Annotation de Marqueurs Discursifs : le cas de la désambiguïsation de *après*.

Paola Herreño Castañeda[1] Maeva Sillaire[1]
(1) Université de Lorraine, CNRS, ATILF, 44 avenue de la Libération, 54000 Nancy, France
paola.herreno-castaneda@univ-lorraine.fr
maeva.sillaire@univ-lorraine.fr

RÉSUMÉ

Les marqueurs discursifs (désormais MD) sont des expressions souvent polysémiques, voire polyfonctionnelles dans la langue (*quoi*, *enfin*, *bon*, *mais*, *voilà*, *là*, etc.). Dans ce dernier cas, une tâche consiste d'abord à distinguer leurs emplois comme MD et non-MD, en fonction notamment du contexte d'apparition. Dans le cadre de CODIM, un corpus de français a été constitué et annoté semi-automatiquement pour identifier les expressions potentiellement employées comme MD, non-MD, ou MD-CAND (étiquette regroupant les cas ambigus qui n'ont pas pu être déterminés par l'annotation). Nous cherchons à enrichir le processus d'annotation pour les cas où *après* a été classé comme MD-CAND. Pour cela, nous proposons un protocole d'annotation manuelle supplémentaire visant à trier, parmi ces candidats, les emplois contrastifs et non-contrastifs de *après*. Nos résultats initient des réflexions plus larges sur les enjeux théoriques et méthodologiques liés à l'annotation des MD.

ABSTRACT

Annotation of Discourse Markers : the case of the disambiguation of *après*

Discourse markers (henceforth DMs) are expressions often polysemous, and even multifunctional in the language (in French : *quoi*, *enfin*, *bon*, *mais*, *voilà*, *là*, etc.) In this latter case, one task is first to distinguish their uses as DM and non-DM, particularly based on the context in which they appear. Within the framework of CODIM, a corpus of French has been created and semi-automatically annotated to identify expressions potentially used as DM, non-DM, or DM-CAND (a label grouping ambiguous cases that could not be determined by the annotation). We seek to enhance the annotation process where the french expression *après* has been classified as DM-CAND. To achieve this, we propose an additional manual annotation protocol aimed at sorting, among these candidates, the contrastive and non-contrastive uses of *après*. Our results initiate broader reflections on the theoretical and methodological issues related to the annotation of DMs.

MOTS-CLÉS : Marqueur discursif, annotation, corpus, méthodologie.

KEYWORDS: Discourse marker, annotation, corpus, methodology.

1 Introduction

Les études concernant les marqueurs discursifs (désormais MD) sont nombreuses (pour une synthèse voir Dargnat, 2023, 2024; Hansen & Visconti, 2024). Alors que les définitions et propriétés des

MD varient selon les approches, certaines caractéristiques demeurent communes : il s'agit d'expressions généralement invariables qui, dans leur grande majorité, ne contribuent pas au contenu véri-conditionnel de l'énoncé dans lequel elles se trouvent et qui présentent une certaine *polyfonctionnalité*. On peut distinguer trois grands types de fonctions : 1) la structuration logico-temporelle du discours (on parle alors de « connecteurs », pour une liste indicative en français voir Roze, 2009) ; 2) la manifestation des états mentaux du locuteur (attitudes, émotions) ; et 3) la gestion de l'interaction. Ces deux derniers cas sont parfois étiquetés comme « particules énonciatives », « petits mots du discours », « interjections », etc.

Le corpus que nous utilisons est constitué de données disponibles, réannotées dans le cadre du projet CODIM, qui vise l'étude sémantico-pragmatique des combinaisons de MD (ex. *mais enfin*, *donc du coup*, *et alors*, etc.) La nature des données et le processus d'identification et d'annotation des MD entrant dans ces combinaisons sont décrits à la section 2.3. Ici, nous effectuons un retour sur cette première phase d'annotation. Nous nous concentrons sur l'annotation de *après* en français, en particulier les cas que la première phase d'annotation a laissés au stade de candidat MD (étiquette MD-CAND), c'est-à-dire des cas où l'emploi de *après* semble ambigu. *Après* possédant plusieurs sens MD possibles (cf. section 2.1), le protocole d'annotation que nous proposons permet d'améliorer la première tâche d'annotation, en intégrant des contextes qui n'avaient pas été traités, et, d'un point de vue sémantique, il sert à identifier plus finement les cas où *après* correspond à un emploi MD avec un sens contrastif.

La première partie est une présentation de l'objet d'étude *après* (sections 2.1 et 2.2) et des données utilisées (section 2.3). La seconde partie détaille les deux tâches d'annotation réalisées (sections 3.1 et 3.2). La troisième partie concerne la discussion des résultats (section 4). La section 5 conclut en ouvrant quelques perspectives et pistes pour des tâches futures.

2 Objet d'étude

Dans cette section, nous décrivons les caractéristiques de l'expression *après* et les notions nécessaires à l'identification des emplois MD. Nous décrivons ensuite le processus d'annotation semi-automatique de *après* et ses résultats dans les données utilisées.

2.1 *Après*

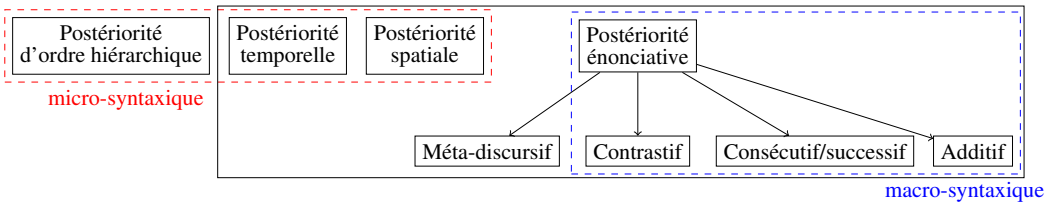

FIGURE 1 – Synthèse des différents emplois de l'expression *après*.

Après était au départ une préposition désignant la proximité spatiale, puis sa valeur sémantique s'est étendue au domaine temporel (Fagard, 2003). En français actuel, l'expression *après* possède

différents sens et la figure 1 reprend les principaux généralement avancés dans les études. En général, la catégorisation des sens repose sur le type de postériorité que *après* désigne : celle-ci peut être spatiale (exemple 1), temporelle (exemple 2) ou énonciative, telle que détaillée ci-dessous.

