DÉPARTEMENT DES LETTRES ET COMMUNICATIONS
Faculté des lettres et sciences humaines
Université de Sherbrooke

# *Identification automatique et analyse sémantique des marqueurs illocutoires du français québécois en contexte de conversation familière*

par
Francis Lapointe
Maître ès arts
de l'Université de Sherbrooke

Thèse
Présentée dans le cadre du programme de doctorat en Études Françaises
(Cheminement en linguistique)
de l'Université de Sherbrooke

Sherbrooke
Août 2017

# Composition du jury

*Identification automatique et analyse sémantique des marqueurs illocutoires du français québécois en contexte de conversation familière*

Francis Lapointe

Cette thèse a été évaluée par un jury composé des personnes suivantes :

Madame Gaétane Dostie, directrice de recherche

(Département des lettres et communications, Université de Sherbrooke)

Monsieur François Lareau, codirecteur de recherche

(Département de linguistique et de traduction, Université de Montréal)

Madame Fouzia Benzakour, examinatrice interne au programme

(Département des lettres et communications, Université de Sherbrooke)

Madame Samia Bouchaddakh, examinatrice interne au programme

(Département des lettres et communications, Université de Sherbrooke)

Madame Agnès Tutin, examinatrice externe

(Université Stendhal - Grenoble 3)

# Remerciements

Je remercie mes professeurs Gaétane Dostie et François Lareau qui ont dirigé la rédaction de ma thèse. Le choix de Mme Dostie comme directrice s'est imposé naturellement suite à mon expérience sous sa direction à la maîtrise. Je ne peux imaginer une directrice plus consciencieuse et disponible. M. Lareau a eu la patience d'orienter mes premiers pas dans le vaste domaine du traitement automatique de la langue malgré ma complète ignorance initiale.

Je remercie Mmes Benzakour, Bouchaddakh et Tutin, les membres du jury qui ont honoré mon travail en acceptant de lire et de commenter ma thèse.

Je remercie mes collègues étudiants et professeurs membres des groupes de recherche CATIFQ, CRIFUQ, OLST et RALI pour avoir écouté mes présentations de recherche et offert leurs commentaires stimulants.

Je remercie les membres de ma famille, mes partenaires de vie Marino et Inaki, mon frère Maxime et mes parents Ginette et Michel, pour leur soutien inconditionnel, tant moral qu'économique.

# Résumé

Les marqueurs illocutoires (MI) sont des unités lexicales indépendantes syntaxiquement qui réalisent des actes illocutoires expressifs, directifs ou assertifs. Ces mots-phrases, comme *wow*, *coudon*, *franchement!* et *mon dieu!*, nous apparaissent comme une des clés de l'expression de la subjectivité à l'oral en contexte de conversation.

L'analyse automatique de ces unités, leur identification et la détermination de leur sens par un système informatique, soulève des problèmes particuliers liés à leur polysémie, à leur comportement syntaxique et à leur relative faible fréquence dans les corpus actuellement disponibles. Dans cette thèse, nous cherchons à résoudre ces problèmes à l'aide du Corpus de français parlé au Québec (CFPQ) comme source de données, des librairies en Python du *Natural Language Toolkit* (NLTK) et de *scikit-learn* comme outils informatiques et des travaux réalisés dans les cadres de la Métalangue sémantique naturelle (MSN) et de la théorie Sens-Texte (TST) comme outils théoriques.

Suite à un état de la question au sujet des MI et du traitement automatique des marqueurs discursifs en général, nous présentons les résultats d'une expérience au sujet de l'identification automatique des MI ambigus présents dans le CFPQ. L'identification de certains MI est triviale parce que ceux-ci se présentent sous des formes qui ne sont pas ambiguës (*chut* et *coudon*, par exemple). L'identification des MI qui sont homonymes avec d'autres classes grammaticales (comme *regarde* et *sérieux*) est plus difficile. Nous voyons qu'il est possible de repérer ceux-ci à l'aide de méthodes automatiques qui obtiennent des f-mesures variant entre 75% et 100% selon les unités, avec une moyenne de 93,98% pour la meilleure méthode. Un étiqueteur à n-grammes et un classifieur de type SVM (*support vector machine*) sont les principaux outils informatiques utilisés par ces méthodes. L'étiqueteur à n-grammes est entraîné sur un ensemble d'étiquettes spécifiquement conçu pour favoriser l'identification des MI. Le classifieur SVM base principalement son entraînement et son analyse sur l'observation des textes et des résultats de l'étiqueteur à n-grammes.

Nous proposons ensuite un système de description sémantique modulaire des MI qui nous permet de décrire leurs signifiés par la combinaison de 17 paraphrases simples en langue naturelle.

Nous terminons notre étude par la présentation d'un exemple d'analyse de texte à l'aide du système d'identification et d'interprétation des MI développé au cours de la thèse.

Mots-clés : Marqueurs discursifs; marqueurs illocutoires; interjections; traitement automatique de la langue; français québécois

# Table des matières

# Liste des tableaux

# Introduction

Cette recherche vise à apporter une contribution aux domaines de la lexicologie, de la sémantique et de l'analyse automatique des textes en ce qui a trait au lexique des marqueurs illocutoires (MI).

Les MI forment une sous-catégorie des marqueurs discursifs (MD). Comme les autres MD, ils sont des unités pragmaticalisées qui subissent peu de variation flexionnelle. Ils sont également des unités extraphrastiques (des mots-phrases), c'est-à-dire qu'ils forment des phrases à eux seuls et n'entrent généralement pas en relation syntaxique avec d'autres unités. Les MI se distinguent des marqueurs d'interaction par leur capacité à réaliser des actes illocutoires en lien avec des éléments propositionnels énoncés (ou suggérés) dans une situation de communication, plutôt qu'avec des éléments interactionnels de cette situation de communication (voir chapitre 2-1).

Au cours d'une conversation, un énonciateur peut avoir recours à des MI afin d'exprimer divers sentiments comme l'étonnement, le dégoût ou la douleur. Dans l'extrait du Corpus de français parlé au Québec (CFPQ) présenté en (1), l'énonciatrice C exprime son approbation devant quelque chose qu'elle juge de manière positive à l'aide du MI COOL.

(1) C : ah tant mieux pour toi [1**cool**
*[CFPQ, sous-corpus 17, segment 3, page 34, ligne 2]*

Le sens des MI est généralement complexe. Plusieurs MI sont non seulement expressifs, mais aussi assertifs ou directifs. En (2) par exemple, l'énonciatrice utilise le MI ÉCOUTE afin d'exprimer un désaccord au sujet d'un état de choses et d'inviter son coénonciateur à y prêter attention.

(2) I : ben là **écoute** là ça avait pas d'allure
*[CFPQ, sous-corpus 7, segment 1, page 4, ligne 5]*

Comme les exemples précédents le montrent, les MI sont typiquement isolés syntaxiquement et sont souvent fortement chargés en contenu émotif. Pour ces raisons, ils pourraient être particulièrement utiles pour le champ d'application de l'analyse automatique que l'on nomme « la fouille d'opinion » ou « l'analyse de sentiment » (*opinion mining* et *sentiment analysis* en anglais). Ce champ d'application vise à repérer et à caractériser les éléments de subjectivité dans les discours (voir Pang et Lee, 2008 pour un inventaire d'études et de techniques liées à l'analyse de sentiment).

Au cours de nos analyses, nous utiliserons la terminologie de la Théorie Sens-Texte (TST), particulièrement bien adaptée à la description des unités lexicales (Mel'čuk, 1997; Mel'čuk, Clas et Polguère, 1995). La description lexicographique des marqueurs discursifs telle que développée par Dostie (2004) ainsi que l'analyse des mots de sentiment telle que menée par Goddard (2013, 2014) à l'aide de la Métalangue sémantique naturelle (Wierzbicka, 1972, 1980, 1999) offrent plusieurs outils pour encadrer notre analyse du lexique à l'étude. Notons, par exemple, que nous adoptons une approche lexicographique polysémique, nous permettant de distinguer les emplois discursifs et non discursifs de nombreux vocables. Nous privilégions également l'usage d'un métalangage naturel plutôt que d'un métalangage artificiel pour décrire les faits sémantiques.

Le chapitre 1 de cette thèse est consacré à sa mise en contexte, sa pertinence et ses objectifs. Nous y donnons également un aperçu du corpus que nous étudions (CFPQ).

Le chapitre 2 propose un court état de la question sur les MI en tant que classe grammaticale et sur les aspects du traitement automatique de la langue en lien avec ceux-ci.

Le chapitre 3 rend compte du processus d'élaboration et de comparaison de quatre méthodes d'identification automatique des MI présents dans le CFPQ.

Le chapitre 4 propose un système de description sémantique des MI à partir d'observations faites sur leur utilisation par les locuteurs du CFPQ.

Le chapitre 5 présente un exemple d'analyse de texte à l'aide du système d'identification et d'interprétation des MI développé au cours de la thèse.

# Chapitre 1 :
# Problématique

## 1 Variété de langue à l'étude

Nous avons choisi d'étudier la langue orale telle qu'utilisée de manière spontanée en contexte de conversation. En accord avec Levinson (1983, p. 284), nous croyons que la conversation est le contexte linguistique humain prototypique. En conséquence, des résultats satisfaisants dans le domaine du traitement automatique de la parole ne pourront naître que par l'étude de la langue orale en contexte de conversation.

Les phénomènes comme la répétition, les amorces de mots, les contractions, les pauses et l'utilisation d'un lexique adapté aux besoins de l'oral ne sont pas des erreurs ou des dysfonctionnements, mais bien des mécanismes de communication adaptés à des contraintes particulières.

En accord avec Hansen (2005), nous considérons que les usagers d'une langue ne possèdent pas des lexiques distincts pour les contextes oraux et écrits, mais plutôt un large éventail de stratégies discursives qu'ils utilisent avec préférence selon les modes de communication. Chaque contexte d'utilisation a son influence sur le discours : on ne laisse pas un message sur un répondeur téléphonique de la même façon que l'on donne une conférence et les usagers d'un forum Internet n'écrivent pas de la même façon que lorsqu'ils envoient un texto ou un email à un ami. Toutes ces utilisations de la langue devraient cependant pouvoir être décrites et comprises avec les mêmes concepts et termes.

Les conditions matérielles et géographiques de la présente étude font de la variété de français parlé au Québec un objet d'analyse qui va de soi.

Le Corpus de français parlé au Québec (CFPQ), développé au Centre d'analyse et traitement informatique de la langue (CATIFQ) à l'Université de Sherbrooke puis au Centre de recherche interuniversitaire sur le français en usage au Québec (CRIFUQ), nous servira comme source de données empiriques (voir 6).

# 2 Pertinence de la recherche

En accord avec Hutchinson (2004), nous estimons qu'une description détaillée des aspects sémantiques et pragmatiques des MD représenterait une ressource utile pour le traitement automatique de la langue.

L'énoncé (3), tiré du CFPQ, exemplifie quelques-uns des phénomènes qui rendent pertinente notre recherche :

(3) VE : ah **mon dieu seigneur heille** on est enregistrées il faut parler des (.) sujets d'intérêt (.) [1<ff<général>>
*[CFPQ, sous-corpus 19, segment 6, page 59, ligne 4]*

Un programme d'analyse automatique mal versé dans l'analyse linguistique pourrait reconnaître les unités *dieu* et *seigneur* de l'extrait (3) et conclure que l'énoncé a comme thème « la religion ». Notre système aura comme objectif de repérer les unités ⌜*mon dieu*⌝, *seigneur* et *heille* dans ce type de contextes syntaxiques et de les caractériser adéquatement comme moyens utilisés par l'énonciateur pour s'exprimer au sujet du caractère hors du commun de quelque chose.

Grâce aux MI, la caractérisation des éléments de subjectivité peut être faite à trois niveaux :

1) L'analyse individuelle des MI nous donnera de l'information au sujet de chacun des contextes d'élocution dans lesquels ils se trouvent. Par exemple, il sera possible de déterminer si un énonciateur est étonné, en colère ou intéressé par certains éléments du contexte.

2) L'analyse collective des MI utilisés par chaque énonciateur nous donnera des informations sur son attitude ou son état d'esprit général au cours d'une conversation.

3) L'analyse globale des MI d'un texte nous donnera de l'information au sujet du contexte d'un échange sur les plans des registres de langue utilisés, du type de relation qu'entretiennent les participants (niveau d'informalité), de leur implication émotive, etc.

# 3 Objectifs de la recherche

Nous nous sommes fixé trois objectifs dans le cadre de cette thèse :
1. Améliorer les connaissances au sujet des MI.
2. Proposer une méthode d'analyse automatique des MI.
3. Démontrer la pertinence d'étudier les MI pour la fouille d'opinion.

De façon à réaliser ces objectifs, nous devrons analyser les MI tels qu'ils sont utilisés par les énonciateurs du CFPQ afin d'identifier les paramètres importants à considérer pour l'analyse automatique de ces unités. Nous comparerons des méthodes d'analyse automatique (avec des applications à l'aide de programmes informatiques) qui permettent de repérer ces MI ainsi que de mesurer et de caractériser les éléments de subjectivité qu'ils communiquent.

Plus précisément, pour chacun des objectifs, nous devrons réaliser les tâches suivantes :

1. Améliorer les connaissances au sujet des MI :
- passer en revue certaines études au sujet des MI et des MD;
- décrire précisément la classe des MI afin de pouvoir dresser une liste de MI présents dans le CFPQ et déterminer leurs fréquences;
- analyser chacune des occurrences des signifiants associés aux MI dans le CFPQ afin de déterminer si ces dernières relèvent ou non de la catégorie des MI.

2. Proposer une méthode d'analyse automatique des MI :

- passer en revue certaines études au sujet du traitement automatique de MI ou de MD;
- mettre au point et comparer différentes méthodes d'identification automatique des MI dans les textes.

3. Démontrer la pertinence d'étudier les MI pour la fouille d'opinion :
- passer en revue certaines études lexicographiques au sujet de MI;
- analyser le sens des unités que nous classerons parmi les MI;
- proposer un système de description sémantique des MI.
- exemplifier notre système d'analyse automatique des MI par l'étude de cas

# 4 Cadres théoriques

La lexicologie explicative et combinatoire issue de la Théorie Sens-Texte (TST) (entre autres, Mel'čuk, 1984, 1997; Iordanskaja et Mel'čuk, 1995) a proposé plusieurs outils et méthodes pour la description des unités lexicales. Sur le plan sémantique, ce cadre théorique favorise l'utilisation de la lexie comme unité de base de la description, une conception polysémique (plutôt que minimaliste ou maximaliste), ainsi qu'une description formelle effectuée à partir du langage naturel.

Les travaux de G. Dostie ont contribué à adapter ce cadre théorique à la description des MD, en tenant compte, entres autres, des présuppositions qui leurs sont associées, des composantes interactionnelles présentes dans leurs signifiés, des composantes déictiques (*je*, *tu*) et des actes de langage qui leurs sont liés. Dans le cadre de l'application lexicographique, tous ces éléments sont représentés par une paraphrase explicative lors de la description des unités lexicales.

La Métalangue sémantique naturelle (MSN), développée entre autres par A. Wierzbicka (1972, 1980, 1986, 1999, 2011) et C. Goddard (2013, 2014) a permis de décrire avec beaucoup de précision et de nuances certaines unités de champs lexicaux proches des MI. Alors que, par principe, la position de la TST est de ne pas décomposer les éléments des paraphrases, la position du MSN est de les décomposer le plus possible à l'aide de primitifs sémantiques. À cet égard,

notre thèse navigue entre deux eaux. Au cours de notre aventure lexicographique, nous respectons la règle de bloc maximal (Mel'čuk *et al.*, 1995), sauf dans certains cas où la décomposition des éléments des paraphrases permet de mettre en lumière des phénomènes intéressants, au sujet de liens sémantiques entre diverses unités, notamment.

Nous décrivons un grand nombre d'unités dans le cadre de cette étude et un traitement complet comme ceux réalisés dans le DEC (Dictionaire explicatif et combinatoire, Mel'čuk *et al.*, 1984; 1988; 1992; 1999), ou comme ceux réalisés à l'aide de la MSN aurait demandé beaucoup de temps et n'aurait pas particulièrement mieux servi nos objectifs. Nous avons donc dû adapter les choix méthodologiques de la TST et de la MSN à nos besoins.

La description de nos unités est orientée vers un but précis : permettre leur analyse automatique de façon à identifier les éléments de subjectivité communiqués par les énonciateurs qui y ont recours. Nous utilisons à cet effet un ensemble limité de courtes paraphrases, qui permettent, par leur combinaison, de caractériser un grand nombre d'unités avec peu de moyens (voir chapitre 4).

# 5 Démarche

Notre démarche constitue un aller-retour entre l'observation d'un corpus et la construction d'outils d'analyse de ce corpus. Elle est empiriste parce qu'elle se base sur des données tirées d'un corpus, mais elle est également rationaliste parce qu'appuyée sur notre compétence linguistique de locuteur natif.

L'étude de corpus à l'aide d'outils informatiques bénéficie d'une longue tradition académique (voir McEnery et Wilson, 2001 pour une introduction à ce sujet). Une multitude de corpus linguistiques à travers le monde, comme le corpus Brown, en langue anglaise, et ses successeurs, ont permis le développement d'outils statistiques que nous pouvons notamment exploiter grâce aux librairies informatiques décrites au chapitre 3.

Plus près de nous, à l'Université de Sherbrooke, la Banque de données textuelles de Sherbrooke (BDTS) (http://catfran.flsh.usherbrooke.ca/catifq/bdts/index.htm) accumule des textes écrits et favorise la production d'études sur le français du Québec depuis la fin des années soixante-dix. Nous avons personnellement exploité la BDTS pendant de nombreuses années dans le cadre du projet lexicographique FRANQUS (devenu *Usito*, https://www.usito.com/).

À la même université, le Corpus de français parlé au Québec (CFPQ) (Dostie, 2006-2015) permet la consultation en ligne de données tirées de l'oral. Le CFPQ est un corpus « multimodal qui intègre les trois dimensions caractéristiques d'une interaction verbale en face-à-face, à savoir ses dimensions verbale, paraverbale et gestuelle » (tiré du site web du CFPQ https://recherche.flsh.usherbrooke.ca/cfpq/). Réalisé entre 2006 et 2015, le corpus regroupe des transcriptions d'enregistrements atteignant plus de 680 000 mots. Le contexte de conversation libre propre à tous les enregistrements de ce corpus signifie que ceux-ci n'incluent pas de contextes « artificiels », comme ceux où un locuteur lit un texte ou utilise un registre de langue caractéristique d'une profession (par exemple, un journaliste à la radio).

Les enregistrements du CFPQ mettent en jeu trois à quatre locuteurs natifs du français du Québec qui discutent librement, laissant ainsi beaucoup de place à l'utilisation d'unités pragmatiques, comme les marqueurs illocutoires qui sont à l'étude.

Le corpus en question nous permettra d'avoir accès à une pluralité de discours, c'est-à-dire à des discours produits par des personnes de différents âges, sexes, bagages culturels, statuts sociaux-économiques, provenances régionales, etc., mais tous transcrits de manière homogène, avec une unité de présentation et un même métalangage.

L'utilisation de transcriptions de discours oraux plutôt que des enregistrements audio non transcrits facilitera grandement l'examen du corpus dans le contexte de l'analyse automatique. Le recours aux enregistrements ayant servi de matériel brut aux transcriptions sera toujours possible afin de vérifier la justesse de certains passages transcrits ou afin d'examiner des éléments qui

n'ont pu être notés (des informations de nature phonétique, prosodique ou gestuelle, par exemple).

Dans cette thèse, nous conserverons la notation des transcriptions du CFPQ pour la plupart des exemples qui y sont tirés (voir l'annexe, pour une présentation des conventions de transcription du corpus).

Nous discutons ici de quelques aspects problématiques inhérents à la linguistique de corpus.

### 5.1 Données limitées

Le problème fondamental des corpus oraux comme sources de données est qu'ils sont limités en taille. Nos observations sur les marqueurs seront ainsi limitées par leurs fréquences dans le corpus retenu. En pratique, l'analyse statistique des unités peu fréquentes devra en partie s'appuyer sur l'analyse d'autres unités similaires plus fréquentes.

### 5.2 Transcription des bandes audiovisuelles

L'enregistrement des fichiers audiovisuels et la transcription qui ont mené à la construction du CFPQ est le travail d'un groupe de recherche dirigé par la professeure Gaétane Dostie.

Chaque sous-corpus du CFPQ a été transcrit par des étudiants et des étudiantes et a fait l'objet d'une première vérification avec visionnement de la bande audiovisuelle et d'une deuxième vérification sans son visionnement.

La problématique de l’accord inter-juges (voir Bolly, Crible, Degand et Uygur-Distexhe, 2015 à ce sujet) concerne le fait que différentes personnes ont parfois des opinions différentes sur tel passage à transcrire ou telle unité à annoter. Il est par conséquent nécessaire d'utiliser des protocoles explicites pour ce type de tâches et de prendre des mesures afin d'assurer une unité de traitement des corpus par les personnes qui les manipulent.

Certains passages des enregistrements du CFPQ offrent plusieurs possibilités de transcriptions et se sont vus attribuer des multitranscriptions. Au cours de la préparation du corpus pour nos besoins, nous avons systématiquement remplacé ces doubles annotations par des marques de passages inaudibles (« inaud. »).

Nous avons comme principe d'utiliser les transcriptions du CFPQ sans les modifier, exception faite des modifications formelles présentées au chapitre 3. Nous avons enfreint cette règle dans un seul cas où la présence d'une pause, qui nous semblait d'une grande importance, n'était pas notée dans la transcription.

### 5.3 Annotation

Les marqueurs discursifs en général sont classés dans des catégories relativement perméables et il n'est pas rare d'observer qu'une même unité assume différents rôles. Aussi, dans certains énoncés du corpus, des unités peuvent être interprétées comme étant plus ou moins pragmaticalisées. Il nous a fallu attribuer une étiquette pour chacune des unités au meilleur de nos moyens et les choix que nous avons faits pourraient être sujets à discussion.

Nous avons consulté avec intérêt les travaux des chercheurs du projet MDMA (voir chapitre 2-2.1.5) qui tentent « d'établir une méthode empirique d'identification et d'annotation des MD en français oral. » (Bolly *et al.*, 2015). L'utilisation de ces techniques n'est cependant pas tout à fait compatible avec les objectifs de cette thèse.

L'étiquetage des unités du corpus tel que décrit au chapitre 3 a été faite par une seule personne (l'auteur de cette thèse) de manière semi-automatique. Comme il s'agit d'un corpus de plus de 680 000 mots contenant 6938 MI, le niveau de fiabilité de cet étiquetage n'est certainement pas de 100%. Nous avons porté une attention particulière à l'identification des MI, en vérifiant lorsqu'il le fallait les véritables conditions d'énonciation à l'aide des bandes audiovisuelles du CFPQ.

# 6 Présentation du CFPQ

Les 30 sous-corpus du CFPQ sont disponibles en format PDF à partir du site Internet du groupe de recherche (https://recherche.flsh.usherbrooke.ca/cfpq/). Nous avons converti et rassemblé ces fichiers dans un seul fichier texte de 74607 lignes, correspondant à autant de tours de parole. Le fichier est composé de 830 797 unités linguistiques (incluant des mots, des marques de silence, des marques de rire, des marques d'intonations et des marques de citations).

## 6.1 Les unités étudiées

Notre objectif est de prendre en compte l'ensemble des MI qui sont utilisés avec une fréquence de plus d'une occurrence dans le CFPQ. Certaines unités pouvant appartenir à cette catégorie ont sans doute échappé à notre vigilance.

Nous prenons 82 vocables de MI comme sujets d'étude. 15 de ces vocables constituent des regroupements de signifiants (voir tableau 2). Les 82 vocables sont répartis en 6938 occurrences dans le CFPQ, ce qui représente un pourcentage de 0,84% du total d'unités. L'ensemble des signifiants associés à ces vocables se répartissent en 9792 occurrences. Nous voyons donc que les MI ont tendance à être peu homonymes, puisque les signifiants auxquels ils sont associés sont MI dans 70,85% des cas.

Les signifiants de certains vocables qui sont des phrasèmes ont des morphèmes communs (comme ceux de VRAIMENT, ⌜PAS VRAIMENT⌝, et ⌜VRAIMENT PAS⌝). Les raisons pour lesquelles nous avons choisi de séparer ces unités en plusieurs vocables sont expliquées aux chapitres 3 et 4.

Le tableau 1 présente les vocables à l'étude, en ordre alphabétique.

*Tableau 1 : Fréquence des marqueurs illocutoires du CFPQ*

| Vocable | Nombre de MI | Nombre de signifiants | MI/Signifiants |
|---|---|---|---|
| AÏE | 7 | 7 | 100% |
| ⌜AÏE AÏE AÏE⌝ | 4 | 4 | 100% |
| ARK* | 63 | 63 | 100% |
| ARRÊTE* | 34 | 132 | 25,76% |
| AYOYE | 66 | 66 | 100% |
| BAPTÊME | 7 | 14 | 50% |
| BATEAU | 2 | 80 | 2,50% |
| BATINSE | 1 | 1 | 100% |
| BOF | 12 | 13 | 92,31% |
| ⌜C'EST ENCORE DRÔLE⌝ | 4 | 4 | 100% |
| CÂLIF | 4 | 5 | 80% |
| CÂLINE | 42 | 42 | 100% |
| CÂLIQUE | 18 | 20 | 90% |
| CÂLISSE | 15 | 22 | 68,18% |
| CALVAIRE | 15 | 16 | 93,75% |
| CALVINCE | 5 | 5 | 100% |
| CHUT | 16 | 16 | 100% |
| CIBOIRE | 14 | 14 | 100% |
| CIBOLE | 25 | 26 | 96,15% |
| COOL | 8 | 67 | 11,94% |
| COUDON | 25 | 25 | 100% |
| CRIF* | 54 | 62 | 87,10% |
| CRIME | 23 | 28 | 82,14% |
| CRISSE* | 126 | 180 | 70% |
| CRISTIE | 4 | 7 | 57,14% |
| ⌜DE LA MARDE⌝ | 7 | 36 | 19,44% |
| ⌜DU TOUT⌝ | 6 | 46 | 13,04% |
| ÉCOUTE* | 217 | 312 | 69,55% |
| ⌜EH BOY⌝* | 20 | 20 | 100% |
| ENVOYE | 42 | 45 | 93,33% |
| FIOU | 12 | 12 | 100% |
| FRANCHEMENT | 22 | 26 | 84,62% |
| GO | 15 | 15 | 100% |
| HEILLE | 1388 | 1388 | 100% |
| HEIN | 2092 | 2092 | 100% |
| ⌜JE COMPRENDS⌝ | 45 | 120 | 37,50% |
| ⌜LET'S GO⌝ | 2 | 2 | 100% |
| MALADE | 3 | 72 | 4,17% |
| MAUDIT | 16 | 54 | 29,63% |
| MAUTADIT | 3 | 9 | 33,33% |

| | | | |
|---|---|---|---|
| MERDE | 10 | 14 | 71,43% |
| ⌜METS-EN⌝ | 63 | 67 | 94,03% |
| ⌜MON DIEU⌝ | 177 | 177 | 100% |
| ⌜MON DOUX⌝ | 78 | 78 | 100% |
| ⌜MY GOD⌝ | 17 | 17 | 100% |
| OSTIE* | 408 | 500 | 81,60% |
| OSTIFIE | 9 | 12 | 75% |
| OSTINE | 2 | 2 | 100% |
| OSTIQUE | 8 | 11 | 72,73% |
| OUF | 53 | 53 | 100% |
| OUPELAILLE | 22 | 22 | 100% |
| OUPS | 38 | 38 | 100% |
| PANTOUTE | 22 | 90 | 24,44% |
| ⌜PAS DU TOUT⌝ | 23 | 36 | 63,89% |
| ⌜PAS VRAIMENT⌝ | 14 | 79 | 17,72% |
| ⌜POUR VRAI⌝* | 51 | 73 | 69,86% |
| REGARDE* | 609 | 803 | 75,84% |
| ⌜REGARDE DONC⌝ | 6 | 6 | 100% |
| SACRE | 5 | 15 | 33,33% |
| SACREMENT | 4 | 9 | 44,44% |
| SACRIFICE* | 7 | 11 | 63,64% |
| SEIGNEUR | 54 | 57 | 94,74% |
| SÉRIEUX | 67 | 90 | 74,44% |
| SIMONAQUE* | 9 | 10 | 90% |
| SUPER | 10 | 187 | 5,35% |
| TABARNACHE | 10 | 12 | 83,33% |
| TABARNAQUE* | 44 | 54 | 81,48% |
| TABARNIQUE | 3 | 4 | 75% |
| TABARNOUCHE | 24 | 37 | 64,86% |
| TELLEMENT | 4 | 395 | 1,01% |
| TIENS | 70 | 86 | 81,40% |
| TORIEU | 4 | 4 | 100% |
| ⌜UNE CHANCE⌝ | 40 | 49 | 81,63% |
| VIARGE | 3 | 3 | 100% |
| VOYONS | 251 | 251 | 100% |
| VRAIMENT | 77 | 1036 | 7,43% |
| ⌜VRAIMENT PAS⌝ | 11 | 89 | 12,36% |
| WÔ* | 38 | 38 | 100% |
| WOW | 83 | 83 | 100% |
| YOUPI | 2 | 2 | 100% |
| YÉ* | 22 | 22 | 100% |
| ZUT | 2 | 2 | 100% |
| Totaux | 6938 | 9792 | 70,85% |

Les vocables marqués par une étoile (*) sont des regroupements de signifiants.

Le tableau 2 présente les différents signifiants qui ont été rassemblés sous un même vocable. Certaines variations sont de nature flexionnelle (ÉCOUTE, REGARDE, ARRÊTE), d'autres sont de nature morphologique (⌜POUR VRAI⌝), phonologique (⌜EH BOY⌝, SACRIFICE) ou orthographique (OSTIE, CRIF). Dans le cas de ARK et de AÏE, il s'agit en réalité de regroupements de différents vocables similaires dont la prise en compte individuelle ne semblait pas utile étant donné leurs synonymies, leurs faibles fréquences et le fait que leurs signifiants ne sont pas homonymes.

*Tableau 2 : Vocables de MI qui regroupent plusieurs signifiants*

| **Type de variation** | **Vocable** | **Fréquence des signifiants** |
|---|---|---|
| Orthographique | CRISSE | 177 *crisse*, 3 *criss* |
| | OSTIE | 500 *ostie*, 2 *osti* |
| | CRIF | 60 *crif,* 1 *criff,* 1 *criffe* |
| | WÔ | 36 wô, 2 wo |
| | TABARNAQUE | 53 *tabarnaque*, 1 *tabarnak* |
| | SIMONAQUE | *10 simonaque*, 3 *simonac* |
| Phonologique | YÉ | 10 *yé*, 12 *yeah* |
| | ⌜EH BOY⌝ | 10 *eh boy*, 2 *ah boy*, 8 *oh boy* |
| | SACRIFICE | 10 *sacrifice*, 1 *sacréfice* |
| Morphologique | ⌜POUR VRAI⌝ | 58 *pour vrai*, 2 *pour le vrai*, 13 *pour de vrai* |
| Flexionnelle | ÉCOUTE | 304 *écoute*, 8 *écoutez* |
| | REGARDE | 795 *regarde*, 8 *regardez* |
| | ARRÊTE | 125 *arrête*, 7 *arrêtez* |
| Lexicale (regroupement de vocables) | AÏE | 5 *aïe*, 1 *ouille*, 1 *ouch* |
| | ARK | 40 *ark*, 4 *ouach*, 3 *ouache*, 5 *yark, 4 eurk*, 3 *yeurk*, 3 *beurk*, 1 *biark* |

Notons que plusieurs des marqueurs étudiés sont représentés par des graphies différentes dans d'autres textes que le CFPQ. Ces mots n'appartiennent en effet pas à un registre qui est sujet à une grande standardisation.

### 6.2 Fréquence des MI

Dans le tableau 1, nous pouvons voir que les vocables de MI les plus fréquents du CFPQ sont de types très différents. Les deux premières places sont occupées par des interjections primaires, HEIN (2092 occurrences) et HEILLE (1388). Suit REGARDE (609) qui est issu d'un verbe et OSTIE (408), issu d'un nom.

Nous remarquons que les sacres sont en général plus fréquents que leurs substituts. Le sacre OSTIE, par exemple, mène le bal avec ses 408 occurrences, tandis que ses substituts OSTIFIE (9 occurrences), OSTIQUE (8 occurrences) et OSTINE (2 occurrences) sont relativement peu nombreux.

### 6.3 Degré d'homonymie

Le degré d'homonymie d'un vocable du CFPQ correspond à la relation entre le nombre d'occurrences de ce vocable et le nombre de signifiants associés à ce vocable dans le corpus. Certains signifiants sont toujours utilisés comme MI dans le CFPQ, d'autres rarement.

Parmi les 82 vocables présentés plus haut, 33 n'ont pas d'homonymes dans le corpus, c'est-à-dire que leurs signifiants sont des MI dans 100% de leurs occurrences dans le corpus. Ces vocables ont un degré d'homonymie nul. Les 49 autres vocables ont des homonymes dans d'autres classes grammaticales.

Certains signifiants sont presque toujours des MI, comme *seigneur* qui est MI dans 94,74% des cas. D'autres, comme *vraiment* (1036 occurrences) et *tellement* (395 occurrences), sont assez fréquents, mais rarement utilisés comme MI : dans 7,43% des occurrences pour *vraiment* et dans 1,01% des occurrences pour *tellement*.

# Chapitre 2 :
# État de la question

Les objectifs de cette thèse font en sorte qu'un état de la question à son sujet doit faire appel à des travaux issus de champs d'études qui touchent à la sémantique, à la pragmatique, à la lexicographie, à la linguistique de corpus et au traitement automatique de la langue.

La première partie de ce chapitre nous permet de résumer les connaissances dont nous disposons au sujet de la classe des MI grâce à différents travaux du domaine de la linguistique, tandis que la seconde partie permet de présenter certains travaux au sujet de l'analyse automatique d'unités similaires aux MI.

## 1 Marqueurs illocutoires

Nous présentons ici un portrait partiel des connaissances disponibles dans la littérature au sujet des marqueurs illocutoires (MI).

### 1.1 Typologie des MI

Le terme marqueur illocutoire (MI) est celui proposé par G. Dostie dans la monographie *Pragmaticalisation et marqueurs discursifs*. Selon la typologie présentée dans cet ouvrage (Dostie, 2004), les MI font partie de la sous-classe des marqueurs discursifs (MD), eux-mêmes membres de la classe des marqueurs pragmatiques (MP). En plus des marqueurs discursifs, les marqueurs pragmatiques rassemblent les connecteurs textuels (comme ⌜EN EFFET⌝, ⌜PAR CONTRE⌝, AINSI...), qui permettent d'indiquer des relations entre des segments de textes.

Les marqueurs discursifs (MD) sont typiquement peu ou pas décrits par les entreprises lexicographiques et les grammaires traditionnelles. Par exemple, dans *Le Petit Robert 2017* (*Le Petit Robert*, 2016), le phrasème ⌜MON DIEU⌝ est bien décrit comme une interjection à

l'article DIEU, mais pas le phrasème ⌜MON CUL⌝, à l'article CUL. Dans le même dictionnaire, certaines utilisations discursives d'adverbes de phrase sont mentionnées, mais l'information à leur sujet est minimale (c'est le cas de l'utilisation « elliptique » de l'adverbe FRANCHEMENT, par exemple).

Chez les linguistes, la profusion des études des dernières décennies au sujet des MD a mené à une grande variété terminologique. Différentes sous-classes de MD ont été identifiées de manières isolées, ou regroupées selon différents critères (voir Hansen, 1997, p. 158 pour une critique des descriptions de Schiffrin, Fraser et Redeker).

Plus récemment, un groupe de chercheurs francophones ont proposé un modèle pour l'identification et l'annotation de MD en corpus (MDMA) qui prend en compte des paramètres syntaxiques, sémantico-pragmatiques, prosodiques et collocationnels (Bolly *et al.*, 2015; Crible et Zufferey, 2015; Crible, 2017). Nous ferons appel à plusieurs de ces paramètres au cours de nos analyses.

Nous considérons toutefois que le classement des MD gagne avant tout à être fait à partir de critères fonctionnels, plutôt que syntaxiques, morphologiques ou d'autres critères comme celui de la classe d'origine (voir Hansen, 1997, p. 155-156). La typologie et le classement présentés dans Dostie (2004) et reproduits au tableau 3 sont révélateurs au sujet du rôle des différents MD dans la communication.

*Tableau 3 : Typologie des marqueurs discursifs*

<table>
<tr><td rowspan="5">M. Pragmatiques (MP)</td><td rowspan="4">M. discursifs (MD)</td><td rowspan="2">M. illocutoires (MI)</td><td>M. d'interprétation</td></tr>
<tr><td>M. de réalisation d'un acte illocutoire</td></tr>
<tr><td rowspan="3">M. d'interaction</td><td>M. d'appel à l'écoute</td></tr>
<tr><td>M. d'écoute</td></tr>
<tr><td>M. de balisage</td></tr>
<tr><td colspan="3">Connecteurs textuels</td></tr>
</table>

Selon la typologie présentée dans ce tableau, les marqueurs d'interprétation sont des guides qui orientent l'interprétation d'un ou de plusieurs actes illocutoires qu'ils accompagnent. Dans l'extrait (4), par exemple, REGARDE introduit la proposition « on sait pas écrire » que l'énonciatrice tient pour vraie.

(4) [...] on est <f<les dirigeants de demain>> pis **regarde** on sait pas écrire [...]
*[CFPQ, sous-corpus 19, segment 3, page 31, ligne 17]*

Les marqueurs de réalisation d'un acte illocutoire, quant à eux, servent à réaliser des actes illocutoires expressifs, directifs ou assertifs; ils apparaissent particulièrement « indépendants » des points de vue pragmatique, sémantique et syntaxique. L'énoncé (5) où ⌜MON DIEU⌝ est utilisé seul présente un emploi typique de cette classe de marqueur.

(5) I : ah (*dit en inspirant bruyamment comme pour imiter la réaction de Michel*) **mon dieu:** (*en haussant les sourcils comme en signe d'étonnement*)
*[CFPQ, sous-corpus 30, segment 6, page 83, ligne 17]*

Les marqueurs d'appel à l'écoute servent à solliciter l'écoute du coénonciateur et à s'assurer de son maintien. En (6), l'énonciatrice utilise T'SAIS afin de signaler à ses coénonciatrices que son énonciation se poursuivra.

(6) MY : comme **t'sais** ma sœur est en secondaire un pis euh <dim<elle vient de commencer **t'sais** pi:s euh>> [...]
*[CFPQ, sous-corpus 19, segment 3, page 23, ligne 13]*

Les énonciateurs utilisent les marqueurs d'écoute afin de manifester leur écoute tout en indiquant souvent, par la même occasion, leur accord ou désaccord. Les marqueurs d'écoute ont la particularité de ne pas interrompre le tour de parole du coénonciateur. L'énonciatrice VI de

l'échange présenté en (7) utilise HUM et OUIN afin d'indiquer à sa coénonciatrice VE qu'elle porte attention à ce qu'elle dit.

(7) VE: [...] des fois des expressions ça (.) [2on peut en employer pis on les met en italique [...]
VI : [2**hum hum** (.) **ouin** (.) **ouin** c'est ÇA là
*[CFPQ, sous-corpus 19, segment 4, page 35, ligne 13-14]*

Les marqueurs de balisage servent à ponctuer le texte, à le découper en différentes séquences. En (8), par exemple, l'énonciatrice VE sépare les propositions qu'elle énonce à l'aide de LÀ.

(8) VE : mais pour les GARS **là**/ (.) il s'achète plein de linge **là** pis ça coûte: ri- lui **là** ça lui coûte rien s'habiller **là** [...]
*[CFPQ, sous-corpus 19, segment 1, page 7, ligne 20]*

En observant les exemples (4) à (8), on remarque que les unités discursives ont souvent des équivalents non-discursifs (*regarde*, *mon dieu*, *t'sais*, *là*). Il est également caractéristique des MD d'être l'objet d'une « érosion » phonologique (voir Denturk, 2008). Ainsi, *tu sais* est devenu *t'sais* et *écoute donc* est devenu *coudon*. Les MD constituent habituellement des unités prosodiques indépendantes (caractéristique fondamentale pour Zwicky, 1985), séparées par des pauses ou mises en relief par des intonations.

En général, les signifiés des MD sont le fruit de conventions précises qui permettent aux énonciateurs d'exprimer des idées complexes en remarquablement peu de moyens (peu de morphèmes). Une autre caractéristique des MD relevée par de nombreux chercheurs (par exemple, Hansen, 1998, p. 74; Dostie, 2004) est qu'ils ne contribuent typiquement pas au contenu propositionnel des énoncés auxquels ils sont joints et que leur présence ou absence ne modifie en général pas la valeur de vérité de ces énoncés.

Les MI tels que nous les définissons regroupent les marqueurs d'interprétation et les marqueurs de réalisation d'un acte illocutoire. Nous considérons qu'il est utile de considérer ces deux classes comme une seule classe d'unités qui peuvent s'utiliser de façons différentes; en lien avec une proposition dans le cas des marqueurs d'interprétation ou en lien avec un élément non nommé de la situation de conversation dans le cas des marqueurs de réalisation d'un acte illocutoire (voir chapitre 2-1.6).

Dans la littérature, le terme *interjection* est utilisé par plusieurs auteurs, notamment des auteurs anglophones (par exemple Goddard, 2013), pour décrire les MI ou une partie des MI.

En conclusion, les MI sont des marqueurs discursifs qui, contrairement aux marqueurs d'interaction qui réalisent des actes illocutoires liés à la gestion de la conversation, réalisent des actes illocutoires au premier plan de la conversation.

### 1.2 Les MI et l'analyse de sentiment

Les MI possèdent plusieurs caractéristiques intéressantes pour le champ d'application de l'analyse de sentiment. Premièrement, ils ont souvent un contenu sémantique fortement expressif. Ensuite, puisque les marqueurs discursifs sont habituellement utilisés hors des phrases, ils n'acceptent pas la négation, l'interrogation ou la modification (à l'aide de mots comme *très* ou *presque*, par exemple) (Iordanskaja et Mel'čuk, 1999, p. 5). Pour cette raison, les sentiments qu'ils expriment ont la propriété d'être toujours attribuables à leur énonciateur (sauf dans le cas des discours rapportés). Ainsi, parmi les trois énoncés suivants, seul le *cool* en (11) est un MI :

(9) C : ouais ça serait **cool** faire du go kart
*[CFPQ, sous-corpus 17, segment 5, page 67, ligne 20]*

(10) D : elle avait posé une question pas **cool** là
*[CFPQ, sous-corpus 3, segment 3, page 49, ligne 18]*

(11) C : <f<ah ouais↑>> **cool** on va bouger
*[CFPQ, sous-corpus 17, segment 1, page 6, ligne 11]*

En tant que MI, *cool* (tel qu'exemplifié en (11)) nous informe sur l'opinion de l'énonciateur indépendamment de son contexte d'énonciation. Les *cool* que l'on trouve en (9) et (10), au contraire, sont par nature en relation de dépendance avec d'autres éléments de la phrase. Le *cool* en (9) est modulé par le mode conditionnel et celui en (10) est modulé par la négation et le temps passé.

Les *cool* de (9) et (10) sont donc moins susceptibles d'être pertinents que celui de (11) pour connaître l'état psychologique de l'énonciateur. Cette relative indépendance des MI par rapport à leurs cotextes facilite leur analyse et est une raison importante de leur choix comme sujet d'étude dans cette thèse.

### 1.3 Unités d'origine des MI

Les MI sont le plus souvent issus, par un processus de pragmaticalisation (Dostie, 2004), d'autres classes grammaticales.

Goddard (2013) distingue trois sous-catégories d'interjections selon le critère de leur unité d'origine. Selon son système de classification, les interjections primaires peuvent être de type « son », comme *Ugh!* et *Psst!*, ou de type « mot », comme *Wow!* et *Yuck!*. Les interjections secondaires, comme *Shit!* et *Fuck!,* sont identiques, sur le plan du signifiant, à d'autres mots. Cette distinction nous apparaît pertinente dans la mesure où, d'un point de vue mécanique, notre système d'analyse automatique gagnera à distinguer les unités qui ont des signifiants identiques à des mots appartenant à d'autres classes grammaticales, comme *franchement* et *vraiment,* de celles qui n'en n'ont pas, comme *beurk* et *oups* (voir chapitre 3).

Les MI du français québécois sont issus de toutes les classes grammaticales traditionnelles. Une unité verbale peut se pragmaticaliser et devenir MI, comme c'est le cas pour ARRÊTE, DISONS,

ÉCOUTE et METTONS. Des adjectifs ont donné naissance aux MI COOL et MALADE, des noms ont produit MERDE et SEIGNEUR, des adverbes sont devenus FRANCHEMENT et TELLEMENT et des locutions se sont figés en phrasèmes, comme c'est le cas pour ⌜BIEN SÛR⌝ et ⌜DIS DONC⌝.

### 1.4 Actes illocutoires associés aux MI

Puisque la classe des MI est principalement conçue à partir de critères fonctionnels, les considérations pragmatiques à son sujet sont de première importance.

Le lexique des MI touche particulièrement à trois des catégories d'actes illocutoires telles que décrites par Searle (1979) : les actes expressifs, directifs et assertifs. Certains marqueurs peuvent servir à réaliser plus d'un acte, appartenant parfois à plus d'une catégorie. Notons que Goddard (2013) utilise les termes correspondants *Emotive*, *Volitive* et *Cognitive* pour décrire les types sémantiques des interjections.

En guise d'exemples, les MI qui réalisent des actes expressifs incluent WOW, ARK, ⌜MON DIEU⌝ et MERDE; ceux qui réalisent des actes directifs incluent ARRÊTE et ÉCOUTE et ceux qui réalisent des actes assertifs incluent METS-EN et ⌜JE COMPRENDS⌝. Nous discuterons amplement des actes illocutoires liés aux différents MI au chapitre 4.

### 1.5 Actants sémantiques des MI

La nature des actants sémantiques des MI est une information indispensable à connaître afin de comprendre les rôles que ceux-ci jouent dans la langue. En raison de leur nature extraphrastique, les MD en général font appel à un large éventail d'éléments de la situation de conversation comme actants sémantiques. La plupart des observations que nous présentons ici s'appliquent aux MD autant qu'aux MI. Nous discutons plus bas des différents éléments présentés dans le tableau 4.

