais complémentaires sur la fiabilité des caractéristiques syntaxiques, sémantiques et contextuelles, quant à leur capacité à prédire le degré de certitude attribué à chaque item. C'est cet indice – correspondant à la moyenne des scores que les quatre annotateurs ont attribués à chaque MD potentiel (pour rappel) – qui sert de variable de

contrôle par rapport aux autres variables, de la même manière qu'un annotateur humain évaluerait l'exactitude d'un système automatique.

Concernant la question du poids respectif des différentes variables par rapport à leur pouvoir de prédiction sur le statut discursif des marqueurs étudiés, une analyse de leur importance conditionnelle révèle la hiérarchie suivante : la position syntaxique apparaît de loin comme la variable la plus prédictive, suivie par la réalisation ou non du sens codé de l'expression, puis par la catégorie syntaxique. À l'inverse, la mobilité syntagmatique, la position dans le tour de parole et la contiguïté avec une pause ne semblent pas être déterminantes dans l'attribution du statut de MD. Ce dernier résultat concernant la fiabilité des pauses comme indices à la présence d'un MD est surprenant et pourrait remettre en question les nombreuses définitions de la catégorie qui font de ce paramètre, et d'autres critères prosodiques, une condition *sine qua non* (*cf.* Vincent, 1993 ; Maschler, 2002). Il pose également la question de la pertinence de cette variable prosodique dans les systèmes d'identification automatique : si une désambiguïsation locale sur des marqueurs spécifiques semble bénéficier d'une caractérisation acoustique du contexte (voir l'exemple de *like* dans Zufferey et Popescu-Belis, 2004), une fois élargie à un échantillon plus hétérogène qui n'aborde les phénomènes de pauses qu'à partir de leur codage perceptif lors de la transcription, sa fiabilité semble amoindrie. Notre suggestion, sur la base de cette première analyse inférentielle, est donc de considérer la position et la catégorie syntaxique (*cf.* Heeman *et al.*, 1998) comme indices de surface à forte valeur prédictive dans une perspective d'annotation (semi-)automatique.

Au sein de ces variables prédictives, certaines valeurs en particulier sont davantage associées à un score élevé de certitude que d'autres. Pour les identifier, un arbre de régression calculé sur l'ensemble des paramètres annotés fournit une visualisation hiérarchisée des faisceaux de valeurs et de leur score de certitude correspondant. Comme on peut le voir sur la Figure 4, des ensembles ou faisceaux de traits spécifiques sont associés à des intervalles de degré de certitude plus ou moins larges et plus ou moins élevés, sur l'échelle de 1 à 3 mentionnée précédemment.