La valeur énonciative se décline en plusieurs traits sémantiques, chacun pouvant se superposer (Le Draoulec, 2017; Skrovec *et al.*, 2022) : le trait additif introduit souvent un élément dans une énumération comme dans l'exemple 6, le trait contrastif concerne les cas où deux situations sont opposées ou bien que le locuteur revoit son discours comme dans l'exemple 5, le trait consécutif désigne les cas où la postériorité introduit un lien causal comme dans l'exemple 4 et le trait méta-discursif se retrouve lorsque que *après* permet un recul sur le discours en cours et le commente ou contribue à son organisation, comme dans l'exemple 7. Les valeurs décrites dans cette approche correspondent à celles contenues dans le cadre noir le plus large.

Une autre approche (Akihiro, 2020) distingue deux *après*, en fonction de ses caractéristiques syntaxiques. Le *après* micro-syntaxique appartient à la dépendance grammaticale et est associé à trois valeurs sémantiques : la postériorité temporelle (exemple 2), la postériorité spatiale (exemple 1) et la postériorité d'ordre hiérarchique (exemple 3). Ces valeurs sont reprises dans l'encart rouge du schéma 1. Le *après* macro-syntaxique connecte deux unités discursives et est associé à trois valeurs sémantiques : *après* contrastif (exemple 5), *après* consécutif (exemple 4) et *après* additif (exemple 6). Ces trois valeurs correspondent à celles de l'encart bleu du schéma 1.

Les cas où *après* est considéré comme un MD correspondent en général aux cas de postériorité énonciative (exemples 4, 5, 6, 7) ou, si l'on considère la distinction macro/micro-syntaxique, les *après* macro-syntaxiques.

(1) ben y a tout tout ce qui est auchan carrefour tout ça on peut trouver un peu de tout **après** orléans centre y a vraiment des des boutiques euh un ouais y a tout aussi c'est sympa on peut trouver quand même pas mal de choses [CODIM-ESLO]

(2) ou juste **après** les vacances de février y a une fille qui tient un cahier de textes je pourrais vous le dire je pourrai vous le dire euh précisément [CODIM-ESLO]

(3) deux ans plus tôt aux mondiaux de rome les bleus étaient arrivés soixante-huit centièmes **après** les américains [CODIM-Le monde 2012]

(4) je voulais changer **après** ce qui m'a axée vers ce lycée là ça a été ces choix là ça a été qu'ils avaient de très bons résultats que euh qu'ils avaient plusieurs choix plusieurs critères de sélection et tout ça [CODIM-CFPP]

(5) je pense que il y en a beaucoup moins par exemple chloé elle est pff elle comprend pas **après** elle a d'autres euh je pense qu'elle a d'autres euh façons de pff de s'exprimer en cachette mais euh [CODIM-MPF]

(6) - et le peu que tu y joues à la en c'est quel type de jeu plutôt ?
- [...]
- Mario Mario Kart euh **après** les Lapins Crétins [CODIM-ESLO]

(7) Héhé c'est clair, je pense que Hollande prendra sa place, **après** je sais pas si ce sera mieux ... [CODIM-REDDIT]

Un lien peut être fait entre le genre discursif et le type d'emploi de *après* : l'emploi MD pragmatique semble beaucoup plus fréquent à l'oral spontané (et dans toutes les situations s'en rapprochant comme dans les écrits produits sur ordinateur Herring, 2005). Le Draoulec élabore une hypothèse pour expliquer ce phénomène : cet emploi est peu stabilisé et, de ce fait, il doit se réaliser en combinaison avec d'autres indices permettant d'inférer son interprétation pragmatique (contrastive) plutôt que temporelle. Cette hypothèse pourrait se réécrire en terme de pragmaticalisation, dans la mesure où la valeur contrastive relève d'une perspective plus énonciative et argumentative et met ainsi en jeu un cadre subjectif et intersubjectif. (Dostie, 2004; Heine, 2013; Closs Traugott, 2010).

2.2 Identification des MD

Désambiguïsation La désambiguïsation est une tâche qui consiste à assigner aux expressions polysémiques une classe correspondant à leur sens le plus probable dans un contexte donné. Les classes possibles correspondent à l'inventaire des sens de l'expression étudiée (Navigli, 2009).

Ici, une expression potentiellement MD (*après* MD-CAND) peut être ambiguë de deux façons : 1) on peut opérer une classification binaire si l'on sépare son emploi MD et son emploi non-MD (ex. *après* de postériorité énonciative, illustré par les exemples exemples 4, 5, 6, 7, et *après* de postériorité temporelle, illustré par l'exemple 2) et 2) on peut opérer une classification multiclasse si l'on considère les différentes valeurs du MD (ex. *après* contrastif correspondant à l'exemple 5, *après* additif correspondant à l'exemple 6, *après* consécutif/successif correspondant à l'exemple 4 et *après* méta-discursif, correspondant à l'exemple 7).

Dans le cadre du projet CODIM, une première désambiguïsation de type MD, non-MD, MD-Candidat pour le français a été faite à partir de graphes et un processus de transduction exploitant des contextes linguistiques prédéterminés. Une description plus détaillée des principes de cette annotation est donnée en section 2.3.

Annotation manuelle L'annotation manuelle des MD est une tâche difficile pour plusieurs raisons : 1) Il n'y pas de consensus sur la définition des MD et, donc, inévitablement sur les expressions considérées comme MD ou non au sein d'une même langue donnée, 2) l'emploi MD ou non-MD d'une expression peut dépendre du genre discursif et un outil d'identification efficace pour les données écrites ne l'est pas forcément pour les données orales. L'annotation doit donc prendre en compte le genre discursif. Le projet MDMA (*Model for Discourse Marker Annotation*) (Bolly *et al.*, 2015) s'était fixé pour objectif de créer une méthode empirique pour l'identification et l'annotation des MD en français oral. Leurs travaux produits indiquent que certaines caractéristiques sont associées aux expressions MD (entre autres : la position initiale, les sens non codé et procédural) et d'autres aux expressions non-MD (entre autres : les positions médiane et finale, les sens codé et procédural-conceptuel). De façon générale, ces travaux établissent que la position syntaxique est l'indicateur le plus prédictif d'un DM.

