THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*
Cotutelle internationnale *Faculté des Sciences Économiques et de Gestion de Sfax (FSEGS)*

Présentée et soutenue le *28/09/2017* **par :**

JIHEN KAROUI

Détection automatique de l'ironie dans les contenus générés par les utilisateurs

---

JURY

LYNDA TAMINE-LECHANI	Professeur, UT3 Paul Sabatier	Présidente du Jury
PATRICK PAROUBEK	Ingénieur de Recherche HDR, LIMSI-CNRS	Rapporteur
PATRICE BELLOT	Professeur, Aix-Marseille Université	Rapporteur
FARAH BENAMARA ZITOUNE	Maître de conférences HDR, UT3 Paul Sabatier	Co-Directrice de thèse
NATHALIE AUSSENAC-GILLES	Directrice de Recherche CNRS, IRIT Toulouse	Directrice de thèse
LAMIA HADRICH BELGUITH	Professeur, FSEGS Tunisie	Directrice de thèse
PAOLO ROSSO	Professeur, Universitat Politècnica de València	Examinateur
AHMED HADJ KACEM	Professeur, FSEGS Tunisie	Examinateur
VÉRONIQUE MORICEAU	Maître de conférences, Université Paris-Sud	Invitée

---

École doctorale et spécialité :
MITT : Domaine STIC : Intelligence Artificielle
Unité de Recherche :
Institut de Recherche en Informatique de Toulouse
Directrice(s) et co-directrice(s) de Thèse :
Farah BENAMARA ZITOUNE, *Véronique MORICEAU*, *Nathalie AUSSENAC-GILLES* et *Lamia HADRICH BELGUITH*
Rapporteurs :
Patrick PAROUBEK et *Patrice BELLOT*

Dédicaces

Du profond de mon cœur, je dédie ce travail

A

Ma Mère Lilia Baklouti

Mon cher Sami Bouaziz

Pour leur compréhension, leurs nombreux sacrifices consentis, pour leur patience, leur soutien moral, leur grand amour et leur confiance en moi.

Pour tout ce qu'ils ont fait pour mon bonheur et ma réussite.

Je n'oublierai jamais leurs conseils prodigieux qui restent toujours dans mon esprit.

A travers ce modeste travail, je leur manifeste mon amour infini et ma gratitude.

Que dieu leur réserve bonne santé et longue vie.

Remerciements

La présente étude n'aurait pas été possible sans le bienveillant soutien de certaines personnes. Et je ne suis pas non plus capable de dire dans les mots qui conviennent, le rôle qu'elles ont pu jouer à mes côtés pour en arriver là. Cependant, je voudrais les prier d'accueillir ici tous mes sentiments de gratitude qui viennent du fond de mon cœur, en acceptant mes remerciements.

Je tiens à remercier Patrick Paroubek, Ingénieur de Recherche CNRS HDR au LIMSI et Président de l'Association pour le Traitement Automatique des Langues, ainsi que Patrice Bellot, Professeur des Universités en Informatique à Aix-Marseille Université CNRS pour l'honneur qu'ils m'ont fait en acceptant d'être rapporteurs de mon travail et en participant à ce jury. Je les remercie pour leurs remarques constructives.

Je tiens à remercier également Paolo Rosso, Professeur à l'Université Polytechnique de Valence-Espagne ainsi que Lynda Tamine-Lechani, Professeur à l'Université Toulouse III Paul Sabatier et Ahmed Hadj Kacem, Professeur à la Faculté des Sciences Économiques et de Gestion de Sfax-Tunisie pour l'intérêt qu'ils ont porté à mes travaux en examinant ce mémoire et pour l'honneur qu'ils m'ont fait en participant à ce jury.

Je tiens à exprimer toute ma reconnaissance à Farah Benamara Zitoune, Maître de conférences HDR à l'Université Toulouse III Paul Sabatier ainsi que Véronique Moriceau, Maître de conférences à l'Université Paris-Sud pour m'avoir encadrée et dirigée mes recherches. Leurs conseils, leur soutien et leurs encouragements m'étaient une aide précieuse. Sans elles, je n'aurais jamais fini ce travail. Je leur reste reconnaissante pour leur soutien sans faille dans la direction de ce travail.

Je remercie Nathalie Aussenac-Gilles, Directrice de recherche CNRS à l'IRIT Toulouse, pour avoir dirigé mes recherches. Je la remercie également pour la confiance qu'elle m'a témoignée tout au long de ces années et pour tous ses conseils et remarques constructives. Son contact a d'ailleurs été très enrichissant tant au niveau humain qu'au niveau de mon travail. Elle peut être assurée de mon sincère respect et de ma profonde gratitude.

Je remercie de nouveau, Farah, Véronique et Nathalie qui m'ont très bien accueillie dès mon arrivée à Toulouse et qui sont devenues ma famille depuis Novembre 2013. Elles m'ont supportée, écoutée, soutenue dans les moments difficiles. Grâce à elles, je n'ai jamais eu le sentiment d'être étrangère. Je remercie infiniment Farah qui a réussi à être à la fois une directrice de thèse dans le cadre du travail et une grande sœur qui m'a toujours offert le soutien moral.

Résumé

L'analyse des sentiments est un domaine de recherche extrêmement actif en Traitement Automatique des Langues (TAL). En effet, ces dernières années ont vu se multiplier les sources de données textuelles porteuses d'opinion disponibles sur le web : avis d'internautes, forums, réseaux sociaux, enquêtes consommateurs, etc. Devant cette abondance de données, l'automatisation de la synthèse des multiples avis devient cruciale pour obtenir efficacement une vue d'ensemble des opinions sur un sujet donné. L'intérêt de ces données est considérable pour les sociétés qui souhaitent obtenir un retour client sur leurs produits comme pour les personnes souhaitant se renseigner pour un achat, ou un voyage.