*Tableau 4 : Éléments de la situation qui peuvent constituer des actants sémantiques des MI*

| | | |
|---|---|---|
| quelque chose | | |
| | quelqu'un | |
| | | je |
| | | tu |
| | ce que tu fais | |
| | | ce que tu dis |

### 1.5.1 (Quelque chose)

Dans le système de description sémantique proposé au chapitre 4, nous utilisons les mots *cela* ou *quelque chose* afin de représenter un actant sémantique non-déterminé qui est un élément de la situation.

Selon Hansen (2005), les MD ne marquent jamais un lien entre l'énoncé dans lequel ils se trouvent et le cotexte linguistique (contrairement aux connecteurs textuels), mais plutôt un lien entre cet énoncé et le « modèle de discours mental en construction ». Ce dernier contiendrait de l'information provenant, entres autres, d'énoncés précédents, mais aussi de l'information provenant du contexte non-linguistique ainsi que de connaissances encyclopédiques pertinentes. Différents marqueurs réfèrent à différents éléments du « modèle de discours mental ».

Dans le contexte d'une description sémantique générale à l'aide d'une paraphrase, il est difficile de nommer explicitement de tels éléments. Dans l'exemple (12) donné plus bas, ce *quelque chose* correspondrait à la fenêtre ouverte qui a été localisée par l'énonciatrice.

### 1.5.2 (Quelqu'un)

Les locuteurs qui participent à une conversation servent très souvent d'actants sémantiques aux MI. D'autres personnes, parfois imaginaires, peuvent également se voir adresser des MI comme si elles étaient des locutrices.

### 1.5.3 (Je)

Les MI ont la caractéristique de toujours exprimer le point de vue de l'énonciateur. Par conséquent, à l'intérieur de chaque MI, se cache un *je*. Parce qu'ils sont des mots-phrases, les MI encapsulent leur premier actant sémantique. Lorsqu'on paraphrase un MI afin de représenter son sens, nous devons utiliser un *je*. Les paraphrases des MI qui ont été proposées par Dostie dans le cadre du MST (2004, 2007; Dostie et Lanciault, 2016) et celles proposées par Goddard dans le cadre du MSN (2013, 2014) suivent cette règle. Cependant, comme le système de description sémantique que nous proposons dans le cadre de cette thèse (chapitre 4) ne tient pas compte du contenu sémantique qui est présupposé par les marqueurs, certaines des paraphrases que nous utilisons ne contiennent pas de *je*.

### 1.5.4 (Tu)

Certains MI ne peuvent être produits qu'en s'adressant à un ou plusieurs coénonciateurs et possèdent donc en plus un *tu/vous* comme actant.

L'analyse de l'extrait (12) permet de mieux comprendre le rôle des actants *je* et *tu* dans le signifié d'un emploi particulier du MI TIENS. L'énonciatrice F y utilise le marqueur TIENS afin d'attirer l'attention de son coénonciateur au sujet d'une fenêtre ouverte.

(12) F : on est bien **tiens** regarde (*en se retournant vers la fenêtre*) on a une belle petite brise ici là
*[CFPQ, sous-corpus 18, segment 6, page 58, ligne 3]*

Cette lexie a été paraphrasée par Dostie (2004, p. 224) dans les termes suivant :

I.2 *Tiens* ≅
Ayant repéré quelque chose //
j'attire ton attention sur le fait que je localise quelque chose dans l'espace qu'il t'est aussi possible de localiser (en suivant mon geste et / ou mon regard).

Le fait que le signifiant de cette lexie devient *tenez* lorsqu'elle est adressée à plusieurs coénonciateurs est révélateur de la présence d'un actant *tu*/*vous* dans son signifié. Les MI qui possèdent des actants *tu*/*vous* sont souvent issus de verbes à la forme impérative (ex : *regarde*, *écoute*), mais pas nécessairement (ex : *sérieux*). Les actants *tu*/*vous* ne sont en général pas représentés par des unités linguistiques dans le discours. Lorsqu'ils le sont, on a affaire à la construction « marqueur + SN vocatif », décrite plus bas.

### 1.5.5 ⟨Ce que tu fais⟩

Certains MI ne sont produits qu'au sujet de gestes d'un interlocuteur. Des MI directifs comme ARRÊTE et ENVOYE mettent nécessairement en jeu l'actant « ce que tu fais ».

Dans l'extrait (13), l'énonciateur R produit ENVOYE afin d'inciter Manon à éteindre une caméra vidéo.

(13) R : **envoye/** [1pèse sur le piton Manon
*[CFPQ, sous-corpus 15, segment 10, page 174, ligne 13]*

Dans un contexte conversationnel, l'actant « ce que tu fais » réfère le plus souvent à l'action de parler. Nous pouvons alors le caractériser plus précisément par la paraphrase « ce que tu dis ».

### 1.5.6 (Ce que tu dis)

L'actant sémantique « ce que tu dis » est parfois soumis à des restrictions. Par exemple, le marqueur VOYONS de l'énoncé (14), tel que décrit dans Dostie (2004, p. 231), porte obligatoirement sur des propos ou des comportements qui se sont produits auparavant.

(14) V : ouin (.) mais avant que j'aille danser nue euh: s:- va falloir qu'on me donne cher là
[...]
M : **voyons** Virginie je te pensais plus open
*[CFPQ, sous-corpus 19, segment 1, page 5, ligne 8-9]*

Notons que les marqueurs d'interaction (brièvement discutés en 1.1) semblent principalement mettre en jeu l'actant sémantique « ce que tu dis ». Le recours à l'actant « ce que tu dis » par certains MI est un des phénomènes qui fait en sorte que la frontière parait floue entre la classe des marqueurs illocutoires et celle des marqueurs d'interaction.

## 1.6 Comportement syntaxique des MI

Une analyse du comportement syntaxique des MI nous aidera à concevoir les systèmes informatiques chargés d'identifier ces unités dans les textes.

Typiquement, les MI se retrouvent sous la forme de mots-phrases, à l'extérieur des structures prédicatives. Très rarement, certains marqueurs s'éloignent de ce comportement syntaxique prototypique et se rattachent à des propositions ou à des syntagmes nominaux.

### 1.6.1 Mots-phrases

Les MI ont la particularité d'être le plus souvent employés de manière isolée d'un point de vue syntaxique. Ils peuvent alors se déplacer plus ou moins librement avant ou après les autres syntagmes. Comme plusieurs autres types de marqueurs pragmatiques, ils semblent ainsi appartenir à la macrosyntaxe du discours en contribuant à l'enchaînement des syntagmes et des

énoncés. Dans l'énoncé (15), par exemple, le mot *ostie* aurait très bien pu être placé à la fin de l'énoncé, comme au début ou avant le *pis*.

(15) ils manifestaient <all<pacifiquement>> pis **ostie** on (.) on leur tire dessus t'sais
*[CFPQ, sous-corpus 9, segment 6, page 81, ligne 11]*

Puisque, dans la plupart des cas, les MI sont utilisés comme mots-phrases, les éléments de la situation sur lesquels ils portent ne sont pas souvent explicités avec précision dans le discours. Les coénonciateurs doivent interpréter l'intention communicative de l'énonciateur à l'aide de l'information qu'ils possèdent sur le contexte non-linguistique. Afin d'éviter des mésinterprétations, les énonciateurs accompagnent souvent les MI qu'ils produisent de phrases qui servent à préciser l'élément de la situation sur lequel ces marqueurs portent. En (14), par exemple, la phrase « je te pensais plus open » offre aux coénonciateurs une piste d'interprétation du mot *voyons* et évite que celui-ci soit mal interprété. Si certains MI sont le plus souvent utilisés de manière isolée, d'autres sont le plus souvent accompagnés de phrases explicatives. Dans plusieurs cas, une même lexie peut être utilisée autant comme marqueur de réalisation d'acte illocutoire que comme marqueur d'interprétation.

### 1.6.2 MI liés à d'autres syntagmes

Exceptionnellement, les MI sont liés grammaticalement à des propositions ou à des syntagmes nominaux par des phénomènes examinés plus bas.

#### 1.6.2.1 MI + *si* P

Certains MI se retrouvent parfois sous la forme « marqueur + *si P* ». Cette construction est mentionnée dans Dostie (2004) et exemplifiée par l'extrait (16).

(16) *Tu parles* (si je lui ai parlé)! Plus fermé que ça, tu meurs!
*(Dostie, 2004, p. 47)*

Nous retrouvons un exemple de cette construction dans le CFPQ, reproduit en (17), où le second marqueur *voyons* est lié à la proposition « si ça a du bon sens ».

(17) S : <all<là Mario lui quand il a il a parti ben imagine-toi ça ça là ça ça le déprimait là vu que (.) partir en autobus voyons donc (.) tu peux pas partir t'en aller en: autobus **voyons si ça a du bon sens**>>
*[CFPQ, sous-corpus 5, segment 10, page 108, ligne 12]*

#### 1.6.2.2 MI + *avec* SN

L'extrait (18) semble exemplifier une utilisation de ARRÊTE lié à un syntagme nominal par la préposition *avec*.

(18) ME : hum (.) la prof d'E.C.C.
MA : **arrête avec** la prof d'E.C.C. là c'est la fin de semaine (*dit en riant*) (RIRE)
*[CFPQ, sous-corpus 3, segment 5, page 82, ligne 16]*

En (18), l'énonciatrice MA produit ARRÊTE afin d'inciter sa coénonciatrice ME d'arrêter de parler d'un sujet en particulier. Cet emploi ressemble à la construction « arrête avec tes conneries », attesté en français d'Europe. Nous croyons que l'énoncé « arrête avec la prof d'E.C.C. » peut être vu comme une contraction d'une phrase comme « arrête de nous embêter avec la prof d'E.C.C. ». Le processus de pragmaticalisation de *arrête* dans ce contexte ne semble pas tout à fait complété et rappelle un usage intraphrastique du verbe *arrêter* à l'impératif.

#### 1.6.2.3 MI + SN vocatif

Presque tous les MI peuvent être accompagnés d'un syntagme nominal « vocatif », c'est-à-dire qui désigne une personne à laquelle l'énonciateur s'adresse. Nous avons vu cette structure avec *voyons* dans l'énoncé (14) plus haut. Dans l'énoncé (19), le SN « Lynda » est lié à *heille* de cette façon :

(19) ah:: **heille** Lynda tu me don- passerais-tu le lait↑
*[CFPQ, sous-corpus 18, segment 4, page 48, ligne 7]*

La plupart des phrases peuvent être accompagnées d'un tel syntagme appellatif et ce phénomène n'est pas particulier aux marqueurs illocutoires.

### 1.6.3 Constructions intraphrastiques non-MI

Les signifiants de plusieurs MI peuvent être utilisés dans des constructions intraphrastiques qui sont très similaires à leur usage comme MI. Nous mentionnons ici certaines de ces constructions afin d'éviter la confusion à leur sujet.

#### 1.6.3.1 Connecteur textuel + *que* P

Certains MI semblent pouvoir se combiner avec des propositions à l'aide de la conjonction *que*. Nous considérons que les MI prennent la forme de connecteurs textuels dans de telles constructions. Dans l'extrait (20), il semble que le MI METS-EN soit lié à la proposition « elle me fait peur » à l'aide de la préposition *que*.

(20) D : **mets-en qu'elle me fait peur *[...]***
*[CFPQ, sous-corpus 3, segment 1, page 3, ligne 5]*

Cette construction semble liée à la réalisation d'actes illocutoires assertifs. En effet, des phrasèmes comme « c'est clair » ou « bien sûr », qui peuvent être utilisés seuls, à la manière de MI affirmatifs, peuvent aussi être utilisés comme connecteurs textuels. L'exemple construit (21) est destiné à montrer la similitude entre le comportement de *mets-en* de l'énoncé (20) et celui de *bien sûr* dans ce contexte.

(21) bien sûr qu'elle me fait peur
*[Exemple construit]*

Nous considérons que le phrasème *mets-en* est suffisamment pragmaticalisé dans la plupart de ses emplois pour atteindre le statut de MI lorsque utilisé seul (contrairement à *bien sûr*, *c'est clair*, *c'est sûr*...).

#### 1.6.3.2 Intensieur phrastique + *que* P

La construction vue plus haut n'est pas à confondre avec l'utilisation des sacres sous la forme d'« intensifieurs phrastiques » tel que discuté par Dostie (2015, p. 20) Le signifiant *ostie* de l'extrait (22) est un exemple d'un tel intensifieur phrastique.

(22) **ostie** qu'elle me fait rire AH:
*[CFPQ, sous-corpus 19, segment 7, page 67, ligne 16]*

Dostie 2015 note qu'un énoncé tel que (22) peut être reformulé de manière à mettre en évidence sa similitude avec d'autres constructions intraphrastiques. Par exemple: *ostie qu'elle me fait rire* ≅ *elle me fait rire en ostie*. Une telle reformulation serait impossible au sujet de l'énoncé (20).

#### 1.6.3.3 Nom de qualité + *de* SN

Les sacres peuvent également être utilisés comme noms de qualité dans des constructions à ne pas confondre avec les MI (voir Dostie, 2015, p. 61 à ce sujet).

Dans l'extrait (23), le SN « Virginie » ne réfère pas à une coénonciatrice, mais à une individue qualifiée par le nom de qualité *ostie*.

(23) **ostie** *de Virginie hein elle a tellement pas changé là-dessus*
*[CFPQ, sous-corpus 19, segment 7, page 63, ligne 5]*

### 1.7 Morphologie des MI

La morphologie des MI n'est pas une question complexe. Les MI sont figés ou, parfois, semi-figés sur le plan morphologique. Le seul contexte où l'on observe une variation morphologique chez les MI est celui où il existe une alternance entre la deuxième personne du singulier d'un verbe et sa deuxième personne du pluriel en contexte de vouvoiement ou de pluriel. Dostie (2004) discute, entre autres, des cas de *regarde*2b et ⌜*dis donc*⌝4, comme en (24) :

(24) [B voit A qui arrive:] Eh ben, *dites donc*! de la grande visite! Bienvenue, madame!
(Dostie, 2004, p. 209)

La lexie *tiens*I.3 présente dans l'énoncé (12) constitue également un exemple où il y a variation morphologique, puisqu'elle peut (de façon optionnelle) se trouver sous la forme [tənе] dans les contextes où il y a plusieurs coénonciateurs ou en contexte de vouvoiement.

Dans le CFPQ, nous observons une telle variation morphologique au sujet des MI REGARDE, ÉCOUTE et ARRÊTE.

### 1.8 Prosodie des MI

Les éléments prosodiques qui accompagnent les MI sont importants à prendre en compte tant pour le repérage de ces unités que pour leur interprétation. Lorsqu'une personne lit des transcriptions de conversations (du CFPQ, par exemple), l'interprétation correcte de certains MI nécessite d'avoir de l'information prosodique à leur sujet. Une intonation montante ou descendante est parfois un paramètre qui permet de distinguer deux lexies d'un même vocable aux sens très différents.

Les MI constituent typiquement des unités indépendantes d'un point de vue prosodique. Ils sont souvent précédés et/ou suivis de pauses ou de micro-pauses. L'énoncé (25) constitue un exemple typique où *wô là* se trouve au début d'un tour de parole et est suivi d'une pause significative (notée par le point entre parenthèses). Le marqueur est donc isolé d'un point de vue prosodique.

(25) F : <p<**wô là** (.) peut-être pas là>>
*[CFPQ, sous-corpus 9, segment 5, page 69, ligne 9]*

Des intonations marquées accompagnent souvent certains MI, parfois de manière quasi-systématique. Par exemple, les lexies *coudon***5** et *coudon***6** (numérotation de Dostie, 2004), qui ont une composante d'inévitabilité dans leurs signifiés, sont habituellement produites avec une intonation descendante, tandis que la lexie *coudon***4**, qui n'a pas une telle composante, est produite avec une intonation montante. Les intonations ne sont pas transcrites systématiquement dans le CFPQ, mais elles sont notées lorsque perçues comme particulièrement significatives.

L'allongement de syllabes semble aussi être un paramètre pouvant distinguer certaines lexies. Dans les deux énoncés suivants, on voit que les deux *heille* ne jouent pas le même rôle. Le *heille* de (26), qui est accentué et bien allongé, semble réaliser un acte directif (même sans contexte), tandis que celui de (27) semble, entre autres, souligner le caractère intense du contenu propositionnel de l'énoncé :

(26) **HEI:::LLE** t'es mieux de pas le dire
*[CFPQ, sous-corpus 9, segment 3, page 32, ligne 3]*

(27) <p<**heille** c'était super>>
*[CFPQ, sous-corpus 9, segment 2, page 19, ligne 7]*

### 1.9 Combinatoire des MI

Les MI peuvent se combiner entre eux ou avec d'autres unités selon des règles précises. Afin de dresser un portrait adéquat de chacun des marqueurs, il nous faudra déterminer dans quelle mesure ils peuvent se combiner avec d'autres unités et quelles conséquences cette combinatoire a sur leur sens. Dostie (2013) décrit les différents types d'associations des MD, à savoir la cooccurrence discursive libre, la locution discursive et la collocation discursive.

### 1.9.1 Cooccurrence discursive libre

Parmi les différents types d'associations des MI, on trouve d'abord la cooccurrence libre, où les marqueurs se suivent dans le texte sans que leurs sens ne soient modifiés par cette juxtaposition. Chacun des marqueurs conserve son sens, comme c'est le cas pour ÉCOUTE, HEILLE et FRANCHEMENT en (28) qui se suivent librement :

(28) ben c'est ça **éCOUte heille franchement**
*[CFPQ, sous-corpus 19, segment 3, page 24, ligne 8]*

### 1.9.2 Locution discursive

Certains MI peuvent être considérés comme des locutions issues de l'assemblage de plusieurs mots-formes. Les mots-formes ainsi assemblés peuvent correspondre à des MD (⌜DIS DONC⌝, *COUDON* (*écoute* + *donc*)) ou être des unités issues d'autres classes grammaticales (⌜*JE COMPRENDS*⌝). Les phrasèmes ainsi formés sont des unités lexicales à part entière.

Quelques marqueurs pragmatiques peuvent se dupliquer (Dostie, 2007) et créer de nouvelles lexies, comme *ok* ou *là*. Il s'agit d'un type particulier de locution. Dans l'énoncé (29), par exemple, la duplication de *là* est significative, puisque *là là* véhicule l'idée d'immédiateté, ce qui n'est pas nécessairement le cas de *là* employé seul.

(29) L : fait que finalement ils se sont organisés puis j'ai récupéré mon poste mais à [1trente heures
F : [1eh::
J : (RIRE)
L : fait que là j'ai un: trente heures [1officiel là
F : [1au lieu d'un trente-cinq
L : au lieu d'un (.) ben je veux pas avoir un trente-cinq **là là** pas tout de suite ben en tout cas (.) je [1 suis bien à trente heures hein/
*[CFPQ, sous-corpus 18, segment 5, page 51, ligne 14]*

### 1.9.3 Collocation discursive

À mi-chemin entre les cooccurrences libres et les locutions, on trouve des constructions semi-figées qu'on peut appeler « collocations ». Selon Dostie (2013), cette forme d'association se caractérise par la présence de deux unités où l'une d'elle, le marqueur-tête, est sémantiquement autonome et sélectionne une autre unité, le marqueur collocatif, qui ne l'est pas dans ce contexte.

Certains marqueurs collocatifs se combinent avec un grand nombre de marqueurs-tête, tel que *là* en (30) avec *heille* et avec *franchement* en (31).

(30) **heille là**
*[CFPQ, sous-corpus 2, segment 2, page 20, ligne 18]*

(31) J-M : ouais mais là c'est du gaspillage là avoue hein avoue **franchement là** t'sais c'est c'est un PLAIsir là t'as
*[CFPQ, sous-corpus 10, segment 7, page 91, ligne 7]*

D'autres marqueurs collocatifs ont une combinatoire plus limitée. C'est le cas de *donc*, qui se combine avec certains marqueurs-têtes comme *voyons* en (32), mais non avec d'autres comme *mets-en* en (33).

(32) ben **voyons donc**
*[CFPQ, sous-corpus 7, segment 3, page 36, ligne 4]*

(33) *mets-en donc

Toutes ces possibilités de combinatoire des MI devront être prises en compte lors de l'élaboration des méthodes d'identification des MI présentées au chapitre 3.

### 1.10 Conclusion au sujet des MI

Les MI en tant que classe ont été relativement peu investigués jusqu'à présent. Nous avons présenté quelques-unes de leurs caractéristiques pragmatiques, sémantiques et syntaxiques à leur sujet. Une réelle compréhension de cette classe ne pourra naître que par l'analyse individuelle d'un grand nombre de marqueurs lui appartenant. Notre étude se veut un pas dans cette direction.

Nous avons établi que les MI réalisent, dans la plupart des cas, des actes illocutoires expressifs et parfois des actes directifs ou assertifs. Pour ce qui est de leurs actants sémantiques, nous savons que les MI renvoient toujours à un énonciateur (un *je*) comme actant sémantique incorporé et peuvent également renvoyer à un ou des coénonciateurs (un *tu*/*vous*). Ils mettent souvent en jeu des éléments du contexte, représentés par des textes ou non. Sur le plan de la syntaxe, les MI sont autonomes et sur le plan de la morphologie, ils sont le plus souvent invariables. Nous avons vu également que la prosodie est un paramètre important du signifiant des MI. Enfin, nous avons présenté le phénomène de collocation discursive, caractéristique de ce type d'unités.

## 2 Traitement automatique des MI

À notre connaissance, il n'existe pas de travaux qui ont spécifiquement été menés au sujet de l'analyse automatique des MI. Des marqueurs similaires, de la grande famille des marqueurs pragmatiques (MP), ont cependant été traités dans plusieurs travaux d'analyse automatique de discours écrits et oraux. Les MP en général sont typiquement vus comme des outils révélateurs

des structures discursives qui peuvent aider à segmenter les discours en unités de sens et à identifier les relations qu'elles ont entre elles (par exemple Marcu, 1997, 2000). Les MI ne sont pas les unités les plus utiles pour ces tâches, ce qui peut expliquer qu'ils sont généralement moins pris en compte que les connecteurs textuels.

Les recherches que nous avons choisies de présenter ici touchent, directement ou indirectement, à au moins un thème parmi les deux suivants : l'identification automatique des MP et l'analyse sémantique automatique des MP.

## 2.1. Identification automatique des MP

Les recherches suivantes offrent plusieurs pistes intéressantes au sujet de l'identification automatique de certains MP de l'anglais et du français.

### 2.1.1 Hirschberg et Litman (1987, 1993) et Litman (1996)

Hirschberg et Litman (1987, 1993) sont, à notre connaissance, les premières à s'être penché sur la question de l'analyse automatique des « cue phrases » en anglais (par exemples, *okay*, *but*, *now*, *anyway*, *by the way*, *in any case* et *that reminds me*). Leur point de départ est l'analyse automatique de la prosodie. Elles établissent un système de règles qui tient compte de l'isolation syntaxique des unités, des intonations hautes et basses et des accentuations, afin de modéliser la façon dont les énonciateurs et les coénonciateurs distinguent les usages discursifs et les usages phrastiques de ce type d'unités. Litman (1996) continue cette démarche par l'utilisation de l'apprentissage machine dans le but de raffiner les règles établies manuellement. Ces recherches indiquent que la position des MP cibles dans les énoncés, la durée de leur prononciation et la présence d'autres MP en pré-position sont les paramètres les plus importants pour leur repérage automatique.

### 2.1.2 Heeman, Byron et Allen (1998), Heeman et Allen (1999), Heeman (1997)

Heeman, Byron et Allen (1998) ont des objectifs similaires à Hirschberg et Litman, mais ils utilisent un système d'apprentissage machine différent qui inclut l'étiquetage systématique des classes grammaticales dans le corpus d'entraînement ainsi que la prise en compte du signal acoustique. Les *discourse markers* qu'ils étudient se classent parmi les différentes sortes de marqueurs pragmatiques selon la typologie présentée en 1. Les chercheurs observent que ces MP peuvent être utilisés efficacement afin d'identifier les structures conversationnelles (par exemple, les énoncés qui résument de l'information ont une forte probabilité d'être introduits par le marqueur *so*) (Heeman *et al.*, 1998, p. 6). Ils remarquent aussi cependant qu'il n'y a pas une forte corrélation entre les unités qu'ils étudient et les actes de langage des énoncés dans lesquels ils se trouvent, à l'exception des paires d'actes illocutoires comme les couples salutation/salutation, question/réponse ou confirmation/infirmation. Cet état de choses s'explique, à notre avis, par le choix des unités étudiées, qui sont le plus souvent des connecteurs textuels ou des marqueurs illocutoires d'interprétation. Nous croyons qu'il existe une corrélation plus forte entre les MI et les actes de langages réalisés par les énoncés dans lesquels ils se trouvent, qu'entre ces derniers et les connecteurs textuels.

Dans Heeman et Allen (1999), les auteurs raffinent leur système par la prise en compte du phénomène propre au discours oral de « speech repairs », par lequel un énonciateur interrompt ce qu'il dit et le redit de façon différente, comme s'il retournait en arrière dans la phrase afin de corriger une erreur. L'utilisation des MP est fortement liée à ce phénomène, puisque plusieurs d'entre eux servent à introduire de tels segments de réparation. Lorsque leur système se penche sur des transcriptions humaines de discours oraux, il est en mesure de repérer 97,3% des MP étudiés avec une précision de 96,3% (Heeman et Allen, 1999, p. 41).

Heeman (1997) a proposé un ensemble d'étiquettes qui a été modifié afin d'inclure des étiquettes pour les différents types de MP et des informations additionnelles au sujet de la syntaxe avec comme objectifs principaux la segmentation en énoncés et la détection de MP dans des transcriptions de discours oraux. Nous effectuons un travail similaire au chapitre 3.

### 2.1.3 Popescu-Belis et Zufferey (2004, 2011)

Popescu-Belis et Zufferey (2004, 2011) inventorient les principales études qui traitent de l'identification automatique des MD et proposent leur propre méthode d'analyse automatique, basée sur l'apprentissage machine et sur des arbres à décision. Ils ont mesuré l'importance de différents paramètres pour le repérage et la désambiguïsation automatique des unités *like* et *well* en anglais, comme les collocations lexicales, les propriétés prosodiques, la position des marqueurs dans l'énoncé ainsi qu'un ensemble de propriétés sociolinguistiques.

Leurs expérimentations ont montré que le mot immédiatement avant les MD *like* et *well* (ou son absence) semble être le critère le plus pertinent pour l'identification des ces unités (Popescu-Belis et Zufferey, 2011, p. 13). Le premier mot qui suit ces unités serait également un critère pertinent, mais dans une moindre mesure. Enfin, les deuxième et troisième mots avant et après les unités cibles ne seraient pas d'une grande utilité pour la tâche de désambiguïsation.

Leur étude a aussi démontré que le système avait de meilleures performances lorsque les unités *well* et *like* étaient étiquetées différemment (plutôt que comme une classe unique) dans les corpus d'apprentissage, ce qui indique qu'elles ont des comportements syntaxiques différents (Popescu-Belis et Zufferey, 2011, p. 18).

Par la prise en compte de paramètres non-linguistiques au sujet des énonciateurs de leur corpus (genre, âge, niveau d'éducation, compétence linguistique et région d'origine), l'étude suggère également que certains groupes de locuteurs sont plus susceptibles d'utiliser certains MD. Par exemple, les habitants de la côte ouest des États-Unis auraient tendance à utiliser davantage le marqueur *like* que ceux de la côte est (Popescu-Belis et Zufferey, 2011, p. 9). Ces facteurs non-linguistiques sont évidemment d'une importance moins grande pour un système d'identification automatique que d'autres facteurs, comme l'environnement lexical et la prosodie.

### 2.1.4 Petukhova, Geertzen et Bunt (2007) et Petukhova et Bunt (2009)

Les articles de Petukhova, Geertzen et Bunt (2007) et Petukhova et Bunt (2009) offrent beaucoup d'information sur plusieurs aspects du repérage des MD. Par exemple, les auteurs remarquent que les MD sont souvent entourés d'unités distinctives comme *um*, *uh*, *so*, *then* et *also* (Petukhova et Bunt, 2009, p. 165).

Les questions liées à la prosodie sont également traitées, avec des conclusions similaires à celles de Hirschberg et Litman. Ainsi, les MD étudiés ont une durée de production presque deux fois plus longue que leurs versions phrastiques, ils sont précédés de pauses plus longues (entre 59 et 228 millisecondes) et ont une tonalité moyenne plus élevée (>12Hz) (Petukhova et Bunt, 2009, p. 165).

### 2.1.5 Bolly *et al.* (2015), Crible et Zufferey (2015), Crible (2017)

Selon ses concepteurs (Bolly *et al.*, 2015; Crible et Zufferey, 2015; Crible, 2017), le projet MDMA (Model for Discourse Marker Annotation) cherche à « développer une méthode empirique inédite pour l'identification et l'annotation systématique – potentiellement automatisable – des MD en contexte. » (Bolly *et al.*, 2015, p. 3). Les chercheurs utilisent une longue liste de paramètres afin de caractériser les MD qu'ils repèrent dans des échantillons de corpus oraux du français.

Parmi les nombreuses observations sur les caractéristiques de l'utilisation des MD que font les auteurs, l'une d'elles semble particulièrement intéressante pour notre propos :

> [Certaines variables comme] 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 (voir Vincent, 1993 ; Maschler, 2002).
> (Bolly *et al.*, 2015, p. 23)

Nous aurons amplement l'occasion d'observer ce phénomène où des MI sont utilisés sans pauses perceptibles à la suite d'une phrase. La prise en compte de ces contextes fait partie des nombreux défis qu'il nous faudra relever afin de mettre au point une méthode de détection automatique des MI.

## 2.2 Analyse sémantique automatique des MD

Les études suivantes nous offriront des pistes de solutions à certains problèmes relevés au chapitre 4 de cette thèse.

### 2.2.1 Hutchinson (2004)

Hutchinson (2004) utilise des techniques d'apprentissage machine afin de recueillir des données permettant de caractériser certains MP (surtout des connecteurs textuels) de l'anglais selon trois paramètres liés à la nature des propositions introduites ou mises en relation par ces marqueurs. Les unités étudiées sont caractérisées selon leur *polarité* négative ou positive, en fonction de leur capacité à introduire un élément de concession, de contraste ou de négation (par exemple, *yet*, '*even tough'*) ou celle à introduire un élément de confirmation (par exemple, *and*, *whenever*). Les unités sont également caractérisées selon leur *véridicité*, dans les cas où elles impliquent la vérité des arguments qu'elles mettent en relation (par exemple, *whereas*, '*seeing as'*) ou non (par exemple, *if*, '*supposing that'*). Enfin, les unités sont caractérisées selon leur *type*, qui peut être *causal*, *temporel* ou *additif*. Ce modèle de description des MP, qui reflète une tendance substantielle dans la littérature selon Hutchinson (2004, p. 1), ne peut pas être directement appliqué aux MI puisque ceux-ci mettent rarement en relation plusieurs propositions, contrairement aux connecteurs textuels. Les MI qui réalisent des actes assertifs (comme ⌜JE COMPRENDS⌝) pourraient cependant être décrits à l'aide de critères similaires.

### 2.2.2 Petukhova *et al.* (2007) et Petukhova et Bunt (2009)

Petukhova et Bunt (2009) utilisent des techniques d'apprentissage machine afin de démontrer le multi-fonctionnalisme de certains MD. Les auteurs utilisent une liste de « fonctions

communicatives » tirée de la Dynamic Interpretation Theory (DIT, développée entre autres par Bunt, 2000) afin de décrire les rôles des unités en discours. Cette description sémantique des MD est un aspect très intéressant pour nous. L'étude démontre l'importance des « actes communicatifs » dans l'analyse des MD. La taxinomie de la DIT distingue 11 « dimensions » dans lesquelles peuvent être classés les éléments du discours. Par exemple, certaines unités linguistiques permettent la transmission d'information, d'autres permettent la gestion des tours de parole, la gestion des sujets de conversation, la gestion des obligations sociales, etc. (Petukhova *et al.*, 2007). Nous reviendrons sur plusieurs éléments de ce cadre théorique au chapitre 4, lorsqu'il nous faudra décrire en détail les MI.

### 2.2.3 Fraisse et Paroubek (2015)

Fraisse et Paroubek (2015) ont pris comme sujet d'étude certaines interjections utilisées dans les messages écrits courts de type « microblog » (Twitter) avec pour objectif d'« évaluer leur apport pour les systèmes d'analyse de sentiments et de fouille d'opinions, en particulier pour la tâche de détection des émotions ». Ils ont utilisé des unités primaires et secondaires afin d'étiqueter certains messages à l'aide de 8 classes émotionnelles.

La classe COLÈRE regroupe les interjections *argh*, *pff* et *voyons!*. La classe PLAISIR regroupe *hihi*, *haha*, *lol* et *youpi*. La classe TRISTESSE regroupe *aïe*, *ouille*, *zut* et *hélas*. La classe PEUR n'inclut qu'un seul membre, *aah*, tout comme la classe APAISEMENT qui n'inclut que *ouf*, la classe INSATISFACTION qui inclut *bof*, la classe MÉPRIS qui inclut *beurk* et la classe SURPRISE NÉGATIVE qui inclut *oups*.

Les messages contenants les différentes interjections considérées ont été étiquetés afin de former un corpus d'apprentissage. Un classifieur a été entraîné pour chaque émotion indépendamment, avec des résultats satisfaisants pour les catégories « colère » et « plaisir ». Un lexique d'émotions a aussi pu être construit à l'aide de ce corpus qui a lui-même servi à construire un système de classement ayant obtenu des résultats comparables à d'autres systèmes similaires pour certaines catégories d'émotions (« mépris » et « plaisir »).

### 2.3 Conclusion au sujet du traitement automatique des MI

Bien que les recherches que nous venons de présenter portent la plupart sur l'anglais et n'ont pas spécifiquement les MI comme objet d'étude, elles nous offrent plusieurs outils d'une grande utilité pour l'analyse automatique des MI. La pertinence de l'environnement lexical, de la position des marqueurs dans les tours de parole et de certains éléments prosodiques a été établie pour la désambiguïsation des MP et ces aspects ont été quantifiés avec une relative précision. Ces études ont également permis de démontrer qu'il était possible pour des étiqueteurs séquentiels de prendre en compte des unités de type « mots-phrases » et de les traiter de manière satisfaisante, pourvu que les informations pertinentes étaient fournies au système.

# Chapitre 3 :
# Identification automatique des MI

Dans ce chapitre, nous présentons le processus d'expérimentation par lequel nous avons conçu et comparé quatre méthodes d'identification automatique des MI. Dans le cadre de cette expérience, une méthode d'identification est un système informatique chargé de dresser avec le plus de précision possible la liste des signifiants qui sont des MI dans un texte cible.

*texte cible* => Méthode d'identification => *liste de MI*

En guise d'exemple, pour un texte cible comme celui reproduit en (34), une méthode d'identification des MI qui fonctionnerait correctement identifierait l'unité *heille* comme étant un MI, mais pas l'unité *super*.

(34) C : **heille** c'était **super** bon j'ai pleuré en plus
*[CFPQ, sous-corpus 17, segment 4, page 52, ligne 18]*

Après avoir présenté notre méthodologie de recherche, nous décrivons différents agencements de modules informatiques puis analysons leur performance.

## 1 Méthodologie de recherche

Dans cette section, nous précisons la méthodologie que nous avons employée afin de déterminer les forces et les faiblesses de différentes méthodes d'identification automatique des MI.

### 1.1 Structure des méthodes d'identification

Les méthodes d'identification que nous analysons sont composées de deux types de modules informatiques. Les modules étiqueteurs prennent un texte cible en entrée et produisent un texte

étiqueté en sortie. Le module classifieur prend un texte étiqueté en entrée et produit une liste de signifiants qui sont des MI en sortie.

*texte cible* => Étiqueteur => *texte étiqueté*
*texte étiqueté* => Classifieur => *liste de MI*

L'utilisation d'un étiqueteur comme intermédiaire entre le texte cible et le classifieur permet un premier niveau d'abstraction à faible coût en termes de ressources informatiques.

Le tableau 5 indique quels modules informatiques sont utilisés par quelles méthodes d'identification.

*Tableau 5 : Modules utilisés par les différentes méthodes d'identifications des MI*

| **Méthodes** | **Modules utilisés** |
|---|---|
| Méthode minimum | *aucun* |
| Méthode n-grammes | Étiqueteur à n-grammes |
| Méthode Brill | Étiqueteur Brill |
| Méthode SVM | Étiqueteur à n-grammes + Classifieur SVM |

### 1.2 Extraction du corpus test

74607 lignes, ou tours de parole, constituent la totalité du CFPQ. Nous avons extrait de cet ensemble toutes les lignes qui contiennent au moins un des signifiants que nous étudions (la liste de ces signifiants est présentée au chapitre 1-1.6). Ces 8095 lignes forment un corpus **circonscrit** qui ne contient que les énoncés qui nous apparaissent pertinents pour notre étude.

Nous avons extrait aléatoirement 20% des tours de parole de ce corpus circonscrit pour former un **corpus test**. 1619 lignes ont ainsi été sélectionnées (une ligne sur cinq à intervalle régulier). Nous n'avons pas eu accès à ce corpus au cours du paramétrage final des différentes méthodes d'identification que nous avons développées.

Les 6476 autres tours de parole du corpus circonscrit forment notre **corpus de travail**. Nous avons librement eu accès à celui-ci. Ce corpus de travail a été lui-même divisé en une multitude de corpus d'entraînement et de corpus tests au cours de nos expérimentations. Nous donnerons plus de détails au point 1.7 sur les méthodes de validation croisée que nous avons utilisées.

### 1.3 Mesure des performances

Nous comparons les performances des différentes méthodes d'identification en mesurant les niveaux de rappel et de précision qu'elles obtiennent lorsqu'elles tentent de repérer les MI présents dans un texte.

Le **rappel** indique la proportion des MI identifiés correctement par rapport au total des unités qui sont réellement des MI dans le texte cible.

rappel = MI corrects / total MI

La **précision** indique la proportion des unités identifiées correctement parmi celles identifiées comme MI :

précision = MI corrects / MI repérés

La **f-mesure** indique un pourcentage qui représente la combinaison de la précision et du rappel. La f-mesure permet de mesurer la performance globale d'une méthode d'identification au sujet d'une unité ou d'un groupe d'unités. Le calcul suivant permet de déterminer la f-mesure d'une tâche de classification :

f-mesure = 2 • [(précision • rappel) / (précision + rappel)]

### 1.4 Problématique du sur-ajustement

Le sur-ajustement (*over fitting*) est un phénomène qui survient lorsqu'une méthode d'identification est tellement bien adaptée au traitement d'un ensemble de données qu'elle devient inefficace dans le traitement de données nouvelles.

C'est en partie pour éviter ce phénomène que nous avons isolé 20% des énoncés du CFPQ afin de constituer un sous-ensemble destiné à un test de performance des différentes méthodes. Le paramétrage final des modules informatiques que nous étudions a donc été effectué sans tenir compte de ces énoncés tests.

La conception de l'étiqueteur à n-grammes a cependant échappé en partie à cette précaution puisque dans la phase d'exploration de cette thèse plusieurs tests préliminaires ont été effectués sur l'ensemble des énoncés du CFPQ. Le processus d'étiquetage des unités de l'ensemble du CFPQ a également nécessité la prise en compte et l'analyse de l'ensemble du corpus. Il est donc possible que le choix des étiquettes utilisées par l'étiqueteur à n-grammes ait souffert de sur-ajustement occasionné par les choix subjectifs que nous avons dû faire suite à l'analyse du corpus.

### 1.5 Conception des méthodes d'identification des MI

Parmi les études qui nous ont inspiré, celle de Hutchinson (2004) a eu une influence importante sur notre stratégie méthodologique globale. Comme l'auteur de cette étude, nous avons opté pour un étiquetage large des unités linguistiques qui composent les textes cibles, suivi d'une analyse statistique à partir de différents traits tirés de cet étiquetage large.

#### 1.5.1 Méthode minimum

La méthode minimum n'utilise pas de système d'apprentissage. Elle identifie toutes les unités qui se trouvent dans la liste de signifiants pouvant être des MI comme étant des MI. Elle a un taux de rappel de 100%, mais une précision qui varie (énormément) selon le pourcentage de chacun des MI par rapport au total de signifiants dans le corpus.

Les résultats de cette méthode servent de barème de base destiné à être comparé aux résultats des autres méthodes. Le calcul de la f-mesure de cette méthode utilise toujours un rappel de 100% et une précision qui équivaut à la proportion des MI par rapport au total de signifiants de chaque unité.

### 1.5.2 Méthode n-grammes

La méthode n-grammes utilise un étiqueteur à n-grammes comme mécanisme d'identification des MI. Dans ce contexte-ci, un n-gramme est une séquence de n unités tirée d'un énoncé. En guise d'exemple, l'énoncé (35) peut être divisé en trois 2-grammes (« le-chien », « chien-jappe », « jappe-fort »), deux 3-grammes (« le-chien-jappe », « chien-jappe-fort ») ou un 4-gramme (« le-chien-jappe-fort »).

(35) le chien jappe fort

Les étiqueteurs à n-grammes ont souvent été utilisés avec succès en linguistique computationnelle, notamment pour identifier les classes grammaticales (*POS-Taggers*) auxquelles appartiennent les unités qui composent les textes d'un corpus. Leur fonctionnement est relativement simple, rapide et transparent.

Puisqu'un étiqueteur permet d'attribuer des étiquettes à tous les mots d'un énoncé, il peut fournir de l'information pertinente à toutes les méthodes d'identification. La méthode n-grammes se base entièrement sur les résultats d'un étiqueteur à n-grammes pour déterminer la classe à laquelle appartient un signifiant. L'étiqueteur Brill et le classifieur SVM utilisent également plusieurs informations tirées du travail de ce même étiqueteur à n-grammes.

Dans la phase d'exploration de cette thèse, nous avons fait quelques tests avec l'étiqueteur TreeTagger (Schmid, 1994). Le système d'étiquettes proposé par Achim Stein destiné au traitement du français par TreeTagger ne permet pas (sans modification) de prendre en compte les

MI de façon satisfaisante. Par exemple, des signifiants comme *coudon* ou même *wow* sont classés comme « inconnu » par TreeTagger. Également, une locution comme « mets-en » est analysée comme étant un verbe au présent (VER:pres) suivi d'un pronom personnel (PRO:PER) et jamais comme une unité extra-phrastique.

L'ensemble d'étiquettes utilisé par Hutchinson (2004), qui exploite avec succès une version large de l'ensemble d'étiquettes du Penn Treebank pour identifier des marqueurs pragmatiques, aurait probablement pu servir de bon point de départ pour notre projet. Malheureusement, ceci aurait nécessité un étiquetage manuel (trop demandant en ressources humaines et budgétaires) des énoncés du CFPQ.

Pour ces raisons, nous avons décidé de créer notre propre système d'étiquettes destiné à faire ressortir les traits pertinents à l'identification des MI (voir 2.1.3).

### 1.5.3 Méthode Brill

La méthode Brill a recours à un étiqueteur Brill (tel que décrit en 2.2) pour repérer les MI.

La faiblesse la plus évidente de la méthode n-grammes est le fait qu'elle tient uniquement compte des unités lexicales à gauche des unités qu'elle cherche à classer. La première solution que nous avons envisagée et testée pour corriger ce problème a été le recours à la méthode développée par Brill (1992, 1995). Cette méthode consiste à utiliser un programme qui examine automatiquement les résultats d'un étiqueteur à n-grammes afin de déterminer des règles de transformation qui permettent de corriger les erreurs que celui-ci commet.

### 1.5.4 Méthode SVM

Nous avons étudié le fonctionnement des SVM (Support Vector Machines) et développé une méthode d'identification qui se base sur un classifieur SVM. Les principes de base des SVM ont été introduits par Boser, Guyon et Vapnik (1992). Le recours aux SVM a connu une grande

hausse de popularité ces dernières années en linguistique et dans plusieurs autres domaines de recherche.

Les SVM ont l'avantage d'être efficaces même lorsqu'ils ne disposent que d'un faible nombre d'exemples avec lesquels s'entraîner. La puissance des SVM repose cependant sur la qualité de l'information qu'on leur fournit. Le classifieur SVM que nous avons développé tire une grande partie de cette information des résultats de l'étiqueteur à n-grammes.

### 1.6 Optimisation des méthodes d'identification

Plusieurs variables entrent en jeu dans la construction d'un système d'apprentissage machine. Le choix des paramètres propres à chaque méthode résulte en partie d'un processus d'aller-retour entre la programmation des modules informatiques et la validation sur notre corpus de travail.

La validation croisée à replis est le principal procédé par lequel il nous a été possible d'estimer puis d'optimiser la fiabilité des différents étiqueteurs et classifieurs que nous avons développés. Ce procédé permet d'exploiter pleinement le corpus de travail dont nous disposons en testant successivement différents agencements de corpus d'entraînement et de corpus tests. La thèse de Heeman (1997, p. 126), discutée au chapitre 2, présente un exemple d'utilisation de ce procédé développé au cours du temps par de nombreux chercheurs (par exemple, Stone, 1974).

Plus précisément, pour chaque ensemble particulier de paramètres propre à un module informatique, nous avons procédé à une validation croisée à 5 replis en suivant ces étapes :

1. Extraction d'un échantillon test (20% des lignes) et d'un échantillon d'entraînement (80% des lignes) du corpus de travail.
2. Entraînement du module informatique sur l'échantillon d'entraînement.
3. Test du module informatique sur l'échantillon de test et compilation des résultats.

4. Répétition des étapes 1 à 4 avec 4 agencements supplémentaires différents d'échantillons d'entraînement et de test en faisant en sorte que chacune des lignes du corpus de travail fasse partie exactement une fois du corpus test.
5. Compilation des scores obtenus et analyse de la performance moyenne du module.

Nous avons répété ce processus en faisant en sorte qu'un grand nombre de combinaisons de paramètres soient testées (plus de détails au point 2). En comparant les résultats de ces tests, nous avons pu déterminer l'effet de certains paramètres sur les scores des différents modules informatiques et ajuster ceux-ci de manière à produire les résultats les plus satisfaisants possible.