2.3 Le cas de *après* dans les données

Le corpus constitué dans le cadre du projet CODIM a été construit à partir de plusieurs corpus existants et disponibles, représentant une diversité de genres discursifs et de médiums, en particulier des données orales, écrites et numériques. Ce dernier ensemble correspond aux textes récupérés sur la

plateformes Reddit et au corpus Wikiconflit (Poudat *et al.*, 2017). Après le traitement automatique des corpus et une tokenisation à l'aide de Python et de la bibliothèque NLTK, nous avons obtenu un corpus final d'environ 110 millions de tokens. La proportion de tokens par type de corpus est hétérogène. Comme cela est indiqué dans la figure 2, le corpus oral est plus petit que les aux autres types, ce qui s'explique par la difficulté de récupération et de traitement de ce type d'interaction.

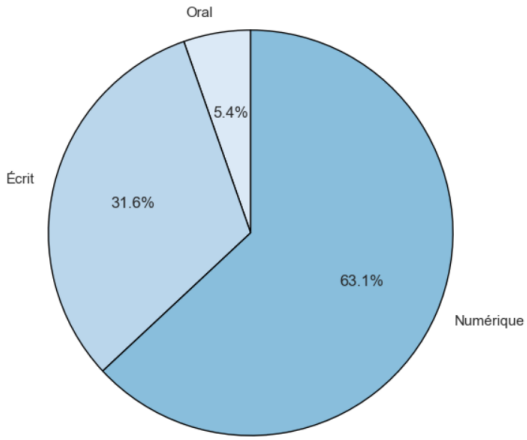

FIGURE 2 – Proportion de tokens par type de données

L'objectif d'avoir un grand corpus est de pouvoir analyser de façon approfondie les MD et leurs différentes utilisations et combinaisons, en particulier en adoptant une approche quantitative. Pour ce faire, une tâche d'annotation semi-automatique a été conçue. Cette tâche a été réalisée en utilisant le logiciel Unitex-GramLab (Paumier *et al.*, 2021) qui, à l'aide de dictionnaires, de graphes et de cascades de graphes, parcourt les données en ajoutant des étiquettes selon le contexte d'apparition des expressions d'intérêt.

Les graphes utilisés pour l'annotation des MD ont été construits à partir d'une liste de ces expressions et de leurs possibles contextes d'apparition afin de différencier leurs emplois. Trois annotations possibles ont été définies pour la tâche d'annotation des MD : a) MD (MDC pour les emplois connecteurs et MDP pour les emplois particules énonciatives), b) non-MD et c) MD-CAND. Cette dernière étiquette a été créée pour les cas où le contexte d'apparition n'apporte pas les informations nécessaires pour classer l'expression en tant que MD ou non.

La figure 3 montre la proportion des trois étiquettes par type de données. Dans le cas de l'oral, la proportion des expressions étiquetées en tant que MD-CAND représente moins du 30% des annotations faites, ce qui nous laisse penser que les contextes d'apparition sont moins ambigus que dans le cas des données écrites et numériques. La proportion assez élevée de MD-CAND dans le corpus total est l'une des motivations de la présente étude. L'annotation réalisée par des automates implique une réflexion particulière pour chacun des MD et de leurs possibles contextes, c'est pourquoi une tâche d'annotation à l'aide d'un protocole semble requise pour améliorer les paramètres retenus dans le but de rendre les automates plus robustes et efficaces.

Nous nous sommes d'abord intéressés aux MD les plus fréquents dans les données, par exemple *et*, *mais*, *donc*, etc. Cependant, le cas de *après* nous a intéressé le plus en raison de sa complexité sémantique (*cf* section 2.1).

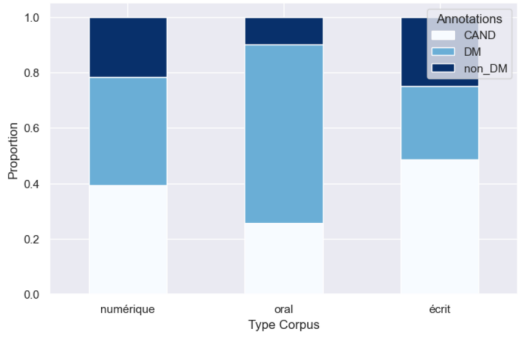

FIGURE 3 – Distribution proportionnelle d'annotations par type de corpus

Afin de faire l'annotation semi-automatique de *après*, deux graphes ont été créés. Ils s'appliquent en cascade, c'est-à-dire l'un à la suite de l'autre. Le principe repose sur l'élimination de certains emplois. Le premier graphe (repris en figure 4) concerne les cas où *après* n'est pas un MD. Nous trouvons dans ce graphe des contextes où, par exemple, *après* est précédé de mots comme *jours*, *minutes*, *années*, etc.

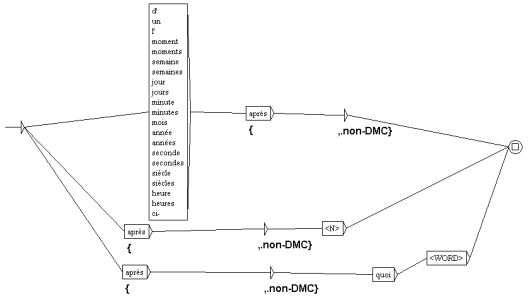

FIGURE 4 – Graphe 1 pour l'annotation de *après* en tant que non-DM.

Le deuxième graphe (indiqué en figure 5) est lancé après le premier dans l'objectif de mettre de côté les cas où *après* est annoté comme non-MD, pour ensuite analyser d'autres contextes d'apparition possibles et d'annoter *après* en tant que MD ou MD-CAND.

Les résultats de cette annotation semi-automatique reflètent la difficulté d'établir des contextes spécifiques pour le cas de *après*. En effet, le nombre d'occurrences annotées MD est 848, contre 75 151 non-MD et 70 059 MD-CAND. La figure 6 montre que dans le cas des données orales et numériques, la quantité de candidats est supérieure à 50%, tandis que pour les données écrites, elle représente environ 30%.