Depuis les années 2000, un grand nombre de travaux ont été publiés sur le sujet, faisant de l'extraction d'opinion un domaine très actif dans la recherche en TAL. Globalement, les systèmes actuels ont obtenu de bons résultats sur la classification automatique du caractère subjectif ou objectif d'un document. En revanche, ceux obtenus sur la tâche d'analyse de polarité (qui consiste à classer le document sur une échelle de subjectivité allant du plus positif au plus négatif) restent encore peu concluants. La raison principale de cet échec est l'incapacité des algorithmes actuels à comprendre toutes les subtilités du langage humain, telles que l'usage du langage figuratif. Contrairement au langage littéral, le langage figuratif exploite des dispositifs linguistiques tels que l'ironie, l'humour, le sarcasme, la métaphore et l'analogie qui entrainent une difficulté au niveau de la représentation linguistique ainsi qu'au niveau du traitement automatique du langage figuratif. Dans le cadre de cette thèse, nous nous focalisons sur l'ironie et le sarcasme dans un type particulier de données à savoir les tweets.

Dans ce cadre, nous proposons une approche par apprentissage supervisé afin de prédire si un tweet est ironique ou pas. Pour ce faire, nous avons suivi une démarche en trois étapes. Dans un premier temps, nous nous sommes intéressés à l'analyse des phénomènes pragmatiques utilisés pour exprimer l'ironie en nous inspirant des travaux en linguistique afin de définir un schéma d'annotation multi-niveaux pour l'ironie. Ce schéma d'annotation a été exploité dans le cadre d'une campagne d'annotation d'un corpus formé de 2000 tweets français. Dans une deuxième étape, en exploitant l'ensemble des observations faites sur le corpus annoté, nous avons développé un modèle de détection automatique de l'ironie pour les tweets en français qui exploite à la fois le contexte interne du tweet à travers des traits lexicaux et sémantiques et le contexte externe en recherchant des informations disponibles sur le web. Enfin, dans la troisième étape, nous avons étudié la portabilité du modèle pour la détection de l'ironie dans un cadre multilingue (italien, anglais et arabe). Nous avons ainsi testé la performance du schéma d'annotation proposé sur l'italien et l'anglais et nous avons testé la performance du modèle de détection automatique à base de traits sur la langue arabe. Les résultats obtenus pour cette tâche extrêmement complexe sont très encourageants et sont une piste à explorer pour l'amélioration de la détection de polarité lors de l'analyse de sentiments.

Abstract

Sentiment analysis is an extremely active area of research in Natural Language Processing (NLP). Recently, many textual resources containing opinions are available on the web : Internet user's opinions, forums, social networks, consumer's surveys, etc. Given this data abundance, automating the synthesis of multiple opinions becomes crucial to get an overview of opinions on a given subject. This synthesis is very interesting for companies who want to have an idea about customer feedback on their products as well as for people wishing to inquire about a purchase or a trip.

Since the 2000s, a significant number of papers were published in this field, making sentiment analysis one of the most attractive applications in Natural Language Processing. Overall, current systems have achieved good results on the automatic classification of a document according to the subjective or objective nature. However, those obtained on polarity analysis (which consists in classifying a document on a scale of subjectivity ranging from the most positive to the most negative) are still inconclusive. The main reason for this failure is the inability of current algorithms to understand all the subtleties of human language, such as the use of figurative language. Figurative language makes use of figures of speech to convey non-literal meaning, i.e., meaning that is not strictly the conventional or intended meaning of the individual words in the figurative expression. Figurative language encompasses a variety of phenomena, including irony, humor, sarcasm, metaphor and analogy. Figurative language detection has gained relevance recently, due to its importance for efficient sentiment analysis. This thesis focuses on irony and sarcasm detection in social media.

To this end, we propose a supervised learning approach to predict whether a tweet is ironic or not. For this purpose, we followed a three-step approach. For the first step, drawing on linguistic studies, we investigated the pragmatic phenomena used to express irony to define a multi-level annotation scheme for irony. This annotation scheme was exploited for an annotation campaign to annotate a corpus of 2000 French tweets. In a second step, by exploiting all observations made on the annotated corpus, we developed an automatic detection model for French tweets that exploits internal context of the tweet through lexical and semantic features and external context by looking for available information on the web. Finally, in the third step, we studied the portability of the model for irony detection task in a multilingual corpus (Italian, English and Arabic). We tested the performance of the proposed annotation scheme on Italian and English and we tested the performance of the automatic detection model on the Arabic language. The obtained results for this extremely complex task are very encouraging and might be worth exploring to improve the polarity detection in sentiment analysis.

Table des matières

Liste des tableaux

Table des figures

Introduction

Contexte et motivations

De nos jours, le Web est devenu une source d'information incontournable grâce à la quantité et à la diversité des contenus textuels porteurs d'opinions exprimées par les internautes. Ces contenus sont multiples : blogs, commentaires, forums, réseaux sociaux, réactions ou avis, de plus en plus centralisés par les moteurs de recherche. Devant cette abondance de données et de sources, le développement d'outils pour extraire, synthétiser et comparer les opinions exprimées sur un sujet donné devient crucial. L'intérêt de ce type d'outils est considérable, pour les entreprises qui souhaitent obtenir un retour client sur leurs produits ou leur image de marque comme pour les particuliers souhaitant se renseigner pour un achat, une sortie, ou un voyage. Actuellement, les instituts de sondage s'intéressent à ces outils aussi pour l'évaluation d'un produit sur le marché ou pour prévoir les résultats lors des élections présidentielles par exemple.

C'est dans ce contexte que l'analyse d'opinions (communément appelée *sentiment analysis* ou *opinion mining* en anglais) a vu le jour. Les premiers travaux de recherche en extraction automatique d'opinion remontent à la fin des années 1990 avec en particulier les travaux de Hatzivassiloglou and McKeown (1997) traitant de la détermination de la polarité d'adjectifs, et ceux de Pang et al. (2002), Littman and Turney (2002) sur la classification de documents suivant leur polarité positive ou négative. Depuis les années 2000, un grand nombre de travaux ont été publiés sur le sujet, faisant de l'extraction d'opinion un domaine très actif dans la recherche en Traitement Automatique des Langues (TAL) (Liu, 2015; Benamara *et al.*, 2017b; Benamara, 2017). De nombreuses campagnes d'évaluation sont également consacrées à ce sujet, telles que la campagne TREC (Text REtrieval Conference) (Ounis *et al.*, 2008), la campagne DEFT (Défi Fouille de Textes) pour le français avec une première édition en 2005 (Azé & Roche, 2005) et la campagne SemEval (Semantic Evaluation) avec une première édition en 1998[3].