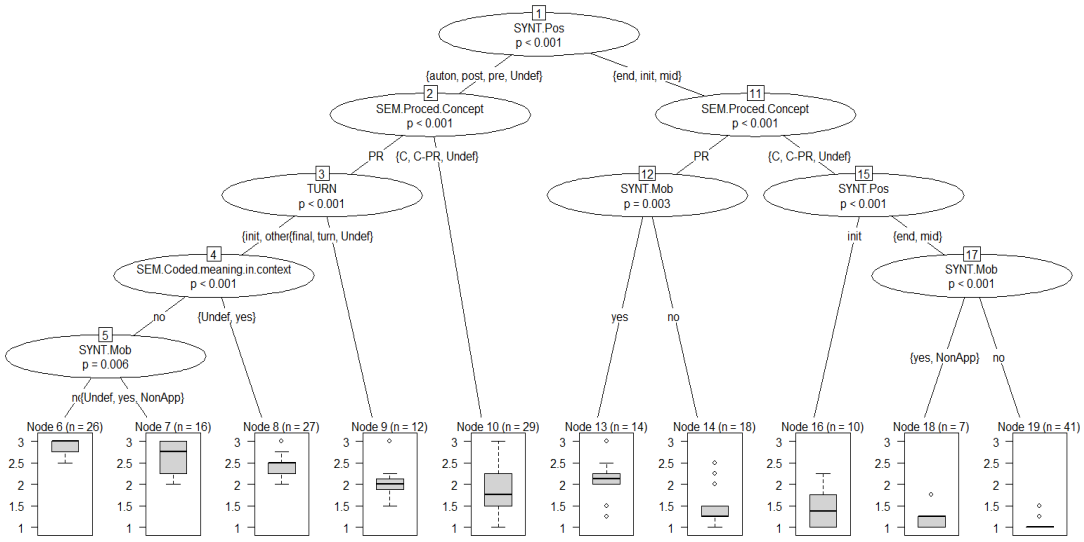

Figure 4. Arbre de régression conditionnelle des paramètres prédicteurs

Par exemple, l'intervalle le plus élevé de certitude (*node* 6, le premier en partant de la gauche sur le graphique) correspond aux items (voir [22] et [23]) ayant reçu l'ensemble des annotations suivantes (cf. Tableau 1 *supra*) :

- position autonome, finale non-régie, initiale non-régie ou non interprétable,
- sens procédural,
- position initiale ou médiane dans le tour de parole,
- éloignement du sens codé, et
- non déplaçable.

[22] bon moi heureusement je n'ai pas eu de choses très graves mais **enfin** quand même / on sent / on commence à se sentir qu'on / qu'on / qu'on diminue / qu'on ne sait pas faire ce qu'on veut (Corpage : corpus ageNM1)

[23] [oui/ c'est un bien propre mais y'a eu des rembours'ments pendant la communauté H exact NH (inaudible) des (deniers) d'communauté hein H d'accord NF **voilà** j'ai vérifié (.) la cour d'cassation depuis le:- dans un arrêt du 31 mars 92 (CLAPI : corpus Naldo)

À l'autre extrême, un item non-MD d'après son degré de certitude correspond aux valeurs régies, intégrées de la position syntaxique (initiale, médiane ou finale) et un sens au moins partiellement conceptuel (*node* 19, le dernier à droite du graphique) (voir exemple [24]) .

[24] alors euh ces immigrés viennent et ils ont souvent des problèmes **un peu** complexes (VALIBEL : corpus ilcDA1r)

On constate également que certaines oppositions de valeurs entraînent une augmentation du degré de certitude plus nette que d'autres. Ainsi, pour ce faisceau de variables au score minimal (*node* 19), le trait de mobilité est statistiquement significatif, mais les intervalles de certitude sont proches et se chevauchent. En revanche, lorsque le degré de certitude est intermédiaire, c'est-à-dire à un niveau hiérarchique plus élevé (*nodes* 13 et 14), le même trait de mobilité syntaxique entraîne une amélioration considérable du score de certitude selon sa valeur (en cas de mobilité syntaxique – codée *yes*, le degré de certitude augmente [25] alors qu'il diminue sur la branche indiquant l'absence de mobilité syntaxique – codée *no* [26]).

[25] ben oui enfin bon **disons que** je (bruits de voix) / je ferai un peu ce qui ce qui se présentera euh ce que je trouverai comme boulot quoi (VALIBEL : corpus ilrCP1r)

[26] y a des différences entre les pays justement/ [y a des différences aussi& SEV [voilà c'est ça CG &**comme** dit sophie quand un des parents est français comme c'est son cas (CLAPI : corpus Adi étudiants)

Ce type d'analyses statistiques permet donc de faire émerger visuellement les valeurs de variables prédictives dans le processus de désambiguïsation des MD potentiels, en présentant aux nœuds les plus élevés du graphique les traits les plus déterminants. Ce graphique confirme notamment la primauté de la position syntaxique, qui sépare d'une part les positions non régies (autonome, non intégrée ou

non interprétable) avec un degré de certitude intermédiaire et élevé, des positions régies d'autre part, c'est-à-dire intégrées dans la syntaxe de dépendance du verbe (Blanche-Benveniste *et al.*, 1990 ; Blanche-Benveniste, 2003), avec un degré de certitude faible. Ces résultats témoignent de l'ancrage principalement macro-syntaxique des MD, qui s'explique par leur portée énonciative plus large, au niveau du discours et non au niveau micro-syntaxique plus restreint de la proposition. Dans son acception cognitiviste-constructiviste, la macro-syntaxe permet en effet d'établir le lien entre le système linguistique, la situation de production langagière et l'ensemble des savoirs individuels et partagés des interlocuteurs (*cf.* Berrendonner 2004).

