Quand les questions en disent plus que les réponses : classification automatique des intentions dans les questions

Angèle Barbedette, Iris Eshkol-Taravella

Quand les questions en disent plus que les réponses : classification automatique des intentions dans les questions

Angèle Barbedette

Université Sorbonne Nouvelle – Paris 3

Iris Eshkol-Taravella

Modèles, Dynamiques, Corpus (MoDyCo) - UMR 7114

Université Paris Nanterre

Résumé

Le travail présenté vise la détection automatique des intentions des locuteurs dans les questions posées au cours de repas. Le corpus est composé de transcriptions d'échanges oraux spontanés. Nous proposons une typologie de ces intentions basée sur les travaux existants et l'analyse du corpus. Nous procédons à l'implémentation d'un modèle de classification automatique supervisée s'appuyant sur les données annotées et sur des traits linguistiques choisis, dont nous évaluons et interprétons les résultats et performances.

Mots-clés : détection d'intention, acte de dialogue, transcriptions de parole spontanée, classification de questions, fouille d'opinion, apprentissage automatique

Abstract

This work aims to automatically detect the intents of speakers in questions asked during meals. The corpus is composed of transcripts of spontaneous oral conversations. We suggest a typology of these intents based on existing work and the corpus analysis. We implement a supervised automatic classification model based on annotated data and selected linguistic features, whose results and performance are evaluated and interpretated.

Keywords : intention detection, dialog act, spontaneous speech transcriptions, questions classification, opinion mining, machine learning

1. Introduction

1.1. Contexte

La recherche présentée dans cet article s'inscrit dans le domaine du Traitement Automatique du Langage Naturel (TAL). Les travaux actuels en TAL ne se concentrent plus uniquement sur ce que les phrases *veulent dire* et donc sur les indices présents directement dans les données, mais s'intéressent de plus en plus à ce que les locuteurs *veulent dire* de façon implicite à travers les énoncés qu'ils produisent. Les applications comme les chatbots ou autres agents conversationnels sont toutes basées sur ce principe de détecter automatiquement une intention de l'utilisateur.

Le travail présenté dans cet article s'intéresse aussi à ce que le locuteur veut exprimer lorsqu'il prend la parole, c'est-à-dire ses intentions, et se concentre sur les questions posées dans le cadre d'un discours spontané. Pour mieux comprendre le contexte de ce travail, nous proposerons d'abord un petit aperçu des travaux en TAL. Après avoir proposé une typologie des intentions dans les questions fondée sur l'exploration du corpus, nous présenterons les données utilisées et le processus d'élaboration du corpus de référence. Nous expliquerons ensuite les étapes du traitement automatique réalisé : le prétraitement des données, les traits linguistiques pris en compte, le module de classification automatique. Enfin, nous montrerons nos résultats et tenterons de les interpréter.

1.2. Travaux en TAL

Dans les travaux en TAL, la notion d'implicite est abordée selon plusieurs axes de recherches : fouille d'opinion et analyse de sentiments, classification automatique en actes de dialogue ou encore classification automatique de questions. Pour réaliser ces tâches, les méthodes sont fondées sur les algorithmes de machine learning ou sur les règles symboliques qui s'appuient sur des marqueurs ou indices présents dans le corpus pour repérer ce qui est recherché (opinions, actes de dialogue, etc.).

1.2.1. Fouille d'opinion

L'analyse de sentiments et d'opinions est un domaine très actif en TAL puisqu'il est de plus en plus facile de récolter sur le web des données comportant des opinions (des données subjectives), notamment grâce aux réseaux sociaux. L'analyse de ces données et de leur polarité positive ou négative constitue un enjeu majeur pour pouvoir mieux appréhender l'opinion générale sur un sujet précis au niveau du document (Pang et coll., 2002 ; Turney, 2002) ou au niveau de la phrase (Hatzivassiloglou et Wiebe, 2000 ; Riloff et Wiebe, 2003 ; Kim et Hovy, 2004 ; Wiebe et coll., 2004 ; Wilson et coll., 2004 ; Riloff et coll., 2006). D'autres travaux ne s'arrêtent plus à la détection de la polarité (positive, négative, neutre) d'une opinion ou d'un sentiment mais introduisent d'autres notions comme les suggestions (Brun et Hagege, 2013 ; Negi et Buitelaar, 2015 ; Negi et coll., 2016 ; Negi et coll., 2019 ; Eshkol-Taravella et Kang, 2019 ; Kang et Eshkol-Taravella, 2020 ; Ramanand et coll., 2010) ou encore les intentions (Carlos et Yalamanchi, 2012 ; Chen et coll., 2013 ; Eshkol-Taravella et Kang, 2019 ; Kang et Eshkol-Taravella,

2020; Benamara et coll., 2017). Karoui et coll. (2019) gardent le terme d'opinion mais font la distinction entre une opinion explicite qui peut être repérée à l'aide d'indices textuels (mots, symboles ou expressions subjectives du langage), et une opinion implicite qui s'appuie sur des connaissances culturelles ou pragmatiques communes à l'émetteur du message et à son récepteur. Les opinions implicites peuvent notamment comporter des éléments de langage dit figuratif, qui « détourne le sens propre pour lui conférer un sens dit figuré ou imagé » (Karoui et coll., 2014). Enfin, une autre notion plus générale, celle de la subjectivité, figure dans certains travaux en TAL. Les chercheurs essaient de détecter les passages considérés comme subjectifs dans les données traitées (Vernier, 2011 ; Flamein, 2019). Les méthodes de détection de toutes ces informations (opinions, sentiments, suggestions, intentions, subjectivité, etc.) sont fondées sur les règles linguistiques élaborées manuellement, l'apprentissage supervisé de surface ou profond.

1.2.2. Annotation manuelle et automatique en actes de dialogue

Interpréter automatiquement une discussion est une tâche difficile. C'est la raison pour laquelle certains travaux en TAL s'intéressent à la classification en actes de dialogue, consistant à repérer et à associer des étiquettes sémantiques aux énoncés et à caractériser ainsi les intentions des locuteurs.

Ces dernières années, des modèles sont apparus en proposant d'attribuer plusieurs étiquettes à un énoncé produit. Le DAMSL, *Dialog Act Markup in Several Layers,* (Allen et Core, 1997) s'attaque à cette problématique en proposant d'appliquer des étiquettes en plusieurs niveaux à un énoncé en tenant compte ainsi de la pluralité des intentions qu'un locuteur peut chercher à y exprimer. Les niveaux principaux, qui permettent de caractériser un énoncé selon l'intention qu'il contient et selon son contenu, sont au nombre de quatre :

– Communicative Status : évaluation de l'intelligibilité de l'énoncé

– Information Level : type de contenu sémantique

– Forward Looking Function : évaluation des effets de l'énoncé sur les actions de l'interlocuteur

– Backward Looking Function : évaluation du rapport entre l'énoncé et les énoncés précédents

Chacun de ces niveaux comprend des étiquettes plus fines, rendant le modèle générique puisqu'il peut être appliqué à différents types de dialogues. Un autre modèle, la théorie de l'interprétation dynamique, *dynamic interpretation theory,* ou DIT, basée sur DAMSL, distingue le contenu sémantique de l'énoncé et sa fonction communicative (sa force illocutoire, son intention) (Bunt, 2005 ; Bunt, 2009).

L'annotation manuelle en actes de dialogue consiste à classer des énoncés dans une catégorie choisie parmi un ensemble de catégories prédéfinies qui remplissent des fonctions particulières du discours social (Moldovan et coll., 2011). Pour permettre la réalisation de cette tâche, plusieurs taxonomies sont proposées : SWBD-DAMSL (Jurafsky et coll., 1997), HCRC Map Task (Anderson et coll., 1991), VERBMOBIL (Alexandersson et coll., 2011), ou encore DIT ++ (Bunt, 2005 ; Bunt, 2009). D'autres travaux portent sur la définition d'un ensemble ou d'une hiérarchie d'actes de dialogue comme le jeu de balises des actes de dialogue

Switchboard (Stolcke et coll., 2000) ou la norme ISO 24617-2 d'annotation des dialogues (Bunt et coll., 2010 ; Bunt et coll., 2012)

1.2.3. Approche de la sémantique formelle

Le dialogue est étudié également en TAL du point de vue de la sémantique formelle où l'identification de l'incompréhension entre les locuteurs se réduit à la recherche d'incohérences logiques dans les combinaisons entre les actes de parole calculées à partir des algorithmes compositionnels fondés sur une représentation logique et dynamique d'une question et d'une réponse (Amblard et coll., 2019 ; Boritchev et Amblard, 2018 ; Boritchev et Amblard, 2019).

Les questions sont étudiées notamment chez Ginzburg et coll. (2019) qui distinguent sept catégories de questions en tant que réponses à des requêtes : des demandes de clarification, des questions dépendantes du contexte, des questions avec une motivation sous-jacente, des questions qui visent à changer de sujet, des questions permettant de préciser la réponse à apporter, des questions avec une réponse déjà présupposée et enfin des questions hors-sujet, ignorant la requête d'origine.