Globalement, les systèmes actuels ont obtenu de bons résultats sur la tâche d'analyse de subjectivité qui consiste à déterminer si une portion de texte véhicule une opinion (i.e. est

[3] http://www.senseval.org/

subjective) ou ne fait que présenter des faits (i.e. est objective) (Turney, 2002). Par exemple, l'utilisation de lexiques de subjectivité couplés éventuellement à des techniques de classification permet de détecter que l'auteur exprime une opinion positive envers le premier ministre dans la phrase (1) (via l'utilisation de l'adjectif *excellent* de polarité positive).

(1) *Le premier ministre a fait un excellent discours.*

En revanche, les résultats des systèmes d'analyse d'opinions sur la tâche d'analyse de polarité, qui consiste à déterminer la polarité globale et/ou le score de l'opinion effectivement véhiculée par une portion de texte que l'on sait subjective, restent encore peu concluants. Les trois exemples ci-dessous, extrait de (Benamara, 2017), illustre parfaitement la difficulté de la tâche :

(2) *[J'ai acheté un iPhone 5s d'occas il y a trois mois.]$_{P1}$ [La qualité d'image est* ***exceptionnelle****.]$_{P2}$ [Par contre, la protection en verre trempée n'est* ***pas de bonne qualité****]$_{P3}$ [et la batterie* ***m'a lâché au bout de 15 jours****!!]$_{P4}$*

L'exemple (2) contient quatre propositions, délimitées par des crochets. Seules les trois dernières sont porteuses d'opinions (en gras). Parmi ces opinions, les deux premières sont explicites, c'est à dire repérables par des mots, symboles ou expressions subjectives du langage, comme l'adjectif *exceptionnelle*. La dernière est cependant implicite car elle repose sur des mots ou groupes de mots qui décrivent une situation (fait ou état) jugée désirable ou indésirable sur la base de connaissances culturelles et/ou pragmatiques communes à l'émetteur et aux lecteurs.

Les commentaires (3) et (4) ci-dessous, où l'auteur utilise du langage figuratif pour exprimer son opinion, illustrent aussi la difficulté de la tâche d'analyse de polarité. En effet, ces derniers expriment des opinions négatives bien que les auteurs utilisent des mots d'opinion positifs (*adorer, merci, magnifique*).

(3) *J'adore la façon dont votre produit tombe en panne dès que j'en ai besoin.*

(4) *Merci une fois de plus la SNCF. Ça annonce une magnifique journée ça encore.*

Parfois, les opinions implicites peuvent s'exprimer ironiquement, ce qui complique davantage l'analyse de polarité. Dans le tweet (5), extrait du corpus FrIC (Karoui, 2016) construit pendant cette thèse, l'utilisateur emploie une fausse assertion (texte souligné) qui rend de ce fait le message très négatif envers Valls. On remarquera ici le recours au hashtag #ironie qui permet d'aider le lecteur à comprendre que le message est ironique.

(5) *#Valls a appris la mise sur écoute de #Sarkozy en lisant le journal. Heureusement qu'il n'est pas ministre de l'intérieur #ironie*

Il est important de noter que bien que l'extraction des opinions dans ces exemples est d'une simplicité presque enfantine pour un humain, son extraction automatique est extrêmement complexe pour un programme informatique. En effet, au delà de la détermination d'expressions subjectives du langage, le problème de la distinction entre opinions explicites/implicites ou encore l'identification de l'usage du langage figuratif est encore non résolu du fait de l'incapacité des systèmes actuels à appréhender le contexte dans lequel les opinions sont émises.

Dans cette thèse, nous nous proposons de travailler sur la détection automatique du langage figuratif, un phénomène linguistique extrêmement présent dans les messages postés sur les réseaux sociaux. Depuis quelques années, la détection de ce phénomène est devenu un sujet de recherche extrêmement actif en TAL principalement en raison de son importance pour améliorer les performances des systèmes d'analyse d'opinions (Maynard & Greenwood, 2014; Ghosh *et al.*, 2015).

Vers la détection du langage figuratif

Contrairement au langage littéral, le langage figuratif détourne le sens propre pour lui conférer un sens dit figuré ou imagé, comme la métaphore, l'ironie, le sarcasme, la satire et l'humour. L'ironie est un phénomène complexe largement étudié en philosophie et en linguistique (Grice *et al.*, 1975; Sperber & Wilson, 1981; Utsumi, 1996). Globalement, l'ironie est définie comme une figure de rhétorique par laquelle on dit le contraire de ce qu'on veut faire comprendre (cf. exemples 3 et 4). En linguistique computationnelle, l'ironie est un terme générique employé pour désigner un ensemble de phénomènes figuratifs incluant le sarcasme, même si ce dernier s'exprime avec plus d'aigreur et d'agressivité (Clift, 1999).

Chaque type de langage figuratif a ses propres mécanismes linguistiques qui permettent de comprendre le sens figuré. L'inversion de la réalité/vérité pour exprimer l'ironie (Grice *et al.*, 1975), la présence des effets amusants pour exprimer l'humour (van de Gejuchte, 1993; Nadaud & Zagaroli, 2008), etc. Dans la plupart des cas, l'ensemble des phénomènes figuratifs nécessite le recours au contexte de l'énonciation afin que le lecteur ou l'interlocuteur réussisse à interpréter le sens figuré d'un énoncé donné. Par conséquent, il est important de pouvoir inférer des informations au delà des aspects lexicaux, syntaxiques voir même sémantiques d'un texte. Ces inférences peuvent varier selon le profil du locuteur (comme le genre) ou encore son contexte culturel.