Ce modèle graphique de classification des traits de variables en fonction de leur importance conditionnelle fournit une validation quantitative supplémentaire du protocole d'annotation proposé dans cette étude : d'une part, l'indice de significativité statistique est relativement élevé (pseudo-R^2 = 0,78) compte tenu du caractère exploratoire de l'approche ; d'autre part, le modèle fait émerger des profils de MD potentiels représentés par des faisceaux de paramètres qui corroborent l'attribution intuitive du degré de certitude quant au statut plus ou moins discursif de l'item en contexte. On peut donc en conclure que le protocole présenté est validé par les données et leur traitement statistique, ce qui rapproche notre démarche des études menées en linguistique cognitive sur corpus et, plus particulièrement, des études relevant du champ de la *corpus-driven cognitive semantics* (Glynn, 2010) où les données de corpus sont utilisées « to produce testable and falsifiable results for the description of meaning » (Glynn, 2010 : 1). Dans ce cadre méthodologique, approche empirique et traitement statistique (notamment multifactoriel) se combinent pour établir des catégories linguistiques fiables et valables sur les plans méthodologique et cognitif.

Conclusion

Partant du constat que la catégorie des MD ne possède pas, à ce jour, de définition satisfaisante qui fournisse les critères nécessaires à leur identification – manuelle ou automatique – de manière stable et exhaustive, nous développons le modèle MDMA proposant une approche fondée sur corpus (*corpus-based*) qui allie un effort théorique de paramétrisation des variables définitoires et une étude empirique sur un corpus de français oral (semi-)spontané. Dans cet article, nous avons présenté la méthode et les résultats d'une expérience d'annotation sur corpus par quatre juges visant à faire émerger, par le biais d'analyses statistiques (arbre de régression et analyses multifactorielles), les critères les plus fiables pour la désambiguïsation de MD potentiels en contexte. L'originalité de la démarche réside principalement dans la flexibilité du protocole d'annotation, qui a été élaboré dans le but d'être appliqué à des MD potentiels, c'est-à-dire des expressions linguistiques très différentes regroupant à la fois des MD consensuels (*tu vois* MD), leurs formes équivalentes dans des usages propositionnels ou non-discursifs (*tu vois* VP), ainsi que des items « candidats » qui peuvent être considérés ou non comme MD selon le cadre épistémologique adopté (*etcetera, euh*).

Nous avons présenté diverses manières, tant qualitatives que quantitatives, de confirmer la validité et la réplicabilité de ce protocole d'annotation. À l'aide de diverses méthodes statistiques de visualisation des données, ce protocole nous a permis d'identifier les faisceaux de traits qui, en combinaison, correspondent à des niveaux de certitude différents. La position syntaxique – qui distingue les positions

régies ou intégrées, d'une part, et non-régies ou macro-syntaxiques, d'autre part – apparaît comme la variable prédictive principale du statut de MD. Le codage de ce paramètre, bien qu'indice « de surface », reste néanmoins un défi pour l'annotation, en raison de l'inévitable part d'interprétation sémantique qu'implique son analyse, malgré son opérationnalisation et la systématicité visée.

Par ailleurs, l'analyse distributionnelle et surtout inférentielle (analyses des correspondances multiples) a permis d'associer aux différents degrés de certitude des faisceaux de traits linguistiques plus ou moins caractéristiques. Ainsi, les MD « confirmés » (c'est-à-dire à haut score de certitude) forment une classe homogène avec un panel de profils restreints et clairement identifiables alors que les cas limite (certitude intermédiaire) ne peuvent être réduits à quelques étiquettes précises, mais couvrent plutôt des expressions très diverses. Cette hétérogénéité des cas limite est une piste pour explorer davantage les membres de ce groupe intermédiaire : nous y avons identifié des éléments appartenant à d'autres catégories pragmatiques (expressions modales, particules d'hésitation, signaux de réponse) qui peuvent aussi être des MD (Bolly *et al.*, 2014).