Notons que nous avons observé avec une plus grande attention les performances des modules au sujet des vocables ambigus qu'au sujet des vocables non-ambigus. En effet, la détection des vocables non-ambigus est en théorie très facile et les erreurs à ce sujet sont dues à des problèmes techniques extra-linguistiques (comme une fréquence trop faible du vocable).

## 2 Description des modules informatiques

Dans cette section, nous décrivons le fonctionnement des modules informatiques qui composent les quatre méthodes d'identification des MI qui font l'objet de notre étude. Nous précisons également les différents paramètres utilisés par chacun de ceux-ci ainsi que le processus par lequel il nous a été possible d'optimiser ces paramètres.

Nous utilisons les librairies du NLTK (Natural Language Toolkit, Bird, Loper et Klein, 2009) écrit en Python pour la programmation des deux étiqueteurs et la librairie Scikit-learn (Pedregosa *et al.*, 2011) pour la programmation du classifieur SVM.

Notons que l'entraînement et l'utilisation de ces modules informatiques sont très rapides et demandent un effort presque négligeable de la part d'un processeur moderne. Nous ne nous sommes par conséquent que très peu attardé sur les considérations d'efficacité informatique dans le cadre de ce projet.

## 2.1 Étiqueteur à n-grammes

Un étiqueteur séquentiel à n-grammes assigne des étiquettes aux mots d'un texte, de gauche à droite, grâce à de l'information statistique qu'il a récoltée dans un corpus d'entraînement dont tous les mots ont été correctement étiquetés.

Dans le contexte qui nous intéresse, une « étiquette » est une chaîne de caractères que l'on associe à un mot afin de le caractériser. Les étiquettes sont souvent utilisées pour indiquer les classes grammaticales auxquelles appartiennent les mots. Dans le cadre de l'analyse automatique des textes, un étiqueteur est un programme informatique chargé d'attribuer automatiquement des étiquettes aux mots d'un texte.

L'objectif des étiqueteurs que nous présentons ici est de déterminer avant tout si une unité appartient ou non à la catégorie des MI. Ils n'ont pas été conçus pour déterminer avec précision les classes grammaticales des unités qui ne sont pas des MI.

Dans le cadre de cette expérience, même si l'étiqueteur à n-grammes propose des étiquettes pour tous les mots d'un énoncé, nous mesurons sa performance en examinant uniquement l'étiquette qu'il attribue aux signifiants qui peuvent être des MI. Les signifiants qui peuvent être des MI sont les seuls au sujet desquels l'étiqueteur peut se tromper puisqu'ils sont les seuls à pouvoir être identifiés par plus d'une étiquette.

Le processus d'optimisation de l'étiqueteur à n-grammes nous a incité à distinguer les constructions *vraiment*, *vraiment pas*, *pas vraiment*, ainsi que *pas du tout* et *du tout*. Ces constructions sont utilisées en tant que MI dans des contextes différents, elles ont des sens différents et les analyser de manière distincte permet de meilleures performances de la part de l'étiqueteur n-grammes.

La construction *pas pantoute* est généralement utilisée de manière intraphrastique (dans 30 occurrences sur 34). Nous avons essayé d'entraîner l'étiqueteur en prenant comme cible le phrasème *pas pantoute* (plutôt que *pantoute* seul), sans résultat satisfaisant.

Les types de paramètres qui ont entré en jeu dans la configuration de l'étiqueteur à n-grammes sont :

1. les librairies informatiques utilisées;
2. l'enchaînement particulier des différents modules n-grammes;
3. le système d'étiquettes utilisé pour la préparation des corpus d'entraînement.

### 2.1.1 Librairies utilisées

Nous avons utilisé les classes *UnigramTagger*, *BigramTagger* et *TrigramTagger* qui se basent sur la classe *ContextTagger* de la librairie NLTK afin de créer l'étiqueteur à n-grammes.

Nous avons dû modifier le comportement de la fonction *_train()* de la classe *ContextTagger* qui a pour rôle de déterminer l'étiquette la plus probable pour chacun des contextes. Celle-ci avait un comportement instable lorsqu'elle rencontrait des contextes avec des probabilités égales. Le programme choisissait alors l'étiquette de manière aléatoire. Ce phénomène n'avait pas d'impact lors des tests sur la plupart des unités, mais était problématique pour certaines unités, particulièrement celles dont on ne dispose que de peu d'occurrences. Afin d'éviter des résultats fluctuants pour ces unités, nous avons modifié le programme pour qu'il ignore les contextes où les probabilités sont égales entre les étiquettes et privilégie plutôt l'étiquette qui est attribuée par l'étiqueteur suivant dans la chaîne (*backoff tagger*).

### 2.1.2 Enchaînement d'étiqueteurs

L'étiqueteur à n-grammes utilise un enchaînement de quatre étiqueteurs qui ont des portées décroissantes. Le programme a d'abord recours à l'étiqueteur qui a la plus grande portée syntaxique. Si celui-ci est incapable de déterminer l'étiquette la plus probable pour un mot, le programme a recours à l'étiqueteur suivant dans la chaîne.

**I. Étiqueteur à trigrammes**

L'étiqueteur à trigrammes détermine l'étiquette probable d'un mot cible à partir des étiquettes des deux mots qui le précèdent.

Durant sa phase d'entraînement, le programme compte tous les trigrammes associés à chacun des mots, c'est-à-dire toutes les combinaisons de 3 étiquettes associées à un mot cible en position finale. Il est ainsi capable de déterminer la probabilité qu'une étiquette particulière se trouve associée à un certain mot lorsque celui-ci se trouve à la suite de 2 autres étiquettes particulières.

Puisqu'il n'a d'information qu'au sujet des trigrammes qu'il a rencontrés dans le corpus d'entraînement, un tel étiqueteur n'est efficace que pour les trigrammes fréquents et bénéficie grandement d'être entraîné sur de très gros corpus. En pratique, un faible pourcentage de MI sont reconnus et étiquetés grâce à l'étiqueteur à trigrammes. En cas d'échec, le programme a recours à l'étiqueteur numéro II.

**II. Étiqueteur à bigrammes**

L'étiqueteur à bigrammes détermine l'étiquette d'un mot à partir de ce mot et de l'étiquette du mot qui le précède.

Comme il y a beaucoup moins de bigrammes possibles que de trigrammes, un tel étiqueteur n'est pas aussi demandant quant à la taille de son corpus d'entraînement que l'étiqueteur à trigrammes. Pour la plupart des mots, il a de l'information sur la probabilité qu'une étiquette particulière se trouve à la suite d'une autre.

En cas d'échec, le programme a recours à l'étiqueteur numéro III.

**III. Étiqueteur à unigrammes**

Au cours de l'entraînement, l'étiqueteur à unigrammes compte le nombre de fois que chacun des mots porte chacune des étiquettes. Lorsqu'il est temps d'exécuter sa tâche d'étiquetage, il

attribuera l'étiquette la plus populaire pour chacun des mots. Si un mot X a l'étiquette 'A' 51% du temps dans le corpus d'entraînement, l'étiqueteur à unigrammes lui attribuera cette étiquette 100% du temps dans les nouveaux textes qui lui seront soumis.

**IV. Étiqueteur par défaut**

Un étiqueteur par défaut attribue à tous les mots la même étiquette ('S' dans notre cas). Il est utile dans le cas où un mot n'a pas été rencontré au cours de l'entraînement.

### 2.1.3 Ensemble d'étiquettes

Afin de construire le corpus d'entraînement nécessaire à l'utilisation des étiqueteurs, nous avons converti les textes du CFPQ (disponibles en format PDF) en fichiers textes où les énoncés occupent une ligne chacun. Suite à un processus de nettoyage de base des textes qui consiste à y enlever les éléments qui ne sont pas pertinents pour les besoins de l'opération (caractères techniques, références, indicateurs de chevauchement de tours de parole, descriptions des gestes), les unités du corpus ont ensuite été étiquetées de manière semi-automatique selon les paramètres décrits ici.

La faible fréquence des MI, en comparaison avec la plupart des autres classes grammaticales, fait en sorte que nous disposons en pratique d'un très petit corpus d'entraînement pour chacun d'eux. Un signifiant comme *cool*, par exemple, ne se trouve que 56 fois dans notre corpus de travail et n'est MI que dans 7 occurrences. Nous avons donc eu recours à différentes stratégies afin d'atténuer ce problème. La première est l'utilisation d'un nombre limité d'étiquettes, en priorisant celles qui nous apparaissent pertinentes au repérage des MI. En effet, un système qui utilise un petit nombre d'étiquettes a plus de chances d'avoir des statistiques significatives à leur sujet. L'utilisation d'un trop grand nombre d'étiquettes différentes ferait en sorte que chacune d'elles ne se retrouverait qu'un petit nombre de fois dans les contextes pertinents des corpus d'entraînement. Notons que Hutchinson (2004) utilise ce principe en transformant le système d'étiquettes du corpus TreeBank en un système aux classes plus larges.

Le processus d'optimisation de l'étiqueteur à n-grammes à l'aide de validations croisées nous a graduellement mené à utiliser des étiquettes de plus en plus larges. L'étiquette 'M', par exemple, regroupe des éléments qui semblent très différents les uns les autres. Ces éléments partagent cependant la caractéristique d'être des indicateurs de ruptures syntaxiques. Nous estimons que le recours à un corpus de travail plus volumineux aurait probablement orienté le processus d'optimisation de l'étiqueteur vers l'utilisation d'étiquettes plus précises.

Notons également que nous n'avons pas les ressources humaines nécessaires à l'étiquetage précis de toutes les classes grammaticales de tous les mots du CFPQ. Les signifiants qui peuvent appartenir à plusieurs classes grammaticales ne sont pas désambiguïsés dans le corpus d'entraînement (sauf ceux qui peuvent être MI). La frontière entre les unités identifiées comme déterminants et celles identifiées comme des pronoms, par exemple, a été établie lors du processus d'optimisation de l'étiqueteur (voir plus bas).

Le tableau 6 dresse la liste des étiquettes que nous avons utilisées. Nous présentons ensuite celles-ci en les mettant en relation avec certains phénomènes, identifiés au chapitre 2, qui semblent déterminants dans la production des MI.

*Tableau 6 : Étiquettes utilisées par l'étiqueteur à n-grammes*

| **Catégories** | **Unités** | **Étiquettes** |
|---|---|---|
| MI | *écoute*, *crisse, coudon, regarde, ...* | M |
| Lieurs syntaxiques | *que, qu'* | QUE |
| | *si* | SI |
| | *à, au* | A |
| Conjonctions | *donc* | DONC |
| | *mais* | MAIS |
| Indicateurs d'hésitations, remplisseurs de pauses | *hi*, *oh*, *eh*, *ah*, *ouais*, *bah*, *hein, hum, euh, pff, ouh, ...* | M |
| | *ben* | BEN |
| Mots phrases | *ouin*, *oui*, *non, ...* | M |
| Amorces de mots | *s-* | AM |
| MDs baliseurs | *ok*, *t'sais, là, ...* | M |
| Pauses | . | M |
| Rires | *<rire>* | M |
| Citations | Début et fin de citation | M |
| Intonations | Intonation montante | IM |
| | Intonation double montante | IM |
| | Intonation descendante | ID |
| | Intonation double descendante | ID |
| Pronoms sujets et réfléchis | *je*, *j'*, *me*, *m'*, *tu*, *te, t'*, *il*, *elle*, *on*, *se*, *s'*, *qui, y, vous* | PRO |
| Déterminants | *le*, *la*, *les*, *un*, *une*, *du*, *ce*, *cet*, *cette, mon, ma, mes, ton, ta, tes, son, sa, ses* | DET |
| Lieur de phrasème adverbial | *en* | EN |
| | Autres unités | S |

#### 2.1.3.1 Les débuts et les fins de tour de parole

Nous avons vu que la caractéristique syntaxique principale des MI est qu'ils entretiennent rarement des liens de dépendance avec d'autres unités lexicales ou d'autres syntagmes. Une autre façon de décrire le phénomène est de dire qu'un MI est souvent précédé et suivi de ruptures de dépendance syntaxique.

Le premier mot qu'un locuteur prononce lorsqu'il commence à parler ne peut pas être lié syntaxiquement à un mot prononcé plus tôt dans la conversation, sauf dans des cas de reprise après une interruption. De façon similaire, un locuteur interrompt habituellement son tour de parole à la fin d'une phrase ou à l'aide d'un mot-phrase. Un mot prononcé seul a également toutes les chances d'être un mot-phrase. En (36) par exemple, le mot *malade* nous apparaît comme un MI, même sans connaître le contexte de la conversation:

(36) J-M : [...]
D : **malade**
*[CFPQ, sous-corpus 10, segment 8, page 104, ligne 20]*

Les résultats de plusieurs études confirment que les MD en tête et en fin d'énoncés sont beaucoup plus faciles à repérer de façon automatique (notamment, Litman, 1996; Popescu-Belis et Zufferey, 2011).

Dans le corpus d'entraînement, chaque tour de parole se termine par un caractère de fin de ligne. Les débuts et les fins de tours de parole sont donc pris en considération par les programmes étiqueteurs.

#### 2.1.3.2 Les pauses dans le débit de la parole

Les syntagmes sont habituellement émis d'une traite, sans pauses longues entre les mots. Les pauses dans le débit de parole correspondent ainsi souvent à des intermissions entre deux syntagmes. Dans leur étude, Popescu-Belis et Zufferey (2004) ont déterminé que des pauses de

60 millisecondes (ms) avant ou après les unités tendaient à caractériser les MD qu'ils étudiaient. Petukhova et Buny (2009) ont pour leur part pu déterminer que des pauses de 59 ms à 228 ms se trouvaient habituellement avant les MD.

L'énoncé (37) exemplifie la production du mot-phrase SUPER intercalé à l'intérieur d'une phrase, grâce à l'usage de pauses :

(37) N : ça va être un automatisme ok **(.) super (.)** mais si […]
*[CFPQ, sous-corpus 2, segment 9, page 103, ligne 4]*

Dans le corpus d'entraînement, nous donnons aux pauses l'étiquette 'M' à l'instar des autres unités qui marquent une rupture syntaxique. Notons que nous avons converti les symboles du corpus qui désignaient des pauses de différentes longueurs en un seul symbole : « *(.)* » . Toutes les pauses significatives sont donc considérées comme étant un même type d'unité, peu importe leur longueur.

#### 2.1.3.3 Les citations

À l'oral, les citations s'insèrent rarement sans rupture à l'intérieur de syntagmes. Lorsqu'un locuteur en cite un autre (ou se cite lui-même), il le fait en introduisant une rupture syntaxique. Cette rupture est souvent signalée par une coloration particulière de la voix (Prsir, 2012) qui peut prendre des formes très variées. Un grand nombre de citations sont annotées dans le CFPQ par des caractères spéciaux, ce qui permet de les repérer et de pallier l'absence d'informations prosodiques détaillées dans le texte du corpus. Ces indicateurs de début et de fin de citations reçoivent également l'étiquette 'M' dans le corpus d'entraînement.

#### 2.1.3.4 Les marques d'intonation

À l'oral, l'intonation sert notamment à délimiter les syntagmes. Une analyse sommaire des marques d'intonation dans le CFPQ suffit à se rendre compte qu'elles sont pratiquement toutes associées à des fins de phrases ou de syntagmes. La plupart des études sur les MP présentées au

chapitre 2 considèrent la prosodie comme un phénomène important pour distinguer les MP de leurs équivalents intraphrastiques (par exemple, Hirschberg et Litman, 1987; Petukhova et Bunt, 2009; Popescu-Belis et Zufferey, 2011).

Dans le corpus d'entraînement, les deux types d'intonations montantes reçoivent l'étiquette 'IM', tandis que les deux types d'intonations descendantes reçoivent l'étiquette 'ID'.

Au cours du processus d'optimisation de l'étiqueteur à n-grammes, nous avons conclu qu'il était préférable de ne pas tenir compte de la distinction entre les intonations légères et fortes. La relative faible fréquence des marques d'intonation dans le corpus fait en sorte qu'il est utile de ne pas diluer les statistiques à leur sujet.

#### 2.1.3.5 Les rires

Les épisodes de rire interrompent souvent les conversations à bâtons rompus que nous examinons et constituent plausiblement un bon indice de ruptures syntaxiques. En (38) par exemple, le marqueur *voyons* est précédé d'une pause dans le discours occupé par un épisode de rire.

(38) J-M : [1mais elle a même pas ca- el- el- elle avait même pas catché encore elle pensait encore j'étais un étudiant [2**(RIRE) voyons** donc (*dit en riant*)
*[CFPQ, sous-corpus 10, segment 6, page 78, ligne 13]*

Les indicateurs de rire reçoivent l'étiquette 'M' dans le corpus d'entraînement.

#### 2.1.3.6 Hésitations et autocorrections

Les hésitations et les autocorrections ne posent pas de grands problèmes pour le repérage des MI (mais posent d'énormes problèmes pour le repérage d'autres classes grammaticales). Selon Heeman (1997) et Heeman et Allen (1999), ces phénomènes se révèlent au contraire être de bons outils pour déterminer le caractère discursif ou non de plusieurs unités. Les mots qui servent à remplir des pauses liées à l'hésitation (*euh*, *hum*) et les amorces de mots (se terminant par un tiret

dans le CFPQ, par exemple : s-, *pou*-) sont les indices les plus facilement détectables d'hésitations et d'autocorrections. En (39) par exemple, nous pouvons observer que les hésitations de l'énonciateur au sujet du mot « contexté » mènent à la production de *voyons*.

(39) G : [1ça réussit à rentrer dans le contexte en [2plus
S : [2oui en plus c'est très <p<très très>> **conte:s- con- contes- (.) voyons**
G : <p<RP>>
S : contexté [1oui c'est très RP
*[CFPQ, sous-corpus 9, segment 3, page 37, lignes 10-13]*

Au cours de la préparation d'un corpus d'entraînement de l'étiqueteur à n-grammes, le système remplace chacun des tirets qui marquent les amorces de mots par une chaîne de caractère spéciale destinée à marquer l'hésitation. Cette marque reçoit l'étiquette 'AM'.

Nous donnons l'étiquette 'M' aux petits mots qui servent à remplir des pauses liées à l'hésitation (*euh*, *hum*).

#### 2.1.3.7 Unités lexicales extraphrastiques

Plusieurs MD sont des mots-phrases et signalent par conséquent des ruptures syntaxiques. Les MI eux-mêmes, les unités *oui, ouais* et *non*, les marqueurs d'appel à l'écoute (comme *t'sais*), les marqueurs d'écoute (*ouin*, *ok*), les marqueurs de balisage (*là*) sont tous des unités fortement susceptibles de se côtoyer les unes les autres. En (40) par exemple, nous pouvons voir s'enchaîner plusieurs MI. Dans cet énoncé, le marqueur de balisage *là* ne laisse pas de doute sur le statut de mot-phrase de *vraiment* et *sérieux* :

(40) S: câ- **heille vraiment là sérieux là heille** (*inaud.*) je me suis brossé les dents le matin (.) arrive le soir (*en ouvrant ses paumes vers le haut comme pour représenter le vide*) (.) plus de brosse à dents (*en écartant les mains comme pour exprimer son incompréhension*)
*[CFPQ, sous-corpus 21, segment 2, page 27, ligne 4]*

Plusieurs de ces unités sont classées parmi les *editing terms* par Heeam (1997, p. 13) et se voient attribuées, dans son corpus d'entraînement, des étiquettes particulières selon leurs rôles discursifs (*AC* pour « Single word acknowledgments », *UH_D* pour « Interjections with discourse purpose » et *CC_D* pour « Co-ordinating conjuncts used as discourse markers »).

Dans notre corpus d'entraînement, les unités lexicales qui sont en pratique toujours extraphrastiques, comme plusieurs MI et MP baliseurs, reçoivent l'étiquette 'M'.

Quant aux signifiants des MI qui sont homonymes avec des unités intraphrastiques, ils reçoivent l'étiquette 'M' s'ils sont MI et l'étiquette 'S' dans les autres cas. Rappelons que, selon notre système d'étiquettes, les signifiants associés aux MI sont les seuls à pouvoir être identifiés par plus d'une étiquette.

#### 2.1.3.8 Lieurs syntaxiques *que*, *si*

Nous avons vu au chapitre 2 que les signifiants liés aux MI ne sont pas toujours utilisés comme mots-phrases et peuvent être produits en lien avec des propositions, parfois à l'aide des mots *que, de* et *si*. La présence de ces mots à la suite de ces signifiants sera naturellement un phénomène pertinent à prendre en compte lors de l'entraînement de l'étiqueteur Brill et du classifieur SVM. En (41), le signifiant *crisse* est lié syntaxiquement à la proposition qu'il précède (par *que*).

(41) Y: [2c’est du temps man mai:s [3**crisse que** des fois tu sauves [4de l’argent
*[CFPQ, sous-corpus 21, segment 3, page 43, ligne 12]*

Dans le corpus d'entraînement, les conjonctions *que* et *si* ont leurs propres étiquettes ('QUE' et 'SI').

#### 2.1.3.9 *De*

Les sacres qui sont des MI peuvent être liés entre eux par la préposition *de*, dans un enchaînement de sacres. L'extrait (42) montre l'enchaînement de *ostie* et de *tabarnaque* à l'aide d'un *de*.

(42) Y : [1t'arrives crisse **ostie de tabarnaque** euh (*en reculant sur sa chaise et en levant les mains comme pour exprimer une évidence*)
*[CFPQ, sous-corpus 21, segment 6, page 86, ligne 17]*

Les sacres peuvent également être employés comme intensifieurs adjectivaux (Dostie, 2015, p. 62). Le signifiant *crisse* de l'extrait (43), agit comme intensifieur d'une propriété de l'adjectif nominalisé *folle*.

(43) [...] je l'ai traitée de **crisse de folle** [...]
*[CFPQ, sous-corpus 2, segment 8, page 97, ligne 3]*

Nous voyons ainsi que l'unité *de* n'est pas une marque certaine du caractère extraphrastique ni du caractère intraphrastique des MI. Après expérimentation, nous avons décidé de ne pas étiqueter de manière distincte la préposition *de* (nous lui donnons l'étiquette générale intraphrastique 'S').

#### 2.1.3.10 Conjonctions *donc* et *mais*

Les conjonctions de coordination *donc* et *mais* sont souvent utilisées de manière discursive et se voient ainsi parfois accolées à des MI. En (44) par exemple, *donc* est accolé à *voyons* pour former un phrasème typiquement utilisé de manière discursive.

(44) AN : pour faire peur au monde (.) **mais voyons donc**
*[CFPQ, sous-corpus 20, segment 4, page 35, ligne 10]*

Dans le corpus d'entraînement, nous donnons aux unités *donc* et *mais* leurs propres étiquettes ('DONC' et 'MAIS').

### 2.1.3.11 Déterminants

Les unités intraphrastiques sont le plus souvent étroitement accompagnées d'autres unités qui trahissent leur caractère intraphrastique. Les déterminants, les pronoms et les prépositions jouent des rôles importants à cet effet.

Les noms sont typiquement liés à des déterminants qui les précèdent. En (45) par exemple, le déterminant *la* permet de savoir que *merde* n'est pas un MI.

(45) J : ben il dit •il m'ont dit de la tester je l'ai testé ça valait pas de **la merde**↑°
*[CFPQ, sous-corpus 15, segment 8, page 134, ligne 22]*

Notons que nous distinguons la forme *merde* de la forme *marde* qui entre en jeu dans le phrasème ⌜DE LA MARDE⌝ que nous considérons comme un vocable à part entière (voir chapitre 4-2.25).

Dans le corpus d'entraînement, nous donnons aux signifiants qui peuvent être des déterminants (*le*, *la*, *un*, *une*, *du*, *ce*, *cet*, *cette, mon, ma, mes, ton, ta, tes, son, sa, ses*) l'étiquette 'DET', même si plusieurs de ceux-ci sont en réalité des pronoms. Nous avons testé différentes façons d'étiqueter les unités *le*, *la*, *les et l'* sans avoir détecté une influence à ce sujet sur les performances du système.

### 2.1.3.12 Pronoms

Les pronoms personnels susceptibles d'être sujets et les pronoms réfléchis (*je*, *j'*, *me*, *m'*, *tu*, *te, t'*, *il*, *elle*, *on*, *se*, *s'*, *qui, vous*) précèdent souvent des verbes. En (46) par exemple, la présence du pronom *il* permet de savoir que le signifiant *écoute* n'est pas un MI.

(46) I : mon chum **il écoute** le hochey:/ [...]
*[CFPQ, sous-corpus 26, segment 6, page 93, ligne 4]*

Les pronoms personnels se voient attribuer l'étiquette 'PRO' dans les corpus d'entraînement. Cet étiquetage des pronoms personnels a permis une amélioration notable des performances de l'étiqueteur à n-grammes au cours du processus d'optimisation. Par exemple, l'identification de *se* comme pronom dans l'énoncé (47) a permis l'étiquetage adéquat du verbe *crisse* par l'étiqueteur à n-grammes :

(47) MC : oui: mais (1") t'sais (.) c'est parce que Nathalie a comme pris euh le bord de: (.) <p<elle **se crisse** d'elle là>>
*[CFPQ, sous-corpus 22, segment 6, page 97, ligne 9]*

Notons que la particule interrogative *-tu*, comme dans l'énoncé (48), n'est pas étiquetée comme un pronom, mais comme une unité phrastique ('S').

(48) C : t'es**-tu** malade/ (RIRE)
*[CFPQ, sous-corpus 25, segment 1, page 9, ligne 14]*

### 2.1.3.13 EN

La préposition EN accompagne parfois un nom de manière à construire un phrasème adverbial. En (49) par exemple, la présence de *en* indique que *maudit* n'est pas un MI.

(49) JN : [1oui pis quand t'as les pieds accrochés après la planche ça doit être différent **en maudit** de quand t'es sur un surf les pieds nus euh: (.) lousses
*[CFPQ, sous-corpus 28, segment 2, page 26, ligne 15]*

Nous donnons à *en* sa propre étiquette 'EN' dans le corpus. Ceci permet notamment à l'étiqueteur à n-grammes de repérer certains emplois extra-phrastiques de sacres qui concernent la construction adverbiale « *en* + sacre ». Par exemple, quatre emplois intraphrastiques du signifiant *crime* sur cinq dans le corpus de travail se trouvent dans la construction « en crime » que l'étiqueteur n'a aucun mal à repérer grâce à la présence du mot *en* et à l'étiquette 'EN' que nous lui avons attribuée.

#### 2.1.3.14 BEN

L'unité *ben* est souvent utilisée en compagnie d'un MI, à sa gauche ou à sa droite. L'énoncé (50) exemplifie un cas où le MI *écoute* est suivi de *ben*.

(50) J : ben une chance maudit je serais devenue folle trois fois/ là **écoute ben/** wô là↑
*[CFPQ, sous-corpus 15, segment 8, page 144, ligne 1]*

Nous donnons à *ben* sa propre étiquette ('BEN') dans le corpus d'entraînement.

#### 2.1.3.15 *À*

Après expérimentation, nous avons constaté que les performances des étiqueteurs sont plus élevées quand nous attribuons à la préposition *à* et sa contraction *au* une étiquette qui leur est propre 'A'.

#### 2.1.3.16 Autres unités intraphrastiques

Il est présumé que la plupart des unités lexicales (noms, verbes, adverbes, adjectifs) sont utilisées de manière intraphrastique. Lors du processus d'étiquetage des textes du CFPQ, nous avons regroupé toutes les unités qui n'avaient pas d'étiquettes sous l'étiquette 'S'.

Comme nous allons le voir en 3, l'étiquette 'S' et les unités phrastiques auxquelles elle est associée interviennent de manière particulièrement pertinente dans l'entraînement de l'étiqueteur à tri-grammes.

## 2.2 Étiqueteur Brill

Un étiqueteur Brill fonctionne à l'aide d'une liste de règles de transformation qu'il acquiert à partir de l'observation des erreurs commises par un autre étiqueteur sur un corpus d'entraînement (Brill, 1992, 1995).

Brill décrit son système comme étant « Transformation-based Error-driven Learning » (Brill, 1995). En prenant comme point de départ les résultats de l'étiqueteur à n-grammes lorsque appliqué sur le corpus d'entraînement, le module Brill détermine la règle de transformation qui produit le meilleur résultat lorsqu'elle est appliquée sur le corpus étiqueté. Le corpus étiqueté est ensuite modifié par cette règle. Le module détermine la règle qui produit le meilleur résultat lorsqu'elle est appliquée sur cette nouvelle version du corpus et ainsi de suite jusqu'à ce que le système ne puisse plus déterminer de règle qui améliore l'étiquetage du corpus (ou qu'il atteigne un nombre de règles maximal déterminé). Les règles de transformation ainsi déterminées peuvent ensuite être testées sur de nouveaux textes.

Alors que la chaîne d'étiqueteurs séquentiels à n-grammes ne considère que les étiquettes des unités qui précèdent un mot à étiqueter, l'étiqueteur Brill permet de prendre en compte les mots eux-mêmes (plutôt qu'uniquement leurs étiquettes) qui précèdent et qui suivent un mot à étiqueter.

Les paramètres propres au programme qui implémente l'étiqueteur Brill sont de deux types :

1. la librairie utilisée;
2. les *templates* utilisés.

### 2.2.1 Librairie utilisée

Nous utilisons la classe *BrillTagger* de la librairie NLTK et lui donnons comme étiqueteur de départ l'enchaînement d'étiqueteurs à n-grammes présenté plus haut. L'étiqueteur Brill teste ce dernier sur son propre corpus d'entraînement et note les erreurs produites à partir d'une liste de positions particulières (*templates*).

### 2.2.2 Templates utilisés

Les *templates* que nous utilisons sont minimalistes. Nous voulons que le système prenne en compte les signifiants des unités qui suivent et qui précèdent immédiatement les signifiants de MI que nous devons identifier. Seulement 2 règles sont nécessaires afin de prendre en compte ces contextes. Le tableau 7 présente ces règles comme elles se retrouvent dans le programme et donne des explications en langue naturelle.

*Tableau 7 : Templates utilisés par le module Brill*

| **Templates** | **Explications** |
|---|---|
| Template(Word([0]), Word([-1])) | Si :<br>pour tel mot X, le mot précédant est Y.<br>Alors :<br>changer l'étiquette A du mot X pour l'étiquette B. |
| Template(Word([0]), Word([1])) | Si :<br>pour tel mot X, le mot suivant est Y.<br>Alors :<br>changer l'étiquette A du mot X pour l'étiquette B. |

Nous avons testé une grande quantité de listes de *templates*, proposées dans différentes études sans qu'il y ait d'amélioration appréciable de la performance de l'étiqueteur. Le contexte lexical immédiat semble donc, sans surprise, être le plus pertinent pour nos besoins.

### 2.3 Classifieur SVM

En termes généraux, une machine à vecteurs de support (*support vector machine* ou SVM en anglais) est un regroupement d'outils statistiques qui permet à un programme informatique d'accomplir des tâches de classification à l'aide d'un apprentissage supervisé.

Plusieurs études et projets ont démontré l'efficacité des SVM dans différents domaines d'application de l'apprentissage machine. Par exemple, Li, Ong, Suganthan et Thing (2010) utilisent les SVM pour déterminer automatiquement la nature de fichiers informatiques. Les SVM peuvent être utilisés pour reconnaître automatiquement des images, des fichiers audios, etc.

Le fonctionnement général des SVM est un phénomène complexe au sujet duquel nous ne nous attarderons pas plus que nécessaire. En ce qui a trait à cette thèse, le classifieur SVM que nous utilisons nécessite un entraînement au cours duquel il détermine le tracé idéal d'une frontière entre deux classes (dans un espace à nombreuses dimensions) en examinant plusieurs exemplaires de ces classes. Une fois entraîné, le classifieur peut prédire, avec plus ou moins de précision, la classe d'une unité (MI ou non-MI) présente dans un texte cible. Les calculs du classifieur s'appuient sur un ensemble de traits tirés de l'observation des textes d'entraînement, puis du texte cible (tel qu'annoté par un étiqueteur à n-grammes).

Nous aurons l'occasion de commenter les paramètres suivants qui entrent en jeu dans le fonctionnement du classifieur SVM :

1. librairies utilisées;
2. étiqueteur utilisé;
3. type de kernel;
4. poids relatif des classes;

5. traits utilisés.

Nous avons procédé à de nombreux tests d'optimisation qui ont permis de « quadriller » ces paramètres, c'est-à-dire de comparer les performances de différentes versions du système en faisant varier les paramètres selon un large éventail de valeurs, tel que nous le décrivons plus bas.

### 2.3.1 Librairies utilisées

Nous utilisons la classe *sklearn.svm.SVC* du projet scikit-learn (Pedregosa *et al.*, 2011) qui offre une librairie majoritairement écrite en Python destinée à l'implémentation de tâches liées à l'apprentissage machine.

Nous utilisons également la classe *DictVectorizer()* de la librairie *sklearn.feature_extraction* afin de transformer les traits à multiples dimensions en traits binaires.

### 2.3.2 Étiqueteur

La méthode d'identification SVM utilise deux modules informatiques, un classifieur SVM et un étiqueteur. L'étiqueteur sert d'intermédiaire entre le texte cible et le texte étiqueté.

*texte cible* => Étiqueteur => *texte étiqueté* => Classifieur SVM => *liste de MI*

Nous avons testé le classifieur SVM en lui fournissant alternativement les résultats des deux étiqueteurs décrits plus haut, à n-grammes et Brill. Les performances étaient très similaires, mais bénéficiaient légèrement de l'utilisation de l'étiqueteur à n-grammes.

### 2.3.3 Type de kernel

Nous avons sommairement testé sans succès les différents types de kernels non-linéaires disponibles dans librairie *sklearn.svm* (*polynomial*, *radial basis function* et *sigmoid*). Les

meilleurs résultats que nous avons obtenus étaient produits par le kernel à discrimination linéaire (valeur 'linear' pour le paramètre *kernel*).

### 2.3.4 Poids relatif des classes

Les paramètres *C* et *class_weight* permettent de contrôler l'importance des erreurs de classification au cours du processus d'entraînement du classifieur. Une haute valeur de *C* pour une classe donnée favorise la reconnaissance des contextes marginaux de cette classe mais défavorise la capacité de généralisation du système. Nos expérimentations ont démontré qu'il était bénéfique pour le classifieur d'accorder plus de poids aux contextes marginaux de la classe MI qu'aux contextes marginaux de la classe non-MI, tant pour les signifiants qui sont souvent des MI que pour ceux qui le sont rarement.

Nous avons testé des valeurs de *C* variant entre 0.1 et 100 en combinaison avec des valeurs de *class_weight* pour les deux classes variant entre 1 et 10. Nous avons aussi testé sans succès la valeur 'balanced' de *class_weight*.

Même si d'autres paramètres influencent les performances du système, les observations suivantes s'appliquent de manière générale à la plupart des configurations que nous avons testées.

Une valeur de *C* de 1 pour les deux classes donne généralement des résultats très équivalents à ceux de la méthode d'identification à n-grammes. Le système accorde alors beaucoup d'importance à l'étiquette de l'unité cible fournie par l'étiqueteur à n-grammes.

Une augmentation de la valeur de *C* jusqu'à 50 de manière égale pour les deux classes a généralement eu un effet légèrement positif sur la f-mesure, en conservant l'équilibre entre la précision et le rappel (+-93%).

L'augmentation du poids de la classe MI par l'entremise du paramètre *class_weight* a également permis de graduellement augmenter les capacités de rappel du système. Un multiplicateur

*class_weight* plus élevé que 3 en faveur de la classe MI occasionne cependant une perte de précision du système assez importante pour influencer de manière négative la f-mesure. À l'inverse, l'augmentation du poids de la classe non-MI a occasionné une amélioration de la précision du système, au détriment du rappel.

Les expérimentations qui utilisaient une valeur de C autour de 18 et une valeur de *class_weight* de {1:3} avaient tendance à donner les meilleurs résultats en ce qui a trait à la f-mesure, avec une précision moyenne de 92% et un rappel moyen de 96% au sujet des MI ambigus.

### 2.3.5 Traits utilisés

Le classifieur SVM utilise différents types d'information au sujet du contexte linguistique de l'unité qu'il cherche à identifier. Nous avons testé un grand nombre de ces traits au cours du processus d'optimisation du classifieur. La combinaison des six traits présentés ici s'est avérée avoir l'influence la plus positive sur les performances de la méthode d'identification.

Le tableau 8 montre comment différents traits permettent au système de considérer le problème de classification à partir de différentes sources d'information. Notons que le terme *token* est ici utilisé afin de regrouper les différents types d'unités que l'on retrouve dans les transcriptions, comme les mots, les pauses, les intonations, les marques de citation et les indicateurs de rire.

*Tableau 8 : Traits utilisés pour l'entraînement du classifieur SVM*

| **Sources d'informations** | **Traits** | **Dimensions potentielles** |
|---|---|---|
| Texte cible | Signifiant du token cible | 85 (nb de signifiants qui peuvent être des MI) |
| | Signifiant du token suivant | 27899 (nb de signifiants distincts dans le corpus) |
| Texte étiqueté | Étiquette du token cible | 2 (nb de classes, 'M' ou 'S') |
| | Étiquette du token suivant | 14 (nb d'étiquettes possibles) |
| | Étiquette du token précédent | 14 (nb d'étiquettes possibles) |
| Dictionnaire | Regroupement syntaxique du token cible | 7 (nb de catégories) |

L'exemple (51) permet de situer ces traits de manière concrète à partir d'un extrait d'un énoncé étiqueté. Chacun des couples de parenthèses liste un signifiant et l'étiquette que l'étiqueteur à n-gramme lui a attribué.

(51) [... ('ah', 'M'), ('non', 'M'), ('non', '**M**'), ('**écoute**', '**M**'), ('**j_**', '**PRO**'), ('ai', 'S'), ('fait', 'S'), ('_debut_citation', 'M'), ('hi', 'M'), ('_fin_citation', 'M'), ('j-', 'AM'), ('ça', 'S'), ('ça', 'S'), ('fausse', 'S'), ('là', 'M'), ('ç-', 'AM'), ('ça', 'S'), ('fausse', 'S')]
*[Texte original : CFPQ, sous-corpus 26, segment 6, page 99, ligne 9]*

On voit que le texte cible fournit au système les signifiants du token cible (*écoute*) et du token suivant (*j'*). L'étiqueteur à n-grammes fournit les étiquettes du token cible ('M'), du token précédent ('M') et du token suivant ('PRO'). Le regroupement syntaxique du mot cible ('verbes' dans ce cas) est une information invariable fournie par une simple liste de valeurs dans le but de tenir compte de certaines similitudes de comportement syntaxiques entre certains vocables. Nous donnons plus d'explication sur chacun de ces traits plus bas.

Les nombres de la colonne de droite équivalent au nombre de valeurs que peuvent prendre les traits. Par exemple, 85 unités différentes sont dans la liste des signifiants qui peuvent être des

MI : le trait du signifiant du token cible peut donc être caractérisé par 85 valeurs différentes et exclusives.

#### 2.3.5.1 Signifiant du token cible

Le signifiant de l'unité cible permet au système de différencier les vocables et de prendre en compte cette information lors de la classification. La prise en compte de ce trait a permis une légère amélioration des performances du classifieur.

Ce trait permet également au système de calculer précisément les scores du classifieur pour chaque vocable.

#### 2.3.5.2 Signifiant du token suivant

Ce trait permet d'adresser un problème fondamental de l'étiqueteur à n-grammes qui ne prend pas en compte le contexte syntaxique à droite du token cible.

Comme un grand nombre de valeurs sont possibles pour ce trait, le système a besoin d'un plus grand nombre de bits pour le représenter. Ce trait est par conséquent le plus demandant en termes de ressources informatiques (besoins qui restent très faibles).

#### 2.3.5.3 Étiquette du token cible

Comme nous allons le voir au point 3, l'étiqueteur à n-grammes identifie le plus souvent correctement les MI. Nous estimons que le trait le plus déterminant pour l'entraînement du classifieur SVM est l'étiquette attribuée au token cible par l'étiqueteur à n-grammes. Parmi toutes les informations dont le système dispose, aucune autre n'est plus directement liée à la classe du token cible que cette étiquette.

#### 2.3.5.4 Étiquette du token suivant

L'étiquette du token qui suit le token cible permet un niveau de généralisation supplémentaire quant au contexte syntaxique à droite de celui-ci. Il permet par exemple de regrouper les intonations de différentes forces en un seul trait.

#### 2.3.5.5 Étiquette du token précédent

L'étiquette du token qui précède le token cible est une information qui semble redondante puisque déjà prise en compte par l'étiqueteur à n-grammes (par l'étiqueteur à bi-grammes). Le système bénéficie pourtant nettement de l'utilisation de ce trait.

#### 2.3.5.6 Regroupement syntaxique du token cible

Le groupement syntaxique du token cible est un trait qui est déterminé à partir d'informations extérieures au contexte linguistique du texte cible. Nous avons regroupé les signifiants qui peuvent être des MI en différentes catégories suite à une expérience avec l'étiqueteur à n-grammes qui a permis de démontrer que certains MI ont des comportements syntaxiques similaires. Le processus qui a mené à la détermination des regroupements de signifiants présentés au tableau 9 est expliqué plus bas.

*Tableau 9 : Regroupements syntaxiques des MI*

| **Regroupements** | **Signifiants** |
|---|---|
| sacres | ostie, ostique, ostifie, ostine, crisse, crif, crime, cristie, câlisse, câlique, câline, câlif, tabarnaque, tabarnache, tabarnouche, tabarnique, calvaire, calvince, ciboire, cibole, viarge, sacrement, sacre, sacréfice, simonaque, maudit, mautadit, baptême, batinse, torieu |
| infirmatifs | ⌜pas du tout⌝, pantoute, ⌜pas vraiment⌝, ⌜vraiment pas⌝, ⌜du tout⌝ |
| affirmatifs | ⌜je comprends⌝, ⌜une chance⌝ |
| expressifs | super, malade, cool |
| verbes | regarde, écoute, tiens, arrête, envoye, arrêtez, regardez, écoutez |
| adverbes | ⌜vraiment⌝, ⌜pour vrai⌝, franchement, tellement |

L'étiquetage automatique des MI par l'étiqueteur à n-grammes présente plusieurs difficultés, principalement en raison de la faible fréquence de plusieurs d'entre eux. Une façon de contourner ce problème est de rassembler des signifiants aux comportements syntaxiques similaires afin de constituer des corpus d'entraînement de plus grandes tailles.

Afin de déterminer la pertinence de regrouper certains vocables, nous avons testé et comparé (de manière automatique) une grande quantité de regroupements de vocables possibles et avons retenu les regroupements qui obtenaient les meilleurs résultats lorsque l'étiqueteur à n-grammes les prenait pour cibles.

Le regroupement des vocables SUPER, MALADE et COOL est un exemple où les performances de l'étiqueteur augmentent significativement si on les compare avec celles qu'il obtient lors de l'analyse individuelle de chacune de ces unités.

Il est intéressant de constater que les unités qui gagnent le plus à être regroupées sont souvent issues de classes grammaticales similaires et ont des sens similaires. Ainsi, le regroupement des « infirmatifs » concerne des vocables issus de locutions adverbiales, tandis que le regroupement des expressifs concerne des unités dont les signifiants peuvent jouer le rôle d'adjectifs.

En pratique, l'information au sujet du regroupement syntaxique auquel appartient tel ou tel vocable est communiquée au classifieur SVM à l'aide d'une liste de variables. Les unités qui sont absentes du tableau 9 se voient attribuer la valeur « autres » pour ce trait.

### 2.3.6 Traits impertinents

Nous avons testé un grand nombre de traits au cours de la phase d'optimisation du classifieur SVM et plusieurs d'entre eux se sont révélés avoir un effet négligeable ou néfaste sur les performances du système.

Nous avons sans succès essayé d'utiliser des traits précis, au sujet de la présence de certaines étiquettes ou de signifiants particuliers avant ou après le token cible, comme la présence d'intonation montante ou descendante à la droite du token cible ou la présence de prépositions. Les traits généraux que nous avons présentés en 2.3.5 se sont révélés plus efficaces.

Le trait de l'étiquette du deuxième token qui précède l'unité cible ainsi que le trait de l'étiquette du deuxième token qui suit l'unité cible s'avèrent légèrement bénéfiques pour le classement de certaines unités, mais entraînent globalement une diminution de la performance. Il nous est cependant difficile d'identifier avec précision quels phénomènes linguistiques sont derrière ces fluctuations.

Le signifiant du token précédant le mot cible est un exemple étonnant de trait dont la prise en compte a eu un effet négatif sur les performances du classifieur. Ce trait introduit probablement trop de détails impertinents (bruits) et peu de détails pertinents dans les vecteurs d'entraînement pour occasionner une amélioration du classifieur.

# 3 Évaluation

Après le processus d'optimisation des quatre méthodes d'identification présentées en 2, nous avons testé leur performance au sujet des énoncés du CFPQ préalablement isolés dans le corpus test (voir 1). Les résultats obtenus sont similaires à ceux observés lors de nos tests sur le corpus de travail.

Les performances des différentes méthodes d'identification ont été comparées en examinant les 26 MI ambigus les plus fréquents du corpus de travail qui se retrouvent dans le corpus test. Il s'agit de tous les MI ambigus qui ont une fréquence de plus de dix dans le corpus de travail. Nous ne pouvons pas comparer de manière équitable le traitement des unités très peu fréquentes, puisque la méthode d'identification SVM possède un avantage net sur les autres à ce sujet. Au total, 796 signifiants ont été la cible des méthodes d'identification. 431 de ceux-ci sont des MI (une proportion de 54,15%).

Le tableau 10 présente les différents scores obtenus par chacune des méthodes d'identification sujet du corpus test.

*Tableau 10 : Scores des méthodes d'identification des MI au sujet du corpus test*

| **Méthode** | **Précision** | **Rappel** | **f-mesure** |
|---|---|---|---|
| minimum | 54,15% | 100% | 70,25% |
| n-grammes | 90,37% | 95,82% | 93,02% |
| Brill | 92,17% | 95,59% | 93,85% |
| SVM | 92% | 96,06% | 93,98% |

Nous voyons que la méthode SVM a la plus haute performance globale, grâce à son taux de rappel légèrement plus élevé que celui des autres méthodes et un taux de précision presque aussi élevé que celui de la méthode Brill.