En conclusion, détecter automatiquement l'intention dans un discours, c'est-à-dire repérer automatiquement une information implicite, est un vrai défi pour le TAL, ce qui explique le nombre important de travaux cités ci-dessus. Tous ces travaux varient en fonction de la tâche visée et des techniques utilisées. Certains s'appuient sur des corpus, d'autres sur des modèles, mais tous distinguent le contenu sémantique de l'énoncé et ses fonctions communicatives. Notre travail s'inscrit également dans cette lignée mais aborde les intentions sous un autre angle. En premier lieu, les données proviennent des enregistrements oraux du discours spontané qui traite des sujets divers et variés et ne se limite pas à un domaine particulier, ce qui est le cas de plusieurs travaux évoqués ci-dessus comme ceux de Chen et coll. (2013), Pang et coll. (2002) ou encore Turney (2002). Notre travail met également au centre de ses recherches les intentions et leurs différentes catégories et ne s'arrête pas à la présence de l'intention comme c'est le cas chez Eshkol-Taravella et Kang (2019) et Kang et Eshkol-Taravella (2020). Il vise la détection des types d'intentions dans les questions fondée sur une typologie proposée et inspirée de l'exploration du corpus.

2. Typologie des intentions dans les questions

Ce travail porte sur les questions posées au cours des repas et les intentions qui y sont exprimées.

Les questions font partie, pour nous, des énoncés performatifs au sens d'Austin (1962) pour lequel lorsqu'un locuteur dit quelque chose, il fait quelque chose. Nous considérons que lorsqu'un locuteur pose une question, il veut toujours dire quelque chose de plus que ce que la question exprime littéralement. Austin distingue les énoncés constatifs qui servent à décrire le monde (exemple [1]), les énoncés performatifs explicites (exemple [2]) et les énoncés performatifs implicites (exemple [3]). Les énoncés performatifs ont pour caractéristique de n'être ni vrais, ni faux

mais plutôt heureux ou malheureux, réussis ou non, puisqu'ils ne décrivent rien de particulier. Ils correspondent à l'accomplissement d'une action de la part du locuteur de l'énoncé qui dit quelque chose.

[1] La fenêtre est ouverte.

[2] Je te promets de venir te voir demain.

[3] Je viendrai te voir demain.

Afin de pouvoir mieux définir ce qui constitue les performatifs et les constatifs, Austin distingue trois actes qui s'accomplissent selon lui au moment de l'énoncé :

– l'acte locutoire, la production de l'énoncé en elle-même, c'est-à-dire des sons qui le constituent ;

– l'acte illocutoire, l'intention communicative du locuteur lorsqu'il dit quelque chose, la force attribuée à l'énoncé ;

– l'acte perlocutoire, l'effet attendu sur le ou les interlocuteur(s) ou sur la situation de communication en elle-même lorsque le locuteur dit quelque chose.

Les questions entrent par ailleurs dans le cadre des travaux de Searle (1975) et de son concept d'acte de langage indirect : un acte illocutoire accompli indirectement par un autre acte. Cet acte de langage indirect se compose d'un acte illocutoire primaire, c'est-à-dire non littéral, et d'un acte illocutoire secondaire, c'est-à-dire littéral. Dans l'énoncé « Peux-tu me passer le sel ? », l'acte de langage secondaire (le sens littéral) serait « As-tu la capacité de me passer le sel ? » et l'acte de langage primaire (le sens non littéral) serait « Donne moi le sel ».

Pour déterminer ce que le locuteur veut dire, il faut donc se demander s'il a souhaité, à travers son énoncé, communiquer quelque chose de façon indirecte, autrement dit si son énoncé correspond à une implicature. Le principe de coopération introduit par Grice (1975) suppose que les locuteurs participent à la conversation d'une manière efficace pour permettre la réussite de celle-ci, c'est-à-dire qu'ils suivent certaines règles et normes implicites dont ils ont connaissance.

Grice développe ces différentes règles sous la forme de quatre maximes conversationnelles : quantité (informativité de la contribution), relation (pertinence de la contribution), qualité (caractère vrai de la contribution) et manière (clarté de la contribution). Lorsqu'un locuteur produit un énoncé, son interlocuteur le décode soit en faisant la supposition que le locuteur respecte le principe de coopération et donc qu'il suit et applique les maximes conversationnelles présentées par Grice, soit en partant de l'idée qu'il y a eu violation des maximes, ce qui peut arriver lorsque l'énoncé est défini par un acte de langage indirect.

Les questions peuvent donc accomplir, selon nous et tels que les définit Searle, un acte illocutoire primaire (non littéral) et un acte illocutoire secondaire (littéral), que l'interlocuteur ou le destinataire doit savoir interpréter selon le contexte d'énonciation et en fonction des règles conversationnelles dont ils ont connaissance et qui sont décrites chez Grice (1975).

Les questions portent une valeur illocutoire, c'est-à-dire une intention de la part du locuteur. L'intention correspond à l'activité illocutoire exprimée par un énoncé (au sens de Ducrot (1972), « l'ensemble des actes qui s'accomplissent, immédiatement et spécifiquement, par l'exercice de la parole »), qui permet de caractériser celui-ci selon son but, que ce dernier soit explicite (repérable directement dans l'énoncé) ou implicite (s'appuyant sur les connaissances

communes des acteurs de la conversation). Cette distinction est poursuivie chez Chen et coll. (2013), selon lesquels les intentions peuvent être explicites et clairement énoncées ou implicites. Par exemple dans la question « Quelqu'un connaît l'autonomie de l'iPhone ? », le locuteur a possiblement l'idée d'acheter un nouveau téléphone, mais il n'exprime pas son souhait d'une manière directe. Cette distinction entre les intentions explicites et implicites est reprise dans la modélisation des intentions proposée dans ce travail.

Notre démarche est empirique et fondée sur les observables du corpus. La typologie présentée s'appuie sur les données, pour lesquelles elle est plus adaptée et pertinente. Nous nous intéressons aux données orales transcrites et recueillies dans un contexte particulier : des repas de familles, où la conversation entre les personnes est réalisée de manière spontanée. La notion d'intention apparaît dans ces données sous un angle nouveau et plus large que celui évoqué dans les travaux évoqués précédemment.

La typologie proposée distingue deux niveaux correspondant à deux dimensions : la première, T1, s'intéressant au type de réponse attendu pour chacune des questions et donc à leur dimension explicite, et la seconde, T2, portant sur l'intention exprimée de façon non littérale à travers chacune des questions par le locuteur et donc plutôt sur leur dimension implicite (voir tableau 1).

Tableau 1 : Exemples de questions annotées selon les deux niveaux d'annotation T1 (avec deux étiquettes : *demande d'accord* et *demande d'information*) et T2 (avec trois étiquettes : *avis*, *volonté* et *doute*)

QUESTIONS	T1	T2
ils sont vraiment *bêtes* hein ?	demande d'accord	avis
bah non c'est quoi ce Sopalin de *brin* là ?	demande d'information	avis
pourquoi tu es *malpolie* ?	demande d'information	avis
tu nous *sers* à boire mon chéri ?	demande d'accord	volonté
tu en *veux* toi ?	demande d'accord	volonté
tu me *donnes* un petit peu maman ?	demande d'accord	volonté
ou *ailleurs* c'est quoi *ailleurs* ?	demande d'information	doute
sûrement non ?	demande d'accord	doute
il marche euh *vraiment* le truc ?	demande d'accord	doute

Explicite (T1)

Suite à Borillo (1978) et Ahmadvand et coll. (2019), nous distinguons deux types d'intentions dans les questions au niveau explicite :

– les « Yes-No-Question » ou *demandes d'accord*, c'est-à-dire les interrogations totales;

– les « Open-Question » ou *demandes d'information*, correspondant aux interrogations partielles.

Des exemples pour ces deux classes peuvent être trouvés dans le tableau 1.

Implicite (T2)

Le niveau implicite est plus complexe à déterminer car les intentions de ce niveau dépendent du contexte extralinguistique et de la situation de communication et nécessitent une certaine interprétation. En effet, elles correspondent à un message non littéral que l'émetteur laisse entendre au récepteur du message et que ce dernier doit parvenir à décoder et comprendre. Grâce aux différentes étapes ayant permis la constitution de notre corpus de référence que nous allons développer dans la partie suivante, nous avons pu dégager trois classes représentant l'intention exprimée par le locuteur produisant la question : *avis*, *volonté* et *doute*. Ces trois types d'intentions répondent à des critères précis et se caractérisent par différents marqueurs possibles. Nous en trouverons des exemples dans le tableau 1.

Lorsque l'intention du locuteur est d'exprimer un *avis*, les questions portant cette intention répondent aux critères suivants :

– elles expriment des jugements ou catégorisations positifs ou négatifs ;

– elles n'impliquent pas nécessairement une action de la part d'un des locuteurs.

Les marqueurs possibles pour l'expression d'un avis sont des adjectifs, des adverbes ou encore des locutions verbales qui aident à l'expression d'opinions tels que « je trouve que », « j'adore », « ennuyeux » ou « honnêtement ». Ces éléments font partie des lexiques des émotions, des sentiments et des opinions.

La question « ils sont vraiment bêtes hein ? » illustre cette classe puisqu'elle contient l'adjectif « bête » qui porte une valeur de jugement, dans ce cas-ci plutôt négatif. Le locuteur exprime alors de façon implicite une opinion à travers sa question, qui correspond par ailleurs d'un point de vue explicite à une demande d'accord.

Lorsque l'intention du locuteur est d'exprimer une *volonté*, les questions portent les caractéristiques suivantes :

– elles correspondent à la volonté d'une action ou d'un comportement de la part du locuteur ou de son/ses interlocuteur(s) (nous ne savons pas forcément ce que le locuteur veut mais nous savons qu'il veut quelque chose) ;

– elles impliquent une réponse correspondant à une action dans le présent, dans un futur proche[1].

Les marqueurs possibles pour cette classe peuvent être :

[1]Selon Benamara et coll. (2017), la notion d'intention est liée au futur, mais l'étude de notre corpus montre qu'on ne peut pas caractériser cette notion seulement par le futur. La typologie proposée révèle d'autres de ses facettes.