Enfin, d'autres perspectives sont envisagées pour poursuivre l'approche présentée ici : on pourra tester l'applicabilité du modèle MDMA à d'autres types de données (genres, corpus, langues, modalités) afin de tendre vers une validité plus générale du protocole, ainsi que pour vérifier dans quelle mesure les résultats eux-mêmes (par exemple, la primauté de la position syntaxique) dépendent de facteurs situationnels et/ou linguistiques. Pour finir, un projet en cours (voir Bolly et Crible, 2015) s'attèle à l'annotation fonctionnelle et multimodale des MD confirmés (ainsi que certains cas limite, dans une acception plus large de la catégorie) afin d'examiner l'influence réciproque des variables formelles et des paramètres sémantico-pragmatiques, faisant ainsi le pont entre sélection, contexte et interprétation.

Remerciements

Les auteures tiennent à remercier très sincèrement Dr. Natalia Levshina (Université catholique de Louvain) pour son aide experte apportée lors du traitement statistique des données. Une partie du travail de réflexion et d'annotation a été menée en collaboration avec Federica Ciabarri (Université catholique de Louvain jusque septembre 2013). Les recherches de Catherine T. Bolly bénéficient d'une bourse Marie Curie [PIEF-GA-2012-328282] accordée par l'Union Européenne dans le cadre du Seventh Framework Programme ([FP7/2007-2013]). Ludivine Crible et Liesbeth Degand bénéficient du soutien financier de l'Action de Recherche concertée n°12-17-044 de la Fédération Wallonie-Bruxelles. Les recherches de Deniz Uygur-Distexhe ont été financées par une bourse interne à l'Université catholique de Louvain (FSR). Les trois premières auteures (Bolly, Crible et Degand) sont membres de l'Action COST ISCH 1312 TextLink.

Références bibliographiques

AIJMER, K. 1997. *I think* – an English modal particle. In T. SWAN, O. WESTVIK (eds), *Modality in Germanic languages. Historical and comparative perspectives*. Berlin : Mouton de Gruyter, 1-47.

AIJMER, K. 2013. *Understanding pragmatic markers. A variational pragmatic approach*. Edinburgh : Edinburgh University Press.

AIJMER, K., SIMON-VANDENBERGEN, A.-M. 2011. Pragmatic markers. In J. ZIENKOWSKI, J.-O. ÖSTMAN, J. VERSCHUEREN (eds), *Discursive pragmatics*. Amsterdam : John Benjamins, 223-247.

BALTHASAR, L., BERT, M. 2005. La base de données « Corpus de langues parlées en interaction » (CLAPI) : Genèse, état des lieux et perspectives. In M. SAVELLI (ed.), *Corpus oraux et diversité des approches*. *Lidil* 31.

BARNES, B. K. 1995. Discourse particles in French conversation : (*eh*) *ben*, *bon*, and *enfin*. *The French Review* 68(5) : 813-21.

BERRENDONNER, A. 2004. Grammaire de l'écrit vs grammaire de l'oral : le jeu des composantes micro- et macro-syntaxiques. In A. RABATEL (éd.), *Interactions orales en contexte didactique*. Lyon : Presses Universitaires de Lyon, 249-264.

BIBER, D., CONRAD, S., REPPEN, R. 1994. Corpus-based approaches to issues in applied linguistics. *Applied Linguistics* 15(2) : 169-89.

BLANCHE-BENVENISTE, C., BILGER, M., ROUGET, C., VAN DEN EYNDE, K. 1990. *Le français parlé. Etudes grammaticales*. Paris : CNRS.