Les différences de performance entre les diverses méthodes d'identification semblent minimes à cause des valeurs élevées des pourcentages comparés. La méthode SVM permet l'élimination d'une erreur sur sept parmi celles produites par la méthode n-grammes.

L'analyse des performances des méthodes d'identification est plus intéressante lorsque réalisée spécifiquement pour chacun des vocables. Le tableau 11 présente les résultats du test pour chacun des vocables cibles en commençant par les MI les plus fréquents du corpus test.

*Tableau 11 : F-mesures obtenues pour chacun des vocables*

| Signifiant | Nb. de signifiants | Nb. de MI | Proportion MI/signif. | Méthode Minimum | Méthode ngrammes | Méthode Brill | Méthode SVM |
|---|---|---|---|---|---|---|---|
| *regarde* | 159 | 121 | 76,1 | 86,43 | 94,86 | **95,24** | 93,55 |
| *ostie* | 95 | 74 | 77,89 | 87,57 | 95,48 | **98,67** | 97,37 |
| *écoute* | 44 | 29 | 65,91 | 79,45 | 100 | 100 | 100 |
| *crisse* | 32 | 24 | 75 | 85,71 | 96 | **97,96** | 95,83 |
| *tiens* | 25 | 22 | 88 | 93,62 | 100 | 100 | 100 |
| *sérieux* | 23 | 20 | 86,96 | 93,02 | 93,02 | 92,68 | 90,48 |
| *seigneur* | 21 | 18 | 85,71 | 92,31 | 92,31 | 92,31 | **94,44** |
| *vraiment* | 195 | 16 | 8,21 | 15,17 | 56,25 | 56,25 | **76,47** |
| *crif* | 17 | 14 | 82,35 | 90,32 | 96,55 | 96,55 | 96,3 |
| *je comprends* | 31 | 12 | 38,71 | 55,81 | 66,67 | 70 | **75,86** |
| *arrête* | 25 | 8 | 32 | 48,48 | 71,43 | 71,43 | **87,5** |
| *mets-en* | 8 | 8 | 100 | 100 | 93,33 | 93,33 | 93,33 |
| *pour vrai* | 12 | 8 | 66,67 | 80 | 100 | 100 | 100 |
| *tabarnaque* | 9 | 8 | 88,89 | 94,12 | 100 | 100 | 100 |
| *une chance* | 8 | 7 | 87,5 | 93,33 | 100 | 100 | 93,33 |
| *envoye* | 8 | 6 | 75 | 85,71 | 85,71 | 85,71 | 85,71 |
| *franchement* | 6 | 6 | 100 | 100 | 100 | 100 | 90,91 |
| *pantoute* | 22 | 6 | 27,27 | 42,86 | 100 | 100 | 100 |
| *pas du tout* | 11 | 6 | 54,55 | 70,59 | 83,33 | 83,33 | 83,33 |
| *tabarnouche* | 7 | 5 | 71,43 | 83,33 | 100 | 100 | 100 |
| *crime* | 5 | 4 | 80 | 88,89 | 100 | 100 | 100 |
| *maudit* | 8 | 3 | 37,5 | 54,55 | 85,71 | 85,71 | **100** |
| *câlisse* | 4 | 2 | 50 | 66,67 | 100 | 100 | 100 |
| *calvaire* | 2 | 2 | 100 | 100 | 100 | 100 | 100 |
| *câlique* | 3 | 1 | 33,33 | 50 | 50 | 50 | **100** |
| *vraiment pas* | 16 | 1 | 6,25 | 11,76 | 100 | 100 | 100 |

Suite à l'analyse de ce tableau, nous remarquons que la méthode n-grammes permet une amélioration généralisée des performances par rapport à la méthode minimum.

Le module Brill obtient les meilleurs scores pour les signifiants les plus fréquents, notamment *regarde*, *ostie* et *crisse*.

En général, la méthode SVM est supérieure à la méthode n-grammes en ce qui concerne l'identification des MI qui se présentent dans une faible proportion par rapport au total des signifiants auxquels ils sont associés (par exemple : *vraiment*, *arrête*, *je comprends*, *maudit*).

Nous discutons plus bas des phénomènes linguistes qui semblent être à la source des différences de performance des quatre méthodes d'identification.

### 3.1 Méthode minimum

Plus un score de la méthode minimum est élevé, moins on devrait s'attendre à une amélioration de ce score de la part des autres méthodes d'identification.

Il est significatif que la méthode minimum obtienne des f-mesures équivalentes ou meilleures que celles des autres méthodes au sujet de certains signifiants comme *mets-en*, *envoye* et *sérieux*. Dans les cas de *mets-en* et *sérieux*, les scores relativement bas des méthodes d'identification plus complexes pourraient s'expliquer par le fait que ces signifiants se trouvent être des MI dans une proportion moins grande dans le corpus d'entraînement que dans le corpus test. *Sérieux*, par exemple, est MI dans 70,15% des cas dans le corpus d'entraînement, mais dans 86,96% des cas dans le corpus test. Les méthodes d'identification qui bénéficient d'un entraînement ont donc tendance à sous-estimer la probabilité que ce signifiant soit un MI.

### 3.2 Méthode n-grammes

La méthode n-grammes permet une nette amélioration des scores pour presque toutes les unités par rapport à la méthode minimum. Elle obtient un score parfait pour les unités *écoute*, *tiens*, *pour vrai*, *tabarnaque*, *une chance*, *franchement*, *pantoute*, *tabarnouche*, *crime, câlisse, calvaire* et *vraiment pas*. Cette méthode semble en général particulièrement efficace pour l'identification des sacres.

#### 3.2.1 Avantages de la méthode n-grammes

Grâce à son utilisation de l'étiqueteur à n-grammes, la méthode n-grammes possède de nombreux avantages sur la méthode minimum.

#### 3.2.1.1 Pauses

L'étiqueteur à n-grammes repère habituellement correctement les contextes où les MI sont en tête de phrase ou produits à la suite de pauses dans le débit de parole. En (52) par exemple, le signifiant *écoute* est précédé d'une pause et est correctement identifié comme un MI :

(52) C : [1ouin c'est SÛR que ça doit être dans l'alimentation **(.)** [2**écoute** (.) c'est SÛR
*[CFPQ, sous-corpus 1, segment 8, page 116, ligne 21]*

#### 3.2.1.2 Étiquette 'EN'

La méthode n-grammes permet d'identifier les utilisations phrastiques des sacres avec la préposition *en*, comme dans l'énoncé (53) où *ostie* s'insère dans une locution adverbiale avec une valeur d'intensification.

(53) Y: [...] Cindy m'a montré des photos aujourd'hui regarde heille c'est beau là (.) ça fait capoter en **ostie** [...]
*[CFPQ, sous-corpus 21, segment 2, page 32, ligne 13]*

#### 3.2.1.3 Étiquette 'DET'

Lorsque qu'un MI est précédé par une étiquette 'DET', comme en (54), il est nom ou intensifieur et est correctement identifié comme une unité intraphrastique par la méthode n-grammes.

(54) Y : [1oui oui je l'ai vue une **ostie** de grosse table là (*en hochant la tête affirmativement*)
*[CFPQ, sous-corpus 21, segment 3, page 36, ligne 16]*

#### 3.2.1.4 Étiquette 'PRO'

Grâce à l'étiquetage des pronoms ('PRO') qui peuvent servir de sujets et des pronoms réfléchis, l'étiqueteur reconnaît le caractère intraphrastique de sacres utilisés comme verbes. Dans l'extrait

(55), le pronom *se* précède le signifiant *crisse* et permet l'étiquetage de ce dernier comme unité intraphrastique.

(55) MC : oui: mais (1") t'sais (.) c'est parce que Nathalie a comme pris euh le bord de: (.) <p<elle **se crisse** d'elle [1là>>
*[CFPQ, sous-corpus 22, segment 6, page 97, ligne 9]*

La plupart des utilisations intraphrastiques du signifiant *écoute* sont des verbes. L'étiqueteur identifie correctement ces unités à l'aide de l'étiquette 'PRO' du corpus d'entraînement qui permet de prendre en compte les pronoms susceptibles d'être des sujets à la première ou troisième personne du singulier. En (56) par exemple, la présence du *il* avant *écoute* fait en sorte que le verbe est étiqueté correctement.

(56) C : [3**il écoute** rien
*[CFPQ, sous-corpus 3, segment 2, page 37, ligne 20]*

#### 3.2.1.5 Intonation

La présence d'une marque d'intonation à gauche d'un mot cible semble faire en sorte que l'étiqueteur à n-grammes identifie correctement plusieurs MI. Dans l'énoncé (57), le signifiant *écoute* suit une marque d'intonation descendante et est correctement classé comme un MI par la méthode n-grammes.

(57) G : AH ben [1là\ **écoute** euh:: ben l'eau est JAUNE e:lle est JAUNE là\ elle est pas euh:
*[CFPQ, sous-corpus 24, segment 3, page 39, ligne 2]*

#### 3.2.1.6 Trigrammes

Grâce à la prise en compte des deux unités qui précèdent chacun des signifiants à classer, les séquences où un adjectif s'insère entre un déterminant et un sacre intraphrastique sont repérées. Lorsqu'il étiquette l'extrait (58), l'étiqueteur identifie correctement *criss* comme n'étant pas un MI

même si l'adverbe *mêmes* se trouve intercalé entre celui-ci et le déterminant *les* qui indique son caractère intraphrastique.

(58) S : il nourrit s- qui le monde qui sont déjà là •oui mais ça va créer des jobs° (.) •ouin regarde (*en inclinant la tête et en fermant les yeux comme en signe d'exaspération*) ça s- tu <f<pour[1rais/>> <ff<tu pourrais tu s- tu pourrais créer **les mêmes criss**>> de jobs pis l'exploiter par nous-autres/ [...]
*[CFPQ, sous-corpus 28, segment 7, page 89, ligne 10]*

### 3.2.2 Inconvénients de la méthode n-grammes

Nous examinons sommairement quelques-uns des phénomènes qui font en sorte que la méthode n-grammes commet des erreurs dans ses efforts d'identification des MI

#### 3.2.2.1 Non prise en compte du contexte à droite du mot cible

La principale faiblesse de la méthode n-grammes est qu'elle ne prend pas en compte les unités à droite de l'unité cible.

Dans l'extrait (59), l'utilisation de la construction « *je comprends* » de manière intraphrastique est évidente grâce à la particule *pas* qui la suit. Puisqu'elle n'a pas accès à cette information, la méthode n-grammes identifie erronément la construction comme étant un MI, probablement parce qu'elle se trouve au début d'un tour de parole.

(59) D : **je comprends pas**
*[CFPQ, sous-corpus 3, segment 1, page 3, ligne 13]*

La méthode n-grammes ne permet également pas de tenir compte des intensifieurs de phrases (tel que discuté au chapitre 2-1.6.3.2). Dans l'extrait (60), l'étiqueteur confond le signifiant *ostie* avec un MI, malgré la conjonction *que* placée à sa droite.

(60) Y :[2AH ton- ben pesant ah ouais (*en levant ses bras comme s'il tenait une grosse scie dans ses mains*) (.) **ostie que** ça travaille pas ben ça
*[CFPQ, sous-corpus 21, segment 8, page 134, ligne 4]*

Nous allons voir plus loin que plusieurs erreurs de ce type sont corrigées par les méthodes Brill et SVM.

#### 3.2.2.2 Analyse individuelle des vocables

Une autre faiblesse de la méthode n-grammes est qu'elle ne peut tirer de conclusions au sujet d'un vocable à partir de l'observation d'un autre vocable similaire.

Étant donné la basse fréquence de certaines unités dans le corpus d'entraînement, l'étiqueteur ne possède souvent pas d'information sur plusieurs contextes d'utilisation de ces unités. L'extrait (61) met en scène la construction « déterminent + nom », caractéristique à plusieurs sacres.

(61) J : [1ah ok **le CRIF** a décidé [2ça/ (*en pointant vers le côté comme pour désigner la décision dont elle parle*)
*[CFPQ, sous-corpus 17, segment 6, page 79, ligne 16]*

L'étiqueteur à n-grammes échoue cependant à identifier le signifiant *crif* de ce tour de parole comme étant une unité intraphrastique. Nous pouvons imaginer que si l'étiqueteur avait pu tenir compte des nombreux cas où le signifiant *crisse*, qui a un comportement très similaire à *crif*, se trouve utilisé de manière intraphrastique à la suite d'un déterminant dans le corpus d'entraînement, il aurait pu correctement étiqueter le signifiant *crif* de l'extrait (61).

Cette étroitesse de l'étiqueteur à n-grammes fait évidemment aussi en sorte qu'il ne peut pas étiqueter convenablement les signifiants qu'il n'a pas rencontrés du tout dans son corpus d'entraînement. Nous allons voir que la méthode SVM échappe à cette faiblesse.

#### 3.2.2.3 Absence d'analyse syntaxique profonde

Dans l'extrait (62), la séquence « •ah:° » qui est intercalée juste avant l'adverbe *vraiment* fait en sorte que l'étiqueteur à n-grammes identifie erronément *vraiment* comme un MI, puisque les marques de discours direct sont considérées comme des indicateurs de ruptures syntaxiques par notre système d'étiquettes.

(62) M : [2c'est pas bon fait que t'sais c'est comme **•ah:° vraiment** une business de de (.) c'est dégueulasse là t'sais [3c'est (.) vraiment exploiter du monde ¤
*[CFPQ, sous-corpus 10, segment 9, page 116, ligne 13]*

Cette erreur peut être vue comme un exemple de l'incapacité de la méthode n-grammes à prendre en compte les structures syntaxiques complexes qui dépassent le contexte immédiat du mot cible. La portée de l'étiqueteur à n-grammes est limitée à deux unités à gauche du mot cible, ce qui est évidemment trop court pour englober plusieurs types de relations syntaxiques.

### 3.2.3 Conclusion au sujet de la méthode n-grammes

Suite à l'examen des performances de la méthode n-grammes au sujet de l'identification des MI du corpus test, nous pouvons tirer quelques conclusions.

1. En général, un grand nombre d'occurrences d'un signifiant dans un corpus d'entraînement favorise de bonnes performances de la part de l'étiqueteur à n-grammes. Ce phénomène est inévitablement lié à n'importe quelle méthode d'analyse statistique, mais l'étiqueteur à n-grammes est particulièrement fragile à ce sujet (en comparaison avec le classifieur SVM) à cause de son incapacité à tisser des liens entre les différents vocables. L'étiqueteur à n-grammes considère chacun des vocables de manière isolée et est incapable de traiter des vocables nouveaux.

2. Parmi les raisons qui ont justifié l'ensemble d'étiquettes que nous utilisons pour nos corpus d'entraînement (2.1.3), six semblent jouer des rôles particulièrement importants dans les performances de l'étiqueteur n-grammes. Les débuts et les fins de tour de parole, la présence de

pauses dans le débit de la parole et la proximité d'unités lexicales extraphrastiques (comme les MI et les marqueurs de balisage) ont une influence positive sur l'identification des MI, tandis que la proximité de la préposition *en*, de déterminants et de pronoms ont une influence positive sur l'identification des unités intraphrastiques. La séquence « *en* + sacre » est particulièrement importante pour le repérage des sacres intraphrastiques.

3. Les sacres en général sont bien traités par l'étiqueteur n-grammes. La relative régularité de leurs usages, tant intraphrastiques qu'extraphrastiques, est la raison principale de ce phénomène.

Nous allons maintenant voir les avantages qu'offre l'utilisation du module Brill en plus du module n-grammes pour l'étiquetage des MI.

### 3.3 Méthode Brill

Certaines faiblesses de l'étiqueteur à n-grammes peuvent être contrées par l'utilisation du module Brill tel que décrit en 2.2. Par la nature même de son fonctionnement, plus l'étiqueteur à n-grammes produit d'erreurs, plus le module Brill a la possibilité de résoudre des erreurs.

La méthode Brill obtient un meilleur résultat que la méthode n-grammes au sujet des signifiants *regarde*, *ostie*, *crisse* et *je comprends*. Nous examinons plus bas les résultats du module Brill au sujet des unités *ostie*, *crisse* et *je comprends*.

#### 3.3.1 Identification de OSTIE par la méthode Brill

La prise en compte du signifiant à droite de l'unité cible permet à la méthode Brill d'améliorer sa performance au sujet du signifiant *ostie.* Plus précisément, la méthode Brill commet 5 erreurs de précision de moins que la méthode n-grammes grâce à l'identification des contextes où *ostie* est utilisé comme intensifieur de phrase avec *que*.

Nous pouvons retracer le cheminement informatique que la méthode Brill doit accomplir pour arriver à ce résultat en examinant les contextes linguistiques en œuvre dans le corpus d'entraînement et le corpus test.

Grâce à l'analyse des tours de parole (63) et (64) du corpus d'entraînement, l'étiqueteur Brill a pu construire les règles présentées au tableau 12.

(63) J : [1<pp<**ostie que** ça a l'air [2bon ça>> (*dit avec émotion*)
[CFPQ, sous-corpus 22, segment 6, page 88, ligne 18]

(64) MC : ah: **ostie qu'**il est [1moumoune (*dit avec exaspération*)
[CFPQ, sous-corpus 22, segment 5, page 73, ligne 6]

*Tableau 12 : Règles de transformation du module Brill au sujet du signifiant* ostie

| **Règles de transformation** | **Nb de cas** | **Explications** |
|---|---|---|
| M->S if Word:ostie@[0] & Word:qu_@[1] | 10 | L'unité *qu'* à droite du signifiant *ostie* indique que ce dernier est instraphrastique. |
| M->S if Word:ostie@[0] & Word:que@[1] | 10 | L'unité *que* à droite du signifiant *ostie* indique que ce dernier est instraphrastique. |

Dans le tableau 12, nous voyons que chacune des règles de transformation construites par le module Brill est basée sur l'observation de 10 contextes du corpus de travail. La méthode Brill a ensuite pu utiliser ces règles afin d'identifier correctement le signifiant *ostie* lorsque utilisé dans des contextes similaires du corpus test.

### 3.3.2 Identification de CRISSE par la méthode Brill

Par un procédé très similaire à celui vu plus haut, le module Brill permet l'amélioration de la précision au cours de l'identification des signifiants de *crisse*.

*Tableau 13 : Règle de transformation du module Brill au sujet du signifiant* crisse

| **Règle de transformation** | **Nb de cas** | **Explication** |
|---|---|---|
| M->S if Word:crisse@[0] & Word:que@[1] | 4 | L'unité *que* à droite du signifiant *crisse* indique que ce dernier est intraphrastique. |
| M->S if Word:crisse@[0] & Word:de@[1] | 3 | L'unité *de* à droite du signifiant *crisse* indique que ce dernier est intraphrastique. |

La deuxième règle de transformation du tableau 13 permet de repérer les contextes où *crisse* est utilisé comme intensifieur adjectival (comme en (65)), mais pourrait occasionner des erreurs de rappel dans les cas où le MI est utilisé dans des enchaînements de sacres liés avec *de*.

(65) Y : de la **crisse de** marde des ost- (.) [...]
*[CFPQ, sous-corpus 21, segment 1, page 13, ligne 18]*

### 3.3.3 Identification de ⌜JE COMPRENDS⌝ par la méthode Brill

La prise en compte du signifiant à gauche de l'unité cible permet d'améliorer la précision du système lors de l'identification du MI ⌜JE COMPRENDS⌝. La règle de transformation du tableau 14 permet l'identification adéquate du signifiant *je comprends* utilisé de manière intraphrastique lorsque précédé du pronom *moi*.

*Tableau 14 : Règle de transformation du module Brill au sujet du signifiant* je comprends

| **Règle de transformation** | **Nb de cas** | **Explication** |
|---|---|---|
| M->S if Word:je_comprends@[0] & Word:moi@[-1] | 2 | L'unité *moi* à gauche du signifiant *'je comprends'* indique que ce dernier est intraphrastique. |

Dans l'extrait (66) du corpus test, l'unité *moi* à gauche de *je comprends* indique à la méthode d'identification Brill que cette construction n'est pas un MI.

(66) MY : ouais Le vieux dans le bas du fleuve c'est ça •il est il est assis à: compter les° **moi je comprends** •assis à compter les DIVANS°
*[CFPQ, sous-corpus 19, segment 9, page 94, ligne 11]*

### 3.3.4 Conclusion au sujet de la méthode d'identification Brill

Suite à l'analyse des résultats du module Brill au sujet des unités qui en bénéficient le plus, nous avons quelques conclusions à tirer.

1. Le fait que l'identification de plusieurs MI ne bénéficie pas de la méthode Brill laisse croire que le contexte lexical à droite est généralement moins important que le contexte lexical à gauche pour l'identification des unités extraphrastiques.

2. Rappelons cependant que puisque le module Brill fonctionne en analysant les erreurs d'un autre étiqueteur, il ne peut pas être productif dans un contexte où l'étiqueteur n-grammes ne fait pas ou presque pas d'erreurs, comme c'est le cas pour plusieurs vocables.

3. L'analyse des résultats du module Brill fait ressortir la principale faiblesse de notre système d'étiquetage, particulièrement le fait que nous attribuons une même étiquette à la plupart des lexèmes intraphrastiques, comme les noms, les adjectifs, les verbes et les adverbes. Il se pourrait que le module Brill soit plus efficace s'il bénéficiait d'un ensemble d'étiquettes plus précises, mais nous n'avons jusqu'à maintenant pas réussi à déterminer un tel ensemble. Rappelons que l'utilisation d'étiquettes précises a tendance à diminuer les performances de l'étiqueteur à n-grammes au sujet des signifiants peu fréquents.

## 3.4 Méthode SVM

La méthode d'identification SVM permet une amélioration de la précision et du rappel par rapport à la méthode n-grammes au sujet des signifiants *seigneur*, *vraiment*, *je comprends*, *arrête*, *maudit* et *câlique*.

Il est difficile de reconnaître exactement les phénomènes qui expliquent la performance supérieure de la méthode SVM comparativement aux autres méthodes. La nature complexe du mécanisme de classification de cette méthode rend hasardeuse l'analyse d'exemples particuliers. Nous tenterons plus bas d'expliquer certains phénomènes qui aident (ou nuisent) à la performance de la méthode SVM par l'analyse des erreurs qu'elle évite ou de celles qu'elle occasionne.

### 3.4.1 Identification de SEIGNEUR par la méthode SVM

La conjonction du trait qui concerne l'étiquette du mot qui précède l'unité cible et celui qui concerne le signifiant qui suit l'unité cible permet l'identification adéquate de la construction « seigneur des anneaux » présente à deux reprises dans le corpus test.

Dans l'extrait (67), la construction « DET + *seigneur* + *des* » est correctement identifiée comme une utilisation intraphrastique du signifiant *Seigneur* par la méthode SVM. Ce signifiant était incorrectement identifié par la méthode n-grammes.

(67) [...] raconte-moi le le (.) l'histoire **du Seigneur des** anneaux là (1,3") c'est pas terrible comme histoire
*[CFPQ, sous-corpus 13, segment 4, page 51, ligne 15]*

### 3.4.2 Identification de VRAIMENT par la méthode SVM

Le contexte à droite, en combinaison avec l'étiquette à gauche de l'unité cible, semble avoir permis au classifieur SVM de repérer quatre contextes où le MI *vraiment* est précédé et suivi de ruptures syntaxiques. Le signifiant *vraiment* de l'exemple (65) est incorrectement classé comme une unité intraphrastique par la méthode n-grammes, mais correctement classé comme MI par la méthode SVM.

(68) V : [1en plus à Katimavik **là vraiment t'sais** Val les exemples qu'elle m'a donnés mon amie qui travaillait à Katimavik c'est genre ses participants à un moment donné ils TRIpaient c'était comme •on va te faire notre MEILLEURE bouffe là notre recette de SOUPE surprise° qu'ils avaient fait en Alberta avant d'arriver à Drummond leur soupe surprise c'est (RIRE) tu prends tous les restants de la semaine [2dans le frigo ¤
*[CFPQ, sous-corpus 10, segment 11, page 136, ligne 21]*

Les unités *là* et *t'sais* qui entourent le *vraiment* de l'extrait (68) portent la même étiquette ('M') dans le corpus d'entraînement de l'étiqueteur à n-grammes. Grâce à cette étiquette large, le classifieur SVM peut tirer de l'information au sujet de contextes où le signifiant *vraiment* est inséré entre des unités différentes mais aux rôles syntaxiques similaires, comme *ouais*, *ah*, *hum* ou des épisodes de rire. Ces unités portent toutes la même étiquette ('M').

### 3.4.3 Identification de ⌜JE COMPRENDS⌝ par la méthode SVM

Comme nous allons d'ailleurs le voir en détail au chapitre 4, il peut être difficile, même pour un humain, de faire la distinction entre les usages intraphrastiques et extraphrastiques du signifiant *je comprends*, probablement à cause d'une frontière sémantico-pragmatique floue entre certains de ces emplois.

Notons par exemple que, en raison de l'utilisation intransitive du verbe COMPRENDRE, la production de *je comprends* à la fin d'un tour de parole n'est pas nécessairement liée à son utilisation extraphrastique. La construction *Je comprends.* peut par ailleurs s'utiliser par elle-même et former une phrase toute seule. Cette phrase n'a cependant pas suivi le processus de pragmaticalisation qu'a subi le MI ⌜JE COMPRENDS⌝. Celui-ci a notamment une composante assertive.

Nous avons vu que plusieurs phénomènes de prosodie sont importants dans l'énonciation des MI. Des irrégularités phonologiques de constructions comme *je comprends* peuvent parfois être une source de confusion pour les systèmes statistiques.

Dans l'extrait (69) par exemple, une intonation montante suit la construction *je comprends* et laisse erronément croire à la méthode SVM qu'il s'agit d'un MI.

(69) J : [1heille [2mais **je comprends/** [3pas à peu près
*[CFPQ, sous-corpus 15, segment 6, page 105, ligne 13]*

En effet, une marque d'intonation montante suit cinq fois plus souvent cette construction lorsqu'elle est utilisée comme mot-phrase dans le corpus d'entraînement. Il est intéressant de constater que la méthode n-grammes n'a pas commis d'erreur au sujet de l'extrait (69).

### 3.4.4 Identification de ARRÊTE par la méthode SVM

La méthode SVM repère deux occurrences du MI ARRÊTE de plus que la méthode n-grammes. Une de ces occurrences concerne l'unité *donc* qui accompagne parfois l'énonciation de certains MI par phénomène de collocation discursive (voir chapitre 2-1.9.3).

Dans l'extrait (70) du corpus test, où l'énonciateur parle en imitant un enfant, le MI ARRÊTE est produit en compagnie de *donc*. Probablement parce qu'elle tient compte de l'unité à droite de l'unité cible, la méthode SVM classifie adéquatement le MI ARRÊTE de cet extrait, contrairement à la méthode n-grammes.

(70) G : •**arrête donc** [1 ben oui je le sais (inaud.)°

### 3.4.5 Conclusion au sujet des résultats de la méthode SVM

Suite à l'analyse des résultats de la méthode d'identification SVM, il nous est permis de tirer les conclusions suivantes :

1. Dans plusieurs cas, les améliorations apportées par la méthode SVM sont similaires à celles apportées par la méthode Brill. La prise en compte du contexte à droite de l'unité cible semble être le phénomène en jeu le plus important à ce sujet.

2. Contrairement au module n-grammes qui recueille de l'information pour chacun des signifiants de manière isolée, le module SVM peut apprendre de chacun des exemples qu'on lui donne. Cette capacité permet notamment à la méthode SVM de classifier avec succès des signifiants qu'elle n'a jamais rencontrés dans sa phase d'entraînement.

3. Nous croyons que l'ensemble d'étiquettes larges de l'étiqueteur à n-grammes et son utilisation lors de l'entraînement du classifieur SVM permet un niveau d'abstraction supplémentaire au sujet du rôle syntaxique de plusieurs unités et entraîne une amélioration de la performance de la méthode SVM. Le regroupement de plusieurs unités qui marquent des ruptures syntaxiques sous l'étiquette 'M' semble jouer un rôle important à cet effet.

## 4 Conclusion au sujet de l'identification automatique des MI

L'identification de certains MI dans un texte oral transcrit peut être difficile pour un être humain comme pour un système informatique. Nous pourrions sans doute utiliser des systèmes à l'analyse beaucoup plus fine si nous avions accès à plus de données. Le haut niveau de généralisation des méthodes d'identification est nécessaire dans un contexte où nous utilisons un corpus relativement limité. Rappelons que les MI ne sont pas tous des unités fréquentes, surtout les MI qui sont ambigus.

# Chapitre 4 :
# Caractérisation sémantique des MI

Dans ce chapitre, nous porterons principalement notre attention sur les caractéristiques sémantiques du champ lexical des MI dans l'objectif de proposer une méthode de description sémantique formelle et cohérente à leur sujet.

L'identification des MI ne serait pas utile à la compréhension d'un texte sans moyens de caractériser sémantiquement ces MI. Plusieurs classements sémantiques ont été proposés au sujet de champs lexicaux liés aux interjections ou aux marqueurs pragmatiques (MP) en général. Nous avons présenté certains de ceux-ci au chapitre 2 dans le cadre de notre revue d'études sur l'analyse sémantique automatique des MP. Dans cette partie, nous aurons également l'occasion de présenter plusieurs travaux qui ne sont pas nécessairement liés aux MP, mais qui portent sur l'expression de sentiments par différents moyens linguistiques.

Comme il a été mentionné au chapitre 1, nous nous appuyons sur les principes généraux de la lexicologie explicative et combinatoire (LEC) (Mel'čuk *et al.*, 1995) et ceux de la Métalangue sémantique naturelle (MSN) (Wierzbicka, 1972, 1980, 1986, 1997, 1999, 2011, Goddard, 2013, 2013, 2014, 2015; Goddard et Ye, 2014) pour la description sémantique des MI. Nous considérons ainsi que la lexie est l'unité de base du lexique et que le sens d'une lexie est le plus exactement représenté à l'aide d'une paraphrase en langue naturelle.

Nous avons cependant dû adapter ces méthodes lexicographiques de manière à simplifier considérablement l'analyse computationnelle des signifiés que nous étudions. Nous proposons un système qui caractérise sémantiquement les MI à partir d'un ensemble limité de courtes paraphrases, construites à l'aide d'un vocabulaire réduit. Ce choix entraîne forcément une perte de précision dans l'analyse sémantique et une perte de précision dans le découpage des lexies. Les lexies de plusieurs vocables sont ainsi regroupées lorsque les distances sémantiques entre celles-ci sont trop faibles pour entraîner une différenciation dans leur traitement computationnel.

La polysémie est le plus grand défi qu'un système d'analyse automatique doit relever au sujet des MI. 37 des vocables que nous avons sélectionnés n'ont qu'un seul sens, 39 ont deux sens et 6 ont plus de deux sens. Malheureusement, tant pour un humain que pour un système informatique, la distinction des différentes lexies des MI n'est pas toujours possible, à plus forte raison quand il s'agit de transcriptions de discours oraux. L'information au sujet de la prosodie et de la gestuelle des énonciateurs est en effet limitée dans le CFPQ, comme dans tout corpus oral. Nous allons suggérer quelques solutions au problème de la polysémie en conclusion de cette thèse.

La première partie de ce chapitre servira à présenter les réflexions qui ont mené à la mise en place de notre système de description sémantique, tandis que la seconde partie servira à décrire chacun des MI à l'étude à l'aide de ce système.

# 1 Système de description sémantique des MI

Le tableau 15 présente les différents mots-clés de notre système de description sémantique et les paraphrases qui leur correspondent :

*Tableau 15 : Description des actes illocutoires et connotations liés aux MI*

| **Types d'actes illocutoires** | **Mots-clés** | **Paraphrases** |
|---|---|---|
| Expressifs | Bien | c'est bien |
| | Mauvais | c'est mauvais |
| | Se sentir bien | je ressens quelque chose de bien |
| | Se sentir mal | je ressens quelque chose de mauvais |
| | Inattendu | je ne savais pas que cela allait être comme cela |
| | Hors du commun | je me sens comme quelqu'un qui voit quelque chose de très grand |
| | Douleur | je ressens de la douleur |
| | Dégoût | je ressens du dégoût |
| | Forte émotion | je ressens une forte émotion |
| Assertifs | Affirmatif | c'est vrai |
| | Infirmatif | ce n'est pas vrai |
| | Infirmatif partiel | ce n'est pas complètement vrai |
| Directifs | Attention | pense à cela |
| | Arrêt | ne fais pas cela |
| | Question | est-ce que c'est vrai? |
| | Encouragement | fais cela |
| **Types de connotations** | **Mots-clés** | **Paraphrases** |
| Stylistiques | Tabou | certaines personnes peuvent ressentir quelque chose de mauvais quand elles entendent ce mot |

Nous aurons graduellement l'occasion de présenter les observations et réflexions qui ont mené à la construction de ce système de description. Des travaux entrepris dans les cadres théoriques de la Métalangue sémantique naturelle et du Modèle Sens-Texte seront les sources principalement citées. Il semble que les concepts associés aux émotions se construisent souvent autour de « scénarios cognitifs prototypiques » (Goddard, 2014, p. 9). Nous tâcherons de décrire ces scénarios, parfois à l'aide d'explications proposées par d'autres chercheurs.

### 1.1 Les expressifs

Pour plusieurs MI, la fonction illocutoire expressive est évidente. Les actes expressifs ont pour but de communiquer des informations au sujet de la subjectivité d'un énonciateur à propos d'états du monde (dont la vérité est présupposée). Searle (1979) définit cette classe de façon à ce qu'elle regroupe plusieurs actes pouvant être nommés par des verbes performatifs (comme l'excuse, les félicitations, le souhait de bienvenue...).

Un bon nombre de systèmes de classement des émotions ont été proposés en linguistique et en psycholinguistique. En guise d'exemple, l'article de Alm, Roth et Sproat (2005) propose une annotation sémantique automatique des émotions à l'aide de 7 classes : colère, dégoût, peur, joie, tristesse, surprise positive et surprise négative. Les MI associés à ces classes peuvent être grossièrement caractérisés à l'aide des paraphrases expressives du tableau 15. Une unité qui exprime la peur par exemple, peut réaliser des actes illocutoires qu'il est possible de caractériser (peut-être avec plus de précision) à l'aide d'au moins une des paraphrases associées aux mots-clés **inattendu** ou **hors du commun** suivi de la paraphrase associée au mot-clé **se sentir mal**. Ces simples agencements de mot-clés permettent ainsi de décrire trois types de peur, celle qui est causée par la surprise, celle qui est causée par une vision effrayante et celle qui est causée par la combinaison de ces deux phénomènes.

#### 1.1.1 Bien

Dans Goddard et Ye (2014), les auteurs explorent différents cadres sémantiques qui mettent en jeu les concepts de *bien* et de *mal* en lien avec les verbes de base *penser* et *ressentir*. Ces propositions construites à l'aide de primitifs sémantiques forment la base de notre description des actes illocutoires propres aux MI.

Plusieurs MI servent à exprimer une appréciation positive au sujet d'états du monde. Le premier cadre sémantique que nous tirons de Goddard et Ye (2014) peut être associé aux actes expressifs approbatifs que nous identifions par le mot-clé **bien**.

quelqu'un PENSE quelque chose de (bien/mal) au sujet de (quelqu'un/quelque chose)
(Traduit de Goddard et Ye, 2014, p. 7)

Puisque les MI expriment toujours le point de vue de leurs énonciateurs, nous pouvons utiliser la paraphrase déclarative « c'est bien » pour représenter les actes expressifs réalisés par ceux-ci dans ce cadre sémantique.

Les unités COOL, SUPER, WOW, YOUPI, ⌜UNE CHANCE⌝, FIOU et OUF sont des exemples de MI dont le signifié d'au moins une lexie inclut la composante **bien**. Dans l'extrait (71), l'énonciatrice A produit le MI SUPER pour indiquer qu'elle approuve l'état de la situation décrit par sa coénonciatrice AN.

(71) A : pis elle elle a quoi un entreprise/
AN : elle a un dépanneur (*dit comme pour exprimer une évidence en se penchant vers Agathe comme pour insister sur ses propos*)
R : un dépanneur
A : ah ben oui (*en levant un bras en l'air comme pour signifier que c'est un endroit idéal*)
[1**super** (*dit en souriant*)
R : [1c'est parfait (.) parfait [...]
*[CFPQ, sous-corpus 20, segment 9, page 121, lignes 11-15]*

Le SUPER de l'extrait (71) peut en partie être remplacé par l'énoncé « c'est bien » qui rend compte de l'acte expressif qu'il réalise.

### 1.1.2 Mauvais

Plusieurs MI servent à exprimer une appréciation négative au sujet d'états du monde. Le mot-clé **mauvais** est le pendant négatif du mot-clé **bien** et s'inscrit dans le cadre sémantique présenté en 1.1. La paraphrase déclarative « c'est mauvais » représente cet acte illocutoire.

Les unités COUDON, ÉCOUTE, FRANCHEMENT, ⌜EH BOY⌝, OUPS, OUPELAILLE, VOYONS et ZUT sont des exemples de MI dont le signifié d'au moins une lexie inclut la composante **mauvais**. Dans l'extrait (72), l'unité OUPS produite par l'énonciatrice L véhicule certainement une composante désapprobative au sujet de l'état de la situation décrit par sa coénonciatrice J, ce qui n'implique pas que L désapprouve sa coénonciatrice J ou ce qu'elle a dit. Elle fait savoir qu'il est comiquement mauvais qu'un chat s'attaque à des beaux oiseaux.

(72) J : {pis;ok} fait que ça c'est nouveau cette année fait qu'on a beaucoup de pics (.) on a un chat aussi alors [1euh:
L : [1**oups::**
*[CFPQ, sous-corpus 18, segment 3, page 28, lignes 15-16]*

Notons que la composante **inattendu** est également associée à OUPS.

### 1.1.3 Se sentir bien

Nous avons vu que des MI servent à exprimer une appréciation positive ou négative au sujet d'états du monde. Les deux paraphrases suivantes concernent l'expression d'une humeur positive ou négative.

Le deuxième cadre sémantique que nous tirons de Goddard et Ye (2014) peut être associé aux actes expressifs que nous identifions par les mots-clés **se sentir bien** et **se sentir mal**.

quelqu'un RESSENT quelque chose de (bien/mal)
(Traduit de Goddard et Ye, 2014, p. 7)

Afin de représenter les actes illocutoires réalisés par les MI, nous avons recours au pronom JE qui forme les paraphrases « je ressens quelque chose de bien » et « je ressens quelque chose de mauvais ». Rappelons-le, les MI ont la caractéristique de toujours exprimer le point de vue de leur énonciateur.

Il est pertinent de distinguer l'expression du fait que quelque chose est considéré bien ou mal et l'expression du fait qu'une personne ressent un sentiment positif ou négatif dans la mesure où, comme nous allons le voir au point 2, certaines unités ne réalisent qu'un de ces actes à la fois.

Les MI COOL, FIOU, OUF, SUPER, TIENS, WOW et YOUPI ont au moins une lexie qui réalise un acte expressif qu'il est possible de caractériser avec le mot-clé **se sentir bien**. Le MI SUPER de l'extrait (71) a une telle composante.

### 1.1.4 Se sentir mal

Le cadre sémantique présenté en 1.1.3 peut servir à caractériser l'expression d'une humeur négative. Les MI COUDON, MERDE, OUPS, VOYONS, ZUT ainsi que les sacres et les substituts de sacres sont des exemples d'unités dont au moins une lexie peut être caractérisée par la paraphrase «  je ressens quelque chose de mauvais ».

Dans l'extrait (73), l'énonciateur JN utilise le MI MERDE dans un discours rapporté afin de représenter son émotion négative suite à la fin d'un film.

(73) JN : oui quand c’est fini j’étai:s •ah **merde/** déjà fini/ (.) encore/° (RIRE)
*[CFPQ, sous-corpus 28, segment 5, page 72, ligne 16]*

### 1.1.5 Inattendu

Goddard (2013, 2014) a décrit des adjectifs et des interjections liés à la surprise et à l'étonnement. Ses définitions à l'aide de la MSN permettent de distinguer les éléments qui sont tout simplement nouveaux, des éléments qui sont difficiles à croire ou difficiles à comprendre et ceux qui suscitent un sentiment d'émerveillement.

Les composantes centrales de l'adjectif *surprised* sont expliquées à l'aide du scénario cognitif suivant par Goddard (2014) :

[A] Someone X was surprised (at that time) :
[...]
a short time before it was like this:
– something happened
[...]
after this, this someone thought about it like this:
I didn't know before that it will be like this
I know it now
(Goddard, 2014, p. 6)

La composante principale de cette définition est l'énoncé qui se traduit par « je ne savais pas plus tôt que cela allait être comme cela ». Ainsi, la définition de Goddard décrit l'état d'une personne mise en face d'un élément nouveau. La plupart des MI sont produits en réaction à quelque chose, souvent en réaction à quelque chose de nouveau. Mais la particularité des marqueurs liés à l'acte illocutoire identifié par le mot-clé **inattendu** est qu'ils expriment un sentiment par rapport à ce fait.

Les composantes qui figurent dans l'explication sémantique de Goddard peuvent être regroupées de façon à décrire plus particulièrement l'acte expressif propre aux MI : « je ne savais pas que cela allait être comme cela ». L'utilisation du verbe *savoir*, plutôt que le verbe *penser*, implique que l'énonciateur n'avait pas nécessairement une opinion préalable au sujet de l'élément nouveau.

Les MI ARRÊTE, COUDON, FIOU, FRANCHEMENT, HEIN, MALADE, ⌜MON DIEU⌝, OUF, OUPELAILLE, OUPS, ⌜POUR VRAI⌝, ⌜REGARDE DONC⌝, SÉRIEUX, TIENS, ⌜UNE CHANCE⌝, VOYONS et WOW ont tous au moins une lexie liée à un acte illocutoire qui exprime une émotion face à quelque chose d'inattendu.

Dans l'extrait (74), l'énonciateur B produit ⌜REGARDE DONC⌝ suite à l'état de fait décrit par sa coénonciatrice S. Il est étonné par le contexte inusité d'une exposition.

(74) S : une exposition/ là [1en dessous de l'eau
[...]
S : [3fait que tu peux louer l'équipement de de plongée/
B : ah **regarde donc** [1ça/
*[CFPQ, sous-corpus 15, segment 10, page 170, ligne 7]*

### 1.1.6 Hors du commun

Un élément lié au mot-clé **hors du commun** épate non par son caractère improbable, mais par sa « grandeur ». La dernière ligne de la définition de l'interjection *wow!* de Goddard (2013) fait appel à un scénario physique afin de décrire l'émotion de l'énonciateur devant quelque chose d'impressionnant :

Wow!
I think like this: "this is very good"
I didn't know before that it can be like this
I feel something very good because of this
**I feel like someone can feel when this someone sees something very big**
(Goddard, 2013, p. 5)

Le sentiment d'épatement exprimé par *wow* en anglais se retrouve dans le signifié de plusieurs MI du français que nous analysons ici. Nous décrivons l'acte expressif associé à ce sentiment à l'aide d'une paraphrase inspirée de la définition de Goddard, mais légèrement simplifiée : « je me sens comme quelqu'un qui voit quelque chose de très grand ».

Plusieurs sons peuvent être utilisés comme interjections primaires pour exprimer l'étonnement devant quelque chose de hors du commun. Avec ce type d'unités, nous sommes à la limite entre la langue et les autres formes d'expression orale. Dans le cadre de cette étude, nous portons notre attention sur les unités linguistiques, plutôt que les cris et productions sonores non-linguistiques, mais la distinction n'est, dans certains cas, pas facile à faire.

Les MI AYOYE, ⌜EH BOY⌝, HEILLE, MALADE, ⌜MON DIEU⌝, OUF, ⌜POUR VRAI⌝, SEIGNEUR, VOYONS, WÔ, WOW ont au moins une lexie qui inclut la composante **hors du commun**.

Dans l'extrait (75), l'énonciatrice I produit AYOYE en réaction à la situation hors du commun décrite par sa coénonciatrice É.

(75) É : mais là: LUI il est architecte il a jamais été capable de retrouver euh une job (.) il a cherché pendant UN an ils ont été un an là-bas
I : **ayoye**
*[CFPQ, sous-corpus 16, segment 6, page 51, lignes 9-10]*

### 1.1.7 Douleur

Nous avons vu plus haut que le bien-être ou le mal-être peuvent être décrits à l'aide des primitifs sémantiques. Goddard et Ye (2014) décrivent de manière similaire un cadre sémantique spécifiquement lié à la douleur.

quelqu'un RESSENT quelque chose de (bien/mal) dans une partie de son corps
(*Traduit de Goddard et Ye, 2014 : 7)*

Dans le même article, les auteurs décrivent de façon plus précise le nom *pain* en anglais :

She felt pain.
a. she felt something bad at that time
like someone can feel when it is like this:
b. something bad is happening to a part of this someone's body
c. this someone feels something bad in this part of the body because of this
d. this someone can't not think like this at this time: “I don't want this”
(Goddard et Ye, 2014, p. 10)

La deuxième ligne de la description permet de tenir compte des emplois métaphoriques du mot *pain*. Les MI qui expriment la douleur peuvent également être utilisés de manière métaphorique, suite à un événement désagréable sur le plan psychologique, par exemple.

Afin d'éviter la difficile question de la définition précise des différents actes expressifs liés à la douleur, nous préférons avoir recours au principe du bloc maximal ici. La paraphrase « je ressens de la douleur » servira à représenter l'acte expressif de tous les MI de cette catégorie.

Les vocables AYOYE et AÏE, qui incluent les signifiants *aïe*, *ouille* et *ouch* pour les raisons mentionnées au chapitre 1, sont utilisées par des énonciateurs pour exprimer la douleur. Dans le cadre de conversations où les interlocuteurs n'ont pas souvent l'occasion de se blesser, ces marqueurs sont surtout utilisés de manière fictive, dans les discours rapportés.

Dans l'extrait (76), l'énonciateur S feint une utilisation du MI AYOYE pour illustrer une situation où quelqu'un se blesse.

(76) *[S se lève et se rassoit pour mimer quelqu'un qui s'assoit sur un siège muni de clous.]*
S : [3•**ayoye** [4tabarnaque°
*[CFPQ, sous-corpus 21, segment 3, page 42, ligne 9]*

### 1.1.8 Dégoût

Les travaux de Wierzbicka (notamment 1986, 2011) ont fait état de la grande variété des unités lexicales liées au dégoût dans plusieurs langues. Il est possible de comprendre le sentiment de dégoût comme une extension du besoin primaire du corps de rejeter une substance néfaste.