La première annotation par graphes permet seulement de distinguer les emplois MD et non-MD de *après*. Or, *après* présente un certain nombre de nuances sémantiques, et nous souhaitons affiner encore plus cette catégorisation en portant notre attention sur l'emploi MD contrastif, souvent présenté comme le plus pragmatique. Cet emploi est particulièrement intéressant, car il est apparemment peu stabilisé. Ainsi, en créant un protocole d'annotation supplémentaire permettant de distinguer les emplois contrastifs et non-contrastifs de *après* pour les cas annotés MD-CAND (c'est-à-dire dans des

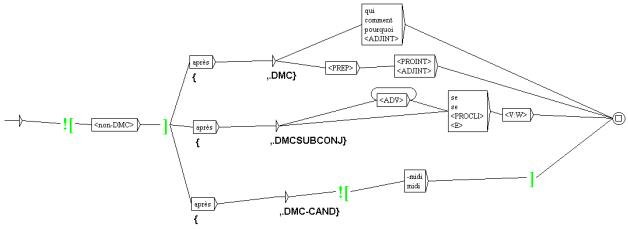

FIGURE 5 – Graphe 2 pour l'annotation de *après* en tant que MD ou MD-CAND.

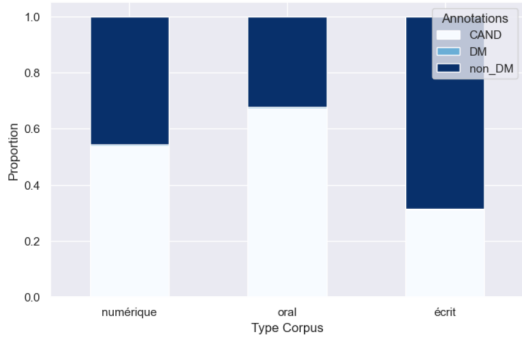

FIGURE 6 – Proportion d'annotations sur *après* par type de corpus.

contextes non prévus par l'annotation par graphes), nous pourrions étudier ce phénomène de façon plus précise et enrichir, a posteriori, lesdits graphes.

3 Présentation de la tâche

Nous avons réalisé une tâche d'annotation complémentaire en explicitant un protocole. Cette tâche a pour objectif de distinguer les emplois comme MD contrastif des emplois MD non-contrastif de *après*. Pour cela, nous avons réalisé une première annotation manuelle sur une sélection aléatoire de données, qui a soulevé les questions à l'origine de la constitution du protocole permettant l'identification de l'emploi de *après* MD contrastif. Dans un second temps, nous avons appliqué ce protocole d'annotation à une autre sélection de données.

3.1 Première étape d'annotation

Description Nous avons d'abord réalisé une annotation « naïve ». Nous avons choisi de nous baser sur l'optionnalité syntaxique, c'est-à-dire le fait que la suppression du DM n'entraîne pas l'agrammaticalité de l'énoncé, pour tester le caractère contrastif de *après*. En effet, le contraste est une relation qui en général nécessite d'être marquée explicitement, c'est-à-dire qu'elle doit être associée à un connecteur (Asr & Demberg, 2012a). En partant de ce principe, nous avons réalisé des tests de suppression de *après*. Si la suppression était possible, alors nous avons attribué l'étiquette « non-

contrastif » et, dans le cas contraire, l'étiquette « contrastif ». Une troisième étiquette « indéterminé » a été utilisée pour les cas où l'énoncé semblait incomplet et ne nous permettait pas de trancher.

Pour cette première étape, l'annotation a été conduite par deux annotateurs experts et nous avons choisi au hasard un total de 450 (150 par type de corpus) occurrences de *après* étiquetées comme MD-CAND.

Résultats Pour la première étape d'annotation, le coefficient Kappa de Cohen a une valeur peu satisfaisante de $0,54$.

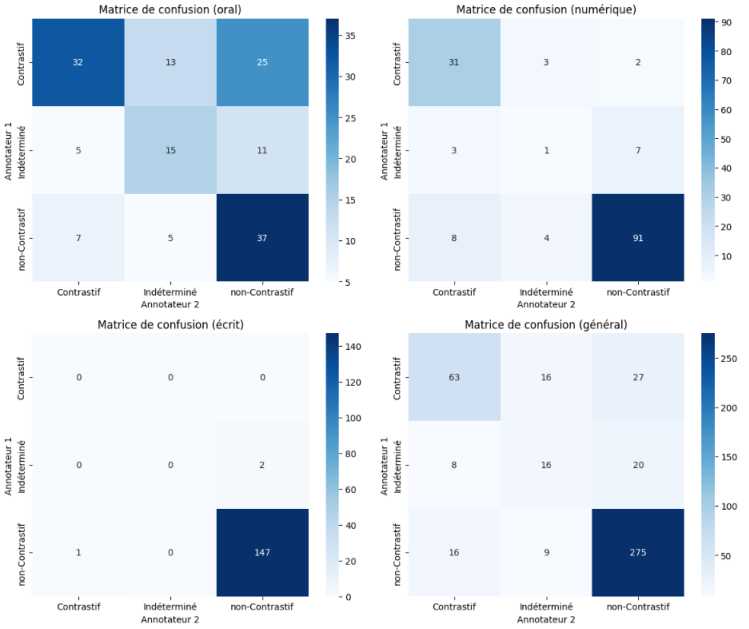

FIGURE 7 – Matrices de confusion pour la première étape d'annotation.

La figure 7 représente les matrices de confusion des annotations obtenues. Les trois premières représentent chaque type de données (orales, numériques et écrites) et la dernière (en bas à droite) récapitule les annotations obtenues sur l'ensemble.

Nous observons plus de désaccords dans le corpus oral que dans le corpus numérique, et plus dans le corpus numérique que dans le corpus écrit. Cette remarque est conforme à nos attentes, étant donné les caractéristiques de ces trois genres discursifs : l'écrit est en généralement plus normé (d'autant plus qu'il s'agit ici d'écrits journalistiques) que l'oral, le numérique reprenant des caractéristiques aux deux genres. Or, les phénomènes oraux comme les amorces, les ruptures de constructions ou les chevauchements compliquent l'application du critère d'optionnalité.