BLANCHE-BENVENISTE, C. 2003. Le recouvrement de la syntaxe et de la macro-syntaxe. In A. SCARANO (ed.), *Macro-syntaxe et pragmatique. L'analyse linguistique de l'oral*. Rome : Bulzoni, 53-75.

BOLLY, C. T. 2010. Pragmaticalisation du marqueur discursif 'tu vois'. De la perception à l'évidence et de l'évidence au discours. In F. NEVEU, J. DURAND, T. KLINGLER, S. PREVOST, V. MUNI-TOKE (eds), Actes du *CMLF 2010 (2e Congrès Mondial de Linguistique Française*, 12-15 juillet 2010, New Orleans, United States of America), <http://dx.doi.org/10.1051/cmlf/2010243>.

BOLLY, C. T. 2012a. Du verbe de perception visuelle au marqueur parenthétique 'tu vois': Grammaticalisation et changement linguistique. *Journal of French Language Studies* 22(2) : 143-164.

BOLLY, C. T. 2012b. Flou phraséologique, quasi-grammaticalisation et pseudo marqueurs de discours. Un no man's land entre syntaxe et discours? In V. Conti, G. Corminbœuf, L. A. Jonhsen (eds), *Entre syntaxe et discours. Eclairages épistémologiques et descriptions linguistiques* (*LINX* 62-63, 2010) : 11-39.

BOLLY, C. T. 2014. Gradience and gradualness of parentheticals. Drawing a line in the sand between phraseology and grammaticalization. *Yearbook of Phraseology* 5(1) : 25-56.

BOLLY, C. T., CRIBLE, L. 2015. From context to functions and back again : Disambiguating pragmatic uses of discourse markers. Communication présentée au panel *Anchoring utterances in co(n)text, argumentation, common ground*, *14th International Pragmatics Conference (IPra)*, 26-31 Juillet 2015, Antwerp, Belgique.

BOLLY, C. T., DEGAND, L. 2009. Quelle(s) fonction(s) pour 'donc' en français oral? Du connecteur conséquentiel au marqueur de structuration du discours. *Lingvisticae Investigationes* 32(1) : 1-32.

BOLLY, C. T., DEGAND, L. 2013. Have you seen what I mean? From verbal constructions to discourse structuring markers. *Journal of Historical Pragmatic* 14(2) : 210-235.

BOLLY, C. T., MASSE, M., MEIRE, P. 2012. *Corpage. Reference corpus for the elderly's language*. Louvain-la-Neuve : Université catholique de Louvain.

BOLLY, C. T., CRIBLE, L. DEGAND, L., UYGUR-DISTEXHE, D. 2014. Towards a Model for Discourse Marker Annotation in spoken French : From potential to feature-based discourse markers. Communication présentée au workshop international *Pragmatic Markers, Discourse Markers and Modal Particles : What do we know and where do we go from here?*, 16-17 octobre 2014, Como, Italie.

BRINTON, L. 1996. *Pragmatic markers in English. Grammaticalization and discourse functions*. New York : Mouton de Gruyter.

BRINTON, L. 2008. *The comment clause in English*. Cambridge : CUP.

CIABARRI, F. 2013. Italian Reformulation Markers : A study on spoken and written language. In C. T. BOLLY, L. DEGAND (eds), *Across the line of speech and writing variation* (Corpora and Language in Use – Proceedings 2). Louvain-la-Neuve : Presses universitaires de Louvain, 113-128.

CRIBLE, L. Soumis. Towards an operational category of discourse markers : A definition and its annotation model.

CUENCA, M. J. 2008. Pragmatic markers in contrast : The case of *well*. *Journal of Pragmatics* 40 : 1373-1391.

DEFOUR, T., D'HONDT, U., VANDENBERGEN, A.-M., WILLEMS, D. 2010. *In fact*, *en fait*, *de fait*, *au fait* : A contrastive study of the synchronic correspondences and diachronic development of English and French cognates. *Neuphilologische Mitteilungen* 111(4) : 433-463.

DEGAND, L., FAGARD, B. 2011. *Alors* between Discourse and Grammar. The role of syntactic position. *Functions of Language* 18(1) : 29-56.