Goddard (2013) distingue deux interjections liées au dégoût en anglais, *Ugh!* et *Yuck*! et les décrit avec beaucoup de nuances. Les sens des deux interjections s'articulent autour de l'idée que « quelque chose de mauvais est à l'intérieur de ma bouche ».

Les unités utilisées pour exprimer un dégoût physique peuvent typiquement être aussi employées pour exprimer un dégoût moral. Pour les besoins de notre étude, nous pouvons donner une explication très générale aux actes illocutoires qui expriment le dégoût. La paraphrase « je ressens du dégoût » caractérisera les MI de ce type.

Il est possible que, comme pour les actes illocutoires de la thématique du bien et du mal, il s'avérerait pertinent de distinguer l'expression du sentiment de dégoût de l'expression d'un jugement sur le caractère dégoûtant de quelque chose. Nous estimons manquer de données pour tirer une conclusion à ce sujet.

Les signifiants *ark*, *ouach*, *ouache*, *yark, eurk*, *yeurk*, *beurk* et *biark* que l'on regroupe, pour des raisons de simplicité, sous le vocable ARK servent à exprimer le dégoût. Ces signifiants, qui sont des interjections primaires, rappellent les sons produits par l'action de vomir et sont pour cette raison considérés comme des onomatopées.

Les unités de cette classe peuvent s'appliquer à un dégoût physique ou à un dégoût moral. Dans l'extrait (77), l'énonciateur J-M produit *beurk* en se rappelant le goût désagréable d'un aliment.

(77) J-M : pis ostie que [1ça goûtait TELLEMENT pas bon les foutues beans (.) **beurk** en tout cas [2c'étai:t
*[CFPQ, sous-corpus 10, segment 2, page 13, ligne 7]*

En (78), c'est sur le plan moral que l'énonciatrice MY semble dégoûtée lorsqu'elle produit *ouach*.

(78) MY : les deux sœurs [1s'embrassent/ [2**ouach**
*[CFPQ, sous-corpus 19, segment 1, page 2, ligne 6]*

### 1.1.9 Forte émotion

Certains MI sont associés à l'expression vive d'une forte émotion et ce phénomène semble faire partie de leur sens dénotatif (et non simplement connotatif). Les exemples les plus évidents de telles unités sont les interjections primaires plus ou moins involontaires liées à l'expression de la surprise, la frustration ou la frayeur (par exemple, *Ah!*). Ces unités ne font cependant pas partie de celles que nous étudions dans le cadre de cette thèse.

Les sacres, que nous allons analyser plus bas (1.5), sont également liés à cet acte expressif. La paraphrase que nous attribuons à cet acte illocutoire « je ressens une forte émotion » est inspirée des définitions de ces sacres par Dostie (2015).

À notre avis, les MI YOUPI et YÉ sont les seuls vocables à part les sacres qui méritent la composante « forte émotion » parmi ceux que nous étudions.

Dans l'extrait (79), l'énonciatrice ME exprime son vif bonheur de manquer trois cours de maths physique à l'aide de YÉ.

(79) C : mais [1comme euh hum histoire euh pas histoire mais euh Halloween Noël pis Saint-Valentin c'est tous le même jour fait que on manque toutes les fois maths physique pis ils (*en pointant Magalie*) manquent trois cours [2d'E.C.C.
ME : [1ah: c'est plate
ME : [2OH: **yé** [...]
*[CFPQ, sous-corpus 3, segment 5, page 83, ligne 8]*

Peut-être à cause de la composante **forte émotion**, les marqueurs YOUPI et YÉ sont souvent utilisés de manière ironique. En (80) par exemple, l'énonciatrice ME produit YOUPI pour exprimer son approbation de manière ironique suite à la situation qu'elle vient de décrire.

(80) ME : c’est triste (.) mais dis-toi que t’es pognée avec toute l’année prochaine
[...]
ME : ouais elle nous a annoncé ça cette année
D : (RIRE) [1•vous allez me revoir l’an prochain::°
C : [1(RIRE) (inaud.)
ME : **youpi**[1:
*[CFPQ, sous-corpus 3, segment 5, page 84, ligne 20-22]*

### 1.2 Les assertifs

Certains MI sont liés à des actes assertifs et mettent en jeu les notions de ⁽vrai⁾, d'⁽accord⁾ ou de ⁽désaccord⁾. Selon Searle (1979), le but des actes assertifs est de situer l'énonciateur par rapport à la réalité de quelque chose ou à la vérité d'une proposition. Contrairement aux actes expressifs, la vérité de la proposition (ou la réalité de la chose) à laquelle le marqueur est rattaché n'est pas présupposée dans le cas des actes assertifs. Ainsi, les états du monde liés aux marqueurs assertifs peuvent être sujets à des questionnements et les marqueurs assertifs peuvent eux-mêmes être utilisés en réponse à des questions totales.

Le tableau 16 rappelle les mots-clés et paraphrases associés aux actes illocutoires assertifs réalisés par les MI.

*Tableau 16 : Actes illocutoires assertifs des MI*

| **Mots-clés** | **Paraphrases** |
|---|---|
| Affirmatif | c'est vrai |
| Infirmatif | ce n'est pas vrai |
| Infirmatif partiel | ce n'est pas complètement vrai |

Les trois catégories présentées ici sont très larges. Elles ne tiennent par exemple pas compte des différentes niveaux de certitude et d'évidence. En pratique, plusieurs MI affirmatifs, tels que ceux

présentés plus bas, semblent interchangeables la plupart du temps. Une caractérisation sémantique plus précise de ceux-ci serait peut-être possible, mais demanderait une analyse beaucoup plus profonde.

Plusieurs unités, comme *oui* ou *non* et des phrasèmes comme *c'est clair* et *bien sûr* sont utilisés d'une manière qui ressemble fortement aux MI assertifs. Ces unités peuvent jouer un grand nombre de rôles discursifs qui dépassent le contexte de cette thèse et, pour les raisons expliquées au point 3 de ce chapitre, nous avons choisi de ne pas les traiter ici.

### 1.2.1 Affirmatif

Les MI affirmatifs répondent souvent à des questions ou confirment une affirmation du coénonciateur. Nous paraphrasons l'acte illocutoire affirmatif réalisé par ceux-ci de cette façon : « c'est vrai ».

Les vocables ÉCOUTE, ⌜JE COMPRENDS⌝, METS-EN, TELLEMENT, VRAIMENT, TIENS et REGARDE ont tous au moins une lexie qui a une composante affirmative.

Dans l'extrait (81), l'énonciatrice F produit VRAIMENT pour signaler à sa coénonciatrice L qu'elle est d'accord avec ce qu'elle dit.

(81) L : ici on: [1on trouve que c'est juste normal [2là t'sais moi [3ça me prend quinze minutes pis c'est juste normal même je trouve ça loin là (RIRE) <f<avant ça me prenait huit minutes là ça [4m'en prend quinze
F : [1<len<**vraiment**>> (*dit en séparant chacune des syllabes comme pour insister sur le fait qu'elle approuve ce que Lynda dit*)
*[CFPQ, sous-corpus 18, segment 4, page 46, lignes 10-11]*

La marque double « <*len*<>> » et le commentaire entre parenthèses à la suite du signifiant *vraiment* sont des indices du rôle crucial de la phonologie dans la désambiguïsation des MI.

Notons que, bien que les unités *ok* et *d'accord* soient très fréquentes dans le CFPQ (1510 occurrences pour *ok*), leur utilisation en tant que MI est pratiquement inexistante. Les deux vocables servent de marqueurs d'interaction, ils sont utilisés pour signaler l'écoute et pour baliser le discours.

### 1.2.2 Infirmatif

Les MI infirmatifs indiquent qu'une proposition n'est pas vraie ou qu'une situation n'est pas réelle. Comme les autres MI assertifs, ils peuvent servir à répondre à une question d'un coénonciateur. Nous paraphrasons l'acte illocutoire infirmatif de cette façon : « ce n'est pas vrai ».

Les vocables ⌜PAS DU TOUT⌝, ⌜DU TOUT⌝ et PANTOUTE ont une composante infirmative.

Dans l'extrait (82), l'énonciatrice J produit PANTOUTE afin de rejeter le doute émis par JN au sujet de son récit.

(82) J : t'sais à quelque part euh moi là tous mes amis présentement que: que je suis le plus proche au Saguenay c'était pas de mes amies quand je suis partie là (*en hochant la tête négativement*)
JN : non/
J : **pantoute** (*en hochant la tête négativement*)
*[CFPQ, sous-corpus 28, segment 8, page 106, ligne 14]*

### 1.2.3 Infirmatif partiel

Certaines unités servent à indiquer un désaccord partiel d'un énonciateur au sujet de quelque chose. Alors que les infirmatifs indiquent que quelque chose n'est pas vrai, ces unités indiquent que quelque chose n'est pas *complètement* vrai. Elles sont souvent utilisées comme euphémismes : elles sont une façon polie de contredire un coénonciateur. Comme pour les autres

MI qui réalisent des actes assertifs, ces unités répondent souvent à des questions ou à des affirmations de coénonciateurs.

Les MI BOF, ⌜C'EST ENCORE DRÔLE⌝ et ⌜PAS VRAIMENT⌝ ont au moins une lexie qui réalise un acte assertif infirmatif partiel.

En (83), l'énonciateur ME utilise BOF en réponse à une question totale. *Bof* semble alors être « presque » un *non.*

(83) M : [...] euh ben si on revient un peu au cinéma est-ce que vous écoutez des films québécois des fois/
D : [1pas vraim-
ME : [1**bo:f**
*[CFPQ, sous-corpus 3, segment 3, page 48, lignes 12-13]*

**1.3 Les directifs**

En utilisant des marqueurs qui réalisent des actes directifs, les énonciateurs cherchent à faire en sorte que leur coénonciateur agisse d'une certaine manière. Un acte directif est toujours adressé à un (ou plusieurs) coénonciateur, qui peut être fictif. Nous pouvons également observer dans le CFPQ des énonciateurs qui se parlent à eux-mêmes, à l'aide de MI directifs.

Nous distinguons quatre catégories larges d'actes directifs associés aux MI. Ceux-ci sont reproduits dans le tableau 17. Les actes identifiés par le mot-clé **attention** invitent une personne à utiliser ses capacités cognitives. Les actes identifiés par le mot-clé **arrêt** invitent une personne à cesser un comportement. Le mot-clé **question** identifie les actes qui invitent quelqu'un à donner plus d'information au sujet de quelque chose. Le mot-clé **encouragement** identifie les actes qui invitent quelqu'un à entreprendre une action d'un autre ordre que celles que l'on vient de mentionner.

*Tableau 17 : Actes illocutoires directifs des MI*

| **Mot-clé** | **Paraphrase** |
|---|---|
| Attention | pense à cela |
| Arrêt | ne fais pas cela |
| Question | est-ce que c'est vrai? |
| Encouragement | fais cela |

Le nombre de marqueurs directifs est assez restreint dans le CFPQ en raison de la nature des discussions qu'on y retrouve. D'autres types d'interaction entre les locuteurs susciteraient la production de marqueurs directifs bien différents.

Souvent, les marqueurs directifs réalisent également des actes expressifs ou assertifs. La dimension directive apparaît alors comme étant d'une faible importance d'un point de vue pragmatique. On peut, par exemple, poser une question dans le but d'exprimer son étonnement. Pour cette raison, et le fait mentionné plus haut que les marqueurs directifs s'adressent souvent à un public fictif, les actes directifs ne nous renseignent en général pas directement sur le réel contenu sémantique d'une conversation.

### 1.3.1 Attention

Certains MI invitent un coénonciateur à se pencher sur un problème, à utiliser ses capacités cognitives afin de comprendre ce dont il est question. Plusieurs de ces unités sont issues, par un processus de pragmaticalisation, de verbes de perception. C'est le cas des MI ÉCOUTE, REGARDE et VOYONS qui ont été analysées en détail dans l'ouvrage de Dostie (2004). Parmi les définitions proposées dans cet ouvrage pour les lexies illocutoires de ces vocables, on trouve la composante « je t'invite à user de tes capacités cognitives ». Dans le cas d'une lexie de REGARDE, la partie de la définition qui sert à d'écrire cet acte illocutoire directif se lit ainsi : « je t'invite à […] regarder2 [quelque chose qui vient d'être dit1 ou d'être fait] », ce qui représente à notre avis un cas plus étroit de l'acte qui pourrait être décrit de manière large par la paraphrase « pense à cela ».

Le mode impératif du verbe *penser* dans la paraphrase « pense à cela » rend compte du fait qu'il s'agit d'un acte directif dirigé vers un coénonciateur. *Pense* devrait être transformé en *pensez* en présence de coénonciateurs multiples ou en contexte de vouvoiement.

Dans l'extrait (84), l'énonciatrice C produit VOYONS afin d'indiquer qu'elle trouve insensé ou déplacé ce que dit son coénonciateur P et l'inviter à y penser davantage.

(84) C : moi je [1sais pas là mais (.) pas de char/ c'est correct\ mais pas de permis là [2pff (*dit avec découragement*)
P : [1hum
M : [2mais ça c'est tout le monde de Québec c'est de même
C : ben [1**voyons**
*[CFPQ, sous-corpus 25, segment 6, page 79, ligne 1]*

### 1.3.2 Arrêt

Nous croyons pertinent de distinguer un acte directif qui invite un coénonciateur à arrêter de faire quelque chose d'un autre qui encourage un coénonciateur à faire quelque chose. La simple paraphrase « ne fais pas cela » décrit sommairement cette catégorie d'actes directifs. Le mode impératif implique que l'acte est dirigé vers un coénonciateur.

Les MI ARRÊTE, CHUT et WÔ ont tous une composante directive par laquelle un énonciateur invite un coénonciateur à arrêter de faire quelque chose.

Dans l'extrait (85), l'énonciateur S est mis devant une information inattendue et produit ARRÊTE. Il réalise ainsi un acte directif de demande d'arrêt et exprime son étonnement.

(85) J : [1pis il va y avoir U2 qui va être là/
S : [2**arrête**/
*[CFQP, sous-corpus 15, segment 8, page 133, ligne 10-11]*

S semble ici demander à son amie J d'arrêter de parler, comme si le contenu émotif des paroles de J était trop fort pour être tolérable. Dans ce contexte, comme dans plusieurs autres, cet acte directif est davantage feint que sincère.

### 1.3.3 Question

Le mot-clé **question** identifie les actes qui invitent quelqu'un à donner plus d'information au sujet de quelque chose. Il est difficile de paraphraser l'ensemble des actes interrogatifs à l'aide d'une seule proposition. La paraphrase très générale « est-ce que c'est vrai? » semble appropriée dans la plupart des contextes.

Les MI HEIN, ⌜POUR VRAI⌝ et SÉRIEUX ont tous une composante interrogative dans au moins un de leurs signifiés.

Dans l'extrait (86), l'énonciateur F produit le MI SÉRIEUX suite à l'information inattendue offerte par son coénonciateur G. F demande à G de confirmer son affirmation et réalise ainsi un acte directif de type interrogatif. Suite à cette demande, G confirme la vérité de l'information avec « oui ».

(86) G : ben: à côté du cégep il y a une garderie qui s'appelle Manche de pelle
F : **SÉRIEUX↑**
G : oui
*[CFPQ, sous-corpus 9, segment 4, page 47, ligne 12]*

L'accentuation de SÉRIEUX, indiqué dans la transcription par des majuscules, est un indice de l'acte illocutoire expressif associé à cette lexie (inattendu). En outre, l'intonation fortement montante est un indice de l'acte interrogatif qui lui est associé.

### 1.3.4 Encouragement

La catégorie des actes directifs d' « encouragement » rassemble les marqueurs qui incitent un ou des coénonciateurs à faire quelque chose qui n'est pas couvert par les trois autres catégories que nous venons de présenter. La paraphrase « fais cela » permet de représenter de manière large cet acte directif.

ENVOYE et GO sont les deux seuls MI parmi ceux que nous analysons à réaliser le type d'acte illocutoire que nous désignons par le mot-clé **encouragment**.

En (87), l'énonciatrice VE produit ENVOYE afin d'inciter sa coénonciatrice à participer à un projet de musique :

(87) VI : [...] mon collègue Félix qui est rentré six mois avant moi il a embarqué dans les Funky Boys pis ils on:t ils ont essayé de m'embarquer là mais j'étais comme pas sûre [3{c'étai[t] full drôle;(inaud.)} (*dit en riant*)
[...]
VE : [3ah: **envoye** embarque [4ça va être drôle
*[CFPQ, sous-corpus 19, segment 5, page 47, ligne 15]*

## 1.4 Les connotations

Certaines unités sont porteuses de sens connotatifs qu'il convient de prendre en compte, c'est-à-dire qu'elles possèdent des caractéristiques sémantiques importantes qui n'appartiennent pas à leurs définitions (Mel'čuk *et al.*, 1995).

Nous n'avons relevé qu'une seule grande catégorie de signifiés connotatifs associés aux MI du corpus : le **tabou**. Kerbrat-Orecchioni (1977) classerait celle-ci parmi les connotations stylistiques, c'est-à-dire les connotations « dont la fonction consiste à signaler que le message procède d'un certain code ou sous-code linguistique particulier » (Kerbrat-Orecchioni, 1977, p. 94).

Rappelons que, par défaut, les unités linguistiques sont considérées comme neutres sur le plan connotatif et qu'une unité qui n'est pas neutre est perçue comme telle par les coénonciateurs. Une connotation implique une marque sémantique supplémentaire (Kerbrat-Orecchioni, 1977, p. 96) qui s'ajoute au sens d'une unité.

Il serait peut-être possible d'associer un autre type de connotation à certains marqueurs : leur caractère « impoli ». Nous pensons par exemple aux marqueurs *Chut* et *Wô*. La considération de ce type de trait connotatif demanderait cependant un travail d'investigation qui dépasse le cadre de cette thèse.

### 1.4.1 Tabou

Les mots tabous ont souvent des référents associés à la religion, à la sexualité et à la scatologie, mais l'interdit social qui leur est associé n'est pas entièrement déterminé par leur fonction référentielle. Plusieurs mots font en effet référence à ces réalités sans être perçus pour autant comme des mots grossiers.

Dans une présentation de certains gros-mots australiens, Wierzbicka (1997) définit ceux-ci en faisant appel au concept de « mauvais mots ». Sa définition de l'interjection *bugger*! décrit l'aspect psychologique de la production d'un gros-mot :

*bugger!* (interjection)
(a) I think: something bad happened
(b) I don't want to say "very bad"
(c) because of this, I think:
(d) I want to do something
(e) I can' do it
(f) because of this, I feel something bad
(g) because of this, I want to say something
(h) some people say that some words are bad words
(i) I want to say something of this kind
(Wierzbicka, 1997, p. 226)

Dans un article plus récent, Goddard (2015) définit certains « swear words » et « curse words » en anglais américain et australien. Il décrit la dimension sociale de leur utilisation, à l'aide de la paraphrase « some people can feel something bad when they hear this word ».

En accord avec ces deux auteurs, nous considérons qu'un mot qui possède la composante **tabou** est un « mauvais mot », c'est-à-dire que « certaines personnes peuvent ressentir quelque chose de mauvais quand elles entendent ce mot ».

### 1.5 Notes sur les sacres et leurs substituts

Les sacres sont des gros mots issus du vocabulaire sacré. Pour chaque sacre, il existe au moins un substitut de sacre. Les substituts de sacre sont des formes de remplacement : des mots qui ne sont pas des sacres, mais les rappellent par leur prononciation. Ces formes de remplacement peuvent être vues comme des euphémismes, elles ne sont pas ou peu stigmatisées, elles expriment des sentiments moins intenses (Dostie, 2015, p. 67) et ne sont ainsi pas synonymes des formes sacrilèges. La distinction entre ces deux classes de mots n'est pas tranchée au couteau et les unités appartiennent à des degrés plus ou moins élevés à l'une ou l'autre. Comme l'a démontré Dostie (2015), il existe des prototypes meilleurs exemplaires parmi les sacres (*câlisse*, *crisse*, *ostie* et *tabarnaque*), ainsi que des unités qui sont seulement parfois perçues comme étant des sacres (*calvaire*, *ciboire*, *sacrement*, *maudit*, *baptême*).

Le tableau 18 met en évidence les liens entre les signifiants des sacres et des substituts de sacres qui ont été relevés dans le CFPQ. Les signifiants de la colonne de droite rappellent ceux de la colonne de gauche.

*Tableau 18 : Les sacres et leurs substituts utilisés comme MI dans le CFPQ*

| **Formes stigmatisées** | **Fréquences** | **Formes non-stigmatisées** | **Fréquences** |
|---|---|---|---|
| câlisse | 15 | câlique | 18 |
| | | câline | 42 |
| | | câlif | 4 |
| ostie | 408 | ostique | 8 |
| | | ostifie | 9 |
| | | ostine | 2 |
| crisse | 126 | crif | 54 |
| | | crime | 23 |
| | | cristie | 4 |
| tabarnaque | 44 | tabarnache | 10 |
| | | tabarnouche | 24 |
| | | tabarnique | 3 |
| calvaire | 15 | calvince | 5 |
| ciboire | 14 | cibole | 25 |
| | | viarge | 3 |
| sacrement | 4 | | |
| | | sacre | 5 |
| | | sacrifice | 7 |
| | | simonaque | 9 |
| maudit | 16 | mautadit | 3 |
| baptême | 7 | bateau | 2 |
| | | batinse | 1 |
| | | torieu | 4 |

### 1.5.1 Unités d'origines des sacres et substituts MI

Une caractéristique bien connue des sacres est d'avoir des emplois qui correspondent à plusieurs catégories du discours, quelles soient lexicales, grammaticales ou discursives. Nous nous concentrons sans surprise sur les MI, c'est-à-dire sur les emplois isolés syntaxiquement qui réalisent par eux-mêmes des actes illocutoires expressifs. Ainsi, les constructions comme « crisse de N », « crisse que P », « en crisse » mettent en scène des emplois du signifiant de *crisse* que notre système d'analyse automatique devra identifier comme n'étant pas des MI.

Les unités issues de différentes catégories grammaticales se comportent de façons différentes lorsqu'elles ne sont pas des MI. Par exemple, le vocable MAUDIT peut être utilisé comme MI en (88), mais aussi comme adjectif en (89). Les sacres issus de noms (comme *crisse*) ne pourraient pas être substitués à *maudit* en (89), sans l'ajout de la préposition *de*.

(88) A : [1eh: **maudit** c’est pas le Québec qui ferait ça sont endettés [2de je sais pas de comment de milliards
*[CFPQ, sous-corpus 5, segment 8, page 83, ligne 15]*

(89) V : toujours [1des: CRÊpes dégueu là comme [2au McDo pis [3des pains dorés (.) [4tout le temps j’étais plus capable de: m:anger dans le **maudit** sirop d’érable à la fin là
*[CFPQ, sous-corpus 10, segment 6, page 72, ligne 1]*

### 1.5.2 Signifiés dénotatifs des sacres et substituts MI

Malgré certaines variations dans l'utilisation des sacres et des substituts, les signifiants de ceux-ci sont largement interchangeables en discours. Nous avons vu que les sacres étaient associés à une connotation de « mauvais mots », contrairement aux substituts de sacres. Les substituts de sacres réalisent également des actes expressifs plus faibles que les sacres.

Goddard (2015) qui analyse des *swear words* et *curse words* de l'anglais australien et de l'anglais américain, fait une distinction au sujet de versions faibles (comme *shit*) et fortes (comme *fuck*) de

gros mots. La distinction se réduit essentiellement au mot intensifieur « *very* », associé aux gros-mots forts, mais pas aux gros-mots faibles. Dostie (2015) fait la même distinction dans les définitions des lexies de CRISSE et CRIME. Alors que les deux lexies MI de CRISSE sont utilisées pour exprimer « avec une vive intensité une forte émotion », celles de CRIME sont utilisées pour exprimer « avec une faible intensité une légère émotion ». Le mot-clé **forte émotion,** introduit en 1.1.9, nous permettra de rendre compte de cette distinction sémantique.

Nous distinguons deux catégories d'emplois des sacres comme MI et deux catégories d'emplois des substituts de sacres comme MI. Les classes de lexies SACRE1, SACRE2, SUBSTITUT1 et SUBSTITUT2 sont une manière de regrouper les lexies similaires de tous les sacres et substituts. La description des vocables CRISSE et CRIME qui suit servira de modèle de description pour les autres sacres et substituts de sacres.

#### 1.5.2.1 SACRE1

Le premier sens des sacres en tant que MI exprime le sentiment que qqch. est **inattendu**. Dans l'extrait (90), l'énonciateur qui produit CRISSE exprime son étonnement devant un événement qui n'est pas particulièrement hors du commun.

(90) Y: ben: j'ai vu ça à la tv là (*en pointant son menton vers la droite comme pour désigner l'emplacement de la télévision dont il parle*)
[...]
Y: c'est le film que t'avais enregistré là
O: ah ouais **crisse** (*en haussant les sourcils comme pour exprimer sa surprise*)
*[CFPQ, sous-corpus 21, segment 2, page 17, lignes 4-14]*

Voici la définition proposée par Dostie (2015) pour cette lexie :

**3b.** *Crisse* ≅

Étant donné un état de choses α dont je prends tout à coup conscience en raison de ce qui vient d'être dit ou en raison d'un événement qui survient ||
j'exprime avec une vive intensité une forte émotion à l'endroit de α qui n'est pas (parfaitement) conforme à ce que j'aurais pu imaginer ou anticiper a priori.
(Dostie, 2015, p. 70)

La partie de la définition « qui n'est pas (parfaitement) conforme à ce que j'aurais pu imaginer ou anticiper a priori » rappelle la composante que nous identifions par le mot-clé **inattendu**. Nous reconnaissons aussi une composante au sujet de l'expression d'une **forte émotion** avec laquelle, comme nous l'avons vu, les sacres sont intrinsèquement liés.

Nous considérons ainsi que les composantes **inattendu** et **forte émotion** forment le signifié dénotatif des marqueurs de la classe SACRE1 et que la composante **tabou** fait partie de son signifié connotatif.

#### 1.5.2.2 SACRE2

Le deuxième sens des sacres en tant que MI exprime en quelque sorte la colère, le sentiment que quelque chose est mal et ne correspond pas au désir de l'énonciateur. En (91), il n'est pas question d'une situation inattendue, mais plutôt d'une situation qui ne correspond pas aux désirs de l'énonciateur.

(91) Y : {mais;ben} t'sais je peux pas chialer mais jamais que (.) si le Québec se sépare je câlisse (*en tournant la tête vers la droite comme pour représenter le mouvement de départ dont il s'apprête à parler*) mon camp ailleurs [1ça c'est clair eh **crisse** (*dit avec un petit rire*)
*[CFPQ, sous-corpus 21, segment 1, page 14, ligne 2]*

L'énonciateur de (91) ne peut possiblement pas être étonné par l'événement hypothétique qu'il décrit avant de produire CRISSE.

De manière similaire, l'énonciateur M ne réagit pas à une situation inattendue lorsqu'il produit CRISSE dans l'extrait (92), mais plutôt à sa propre histoire.

(92) M: mais euh: genre euh: à ce qu'il paraît il fait vraiment (.) plein de calls sexuels dans ses cours (.) AU secondaire [1genre
C: [1ben voyons
M: ben oui/ (.) mais t'sais OUI t'es en secondaire quatre oui ils sont plus vieux (e*n dessinant des guillemets avec ses doigts en disant* plus vieux)
MÈ: ben non [1mais tu dis pas ça
M: [1mais **crisse** ça ça: ça l'a pas sa place là
*[CFPQ, sous-corpus 25, segment 8, page 116, ligne 15]*

La définition que propose Dostie (2015) pour cette lexie est différente de celle vue plus haut pour *crisse***3b** sur quelques aspects.

**3c.** *Crisse* ≅
Étant donné un état de choses α très dérangeant dont je prends tout à coup conscience en raison de ce qui vient d'être dit ou en raison d'un événement qui survient ||
j'exprime avec une vive intensité une forte émotion négative à l'endroit de α qui m'atteint personnellement et qui perturbe, à des degrés variables, mon bien-être physique et/ou mental. (Dostie, 2015, p. 77-78)

Malgré la composante présupposée de cette définition où un énonciateur prend tout à coup conscience d'un état de choses, les marqueurs de la classe SACRE2 ne semblent pas particulièrement liés à l'expression de la surprise.

La composante « j'exprime avec une vive intensité une forte émotion » est la même que pour *crisse***3b** et peut également être représentée par le mot-clé **forte émotion**.

La partie de la définition sur l'expression d'une « émotion négative à l'endroit de » est représentée dans notre système de description par le mot-clé **mauvais**. La partie de la définition « qui

m'atteint personnellement et qui perturbe, à des degrés variables, mon bien-être physique et/ou mental. » est représenté par le mot-clé **se sentir mal**.

Comme pour les marqueurs de la classe SACRE1, ceux de la classe SACRE2 sont de plus caractérisés par la composante **tabou**.

#### 1.5.2.3 SUBSTITUT1

Les formes de remplacement peuvent être définies en parallèle aux sacres. La composante **forte émotion** n'est pas présente dans leur signifié, mais la distinction entre les emplois qui expriment l'étonnement et ceux qui expriment la colère persiste.

En (93), l'énonciateur C utilise CRIME pour exprimer son étonnement :

(93) ME : mais euh ils se battaient [1à coups de poing là/ vraiment fort/
C : [1**crime** {il a/;ils ont fait} ça sérieux/ euh
*[CFPQ, sous-corpus 3, segment 6, page 101, lignes 10-11]*

La définition de Dostie pour la lexie *crime***2b** correspond à la classe de lexies que nous nommons « Substitut1 » :

**2b.** *Crime* ≅
Étant donné un état de choses α dont je prends tout à coup conscience en raison de ce qui vient d'être dit ou en raison d'un événement qui survient ||
j'exprime avec une faible intensité une légère émotion à l'endroit de α qui n'est pas (parfaitement) conforme à ce que j'aurais pu imaginer ou anticiper a priori.
(Dostie, 2015, p. 81)

La seule différence entre cette définition et celle de *crisse***3b** se trouve dans les mots « avec une faible intensité une légère émotion » qui distinguent le signifié dénotatif des substituts des sacres. La classe SUBSTITUT1 sera donc dénuée de la composante **forte émotion**. Elle n'est également pas connotée par l'interdit qui caractérise les sacres, représenté par le mot-clé **tabou**.

#### 1.5.2.4 SUBSTITUT2

En (94), l'énonciateur M utilise CRIME pour exprimer son sentiment négatif au sujet d'une situation déplaisante qui concerne un locuteur qui ne comprend pas une langue étrangère.

(94) M : tu comprends pas un sacrement de mot [1de ce qu'il dit là <all<la seule affaire que tu comprends>> c'est c'est •vino° pis euh (RIRE)
[...]
M : [1<f<tandis que quand t'es dans la la bonne [2comme là en FRANCE c'est pas pire t'es dans la bonne langue>> <dim<là je veux dire [3ben là **crime**>>
*[CFPQ, sous-corpus 6, segment 7, page 94, ligne 17 - page 95, ligne 4]*

La définition proposée dans Dostie (2015) pour cet emploi précise que l'émotion négative exprimée est légère :

**2c.** *Crime* ≅
Étant donné un état de choses α dont je prends tout à coup conscience en raison de ce qui vient d'être dit ou en raison d'un événement qui survient ||
j'exprime avec une faible intensité une émotion plutôt négative à l'endroit de α qui n'est pas tel que je l'aurais voulu.
(Dostie, 2015, p. 82)

La dernière partie de cette définition peut être représentée par les composantes **mauvais** et **se sentir mal**.

### 1.6. Ordonnancement des sens

Il semble que l'ordre dans lequel les paraphrases doivent se combiner pour décrire adéquatement les scénarios cognitifs mis en jeu dans la production des MI soit assez régulier. Les paraphrases qui décrivent des actes expressifs doivent précéder celles qui décrivent des actes directifs qui doivent elles-mêmes précéder celles qui décrivent des actes assertifs.

Dans le tableau 19, nous suggérons un ordre de priorité des paraphrases explicatives liées aux MI, de la plus haute à la plus basse. Les paraphrases en haut de la liste semblent naturellement se placer avant celles introduites plus bas lorsqu'on les utilisent pour décrire les signifiées des unités.

*Tableau 19 : Priorité d'application des paraphrases explicatives*

| **Types d'actes illocutoires** | **Mots-clés** | **Paraphrases** |
|---|---|---|
| Expressifs | Forte émotion | je ressens une forte émotion |
| | Inattendu | je ne savais pas que cela allait être comme cela |
| | Hors du commun | je me sens comme quelqu'un qui voit quelque chose de très grand |
| | Bien | c'est bien |
| | Mauvais | c'est mauvais |
| | Se sentir bien | je ressens quelque chose de bien |
| | Se sentir mal | je ressens quelque chose de mauvais |
| | Douleur | je ressens de la douleur |
| | Dégoût | je ressens du dégoût |
| Directifs | Attention | pense à cela |
| | Arrêt | ne fais pas cela |
| | Question | est-ce que c'est vrai? |
| | Encouragement | fais cela |
| Assertifs | Affirmatif | c'est vrai |
| | Infirmatif | ce n'est pas vrai |
| | Infirmatif partiel | ce n'est pas complètement vrai |
| **Types de connotations** | **Mots-clés** | **Paraphrases** |
| Stylistiques | Tabou | certaines personnes peuvent ressentir quelque chose de mauvais quand elles entendent ce mot |

L'élocution vive et le haut volume de la voix qui accompagnent typiquement l'acte illocutoire lié à l'expression d'une **forte émotion** sont des indicateurs immédiatement sensibles de celui-ci. La paraphrase « je ressens une forte émotion » est tout en haut du tableau 19 parce que nous considérons qu'elle précède séquentiellement toutes les autres paraphrases lorsqu'elles s'assemblent pour former la description d'un signifié.

La paraphrase liée au mot-clé **inattendu**, qui met elle-même en jeu un enchaînement d'événements, semble également devoir se placer avant les autres. Les deux sens de ⌜POUR VRAI⌝, décrits au point 2.56, sont des exemples où la composante **inattendu** précède une composante directive (⌜POUR VRAI⌝1 ≅ **inattendu**, **question**) ou une autre composante expressive (⌜POUR VRAI⌝2 ≅ **inattendu**, **mauvais**).

Comme la composante **inattendu**, la composante **hors du commun** décrit un scénario où un énonciateur réagit à un élément externe (⁽cela⁾ ou ⁽quelque chose⁾). Ces deux composantes servent de justification à plusieurs autres actes expressifs, directifs ou assertifs.

Les actes liés à l'expression d'un état du locuteur sont généralement produits après ceux qui concernent l'expression d'un jugement du locuteur au sujet d'une chose qui lui est extérieure. Ainsi, les mots-clés **bien** et **mauvais** précèdent **se sentir bien** et **se sentir mal**.

Le sens de ÉCOUTE3 (voir 2.27) est un exemple où une composante directive précède une composante assertive (ÉCOUTE3 ≅ **attention**, **affirmatif**).

Le sens de COUDON4 (voir 2.20), pour sa part, met en jeu une composante expressive suivie d'une composante assertive (COUDON4 ≅ **inattendu**, **affirmatif**).

## 2 Description des unités

Dans cette partie, nous offrons une description sommaire des MI de la liste établie au chapitre 1 (point 6.1). Pour chacun des vocables, nous précisons le nombre de signifiants qui lui sont associés dans le CFPQ ainsi que le nombre de MI parmi ces signifiants.

Pour chacune des sens de ces vocables, nous donnons un certain nombre de mot-clés qui caractérisent son signifié. Les actes illocutoires liés à ces mots-clés sont décrits par les paraphrases présentées au tableau 18. Le sens de chacune des lexies peut ainsi être représenté par une série de paraphrases indépendantes.

Nous fournissons également les quasi-synonymes de chacune des lexies en vertu du système de description sémantique que nous utilisons. Une unité est habituellement quasi-synonyme à une autre si elle est décrite par les mêmes mots-clés.

Nous faisons également des commentaires au sujet d'exemples d'utilisations particulières de certaines lexies.

Le tableau 20 présente la liste des lexies des MI, distribuées selon les actes illocutoires et connotations auxquels elles sont associées. Notons que, pour plus de lisibilité, tous les sacres sont regroupés sous les classes de lexies SACRE1 et SACRE2, tandis que leurs substituts sont regroupés sous les classes SUBSTITUT1 et SUBSTITUT2.

*Tableau 20 : Lexies des MI et les actes illocutoires qu'elles réalisent*

| **Types d'actes illocutoires** | **Mots-clés** | **Lexies dont le signifié inclut cette composante** |
|---|---|---|
| Expressifs | Bien | WOW, FIOU1, OUF1, ⌜UNE CHANCE⌝, COOL, SUPER, TIENS3, YOUPI |
| | Mauvais | ⌜DE LA MARDE⌝, OUPELAILLE1, OUPS, COUDON5, SUBSTITUT2, SACRE2, MERDE, ZUT, ÉCOUTE2, VOYONS2, COUDON2, ⌜AÏE AÏE AÏE⌝, FRANCHEMENT2, ⌜POUR VRAI⌝2, SEIGNEUR2 |
| | Se sentir bien | WOW, FIOU1, OUF1, ⌜UNE CHANCE⌝, COOL, SUPER, TIENS3, YOUPI, YÉ |
| | Se sentir mal | OUPELAILLE1, OUPS, SUBSTITUT2, SACRE2, COUDON5, MERDE, ZUT, VOYONS4 |
| | Inattendu | OUPELAILLE1, OUPS, COUDON2, WOW, FIOU1, OUF1, ⌜UNE CHANCE⌝, ARRÊTE2, COUDON4, AYOYE2, MALADE, ⌜MON DIEU⌝, ⌜MON DOUX', ⌜MY GOD⌝, OUPELAILLE2, VOYONS1, ⌜POUR VRAI⌝1, SÉRIEUX2, SUBSTITUT1, SACRE1, COUDON3, HEIN2, ⌜REGARDE DONC⌝, TIENS4 |
| | Hors du commun | WOW, AYOYE2, MALADE, ⌜MON DIEU⌝, ⌜MON DOUX⌝, ⌜MY GOD⌝, OUPELAILLE2, VOYONS1, ⌜AÏE AÏE AÏE⌝, FRANCHEMENT2, ⌜POUR VRAI⌝2, SEIGNEUR2, WÔ, HEILLE, ⌜EH BOY⌝, FIOU2, OUF2, SEIGNEUR1 |
| | Douleur | AÏE, AYOYE1 |
| | Dégoût | ARK |
| | Forte émotion | SACRE1, SACRE2, YOUPI, YÉ |
| Assertifs | Affirmatif | COUDON4, ÉCOUTE3, REGARDE, FRANCHEMENT1, ⌜JE COMPRENDS⌝, METS-EN, ⌜POUR VRAI⌝3, SÉRIEUX1, TELLEMENT, TIENS5, VOYONS3, |

| | | |
|---|---|---|
| | | VRAIMENT |
| | Infirmatif | ⌜DE LA MARDE⌝, ⌜DU TOUT⌝, PANTOUTE, ⌜PAS DU TOUT⌝, ⌜VRAIMENT PAS⌝ |
| | Infirmatif partiel | BOF, ⌜C'EST ENCORE DRÔLE⌝, ⌜PAS VRAIMENT⌝ |
| Directifs | Attention | ÉCOUTE3, REGARDE, HEILLE, ÉCOUTE2, VOYONS2, COUDON1, ÉCOUTE1, TIENS1, TIENS2 |
| | Arrêt | WÔ, ARRÊTE2, ARRÊTE1, CHUT |
| | Question | ⌜POUR VRAI⌝1, SÉRIEUX2, HEIN1 |
| | Encouragement | ENVOYE, GO, ⌜LET'S GO⌝ |
| **Types de connotations** | **Mots-clés** | **Unités** |
| Stylistiques | Tabou | SACRE1, SACRE2, MERDE |

**2.1 AÏE**

Signifiants : 7 (5 *aïe*, 1 *ouille*, 1 *ouch*)
MI : 7
Mot-clé : **douleur**
Quasi-synonymes : AYOYE1

Nous avons choisi de regrouper les différentes interjections primaires (*aïe*, *ouch* et *ouille*) qui expriment la douleur dans une seule classe de marqueur que nous nommons AÏE.

En (95), l'énonciatrice K dit *aïe* après avoir été touchée par sa coénonciatrice, comme si le contact physique lui avait fait mal.

(95) *K agite la main vers la caméra en signe de salut et touche à C par accident.*
K : **aïe** (*en regardant C et en se touchant le coude*)
C : ah excuse (*en se retournant vers K*)
*[CFPQ, sous-corpus 17, segment 10, page 144, ligne 14]*

Dans un contexte de discussion sécuritaire comme celui du CFPQ, il n'est pas étonnant que plusieurs exemples de marqueurs qui expriment la douleur se trouvent à l'intérieur de discours rapportés. En (96), par exemple, l'énonciateur MA cite une autre personne qui utilise *aïe* pour exprimer sa douleur :

(96) MA : 1elle était là •**aïe** ça fait mal° c'est parce qu'elle avait trop de barrettes sur la tête fait que ça-
*[CFPQ, sous-corpus 3, segment 9, page 128, ligne 13]*

Bien que la forme *aïe* semble être une contraction de la forme *ayoye*, nous avons choisi de distinguer en deux classes AYOYE (qui peut exprimer l'étonnement) et AÏE (qui n'exprime que la douleur).

Pour rajouter à la confusion, le marqueur *heille* est parfois prononcé d'une manière qui fait penser à *aïe*. De plus, il existe un autre marqueur formé par la répétition du morphème /aj/ que nous pouvons appeler ⌜AÏE AÏE AÏE⌝.

### 2.2 ⌜AÏE AÏE AÏE⌝

Signifiants : 4 (*aïe aïe aïe*)
MI : 4
Mot-clés : **hors du commun**, **mauvais**
Quasi-synonymes : FRANCHEMENT2, ⌜POUR VRAI⌝2, SEIGNEUR2

Le MI ⌜AÏE AÏE AÏE⌝ semble le plus souvent employé pour exprimer un sentiment légèrement négatif au sujet d'une situation problématique qui sort de l'ordinaire. Dans l'exemple (97), l'énonciatrice V utilise ⌜AÏE AÏE AÏE⌝ afin de signifier sa désapprobation au sujet de la situation décrite par MA.

(97) MA : ils ont plus le droit d'amener les é- les enfants du primaire/ (.) glisser/ (.) parce {que euh/;que:/} (.) c'est pas COUvert par l'assurance de la commission scolaire les sports de glisse
V : <all<**aïe aïe aïe**>> (*dit avec découragement*)
*[CFPQ, sous-corpus 30, segment 2, page 14, lignes 1-2]*

### 2.3 ARK

Signifiants : 63 (40 *ark*, 4 *ouach*, 3 ouache, 5 *yark, 4 eurk*, 3 *yeurk*, 3 *beurk*, 1 *biark*)
MI : 63
Mot-clé : **dégoût**

ARK est un regroupement de signifiants utilisés pour exprimer le dégoût. Dans l'extrait (98), l'énonciatrice MC produit le signifiant *ouache* après avoir observé une saleté sur son verre.

(98) MC : [1Marina: (*dit sur un ton moqueur*) (.) (*elle regarde sa coupe*) [2<p<**ouache**>> (*elle essuie le pied de sa coupe*)
*[CFPQ, sous-corpus 22, segment 6, page 94, ligne 21]*

Le fait que *ouache* soit entouré de la marque « piano », qui indique un faible volume de parole, laisse croire que l'énonciatrice se parle à elle-même dans ce contexte.

### 2.4 ARRÊTE

Signifiants : 132 (125 *arrête*, 7 *arrêtez*)
MI : 34 (30 *arrête*, 4 *arrêtez*)

Le MI ARRÊTE est similaire aux marqueurs directifs ÉCOUTE et REGARDE par son origine verbale et les liens qu'il entretient avec son unité source. Il a le plus souvent une fonction expressive, mais peut aussi être directif.

**ARRÊTE1**
Mot-clé : **arrêt**

Le sens le plus évident de ARRÊTE est directif. À l'aide de ARRÊTE1, un énonciateur demande simplement au coénonciateur d'arrêter de faire quelque chose. Cet emploi de *arrête* est assimilable à l'utilisation du verbe *arrêter* à l'impératif. Utilisé seul, le verbe concerne un comportement du coénonciateur, le plus souvent le fait qu'il parle.

En (99), l'énonciatrice É demande à sa coénonciatrice d'arrêter son explication :

(99) R : [2ben t'enlè- t'enlèves ton crochet si c'est la même chose qu'on parle t'enlèves ton crochet tu cloues ton clou pis tu remets ton crochet (.) [3si c'est le même crochet
[...]
É : [1**arrête** heille que ça s'en vient compliqué hein ah:::
[CFPQ, *sous-corpus 16, segment 9, page 87, ligne 20*]

**ARRÊTE2**

Mots-clés : **inattendu**, **arrêt**

ARRÊTE2 est beaucoup plus fréquent que ARRÊTE1 dans le CFPQ. Le MI ARRÊTE2 simule un acte directif de manière à appuyer un acte expressif. Un énonciateur qui utilise ARRÊTE2 fait mine de demander à son coénonciateur d'arrêter de dire ce qu'il dit, comme si ses paroles étaient trop difficiles à croire. Il exprime ainsi son étonnement.

L'ordre de priorité des paraphrases (présenté au tableau 19) fait en sorte que le scénario cognitif qui mène à la production de ARRÊTE2 est adéquatement décrit en commençant par l'élément expressif, suivi de l'élément directif.

A : X
B : je ne savais pas que X allait être comme cela
ne fais pas cela :
dire X

L'inversion de ces paraphrases ferait en sorte que l'acte directif perdrait sa justification. C'est parce qu'un énonciateur est étonné qu'il demande à son coénonciateur d'arrêter de parler, et non l'inverse.

Le marqueur peut être accompagné d'une phrase explicative, comme en (100), où la phrase « je fais trente et un mille brut » met en contexte ARRÊTE2.

(100) VE : [2parce qu'on doit avoir un salaire quand même équivalent là toi [3pis moi/

[...]

VI : je le sais pas (.) vingt-six mille

[...]

VE : BRUT//

VI : ouais\

[...]