La majorité des travaux en détection de l'ironie en TAL concerne des corpus de tweets car les auteurs peuvent explicitement indiquer le caractère ironique de leurs messages en employant des hashtags spécifiques, comme #sarcasme, #ironie, #humour. Ces hashtags sont alors utilisés pour collecter un corpus annoté manuellement, ressource indispensable pour la classification supervisée de tweets comme ironiques ou non ironiques. Les travaux

de l'état de l'art concernent majoritairement des tweets en anglais, mais des travaux existent également pour la détection de l'ironie et/ou du sarcasme pour l'italien, le chinois ou encore le néerlandais (Farias *et al.*, 2015; jie Tang & Chen, 2014; Liebrecht *et al.*, 2013).

Globalement, les approches qui ont été proposées reposent presque exclusivement sur l'exploitation du contenu linguistique du tweet. Deux principales familles d'indices ont été utilisées :

- Indices lexicaux (n-grammes, nombre de mots, présence de mots d'opinion ou d'expressions d'émotions) et/ou stylistiques (présence d'émoticônes, d'interjections, de citations, usage de l'argot, répétition de mots). (Kreuz & Caucci, 2007; Burfoot & Baldwin, 2009; Tsur *et al.*, 2010; Gonzalez-Ibanez *et al.*, 2011; Gianti *et al.*, 2012; Liebrecht *et al.*, 2013; Reyes *et al.*, 2013; Barbieri & Saggion, 2014b)

- Indices pragmatiques afin de capturer le contexte nécessaire pour inférer l'ironie. Ces indices sont cependant extraits du contenu linguistique du message, comme le changement brusque dans les temps des verbes, l'usage de mots sémantiquement éloignés, ou encore l'utilisation de mots fréquents vs. mots rares. (Burfoot & Baldwin, 2009; Reyes *et al.*, 2013; Barbieri & Saggion, 2014b)

Ces approches ont obtenu des résultats encourageants[4]. Nous pensons cependant que ce type d'approche, bien qu'indispensable, n'est qu'une première étape et qu'il est primordial d'aller plus loin en proposant des approches plus pragmatiques qui permettent d'inférer le contexte extra-linguistique nécessaire à la compréhension de ce phénomène complexe.

Contributions

Dans ce cadre, nous nous focalisons pour la première fois sur des tweets en français et proposons une approche par apprentissage supervisé afin de prédire si un tweet est ironique ou pas. Nos contributions peuvent être résumées en trois principaux points.

(1) *Un modèle conceptuel permettant d'appréhender les phénomènes pragmatiques mis en œuvre pour exprimer l'ironie dans les messages postés sur Twitter*. En nous inspirant des travaux en linguistique sur l'ironie, nous proposons le premier schéma d'annotation multi-niveaux pour l'ironie. Ce schéma, publié dans l'atelier ColTal@TALN2016, a été exploité dans le cadre d'une campagne d'annotation d'un corpus formé de 2 000 tweets français (Karoui, 2016). Une version étendue de ce corpus a été utilisée comme données d'entrainement dans le cadre de la première campagne d'évaluation sur l'analyse d'opinion et le langage

[4]Par exemple, Reyes et al. (2013) ont une précision de 79% pour des tweets anglais. Voir chapitre 2 pour un état de l'art détaillé et résultats des approches existantes.

figuratif DEFT@TALN 2017 [5] (Benamara *et al.*, 2017a). Le schéma d'annotation ainsi que les résultats quantitatifs et qualitatifs de la campagne d'annotation sont décrits dans le chapitre 3.

(2) *Un modèle computationnel permettant d'inférer le contexte pragmatique nécessaire à la détection de l'ironie*. En exploitant l'ensemble des observations faites sur le corpus annoté, nous avons développé un modèle de détection automatique de l'ironie dans les tweets en français qui exploite à la fois le contexte interne du tweet à travers des traits lexicaux et sémantiques et le contexte externe en recherchant des informations disponibles dans des ressources externes fiables. Notre modèle permet, en particulier, de détecter l'ironie qui se manifeste par des fausses assertions (cf. exemple (5)). Ce modèle, qui a été publié à TALN 2015 (Karoui *et al.*, 2015b) et ACL 2015 (Karoui *et al.*, 2015a), est présenté dans le chapitre 4.

(3) *Etude de la portabilité à la fois du modèle conceptuel et computationnel pour la détection de l'ironie dans un cadre multilingue*. Nous avons d'abord testé la portabilité de notre schéma d'annotation sur des tweets en italien et en anglais, deux langues indo-européennes culturellement proches du français. Nos résultats, publiés à EACL 2017, montrent que notre schéma s'applique parfaitement sur ces langues (Karoui *et al.*, 2017). Nous avons ensuite testé la portabilité de notre modèle computationnel pour la langue arabe où les tweets sont à la fois écrits en arabe standard et en arabe dialectal. Nos résultats montrent que notre modèle, là encore, se comporte bien face à une famille de langue différente. La portabilité de nos modèles est discutée au chapitre 5.

Avant de détailler nos contributions, nous commençons par présenter dans les deux premiers chapitres de ce manuscrit un état de l'art complet sur les approches linguistiques et computationnelles de détection de l'ironie. En fin de manuscrit, une conclusion synthétise les résultats obtenus et ouvre des pistes de recherches futures.

[5] https://deft.limsi.fr/2017/

Chapitre 1

De l'analyse d'opinion au traitement du langage figuratif

1.1 Introduction

Les premiers travaux en extraction automatique d'opinions remontent à la fin des années 1990 avec, en particulier, les travaux de Hatzivassiloglou and McKeown (1997) traitant de la détermination de la polarité des adjectifs dans les documents, c'est-à-dire la détermination du caractère positif ou négatif de l'opinion véhiculée par les adjectifs, et ceux de Pang et al. (2002) et Littman and Turney (2002) sur la classification de documents suivant leur polarité positive ou négative.