DEGAND, L., CORNILLIE, B., PIETRANDREA, P. 2013. Discourse markers and modal particles : Two sides of the same coin? In L. DEGAND, B. CORNILLIE, P. PIETRANDREA (eds), *Discourse markers and modal particles. Categorization and description*. Amsterdam : John Benjamins, 1-18.

DISTER, A., FRANCARD, M., HAMBYE, P., SIMON, A.-C. 2009. Du corpus à la banque de données. Du son, des textes et des métadonnées. L'évolution de banque de données textuelles orales VALIBEL (1989-2009). *Cahiers de Linguistique* 33(2) : 113-129.

DOSTIE, G. 2007. La réduplication pragmatique des marqueurs discursifs. De *là* à *là là*. *Langue française* 154 : 45-60.

DOSTIE, G. 2014. De la liberté d'association des mots lexicaux et grammaticaux au quasi-figement. La locution polycatégorielle et polysémique *c'est ça*. Communication présentée au colloque *À l'articulation du lexique, de la grammaire et du discours : Marqueurs grammaticaux et marqueurs discursifs*, 3-5 avril 2014, Paris, France.

DOSTIE, G., PUSCH, C. D. 2007. Présentation. Les marqueurs discursifs. Sens et variation. *Langue française* 154(2) : 3-12.

FISCHER, K. 2000. Discourse particles, turn-taking, and the semantics-pragmatics interface. *Revue de Sémantique et Pragmatique* 8 : 111-137.

FLEISS, J. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76 : 378-382.

FRASER, B. 1999. What are discourse markers? *Journal of Pragmatics* 31 : 931-952.

GEERTZEN, J. 2012. Inter-rater agreement with multiple raters and variables. Accessible en ligne sur <https://mlnl.net/jg/software/ira/>, consulté le 29 Novembre 2014.

GEORGAKOPOULOU, A., GOUTSOS, D. 1998. Conjunctions versus discourse markers in Greek : The interaction of frequency, position, and functions in context. *Linguistics* 36(357) : 887-917.

GILQUIN, G. 2008. What you see ain't what you get : Highly polysemous verbs in mind and language. In J.-R. LAPAIRE, G. DESAGULIER, J.-B. GUIGNARD, *Du fait grammatical au fait cognitif. From Gram to Mind: Grammar as Cognition. Volume 2*. Pessac : Presses Universitaires de Bordeaux, 235-255.

GLYNN, D. 2010. Corpus-driven cognitive semantics. Introduction to the field. In D. GLYNN, K. FISCHER (eds), *Corpus-driven cognitive semantics. Quantitative approaches*. Berlin : Mouton de Gruyter, 1-42.

GONZÀLEZ, M. 2005. Pragmatic markers and discourse coherence relations in English and Catalan oral narrative. *Discourse Studies* 7(1) : 53-86.

HANSEN, M.-B. M. 1995. Marqueurs métadiscursifs en français parlé : L'exemple de *bon* et de *ben*. *Le français moderne* LXIII(1) : 20-41.

HANSEN, M.-B. M. 1998. *The function of discourse particles. A study with particular reference to spoken standard French*. Amsterdam, Philadelphia : John Benjamins.

HANSEN, M.-B. M. 2006. A dynamic polysemy approach to the lexical semantics of discourse markers (with an exemplary analysis of French *toujours*). In K. FISCHER (ed.), *Approaches to discourse particles*. Amsterdam : Elsevier, 21-41.

HEEMAN, P., BYRON, D., ALLEN, J. 1998. Identifying discourse markers in spoken dialog. In *American Association for Artificial Intelligence – Spring symposium on applying machine learning to discourse processing (AAAI 1998)*. Stanford : Stanford University Press, 44-51.

KRIPPENDORFF, K. 1980. *Content analysis : An introduction to its methodology*. Beverly Hills, US : Sage Publications.

LEWIS, D. 2006. Discourse markers in English : A discourse-pragmatic view. In K. FISCHER (ed.), *Approaches to discourse particles*. Amsterdam : Elsevier, 21-41.