VE : ben voyons donc **arrête** là (.) je fais trente et un mille brut

*[CFPQ, sous-corpus 19, segment 1, page 9, ligne 15]*

Comme pour les marqueurs ÉCOUTE1, ÉCOUTE2 et REGARDE, nous considérons que ARRÊTE2 est très peu directif ou que l'acte directif qu'il réalise est simulé. La composante directive de la définition pourrait être approximativement paraphrasée par « je fais comme si je te demandais d'arrêter de dire quelque chose ».

### 2.5 AYOYE

Signifiants : 66 (*ayoye*)

MI : 66

**AYOYE1**

Mot-clé : **douleur**

Quasi-synonyme : AÏE

Dans son premier sens, AYOYE est utilisé pour exprimer la douleur et semble être équivalent aux signifiants regroupés dans la classe AÏE (*aïe*, *ouille*, *ouch*). Le signifiant *ayoye* n'est pas inclus dans la classe AÏE parce qu'il peut prendre un sens supplémentaire en plus de celui qui est lié à la douleur.

**AYOYE2**

Mots-clés : **inattendu**, **hors du commun**

Quasi-synonymes : MALADE, ⌜MON DIEU⌝, ⌜MON DOUX⌝, ⌜MY GOD⌝, OUPELAILLE2, VOYONS1

AYOYE est le plus souvent utilisé pour exprimer un sentiment en réaction à quelque chose qui est inattendu et hors du commun. En (101), l'énonciateur I réagit à l'affirmation surprenante de É en utilisant AYOYE2 :

(101) É : mais là: LUI il est architecte il a jamais été capable de retrouver euh une job (.) il a cherché pendant UN an ils ont été un an là-bas
I : **ayoye**
*[CFPQ, sous-corpus 16, segment 6, page 51, lignes 10-11]*

Comme c'est le cas avec tous les MI, le sentiment exprimé par AYOYE2 peut être feint ou ironique. En (102), l'énonciatrice semble utiliser AYOYE2 à retardement, comme pour indiquer qu'elle a été étonnée au moment où elle a pris connaissance du fait en question. Cette utilisation d'un MI est similaire à l'énonciation d'un discours rapporté.

(102) F : ah oui 1tes tes nièces elles ont voyagé là/ 2**ayoye** 3heille il y a une de ses nièces qui est allée en INDE
*[CFPQ, sous-corpus 6, segment 5, page 67, ligne 9]*

**2.5 BAPTÊME**

Signifiants : 14 (*baptême*)

MI : 7

BAPTÊME est un sacre non-prototypique, parfois perçu comme état stigmatisé par les locuteurs québécois. Nous croyons qu'il peut être utilisé à la manière d'un sacre ou à la manière d'un

substitut de sacre. Dans ce dernier cas, les composantes **forte émotion** et **tabou** sont absentes de son signifié.

**BAPTÊME1**

Mot-clé: **inattendu** (**forte émotion**, **tabou**)

Quasi-synonymes : Les membres des classes SACRE1 et SUBSTITUT1

**BAPTÊME2**

Mots-clés: **mauvais**, **se sentir mal (forte émotion**, **tabou)**

Quasi-synonymes : Les membres des classes SACRE2 et SUBSTITUT2

### 2.6 BATEAU

Signifiants : 80 (*bateau*)

MI : 2

BATEAU semble être utilisé comme substitut du sacre non-prototypique BAPTÊME, malgré le caractère peu stigmatisé de ce dernier.

**BATEAU1**

Mot-clé: **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**BATEAU2**

Mots-clés: **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

### 2.7 BATINSE

Signifiants : 1 (*batinse*)

MI : 1

BATINSE paraît être utilisé comme substitut du sacre non-prototypique BAPTÊME.

**BATINSE1**

Mot-clé: **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**BATINSE2**

Mots-clés: **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

**2.8 BOF**

Signifiants : 13 (*bof*)

MI : 12

Mot-clé : **infirmatif partiel**

Quasi-synonymes : ⌜C'EST ENCORE DRÔLE⌝, ⌜PAS VRAIMENT⌝

Le signifié de BOF est difficile à paraphraser, mais il apparaît que l'unité est utilisée par les énonciateurs afin d'indiquer leur désaccord, leur réticence ou leur scepticisme au sujet de quelque chose. En (103), par exemple, l'énonciateur J produit BOF et indique ainsi qu'il n'est pas tout à fait en accord avec ce que dit N.

(103) N : ah non ça a été ça euh: c'est plate dans durant nos vacances là une journée euh: sur le dos
J : **bof** (1,4") c'est une journée qui passe [1(*il sourit à Nadia*) qui passe vite
*[CFPQ, sous-corpus 6, segment 1, page 3, ligne 20]*

Le seul emploi du signifiant *bof* qui n'est pas un MI dans le CFPQ est reproduit dans l'exemple (104). On y voit que *bof* est utilisé comme adverbe de manière et qu'il pourrait être substitué par *mal*.

(104) J : m- mais j:- il m'a juste dit {que: il;que euh il;que i:l} allait **bof** [...]
*[CFPQ, sous-corpus 17, segment 8, page 104, ligne 8]*

### 2.9 ⌜C'EST ENCORE DRÔLE⌝

Signifiants : 4 (*c'est encore drôle*)
MI : 4
Mot-clé : **infirmatif partiel**
Quasi-synonymes : BOF, ⌜PAS VRAIMENT⌝

En (105), l'énonciateur ME utilise ⌜C'EST ENCORE DRÔLE⌝ pour indiquer que la proposition *les filles sont plus critiques* n'est pas complètement vraie.

(105) MA : [1moi ça dépend devant les gens comme cette année je vais peut-être être plus gênée là parce que t'sais c'est plus (.) des filles (.) pis je trouve les filles sont plus critiques fait que
ME : non **c'est encore drôle** (.) [1Gingras pis Francis là [2ensemble là
*[CFPQ, sous-corpus 3, segment 4, page 60, ligne 17]*

### 2.10 CÂLIF

Signifiants : 5 (*câlif*)
MI : 4

CÂLIF est vraisemblablement utilisé comme substitut du sacre CÂLISSE.

**CÂLIF1**

Mot-clé : **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**CÂLIF2**

Mots-clés : **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

## 2.11 CÂLINE

Signifiants : 42 (*câline*)

MI : 42

CÂLINE semble être utilisé comme substitut du sacre CÂLISSE. La construction « câline de bine » se trouve une fois dans le CFPQ.

**CÂLINE1**

Mot-clé : **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**CÂLINE2**

Mots-clés : **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

## 2.12 CÂLIQUE

Signifiants : 20 (*câlique*)

MI : 18

CÂLIQUE semble être utilisé comme substitut du sacre CÂLISSE.

**CÂLIQUE1**

Mot-clé : **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**CÂLIQUE2**

Mots-clés : **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

## 2.13 CÂLISSE

Signifiants : 22 (*câlisse*)

MI : 15

CÂLISSE est un sacre prototypique.

**CÂLISSE1**

Mots-clés : **forte émotion, inattendu, tabou**

Quasi-synonymes : Les membres des classes SACRE1

**CÂLISSE2**

Mots-clés : **forte émotion**, **mauvais**, **se sentir mal**, **tabou**

Quasi-synonymes : Les membres des classes SACRE2

## 2.14 CALVAIRE

Signifiants : 16 (*calvaire*)

MI : 16

CALVAIRE est généralement tenu pour un sacre (Dostie, 2015, p. 60)

**CALVAIRE1**

Mots-clés : **forte émotion, inattendu, tabou**

Quasi-synonymes : Les membres des classes SACRE1

**CALVAIRE2**

Mots-clés : **forte émotion**, **mauvais**, **se sentir mal**, **tabou**

Quasi-synonymes : Les membres des classes SACRE2

### 2.15 CALVINCE

Signifiants : 5 (*calvince*)

MI : 5

CALVINCE est vraisemblablement utilisé comme substitut du sacre CALVAIRE.

**CALVINCE1**

Mot-clé : **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**CALVINCE2**

Mots-clés : **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

### 2.16 CHUT

Signifiants : 16 (*chut*)

MI : 16

Mot-clé : **arrêt**

Un énonciateur qui utilise CHUT réalise un acte directif par lequel il demande à quelqu'un d'arrêter de parler d'un sujet ou d'arrêter de faire du bruit. Il s'agit d'un cas particulier des actes directifs qui sont regroupés par le mot-clé **arrêt**.

En (106), l'énonciatrice produit CHUT afin de demander à M de ne pas aborder la question de la présence d'un chat dans la résidence :

(106) P: [1non il faut pas le dire Héloïse [2serait pas contente
MÈ: [2non j'ai pas le droit
M: [2pas le droit de chat il y a il y a-tu un chat↑
P: [1non
MÈ: [1**chu:t** (*en plaçant son doigt devant sa bouche*)
*[CFQP, sous-corpus 25, segment 3, page 33, ligne 15]*

**2.17 CIBOIRE**

Signifiants : 14 (*ciboire*)
MI : 14

CIBOIRE est généralement tenu pour un sacre (Dostie, 2015, p. 60).

**CIBOIRE1**

Mots-clés : **forte émotion**, **inattendu**, **tabou**
Quasi-synonymes : Les membres des classes SACRE1

**CIBOIRE2**

Mots-clés : **forte émotion**, **mauvais**, **se sentir mal**, **tabou**
Quasi-synonymes : Les membres des classes SACRE2

### 2.18 CIBOLE

Signifiants : 26 (*cibole*)

MI : 25

CIBOLE semble être utilisé comme substitut du sacre CIBOIRE.

**CIBOLE1**

Mot-clé : **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**CIBOLE2**

Mots-clés : **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

### 2.19 COOL

Signifiants : 67 (*cool*)

MI : 8

Mots-clés : **bien**, **se sentir bien**

Quasi-synonyme : SUPER

Le MI COOL est issu de l'adjectif *cool*, il peut être vu comme la contraction de la phrase « *C'est cool.* ». En utilisant COOL comme MI, un énonciateur indique que quelque chose est positif et qu'il ressent un sentiment positif. En (107), l'énonciateur A exprime ainsi son approbation à l'aide de COOL :

(107) H : ah c'était pas pire (.) j'ai euh (.) combattu ma peur pis j'ai été danser avec la mariée
A : [1**COOL** (*en se reculant pour le regarder bien dans les yeux*)
*[CFPQ, sous-corpus 14, segment 8, page 83, ligne 25-26]*

COOL n'est pas forcément utilisé en réaction à une affirmation d'un coénonciateur. Par exemple, on peut imaginer quelqu'un dire « *cool* » après avoir reçu un cadeau qu'il apprécie.

**2.20 COUDON**

Signifiants : 25 (*coudon*)
MI : 25

Dostie (2004) a décrit en détail le vocable COUDON, dont ses nombreuses utilisations en tant que MI et marqueur d'appel à l'écoute. Nous ne jugeons pas nécessaire ici de reproduire les définitions et les nombreux commentaires de cet ouvrage au sujet de ce vocable. Notons que, étant donné la caractérisation sommaire des unités que nous effectuons, il nous est possible de rassembler les lexies *coudon***2** et *coudon***4** de Dostie (2004) en une seule lexie que nous nommons COUDON2.

**COUDON1**

Mot-clé : **attention**
Quasi-synonymes : ÉCOUTE1, TIENS1, TIENS2

COUDON1 sert à solliciter l'attention d'un coénonciateur, le plus souvent afin d'introduire un nouveau sujet de conversation.

Dans l'extrait (108), l'énonciateur produit COUDON au cours d'une partie de cartes afin d'introduire une question hors contexte au sujet d'une main antérieure.

(108) G : **coudon** c'était qu'est-ce que j'avais tantôt/ (*en fronçant les sourcils comme en signe d'incertitude*)
*[CFPQ, sous-corpus 27, segment 4, page 59, ligne 18]*

**COUDON2**

Mots-clés : **inattendu**, **mauvais**

Avec COUDON2, un énonciateur indique qu'il considère comme négative une situation inattendue. À l'écoute de la bande audiovisuelle qui a mené à la transcription de l'exemple (109), on observe que le *coudon* y est prononcé avec une intonation montante qui marque clairement la surprise et la désapprobation. L'énonciatrice F utilise ainsi COUDON2 en prenant conscience d'une situation dangereuse, à savoir qu'un sentier donne sur une carrière en fonction.

(109) J: [...] on avait fait le tour là des autres sentiers pis ça menait nulle part ben il fallait passer à travers la machi[9nerie
F : [9l'espèce de carrière [10là c'est une carrière heille c'est
J : [10 ouais il y a pas c'est pas c'est pas réglementé [11t'sais
F : [11pis c'est pas annoncé non [12plus **coudon**
*[CFPQ, sous-corpus 18, segment 7, page 68, ligne 13]*

**COUDON3**

Mot-clé : **inattendu**

COUDON3 est également produit en réaction à un état de chose inattendu, mais n'exprime pas de désapprobation.

Dans l'extrait (110), l'énonciatrice T produit COUDON3 suite à la mention du mot *babiche* de la part de sa coénonciatrice.

(110) S : [...] AH c'est de la baBICHE qu'on dit
[...]
T : [1t'as-tu↑ du sang indien **coudon** toi de la BAbich:e
*[CFPQ, sous-corpus 23, segment 4, page 71, ligne 9]*

**COUDON4**

Mots-clés : **inattendu**, **affirmatif**

COUDON4 a une composante assertive par laquelle un énonciateur indique qu'il considère que quelque chose est vrai et va de soi, malgré son apparence inattendue.

Dans l'extrait (111), l'énonciateur SA utilise COUDON4 afin d'affirmer le caractère inévitable du futur rôle d'artiste d'une enfant tout en le caractérisant comme inattendu.

(111) SA : tu vois à matin je me lève à matin ou hier matin (.) elle était assis elle s'était levée le matin là pis elle écrivait ses ses personnages elle les décrivait comment c'était quels types de personnages pis tout ça\ [..]
I : [2e::lle prend ça au sérieux ah ouais (.) ben (.) **coudon** [3on a une future euh
[...]
I : on a une future euh [1artiste euh (inaud.) pis ouais c'est vrai
*[CFPQ, sous-corpus 7, segment 10, page 105, ligne 10]*

**COUDON5**

Mots-clés : **mauvais**, **se sentir mal**

Le cinquième sens de COUDON a une composante expressive par laquelle un énonciateur indique qu'il est insatisfait d'un élément négatif d'une situation. En (112), par exemple, le COUDON5 produit par l'énonciatrice H laisse entendre qu'elle n'est pas satisfaite de la situation, même si elle se résout à l'accepter.

(112) A : ah ben ça c'est (.) ça c'est le sexe [1ça c'est pas c'est pas pareil (.) ça c'est passé ça (RIRE)
H : [1ah: (*dit en riant*)
A : [1non c'est
H : [1mais ils en ont profité ben **coudon** hein ça ça peut pas durer [2éternellement
*[CFPQ, sous-corpus 4, segment 1, page 7, ligne 8]*

Seule la composante « situation […] non conforme à mes attentes ou à mes désirs » de la définition de Dostie (2004, p. 203) pour cette lexie est prise en compte par notre système de description :

> 6. Coudon ≅
> Réagissant à une situation//
> j'indique que je me résous à accepter cette situation, non conforme à mes attentes ou à mes désirs, que je ne pourrai pas changer.

Nous n'avons pas prévu de mot-clé afin de rendre compte de la composante « je me résous à accepter cette situation […] que je ne pourrai pas changer ». C'est un des points faibles de la méthode de description componentielle que nous avons choisie d'adopter de ne pouvoir rendre compte en détail de toutes les composantes particulières de certains marqueurs.

**2.21 CRIF**

Signifiants : 62 (60 *crif,* 1 *criff,* 1 *criffe*)
MI : 54

CRIF paraît être utilisé comme substitut du sacre CRISSE.

**CRIF1**

Mot-clé : **inattendu**
Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**CRIF2**

Mots-clés : **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

### 2.22 CRIME

Signifiants : 28 (*crime*)

MI : 24

CRIME semble être utilisé comme substitut du sacre CRISSE.

**CRIME1**

Mot-clé : **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**CRIME2**

Mots-clés : **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

### 2.23 CRISSE

Signifiants : 180 (177 *crisse,* 3 *criss*)

MI : 132

CRISSE est un sacre prototypique. Le vocable nous a servi de modèle dans notre analyse des sacres au point 1.5.2.

**CRISSE1**

Mots-clés : **forte émotion**, **inattendu**, **tabou**

Synonymes : Tous les membres de la classe SACRE1

**CRISSE2**

Mots-clés : **forte émotion**, **mauvais**, **se sentir mal**, **tabou**

Synonymes : Tous les membres de la classe SACRE2

### 2.24 CRISTIE

Signifiants : 7 (*cristie*)

MI : 5

CRISTIE paraît être utilisé comme substitut du sacre CRISSE.

**CRISTIE1**

Mot-clé : **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**CRISTIE2**

Mots-clés : **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

### 2.25 ⌜DE LA MARDE⌝

Signifiants : 36 (*de la marde*)

MI : 7

Mots-clés : **mauvais**, **infirmatif**

Le phrasème ⌜DE LA MARDE⌝ est produit afin d'exprimer un désaccord au sujet d'un état du monde qui est jugé négativement.

Dans l'extrait (113), l'énonciateur M produit ⌜DE LA MARDE⌝ lorsqu'il relate une situation où il a refusé de modifier ses habitudes de vie pour perdre du poids.

(113) MA : [3fait que t'avais le choix en:tre [4manger comme du monde ou faire beaucoup de vélo:/
[...]
M : [1ouin j'ai dit •**de la marde**° {mais;(inaud.)}
*[CFPQ, sous-corpus 30, segment 4, page 47, ligne 4]*

### 2.26 ⌜DU TOUT⌝

Signifiants : 46 (*du tout*)
MI : 6
Mot-clé : **infirmatif**
Quasi-synonymes : ⌜PAS DU TOUT⌝, PANTOUTE, ⌜VRAIMENT PAS⌝

⌜DU TOUT⌝ est issu de ⌜PAS DU TOUT⌝ et ces deux phrasèmes sont de toute évidence très similaires. Nous les distinguons en deux classes parce que cette façon de faire permet d'améliorer les performances de certaines des méthodes d'identification automatique des MI présentées au chapitre 3.

### 2.27 ÉCOUTE

Signifiants : 312 (304 *écoute*, 8 *écoutez*)
MI : 217 (215 *écoute*, 2 *écoutez*)

Le marqueur discursif ÉCOUTE est issu, par pragmaticalisation, de la forme impérative du verbe *écouter*. En accord avec Dostie (2004), nous distinguons trois utilisations du marqueur : la première réalise principalement un acte directif, la deuxième réalise principalement un acte affirmatif en plus d'un acte directif et la troisième réalise principalement un acte expressif en plus d'un acte directif.

La variation *écoutez* est nécessaire en contexte de vouvoiement et de pluriel (Dostie, 2004, p. 210).

**ÉCOUTE1**

Mot-clé : **attention**

L'acte directif de la lexie *Écoute1* telle que décrite par Dostie concerne un texte que l'énonciateur s'apprête à produire :

1. *Écoute*, T ≅
   Voulant m'assurer que tu écouteras**2** bien ce que je dirai**1** au moyen du texte T //
   je t'invite à user de tes capacités cognitives afin que cet objectif soit atteint.
   (Dostie, 2004, p. 210)

Cette utilisation semble la plus plus proche du verbe *écouter*.

**ÉCOUTE2**

Mots-clés : **mauvais**, **attention**

Quasi-synonyme : VOYONS2

La deuxième utilisation de ÉCOUTE en tant que MI correspond à la lexie *écoute***2** telle que décrite dans Dostie :

2. *Écoute* ≅
   Réagissant à des propos ou à ton comportement//
   je t'indique qu'ils ne sont pas conformes à mes attentes ou à mes désirs et je t'invite à user de tes capacités cognitives pour comprendre, grâce aux connaissances dont tu disposes, que tu dois agir différemment.
   (Dostie, 2004)

Un acte expressif de désapprobation est ici associé au marqueur en plus d'un acte directif. En (114), l'énonciateur I utilise ÉCOUTE2 afin d'indiquer qu'il trouve « mal » la situation et d'inviter un coénonciateur à « penser » :

(114) I : ben là **écoute** là ça avait pas d'allure

*[CFPQ, sous-corpus 7, segment 1, page 4, ligne 5]*

**ÉCOUTE3**

Mots-clés : **attention, affirmatif**

Quasi-synonyme : REGARDE

En utilisant cette lexie, un énonciateur réalise d'abord un acte directif qui peut être classé dans la catégorie d'actes que l'on identifie par le mot-clé « attention ». L'énonciateur réalise également un acte assertif par lequel il caractérise un élément du contexte comme étant évident, ce qui représente un cas particulier de la catégorie d'actes que l'on identifie par le mot-clé « affirmatif ».

Dans l'extrait (115), l'énonciatrice A produit ÉCOUTE afin de signifier son accord avec son coénonciateur.

(115) R : parce qu'un œuf ça doit fesser certain↓

[...]

R : mais imagine-toi que tu l'aurais eu en dedans de la figure là

A : ben **écoute**

*[CFPQ, sous-corpus 20, segment 3, page 33, ligne 10]*

La composante directive du MI semble très faible dans cet exemple.

L'acte affirmatif de cette lexie telle que décrite par Dostie porte sur un texte déjà introduit dans la conversation :

3. T1, *Écoute* (T2) ≅
Le texte T1 ayant été produit//
je t'invite à user de tes capacités cognitives pour comprendre en quoi celui-ci fait référence à quelque chose qui va de soi (en fonction de la raison que j'explicite au moyen du texte T2).
(Dostie, 2004, p. 213)

La composante facultative (entre parenthèses) de la définition de ÉCOUTE3 rend compte du fait que cette lexie puisse être accompagnée d'un texte explicatif. Dans un tel cas, il est un marqueur d'interprétation.

**2.28 ⌜EH BOY⌝**

Signifiants : 20 (10 *eh boy*, 2 *oh boy*, 9 *ah boy*)
MI : 20
Mot-clé : **hors du commun**
Quasi-synonymes : SEIGNEUR1, OUF2, FIOU2

Le MI ⌜EH BOY⌝ est utilisé en réaction à quelque chose qui est hors du commun. Il n'est pas nécessairement lié à quelque chose d'inattendu, de positif ou de négatif. En (116), l'énonciateur E signale à l'aide de *oh boy* qu'il est impressionné par l'événement plutôt négatif décrit par N :

(116) N : ben là on s- s- il fallait il fallait que tu euh tu fasses attention pour pas t- pour pas que le vent t'emporte à force qu'il ventait
[...]
E : <p<**oh boy**>>
*[CFPQ, sous-corpus 8, segment 2, page 14, ligne 11]*

En (117), nous savons que ⌜EH BOY⌝ n'exprime pas la surprise puisque l'énonciateur E connaît l'endroit dont il est question dans la conversation :

(117) N : ouin mais là c'est GRAND là
E : **EH:::// boy** que c'est grand (.) c'est [1pas croyable
*[CFPQ, sous-corpus 8, segment 7, page 71, lignes 17-18]*

**2.29 ENVOYE**

Signifiants : 45

MI : 42

Mot-clé : **encouragement**

Quasi-synonymes : GO, ⌜LET'S GO⌝

ENVOYE est utilisé pour encourager quelqu'un à faire quelque chose ou pour signaler le moment de faire quelque chose.

En (118), l'énonciatrice J s'adresse à un coénonciateur inconnu, comme si elle cherchait à encourager la tranche de fruit à tomber dans son verre :

(118) J : <p<attends un peu>> (.) non: il y a-tu/ ah ok la petite tranche [1(inaud.) (*en versant de la sangria*)
L : [1tu veux-tu de la glace↑
J : je vais [1aller en chercher
F : [1ta petite tranche/ t'as-tu peur que ta petite tranche a-:/
J : **envoye** donc [1elle va-tu y aller↑
F : [1il y a pas rien dessus là↑ (*en prenant un couteau et en aidant Julie à mettre la tranche de fruit dans son verre*) tu la veux-tu/
*[CFPQ, sous-corpus 18, segment 7, page 79, ligne 18]*

L'expression « envoye à la maison » se trouve à trois reprises dans le CFPQ. Il s'agit sans doute d'un emploi figé du MI ENVOYE dans un phrasème.

**2.30 FIOU**

Signifiants : 12 (*fiou*)

MI : 12

**FIOU1**

Mots-clés : **inattendu**, **bien**, **se sentir bien**

Quasi-synonymes : ⌜UNE CHANCE⌝, OUF1

Dans son premier sens, FIOU est utilisé par un énonciateur pour exprimer un soulagement, pour indiquer que l'état de la situation est différent de ce qu'il pensait à l'origine et que cet état réel de la situation est positif.

En (119), l'énonciateur É utilise FIOU1 afin de communiquer qu'il trouve bien que ce ne soit que les poignées du quatre-roues qui soient chauffantes, contrairement ce qu'il pensait :

(119) É : le le quatre-roues est chauffant// (.) il y a du chauffage//(inaud.) (*dit avec surprise en s'adressant à Gilles*)
G : les [1non non les poignées et les pouces les poignées/ là// (*en faisant comme s'il tenait des poignées de quatre-roues*)
[...]
É : ah OK **fiou** (.) (inaud.) comment qu'ils font [1pour mettre du chauffage (*dit en riant*)
*[CFPQ, sous-corpus 1, segment 2, page 23, ligne 15]*

Ce sens de FIOU est souvent précédé par l'unité *ok* ou les unités *ah ok*, ce qui est une indication qu'il est produit suite à la réception d'une information nouvelle.

FIOU1 est parfois utilisé comme marqueur d'interprétation pour introduire une explication du soulagement qu'il exprime. Dans l'exemple (120), l'énonciatrice est soulagée de constater d'être hors de danger. Le contexte décrit ne semble pas comporter de situation inattendue, le *fiou* de l'énonciatrice laisse donc entendre qu'elle a subi un moment de crainte. La personne a cru un instant être en danger, puis a constaté que ce n'était pas le cas. Notons que le *fiou* de l'exemple et la proposition qu'il introduit (« on est dans la maison ») aurait pu être transcrits comme étant du discours rapporté.

(120) L : pis Bernard justement il s'informait aux garde: s-pê- gardes-chasses [1en tout cas ceux qui font visiter euh: au sujet des coyotes pis il dit •on en a entendu hurler mais vraiment HURler là° ça te glace↓ le sang↓ [2nous-autres on était couchés rendu là ça **fiou** on est dans la maison pis c'est quand ils ont pogné quelque chose là (.) [3t'sais c'est comme un cri de:
*[CFPQ, sous-corpus 18, segment 6, page 64, ligne 17]*

**FIOU2**

Mot-clé : **hors du commun**

Quasi-synonymes : OUF2, SEIGNEUR1, ⌜EH BOY⌝

Dans son deuxième sens (une seule occurrence dans le CFPQ), le MI FIOU est utilisé pour exprimer le sentiment qu'un état du monde est hors du commun. Le caractère inattendu de cet état du monde ne semble pas être obligatoire. En (121), l'énonciatrice signale à l'aide de FIOU2 qu'elle est impressionnée par la quantité de vis achetée. La proposition « il va en avoir du cadre chez vous » guide l'interprétation de *fiou* dans ce sens.

(121) I : ça c'est [1la DEUxième étape de (.) [2une fois qu'on a eu fini nos petites boîtes là (RIRE) la fois d'après on a acheté un CENT (.) paquet de cent vis avec les [3chevilles (.) qui fittent [4ensemble (*en faisant semblant de tenir un des paquets dont elle parle*)
[...]
É : [4**fiou**: il va en avoir du cadre chez vous [5(RIRE)
*[CFPQ, sous-corpus 16, segment 9, page 85, ligne 8]*

**2.31 FRANCHEMENT**

Signifiants : 26 (*franchement*)

MI : 22

**FRANCHEMENT1**

Mot-clé : **affirmatif**

Quasi-synonymes : ⌜POUR VRAI⌝3, SÉRIEUX1, VRAIMENT, VOYONS3

FRANCHEMENT1 sert à présenter une affirmation ou une question comme vraie. Il peut être confondu avec ce qui est parfois décrit comme étant un « adverbe de phrase » dans les grammaires normatives (dans Grevisse et Goosse, 2007, p. 469 par exemple). Aussi appelé « modalisateur d'énonciation » par Pop (2001, p. 13-19), ce type d'adverbe porte sur une énonciation d'une manière très similaire à un marqueur d'interprétation. Le *franchement* de (122), qu'il soit adverbe de phrase ou MI, veut dire à peu près « je suis franc en disant que P ».

(122) É : ah oui le Québec là **franchement** côté embonpoint on est corrects encore
*[CFPQ, sous-corpus 1, segment 8, page 116, ligne 6]*

**FRANCHEMENT2**

Mots-clés : **hors du commun**, **mauvais**

Quasi-synonymes : ⌜POUR VRAI⌝2, ÉCOUTE2

FRANCHEMENT2 réalise des actes illocutoires expressifs liés à la désapprobation devant quelque chose de hors du commun. Il est souvent émis en réaction à ce qu'un coénonciateur a dit. En (123), l'énonciatrice produit FRANCHEMENT2 en réaction à une affirmation qu'on lui rapporte et qu'elle juge négativement.

(123) J : [...] je disais que je trouvais ça BEAU la Gaspésie tout ça [1pis je parlais avec une copine française qui <len<•mais il y a pas de magasins°>> (*dit avec exaspération*) [2 elle dit ça de même j'ai dit [...]
L : [1ah c'est beau c'est BEAU
F : [1hum hum
F : [2ah **franchement**
*[CFPQ, sous-corpus 18, segment 6, page 60, ligne 20]*

**2.32 GO**

Signifiants : 15 (*go*)
MI : 15
Mot-clé : **encouragement**
Quasi-synonymes : ENVOYE, ┌LET'S GO┐

GO peut être utilisé pour encourager quelqu'un à faire quelque chose ou pour signaler le moment précis où il faut faire quelque chose, comme le départ d'une course ou le début d'une discussion.

Dans l'extrait (124), l'énonciatrice S produit GO pour indiquer à ses coénonciateurs de commencer à converser.

(124) G : <ff<c'est parti>>
S : <f<on commence
L : <f<à l'ordre>>
S : un deux trois **go**>> (*en regardant Émilie*)
*[CFPQ, sous-corpus 1, segment 1, page 7, ligne 15]*

Les énonciateurs du CFPQ se disent parfois GO à eux-mêmes, comme pour s'encourager. Le GO de l'énoncé (125) aurait pu être transcrit comme étant du discours direct :

(125) H : 1t'sais dernière minute là vraiment dernière minute fait que là: euh ok t'sais je me lance là t'sais euh **go** t'sais j'étais pas prêt [...]
*[CFPQ, sous-corpus 14, segment 6, page 58, ligne 1]*

GO est utilisé une fois dans l'expression « go go go » dans le CFPQ. Nous croyons qu'il s'agit ici d'un phrasème, mais nous manquons de données pour en déterminer le sens exact.

**2.33 HEILLE**

Signifiants : 1387 (*heille*)
MI : 1387
Mots-clés : **hors du commun**, **attention**

L'unité *heille* joue un grand nombre de rôles pragmatiques en français québécois oral qu'il peut être difficile de distinguer. Il peut être un marqueur d'appel à l'écoute, un marqueur d'interprétation ou un marqueur de réalisation d'actes illocutoires.

En tant que MI, *heille* sert à inviter un coénonciateur à porter son attention sur quelque chose de hors du commun. En (126), par exemple, l'énonciateur O cherche à faire en sorte que son coénonciateur Y réfléchisse à la chose hors du commun qu'il a dit.

(126) Y: [3Signature Mario Tétrault <p<{oui/;(inaud.)}>>
O: <f<non non non>> **heille** (*en levant le menton comme en signe d'incrédulité*) t'es-tu malade crisse [1Signature Mario Tétrault
Y: [1là tu là tu vas payer
*[CFPQ, sous-corpus 21, segment 3, page 37, ligne 15]*

La composante **inattendu** n'est pas nécessairement associée à ce marqueur, comme en témoigne l'exemple (127) où l'énonciateur raconte des événements avec lesquels il est familier :

(127) S: câ- **heille** vraiment là sérieux là **heille** (*inaud.*) je me suis brossé les dents le matin (.) arrive le soir (*en ouvrant ses paumes vers le haut comme pour représenter le vide*) (.) plus de brosse à dents (*en écartant les mains comme pour exprimer son incompréhension*)
*[CFPQ, sous-corpus 21, segment 2, page 27, ligne 4]*

**2.34 HEIN**

Signifiants : 2092 (*hein*)
MI : 2092

**HEIN1**
Mot-clé : **question**

Le premier sens de HEIN est similaire au mot *quoi?* utilisé seul. Il sert à indiquer que quelque chose n'a pas été compris par l'intermédiaire d'un acte directif interrogatif.

En (128), l'énonciatrice C utilise HIEN1 afin de signaler qu'elle n'a pas compris ce qu'a dit sa coénonciatrice et de l'interroger à ce sujet.

(128) C: [1j'avais trouvé ça ben drôle (.) [2bref
MÈ: [2hum: ouin (1") <all<un peu l'opposé là/>>
C: **hein**/
MÈ: un peu l'opposé [1là
*[CFPQ, sous-corpus 25, segment 2, page 19, ligne 11]*

Lorsque utilisé comme marqueur d'interprétation, HEIN1 peut être paraphrasé par « *N'est-ce pas?* ». En (129), l'énonciateur indique qu'il recherche une confirmation de ce qu'il dit avec HEIN1 :

(129) A : là il a pris <f<conscience>> de la <f<vérité\>> (3") mais il est tard **hein**
*[CFPQ, sous-corpus 20, segment 3, page 23, ligne 1]*

**HEIN2**
Mot-clé : **inattendu**

HEIN peut aussi être utilisé de manière non-interrogative, pour exprimer l'incompréhension en réaction à quelque chose d'inattendu. En (130), par exemple, l'énonciateur I trouve intriguant la mention de « genou de cochon mort » de la part de son coénonciateur.

(130) M : [3oui j'ai des guimauves avec t- euh du genou de cochon MORT/
[...]
MA : oui
I : [1<pp<**hein**>> (*dit en s'adressant à Marc-André en plissant le nez et en fronçant les sourcils comme en signe d'incompréhension*)
*[CFPQ, sous-corpus 30, segment 4, page 54, lignes 20-23 et page 55, lignes 1-2]*

La prosodie de HEIN2 est très différente de celle de HEIN1, le premier étant prononcé avec une voyelle allongée et parfois descendante.

### 2.35 ⌜JE COMPRENDS⌝

Signifiants : 120 (*je comprends*)
MI : 45
Mot-clé : **affirmatif**
Quasi-synonymes : METS-EN, TELLEMENT, VRAIMENT

Le MI ⌜JE COMPRENDS⌝ est toujours produit en réaction à une affirmation ou à une interrogation, habituellement d'un coénonciateur. Avec *je comprends* en (131), F signifie qu'il est en accord avec ce que P dit, à savoir que *c'était vraiment fou*.

(131) P : [1ah mais c'était c'était vraiment euh: c'était vraiment fou les matchs qu'on a joués là c'était incroyable là (.) la pluie euh::: elle nous tombait vraiment (.) ben vraiment [2souvent dessus mais je veux dire [3un moment donné tu la sentais plus là/
F : [2ben (.) **je comprends** (*dit en riant*)
*[CFPQ, sous-corpus 9, segment 1, page 6, ligne 14]*

Le MI ⌜JE COMPRENDS⌝ est prononcé d'une manière caractéristique, avec un allongement de la syllabe /kõ/ ou une intonation montante. Il est aussi typiquement précédé de *ben* ou *ben là*, comme en (132) :

(132) S : [11ben là/ **je comprends**/
*[CFPQ, sous-corpus 15, segment 6, page 113, ligne 2]*

Parce que dans le phrasème ⌜JE COMPRENDS⌝ se trouve le pronom *je*, il très difficile de le distinguer des emplois du verbe *comprendre* sur le plan lexical, comme dans la phrase *Je comprends*, au sens littéral. Cette phrase n'a cependant pas suivi le processus de pragmaticalisation qu'a subi le MI ⌜JE COMPRENDS⌝. Celui-ci a notamment une composante assertive.

**2.36 ⌜LET'S GO⌝**

Signifiants : 2 (*let's go*)
MI : 2
Mot-clé : **encouragement**
Quasi-synonymes : GO, ENVOYE

Le phrasème ⌜LET'S GO⌝, issu de l'anglais, est utilisé pour encourager une personne à faire quelque chose. Dans l'exemple (133), l'énonciateur M utilise ⌜LET'S GO⌝ pour inciter le coénonciateur Germain à suivre la mise au cours d'une main de poker.

(133) A : heille RiCHARD vas-y mon [2Richard
M : [2(*il pose des jetons au centre de la table*)
A : on [1retombera pas ici (*inaud.*)
M : [1**let's go** mon [2GerMAIN::\
*[CFPQ, sous-corpus 27, segment 6, page 87, ligne 16]*

**2.37 MALADE**

Signifiants : 72 (*malade*)
MI : 3
Mots-clés : **hors du commun**, **inattendu**
Quasi-synonymes : ⌜MON DIEU⌝, ⌜MON DOUX⌝, ⌜MY GOD⌝, OUPELAILLE2

Le mot-phrase MALADE peut être vu comme la contraction de la phrase « C'est malade! ». Le MI exprime l'étonnement devant quelque chose d'inattendu et hors du commun.

Dans l'exemple (134), l'histoire de F, qui provoque l'énonciation de MALADE par P, est digne d'intérêt par son caractère hors de l'ordinaire :

(134) F : [...] je sais pas si c'est vrai mais (.) j'ai entendu dire que (.) il y a <all<des des>> noirs qui ont comme un muscle plus développé genre dans les jambes pour courir les (.) les longues distances (.) pis que: (.) les les blancs (.) les personnes de race blanche (.) ont plus (.) ils ont comme (.) euh euh une sorte de muscle qui est MEIlleure à la NAge (.) pis t'essaies de voir un noir nager ben ça marche pas [2genre
[...]
P : **malade**
*[CFPQ, sous-corpus 9, segment 9, page 126, lignes 6-8]*

Dans le CFPQ, *malade* est utilisé 2 fois, par la même personne, dans le sens de « t'es malade ». La locutrice en question a tendance à faire des coupures et des contractions morphologiques. Nous croyons que cet emploi de *malade* n'est pas conventionnel.

Ce MI ne doit pas être confondu avec l'emploi adjectival non discursif, tel qu'exemplifié en (135) :

(135) J-M : ouais (RIRE) mais t'sais genre c'est juste si il y a de quoi de SUper bon genre pis là t'sais une bonne sauce là **MALADE** là pis t'sais [...]
*[CFPQ, sous-corpus 10, segment 8, page 85, ligne 8]*

**2.38 MAUDIT**

Signifiants : 54 (*maudit*)
MI : 18

MAUDIT est parfois considéré comme un sacre, mais son utilisation semble davantage similaire à celle des substituts de sacres.

**MAUDIT1**

Mot-clé: **inattendu**
Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**MAUDIT2**

Mots-clés: **mauvais**, **se sentir mal**
Quasi-synonymes : Les membres de la classe SUBSTITUT2.

**2.39 MAUTADIT**

Signifiants : 9 (*mautadit*)
MI : 3

MAUTADIT est vraisemblablement une forme de substitution du quasi-sacre ou sacre non-prototypique MAUDIT.

**MAUTADIT1**

Mot-clé: **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**MAUTADIT2**

Mots-clés: **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

### 2.40 MERDE

Signifiants : 14 (*merde*)

MI : 10

Mot-clé : **mauvais**, **se sentir mal**, **tabou**

Utilisé comme MI, MERDE est un synonyme proche de ZUT, mais l'unité est, en plus, connotée par la composante **tabou**. Contrairement aux sacres, le caractère mauvais du signifiant *merde* n'a pas pour origine un interdit religieux, mais plutôt un interdit lié à son référent scatologique.

En (136), l'énonciatrice utilise MERDE pour indiquer qu'elle éprouve un sentiment négatif et que le fait qu'elle a un cours d'E.C.C. n'est pas bien :

(136) MA : arrête avec la prof d'E.C.C. là c'est la fin de semaine (*dit en riant)*

[...]

C : [1**merde** on a un cours vendre- euh jeudi [2à la pre- dernière période

*[CFPQ, sous-corpus 3, segment 5, page 82, ligne 18]*

La forme *marde* n'est pas employée comme MI dans le CFPQ (sauf dans la construction ⌜DE LA MARDE⌝).

### 2.41 METS-EN

Signifiants : 67 (*mets-en*)
MI : 66
Mot-clé : **affirmatif**
Quasi-synonymes : ⌜JE COMPRENDS⌝, TELLEMENT, VRAIMENT

Le MI METS-EN sert à manifester son accord en réaction à une chose dite par un coénonciateur. Dans l'exemple (137), *mets-en* est utilisé deux fois comme MI. Dans sa seconde occurrence, il est en relation avec une proposition introduite par *que*.

(137) D : il y a une fille qui s'appelle Féline aussi
ME : ouin
MA : elle fait peur
(RIRE GÉNÉRAL)
D : **mets-en**
ME : on a genre la jungle au complet
D: **mets-en** qu'elle me fait peur [...]
*[CFPQ, sous-corpus 3, segment 1, page 2, ligne 20 - page 3, ligne 5]*

Nous pourrions paraphraser METS-EN de la façon suivante :

**Mets-en ≅**
Réagissant à ce que tu dis //
j'indique que c'est vrai et que tu pourrais même en mettre plus.

Le tour de parole exemplifié en (138) montre l'unique exemple où *mets-en* est utilisé de manière intraphrastique dans le CFPQ. On y voit le verbe *mettre* en relation syntaxique avec *pas*.

(138) SA : [3(RIRE) heille **mets-en** pas [4trop Sonia
*[CFPQ, sous-corpus 7, segment 7, page 69, ligne 6]*

**2.43 ⌜MON DIEU⌝**

Signifiants : 177 (177 *mon dieu*)
MI : 177
Mots-clés : **inattendu**, **hors du commun**
Quasi-synonymes : ⌜MON DOUX⌝, ⌜MY GOD⌝, OUPELAILLE2, MALADE, AYOYE2

Employé comme MI, le phrasème ⌜MON DIEU⌝ sert à signaler que quelque chose est inattendu et hors du commun. La lexie peut être utilisée dans une grande variété de contextes, tant neutres, que négatifs ou positifs. En (139), l'énonciateur F exprime son étonnement à l'aide de ⌜*mon dieu*⌝ :

(139) J : [...] une grande grande grande pièce où c'est que tu aurais pu mettre des s- des sleepings pour dix personnes si tu veux (.) [6mais de l'extérieur ça avait l'air petit↑ [...]
F : [6eh **mon dieu**
*[CFPQ, sous-corpus 18, segment 6, page 65, ligne 18]*

Contrairement aux sacres, ⌜MON DIEU⌝ (et ses substituts ⌜MON DOUX⌝ et ⌜MY GOD⌝) n'est pas utilisé pour exprimer la colère.

**2.44 ⌜MON DOUX⌝**

Signifiants : 78 (*mon doux*)
MI : 78
Mots-clés : **inattendu**, **hors du commun**
Quasi-synonymes : ⌜MON DIEU⌝, ⌜MY GOD⌝, OUPELAILLE2, MALADE, AYOYE2

⌜MON DOUX⌝ semble être une forme de substitution de ⌜MON DIEU⌝. Ce dernier a peut-être été associé à une connotation sacrilège par le passé, ce qui aurait mené à la création de ⌜MON DOUX⌝.

La collocation discursive « mon doux seigneur » se trouve 6 fois dans le CFPQ. L'origine de ⌜MON DOUX⌝ est peut-être également liée à cette construction.

**2.45 ⌜MY GOD⌝**

Signifiants : 17 (*my god*)
MI : 17
Mots-clés : **inattendu**, **hors du commun**
Quasi-synonymes : ⌜MON DIEU⌝, ⌜MON DOUX⌝, OUPELAILLE2, MALADE, AYOYE2

Traduction littérale de « mon dieu » en anglais, ⌜MY GOD⌝ joue le rôle d'une forme de substitution de ⌜MON DIEU⌝.

**2.46 OSTIE**

Signifiants : 500 (*ostie*)
MI : 441

OSTIE est un sacre prototypique.

**OSTIE1**
Mots-clés : **forte émotion**, **inattendu**, **tabou**
Synonymes : Tous les membres de la classe SACRE1

**OSTIE2**
Mots-clés : **forte émotion**, **mauvais**, **se sentir mal**, **tabou**
Synonymes : Tous les membres de la classe SACRE2

**2.47 OSTIFIE**

Signifiants : 12 (*ostifie*)
MI : 10

OSTIFIE est apparemment utilisé comme substitut du sacre OSTIE.

**OSTIFIE1**
Mot-clé : **inattendu**
Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**OSTIFIE2**
Mots-clés : **mauvais**, **se sentir mal**
Quasi-synonymes : Les membres de la classe SUBSTITUT2.

**2.48 OSTINE**

Signifiants : 2 (*ostine*)
MI : 2

OSTINE est apparemment utilisé comme substitut du sacre OSTIE.

**OSTINE1**
Mot-clé: **inattendu**
Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**OSTINE2**
Mots-clés: **mauvais**, **se sentir mal**
Quasi-synonymes : Les membres de la classe SUBSTITUT2.

**2.49 OSTIQUE**

Signifiants : 11 (*ostique*)
MI : 8

OSTIQUE semble être utilisé comme substitut du sacre OSTIE.

**OSTIQUE1**
Mot-clé: **inattendu**
Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**OSTIQUE2**
Mots-clés: **mauvais**, **se sentir mal**
Quasi-synonymes : Les membres de la classe SUBSTITUT2.

**2.50 OUF**

Signifiants : 53 (*ouf*)
MI : 53

**OUF1**
Mots-clés : **inattendu, bien**, **se sentir bien**
Quasi-synonymes : FIOU1, ⌜UNE CHANCE⌝

OUF1 est utilisé (très rarement dans le CFPQ) pour exprimer un soulagement, comme FIOU1. En utilisant OUF1 en (140), l'énonciateur S indique en blague qu'il est heureux de se voir confirmer que la fée des glaces existe.

(140) S : <len<hein\>> (.) [1le Père Noël existe pas\

[...]