Depuis les années 2000, un grand nombre de travaux ont été publiés sur le sujet, faisant de l'extraction d'opinions l'un des domaines les plus actifs en TAL et en fouille de données, avec plus de 26 000 publications recensées sur Google Scholar. Citons par exemple, les travaux de Wiebe et al. (2005) autour de l'annotation du corpus d'opinion MPQA (Multi-Perspective Question Answering), les travaux de Taboada et al. (2011) concernant la prise en compte des effets des opérateurs sur l'opinion, comme les intensifieurs, les modalités et les négations ainsi que les travaux de Asher et al. (2009) et Chardon et al. (2013) sur la prise en compte de la structure discursive pour le calcul de l'opinion globale d'un document. Notons enfin l'apparition de nombreuses campagnes d'évaluation, telles que : la campagne TREC (Text REtrieval Conference) (Ounis *et al.*, 2008), la campagne DEFT (Défi Fouille de Textes) pour le français avec une première édition en 2005 (Azé & Roche, 2005) et la campagne SemEval (Semantic Evaluation) avec une première édition en 1998[1].

Il est important de noter qu'avant d'être un domaine de recherche en informatique, l'analyse d'opinions a été largement étudiée en linguistique (Hunston & Thompson, 2000), psy-

[1] http://www.senseval.org/

chologie (Davidson *et al.*, 2003), sociologie (Voas, 2014) et en économie (Rick & Loewenstein, 2008). C'est donc un domaine multidisciplinaire nécessitant des outils et techniques diverses comme nous le verrons tout au long de ce chapitre.

Le développement de systèmes d'analyse d'opinions n'est pas simple et nécessite de se confronter à plusieurs difficultés : *comment reconnaître les parties des textes qui renseignent l'utilisateur sur l'opinion qu'il recherche ? Comment évaluer la qualité des opinions qui en ressort : sont-elles plutôt positives, plutôt négatives ? Comment présenter le résultat de manière pertinente à l'utilisateur ?*.

La plupart des approches s'appuient sur une analyse lexicale au niveau du mot, éventuellement couplée à une analyse syntaxique au niveau de la phrase pour repérer les opérateurs et calculer leurs effets sur les mots d'opinion (Liu, 2012). Il est évident que ce type d'analyse est loin d'être suffisant pour tenir compte de toute la complexité langagière de l'expression des opinions. Le recours à une analyse sémantique fine, voire pragmatique, de ces expressions devient une nécessité, spécialement quand il faut traiter des phénomènes complexes comme celui de l'usage du langage figuratif, phénomène qui est au cœur de notre sujet de recherche pour cette thèse.

Ce chapitre a pour but de fournir une brève introduction au domaine d'analyse d'opinion et de poser les principales définitions de la notion du langage figuratif. Il est important de noter que notre objectif n'est pas de fournir un état de l'art exhaustif de ce domaine de recherche très vaste. Les lecteurs intéressés peuvent se référer aux excellentes synthèses de (Liu, 2015) et (Benamara *et al.*, 2017b).

Ce chapitre est organisé comme suit. Nous commençons d'abord en section 1.2 par présenter la notion d'opinion et les principales approches utilisées dans la littérature. Nous détaillons en section 1.3 les principales limites des systèmes actuels en se focalisant sur l'usage du langage figuratif. La section 1.4 est entièrement consacrée à ce type de langage en se focalisant sur quatre phénomènes figuratifs : l'ironie, le sarcasme, la satire et l'humour. Nous terminons ce chapitre par une discussion autour des principaux défis auxquels le TAL doit faire face pour la détection automatique du langage figuratif.

1.2 Définition de la notion d'opinion

1.2.1 Les multiples facettes de l'opinion

En TAL, le mot opinion est un terme générique utilisé pour désigner un ensemble d'expressions subjectives telles que les sentiments, attitudes, points de vues, jugements, désirs, etc. La définition la plus communément admise est la suivante (Benamara, 2017) :

> Une *opinion* est une **expression subjective** du langage qu'utilise un **émetteur** (une personne, une institution, etc.) pour juger ou évaluer un **sujet** (un objet, une

personne, une action, un événement, etc.) en le positionnant sur une **échelle polarisée** d'après une norme sociale (comme un jugement esthétique) ou morale (comme la distinction entre le bien et le mal).

La phrase (1.1) illustre parfaitement cette définition. En effet, l'auteur exprime une opinion positive envers les plats servis dans le restaurant en utilisant un verbe de polarité positive (*adorer*). En analyse d'opinions, il est important de distinguer entre le caractère subjectif ou objectif d'une expression. La phrase (1.2) n'exprime pas une opinion, mais un événement purement factuel.

(1.1) *J'ai adoré les plats servis dans ce restaurant.*

(1.2) *Le premier ministre a inauguré le nouvel hôpital.*

L'élément le plus important dans cette définition est la notion d'échelle polarisée (positif vs. négatif, bien vs. mal, désirable vs. indésirable, accord vs. désaccord, etc.). Ainsi le sentiment de jalousie dans l'exemple (1.3) exprime une émotion et peut apparaître indépendamment de l'opinion évaluative portée à une entité. De même, certaines expressions de prédiction, relevant de l'opinion dans le langage courant, ne constituent pas des évaluations. Ainsi, dans la seconde phrase de (1.4), l'auteur émet une hypothèse sur la météo du soir, sans que cela constitue une évaluation du temps en question.

(1.3) *Je suis jaloux de mon frère.*

(1.4) *Je ne pourrai pas y aller ce soir, je pense qu'il va pleuvoir.*

Dans la suite de cette thèse, nous nous focalisons exclusivement sur la détection automatique d'opinions exprimées sur une **échelle polarisée** ou encore **opinions évaluatives**.

1.2.2 Opinion vue comme un modèle structuré

Dans le cadre de l'extraction automatique, Liu (2012) a proposé un modèle structuré Ω formé de cinq éléments :

- s est **le sujet** de l'opinion.
- a est **un aspect** de s.
- e est **l'émetteur**.

- $senti$ est **le sentiment** exprimé par e envers s (et éventuellement a). $senti$ est généralement représenté par un triplet $(type,\ p,\ v)$ tel que :

  - $type$ est **le type sémantique** du sentiment exprimé. Ce type est défini en fonction de catégories linguistiques ou psycho-linguistiques prédéfinies. Par exemple, dans *Ce film m'a ennuyé*, l'auteur exprime un sentiment d'ennui alors que dans (1.1) l'auteur exprime un jugement d'évaluation.
  - p est **la polarité** qui peut être positive ou négative.
  - v est **la valence** (aussi appelée **force**) qui indique le degré de positivité ou de négativité. La valence est souvent combinée avec la polarité pour obtenir le score de l'opinion. Ainsi, le score associé à l'adjectif *excellent* ($+2$ par exemple) sera supérieur à celui de l'adjectif *bon* ($+1$).