Cette étape a également permis de découvrir que certains cas annotés MD-CAND (et particulièrement ceux issus des corpus écrits) correspondaient à des cas de *après* non-MD se réalisant dans des contextes non pris en compte dans l'annotation semi-automatique initiale.

Les désaccords d'annotation ont permis de discuter des caractéristiques les plus pertinentes pour la

création d'un protocole, visant à standardiser l'annotation de *après* comme MD contrastif et MD non-contrastif.

3.2 Deuxième étape d'annotation

3.2.1 Présentation du protocole d'annotation

Le protocole a été constitué à partir de nos réflexions sur les différents exemples trouvés lors de la première phase d'annotation. La figure 8 illustre l'arbre de décision correspondant au protocole proposé.

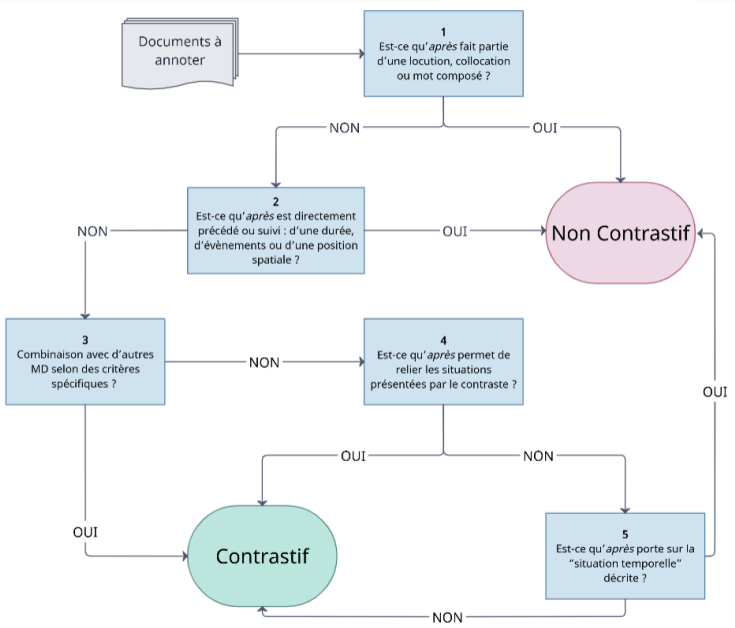

FIGURE 8 – Arbre de décision du protocole d'annotation.

Locutions, collocations et mots composés La première étape consiste à écarter tous les cas où *après* fait partie d'une locution, d'une collocation ou d'un mot composé (*avant et après, les uns après les autres, après-guerre, etc...*). Comme cela a été dit plus haut, cette étape sert à compléter l'annotation semi-automatique initiale puisque certains contextes n'avaient pas été pris en compte. Si *après* ne se trouve pas dans ces cas, le protocole implique de continuer à l'étape 2.

Durée, événements ou position spatiale La seconde étape consiste à annoter les cas où *après* est directement précédé ou suivi :

1. D'une une durée, du type : *70 ans, une semaines, trois années, etc...*
2. D' évènements, notamment des noms à lecture évènementielle comme : *les législatives, le bac, les élections, etc...*

3. D'une position spatiale, qui peut être comprise dans une succession spatiale comme : *après Versailles*

Si *après* ne correspond pas à l'un de ces cas, il est nécessaire de continuer à l'étape 3.

Combinaison avec d'autres MD Les MD peuvent s'agglutiner entre eux (Crible & Degand, 2021; Cuenca, 2024; Dargnat, 2022; Dostie, 2013). Dans la troisième étape du protocole, nous utilisons cette caractéristique pour identifier les cas où l'emploi comme MD contrastif est le plus probable :

1. Est-ce qu'*après* est entouré de façon contiguë à gauche et à droite de MD ? (ex. *bon après euh...*). Si non, continuer.
2. Est-ce qu'*après* apparaît en début de tour de parole ou de phrase et est suivi de MD ? (ex. *après bon...*). Si non, continuer.
3. Si *après* est précédé de *mais* : réaliser un test de combinaison en ajoutant *bon* à la fin de *mais après*. Est-il possible d'opérer cette modification ? (*mais après* = *mais après bon* ?) Si non, continuer.
4. Si *après* est précédé de *et* : appliquer le même test du cas de *mais* (*et après* = *et après bon* ?)

Les combinaisons *mais après* ainsi que *et après* étaient fréquentes dans les données et leur comportement est, par certains aspects, similaires. En ce qui concerne *mais après*, nous avons observé que cette combinaison est souvent utilisée pour introduire des idées opposées, conformément aux caractéristiques sémantiques de *après* et de *mais*. Ainsi, le test d'ajout de *bon* à la fin de la combinaison sert à confirmer son emploi MD contrastif, car son utilisation semble moins probable pour les cas non contrastifs (exemple 8). Si *après* ne correspond à aucun de ces cas, passer à l'étape 4.

(8)
tu m'aideras **après** à faire mon dessin moi [CODIM-TCOF]
? tu m'aideras **après bon** à faire mon dessin moi

Notion de contraste Dans le cas de l'étape 4 du protocole, nous nous interrogeons directement sur l'emploi de *après* dans un contexte de contraste. Afin de prendre une décision pour l'annotation à donner, il faut analyser si les situations présentées s'opposent. Si oui, cela signifie qu'*après* est un MD contrastif. Dans le cas contraire, passer à l'étape suivante.

Portée sur la situation temporelle La dernière étape du protocole consiste à évaluer à quel point *après* se rapproche de son sens initial temporel. L'objectif est de catégoriser la portée de *après* sur ce que nous appelons la *situation temporelle* des évènements décrits dans l'énonciation : est-ce que les évènements présentés se trouvent dans une séquentialité ? est-ce que *après* joue un rôle dans la structuration temporelle des évènements ? Pour répondre à ces questions, nous réalisons un test de suppression : si la phrase est modifiée en le supprimant, alors il est annoté non-contrastif. Dans le cas contraire, il est annoté contrastif.