LIGHT, R. 1971. Measures of response agreement for qualitative data : Some generalizations and alternatives. *Psychological Bulletin* 76 : 365-377.

LINDSTRÖM, J. 2001. Inner and outer syntax of constructions : The case of the *x och x* construction in Swedish. Communication présentée au panel *Pragmatic aspects of frame semantics and construction grammar*, *7th International Pragmatics Conference*, 9-14 juillet 2000, Budapest, Hongrie.

LINDSTRÖM, J., KARLSSON, S. 2005. Verb-first constructions as a syntactic and functional resource in (spoken) Swedish. In *Nordic Journal of Linguistics* 28(1) : 1-35.

LOUWERSE, M. M., MITCHELL, H. H. 2003. Toward a taxonomy of a set of discourse markers in dialog : A theoretical and computational linguistic account. *Discourse Processes* 35(3), 199-239.

MARCUS, M. P., SANTORINI, B., MARCINKIEWICZ, M. A. 1994. Building a large annotated corpus of English : The Penn Treebank. *Computational linguistics* 19(2) : 313-330.

MASCHLER, Y. 2002. The role of discourse markers in the construction of multivocality in Israeli Hebrew talk-in-interaction. *Research on Language and Social Interaction* 35(1) : 1-38.

MORTIER, L., DEGAND, L. 2009. Adversative discourse markers in contrast : The need for a combined corpus approach. *International journal of corpus linguistics* 14(3) : 338-66.

PONS BORDERÍA, S. 2008. Do discourse markers exist? On the treatment of discourse markers in Relevance Theory. *Journal of Pragmatics* 40 : 1411-1434.

SCHIFFRIN, D. 1987. *Discourse markers*. Cambridge : CUP.

SCHOURUP, L. 1999. Discourse markers. *Lingua* 107 : 227-265.

SPOOREN, W., DEGAND, L. 2010. Coding coherence relations : Reliability and validity. *Corpus Linguistics and Linguistic Theory* 6(2) : 241-266.

TLFi (Trésor de la Langue Française informatisé), <http://atilf.atilf.fr>.

UYGUR-DISTEXHE, D. 2012. *Lol*, *mdr* and *ptdr* : An inclusive and gradual approach to discourse markers. *Lingvisticae Investigationes* 35(2) : 389-413.

UYGUR-DISTEXHE, D. 2014. *Lol*, *mdr* and *ptdr* : An inclusive and gradual approach to discourse markers. In L.-A. COUGNON, C. FAIRON (eds), *SMS Communication. A linguistic approach* (Benjamin Current Topics 61). Amsterdam, New York : John Benjamins, 239-263.

VINCENT, D. 1993. *Les ponctuants de la langue et autres mots du discours*. Québec : Nuit Blanche.

WALTEREIT, R. 2007. A propos de la genèse diachronique des combinaisons de marqueurs. L'exemple de *bon ben* et *enfin bref*. *Langue Francaise* 154 : 94-109.

WILSON, D. 2011. The conceptual-procedural distinction : Past, present and future. In V. ESCANDELL-VIDAL, M. LEONETTI, A. AHERN (eds), *Procedural meaning : Problems and perspectives* (Current Research in the Semantics/Pragmatics Interface 25). Emerald : Bingley, 3-31.

ZUFFEREY, S., DEGAND, L., POPESCU-BELIS, A., SANDERS, T. 2012. Empirical validations of multilingual annotation schemes for discourse relations. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turquie : 2716-2720.

ZUFFEREY, S., DEGAND, L. 2014. Representing the meaning of discourse connectives for multilingual purposes. *Corpus Linguistics and Linguistic Theory* 10 : 1-24.

ZUFFEREY, S., POPESCU-BELIS, A. 2004. Towards automatic identification of discourse markers in dialogs : The case of *like*. In *Proceedings of SIGDIAL'04 (5th SIGdial Workshop on Discourse and Dialogue)*. Cambridge, US : 63-71.