– des verbes d'action tels que « aller », « manger », « dormir » ou « regarder » ;

– des verbes exprimant une volonté comme « vouloir », « souhaiter » ou « désirer ».

La question « tu nous sers à boire mon chéri ? » pourrait être un exemple pour cette classe. En effet, il s'agit ici d'une demande d'accord qui comprend un verbe d'action (« sers ») et qui implique donc à la fois une volonté d'une action (ici de la part de l'interlocuteur) et une réponse correspondant à une action (ici l'action de servir quelque chose à boire).

Enfin, lorsque l'intention du locuteur correspond à l'expression d'un *doute*, les questions répondent à d'autres critères :

– elles mettent en doute de ce qui est dit, le caractère vrai ou faux d'une chose ou d'un événement extérieur ;

– elles peuvent s'apparenter à une répétition, une demande de précisions ou de l'étonnement ;

– elles n'impliquent pas nécessairement une action de la part d'un des locuteurs.

Les marqueurs possibles pour exprimer un doute sont :

– la répétition d'une partie du contexte précédent ;

– la présence de mots interrogatifs tels que « qui », « quel » ou « quoi » ;

– la présence d'adverbes d'affirmation/modaux/de certitude comme «sûrement », « peut-être » ou « probablement ».

La question « ou ailleurs c'est quoi ailleurs ? » est un exemple illustrant cette classe : il s'agit ici d'une répétition du terme « ailleurs » qui provient du contexte précédant la question (« tu aimerais aller en Angleterre ? »/« en Angleterre ou ailleurs »). Cette répétition est du point de vue explicite une demande d'information et permet de mettre en doute ce qui a été dit dans le contexte.

Les trois classes présentées ci-dessus rassemblent des questions qui n'expriment pas seulement une demande d'accord ou d'information. L'exemple « pourquoi tu es malpolie ? » est à la fois une demande d'information et une façon d'exprimer à l'interlocuteur qu'il manque de politesse. Il permet donc d'exprimer un avis. L'exemple « tu me donnes un petit peu maman ? » correspond à un ordre et donc à une volonté de la part du locuteur (« donne-moi un petit peu maman ») en plus d'être une demande d'accord. Enfin, l'énoncé « il marche euh vraiment le truc ? » est une demande d'accord d'un point de vue explicite mais suppose indirectement que le locuteur pense qu'il y a un dysfonctionnement.

3. Données utilisées

3.1. Présentation générale

Les données utilisées pour la réalisation de ce travail proviennent des corpus oraux ESLO1 et ESLO2, créés dans le cadre du projet scientifique ESLO, Enquêtes SocioLinguistiques à Orléans, du Laboratoire Ligérien de Linguistique de l'Université d'Orléans (Baude et Dugua, 2011 ; Eshkol-Taravella et coll., 2011).

3.2. Architecture du corpus et format des transcriptions

Le corpus ESLO1 est constitué d'environ 300 heures d'enregistrement et celui d'ESLO2 de 400 heures d'enregistrement. Nous pouvons compter pour ESLO1 un nombre à peu près équivalent entre les interviews et les enregistrements plus variés tels que les communications téléphoniques ou les conversations au cours de repas. ESLO2 comprend plus d'enregistrements divers de conversations dans des lieux publics (boulangeries, marchés, commerces, guichets et cinémas par exemple) ou privés (conversations au cours de repas) que d'entretiens.

Les transcriptions des enregistrements d'ELSO1 et d'ESLO2 sont en grande partie disponibles sur le site web d'ESLO et ouvertes au public. Elles sont au format *.xml* et les métadonnées principales que nous pouvons y trouver sont l'identité des locuteurs et des transcripteurs, les temps de début et de fin de transcription par rapport à l'enregistrement ou encore les indications sur les tours de paroles au cours de l'enregistrement, information qui nous sera utile pour la suite.

3.3. Utilisation des données

Dans le cadre de ce travail et pour en assurer sa reproductibilité, nous utiliserons toutes les transcriptions d'enregistrements effectués au cours de repas et disponibles au public sur le site web du projet ESLO, à la fois dans ESLO1 et dans ESLO2, ce qui correspond à un total de 28 fichiers au format .xml ayant pour extension .trs. Parmi ces 28 fichiers, sept font partie du corpus ESLO1 et 21 sont issus du corpus ESLO2. Ces fichiers forment un tout d'environ 19 heures d'enregistrement.

Le choix de cette catégorie précise est lié au fait de vouloir utiliser les données les plus spontanées possibles. En effet, un des objectifs de ce travail consistant à prédire l'intention du locuteur à travers des questions, il semblait plus logique de ne pas utiliser des transcriptions d'enregistrements réalisés lors d'entretiens ou autres situations plutôt formelles, comme des interviews à l'aide de questionnaires, mais au contraire d'utiliser des données récoltées au cours de conversations naturelles, pendant des moments de détente, dans un contexte souvent informel, entre amis ou en famille et dans des lieux familiers des participants. Les repas sont des contextes favorisant la présence de questions, objets de cette étude, car habituellement une partie des conversations dans ces contextes portent sur des demandes et interrogations quant au contenu des repas servis ou aux actions impliquées par la préparation d'un repas.

3.4. Mise en forme des données

Afin de commencer l'annotation manuelle du corpus pour obtenir le corpus de référence, les données des transcriptions originales ont été mises en forme à l'aide d'un script prévoyant plusieurs étapes.

Le corpus a d'abord été nettoyé et les tours de parole ont été récupérés (repérables dans les fichiers originaux par des balises *Turn*). Nous avons ensuite procédé à la récupération de toutes les questions (que nous appellerons également cibles dans la suite de ce travail) à l'aide des points d'interrogation transcrits.[2]

[2]Certaines questions jugées peu intéressantes telles que « ah oui ? » ou « ah bon » ont été exclues du corpus.

Seules, ces questions ne sont pas suffisamment informatives : il est nécessaire de prendre en compte le contexte autour de chacune d'elles afin de réduire les difficultés pour comprendre la conversation sans écouter l'audio. Nous avons donc récupéré les contextes gauche et droit de chaque question en fixant un seuil à dix tours de parole avant et après la question (dans certains cas, il y a moins de dix tours avant ou après, lorsqu'une question se situe dans les premiers ou derniers tours de parole du fichier, au début ou à la fin de l'enregistrement). Toutes ces informations ont ensuite été enregistrées dans un fichier de sortie (figure 1) et des attributs vides ont été ajoutés pour chaque question dans le but de les remplir lors de l'étape d'annotation manuelle du corpus.

La figure 1 est un extrait de ce fichier de sortie qui va servir pour la tâche d'annotation présentée dans la suite de ce travail. L'image montre une question qui apparaît dans le corpus, entourée d'une balise cible. Plusieurs attributs sont associés à cette balise, dont deux contenant des informations déduites directement des fichiers de transcriptions originaux : *n*, correspondant au compteur de cibles, et *spk*, correspondant aux identifiants des différents locuteurs. Nous pouvons aussi retrouver avec la même valeur l'attribut *n* au niveau de la balise *contexteG*, qui comprend le contexte gauche de la cible, et de la balise *contexteD*, qui comprend le contexte droit de la cible. Ces deux balises encadrent elles-mêmes un ensemble de dix tours de parole, compris chacun dans des balises *tourG* pour un tour de parole appartenant au contexte gauche ou *tourD* pour un tour de parole appartenant au contexte droit, chacune de ces balises possédant également un attribut *spk*.

Figure 1. Extrait du fichier d'annotation correspondant une question ou cible du corpus à annoter, accompagnée de ses contextes gauche et droit, composés de dix tours de paroles chacun

```
<contexteG n="773">
<tourG spk="spk1">['ça va oui']</tourG>
<tourG spk="spk2">['alors euh enfin je vous ai monté']</tourG>
<tourG spk="spk1">['entrez entrez']</tourG>
<tourG spk="spk3">['merci']</tourG>
<tourG spk="spk5">["je l'ai dit à mamie mais faut l'enlever faut l'enlever elle me dit non elle
voulait mettre de l'eau dessus"]</tourG>
<tourG spk="spk1">['il est bête', 'à raconter ça à tout le monde']</tourG>
<tourG spk="spk3">['tu as mangé les', 'pour ce soir ?']</tourG>
<tourG spk="spk1 spk5">["alors ça normalement c'est"]</tourG>
<tourG spk="spk2">['ah on peut']</tourG>
<tourG spk="spk1">['au fait vous connaissez mon frère ? oui']</tourG>
<tourG spk="spk1 spk2">["vous l'avez déjà vu je crois une fois hein ?", 'oui je oui au cinéma à
My Fair Lady']</tourG>
</contexteG>

<cible n="773" spk="spk1 spk2" explicite="" implicite="" doute_plus="">vous l'avez déjà vu je
crois une fois hein ?</cible>

<contexteD n="773">
<tourD spk="spk1 spk2">["vous l'avez déjà vu je crois une fois hein ?", 'oui je oui au cinéma à
My Fair Lady']</tourD>
<tourD spk="spk1">["c'est ça oui"]</tourD>
<tourD spk="spk4">['My Fair Lady']</tourD>
<tourD spk="spk4 spk2">['en français en plus', 'My Fair Lady']</tourD>
<tourD spk="spk1">['y a plus rien à boire']</tourD>
<tourD spk="spk5">["c'est ça qu'on met sur la table ?"]</tourD>
<tourD spk="spk1 spk3">['attends attends on va mettre', 'on va mettre la nappe hein']</tourD>
<tourD spk="spk3">['prends le non non']</tourD>
<tourD spk="spk1">["bon euh moi je m'assois hein", 'oh', 'et si on ouvrait la fenêtre ?']</tourD>
<tourD spk="spk5 spk4">['oh oui', "oui c'est vrai"]</tourD>
<tourD spk="spk1">['oui ce serait une bonne idée hein ?']</tourD>
</contexteD>
```