3.2.2 Application du protocole

Description Étant donné les résultats obtenus lors de la première étape, nous avons décidé de ne plus annoter de données issues du corpus écrit, puisqu'en grande majorité *après* avait un emploi comme MD non-contrastif.

Comme pour la première étape d'annotation, celle-ci a été menée par deux annotateurs experts. Les données de la première étape étant déjà connues, nous avons récupéré 400 nouvelles occurrences de *après* (200 du corpus oral et 200 du corpus numérique). L'annotation a été réalisée en suivant le protocole décrit précédemment. L'étiquette « Indéterminé » était utilisée lorsque les informations à disposition étaient insuffisantes pour appliquer le protocole.

Résultats Après avoir appliqué le protocole à nos données, nous avons mesuré le coefficient Kappa de Cohen qui est de $0,69$., ce qui représente un accord satisfaisant.

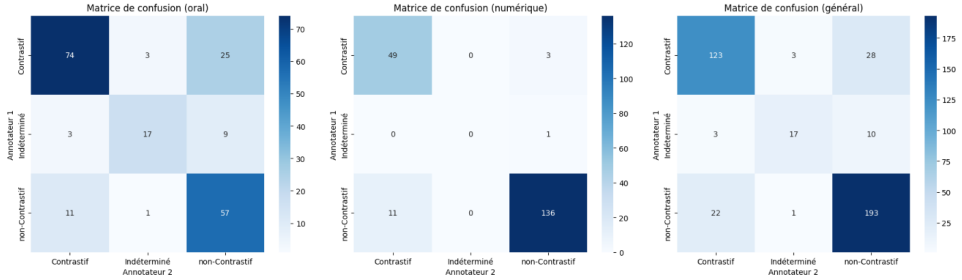

FIGURE 9 – Matrices de confusion pour la deuxième étape d'annotation par type de corpus.

Les résultats de la deuxième tâche d'annotation sont indiqués dans la figure 9. Il y a eu une amélioration par rapport aux premiers résultats.

En ce qui concerne le corpus numérique, le protocole nous a aidé à mieux déterminer les emplois MD contrastif de *après*. Les annotateurs n'ont donné une annotation différente que dans 7% des cas et l'étiquette « Indéterminé » n'a été utilisée que pour une seule occurrence.

L'annotation de l'oral est plus complexe. 18% des cas ont été annotés différemment : l'un des annotateurs a donné l'étiquette MD contrastif tandis que l'autre a opté pour MD non-contrastif. L'étiquette « Indéterminé » a été attribuée plus fréquemment à l'oral que pour le numérique ($16,5\%$ pour l'oral contre $0,2\%$ pour le numérique). Nous expliquons cela en partie par le fait que, pour la plupart des occurrences concernées, le tour de parole était incomplet.

De façon générale, les résultats de ce protocole sont encourageants. Nous avons annoté des étiquettes contraires (contrastif, non-contrastif) dans $12,5\%$ des cas, alors que pour 74% des cas, les étiquettes étaient similaires.

4 Discussion

Annotation semi-automatique par graphes (MD, non-MD, MD-CAND) L'annotation semi-automatique réalisée dans le cadre du projet CODIM utilise une approche descendante, dans la mesure où les contextes sur lesquels se basent la désambiguïsation (MD, non-MD et MD-CAND) sont établis a priori. Étant donné la diversité, le nombre des contextes possibles et le fait que certains emplois MD ne sont pas stabilisés, il est impossible de tous les recenser. Or, si le contexte d'un emploi MD d'une expression n'est pas recensé, ce contexte ne sera pas pris en compte dans l'annotation. Cette

approche reste efficace, puisqu'elle présente les avantages de pouvoir traiter facilement un grand nombre de données et de ne pas dépendre de l'identification d'arguments ou de relations de discours pour identifier les MD, permettant d'annoter également des MD non associés à des relations de discours, comme les particules énonciatives (Dargnat, 2024).

L'annotation manuelle des cas MD-CAND de *après* nous a permis de trouver des exemples de contextes qui n'ont pas été pris en compte lors de l'annotation semi-automatique (comme dans l'exemple 9). C'est pourquoi, à l'avenir, nous comptons enrichir les graphes initiaux avec les contextes que nous avons pu relever dans le cadre du travail mené pour cet article.

(9) *je ne me souviens plus exactement si on l'avait fait dans ma classe mais* ***après*** *le 11 septembre pas mal de jeunes gens refusaient la minute(...)* [CODIM-REDDIT]

Superposition des valeurs sémantiques Le fait que l'emploi *après* MD contrastif soit plus fréquent à l'oral qu'à l'écrit pourrait être expliqué par le fait que cet emploi n'est pas stabilisé et doit dépendre d'un contexte favorable pour se réaliser. En d'autres termes, nous pouvons faire l'hypothèse que la *charge sémantique* de *après* comme MD contrastif n'est pas suffisante pour l'inférence de la relation de discours associée : le contexte joue alors encore un rôle important dans l'interprétation contrastive.

L'importance du contexte, et plus précisément des signaux autres que les MD, dans l'inférence de la relation de discours est souvent discutée dans les études. Une hypothèse avancée est celle de l'existence d'un continuum de marquage discursif (ou de discursivité) allant du marquage implicite au marquage explicite (Bolly, 2014; Crible, 2020). Comme l'emploi de type MD contrastif de *après* nécessite d'autres signaux pour se réaliser, la question de sa position sur ce continuum reste à étudier.

Le travail d'annotation de *après* questionne également quant à l'intérêt d'utiliser d'une classification contrastif/non-contrastif sans autres indications. En effet, la plupart des *après* MD contrastifs, présentent également une valeur temporelle, les deux valeurs étant généralement superposées (exemple 10). L'annotation de caractéristiques sémantiques et contextuelles permettrait une meilleure capture des propriétés de certains MD dont les valeurs se superposent, comme *après*, et un meilleur transfert vers d'autres types de tâches.

(10) *non ils ont hué les américains mais* ***après*** *quand la France a gagné c'était pas si euh je veux dire je le dis moi j'ai pas senti la ferveur de la compétition hein* [CODIM-MPF]

Difficultés d'annotation La tâche d'annotation a permis d'identifier quelques difficultés.