- d est **la date** à laquelle l'opinion a été postée sur internet.

Le but de l'extraction automatique est donc de retrouver dans les textes chaque élément de ce quintuplet. Ce modèle a été conçu pour répondre aux besoins spécifiques des **systèmes d'extraction d'opinions à base d'aspects** (ou encore *feature-based opinion mining systems* en anglais). Ces systèmes, très populaires dans le domaine des commentaires de produits (films, livres, restaurants, ou tout autre produit pouvant se décomposer en parties), ont pour but d'associer chaque opinion extraite $senti$ à un aspect ou partie a. Liu précise que la présence de ces cinq éléments dépend de l'application visée et que certains éléments peuvent être ignorés, comme d ou encore s.

Il est important de noter que bien que l'instantiation du quintuplet $\Omega = (s,\ a,\ senti,\ e,\ d)$ semble simple pour un humain, son extraction automatique est extrêmement complexe pour un programme informatique, principalement à cause de l'incapacité des systèmes actuels à appréhender le contexte dans lequel les opinions sont émises.

Une solution à ce problème consiste à définir l'opinion non pas comme un modèle statique mais comme un modèle dynamique dans lequel chaque élément de Ω dépend de facteurs linguistiques et extra-linguistiques divers, comme la dépendance au domaine, aux opérateurs ou encore au discours au delà de la phrase. Le lecteur peut se référer au nouveau modèle proposé par Benamara et al. (2017b) qui étend celui de Liu (2012) pour prendre en compte la notion de contexte.

1.2.3 Extraction d'opinions : principales approches

En général, les systèmes actuels d'analyse d'opinion se focalisent sur l'extraction d'un ou plusieurs éléments au niveau d'une phrase ou d'un document. Trois principales tâches sont

alors associées à ce processus (Benamara, 2017) :

1. Extraction du sujet et de ses aspects
2. Extraction de l'émetteur
3. Extraction du sentiment. Cette tâche se décompose en deux étapes :
   (a) L'analyse de subjectivité : en déterminant si une portion de texte véhicule une opinion (i.e. est subjective) ou ne fait que présenter des faits (i.e. est objective).
   (b) L'analyse de polarité : en déterminant l'opinion effectivement véhiculée par une portion de texte que l'on sait subjective.

Ces sous-tâches peuvent être réalisées indépendamment les unes des autres ou simultanément. Lorsque les sous-tâches 1 et 2 sont réalisées conjointement, chaque opinion est associée à un couple (sujet, aspect). On parle alors de systèmes d'opinions à base d'aspects.

Dans la plupart des systèmes, les méthodes et techniques employées dans chacune de ces tâches reposent sur quatre hypothèses majeures :

- Les opinions concernent un sujet s unique.
- L'émetteur e est unique.
- Les phrases ou documents analysés sont indépendants les uns des autres.
- Une proposition (voire une phrase) contient au maximum une seule opinion.

Étant données ces hypothèses, ces systèmes se focalisent exclusivement sur l'extraction d'opinions explicites et/ou d'aspects explicites dans une démarche "bottom-up" où le calcul de l'opinion globale du texte dans son ensemble est vu comme un processus d'agrégation des opinions identifiées au niveau local de la proposition ou de la phrase.

Dans la suite de cette section, nous dressons un bref panorama de ces méthodes (pour une présentation détaillée, le lecteur pourra consulter les travaux de Liu (2015), l'ouvrage de référence dans le domaine), en présentant en particulier les travaux publiés dans deux éditions de la campagne **DEFT** (Défi Fouille de Textes) : **DEFT'09** et **DEFT'15**.

La campagne **DEFT'09** comprend, pour le français, l'anglais, et l'italien, une tâche de reconnaissance du caractère subjectif au niveau d'un document. Le corpus utilisé était un corpus d'articles de presse. Parmi les participants, les meilleurs résultats ont été obtenus pour le français et l'anglais par le système de Bestgen et Lories (2009), qui propose une classification SVM standard, basée sur des unigrammes, bigrammes et trigrammes lemmatisés et filtrés par seuil de fréquence. Il est à noter que différents essais d'optimisation

des paramètres n'ont pas montré d'améliorations par rapport à ceux par défaut. Les autres participants ont proposé des approches basées sur l'algorithme des k plus proches voisins (Létourneau & Bélanger, 2009) et sur l'utilisation de lexiques spécialisés comme traits d'apprentissage pour SVM (Toprak & Gurevych, 2009), sans toutefois atteindre les résultats de Bestgen et Lories (2009).

La campagne **DEFT'15** a porté sur la fouille d'opinions et l'analyse des sentiments et des émotions dans les messages postés sur Twitter en relation avec la thématique du changement climatique. Trois tâches ont été proposées : (i) déterminer la polarité globale des tweets, (ii) identifier les classes génériques (opinion, sentiment, émotion, information) et spécifiques (parmi 18 classes) de ces tweets, et (iii) identifier la source, la cible et l'expression porteuse d'opinion, de sentiment ou d'émotion. Douze équipes ont participé. Les meilleurs résultats, en macro-précision, sont de 0,736 (polarité) obtenu par le système proposé par Rouvier et al. (2015), 0,613 (classes génériques) obtenu par le système proposé par Abdaoui et al. (2015) et 0,347 (classes spécifiques) obtenu par le système proposé par Rouvier et al. (2015). Aucun participant n'a soumis de données pour la dernière tâche. Les méthodes utilisées reposent majoritairement sur des approches par apprentissage statistique supervisé (SVM, Naïve Bayes, réseaux neuronaux, PPMC), et utilisent de nombreux lexiques d'opinions (ANEW, Casoar, Emotaix, Feel, Lidilem) et de polarités (Polarimots) comme traits.