4. Corpus de référence

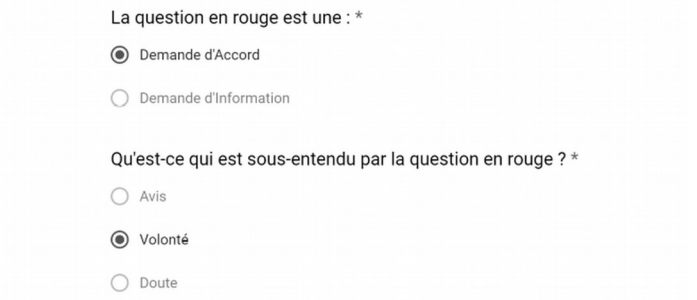
4.1. Processus d'annotation

L'annotation manuelle du corpus correspond à un processus itératif composé de plusieurs phases. Notre démarche s'inspire des sciences participatives (Millour et Fort, 2018). Après avoir élaboré la typologie des intentions dans les questions (partie 2), nous avons préparé, à l'aide d'un formulaire en ligne créé depuis Google Forms (voir figure 2), une tâche d'annotation de quinze questions issues de notre corpus. Ce formulaire comprenait un ensemble de conventions tirées de la typologie, ainsi que des exemples et des contre-exemples pour chacune des catégories d'annotation. Le lien a été diffusé au sein de notre réseau social et 26 participants ont répondu à cet appel.

Ce nombre d'énoncés à annoter et plus généralement le format choisi pour cette tâche sont liés à la difficulté et à la durée plutôt longue de l'exercice demandé : la tâche demandait une forte concentration pour comprendre les conversations de part la nature des données (lecture de transcriptions de l'oral sans pouvoir écouter l'audio) et leur longueur (une question et ses contextes gauche et droit comprenant dix tours de parole chacun). Il nous a été rapporté par les annotateurs une durée d'en moyenne une heure pour l'annotation des quinze questions. Nous avons donc choisi d'alléger la tâche cognitive pour les participants en proposant un nombre restreint de questions à annoter pour assurer des annotations de qualité.

Figure 2. Extrait du formulaire d'évaluation contenant une question (en rouge dans le formulaire original, ici encadrée) entourée de ses contextes gauche et droit, l'ensemble accompagné de l'identifiant du locuteur et suivi de deux questions : la première concernant T1 (l'aspect littéral) et la seconde concernant T2 (l'intention exprimée)

P3	'traînent', 'la mode traîne', 'parce que la'
P3/P1	'oh là là', "tu as vu l'orage ?"
P1	'horrible hein'
P3	"tu crois pas qu'il faut quelque chose de marrant ?"
P1/P3	'il y a', 'oh non non'
P3/P1	-
P2/P1	"ce qui est le plus joli c'est carrément le le le le à mi-mollet", 'tiens tu en veux euh', '?'
P2	'vraiment moi je'
P1	'Suzanne tu en veux ?'
P3	'non merci'
P4	'tu me donnes un petit peu maman ?'
P3	"moi je trouve qu'on on", 'pareil à', 'à mi-mollet là oh moi tu sais je crois que ça reviendra', 'plus ou moins ça reviendra', 'aussi bien', "
P1/P3	'moi je crois que y aura'
P3	"y aura la mode pour l'hiver et puis la mode pour l'été", "le court pour l'été le long pour l'hiver"
P3/P2	'tu te rends compte ?', 'ah oui'
P5	"mais ce qui est le plus marrant c'est qu'en été y a"
P1	'ah oui'
P5	'encore une autre'
P3/P1	'eh bah dis donc', 'dis donc elle va'
P1	'hein'
P5	'encore'



4.2. Évaluation

A partir de ces 26 participations, nous avons mesuré l'accord inter-annotateur de plusieurs façons : d'abord avec un Kappa de Cohen, un coefficient permettant de mesurer l'accord entre deux jugements qualitatifs, entre la référence et les réponses de chacun des participants, puis avec un Kappa de Fleiss qui mesure un accord entre tous les participants.

Dans le premier cas, nous avons obtenu pour 50% des participations (grâce à une mesure de la médiane) un accord inter-annotateur supérieur à 0,73 pour l'explicite et supérieur à 0,6 pour l'implicite. Cela signifie, d'après la table d'interprétation du K de Cohen de Landis et Koch (1977), que nous obtenons pour la moitié des participations un accord fort (entre 0,61 et 0,8) ou un accord presque parfait (entre

0,81 et 1) à la fois pour les classes de T1 et de T2. Nous observons également (grâce au calcul du 3e quartile) qu'environ 25% des participants ont obtenu un accord presque parfait pour les deux typologies, puisqu'il est supérieur à 0,86 pour T1 et supérieur à 0,8 pour T2.

Dans le second cas, nous obtenons pour les classes de T1 un score global de 0,54 avec un pourcentage d'accord global d'environ 77% et pour les classes de T2 un score global de 0,52 avec un pourcentage d'accord global avoisinant les 68%.

De façon générale, les résultats montrent une concordance moyenne entre les annotateurs. Ils reflètent une certaine cohérence mais mettent aussi en évidence les difficultés de compréhension, d'interprétation et donc d'annotation ayant pu être rencontrées au cours de la tâche demandée.

Tableau 2 : Accords inter-annotateurs obtenus pour l'évaluation de la référence avec un Kappa de Cohen (accord calculé entre la référence et chacun des annotateurs ce qui aboutit à 26 mesures obtenues pour 26 participants) et un Kappa de Fleiss (d'abord global, puis entre spécialistes et entre non spécialistes)

		T1	T2
Kappa de Cohen	Médiane	0,73	0,6
	3e quartile	0,86	0,8
Kappa de Fleiss		0,54	0,52
Pourcentage d'accord global		77,1	67,9

5. Détection automatique

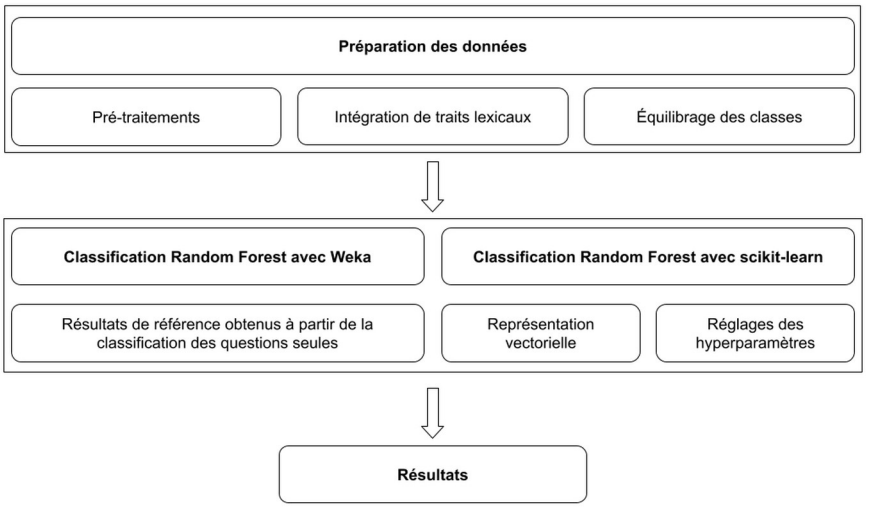
Nous proposons dans cette partie un modèle de classification des questions en fonction de leur intention (*avis*, *volonté* ou *doute*).

Les différents marqueurs possibles présentés pour chaque classe dans notre typologie constituent des indices linguistiques qui peuvent permettre d'aider à catégoriser les questions que ce soit manuellement ou automatiquement, c'est pourquoi nous avons décidé dans la suite de ce travail de favoriser des algorithmes d'apprentissage supervisé.

Le schéma proposé en figure 3 montre les étapes qui ont permis d'annoter automatiquement les questions du corpus et résume les sections qui suivent : nous évoquerons d'abord les pré-traitements qui ont été nécessaires (POS tagging et lemmatisation) ainsi que l'ajout de traits lexicaux servant d'indices pertinents et l'équilibrage des différentes classes pour éviter un biais dans les résultats. Nous expliquerons ensuite les classifications Random Forest opérées sur les données à l'aide de Weka ou du module Python scikit-learn. Nous expliquerons enfin les résultats obtenus.

Dans la suite de cette section, nous comparerons les résultats obtenus avec ceux d'autres classifieurs et nous tenterons de les interpréter.

Figure 3. Schéma global de la démarche de détection des intentions dans les questions



5.1. Pré-traitements et traits linguistiques intégrés

5.1.1. Étiquetage morpho-syntaxique et lemmatisation

Avant de commencer la tâche de classification automatique qui sera présentée plus tard, des pré-traitements ont été effectués sur les transcriptions :

– l'étiquetage morpho-syntaxique (ou Part-Of-Speech tagging), qui consiste en l'attribution d'une étiquette correspondant aux informations grammaticales d'un mot, étiquette qui sera injectée en tant que descripteur à notre algorithme de classification ;

– la lemmatisation, qui consiste en la transformation de chaque mot en son lemme (ou sa forme canonique) pour notamment faciliter l'utilisation de lexiques comme traits linguistiques par la suite.