Une première difficulté est liée à la façon dont l'énoncé est segmenté par l'annotateur. En effet, les données étant orales et numériques, de nombreux facteurs empêchent la bonne segmentation comme la non-prise en compte des traits prosodiques lors de l'annotation, l'irrégularité des signes de ponctuation ou encore les particularités des tours de parole liées à la modalité interactive (réponses et citations dans le cas de *reddit* et amorces, chevauchements et disfluences pour les conversations).

La seconde difficulté concerne les différences d'interprétation du contexte entre les annotateurs. Cette remarque rejoint celle déjà faite sur le continuum du marquage discursif. En effet, lors de l'analyse des annotations, nous avons remarqué que l'interprétation des éléments contextuels (items lexicaux, temps et aspects, ...) différaient beaucoup. Notre protocole s'étant focalisé sur l'identification de *après* comme MD contrastif, il était resté vague par rapport à l'interprétation du contexte linguistique. Le protocole à cet égard n'a pas permis une objectivisation de l'interprétation du contexte suffisante pour

assurer les mêmes annotations. Une amélioration du protocole pourrait être d'ajouter des contraintes liées à l'interprétation contextuelle.

La dernière difficulté est liée à un effet de primauté. Effectivement, dès que la phrase est lue, l'annotateur en construit une première interprétation. Cette première interprétation oriente notre application du protocole et il peut être difficile de s'en défaire ou de constituer une interprétation alternative. Dans les cas où les contributions temporelles et contrastives de *après* sur l'énonciation sont superposées et où l'application du protocole force la classification binaire (contrastif/non-contrastif), cela peut entraîner des différences au niveau de l'annotation. Une annotation prenant en compte plusieurs dimensions sémantiques permettrait de limiter les effets liés à cette première interprétation.

Combinaison (co-occurrence de MD) Notre protocole repose sur les combinaisons de MD pour identifier les emplois MD contrastifs de *après*. Deux combinaisons en particulier nous paraissent pertinentes en raison de leur fréquence d'apparition et de leur difficulté d'annotation : *et après* et *mais après*. Les exemples suivants ont reçu des étiquettes opposées lors du processus d'annotations.

(11) ***et après*** *elle me demande comment je peux être sérieux quand je préfère aller jouer à rocket league plutôt que matter cette daube* [CODIM-MPF]

(12) *ouaip faut que je réponde aux sujets que je poste pour le moment c'est la tete dans le guidon jusque fin juin pour mon année de master* ***mais après*** *il faudra essayer oui merci de ta réponse[...]* [CODIM-REDDIT]

La combinaison de l'exemple (11) illustre un cas où il est difficile d'établir la portée de *après*. L'absence de séquence temporelle explicite entre les éléments présentés permet d'accepter plusieurs interprétations de la proposition selon l'interprétation de l'annotateur. Cette combinaison pose la question du rôle et de la portée des deux MD. Dans cette co-occurrence, comme Badiou-Monferran & Capin (2021) l'expliquent, on observe une superposition des valeurs d'addition et de succession temporelle.

Finalement, le cas de *mais après* dans l'exemple (12) représente un point important de discussion. L'interprétation de l'emploi de *après* est difficile à annoter car plusieurs options sont possibles. La présence de *mais* dans la combinaison pousse à y voir une idée d'opposition qui n'est pas nécessairement explicite mais qui semble nuancer la notion temporelle de *après*.

Retour sur le protocole d'identification des emplois MD contrastif En ce qui concerne le protocole, nous avons détaillé dans cette section les indications pour son amélioration. Nous les résumons ici :

— Il serait raisonnable d'envisager d'ajouter des contraintes par rapport à l'interprétation du contexte. Cela pourrait par exemple se réaliser à travers des contraintes liées à l'aspect fini des évènements mis en relation par *après*.
— Il faudrait changer le critère d'optionnalité par un autre test permettant de mieux estimer la contribution de *après* sur l'énonciation.
— En fonction des objectifs de recherche, la classification binaire contrastif et non-contrastif peut ne pas être optimale sans ajout d'informations supplémentaires. Il serait possible d'utiliser des échelles, permettant de coder la *force* de l'inférence du MD pour différentes relations de discours (cette annotation permettrait en quelque sorte d'appréhender ce que Asr & Demberg

(2012b) appellent la force d'un connecteur). L'annotation sur plusieurs dimensions permettrait ainsi la superposition de plusieurs relations, représentant mieux la réalité complexe du MD *après*. Ce type d'annotation aurait l'avantage d'utiliser les mêmes annotations pour différents objectifs, mais présente l'inconvénient de ne reposer que sur une annotation manuelle, limitant la taille des données pouvant être traitées.

5 Conclusion & perspectives

Dans cet article, nous avons proposé un protocole d'annotation pour identifier les emplois contrastifs du MD *après* dans les occurrences initialement étiquetées MD-CAND par l'annotation semi-automatique du projet CODIM. Ce protocole a été constitué après une première annotation naïve dont l'accord inter-annotateur était moyennement satisfaisant ($0,54$ Kappa de Cohen). Après des discussions et la construction d'un protocole plus précis, une seconde tâche d'annotation manuelle a été réalisée. L'accord inter-annotateur a été amélioré ($0,69$ Kappa de Cohen) et semble satisfaisant. Pour confirmer ces résultats, nous devrons à l'avenir appliquer le protocole à un plus grand nombre d'annotateurs.

Cet article met en lumière les enjeux liés à l'annotation des MD et plus particulièrement pour le cas du MD *après*. Notre travail suggère que l'annotation en deux catégories (contrastif, non-contrastif) n'est pas optimale et qu'une annotation prenant en compte différentes dimensions rendrait plus compte de la complexité sémantique d'un MD comme *après*.

A l'avenir, nous voudrions améliorer le protocole existant à partir des recommandations faites et l'appliquer à des MD aux profils similaires : des temporels qui présentent des emplois comme MD contrastif (ex. *maintenant*, *en même temps*, etc.). Cela permettra de tester le protocole sur d'autres MD et, en cas de succès, nous pourrons étudier plus précisément les liens entre les différents sens d'un MD, avec un focus particulier sur le passage du domaine temporel à celui du raisonnement.