Y : non mais la fée des glaces oui en tout cas

S : câ- **ouf** fiou heille

*[CFPQ, sous-corpus 21, segment 1, page 16, ligne 17]*

**OUF2**

Mot-clé : **hors du commun**

Quasi-synonymes : FIOU2, SEIGNEUR1, ⌜EH BOY⌝

Le sens le plus fréquent de OUF dans le CFPQ correspond au sens 2 de FIOU. En (141), l'énonciatrice E produit OUF2 pour indiquer que la situation dont elle parle est hors du commun :

(141) E : mais ça fait longtemps que les enseignants disent qu'il y a un problème là à plein de niveaux là euh: (.) i:ls **ouf** mon dieu j'ai vu des re- [1des reportages là-dessus (RIRE)

*[CFPQ, sous-corpus 2, segment 7, page 87, ligne 16]*

**2.51 OUPELAILLE**

Signifiants : 22 (*oupelaille*)

MI : 22

**OUPELAILLE1**

Mots-clés : **inattendu**, **mal**, **se sentir mal**

Quasi-synonyme : OUPS

Comme pour OUPS, OUPELAILLE1 est produit lorsqu'un énonciateur se rend compte que quelqu'un a commis une erreur qui a provoqué une situation négative. Dans l'extrait (142), l'énonciatrice R produit OUPELAILLE1 après avoir été informée qu'elle a bu dans un verre qui n'était pas le sien.

(142) [*R boit dans un verre.*]

P : heille tu bois mon eau (*en s'adressant à Raphaëlle*) <P9,L8>

R : **oupelaille**

*[CFPQ, sous-corpus 13, segment 1, page 9, lignes 8-9]*

**OUPELAILLE2**

Mots-clés : **inattendu**, **hors du commun**

Quasi-synonymes : ⌜MON DIEU⌝, ⌜MY GOD⌝, OUPELAILLE2, MALADE, ⌜MON DOUX⌝

OUPELAILLE2 est utilisé pour indiquer que quelque chose est hors du commun et inattendu. Dans l'extrait (143), l'énonciatrice J produit OUPELAILLE2 en rapportant le discours d'une personne étonnée par le comportement exubérant de son enfant.

(143) J :[7<f<heille>> quand qu'elle m'a vue toi une chance qu'il y avait une barrière (.) [8non non mais la gardienne en revenait pas↑ la hein↑ (.) elle dit [9•**oupelaille**↑ ok°

*[CFPQ, sous-corpus 15, segment 9, page 147, ligne 9]*

**2.52 OUPS**

Signifiants : 38

MI : 38

Mots-clés : **inattendu**, **mal**, **se sentir mal**

Quasi-synonyme : OUPELAILLE1

Un énonciateur produit OUPS lorsqu'il se rend compte que quelqu'un a commis une erreur qui a provoqué une situation négative. En (144), l'énonciatrice utilise OUPS pour signaler qu'elle se sent mal suite à son expulsion inattendue de salive :

(144) C: [1mai:s j'ai fait un ostie de speech là j'ai dit •écoute (.) je suis pas une BS je suis une étudiante ok/ cinquante-deux piastres c'est un livre de psychologie° **oups** excuse (*dit en s'adressant à Patrick*) ah [2je postillonne (*dit en riant et en passant sa main sur le bras de Patrick comme pour le nettoyer*)
*[CFPQ, sous-corpus 25, segment 6, page 92, ligne 10]*

**2.53 PANTOUTE**

Signifiants : 90
MI : 22
Mot-clé : **infirmatif**
Quasi-synonymes : ⌜DU TOUT⌝, ⌜PAS DU TOUT⌝, ⌜VRAIMENT PAS⌝

Sur 90 occurrences dans le CFPQ, le signifiant *pantoute* n'est utilisé que 22 fois comme MI. Il est alors produit en réponse à une affirmation dans le sens de « pas du tout », comme en (145) :

(145) J : [2mais t'écoutes-tu la télé pour la peine/
M : **pantoute**
*[CFPQ, sous-corpus 12, segment 8, page 140, ligne 17]*

**2.54 ⌜PAS DU TOUT⌝**

Signifiants : 36 (*pas du tout*)
MI : 23
Mot-clé : **infirmatif**
Quasi-synonymes : ⌜DU TOUT⌝, PANTOUTE, ⌜VRAIMENT PAS⌝

⌜PAS DU TOUT⌝ semble équivalent à ⌜DU TOUT⌝ lorsque utilisé comme MI.

**2.55 ⌜PAS VRAIMENT⌝**

Signifiants : 79 (*pas vraiment*)

MI : 14

Mot-clé : **infirmatif partiel**

Quasi-synonymes : BOF, ⌜C'EST ENCORE DRÔLE⌝

Nous avons traité le vocable ⌜PAS VRAIMENT⌝ dans un article en 2007 (Lapointe, 2007). Les définitions des différentes lexies de ce vocable sont articulées autour des concepts de prototypie et de vérité. La définition que nous avions proposée au sujet de l'utilisation du MI ⌜PAS VRAIMENT⌝ fait appel à un argument « information communiquée par un énoncé » qui ne correspond pas tout à fait à la vérité.

« Q » ⌜*pas vraiment*⌝ ≅
Je signale que l'information [communiquée / mise en question] par l'énonce « Q » est vraie**1** d'une manière qui n'a pas fortement les caractéristiques qui font qu' « être vrai**1** » est vraiment**1** « être vrai**1** ».

Nous pouvons reformuler l'essentiel de cette définition de manière plus concise pour les besoins de cette thèse :

Pas vraiment ≅
Réagissant à ce que tu dis //
j'indique que ce n'est pas complètement vrai.

En (146), l'énonciateur C utilise ⌜PAS VRAIMENT⌝ afin d'indiquer qu'il ne croit pas que l'affirmation « ça vous aide [...] à savoir ce que vous voulez faire plus tard » est complètement vraie :

(146) M: fait que ça vous aide pas à savoir euh ce que vous voulez faire plus tard
ME : non ben [1[...]
C : [1ben **pas vraiment** là
*[CFPQ, sous-corpus 3, segment 1, page 6, lignes 2-4]*

**2.56 ⌜POUR VRAI⌝**

Signifiants : 73 (58 *pour vrai*, 2 *pour le vrai*, 13 *pour de vrai*)
MI : 51

Il existe une multitude d'emplois du phrasème ⌜POUR VRAI⌝. Nous les avons décrits dans le cadre de notre mémoire de maîtrise (Lapointe, 2005). Le phrasème peut être utilisé à l'intérieur d'une phrase comme adverbe, marqueur d'interprétation, marqueur de réalisation d'actes illocutoires ou connecteur textuel. Dans notre mémoire, nous n'avions décrit que 2 emplois de ⌜POUR VRAI⌝ en tant que MI. Nous soupçonnions à l'époque la possibilité d'un autre emploi du vocable qui correspondrait à celui de *vraiment*6 tel que décrit dans le mémoire. Depuis, nous avons observé ce nouvel emploi dans le CFPQ et avons choisi de le décrire ici. Il est présenté plus bas en tant que ⌜POUR VRAI⌝2.

**⌜POUR VRAI⌝1**
Mots-clés : **inattendu**, **question**
Quasi-synonyme : SÉRIEUX2

L'emploi le plus fréquent du phrasème ⌜POUR VRAI⌝ se classe parmi les MI qui expriment l'étonnement. Le marqueur est accompagné d'une intonation montante, caractéristique de l'acte directif interrogatif.

En (147), l'énonciateur F exprime son étonnement et questionne J au sujet de la vérité de son affirmation :

(147) J :[2mon chum il veut pas (inaud.) d'hémérocalles chez nous il aime pas ça
LO : [3ah d'accord
F : [3hein/ [4HEIN:/ **pour VRAI**/
*[CFPQ, sous-corpus 18, segment 2, page 22, lignes 14-16]*

**⌜POUR VRAI⌝2**
Mots-clés : **hors du commun, mauvais**
Quasi-synonymes : SEIGNEUR2, FRANCHEMENT2

Le second sens de ⌜POUR VRAI⌝2 sert à exprimer la désapprobation face à quelque chose de hors du commun. Il n'est pas associé à une intonation montante et n'est pas nécessairement produit en réaction à une affirmation du coénonciateur. Il peut notamment être comparé à l'emploi similaire de FRANCHEMENT2.

En (148), l'énonciateur exprime sa désapprobation à l'aide de ⌜POUR VRAI⌝2 devant l'idée hors du commun de mettre des caméras dans une chambre d'invités :

(148) SA : [1tu peux avoir des caméras sur Internet tu peux voir chez vous qu'est-ce [2qui se passe
[...]
SA :[3ouin mais si t'es ben quand tu mets ça dans la chambre à coucher des invités là t'sais
I : ben là:
(RIRE GÉNÉRAL)
I : **pour de vrai** là
*[CFPQ, sous-corpus 7, segment 2, page 21, ligne 6 à page 22, ligne 3]*

En (148), l'énonciatrice n'est pas mise en face d'une information inattendue. Elle est plutôt outrée devant un geste déplacé.

**┌POUR VRAI┐3**

Mot-clé : **affirmatif**

Quasi-synonymes : SÉRIEUX1, VRAIMENT

Un autre emploi de ┌POUR VRAI┐ n'est également pas associé à une intonation montante et n'exprime pas l'étonnement. Il sert à présenter une affirmation ou une question comme vraie. En (149), l'énonciateur utilise ┌POUR VRAI┐3 afin que son affirmation (déguisée en question) soit interprétée sérieusement.

(149) V : 1ben tu conn- **pour vrai** là↓ concrètement est-ce que tu connais beaucoup d'hommes qui d- aimeraient mieux être sur un étage seulement d'hommes↑
*[CFPQ, sous-corpus 10, segment 3, page 40, ligne 4]*

**2.57 REGARDE**

Signifiants : 803 (795 *regarde*, 8 *regardez*)

MI : 609

Mots-clés : **attention**, **affirmatif**

Quasi-synonyme : ÉCOUTE1

REGARDE est un des marqueurs discursifs les plus fréquents dans le CFPQ. Le vocable REGARDE a été l'objet d'une description détaillée dans Dostie (2004). Il est issu du verbe *regarder* par pragmaticalisation. Comme pour ÉCOUTE, il entretient des relations sémantiques très fortes avec la forme impérative du verbe dont il est issu.

Avec l'utilisation du MI REGARDE en (150), l'énonciatrice VI affirme son accord avec ce que dit sa coénonciatrice. La composante directive de la lexie, qui se paraphraserait par « pense à cela », semble être adressée à un public fictif qui aurait des idées déraisonnables sur le sujet dont il est question.

(150) MA : heille qu'est-ce qu'ils vont faire/ t'sais/ (.) si elle est si elle est pas capable de corriger ses éLÈves là elle va faire quoi/
VI : ben **regarde**
*[CFPQ, sous-corpus 19, segment 4, page 33, ligne 20]*

Voici la définition donnée par Dostie 2004 pour cette lexie :

2b. Regarde ≅
Réagissant à quelque chose qui vient d'être dit1 ou d'être fait //
je t'invite à le regarder**2** parce que cela illustre bien qu'une certaine idée que je tiens pour vraie - comme tu sais - l'est réellement.
(Dostie, 2004, p. 219)

Dans cette définition, le verbe *regarder* est utilisé dans son sens abstrait qui équivaut à « examiner ». Les composantes « je t'invite à regarder2 » et « je t'invite à utiliser à tes capacités cognitives » (présente dans les définitions de ÉCOUTE) sont suffisamment proches pour nous permettre de regrouper les marqueurs ÉCOUTE1, ÉCOUTE2 et REGARDE sous la catégorie des marqueurs directifs d'**attention**.

La variation *regardez* est de mise en contexte de vouvoiement et de pluriel (Dostie, 2004, p. 216).

### 2.58 ⌜REGARDE DONC⌝

Signifiants : 6 (*regarde donc*)
MI : 6
Mot-clé : **inattendu**
Quasi-synonyme : TIENS4

L'utilisation de *regarde* comme verbe à l'impératif a souvent des caractéristiques très similaires à son utilisation comme MI. Comme pour d'autres verbes associés au regard (« check ça », « mate moi ça »), il peut être utilisé dans la formule « regarde donc ça » ou simplement « regarde donc » pour inviter quelqu'un à poser son regard ou son attention sur quelque chose.

La construction « regarde donc (ça) » semble se figer en phrasème dans les quelques contextes où on la retrouve dans le CFPQ. L'acte directif s'efface alors au profit d'un acte expressif.

En (151) par exemple, l'énonciateur B produit la phrase *regarde donc ça* afin d'exprimer son intérêt et son léger étonnement face à ce que S dit.

(151) S : il y avait des ta- que c'est tu m'as d- des tableaux [1qu'il y avait de:
J : [1c'est ça tous des tableaux qui sont en euh: dans l'eau [2là euh: profond (.) [3vraiment profond
B : [2sous l'eau/
R : [2sous l'eau sous l'eau
S : [3fait que tu peux louer l'équipement de de plongée/
B : ah **regarde donc** [1**ça/**
*[CFPQ, sous-corpus 15, segment 10, page 170, ligne 7]*

En (152), ⌜REGARDE DONC⌝ est utilisé dans le même sens, mais le *ça* est maintenant disparu, ce qui suggère un niveau de pragmaticalisation supérieur de la phrase ou du phrasème :

(152) I : [3mais non mais là/ au début c'était ça je me cassais la tête moi [4parce que j'étais là •il me semble qu'on a trop de photos pour mettre sur la page° [5t'sais pis là j'essayais de rentrer (*en mimant les efforts qu'elle a faits pour placer les photos sur une même page*) la photo pour que pis là elle dit •bon ben Irène on peut la découper aussi°
[...]
I : <p<ah ben **regarde donc**>> 1•cours 101 de scrapbooking on peut couper des photos°
*[CFPQ, sous-corpus 16, segment 9, page 81, ligne 5]*

Nous assistons peut-être ici à l'émergence d'un nouveau marqueur que l'on pourrait presque appeler *gardon* (ou *gadon*) à cause de ses similarités avec la forme *coudon*, elle-même issue de la construction « écoute donc ».

Alors que le MI REGARDE employé seul est prononcé en allongeant la deuxième syllabe, ⌜REGARDE DONC⌝ est souvent prononcé avec une tonalité montante et un allongement possible du *donc*.

**2.59 SACRE**

Signifiants : 15 (*sacre*)
MI : 5

SACRE semble être utilisé comme un substitut du sacre non-prototypique SACREMENT.

**SACRE1**
Mot-clé: **inattendu**
Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**SACRE2**
Mots-clés: **mauvais**, **se sentir mal**
Quasi-synonymes : Les membres de la classe SUBSTITUT2.

**2.60 SACREMENT**

Signifiants : 9 (*sacrement*)
MI : 5

SACREMENT est un sacre non-prototypique, parfois perçu comme état stigmatisé par les locuteurs québécois (Dostie, 2015, p. 60).

**SACREMENT1**

Mots-clés : **forte émotion**, **inattendu**, **tabou**
Quasi-synonymes : Les membres des classes SACRE1

**SACREMENT2**

Mots-clés : **forte émotion**, **mauvais**, **se sentir mal**, **tabou**
Quasi-synonymes : Les membres des classes SACRE2

**2.61 SACRIFICE**

Signifiants : 11 (*sacrifice*)
MI : 7

SACRIFICE semble être utilisé comme un substitut du sacre non-prototypique SACREMENT.

**SACRIFICE1**

Mot-clé: **inattendu**
Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**SACRIFICE2**

Mots-clés: **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

## 2.62 SEIGNEUR

Signifiants : 57 (*seigneur*)

MI : 54

Le signifiant *seigneur* est le plus souvent utilisé comme MI dans le CPFQ. Les trois occurrences intraphrastiques du signifiant concernent la construction *seigneur des anneaux*.

**SEIGNEUR1**

Mots-clés : **hors du commun**

Quasi-synonymes : ⌜EH BOY⌝, FIOU2, OUF2

Comme pour ⌜EH BOY⌝, le MI SEIGNEUR est produit en réaction à quelque chose de hors du commun. La composante **inattendu** ne semble pas présente dans plusieurs contextes où SEIGNEUR est employé. En (153), par exemple, l'énonciateur se rappelle un souvenir et ne peut possiblement pas être mis devant une nouvelle information :

(153) S : pis le père avait tellement travaillé fort pour monter la tente à côté de nous-autres [ 1 écoute ah **seigneur** t'sais quand tu vois dans les comiques là (.) une patte une pôle un [2 t'sais là (*en faisant semblant d'accrocher des pièces d'une tente*)
*[CFPQ, sous-corpus 15, segment 2, page 32, ligne 12]*

**SEIGNEUR2**

Mots-clés : **hors du commun**, **mauvais**

Quasi-synonymes : ⌜POUR VRAI⌝2, FRANCHEMENT2

Le MI SEIGNEUR semble également pouvoir être accompagné d'une composante désapprobative. Dans l'extrait (154), l'énonciatrice S produit SEIGNEUR2 à la suite du récit d'une situation déplaisante concernant de la musique trop forte.

(154) S : [9pis moi t'sais c'est même pas des petits jeunes qui me font peur c'est [10des adolescents [11t'sais ça met la musique à pleine [12tête pis là tu dis boum boum boum (*en frappant l'air avec ses poings, comme pour marquer le rythme de la musique*) [13boum [14et:: **seigneur**
*[CFPQ, sous-corpus 15, segment 4, page 66, ligne 16]*

**2.63 SÉRIEUX**

Signifiants : 90 (*sérieux*)
MI : 67

Les emplois du MI SÉRIEUX sont similaires à certains emplois de ⌜POUR VRAI⌝.

**SÉRIEUX1**
Mot-clé : **affirmatif**
Quasi-synonymes : ⌜POUR VRAI⌝3, VRAIMENT

SÉRIEUX1 sert à présenter une affirmation ou une question comme étant sérieuse. La lexie est décrite ainsi dans l'article de Dostie et Lanciault (2016):

> *Sérieux* permet d'abord au locuteur de qualifier de sérieuse, au sens de « vraie », voire de « sincère », sa propre énonciation [...] En ce sens, il s'apparente à une précaution oratoire. Sa présence introduit le présupposé suivant : « *Au cas où tu en douterais*, je suis sérieux en produisant P. »
> (Dostie et Lanciault, 2016, p. 370-371)

En (155), le MI SÉRIEUX1 indique que l'on doit interpréter la phrase « c'est cool des partys de Noël » de façon littérale; qu'elle n'est pas ironique :

(155) VE : [1c'est cool des partys de Noël **sérieux** c'est comme euh (2,8") je sais pas je trouve que c'était si c'est important de faire des beaux [2partys de Noël
*[CFPQ, sous-corpus 19, segment 6, page 52, ligne 2]*

Le même marqueur peur servir à situer quelque chose comme étant grave ou « non léger ». Il a alors un sens qui s'approche de ⌜POUR VRAI⌝2 ou même FRANCHEMENT2. En (156), l'énonciatrice ME produit SÉRIEUX1 sans l'accompagner d'une proposition et laisse ainsi entendre qu'elle a une opinion « sérieuse » au sujet de la pratique de production de fromage hors du commun décrite par sa coénonciatrice.

(156) D: [1ils ils mettent du fromage là
[...]
D : à pourrir dans l'eau là\(RIRE)
ME : ah: **sérieux** (*dit avec dégoût*)
*[CFPQ, sous-corpus 3, segment 8, page 129, ligne 8]*

**SÉRIEUX2**
Mots-clés : **inattendu**, **question**
Quasi-synonyme : ⌜POUR VRAI⌝1

Le marqueur SÉRIEUX2 est similaire à ⌜POUR VRAI⌝1. En (157), l'énonciatrice É réalise un acte interrogatif et exprime son étonnement à l'aide de SÉRIEUX2 après avoir été informée d'un fait auquel elle ne s'attendait pas :

(157) [*En parlant d'un test d'orientation*]

C : moi ça m'a donné aide-laitier (RIRE)

ME : **sérieux**/ (*en pointant Clodine du doigt*)

MA : pour vrai↑

*[CFPQ, sous-corpus 3, segment 1, page 9, lignes 7-9]*

L'article de Dostie et Lanciault (2016) offre une glose pour cette lexie, qui s'articule autour de l'idée du doute et d'un acte directif de confirmation :

*Sérieux***2** :
(Ayant un léger doute quant à la véracité de ce que tu viens de dire, je te demande de me le confirmer)
(Dostie et Lanciault, 2016 : 372)

Cette définition peut être représentée, avec moins de précision, par le mot-clé **inattendu** qui correspond à la paraphrase « je ne savais pas que cela allait être comme ça » et le mot-clé **question** qui correspond à la paraphrase « est-ce que c'est vrai? ».

**2.64 SIMONAQUE**

Signifiants : 10 (7 *simonaque,* 3 *simonac*)

MI : 9

L'origine de SIMONAQUE est mystérieuse. Le MI s'utilise d'une manière équivalente à un substitut de sacre.

**SIMONAQUE1**

Mot-clé : **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**SIMONAQUE2**

Mots-clés: **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

### 2.65 SUPER

Signifiants : 187 (*super*)

MI : 10

Mots-clés : **bien**, **se sentir bien**

Quasi-synonymes : COOL

Le MI SUPER est issu de l'adjectif *super*. On peut le voir comme le résultat de l'élision du « c'est » dans la phrase « C'est super! ». Il est très similaire à COOL. En (158), SUPER est utilisé pour signaler que quelque chose est bien et correspond aux désirs de l'énonciateur :

(158) D : {tu;Ø} fais-tu partie du groupe toi↑

J-M : euh (.) oui

D : **super** alors euh quelle est sont vos impressions euh depuis le début là de cette aventu:re↑

*[CFPQ, sous-corpus 10, segment 1, page 1, ligne 11-13]*

### 2.66 TABARNACHE

Signifiants : (*tabarnache*)

MI : 10

TABARNACHE est utilisé comme substitut du sacre TABARNAQUE.

**TABARNACHE1**

Mot-clé : **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**TABARNACHE2**

Mots-clés : **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

### 2.67 TABARNAQUE

Signifiants : 54 (53 *tabarnaque*, 1 *tabarnak*)

MI : 44

TABARNAQUE est un sacre prototypique.

**TABARNAQUE1**

Mots-clés : **forte émotion**, **inattendu**, **tabou**

Synonymes : Tous les membres de la classe SACRE1

**TABARNAQUE2**

Mots-clés : **forte émotion**, **mauvais**, **se sentir mal**, **tabou**

Synonymes : Tous les membres de la classe SACRE2

### 2.68 TABARNIQUE

Signifiants : 4 (*tabarnique*)

MI : 3

TABARNIQUE est utilisé comme substitut du sacre TABARNAQUE.

**TABARNIQUE1**

Mot-clé: **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**TABARNIQUE2**

Mots-clés: **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

## 2.69 TABARNOUCHE

Signifiants : 37 (*tabarnouche*)

MI : 24

TABARNOUCHE est utilisé comme substitut du sacre TABARNAQUE.

**TABARNOUCHE1**

Mot-clé: **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**TABARNOUCHE2**

Mots-clés: **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

## 2.70 TELLEMENT

Signifiants : 395 (*tellement*)

MI : 4

Mot-clé : **affirmatif**

Quasi-synonymes : ⌜JE COMPRENDS⌝, METS-EN, VRAIMENT

Même si le MI TELLEMENT est issu de l'adverbe *tellement*, le lien sémantique entre le marqueur et l'adverbe n'est pas évident. Le marqueur peut être vu comme une abréviation d'un énoncé tel que « C'est tellement vrai. ».

En (159), l'énonciateur V utilise TELLEMENT pour indiquer son accord avec l'affirmation de J-M :

(159) J-M : pis ça te coûte pas plus cher sinon là t'habites à côté de l'épicerie c'est facile là d'y aller [1 là [...]
V : [1ah **tellement**
*[CFPQ, sous-corpus 10, segment 10, page 124, lignes 13-14]*

Le signifiant *tellement* a évidemment un usage adverbial fréquent, mais celui-ci est relativement facile à distinguer du MI qui est le plus souvent isolé au point de vue syntaxique.

**2.71 TIENS**

Signifiants : 86 (*tiens*)
MI : 70

Le vocable TIENS est très polysémique. Nous avons repéré 5 utilisations différentes de MI dans le CFPQ. Dostie (2004) a décrit avec précision les lexies de ce vocable. La lexie numéro I.1b de cet ouvrage, qui est produite alors qu'un énonciateur touche de manière énergique un coénonciateur, n'a pas été repérée dans le CFPQ.

**TIENS1**

Mot-clé : **attention**

TIENS1 est produit par un énonciateur qui fait un geste de la main pour tendre quelque chose et a pour but d'attirer l'attention sur ce geste. En (160), l'énonciatrice dit TIENS1 en tendant un verre vers sa coénonciatrice :

(160) ME : [6je vas faire passer les (inaud.) qui qui veut celui-là il y en a-tu d'autres qui voulaient/ [7ou il y a juste moi (.) **tiens** (*dit en s'adressant à Magalie*)
*[CFPQ, sous-corpus 3, segment 4, page 70, ligne 12]*

**TIENS2**

Mot-clé : **attention**

TIENS2 est un marqueur directif qui a pour but d'attirer l'attention sur quelque chose qui est à portée de sens. En (161), l'énonciateur H dit TIENS*2* en entendant la sirène d'une ambulance afin de faire remarquer cet événement à ses coéonciateurs :

(161) Mi : [1ah une ambulance (*elle fait un signe de la tête en direction de la fenêtre*) <pp<(inaud.) ambulance>>
M : encore [1l'ambulance qui passe eh mon doux
H : [1**tiens** (.) l'ambulance qui passe
*[CFPQ, sous-corpus 11, segment 4, page 48, ligne 18]*

**TIENS3**

Mots-clés : **bien**, **se sentir bien**

Selon la paraphrase proposée par Dostie (2004, p. 226), le marqueur TIENS3 sert à indiquer qu'un « résultat attendu et / ou désiré d'une action donnée est finalement atteint ». En (162),

l'énonciatrice J utilise TIENS3 afin d'indiquer que son déplacement était un geste délibéré et qu'il visait à améliorer la configuration des chaises :

(162) L : [1<p<je vais me tourner je vais avoir mal au COU (RIRE)>> (*en retournant sa chaise vers Julie et Françoise*)
J : (RIRE) **tiens** (*elle déplace sa chaise*)
*[CFPQ, sous-corpus 18, segment 7, page 69, ligne 13]*

Nous pouvons caractériser ce marqueur à l'aide de la paraphrase associée au mot-clé **se sentir bien**, qui décrit grosso-modo son rôle expressif.

**TIENS4**
Mot-clé : **inattendu**
Quasi-synonyme : ⌜REGARDE DONC⌝

Selon la définition de Dostie 2004, un énonciateur qui produit TIENS4 indique qu'il passe d'un état où il est surpris par quelque chose à un état où il accepte comme normal cette chose :

> **II.2** *Tiens* ≅
> Réagissant à une situation//
> j'indique que j'en prends conscience et que le fait que les choses soient ainsi ne s'oppose pas vraiment à mes attentes.
> (Dostie, 2004, p. 227)

Nous caractérisons ce MI à l'aide du mot-clé **inattendu**, bien que celui-ci soit insuffisant pour rendre complètement compte du sens du marqueur.

En (163), l'énonciateur M utilise TIENS4 afin de marquer le caractère inattendu du moment choisi par Jean pour bâtir une aréna.

(163) M : [...] ça faisait DES années qu'ils voulaient avoir l'aréna\ pis Jean leur répondait toujours •écoutez là (.) on n'a PAS d'argent pour ça (.) quand on aura de l'argent pour ça on fera on bâtira une aréna° <p<pis>> **tiens** tout à coup oups (.) là c'était le temps pis i- il a bâti un aréna
*[CFPQ, sous-corpus 11, segment 7, page 81, ligne 8]*

**TIENS5**

Mot-clé : **affirmatif**

TIENS5 est un marqueur assertif affirmatif avec lequel un énonciateur indique qu'il considère que quelque chose va de soi. En (164), l'énonciatrice A produit TIENS5 afin de signifier son accord avec l'énoncé de R.

(164) R: oui mais là ils doivent leur leur faire penser qu'ils devraient prendre leur carte de guichet pis d'aller au guichet [1automatique
A : [1ben **tiens**
*[CFPQ, sous-corpus 20, segment 5, page 47, ligne 17]*

Selon la paraphrase de Dostie pour cette lexie, ce MI est toujours utilisé en réponse à une question d'un coénonciateur (Dostie, 2004, p. 228). L'occurrence du marqueur présenté en (161) n'est pas produite en réponse à une question explicite, mais l'énonciateur répond affirmativement comme si une question avait été posée.

**2.72 TORIEU**

Signifiants : 4 (*torieu*)

MI : 4

TORIEU semble tirer son origine de l'expression « tord à dieu ». Il est utilisé à la manière d'un substitut de sacre.

**TORIEU1**

Mot-clé: **inattendu**

Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**TORIEU2**

Mots-clés: **mauvais**, **se sentir mal**

Quasi-synonymes : Les membres de la classe SUBSTITUT2.

**2.73 ⌜UNE CHANCE⌝**

Signifiants : 49

MI : 40

Mots-clés : **inattendu**, **bien**, **se sentir bien**

Quasi-synonymes : FIOU1, OUF1

Le phrasème ⌜UNE CHANCE⌝ est utilisé pour exprimer un soulagement devant un état de la situation qui est vu comme étant positif, alors qu'il aurait pu ne pas l'être.

En (165), l'énonciateur C exprime son soulagement suite au tour de parole de M dans lequel celui-ci infirme la proposition de J au sujet du fait qu'il n'aimerait pas les manèges :

(165) J : [3t'aimes pas les manèges/ (*en s'adressant à Maxime*)
M : oui j'aime ça
J : ah [1ok fiou
C : [1ah ok **une chance** (.) heille genre euh (.) il me dit que quand il é- il était à La Ronde là la Pitoune c'était extrême pour lui
*[CFPQ, sous-corpus 17, segment 3, page 37, ligne 23]*

Le phrasème peut également être utilisé comme connecteur textuel en lien syntaxique avec une proposition sous la forme « une chance que P » ou « un chance, P ».

**2.74 VIARGE**

Signifiants : 3 (*viarge*)
MI : 3

VIARGE parait être utilisé comme substitut au sacre non-prototypique, absent du CFPQ, VIERGE.

**VIARGE1**

Mot-clé: **inattendu**
Quasi-synonymes : Les membres de la classe SUBSTITUT1.

**VIARGE2**

Mots-clés: **mauvais**, **se sentir mal**
Quasi-synonymes : Les membres de la classe SUBSTITUT2.

**2.75 VOYONS**

Signifiants : 251 (*voyons*)
MI : 251

VOYONS est un MD à la fois très fréquent et très polysémique. Certains emplois de VOYONS ont une composante directive, comme c'est le cas pour les MI issus de formes impératives de verbes à la deuxième personne (ÉCOUTE et REGARDE).

**VOYONS1**

Mots-clés : **inattendu**, **hors du commun**

VOYONS semble le plus fréquemment utilisé afin d'exprimer l'étonnement. Dans l'extrait (166), l'énonciateur Y produit VOYONS1 en réaction à l'énoncé inattendu et hors du commun de son coénonciateur S :

(166) S: en revenant (*dit avec un petit rire*) on roulait cent soixante mon gars chacun leur bord ostie il tirait sur le voLANT (*en agitant les bras de haut en bas comme s'il tirait sur un volant*) [1 (RIRE)
Y: [1(*il lève le menton comme pour exprimer son étonnement*) **voyons** toi
*[CFPQ, sous-corpus 21, segment 1, page 1, lignes 13-14]*

La définition proposée dans Dostie 2004 pour cette lexie nous semble compatible avec notre analyse. La composante **inattendu** est exprimée par « cette situation est [...] opposée à mes attentes » et la composante **hors du commun** trouve son équivalence avec le mot « diamétralement », qui suggère un état de la situation hors de l'ordinaire :

**4a.** Voyons ≅
Réagissant à une situation //
j'indique que cette situation est diamétralement opposée à mes attentes.
(Dostie, 2004, p. 235)

**VOYONS2**

Mots-clés : **mauvais**, **attention**

Quasi-synonyme : ÉCOUTE2

Le deuxième sens de VOYONS en tant que MI est désapprobatif. En (167), l'énonciatrice exprime qu'elle juge négativement un élément de la situation et lance un appel à la réflexion à l'aide de VOYONS2 :

(167) VE : [1un entrevue de télé- un journaliste te téléphone t'a posé des questions [2sur quelque chose c'est quoi↑ (.) il a du temps à perdre genre↑ ça lui tente de jaser↑ (.) **voyons** donc (.) on a d'autres choses à faire que de jaser là/ (.) [3c'est quoi ça/
*[CFPQ, sous-corpus 19, segment 5, page 45, ligne 2]*

La définition de Dostie 2004 décrit avec beaucoup de précision l'acte expressif et l'acte directif associés à cette lexie :

> **1.** *Voyons* ≅
> Réagissant à des propos ou à un comportement //
> j'indique que ceux-ci m'apparaissent insensés ou déplacés, et je fais appel aux capacités cognitives de la personne à qui je parle afin qu'elle revienne à la façon appropriée de voir les choses ou de se comporter, conformément à ce qui est évident.
> (Dostie, 2004, p. 231)

**VOYONS3**

Mot-clé : **affirmatif**

VOYONS3 est utilisé afin de faire remarquer à quelqu'un que quelque chose est évident, ce qui est un cas particulier de l'acte affirmatif. En (168), l'énonciatrice C produit VOYONS3 afin d'attirer l'attention sur le caractère évident de ses blagues :

(168) C : ben non c'est des jokes **voyons** (RIRE)
*[CFPQ, sous-corpus 17, segment 2, page 19, ligne 8]*

**VOYONS4**

Mot-clé : **se sentir mal**

En utilisant VOYONS4, un énonciateur exprime qu'il ressent un sentiment négatif, plus précisément « une difficulté d'expression inattendue et injustifiée » qui l'embarrasse, comme décrit par la définition de Dostie pour cette lexie :

> **3.** T *voyons* ≅
> Étant contraint à interrompre ce que je suis en train de dire au moyen du texte T//
> j'indique que j'éprouve tout à coup une difficulté d'expression inattendue et injustifiée qui m'embarrasse.
> (Dostie, 2004, p. 234)

En (169), par exemple, l'énonciatrice J essaie de se remémorer ce qui a recommencé à la télévision et produit VOYONS4 pour exprimer sa frustration :

(169) J : ouais euh **voyons** ostine il y a une affaire qui était recommencée à la tv hier c'est quoi/
*[CFPQ, sous-corpus 17, segment 2, page 14, ligne 20]*

**2.76 VRAIMENT**

Signifiants : 1036 (*vraiment*)

MI : 77

Mot-clé : **affirmatif**

Quasi-synonymes : FRANCHEMENT1, ⌜JE COMPRENDS⌝, METS-EN, TELLEMENT

Nous avons décrit les différents sens du vocable VRAIMENT dans le cadre de notre mémoire de maîtrise (Lapointe, 2005). Lorsque VRAIMENT est un MI, comme en (170), il réalise le plus souvent un acte « affirmatif ».

(170) VI : [1c'est une profession qui est très mystérieuse pour Monsieur-Madame- [2Tout-le-Monde
VE : [2 (*elle hoche la tête affirmativement*)
MA : [2<p<ouais>>
VE : hum [1**vraiment** (.) [2c'est drôle
*[CFPQ, sous-corpus 19, segment 3, page 28, ligne 18-21]*

Nous avions relevé un autre sens possible de *vraiment* en tant que MI dans (Lapointe, 2005). Il s'agit d'un emploi désapprobatif auquel nous avions attribué la paraphrase suivante :

(Je signale que le comportement ou la série d'évènements *a* m'exaspère vraiment**4**.)
(Lapointe, 2005, p. 53)

Le sens de cet emploi est similaire à celui de ⌜POUR VRAI⌝2. Nous ne l'avons toutefois pas observé dans le CFPQ, ce qui laisse croire qu'il n'est pas fréquent dans la langue orale.

Dans Lapointe (2005), nous avions relevé un emploi de VRAIMENT prononcé avec une intonation montante interrogative, qui avait un sens similaire à celui de ⌜POUR VRAI⌝1. Avec le mot-phrase VRAIMENT à l'interrogatif, l'énonciateur demande au coénonciateur si ce qu'il dit est vrai et exprime ainsi sa surprise ou son scepticisme. Nous n'avons également pas observé cet emploi dans notre corpus.

### 2.77 ⌜VRAIMENT PAS⌝

Signifiants : 89 (*vraiment pas*)
MI : 11
Mot-clé : **infirmatif**
Quasi-synonymes : ⌜DU TOUT⌝, ⌜PAS DU TOUT⌝, PANTOUTE

Le phrasème ⌜VRAIMENT PAS⌝ utilisé comme MI est très similaire à PANTOUTE. En (171), l'énonciatrice MA utilise ⌜VRAIMENT PAS⌝ pour signaler son désaccord avec les propos de sa coénonciatrice VE :

(171) VE : [2<p<peut-être>> (.) peut-être que t'es une lesbienne refoulée/ (*en se retournant vers Marie-Ève*)
MA : <p<non: **vraiment pas**>> (*en hochant la tête négativement*)
*[CFPQ, sous-corpus 19, segment 8, page 80, ligne 4]*

**2.78 WÔ**

Signifiants : 38 (*wô*)

MI : 38

Mots-clés : **hors du commun**, **arrêt**

WÔ est associé à un acte directif, où un énonciateur demande à un coénonciateur d'arrêter de faire quelque chose, en réaction à quelque chose d'exagéré. Cette composante directive est toutefois souvent feinte ou adressée à un coénonciateur fictif.

En (172), l'énonciatrice É utilise WÔ en réaction au contenu hors du commun du discours de sa coénonciatrice. L'acte directif ne semble pas dirigé vers sa coénonciatrice I, mais plutôt vers le curriculum fictif trop compliqué du cours de scrapbooking dont il est question.

(172) É : [1il faut faire du bricolage t'sais (RIRE)
I : non parce que: euh: (.) parce que j'ai envie de m'amuser [1pis parce que: euh: je suis quelqu'un qui se déc- ben PAS qui se décourage facilement mais que (.) quand ça devient trop de taponna:ge j'haïs ça (.) t'sais là on fait [2on fait du t'sais on on découpe on met nos petits collants on n- m- [3 mais t'sais là éCOUte là c'était des affaires de GRAvures pis de: [4t'sais là c'étai:t là (*elle écarte les mains comme pour montrer que c'était exagéré et démesuré*) [5ben moi là/ commencer à: t'sais [6à:R : [1hum
É : [2ouin/
É : [3on écrit no:s
R : [4ouais (*elle hoche la tête affirmativement*)
É : [5<p<**wô wô wô** (inaud.)>> (RIRE)
*[CFPQ, sous-corpus 16, segment 9, page 78, ligne 15]*

WÔ est souvent répété dans un même énoncé, comme en (172).

**2.79 WOW**

Signifiants : 83 (wow)

MI : 83

Mots-clés : **inattendu**, **hors du commun**, **bien**, **se sentir bien**

Il semble qu'il nous suffisse de postuler l'existence d'une seule lexie pour rendre compte de toutes les utilisations de WOW. En (173), l'énonciatrice F utilise WOW pour exprimer son étonnement et son appréciation positive de la situation :

(173) J : [3ben à à Percé ben là [4(.) bon t'as (.) un: restaurant qui s'appelle Le Gargantua pis c'est en haut d'une montagne (*en levant les deux bras dans les airs comme pour montrer la hauteur de la montagne*) vraiment t- à pic là ben à pic et en arrière pis l-t'as une vue t'as l'impression d'être dans *The sound of music* quand t'es en haut de [5ça là […]
[...]
F : [5**wow** ça doit être magnifique ça
*[CFPQ, sous-corpus 18, segment 6, page 66, lignes 7-9]*

En nous inspirant de la définition suivante de *wow* donnée par Goddard (2013, p. 5), nous considérons que le signifié de cette lexie est composé de quatre composantes.

Wow!
I think like this: "this is very good"
I didn't know before that it can be like this
I feel something very good because of this
I feel like someone can feel when this someone sees something very big
(Goddard, 2013, p. 5)

Wow!
Je pense comme cela : « cela est très bien »
Je ne savais pas avant que cela pouvait être comme cela
Je ressens quelque chose de bien à cause de cela
Je me sens comme quelqu'un peut se sentir quand ce quelqu'un voit quelque chose de très grand
*[Traduction]*

Chaque ligne de la définition de Goddard correspond à une composante de notre système de description des MI. La ligne « Je pense comme cela : « cela est très bien » » correspond à l'acte illocutoire que nous identifions par le mot-clé **bien**. La deuxième ligne correspond grosso-modo au mot-clé **inattendu**, la troisième au mot-clé **se sentir bien** et la quatrième au mot-clé **hors du commun**. Notons que l'ordre des paraphrases proposées par Goddard diffère de celui que nous avons établi au tableau 19.

Le marqueur WOW semble particulièrement approprié pour une usage ironique, souvent en réaction à quelque chose qui n'est pas impressionnant, comme en (174).

(174) J : c’est aussi [1bon qu’un (.) c’est aussi bon qu’un tape de de de de (.) des années soixante et dix (*dit en riant*)
M : [1<f<**wow**>>
*[CFPQ, sous-corpus 6, segment 4, page 59, lignes 16-17]*

**2.80 YOUPI**

Signifiants : 2
MI : 2
Mots-clés : **forte émotion**, **bien**, **se sentir bien**
Quasi-synonyme : YÉ

YOUPI est similaire à SUPER et COOL, mais une composante supplémentaire liée à l'expression d'une forte émotion semble de plus lui être associée. Dans l'extrait (175), l'énonciateur JN produit YOUPI pour exprimer, ou feindre d'exprimer, son grand enthousiasme devant l'idée de déballer un cadeau avec des mitaines et des dents.

(175) JN : on va te mettre une contrainte/ (.) sans les mains:/ [1(RIRE)
S : [1(RIRE)[2(RIRE)
J : [2avec des mitai:nes (*dit d'une voix enjouée*)
S : wou[1hou:
JN : [1avec des mitaines pis tes dents:\ (RIRE) **youpi/**
*[CFPQ, sous-corpus 28, segment 3, page 43, lignes 2-6]*

La substitution de YOUPI par COOL dans l'énoncé (175) ne peut se faire sans modification de sens. YOUPI semble inévitablement lié à l'expression d'un fort enthousiasme, au contraire de COOL.

**2.81 YÉ**

Signifiants : 22 (10 *yé*, 12 *yeah*)
MI : 22
Mots-clés : **forte émotion**, **bien**, **se sentir bien**,
Quasi-synonyme : YOUPI

Le marqueur YÉ est parfois prononcé à l'anglaise dans le CFPQ et souvent transcrit sous la forme *yeah*. Nous soupçonnons que cette variation est liée à une variation sémantique, mais nous n'avons pas investigué le phénomène.

L'extrait (176) offre un exemple typique d'utilisation de YÉ dans un discours rapporté, où l'énonciatrice J produit le MI pour signifier son enthousiasme.

(176) J : [1•yé on va prendre le bateau°
*[CFPQ, sous-corpus 17, segment 5, page 58, ligne 3]*

**2.82 ZUT**

Signifiants : 2 (*zut*)
MI : 2
Mots-clés : **mauvais**, **se sentir mal**

ZUT est utilisé par un énonciateur pour manifester une légère déception. L'unité réalise en quelque sorte des actes illocutoires inverses de ceux réalisés par COOL et SUPER.

En (177), l'énonciatrice ME utilise ZUT pour exprimer son sentiment négatif au sujet du fait que ses amies et elle n'aient pas de petit ami.

(177) M : avez-vous des chums vous-autres↑
ME : [1non (RIRE)
D : [1non (RIRE)
C : [1non (RIRE)
MA : [1non (RIRE)
ME : **zut** alors non
*[CFPQ, sous-corpus 3, segment 3, page 42, ligne 11]*

# 3. Unités non traitées

D'autres candidats au statut de MI que ceux que nous venons de présenter se retrouvent dans le CFPQ, mais certains phénomènes ont fait en sorte que nous ne les avons pas inclus dans la liste des unités que nous étudions (voir chapitre 1-6.1).

## 3.1 Unités peu fréquentes

Nous n'avons pas tenu compte des unités qui n'apparaissaient qu'une seule fois dans le CFPQ, en excluant les discours rapportés. L'unité PUTAIN, par exemple, est utilisée trois fois comme MI

dans le corpus, mais deux de ces occurrences concernent des discours rapportés de locuteurs européens.

Parmi les MI trop peu fréquents pour nous permettre de tirer des conclusions claires à leur sujet, on trouve PUTAIN, DÉGEULASSE, ⌜BONNE CHANCE⌝, ⌜GO GO GO⌝ et BÂTARD.

### 3.2 Connecteurs textuels

Plusieurs connecteurs textuels ont des utilisations comme MI, mais leur prise en compte aurait nécessité un travail de désambiguïsation manuelle beaucoup plus important. De plus, nous soupçonnons que l'identification automatique de ces unités demanderait des systèmes informatiques aux propriétés très différentes que ceux destinés à repérer les MI dont les signifiants ne peuvent pas être utilisés comme connecteurs textuels.

Les unités METTONS, ⌜BIEN SÛR⌝, ⌜C'EST SÛR⌝ et ⌜C'EST CLAIR⌝ ont été exclues de notre étude pour cette raison.

### 3.3 Unités discursives très fréquentes et polycatégorielles

Plusieurs unités discursives très fréquentes remplissent des rôles très variés en conversation et sont parfois utilisées comme MI. Une étude des unités OUI et NON, par exemple, aurait nécessité une thèse à elle seule. Les unités OH, HUM, OK, OUI et NON ont ainsi été exclues de notre étude en raison de la trop grande complexité de leurs usages. Les unités que nous avons choisi d'étudier ont un statut de MI plus clair, ce qui a permis de simplifier considérablement leur analyse.

### 3.4 Salutations

Nous avons décidé de ne pas tenir compte des salutations comme ALLO, HELLO, SALUT et BONJOUR parce ces unités nous semblent appartenir à une classe à part de marqueurs discursifs.

## 4. Conclusion au sujet de la caractérisation sémantique des MI

L'élaboration d'un système modulaire de caractérisation sémantique des MI nous a permis de mettre en lumière quelques phénomènes intéressants. Nous avons vu qu'un grand nombre de MI partagent des composantes sémantiques et que plusieurs entretiennent des relations de synonymie. En contre partie, plusieurs MI ont des particularités sémantiques que notre système de description ne peut pas rendre compte à cause de son manque de finesse.