En ce qui concerne la lemmatisation, nous avons choisi de nous servir de TreeTagger, un outil d'annotation permettant d'obtenir à la fois l'information du lemme et du POS tagging de chaque mot (Schmid, 1994), et plus particulièrement des fichiers de paramètres du projet PERCEO, un Projet d'Étiqueteur Robuste pour l'Écrit et pour l'Oral constitué de ressources dont l'objectif est l'annotation automatique de données orales ou écrites en lemmes et en parties du discours (Benzitoun et coll., 2012). Ces fichiers de paramètres ont été obtenus depuis le module d'entraînement de TreeTagger, à partir de fichiers provenant au départ du projet TCOF, Traitement de Corpus Oraux en Français. Un script nous a permis d'obtenir un fichier contenant les données lemmatisées et un fichier contenant les données étiquetées morpho-syntaxiquement, ces deux fichiers étant exactement au même format que le fichier original de données annotées.

La raison du choix de ces différentes ressources dans le cadre de notre travail s'explique par la volonté d'utiliser des outils adaptés à nos données, qui sont des

transcriptions d'enregistrements, ayant eu lieu au cours de repas généralement informels.

5.1.2. Traits lexicaux

La mise en place de la typologie des intentions a permis de relever des indices linguistiques que nous avons décidé de regrouper sous la forme de six lexiques détaillés ci-dessous :

– les verbes de parole (34 occurrences formant un lexique s'appuyant sur (Eshkol, 2002)) : ils sont généralement utilisés pour rapporter, introduire du discours et des paroles dans un récit et ont leur utilité ici puisqu'ils peuvent permettre d'exprimer une volonté de la part d'un des locuteurs (*dire, demander, proposer, suggérer, expliquer*, etc.) ;

– les verbes de mouvement (1003 occurrences issues de la ressource lexicale DinaVmouv, (Stosic et Aurnague, 2017)) : ils correspondent à des indices permettant de détecter une volonté car comme nous l'avons expliqué dans notre typologie, les cibles de cette classe supposent une action de la part d'un des locuteurs et les verbes de mouvement expriment, comme leur nom l'indique, une idée de mouvement, de déplacement et donc d'action (*accrocher, suivre, s'asseoir, remplir, parcourir*, etc.) ;

– les mots interrogatifs (24 occurrences) : il s'agit des adverbes et pronoms interrogatifs qui peuvent permettre de détecter une mise en doute à travers une question (*qui, combien, lequel, pourquoi, quand*, etc.) ;

– les interjections (73 occurrences) : elles correspondent simplement à la liste d'interjections présentées dans le guide de transcription des corpus ESLO et peuvent, en fonction du contexte, permettre la détection d'une mise en doute, d'une émotion ou d'une opinion (*mouais, hein, miam, bof, bah*, etc.) ;

– les sentiments (190 occurrences formant un lexique s'appuyant sur des entrées du lexique des sentiments et des émotions du NRC (National Reasearch Council Canada) (Mohammad et Turney, 2013)) : ce sont des indices pour détecter des avis qui peuvent être des noms, des verbes, des adjectifs ou des adverbes et qui permettent d'exprimer des émotions ou des opinions (*apprécier, ravi, haine, nul, inquiéter*, etc.) ;

– les adverbes et adjectifs modaux, de certitude ou épistémiques (24 occurrences s'appuyant sur plusieurs travaux (Rossari et coll., 2016 ; Rossari et Salsmann, 2017)) : ils s'agit ici d'indices pour arriver à déterminer des cibles exprimant un doute puisqu'ils permettent d'affirmer ou de contester quelque chose (*vraiment, impossible, certainement, peut-être, vrai*, etc.).

En plus des informations liées aux lexiques, c'est-à-dire la fréquence d'apparition des mots de chacun des lexiques dans chaque cible et contexte gauche ou droit, nous avons intégré à nos traits lexicaux le nombre de mots et le nombre de caractères par cible et par contexte, en nous appuyant notamment sur l'hypothèse selon laquelle la longueur de l'énoncé peut être liée à l'intention exprimée : par exemple, une question exprimant une opinion serait peut-être plus longue qu'une question exprimant une volonté puisque cette dernière n'implique pas a priori de longue argumentation. Ce choix s'appuient aussi sur les descripteurs utilisés dans le travail de Grabar et Eshkol-Taravella (2016) qui traite aussi les données orales transcrites et

qui vise la détection automatique des raisons de reformulations dans les discussions.[3]

5.2. Mise en place des expériences

5.2.1. Choix des classifieurs

Des algorithmes d'apprentissage supervisé ont été utilisés pour la tâche de classification des questions puisqu'ils permettent de contrôler (ajouter, supprimer, modifier) les données et indices fournis et de tester plusieurs combinaisons d'indices pour obtenir les meilleures performances possibles. Ce sont également des modèles qui ont fait leurs preuves dans le domaine de la fouille d'opinion. Pour la suite de l'étude, le principal classifieur qui servira aux expériences sera Random Forest. D'autres classifieurs (Naive Bayes, Decision Tree et SVM) seront testés pour comparer les résultats.

5.2.2. Équilibrage des classes

Toutes nos expérimentations ont été testées en équilibrant les classes à l'aide d'une diminution des données pour éviter un biais dans les résultats : une classe comportant un nombre plus élevé de cibles aurait probablement été favorisée par rapport aux autres lors de l'étape de classification tandis qu'une classe avec un nombre d'occurrences inférieur à celui des autres classes aurait pu être moins représentée dans les résultats de la classification. Le nombre total de questions au départ était de 3647. Les données à supprimer ont été déterminées aléatoirement pour obtenir finalement 2538 questions à classer selon T1, soit 1269 par classe, et 858 questions à classer selon T2, soit 286 pour chacune des trois classes.

5.2.3. Classification dans Weka

Pour commencer nos expériences, nous avons d'abord utilisé le logiciel Weka, qui permet l'utilisation de nombreux algorithmes et outils de machine learning, notamment pour la classification automatique, dans le cadre de l'exploration de données. Nous avons donc pu tester l'algorithme Random Forest pour la classification des questions seulement, représentées par un sac de mots (fréquences brutes), d'une part dans les classes de T1 (*demande d'accord*, ou *DA*, et *demande d'information*, ou *DI*) avec 2 538 cibles à classer et d'autre part dans les classes de T2 (*avis*, *volonté* et *doute*) avec 858 cibles à classer. Nous avons obtenu 82,782% d'éléments bien classés et 17,218% d'éléments mal classés pour le premier cas et 60,14% d'éléments bien classés et 39,86% d'éléments mal classés pour le second cas. Le tableau 4 rassemble les résultats obtenus avec des mesures de précision, rappel et F-mesure pour chaque classe appartenant aux typologies T1 et T2, ainsi qu'une moyenne globale de ces mesures par typologie. Ces chiffres constituent des résultats de référence à essayer d'améliorer en utilisant l'algorithme Random Forest et en ajoutant des traits linguistiques pertinents. Pour la suite de ce travail, nous nous

[3]Il s'agissait dans ce travail d'attribuer automatiquement à l'aide de méthodes supervisées à chaque reformulation une fonction pragmatique déterminée parmi un un ensemble restreint de onze fonctions telles que l'explication, la justification, la paraphrase ou encore la correction référentielle.

concentrerons uniquement sur les classes de T2, c'est-à-dire sur la dimension implicite des données.

Tableau 4 : Mesures de précision, rappel et F-mesure obtenues avec Random Forest dans Weka pour les classes de T1 (*DA* pour *demande d'accord* et *DI* pour *demande d'information*) et de T2 (*avis*, *volonté* et *doute*)

	DA	DI	Moy. T1	Avis	Volonté	Doute	Moy. T2
Précision	0,894	0,781	0,837	0,737	0,636	0,516	0,63
Rappel	0,744	0,912	0,828	0,462	0,622	0,72	0,601
F-mesure	0,812	0,841	0,827	0,568	0,629	0,601	0,599

5.2.4. Classification finale

5.2.4.1. Représentation des questions

Nous avons utilisé en entrée de notre algorithme de classification des traits linguistiques comprenant notamment une représentation des questions et de leur contexte (vectorisation ou normalisation) pour capturer au mieux leur sens. En comparaison de vectorisations censées capturer le sens des mots, notamment word2vec avec les modèles CBOW et Skip-Gram s'appuyant sur le corpus FrWac (Fauconnier, 2015) (200 dimensions) ou les vecteurs pré-entraînés de Flair (Akbik et coll., 2018) (4096 dimensions), le calcul d'un TF-IDF (term frequency-inverse document frequency) nous a permis d'obtenir des résultats de classification légèrement meilleurs (tableau 3), pouvant être liés au fait que les modèles de word2vec et Flair ne sont pas pré-entraînés sur des données orales. Pour ce travail, nous avons donc modélisé les questions du corpus et leurs contextes par un poids TD-IDF, sans mélanger leurs lexiques afin de prendre en compte l'importance, la rareté et la fonction discriminative des mots du corpus.

Tableau 3 : Comparaison des représentation obtenues avec les modèles CBOW et Skip-Gram de word2vec, les vecteurs pré-entraînés de Flair et avec un TD-IDF (les résultats pour les words embbedings ont été obtenus en prenant en compte tous les mots et en moyennant les vecteurs pour chaque mot)

	CBOW	SKIP-GRAM	FLAIR	TF-IDF
F-mesure	0,582	0,57	0,583	0,611

5.2.4.2. Réglages et évaluation

Nous avons utilisé la méthode de validation croisée qui consiste en la séparation du corpus en k échantillons (paramètre pour lequel nous avons choisi la valeur de 8) pour ensuite utiliser tour à tour chacun d'entre eux comme ensemble de test et les

autres comme ensemble d'apprentissage. Cette méthode a été appliquée en suivant l'ordre naturel des instances du corpus.