Nous voudrions également mener une étude plus générale sur les expressions annotées MD-CAND du projet CODIM, afin d'augmenter le nombre de contextes recensés dans les graphes pour l'annotation semi-automatique en MD ou non-MD.

Enfin, une perspective plus lointaine sera d'automatiser le protocole servant à distinguer les cas contrastifs des cas non-contrastifs. Nous pourrions envisager un processus d'annotation mélangeant plusieurs type de tâches (annotation semi-automatique par graphes, puis application d'un autre protocole). Cela permettrait de faire un apport à la tâche d'annotation automatique de corpus, plus précisément dans un contexte d'annotation de MD.

Références

AKIHIRO H. (2020). L'emploi discursif de « après », étude contrastive avec « ato » en japonais. extension contextuelle et pragmaticalisation. *Langages*, **220**, 65–86. DOI : 10.3917/lang.220.0065.

ASR F. T. & DEMBERG V. (2012a). Implicitness of discourse relations. In *Proceedings of COLING 2012*, p. 2669–2684.

ASR F. T. & DEMBERG V. (2012b). Measuring the strength of linguistic cues for discourse relations. In *Proceedings of the Workshop on Advances in Discourse Analysis and Its Computational Aspects*, p. 33–42.

BADIOU-MONFERRAN C. & CAPIN D. (2021). Cooccurrences de « et » + adverbe en diachronie longue : délimitation et enjeux d'un nouveau champ de recherche. *Çédille : Revista de Estudios Franceses*, (19), 89–125.

BOLLY C. (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., DEGAND L. & UYGUR-DISTEXHE D. (2015). Mdma. un modèle pour l'identification et l'annotation des marqueurs discursifs «potentiels» en contexte. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (16).

CLOSS TRAUGOTT E. (2010). *(Inter)subjectivity and (inter)subjectification : A reassessment*, In K. DAVIDSE, L. VANDELANOTTE & H. CUYCKENS, Éds., *Subjectification, Intersubjectification and Grammaticalization*, p. 29–74. De Gruyter Mouton : Berlin, New York. DOI : doi :10.1515/9783110226102.1.29.

CRIBLE L. (2020). Weak and strong discourse markers in speech, chat, and writing : Do signals compensate for ambiguity in explicit relations ? *Discourse processes*, **57**(9), 793–807.

CRIBLE L. & DEGAND L. (2021). Co-occurrence and ordering of discourse markers in sequences : A multifactorial study in spoken french. *Journal of Pragmatics*, **177**, 18–28. DOI : https ://doi.org/10.1016/j.pragma.2021.02.006.

CUENCA M.-J. (2024). *7 Clusters of discourse markers*, In M.-B. M. HANSEN & J. VISCONTI, Éds., *Manual of Discourse Markers in Romance*, p. 193–224. De Gruyter : Berlin, Boston. DOI : doi :10.1515/9783110711202-007.

DARGNAT M. (2022). Mais enfin : construction et association. *Langages*, **225**, 49–63.

DARGNAT M. (2023). *Lexique et discours*. Habilitation à diriger des recherches, Université Paris 8 - Saint-Denis ; Laboratoire Structures Formelles du Langage (UMR 7023). HAL : tel-04452684.

DARGNAT M. (2024). Les particules énonciatives. Encyclopédie grammaticale du français. DOI : DOI : https ://nakala.fr/10.34847/nkl.485f76mm.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

DOSTIE G. (2004). De la pragmaticalisation aux marqueurs discursifs. In *Pragmaticalisation et marqueurs discursifs*, p. 19–63. De Boeck Supérieur.

DOSTIE G. (2013). Les associations de marqueurs discursifs - de la cooccurrence libre à la collocation. *Linguistik Online*, **62**.

FAGARD B. (2003). Après : de l'espace au temps, la sémantique en diachronie. In *Linguagem, cultura e cognição*, Braga, Portugal : Universidade de Braga. HAL : halshs-01242130.

HANSEN M.-B. M. & VISCONTI J., Éds. (2024). *Manual of Discourse Markers in Romance*. Berlin, Boston : De Gruyter. DOI : doi :10.1515/9783110711202.

HEINE B. (2013). On discourse markers : Grammaticalization, pragmaticalization, or something else ? *Linguistics*, **51**. DOI : 10.1515/ling-2013-0048.

HERRING S. C. (2005). Computer-mediated discourse. *The handbook of discourse analysis*, p. 612–634.

LE DRAOULEC A. (2017). « Après moi ce que j'en dis… » L'emploi pragmatique de « après». In G. DOSTIE & F. LEFEUVRE, Éds., *Lexique, grammaire, discours - Les marqueurs discursifs*, volume 52 de Bibliothèque de grammaire et linguistique, p. 23–40. Honoré Champion.

NAVIGLI R. (2009). Word sense disambiguation : A survey. *ACM computing surveys (CSUR)*, **41**(2), 1–69.

PAUMIER S., GUENTHNER F., LAPORTE E., MALCHOK F., MARSCHNER C., MARTINEAU C., MARTÍNEZ C., MAUREL D., NAGEL S., NEME A., PETIT M., STIEHLER J. & VOLLANT G. (2021). UNITEX 3.3 Manuel d'utilisation. working paper or preprint.

POUDAT C., GRABAR N., PALOQUE-BERGÈS C., CHANIER T. & KUN J. (2017). Wikiconflits : un corpus de discussions éditoriales conflictuelles du Wikipédia francophone. In C. R. W. . G. LEDEGEN, Éd., *Corpus de communication médiée par les réseaux : construction, structuration, analyse*. L'Harmattan. HAL : hal-01485427.

ROZE C. (2009). Base lexicale de connecteurs du français. Mémoire de master, Université Paris Diderot.

SKROVEC M., KANAAN-CAILLOL L. & AKIHIRO H. (2022). Le marqueur « après » à l'oral : une approche micro-diachronique, variationniste et interactionnelle. *Langages*, **N° 226**, 117–131. DOI : 10.3917/lang.226.0117.