En observant le glossaire présenté au point 2 de ce chapitre, il est, entre autres, intéressant de constater que :

- d'avantage de marqueurs sont utilisés pour indiquer qu'une chose est mauvaise que pour indiquer qu'une chose est bien;
- il existe une grande quantité de MI dont la fonction est de signaler qu'une chose est inattendue ou hors du commun et relativement peu qui expriment la douleur et le dégoût;
- parmi les MI assertifs, la plupart sont affirmatifs;
- la demande d'attention est l'acte illocutoire le plus fréquemment réalisé parmi les MI directifs.

# Chapitre 5 :
# Application du système d'analyse des MI

Munis d'un moyen d'identifier automatiquement les MI dans les transcriptions de textes oraux et d'un moyen de caractériser sémantiquement ces MI, nous pouvons maintenant procéder à une courte démonstration de l'application de notre système.

L'objectif de ce chapitre n'est pas de mesurer de nouveau l'efficacité du système, mais plutôt de mettre en lumière sa pertinence comme outil d'analyse.

## 1 Textes cibles

Les sous-corpus 10 et 21 du CFPQ nous serviront de textes cibles pour la démonstration. Ces textes ont été choisis arbitrairement, dans le but de mettre en lumière certaines variations dans la production des MI.

Le sous-corpus 10 est la transcription d'une conversation entre 4 locuteurs de 24 et 25 ans (2 hommes et 2 femmes) qui se connaissent bien et qui fréquentent la même université. Cette conversation est constituée de 3208 tours de paroles où sont répartis 36 227 tokens. Rappelons que le terme *token* désigne principalement des mots, mais aussi d'autres types de symboles, comme des marques phonologiques, des pauses et des marques de rire.

Le sous-corpus 21 est la transcription d'une conversation entre 3 locuteurs âgés de 24 à 27 ans (3 hommes) qui se connaissent bien. Cette conversation est constituée de 2483 tours de paroles (26 861 tokens).

## 2 Procédé

Le système d'analyse proposé prend comme textes cibles des fichiers TXT tirés des fichiers PDF disponibles sur le site web du CFPQ. L'extrait (178) illustre le format du texte reçu par le système d'analyse.

(178) V : [1 j'arrête pas de dire à Jean-Marc que Caro ça a l'air d'un nom de chien ¤<432243>
J-M : [1 (RIRE) (inaud.) <P66,L24>
J-M : ben voyons donc (RIRE) ¤<432840> <P66,L25>
*[CFPQ, sous-corpus 10, segment 5, page 66, lignes 23-25]*

Une opération de nettoyage automatique permet d'enlever les éléments non linguistiques du texte et de remplacer certains caractères difficiles à traiter informatiquement. L'extrait (179) illustre l'état du texte une fois nettoyé.

(179) v j_ arrête pas de dire à jean-marc que caro ça a l_ air d' un nom de chien
j-m _rire_ inaud
j-m ben voyons donc _rire_

Notons que toute l'information qui concerne les chevauchements de tours de parole est perdue au cours du nettoyage.

Chaque tour de parole est analysé par le système qui identifie son énonciateur (grâce aux caractères au début de la ligne). Le tour de parole est ensuite étiqueté par un étiqueteur à n-grammes qui est entraîné sur l'ensemble du CFPQ selon la méthode présentée au chapitre 3 (point 2.1). L'extrait (180) illustre l'état du texte une fois étiqueté.

(180) [('j_', 'PRO'), ('arrête', 'S'), ('pas', 'S'), ('de', 'S'), ('dire', 'S'), ('à', 'A'), ('jean-marc', 'S'), ('que', 'QUE'), ('caro', 'S'), ('ça', 'S'), ('a', 'S'), ('l_', 'DET'), ('air', 'S'), ('d', 'S'), ('"', 'S'), ('un', 'DET'), ('nom', 'S'), ('de', 'S'), ('chien', 'S')]
[('_rire_', 'M'), ('inaud', 'S')]
[('ben', 'BEN'), ('voyons', 'M'), ('donc', 'DONC'), ('_rire_', 'M')]

Un classifieur SVM, entraîné sur l'ensemble des MI du CFPQ, selon les paramètres discutés au chapitre 3 (point 2.3), analyse chacun des signifiants qui sont potentiellement des MI dans le texte étiqueté. Les MI ainsi repérés sont compilés.

Plusieurs opérations computationnelles permettent de générer différents types d'informations qui concernent la présence des MI dans le texte. La discussion entreprise à la section suivante s'appuie sur ces informations.

# 3 Résultats

L'objet de cette section est d'illustrer les types d'informations qu'un système d'analyse automatique des MI peut nous fournir au sujet des conversations et de leurs énonciateurs. Dans les chapitres précédents, nous avons amplement eu l'occasion d'analyser individuellement des occurrences particulières de MI. Nous souhaitons maintenant dépasser les contextes de la phrase et du tour de parole pour envisager le phénomène d'un point de vue plus large. Nous procéderons ainsi à l'analyse et à la comparaison de l'usage des MI par chacun des énonciateurs dans le cadre de leurs conversations respectives.

## 3.1 Énonciateurs du sous-corpus 10

Le tableau 21 présente le nombre de MI produits par chaque énonciateur qui intervient dans le sous-corpus 10. La division du nombre de MI produits par un énonciateur par le nombre de tokens qu'il produit nous donne un ratio qui permet d'évaluer la propension pour cet énonciateur à produire des MI dans une conversation donnée.

*Tableau 21 : MI par 1000 tokens des énonciateurs du sous-corpus 10*

| **Énonciateur** | **MI** | **Tokens** | **MI par 1000 tokens** |
|---|---|---|---|
| Daniel | 39 | 2886 | 13,5 |
| Jean-Marc | 126 | 10602 | 11,9 |
| Michèle | 97 | 12578 | 7,7 |
| Valérie | 41 | 10161 | 4,0 |

Nous constatons que, malgré le fait qu'il soit le moins loquace, Daniel est l'énonciateur qui utilise le plus grand nombre de MI par 1000 tokens.

Il est également intéressant d'observer que les deux énonciateurs masculins ont tendance à produire plus de MI par tokens que les leurs partenaires féminines.

Le tableau 22 permet de comparer la longueur moyenne des tours de parole de chaque énonciateur.

*Tableau 22 : Longueur moyenne des tours de parole des énonciateurs du sous-corpus 10*

| **Énonciateur** | **Tokens** | **Tours de paroles** | **Longueur moyenne (tokens / tours de parole)** |
|---|---|---|---|
| Michèle | 12578 | 984 | 12,78 |
| Valérie | 10161 | 811 | 12,53 |
| Jean-Marc | 10602 | 930 | 11,4 |
| Daniel | 2886 | 466 | 6,19 |

Valérie, Jean-Marc et Michèle sont responsables d'un nombre à peu près égal de tours de parole, tandis que Daniel prend la parole à peu près deux fois moins que les autres. Nous remarquons que l'énonciateur qui parle le moins est celui qui a le plus tendance à produire des tours de parole

courts tandis que ceux qui parlent davantage ont tendance à s'exprimer dans des tours de parole plus longs.

### 3.1.1 Daniel

Daniel est l'énonciateur qui parle le moins au cours de la conversation et qui produit les tours de parole les plus courts. Il est cependant celui qui utilise le plus de MI par 1000 tokens.

Cet énonciateur ne produit aucun MI sacre ou substitut de sacre. En guise de MI expressifs, il utilise plutôt de façon prépondérante le MI OUPELAILLE qui lui est propre. Daniel produit en effet 100% des 12 occurrences de ce vocable dans le sous-corpus 10.

Parmi les MI produits par Daniel, très peu sont directifs (par exemple, VOYONS, REGARDE et ÉCOUTE sont absents).

*Tableau 23 : MI produits par Daniel dans le sous-corpus 10*

| **Vocable** | **Fréquence** |
|---|---|
| OUPELAILLE | 12 |
| ⌜MON DIEU⌝ | 5 |
| ⌜POUR VRAI⌝ | 4 |
| HEIN, HEILLE | 3 |
| AYOYE | 2 |
| ENVOYE, MALADE, SUPER, ARRÊTEZ, ⌜PAS VRAIMENT⌝, VRAIMENT, ⌜JE COMPRENDS⌝, WOW, ⌜PAS DU TOUT⌝, ARRÊTE | 1 |

### 3.1.2 Michèle

L'utilisation de ⌜MON DIEU⌝ et de quelques sacres comme MI expressifs est caractéristique du discours de Michèle.

Nous y remarquons également la présence des MI directifs (et potentiellement désapprobatifs) VOYONS, REGARDE et ÉCOUTE.

*Tableau 24 : MI produits par Michèle dans le sous-corpus 10*

| **Vocable** | **Fréquence** |
|---|---|
| HEIN | 15 |
| HEILLE | 13 |
| ⌜MON DIEU⌝ | 10 |
| VOYONS, OUF | 6 |
| METS-EN, REGARDE, VRAIMENT, WOW | 4 |
| AYOYE, CIBOIRE, CALVAIRE | 3 |
| ⌜EH BOY⌝, OSTIE, ARK, CRISSE, CÂLINE, ÉCOUTE | 2 |
| ENVOYE, ⌜VRAIMENT PAS⌝, MAUDIT, TABARNAQUE, SÉRIEUX, ⌜POUR VRAI⌝, WÔ, ⌜PAS VRAIMENT⌝, ⌜PAS DU TOUT⌝, OUPS | 1 |

### 3.1.3 Valérie

L'énonciatrice Valérie est celle qui produit le moins de MI par tokens au cours de la conversation. Elle n'utilise aucun MI sacre ou substitut de sacre et produit davantage de MI affirmatifs (⌜JE COMPRENDS⌝, ⌜POUR VRAI⌝, TELLEMENT et VRAIMENT) que de MI expressifs ou directifs.

*Tableau 25 : MI produits par Valérie dans le sous-corpus 10*

| **Vocable** | **Fréquence** |
|---|---|
| HEIN | 8 |
| HEILLE | 7 |
| OUF | 5 |
| ⌜JE COMPRENDS⌝, ⌜POUR VRAI⌝, TELLEMENT | 3 |
| VRAIMENT | 2 |
| ⌜VRAIMENT PAS⌝, VOYONS, AYOYE, REGARDE, SUPER, GO, WOW, ⌜PAS DU TOUT⌝, CHUT, OUPS | 1 |

### 3.1.4 Jean-Marc

L'énonciateur Jean-Marc est celui qui produit le plus de MI. La présence de MI qui expriment potentiellement un jugement négatif comme SÉRIEUX et les directifs VOYONS, ÉCOUTE et REGARDE, est caractéristique de son discours.

Jean-Marc produit un assez grand nombre de substituts de sacres (comme TABARNOUCHE, CIBOLE et CÂLINE) en guise de MI expressifs.

*Tableau 26 : MI produits par Jean-Marc dans le sous-corpus 10*

| **Vocable** | **Fréquence** |
|---|---|
| HEIN | 50 |
| HEILLE | 12 |
| SÉRIEUX | 9 |
| VOYONS, ÉCOUTE, REGARDE | 5 |
| AYOYE | 4 |
| CIBOLE, TABARNOUCHE, FRANCHEMENT, SACRE, YÉ, TIENS | 3 |
| OUF, CRISSE, METS-EN, CÂLINE | 2 |
| COUDON, TABARNAQUE, MALADE, OSTIE, PANTOUTE, ARK, VRAIMENT, MAUTADIT, MON DIEU, OUPS | 1 |

### 3.1.5 Conclusions au sujet des énonciateurs du sous-corpus 10

Suite à l'analyse des données que nous venons de présenter, nous pouvons émettre quelques hypothèses au sujet de l'attitude des différents énonciateurs du sous-corpus 10.

Rappelons que Michèle, Valérie et Jean-Marc sont plutôt loquaces au cours de la conversation, tandis que Daniel parle environ deux fois moins et produits des tours de parole en moyenne deux fois plus courts que les autres.

Daniel est celui qui produit le plus haut ratio de MI par tokens. Il utilise avec régularité un MI expressif qui lui est propre (OUPELAILLE) et n'utilise pas de sacres ni de MI directifs. Nous postulons qu'il a tendance à adopter une posture « non-confrontationnelle » et polie au cours de la conversation, qu'il parle peu, écoute avec intérêt le discours de ses coénonciataurs et y réagit périodiquement à l'aide de phrases courtes.

Valérie produit très peu de MI expressifs et de MI directifs. Elle produit par contre plusieurs MI affirmatifs. Ceci laisse croire qu'elle adopte une posture réservée et non-confrontationelle au cours de la conversation.

Jean-Marc utilise avec prépondérance des MI qui sont potentiellement désapprobatifs, incluant les MI directifs. Il choisit également d'utiliser des substituts de sacres. Nous postulons que Jean-Marc adopte une posture expressive avec des épisodes confrontationnels.

Michèle est celle qui parle le plus. Elle produit un grand nombre de MI expressifs, dont plusieurs sacres, ainsi que des MI directifs. Une attitude loquace et expressive caractérise vraisemblablement son discours.

### 3.2 Énonciateurs du sous-corpus 21

Le tableau 27 met en lumière une similitude entre les sous-corpus 10 et 21 au sujet de la relation qui existe entre la loquacité d'un énonciateur et le nombre de MI par tokens qu'il produit.

*Tableau 27 : MI par 1000 tokens des énonciateurs du sous-corpus 21*

| **Énonciateur** | **MI** | **Tokens** | **MI par 1000 tokens** |
|---|---|---|---|
| Sylvain | 98 | 3779 | 25,9 |
| Yan | 257 | 12642 | 20,3 |
| Oscar | 123 | 10440 | 11,8 |

Sylvain est celui qui parle le moins et celui qui produit le plus de MI par tokens dans le sous-corpus 21, comme Daniel dans le sous-corpus 10.

Le tableau 28 permet de comparer la loquacité des différents énonciateurs avec la longueur moyenne de leurs tours de parole.

*Tableau 28 : Longueur moyenne des tours de parole des énonciateurs du sous-corpus 21*

| **Énonciateur** | **Tours de paroles** | **Tokens** | **Longueur moyenne (tokens / tours de parole)** |
|---|---|---|---|
| Yan | 1089 | 12642 | 11 |
| Oscar | 902 | 10440 | 8,64 |
| Sylvain | 491 | 3779 | 7,7 |

Comme c'était le cas pour le sous-corpus 10, l'énonciateur qui parle le moins (Sylvain) a tendance à produire des tours de parole plus courts et celui qui parle le plus (Yan) a tendance à user de tours de parole plus longs.

### 3.2.1. Sylvain

Comme Daniel du sous-corpus 10, Sylvain est à la fois l'énonciateur qui parle le moins au cours de la conversation, qui produit les tours de parole les plus courts et qui utilise le plus de MI par tokens.

Comme Daniel également, Sylvain n'utilise presque pas de MI directif et beaucoup de MI expressifs, mais il fait appel à des sacres plutôt qu'à OUPELAILLE.

Par ailleurs, contrairement à Daniel, Sylvain utilise le vocable HEILLE beaucoup plus souvent que HEIN, probablement parce qu'il éprouve d'avantage le besoin d'attirer l'attention de ses coénonciateurs.

*Tableau 29 : MI produits par Sylvain dans le sous-corpus 21*

| **Vocable** | **Fréquence** |
|---|---|
| OSTIE | 30 |
| HEILLE | 28 |
| CRISSE | 9 |
| TABARNAQUE | 7 |
| HEIN | 6 |
| CÂLISSE, OUF | 3 |
| ENVOYE, VOYONS, AYOYE, CIBOLE, SÉRIEUX, WÔ, CIBOIRE, FIOU, CÂLIQUE, VRAIMENT, BAPTÊME, SEIGNEUR | 1 |

### 3.2.2 Yan

Une utilisation très prolifique des sacres et des MI directifs sont les deux caractéristiques principales du discours de Yan. Très loquace, Yan produit plus de la moitié des sacres de la conversation, ainsi que 23 des 37 occurrences de REGARDE et 12 des 14 occurrences de VOYONS.

Yan utilise les MI d'une manière similaire à Jean-Marc du sous-corpus 10, à la différence que ce dernier a recours à des substituts de sacres plutôt qu'à des sacres.

*Tableau 30 : MI produits par Yan dans le sous-corpus 21*

| **Vocable** | **Fréquence** |
|---|---|
| OSTIE | 68 |
| CRISSE | 55 |
| HEILLE | 47 |
| REGARDE | 23 |
| HEIN | 16 |
| TABARNAQUE | 14 |
| VOYONS | 12 |
| CALVAIRE | 6 |
| ENVOYE | 3 |
| CIBOIRE, CÂLISSE, CRIF | 2 |
| CIBOLE, CÂLIQUE, PANTOUTE, ⌜UNE CHANCE⌝, ⌜PAS VRAIMENT⌝, ÉCOUTE, SEIGNEUR | 1 |

### 3.2.3 Oscar

Oscar est l'énonciateur qui produit le moins de MI par tokens au cours de la conversation. Il utilise quelques sacres en guise d'unités expressives et REGARDE en guise d'unité directive.

*Tableau 31 : MI produits par Oscar dans le sous-corpus 21*

| **Vocable** | **Fréquence** |
|---|---|
| HEILLE | 52 |
| REGARDE | 14 |
| HEIN, OSTIE | 12 |
| CRISSE | 9 |
| OSTIFIE | 7 |
| CÂLIF, CÂLIQUE | 3 |
| OUF, CRIF | 2 |
| ENVOYE, TABARNAQUE, ⌜DE LA MARDE⌝ , VOYONS, CIBOIRE, CÂLISSE, TABARNACHE | 1 |

### 3.2.4 Conclusions au sujet des énonciateurs du sous-corpus 21

Suite à l'analyse des données qui concernent les énonciateurs du sous-corpus 21, nous pouvons émettre quelques hypothèses au sujet de leurs attitudes dans la conversation.

Rappelons que Yan est celui qui parle le plus et fait les plus longs tours de parole, que Oscar parle un peu moins que Yan et que Sylvain parle presque trois fois moins que Yan.

Sylvain est peu loquace et n'utilise presque pas de MI directifs. Il semble qu'il ait une attitude réservée au cours de la conversation, réagit aux propos de ses coénonciateurs par des tours de parole brefs et adopte une posture non-confrontationnelle.

Yan est très loquace, il utilise un grand nombre d'unités expressives et directives (potentiellement désaprobatives). Il semble adopter une posture expressive qui n'évite pas la confrontation, de façon similaire à Michèle et à Jean-Marc du sous-corpus 10.

Oscar parle presque autant que Yan et produit à peu près les mêmes MI que Yan, mais à une fréquence beaucoup moins élevée. Nous émettons l'hypothèse que Oscar adopte une posture peu expressive au cours de la conversation.

### 3.3 Comparaison des conversations

Comme nous le voyons au tableau 32, notre système d'analyse a repéré 303 MI dans le sous-corpus 10 (ce qui équivaut à 8,4 MI par 1000 tokens) et 478 MI dans le sous-corpus 21 (17,8 MI par 1000 tokens). Les énonciateurs du sous-corpus 21 utilisent donc deux fois plus de MI que ceux du sous-corpus 10. Le caractère davantage familier de la conversation du sous-corpus 21 est peut-être le principal facteur derrière ces ratios très différents. Nous postulons que, en général, plus une conversation est familière, plus elle sera parsemée de MI.

*Tableau 32 : MI par 1000 tokens des deux sous-corpus*

| **Conversation** | **MI** | **Tokens** | **MI par 1000 tokens** |
|---|---|---|---|
| Sous-corpus 10 | 303 | 36227 | 8,4 |
| Sous-corpus 21 | 478 | 26861 | 17,8 |

Les tableaux 33 et 34 présentent les MI qui ont été repérés par le système dans les deux textes cibles.

*Tableau 33 : MI du sous-corpus 10*

| **Vocables** | **Fréquence** |
|---|---|
| HEIN | 76 |
| HEILLE | 35 |
| MON DIEU | 16 |
| OUF | 13 |
| OUPELAILLE, VOYONS | 12 |
| AYOYE, REGARDE, SÉRIEUX | 10 |
| VRAIMENT, POUR VRAI | 8 |
| ÉCOUTE | 7 |
| WOW, METS-EN | 6 |
| CRISSE, JE COMPRENDS, CÂLINE, CRISSE | 4 |
| FRANCHEMENT, OUPS, TABARNOUCHE, CALVAIRE, YÉ, ⌜PAS DU TOUT⌝, TIENS, TELLEMENT, CIBOLE, SACRE, ARK, OSTIE, CIBOIRE | 3 |
| SUPER, PAS VRAIMENT, EH_BOY, ENVOYE, VRAIMENT PAS, TABARNAQUE, MALADE, ARRÊTE | 2 |
| CHUT, GO, MAUTADIT, MAUDIT, COUDON, WÔ, PANTOUTE | 1 |

*Tableau 34 : MI du sous-corpus 21*

| **Vocables** | **Fréquence** |
|---|---|
| HEILLE | 127 |
| OSTIE | 110 |
| CRISSE | 73 |
| REGARDE | 37 |
| HEIN | 34 |
| TABARNAQUE | 22 |
| VOYONS | 14 |
| OSTIFIE | 7 |
| CÂLISSE, CALVAIRE | 6 |
| ENVOYE, CÂLIQUE, OUF | 5 |
| CIBOIRE, CRIF | 4 |
| CÂLIF | 3 |
| CIBOLE, SEIGNEUR | 2 |
| ⌜DE LA MARDE⌝ , AYOYE, SÉRIEUX, WÔ, FIOU, PANTOUTE, TABARNACHE, ⌜UNE CHANCE⌝ , ⌜PAS VRAIMENT⌝ , VRAIMENT, ÉCOUTE, BAPTÊME | 1 |

Notons d'abord que les vocables HEIN, HEILLE, qui sont employés par tous les énonciateurs, sont utilisés dans des proportions inversées dans les deux conversations étudiées. Prédominant dans le sous-corpus 10, HEIN est le plus souvent utilisé par des énonciateurs qui sont à la recherche d'une confirmation de leurs dires. HEILLE au contraire sert le plus souvent à attirer l'attention d'un coénonciateur et sa prédominance dans le sous-corpus 21 pourrait être une indication d'une tendance à une plus grande compétitivité quant à « l'accès au micro ».

Nous remarquons que les sacres et les substituts de sacres se trouvent en quantité importante dans le sous-corpus 21, tandis que le système a détecté relativement peu de sacres dans le sous-corpus 10. En contrepartie, les vocables expressifs ⌜MON DIEU⌝, OUPELAILLE, OUF, AYOYE et WOW, ainsi que plusieurs substituts de sacres, sont présents dans le sous-corpus 10, mais

presque absents du sous-corpus 21. Ces différents types de vocables expressifs semblent jouer des rôles similaires.

Les vocables directifs VOYONS, REGARDE (et dans une moindre mesure ÉCOUTE) sont communs aux deux conversations, mais répartis inégalement parmi les énonciateurs. En général, ces MI semblent être caractéristiques des énonciateurs loquaces qui expriment facilement leur désapprobation.

Les vocables affirmatifs VRAIMENT, ⌜JE COMPRENDS⌝, TELLEMENT et ⌜POUR VRAI⌝ sont caractéristiques du sous-corpus 10, ce qui laisse croire que cette conversation est davantage consensuelle ou qu'elle met en jeu des participants qui sont plus polis que celle du sous-corpus 20.

## 4 Conclusion au sujet de l'application du système d'analyse des MI

Suite à l'examen des résultats du système d'analyse automatique des MI que nous avons développé et appliqué sur deux sous-corpus du CFPQ, il semble opportun de synthétiser en quelques remarques nos conclusions.

- Le fait que les locuteurs qui parlent le moins sont ceux qui utilisent le plus de MI par tokens s'explique par leur tendance à réagir aux paroles de leurs coénonciateurs par des tours de parole courts ponctués par des MI, notamment pour signifier leur intérêt.

- Les cinq énonciateurs masculins observés produisent chacun plus de MI par tokens que les deux énonciatrices observées. Il serait intéressant d’investiguer l'influence du genre sur la production des MI en général.

- Tous les énonciateurs utilisent les vocables HEIN et HEILLE. Nous postulons qu'une plus grande proportion de HEILLE par rapport à HEIN dans une conversation est indicateur d'une plus grande compétitivité dans la prise de parole.

- Tous les énonciateurs utilisent des MI expressifs, que ce soit des vocables neutres (OUPELAILLE, AYOYE...), des sacres ou des substituts de sacres. Les vocables de chacune de ces catégories semblent cependant être quasi-exclusifs : les énonciateurs ont tendance à ne pas interchanger les catégories et à ne privilégier que les unités appartenant à une seule d'entre elles.

- Les membres du trio directifs-désapprobatifs VOYONS, REGARDE et ÉCOUTE semblent être solidaires dans leur usage. Un énonciateur qui produit un vocable directif a aussi tendance à produire les autres membres du groupe.

- La présence de MI affirmatifs (VRAIMENT, ⌜POUR VRAI⌝, ⌜JE COMPRENDS⌝...) dans une conversation est vraisemblablement un indicateur d'un certain accord entre les énonciateurs.

Il serait pertinent de pousser l'analyse des MI en tenant compte de la chronologie des énoncés, de l'ordre dans lequel les MI sont produits et de leurs regroupements dans le discours. Jumelée à la prise en compte des unités qui ne sont pas des MI, une telle analyse permettrait de faire des liens entre le rôle des MI et le contenu sémantique des textes.

# CONCLUSION

Nous avons entrepris la rédaction de cette thèse dans l'espoir d'améliorer les connaissances générales au sujet des marqueurs illocutoires et de faciliter leur prise en compte par les systèmes de traitement automatique de la langue.

La revue de littérature au sujet des MI que nous avons présentée au chapitre 2 nous a permis de situer la classe des MI parmi les autres classes grammaticales. Nous avons vu que ces unités sont des marqueurs discursifs qui réalisent des actes illocutoires au premier plan d'une conversation. Une revue de certaines études qui traitent du traitement automatique des MD nous a ensuite permis de relever plusieurs caractéristiques pertinentes de cette classe pour leur identification automatique.

Ce double état de la question nous a servi de point de départ pour mettre au point une expérience sur l'identification automatique des MI que nous avons présentée au chapitre 3. Nous y avons testé quatre méthodes d'identification automatique des MI. L'utilisation conjointe d'un étiqueteur à n-grammes et d'un classifieur SVM a permis à la meilleure de ces méthodes d'identifier les MI ambigus d'un corpus test avec une f-mesure de presque 94%.

Nous avons ensuite présenté un système modulaire de caractérisation sémantique des MI au chapitre 4. Nous avons constaté que les MI réalisent des actes expressifs, assertifs et directifs et que certains d'entre eux sont associés à une connotation stylistique négative. Nous avons démontré que la combinaison de 17 paraphrases simples permet une caractérisation sémantique grossière de ces MI.

Le chapitre 5 a été l'occasion de donner un aperçu de la pertinence de la prise en compte des MI dans les systèmes d'analyse de textes, particulièrement en ce qui a trait aux attitudes des énonciateurs et à leurs rôles dans les conversations.

# 1 Points forts de la thèse

Plusieurs leçons ont été apprises au cours de la rédaction de cette thèse. Nous en présentons quelques-unes ici.

- Les unités lexicales et les éléments prosodiques qui indiquent des ruptures syntaxiques, ainsi que les prépositions qui indiquent des liens syntaxiques se sont révélés particulièrement pertinentes à exploiter pour un système informatique chargé de repérer les MI.

- Le relatif succès de notre étiqueteur à n-grammes démontre la pertinence d'utiliser un système d'étiquettes réduit pour l'identification des unités extraphrastiques, à tout le moins dans un contexte où le corpus utilisé est similaire au nôtre quant à sa taille.

- Nous avons constaté qu'un système d'étiquettes réduit permet une annotation semi-automatique rapide des corpus d'entraînement et représente ainsi une économie énorme de ressources humaines.

- Le relatif succès de l'étiqueteur Brill et du classifieur SVM a permis de démontrer la pertinence de prendre en compte le contexte à droite des MI pour leur identification automatique.

- Le système de caractérisation des signifiés de MI présentée au chapitre 4 a mis en lumière les relations lexicales et les parentés sémantiques entre plusieurs MI.

- L'analyse individuelle des MI a mis en lumière la grande variété dans l'utilisation de ceux-ci ainsi que les limites d'un système de caractérisation sémantique qui ne se base que sur la combinaison de 17 paraphrases simples.

- L'observation des MI produits par chacun des énonciateurs de deux conversations a fait ressortir la pertinence de ces unités dans l'analyse de conversations à bâtons rompus, particulièrement en ce qui a trait aux attitudes des locuteurs et à leurs postures conversationnelles.

## 2 Questions non abordées

Plusieurs sujets d'investigation liés aux MI auraient pu être abordés dans le cadre de cette thèse. Mentionnons le problème de l'anti-phrase, de l'ironie et du sarcasme; l'influence des discours rapportés sur la phonologie; la dimension gestuelle; la prise en compte des niveaux d'informalité et de politesse... Mais la question non abordée la plus importante nous semble être le problème de la désambiguïsation automatique des MI qui sont polysémiques.

Certaines caractéristiques qui permettent à un humain de déterminer le sens d'un MI polysémique sont difficilement accessibles à un ordinateur. Le contexte de conversation, le contexte lexical large et les particularités dans l'énonciation des marqueurs (prosodie et gestuelle) nous apparaissent comme des informations particulièrement difficiles à fournir à un système informatique.

Une partie de ce problème pourrait être réglé par la ré-annotation des bandes audiovisuelles en utilisant un système de description de la prosodie plus détaillé (voir Beckman, Hirschberg et Shattuck-Hufnagel, 2004; et Post, 2006, par exemple), possiblement de manière automatique ou semi-automatique (voir Rosenberg, 2009 par exemple).

Deux obstacles risqueraient cependant de faire échouer une telle tâche. Premièrement, puisque le CFPQ est constitué de conversations, ses enregistrements souffrent du problème des chevauchements des tours de parole. Souvent, les MI sont produits dans un contexte où les paroles s'entre-coupent. Par conséquent, l'analyse automatique des différentes caractéristiques des bandes sonores des corpus n'offrent pas une source de données facilement exploitables sur le plan acoustique. Deuxièmement, nous postulons que les tonalités de phrases et les tonalités lexicales se chevauchent dans la prononciation des MI. De plus, puisqu'on a affaire à des mots-phrases, les règles prosodiques qui servent à éviter les collisions d'accents ne s'appliquent pas, ou s'appliquent différemment (voir Martin, 1998).

En bout de ligne, même un système d'analyse linguistique automatique parfait ne suffirait pas à lui seul à constituer une machine capable de rivaliser avec un locuteur humain pour des tâches de

compréhension linguistique. Rappelons simplement que plusieurs informations sont communiquées de manières non-linguistiques au cours d'une conversation. Les moyens visuels, tels que les expressions faciales et les gestes de mains, sont particulièrement importants à ce sujet.

Les différents fichiers informatiques mentionnés au cours de cette thèse sont disponibles dans le répertoire suivant sur GitHub: https://github.com/flapatate/these.

# Bibliographie

Alm, C. O., Roth, D. et Sproat, R. (2005). Emotions from Text: Machine Learning for Text-based Emotion Prediction. Dans *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA : Association for Computational Linguistics. p. 579–586.

Beckman, M. E., Hirschberg, J. B. et Shattuck-Hufnagel, S. (2004). Chapter 2: The Original ToBI System and the Evolution of the ToBI Framework. Dans *Prosodic Models and Transcription: Towards Prosodic Typology*, p. 9-54.

Bird, S., Loper, E. et Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc. 479p.

Bolly, C., Crible, L., Degand, L. et 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, (16), p. 1-32.

Boser, B. E., Guyon, I. M. et Vapnik, V. N. (1992). A Training Algorithm for Optimal Margin Classifiers. Dans *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. New York, NY, USA : ACM. p. 144-152.

Brill, E. (1992). A Simple Rule-based Part of Speech Tagger. Dans *Proceedings of the Third Conference on Applied Natural Language Processing*. Stroudsburg, PA, USA : Association for Computational Linguistics. p. 152–155.

Brill, E. (1995). Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part-of-speech Tagging. *Comput. Linguist.*, 21(4), p. 543–565.

Bunt, H. (2000). Dialogue pragmatics and context specification. Dans W. Black et H. Bunt (dir.), *Abduction, belief and context in dialogue; studies in computational pragmatics*. John Benjamins Publishing. p. 81-105.

Crible, L. (2017). Towards an operational category of discourse markers: A definition and its model. Dans *Discourse markers, Pragmatics Markers and Modal Particles: New Perspectives*. Amsterdam: John Benjamins : C. Fedriani & A. Sanso. p. 99-124.

Crible, L. et Zufferey, S. (2015). Using a unified taxonomy to annotate discourse markers in speech and writing. Dans *Proceedings of the 11th Joint ACL - ISO Workshop on Interoperable Semantic Annotation (isa-11)*. p. 14-22.

Denturk, E. (2008). *Étude des marqueurs discursifs, l'exemple de* quoi. Gand : Universiteit Gent. 153p.

Dostie, G. (2004). *Pragmaticalisation et marqueurs discursifs : analyse sémantique et traitement lexicographique*. Bruxelles : De Boeck Duculot. 294p.

Dostie, G. (dir.) *et al*. (2006-2015). *(CFPQ) Corpus de français parlé au Québec.* CATIFQ-CRIFUQ, Université de Sherbrooke.

Dostie, G. (2007). La réduplication pragmatique des marqueurs discursifs. De *là* à *là là*. *Langue française*, (154), p. 45-60.

Dostie, G. (2013). Les associations de marqueurs discursifs. De la cooccurrence libre à la collocation. *Linguistik online*, 62(4), p. 15-45. (www.linguistik-online.de/62_13)

Dostie, G. (2015). Gros mots et petits mots dans une perspective prototypique. Les sacres et leurs substituts euphémisés en français québécois. *Cahiers de lexicologie*, (106), p. 55-89.

Dostie, G. et Lanciault, L. (2016). Changement catégoriel et développement sémantique. De sérieux adjectival à sérieux discursif dans le parler des jeunes locuteurs québécois. Dans *Modes langagières dans l'histoire*. Paris : Gilles Siouffi. p. 361-378.

Fraisse, A. et Paroubek, P. (2015). Utiliser les interjections pour détecter les émotions. Dans *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*. Caen, France : Association pour le Traitement Automatique des Langues. p. 279-292.

Goddard, C. (2013). Interjections and Emotion (with Special Reference to « Surprise » and « Disgust »). *Emotion Review*, 6(1), p. 53-63.

Goddard, C. (2014). The semantics of « surprise » (and « interest »). Communication présentée au International Symposium on Describing and Expressing Surprise, U. Paris-Diderot. 20p.

Goddard, C. (2015). « Swear words » and « curse words » in Australian (and American) English. At the crossroads of pragmatics, semantics and sociolinguistics. *Intercultural Pragmatics*, 12(2), 32p.

Goddard, C. et Ye, Z. (2014). Exploring « happiness » and « pain » across languages and cultures. *International Journal of Language and Culture*, 1(2), p. 131-148.

Grevisse, M. et Goosse, A. (2007). *Le bon usage : grammaire française*. Paris : Duculot ; Bruxelles : De Boeck, 2007. 1600p.

Hansen, M.-B. M. (1997). *Alors* and *donc* in spoken French: A reanalysis. *Journal of Pragmatics*, (28), p. 153-187.

Hansen, M.-B. M. (1998). *The Function of Discourse Particles: A Study with Special Reference to Spoken Standard French*. John Benjamins Publishing. 430p.

Hansen, M.-B. M. (2005). *A dynamic polysemy approach to the lexical semantics of discourse markers, (with an exemplary analysis of French* toujours*)*. Copenhague : Université de Copenhague. 44p.

Heeman, P. A. (1997). *Speech Repairs, Intonational Boundaries and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialog*. University of Rochester, New York. 280p.

Heeman, P. A. et Allen, J. F. (1999). Speech Repairs, Intonational Phrases, and Discourse Markers: Modeling Speakers' Utterances in Spoken Dialogue. *Comput. Linguist.*, 25(4), p. 527–571.

Heeman, P. A., Byron, D. et Allen, J. F. (1998). Identifying discourse markers in spoken dialog. Dans *AAAI 1998 Spring Symposium on Applying Machine Learning to Discourse Processing*, 8p.

Hirschberg, J. et Litman, D. (1987). Now Let's Talk About Now: Identifying Cue Phrases Intonationally. Dans *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics*. Stroudsburg, PA, USA : Association for Computational Linguistics. p. 163–171

Hirschberg, J. et Litman, D. (1993). Empirical Studies on the Disambiguation of Cue Phrases. *Comput. Linguist.*, 19(3), p. 501–530.

Hutchinson, B. (2004). Acquiring the Meaning of Discourse Markers. Dans *Proceedings of ACL-04*, p. 685–692.

Iordanskaja, L. et Mel'čuk, I. A. (1995). Traitement lexicographique de deux connecteurs textuels du français contemporain: EN FAIT vs EN RÉALITÉ. Dans *Tendances récentes en linguistique française et générale* (volume dédié à David Gaatone). Amsterdam/Philadelphia : Benjamins. p. 211-236

Iordanskaja, L. et Mel'čuk, I. A. (1999). Textual Connectors Across Languages: French EN EFFET vs. Russian V SAMOM DELE. RASK. Dans *E Pluribus Una*, p. 305-347.

Kerbrat-Orecchioni, C. (1977). *La connotation*. 3. ed. Presses universitaires de Lyon. 262p.

Lapointe, F. (2005). *Analyse sémantique et description lexicographique de marqueurs pragmatiques construits avec VRAI en français québécois* vraiment, pas vraiment, pour de vrai, pour dire vrai, à vrai dire *et* à dire vrai. Université de Sherbrooke. 109p.

Lapointe, F. (2007). Analyse sémantique de « pas vraiment » en français québécois. *Communication, lettres et sciences du langage*, 1(1), p. 72-80.

*Le Petit Robert 2017 : dictionnaire de la langue française*. (2016). Sous la direction éditoriale de A. Rey et J. Rey-Debove. Paris : Dictionnaires Le Robert.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press. 442p.

Li, Q., Ong, A., Suganthan, P. et Thing, V. (2010). A Novel Support Vector Machine Approach to High Entropy Data Fragment Classification. Communication présentée au International Workshop on Digital Forensics and Incident Analysis (WDFIA).

Litman, D. J. (1996). Cue Phrase Classification Using Machine Learning. Dans *Journal of Artificial Intelligence Research*, 5, p. 53-94.

Marcu, D. (1997). The Rhetorical Parsing of Natural Language Texts. Dans *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*. Stroudsburg, PA, USA : Association for Computational Linguistics. p. 96–103.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization*. Cambridge, MA, USA : MIT Press. 248p.

Martin, P. (1998). L'intonation : Analyse instrumentale et modèles. *Collezione dei preprint 1997-98, Lablita* (4), 10p.

McEnery, T. et Wilson, A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh University Press. 260p.

Mel'čuk, I. A. *et al.* (1984). *Dictionnaire explicatif et combinatoire du français contemporain : recherches lexico-sémantiques I.* Montréal : Presses de l'Université de Montréal. 172p.

Mel′čuk, I. A. *et al.* (1988). *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques II.* Montréal: Presses de l'Université de Montréal. 338p.

Mel'čuk, I. A. *et al.* (1992). *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques III.* Montréal: Presses de l'Université de Montréal. 356p.

Mel′čuk, I. A. *et al.* (1999). *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques IV.* Montréal: Presses de l'Université de Montréal. 367p.

Mel’čuk, I. A. (1997). *Vers une linguistique sens-texte : leçon inaugurale faite le vendredi 10 janvier 1997*. Paris : Collège de France. 43p.

Mel’čuk, I. A., Clas, A. P. et Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve : Duculot. 256p.

Pang, B. et Lee, L. (2008). Opinion mining and sentiment analysis. Dans *Foundations and Trends in Information Retrieval*, 2(1), p. 1–135.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Dans *Journal of Machine Learning Research*, 12, p. 2825−2830.

Petukhova, V. et Bunt, H. (2009). Towards a Multidimensional Semantics of Discourse Markers in Spoken Dialogue. Dans *Proceedings of the Eighth International Conference on Computational Semantics*. Stroudsburg, PA, USA : Association for Computational Linguistics. p. 157–168.

Petukhova, V., Geertzen, J. et Bunt, H. (2007). A multidimensional approach to utterance segmentation and dialogue act classification. Dans *Proceedings of the 8th SIGdial Workshop on Discourse*. p.140-149.

Pop, L. (2001). Adverbes de texte. Dans *L’Information Grammaticale*, 91(1), p. 13-19.

Popescu-Belis, A. et Zufferey, S. (2004). Towards Automatic Identification of Discourse Markers in Dialogs: The Case of *Like*. Dans *SIGdial 2004 (5th SIGdial Workshop on Discourse and Dialogue)*, p. 63-71.

Popescu-Belis, A. et Zufferey, S. (2011). Automatic identification of discourse markers in dialogues: An in-depth study of *like* and *well*. Dans *Computer Speech & Language*, 25(3), p. 499-518.

Post, B. (2006). IVTS, un système de transcription pour la variation prosodique. Dans *Bulletin de la Phonologie du Français Contemporain*, (6), p. 51-68.

Prsir, T. (2012). La citation théâtralisée : propositions pour une analyse prosodique et polyphonique de la citation à l'oral. Dans *Le discours et la langue : Revue de linguistique française et d'analyse du discours*, 2(2), p. 123-134.

Rosenberg, A. (2009). *Automatic detection and classification of prosodic events* (Columbia University). 381p.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Dans *Proceedings of International Conference on New Methods in Language Processing*. Manchester. p. 44-49.

Searle, J. (1979). *Expression and meaning: Studies in the theory of speech acts*. Cambridge : Cambridge University. 204p.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(2), 111-147. doi:10.2307/2984809

Wierzbicka, A. (1972). *Semantic primitives*. Frankfurt/M : Athenäum-Verl. 235p.

Wierzbicka, A. (1980). *Lingua mentalis : the semantics of natural language*. Sydney ; Academic Press, Toronto. 367p.

Wierzbicka, A. (1986). Human Emotions: Universal or Culture-Specific? Dans *American Anthropologist*, 88(3), p. 584-594.

Wierzbicka, A. (1997). *Understanding Cultures through Their Key Words: English, Russian, Polish, German, and Japanese* (1er edition). New York : Oxford University Press. 328p.

Wierzbicka, A. (1999). *Emotions Across Languages and Cultures: Diversity and Universals*. Cambridge University Press. 366p.

Wierzbicka, A. (2011). What's wrong with « happiness studies ». The cultural semantics of *happiness, bonheur, Gluck* and *scast'e*. Dans *Word and Language* (Slovo i Jazyk), Moscou, p. 155-171.

Zwicky, A. (1985). Clitics and particles. Dans *Language*, (61), p. 283-305.

# ANNEXE :
# Conventions de notation des transcriptions du CFPQ

Tiré de : http://recherche.flsh.usherbrooke.ca/cfpq.

**MATÉRIEL VERBAL**

| | |
|---|---|
| *Amorces de mots* | Trait d'union après l'unité (ex. : des ca- des cases vides). |
| *Chevauchements* | Crochets ouvrants vis-à-vis des passages où les locuteurs réagissent en même temps. |
| *Impossible à orthographier* | En API entre crochets. |
| *Inaudible* | (inaud.). |
| *Conversations parallèles* | Elles sont encadrées. |
| *Discours direct* | Placé entre deux points. Le premier, qui indique le début du discours direct, est noir (•) et le deuxième, qui signale sa fin, est blanc (°). (Ex. : j'ai raccroché (.) <f<•vite partez>> elle s'en vient elle s'en vient°). |

**MATÉRIEL PARAVERBAL**

**Prosodie**

<table>
<tr><td>Accentuation</td><td colspan="2">Lettres majuscules (ex. : ÉPOUvantable)</td></tr>
<tr><td>Allongement</td><td colspan="2">Deux points ou plusieurs fois deux points, selon son importance (ex. : c’est sû::r)</td></tr>
<tr><td>Intonation</td><td>Légèrement montante : /</td><td>Fortement montante : ↑</td></tr>
<tr><td></td><td>Légèrement descendante : \</td><td>Fortement descendante : ↓</td></tr>
<tr><td>Pauses</td><td colspan="2">Les micropauses (inférieures à une seconde) sont notées par un point entre parenthèses (cf. (.)). La durée des pauses supérieures à une seconde est mesurée et notée entre parenthèses (ex. : (3’’))</td></tr>
<tr><td>Volume de la parole</td><td>Forte (fort) : <f<vous pensez>></td><td>Fortissimo (très fort) : <ff<vous pensez>></td></tr>
<tr><td></td><td>Piano (bas) : <p<vous pensez>></td><td>Pianissimo (très bas) : <pp<vous pensez>></td></tr>
<tr><td></td><td>Crescendo (de plus en plus fort) : <cresc<vous pensez vraiment>></td><td>Diminuendo (de plus en plus bas) : <dim<vous pensez vraiment>></td></tr>
<tr><td>Vitesse de la parole</td><td>Allegro (rapide) : <all<vous pensez>></td><td>Lento (lent) : <len<vous pensez>></td></tr>
<tr><td></td><td>Accelerando (de plus en plus rapide) : <acc<vous pensez vraiment>></td><td>Rallentando (de moins en moins rapide) : <rall<vous pensez vraiment>></td></tr>
</table>

**Données vocales**

Entre parenthèses, en petites capitales (ex. : (RIRE)). Les autres productions vocales sont intégrées au texte et notées à l'aide des graphies les plus courantes (ex. : *hum, pff*).

**Gestuelle**

Description du geste présentée entre parenthèses, en caractères italiques, à côté de l'énoncé avec lequel celui-ci est en lien.

**Multitranscription**

Lorsque plusieurs possibilités de transcription se présentent, elles sont séparées par un point-virgule dans une accolade.

Ex.: *hier soir, je suis allé {aux feux ; au feu}*

Dans le contexte considéré, à savoir le soir de la Fête nationale du Québec, le pluriel et le singulier n'ont pas le même sens; le pluriel signifie ⁽des feux d'artifice⁾ et le singulier, ⁽un feu de camp⁾.