Afin d'obtenir les résultats les plus satisfaisants avec Random Forest, plusieurs hyperparamètres ont été testés. Pour cela, la fonction GridSearchCV de sklearn a été utilisée. Elle permet, une fois que nous lui avons fourni des hyperparamètres associés à différentes valeurs possibles, de sélectionner les valeurs les plus optimales. Les hyperparamètres qui ont été testés et leurs valeurs choisies sont les suivants :

– *n_estimators* qui correspond au nombre d'arbres constituant la forêt d'arbres de décision avec ici une valeur de *1000* (choisie parmi les valeurs suivantes : *100, 300, 500, 800* et *1000*) ;

– *criterion* qui correspond au type de mesure choisie pour évaluer la qualité de chaque point de séparation d'un arbre, c'est-à-dire de chaque noeud (gain ou perte d'informations notamment) avec ici la valeur *entropy* (choisie entre *gini* et *entropy*) ;

– *bootstrap* qui correspond à l'utilisation ou non de nouveaux échantillons de données sélectionnés dans l'échantillon initial (dans le cas où cette technique ne serait pas utilisée, la construction des arbres se baserait sur l'ensemble de l'échantillon initial) avec ici la valeur *True*.

La performance du modèle a été évaluée en faisant une moyenne des performances obtenues pour chaque groupe à l'aide de mesures de précision, rappel et F-mesure et en visualisant grâce à une matrice de confusion l'ensemble des prédictions de notre algorithme.

Nous sommes conscients que les résultats obtenus peuvent être surévalués par la méthode choisie. Cependant, l'objectif de cette étude n'est pas de généraliser mais plutôt de réaliser des expériences préliminaires pour tenter de dégager des indices pertinents pour la classification des questions de notre corpus.

5.3. Résultats des expériences

Un des objectifs étant d'avoir une idée des traits permettant d'aider à la tâche de classification automatique de nos cibles, nous avons décidé de tester notre algorithme avec plusieurs combinaisons de traits différentes qui sont présentées dans le tableau 5. Ce tableau est un récapitulatif global de l'ensemble des valeurs de précision, rappel et F-mesure obtenues pour chaque expérience. Chacune d'elle correspond à un ensemble choisi de traits parmi ceux présentés précédemment que nous avons cochés dans le tableau :

– représentation de la question et de ses contextes avec un TF-IDF
– POS tagging de la question (représenté avec les fréquences brutes)
– présence d'un ou plusieurs élément(s) issu(s) des lexiques (représentée par la fréquence d'apparition de ces éléments) :
    – des verbes de parole
    – des verbes de mouvement
    – des mots interrogatifs
    – des interjections
    – des sentiments
    – des adverbes et adjectifs modaux
– indication de la dimension explicite de la question (trait binaire)

– longueur en mots de la question et de ses contextes (trait à valeur continue)

– longueur en caractères de la question et de ses contextes (trait à valeur continue)

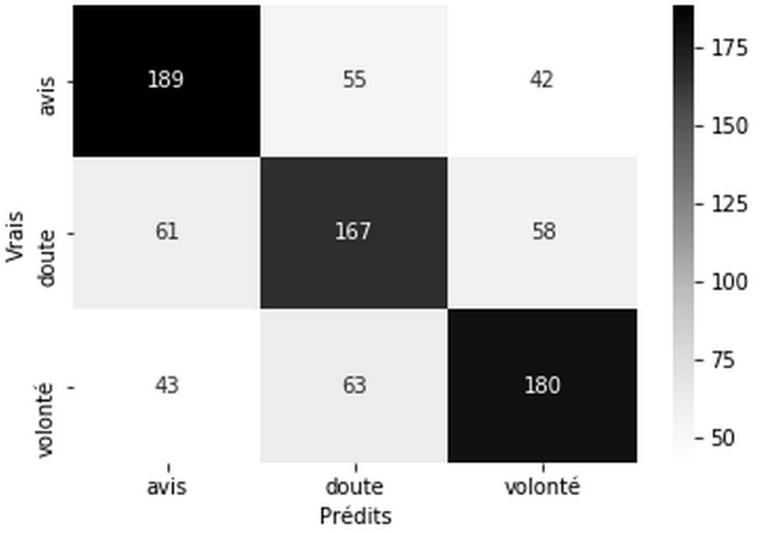
Nous pouvons voir par exemple que toutes les expériences incluent le vecteur de la cible mais que seules les expériences 8 et 9 incluent la longueur des contextes (c'est-à-dire la longueur en termes de mots et de caractères), et que l'expérience 3 ne prend en compte que deux traits correspondant au vecteur de la cible et au POS tagging de la cible.

Ce tableau montre des résultats très proches dans l'ensemble comme nous le voyons pour les expériences 4, 6 et 8 pour lesquelles nous obtenons des scores de F-mesure avoisinant 0,62. La variance associée à ces scores moyens calculés oscillent entre 0,001 et 0,003 ce qui indique une bonne stabilité de ces résultats. Nous pouvons observer la distribution exacte des prédictions de notre algorithme avec une matrice de confusion (figure 4). Certaines mesures semblent cependant se distinguer, telles que celles de l'expérience 2 qui sont inférieures à 0,5 ou celles de l'expérience 5 inférieures à 0,6, ces deux expériences prenant en compte des traits liés au contexte de la question cible (la vectorisation dans le premier cas et la présence de mots des lexiques dans le second).

Tableau 5 : Récapitulatif des expériences et résultats obtenus avec Random Forest selon neuf expériences prenant en compte des traits différents, listés dans la colonne de gauche, où Q = question, C = contextes, lexiques = présence d'un ou plusieurs élément(s) issu(s) des lexiques constitués et longueur = nombre de mots et de caractères

	Expériences								
	1	2	3	4	5	6	7	8	9
vectorisation Q	X	X	X	X	X	X	X	X	X
vectorisation C		X							
POS tagging Q			X	X	X	X	X	X	
lexiques Q				X	X	X	X	X	X
lexiques C					X				
explicite						X	X	X	X
longueur Q							X	X	X
longueur C								X	X
Précision	0,622	0,492	0,618	0,632	0,6	0,63	0,621	0,631	0,613
Rappel	0,612	0,493	0,612	0,623	0,592	0,622	0,617	0,624	0,606
F-mesure	0,611	0,489	0,61	0,622	0,59	0,62	0,616	0,622	0,603

Figure 4. Matrice de confusion permettant de visualiser les prédictions de l'expérience 8



5.4. Comparaison avec d'autres classifieurs

Afin d'avoir un point de comparaison, nous avons étudié les résultats obtenus pour les expériences 1, 2, 5 et 8 en testant d'autres classifieurs qui sont Naive Bayes, Decision Tree et Support Vector Machine (SVM). Ces tests ont été réalisés avec les valeurs par défaut des hyperparamètres. Les expériences choisies pour tester ces algorithmes sont celles comportant peu de traits (expérience 1), ayant obtenu les moins bons scores (expériences 2 et 5) et les meilleurs (expérience 8). Les scores de F-mesure obtenus sont présentés dans le tableau 6, avec les meilleurs indiqués en gras pour chacune des expériences : ceux que Random Forest a fourni sont supérieurs aux autres pour trois expériences sur quatre.

Tableau 6 : Comparaison des F-mesures obtenues pour les expériences 1, 2, 5 et 8 avec les classifieurs Naive Bayes, Decision Tree et SVM

	Exp. 1	Exp. 2	Exp. 5	Exp. 8
Naive Bayes	0,433	0,385	0,437	0,483
Decision Tree	0,532	0,431	0,502	0,528
SVM	0,592	**0,554**	0,579	0,59
Random Forest	**0,611**	0,489	**0,59**	**0,622**

5.5. Vers une interprétation des résultats obtenus : hypothèses et discussion

Pour cette partie, nous nous sommes appuyés sur les résultats obtenus avec Random Forest.

La première observation faite concerne la baisse des performances lorsque des traits liés aux contextes sont ajoutés, comme dans les expériences 2 et 5. Pour vérifier l'importance du contexte des cibles, nous avons reproduit l'expérience 2 et l'expérience 8 avec deux et cinq tours de parole avant et après la question afin de comparer les résultats obtenus initialement avec dix tours de parole. Les scores obtenus (tableau 7) pour l'expérience 2, qui prend en compte uniquement les vectorisations de la cible et des contextes, montrent une amélioration des performances lorsqu'il y a moins de tours de parole dans chaque contexte. Ceci peut s'expliquer par la trop grande quantité d'informations non pertinentes rapportées par la vectorisation des contextes lorsque ceux-ci sont plus larges. En revanche, nous observons pour l'expérience 8 que les scores sont meilleurs lorsque la fenêtre de contexte est plus grande : lorsqu'il s'appuie sur des informations lexicales et notamment le nombre de mots et de caractères, les performances de l'algorithme sont meilleures avec plus de contexte.