L'analyse d'opinion dans les publications sur les réseaux sociaux a joué un rôle majeur dans beaucoup de domaines d'applications tels que : le suivi des résultats des élections américaines en 2009 sur Twitter, Facebook, Google+, Youtube et Instagram (figure 1.1), le suivi des débats politiques en temps réel (figure 1.2) ainsi que la prédiction des résultats des élections présidentielles américaines en novembre 2016 par Google, Twitter et Facebook (figures 1.3 et 1.4), la prédiction de l'état psychologique des personnes sur les réseaux sociaux (Losada & Crestani, 2016), etc.

1.3 Limites des systèmes d'analyse d'opinion

Globalement, les systèmes actuels ont obtenu de bons résultats sur la classification automatique du caractère subjectif ou objectif d'un document (contenant une ou plusieurs phrases) (Section 1.2). En revanche, ceux obtenus sur la tâche d'analyse de polarité (qui consiste à classer le document sur une échelle de subjectivité allant du plus positif au plus négatif) restent encore peu concluants. La raison principale de cet échec est l'incapacité des algorithmes actuels à comprendre toutes les subtilités du langage humain, comme nous le montrons dans les sections suivantes.

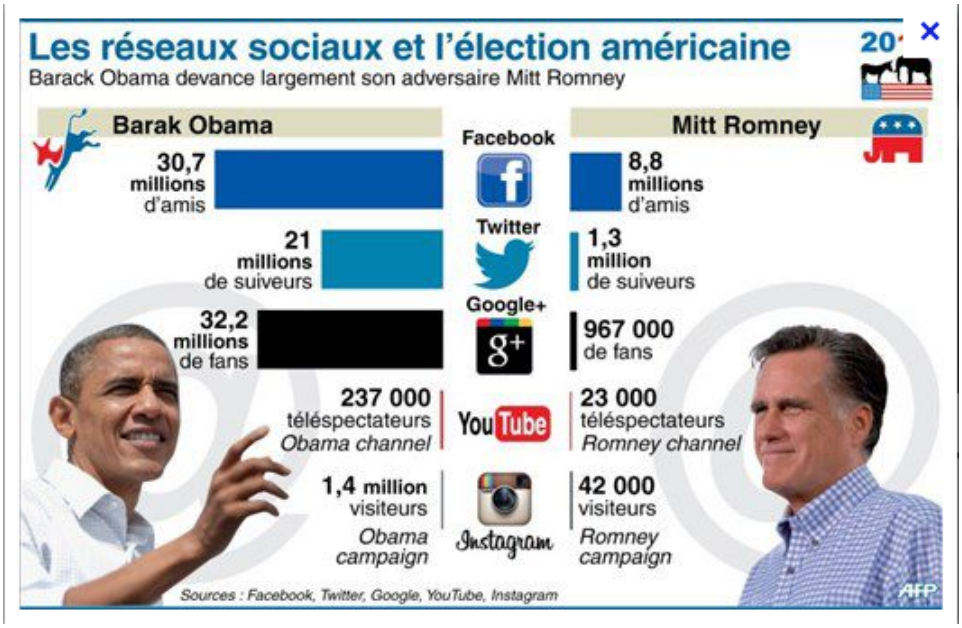

FIGURE 1.1 : Résultats des élections américaines en 2009 par différents réseaux sociaux.

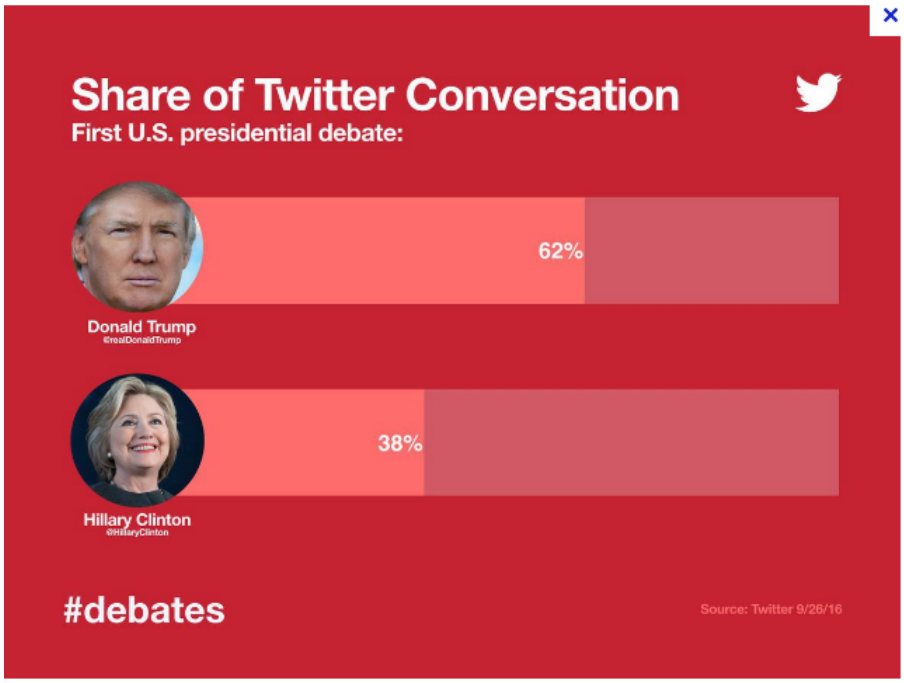

FIGURE 1.2 : Suivi des débats des élections américaines sur Twitter.