Tableau 7 : Comparaison des expériences 2 et 8 avec deux, cinq et dix tours de parole par contexte gauche et droit

	Expérience 2			Expérience 8		
Nombre de tours de parole par contexte	2	5	10	2	5	10
Précision	0,575	0,54	0,492	0,588	0,591	0,631
Rappel	0,572	0,536	0,493	0,584	0,587	0,624
F-mesure	0,571	0,535	0,489	0,582	0,585	0,622

Pour tenter d'interpréter les résultats obtenus, nous nous sommes uniquement concentrés sur l'expérience 8, une des expériences ayant eu les meilleurs scores et prenant en compte le plus de traits : elle comprend tous les traits apportant des informations sur les cibles, ainsi que les informations du nombre de mots et du nombre de caractères pour les contextes gauche et droit, qui sont les seules concernant les contextes des cibles à classer. Pour avoir une meilleure vision des résultats et comprendre la façon dont l'algorithme a fonctionné et pourquoi certaines questions ont été bien classées quand d'autres non, nous avons calculé la médiane et la moyenne du nombre de mots et du nombre de caractères pour chaque cible, contexte gauche et contexte droit. Nous avons également calculé le nombre de mots appartenant aux différents lexiques présents en moyenne dans chaque cible.

Tableau 8 : Mesures pour les questions des classes *avis* (A), *volonté* (V) et *doute* (D) bien et mal identifiées

	A→A	A→D	A→V	V→V	V→A	V→D	D→D	D→A	D→V
méd. mots c.	6	5	5	6	5	5	3	5	6
méd. mots g.	96	92	93,5	87,5	95	88	92	85	101
méd. mots d.	97	97	81	88,5	97	88	84	94	92
méd. caractères c.	21	19	21	21	19	17	11	17	21
méd. caractères g.	331	326	330,5	312,5	322	302	327	312	341
méd. caractères d.	344	338	301	310	345	317	299	331	309,5
verbes de parole	0,03	0,04	0,07	0,04	0,04	0,02	0,05	0,08	0,07
verbes de mouvement	0,06	0,07	0,12	0,19	0,12	0,1	0,08	0,11	0,16
mots interrogatifs	0,34	0,53	0,47	0,33	0,35	0,37	0,58	0,46	0,59
interjections	0,54	0,27	0,25	0,18	0,37	0,3	0,08	0,3	0,24
sentiments	0,41	0,09	0,12	0,03	0,09	0,13	0,04	0,16	0,1
modaux	0,31	0,07	0,09	0,13	0,26	0,13	0,09	0,11	0,07

Pour ce qui est de la classe *doute*, nous avons remarqué que la médiane pour le nombre de caractères des questions y appartenant est de 11. Il s'agit d'une longueur plutôt courte en comparaison des questions de la classe *doute* ayant été mal classées en *avis* ou en *volonté* pour lesquelles nous trouvons respectivement des médianes de 17 et de 21. Les questions de la classe *doute* bien classées sont donc majoritairement plus courtes que les autres questions, ce qui confirme une hypothèse émise précédemment selon laquelle l'intention exprimée est liée à la longueur de la question posée. Pour cette classe, la présence de mots interrogatifs est également significative puisqu'elle est plus élevée en moyenne qu'ailleurs (0,58). Les questions de cette classe ayant été prédites en *volonté* contiennent en moyenne 0,16 verbes de mouvement, un chiffre qui se rapproche de la moyenne du nombre de verbes de mouvement présents dans les bonnes prédictions de *volonté* qui est de 0,19.

Nous observons pour la classe *avis* que les cibles bien classées ont en moyenne une forte présence de mots appartenant aux lexiques des sentiments (0,41), des interjections (0,55) et des modaux (0,31) comparé aux autres classes et aux cibles mal classées de la classe *avis*. En effet, ces dernières, lorsqu'elles sont classées dans *volonté* ont en moyenne 0,12 mots appartenant au lexique des sentiments, 0,25 mots

appartenant aux interjections et 0,09 mots étant des modaux. Ceci se confirme également pour les cibles *avis* classées en *doute*, pour lesquelles nous voyons par ailleurs que le nombre de mots interrogatifs présents en moyenne est de 0,53, un chiffre qui se rapproche de celui obtenu pour les cibles *doute* bien classées qui est de 0,58.

Enfin, lorsque nous nous concentrons sur les prédictions pour la classe *volonté*, nous remarquons une plus forte présence de verbes de mouvement qu'ailleurs, avec une moyenne de 0,19 pour les cibles bien classées en *volonté*. Ce chiffre est plus bas pour les cibles mal classées en *doute* (0,1) et se rapproche notamment de la moyenne du nombre de verbes de mouvement pour les bonnes prédictions de *doute* qui est de 0,08. Pour les cibles de *volonté* classées en *avis*, nous observons une présence plus forte de mots des lexiques d'interjections (0,37) et de modaux (0,26), qui sont des caractéristiques de la classe *avis*.

En plus de cette analyse des résultats obtenus, nous avons pu remarquer que la plupart d'entre eux sont proches puisqu'ils avoisinent souvent 0,6 pour la précision, le rappel et la F-mesure. Ces scores montrent les difficultés rencontrées mais sont également prometteurs au regard de la tâche visée : en effet, ce travail consistait à détecter automatiquement la dimension implicite des questions dans du discours oral spontané sans s'appuyer sur des indices prosodiques et en essayant d'identifier des traits permettant de les distinguer. De plus, les résultats obtenus avec les différents algorithmes testés sont comparables à ceux obtenus suite à la tâche d'annotation collaborative proposée pour évaluer notre typologie, avec un accord inter-annotateur global pour l'aspect implicite de 0,52.

6. Conclusion

Les recherches menées dans le cadre de ce travail s'intéressent aux intentions dans les questions posées par les locuteurs au cours de repas. Les données proviennent de transcriptions d'enregistrements oraux de discours spontanés. L'approche fondée sur l'exploration du corpus a permis de proposer une typologie des intentions dans les questions en deux niveaux : explicite et implicite, définis respectivement par les catégories *demande d'accord* et *demande d'information* et par les catégories *avis*, *volonté* et *doute*. Le corpus a été annoté selon la typologie proposée et suivant une démarche de sciences participatives. L'évaluation de l'annotation et de la typologie a montré un accord inter-annotateur plus élevé pour le niveau explicite que pour le niveau implicite. Pour détecter les différentes catégories de la typologie proposée, plusieurs expériences ont été mises en place avec Random Forest et les meilleurs résultats ont atteint des valeurs d'environ 0,62 pour la précision, le rappel et la F-mesure.

Plusieurs difficultés ont été rencontrées au cours de cette étude et peuvent s'expliquer par plusieurs points. D'abord, ce travail est ambitieux car il s'attaque à la détection des informations implicites dans les questions, qui peuvent être difficile à prédire automatiquement. Ensuite, les données sont des enregistrements oraux transcrits, réalisés dans un cadre spontané. Enfin, le corpus est de taille relativement petite, ce qui complique la tâche de l'apprentissage supervisé. Malgré cela, les résultats obtenus sont encourageants et prometteurs et les techniques de

l'apprentissage supervisé ainsi que les traits linguistiques choisis (lexiques, longueur en mots et en caractères) semblent adaptés à la tâche.

En perspective, il serait intéressant d'ajouter aux traits linguistiques exploités des indices de nature prosodique pour mieux tenir compte de la modalité orale des données.

7. Références bibliographiques

AHMADVAND, A., CHOI, J. et I. et AGICHTEIN, E. 2019. Contextual Dialogue Act Classification for OpenDomain Conversational Agents. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval.* New York : Association for Computing Machinery : 1273–1276.

AKBIK, A., BLYTHE, D. et VOLLGRAF, R. 2018. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics.* Santa Fe : Association for Computational Linguistics : 1638–1649.

ALEXANDERSSON, J., BUSCHBECK–WOLF, B., FUJINAMI, T., KIPP, M., KOCH, S., MAIER, E., REITHINGER, N., SCHMITZ, B. et SIEGEL M. 2011. *Dialogue Acts in VERBMOBIL–2.* Sarrebruck : DFKI.

ALLEN, J. et CORE, M. 1997. *Draft of DAMSL : Dialog Act Markup in Several Layers*.

AMBLARD, M., BORITCHEV, M., CARLETTI, M., DIEUDONAT, L. et TSAI, Y. 2019. A Taxonomy of Real–Life Questions and Answers in Dialogue. In *SemDial 2019 – LondonLogue – 23rd Workshop on the Semantics and Pragmatics of Dialogue.* Londres : SEMDIAL.

ANDERSON, A. H., BADER, M., BARD, E. G., BOYLE, E., DOHERTY, G., GARROD, S., ISARD, S., KOWTKO, J., MCALLISTER, J., MILLER, J. et coll. 1991. The HCRC Map Task Corpus. *Language and speech* 34 (4) : 351–366.

AUSTIN, J. L. 1962. *How to do Things with Words.* William James Lectures. Oxford : Oxford University Press.

BAUDE, O. et DUGUA, C. 2011. (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? *Corpus* 10 : 99–118.

BENAMARA, F., TABOADA, M. et MATHIEU, Y. 2017. Evaluative Language beyond Bags of Words : Linguistic Insights and Computational Applications. *Computational Linguistics* 43 (1) : 201–264.

BENZITOUN, C., FORT, K. et SAGOT, B. 2012. TCOF–POS : un corpus libre de français parlé annoté en morphosyntaxe. In *Actes de la conférence jointe JEP–TALN–RECITAL 2012.* Grenoble : ATALA; AFCP : 99–112.

BORILLO, A. 1978. La construction postposée et le mode interrogatif. *Cahier de linguistique* 8 : 17–42.

BORITCHEV, M. et AMBLARD, M. 2018. Coffee or Tea ? Yes. In L. PREVOT, M. OCHS et B. FAVRE (éd.), *Proceedings of the 22nd Workshop on the Semantics and Pragmatics of Dialogue.* Aix–en–Provence : SEMDIAL.

BORITCHEV, M. et AMBLARD, M. 2019. A compositional view of questions. In *WiNLP – Widening NLP – ACL Workshop*. Florence : Association for Computational Linguistics.

BRUN, C. et HAGEGE, C. 2013. Suggestion Mining : Detecting Suggestions for Improvement in Users' Comments. *Research in Computing Science* 70 (79) : 171–181.

BUNT, H. 2005. A Framework for Dialogue Act Specification. In *Proceedings of SIGSEM WG on Representation of Multimodal Semantic Information*. Tilburg.

BUNT, H. 2009. The DIT++ Taxonomy for Functional Dialogue Markup. In *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts.* Budapest : 13–24.

BUNT, H., ALEXANDERSSON, J., CARLETTA, J., CHOE, J.–W., FANG, A. C., HASIDA, K., LEE, K., PETUKHOVA, V., POPESCU–BELIS, A., ROMARY, L. *et coll.* 2010. Towards an ISO Standard for Dialogue Act Annotation. In *Proceedings of the 7th conference on International Language Resources and Evaluation*. La Valette : European Language Resources Association.

BUNT, H., ALEXANDERSSON, J., CHOE, J.–W., FANG, A. C., HASIDA, K., PETUKHOVA, V., POPESCU–BELIS, A. et TRAUM D. R. 2012. ISO 24617–2 : A Semantically–Based Standard for Dialogue Annotation. In *Proceedings of the 8th conference on International Language Resources and Evaluation*. Istanbul : European Language Resources Association.

CARLOS, C. S. et YALAMANCHI, M. 2012. Intention Analysis for Sales, Marketing and Customer Service. In *Proceedings of the 24th International Conference on Computational Linguistics.* Bombay : Association for Computational Linguistics : 33–40.

CHEN, Z., LIU, B., HSU, M., CASTELLANOS, M. et GHOSH, R. 2013. Identifying Intention Posts in Discussion Forums. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies.* Atlanta : Association for Computational Linguistics : 1041–1050.

DUCROT, O. 1972. *Dire et ne pas dire : principes de sémantique linguistique.* Collection Savoir. Paris : Hermann.

ESHKOL, I. 2002. *Typologie sémantique des prédicats de parole.* Thèse de Doctorat non publiée. Université Sorbonne Paris Nord, Villetaneuse.

ESHKOL–TARAVELLA, I., BAUDE, O., MAUREL, D., HRIBA, L., DUGUA, C. et TELLIER I. 2011. Un grand corpus oral « disponible » : le corpus d'Orléans 1 1968–2012. *Traitement Automatique des Langues* 53 (2) : 17–46.

ESHKOL–TARAVELLA, I., et KANG, H. J. 2019. Observation de l'expérience client dans les restaurants. In *TALN 2019 – Conférence annuelle du Traitement Automatique des Langues Naturelles.* Toulouse : ATALA : 361–370.

FAUCONNIER, J.–P. 2015. *French Word Embeddings*.

FLAMEIN, H. 2019. Étude de la perception d'une ville. Repérage automatique, analyse et visualisation. Thèse de Doctorat non publiée. Université d'Orléans, Orléans.

GINZBURG, J., YUSUPUJIANG, Z., LI, C., et al. 2019. Characterizing the Response Space of Questions : a Corpus Study for English and Polish. In

Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue. Stockholm : Association for Computational Linguistics : 320–330.

GRABAR, N. et ESHKOL–TARAVELLA, I. 2016. Prédiction automatique de fonctions pragmatiques dans les reformulations. In *In TALN 2016 – Conférence annuelle du Traitement Automatique des Langues Naturelles.* Paris : ATALA.

GRICE, H. P. 1975. Logic and Conversation. *Syntax and Semantics* 3. P. COLE, et J. MORGAN (éd.). New York : Academic Press : 41–58.

HATZIVASSILOGLOU, V. et WIEBE, J. M. 2000. Effects of Adjective Orientation and Gradability on Sentence Subjectivity. In *Proceedings of the 18th International Conference on Computational Linguistics.* Saarebruck : Association for Computational Linguistics : 299–305.

JURAFSKY, D., SHRIBERG, E. et BIASCA D. 1997. *Switchboard SWBD–DAMSL Shallow–Discourse–Function Annotation Coders Manual*. Rapport technique 97–02. Boulder : University of Colorado : ICS.

KANG, H. J. et ESHKOL–TARAVELLA, I. 2020. Les avis sur les restaurants à l'épreuve de l'apprentissage automatique. In *Actes de la conférence jointe JEP–TALN–RECITAL 2020*. Nancy : ATALA; AFCP : 249–257.

KAROUI, J., BENAMARA, F. et MORICEAU, V. 2019. *Détection automatique de l'ironie : Application à la fouille d'opinion dans les microblogs et les médias sociaux.* ISTE Group.

KAROUI, J., GILLES, N. A., BENAMARA ZITOUNE, F. et BELGUITH, L. 2014. Le langage figuratif dans le web social : cas de l'ironie et du sarcasme. In *Workshop Fouille d'opinion dans le Web social.* Lyon.

KIM, S.–M. et HOVY, E. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*. Genève : Association for Computational Linguistics : 1367–1373.

LANDIS, J. R. et KOCH, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* : 33 (1) :159–174.

MILLOUR, A. et FORT, K. 2018. À l'écoute des locuteurs: production participative de ressources langagières pour des langues non standardisées. *Traitement Automatique des Langues* 59 (3) : 41–65.

MOHAMMAD, S. M. et TURNEY, P. D. 2013. Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence* 29 (3) : 436–465.

MOLDOVAN, C., RUS, V. et GRAESSER, A. 2011. Automated Speech Act Classification For Online Chat. In S. VISA, A. INOUE, A. RALESCU (éd.), *Proceedings of the 22nd Midwest Artificial Intelligence and Cognitive Science Conference.* Cincinnati : MAICS : 23–29.

NEGI, S., ASOOJA, K., MEHROTRA, S. et BUITELAAR, P. 2016. A Study of Suggestions in Opinionated Texts and their Automatic Detection. In *Proceedings of the 5th Joint Conference on Lexical and Computational Semantics.* Berlin : Association for Computational Linguistics : 170–178.

NEGI, S. et BUITELAAR, P. 2015. Towards the Extraction of Customer–to–Customer Suggestions from Reviews., In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* Lisbonne : Association for Computational Linguistics : 2159–2167.

NEGI, S., DAUDERT, T. et BUITELAAR P. 2019. Semeval–2019 task 9 : Suggestion Mining from Online Reviews and Forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis : Association for Computational Linguistics : 877–887.

PANG, B., LEE, L. et VAITHYANATHAN, S. 2002. Thumbs Up ? : Sentiment Slassification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics : 79–86.

RAMANAND, J., BHAVSAR, K. et PEDANEKAR, N. 2010. Wishful Thinking : Finding Suggestions and « Buy » Wishes from Product Reviews. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.* Los Angeles : Association for Computational Linguistics : 54–61.

RILOFF, E., PATWARDHAN, S. et WIEBE, J. 2006. Feature Subsumption for Opinion Analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing.* Sydney : Association for Computational Linguistics : 440–448.

RILOFF, E. et WIEBE, J. 2003. Learning Extraction Patterns for Subjective Expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing.* Sapporo : Association for Computational Linguistics : 105–112.

ROSSARI, C., HÜTSCH, A., RICCI, C., SALSMANN, M. et WANDEL, D. 2016. Le pouvoir attracteur de mais sur le paradigme des adverbes épistémiques : du quantitatif au qualitatif. In *Actes des 13èmes Journées internationales d'Analyse statistique des Données Textuelles.* Nice.

ROSSARI, C. et SALSMANN, M. 2017. Étude quantitative des propriétés dialogiques des adverbes épistémiques. In *Actes des 9èmes Journées Internationales de la Linguistique de corpus.* Grenoble : 87–93.

SCHMID, H. 1994. Probabilistic Part–of–Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing.* Manchester : 44–49.

SEARLE, J. 1975. Indirect Speech Acts. *Syntax and Semantics* 3 *Speech Acts*. P. COLE, J. MORGAN (éd.). New York : Academic Press.

SEARLE, J. et VANDERVEKEN, D. 1985. *Foundations of Illocutionary Logic*. CUP Archive.

STOLCKE, A., RIES, K., COCCARO, N., SHRIBERG, E., BATES, R., JURAFSKY, D., TAYLOR, P., MARTIN, R., ESSDYKEMA, C. V. et METEER, M. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational linguistics* 26 (3) : 339–373.

STOSIC, D. et AURNAGUE, M. 2017. *DinaVmouv : Description, INventaire, Analyse des Verbes de MOUVement. An Annotated Lexicon of Motion Verbs in French*. Ressource lexicale.

TURNEY, P. D. 2002. Thumbs Up or Thumbs Down ? : Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics.* Philadelphie : Association for Computational Linguistics, : 417–424.

VERNIER, M. 2011. *Analyse à granularité fine de la subjectivité*. Thèse de Doctorat non publiée. Université de Nantes, Nantes.

WIEBE, J., WILSON, T., BRUCE, R., BELL, M. et MARTIN, M. 2004. Learning Subjective Language. *Computational linguistics* 30 (3) : 277–308.

WILSON, T., WIEBE, J. et HWA, R. 2004. Just how Mad are you ? Finding Strong and Weak Opinion Clauses. In *Proceedings of the 19th National Conference on Artificial Intelligence*. San Jose : Association for the Advancement of Artificial Intelligence : 761–769.