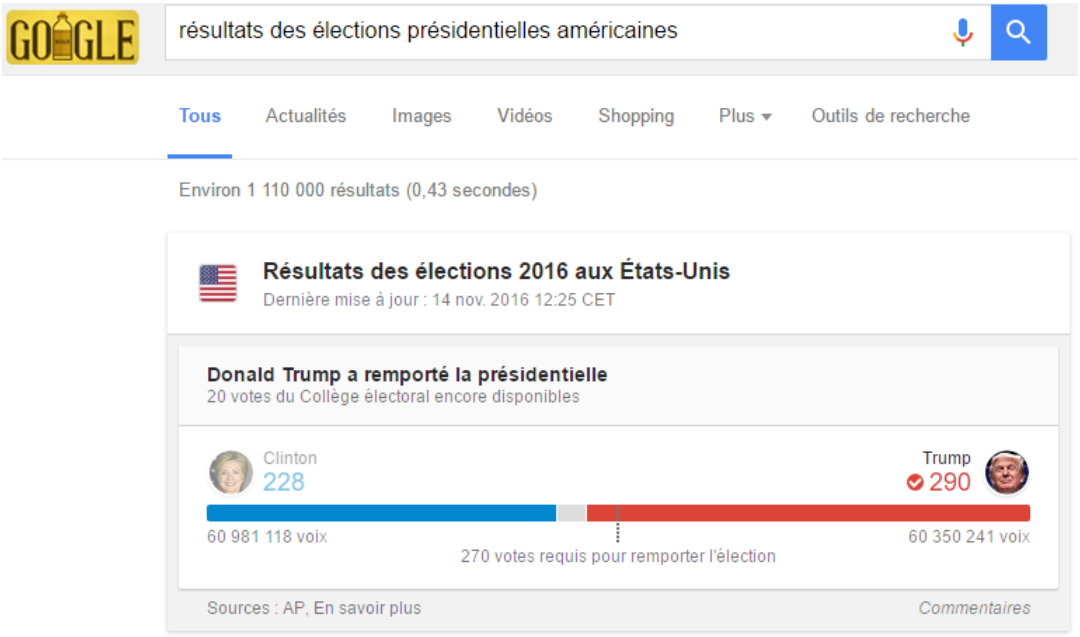

FIGURE 1.3 : Résultats des élections américaines sur Google dès la fin du vote.

1.3.1 Opérateurs d'opinion

Les valeurs de polarité (p) et/ou valence (v), encodées hors contexte dans des lexiques ou des dictionnaires, peuvent être altérées en contexte par la présence d'éléments présents dans la phrase ou le texte. Ces éléments sont appelés **opérateurs**. On distingue trois types majeurs d'opérateurs :

- Les négations, comme *ne...pas*, *jamais*, *rien*, *personne*, etc. Ces derniers ont pour effet de renverser la valeur de p. Cependant, dans certains cas, l'effet peut aussi concerner v. Par exemple, dans *Cet étudiant n'est pas excellent*, l'opinion exprimée n'est pas négative, mais moins intense, tout en restant positive.

- Les intensifieurs, comme *très*, *moins*, *moyennement*, etc., dont l'effet est d'altérer la valeur de v en l'augmentant ou en la diminuant. Ce sont principalement des adverbes. Parfois, la ponctuation, la casse ou la répétition de caractères peuvent avoir le même effet.

- Les modalités, comme *peut être*, *croire*, *devoir*, etc. qui agissent sur la force d'une expression et son degré de certitude. Par exemple, la phrase *Ce restaurant doit être bon*, n'exprime pas une opinion établie. Par contre, dans *Vous devez aller voir ce film*,

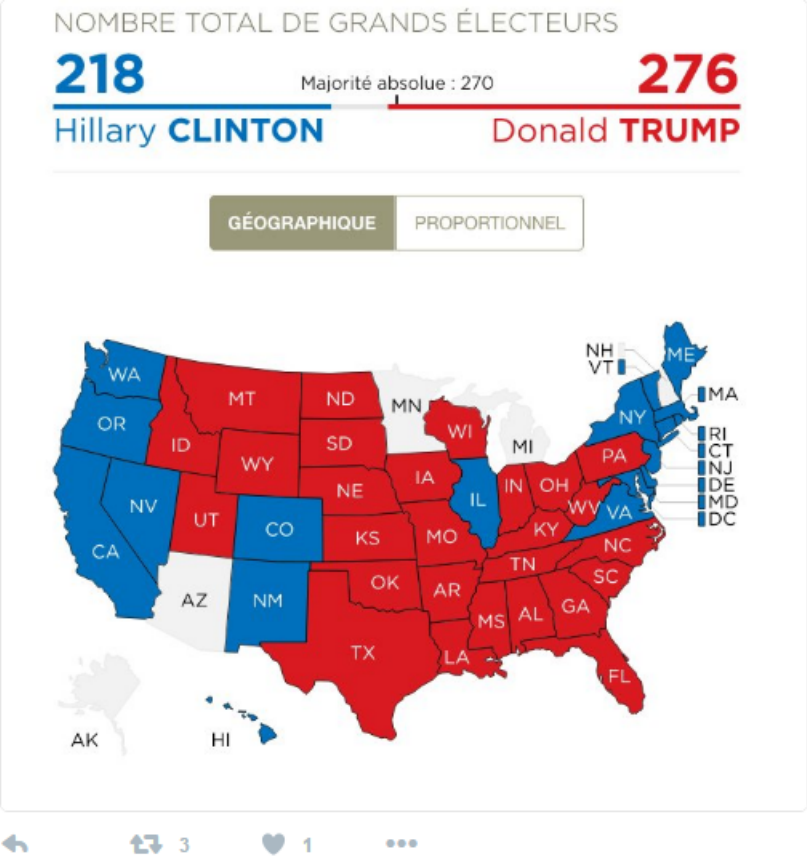

FIGURE 1.4 : Résultats des élections américaines sur Google dès la fin du vote.

la modalité renforce la recommandation.

A l'heure actuelle, la plupart des systèmes traitent les intensifieurs et les négations en tant que phénomène de renversement de polarité (Polanyi & Zaenen, 2006; Shaikh *et al.*, 2007; Choi & Cardie, 2008). Malgré l'importance évidente des modalités dans l'analyse des opinions, elles ne sont pratiquement jamais prises en compte du fait de la difficulté à les traiter automatiquement.

1.3.2 Dépendance au domaine

Un autre facteur qui peut impacter les valeurs de p et v est le **domaine**. Une expression subjective dans un domaine peut être factuelle dans un autre, comme l'adjectif *long* dans la phrase factuelle *Une jupe longue* et la phrase subjective *La durée de vie de ma batterie est*

longue. Même en restant dans un même domaine, la polarité d'une expression peut ne pas être figée. L'opinion dans *Un film horrible* peut être positive pour un film d'horreur mais négative s'il s'agit d'une comédie. Les expressions d'étonnement ou de surprise, comme *Ce film m'a surpris* ont aussi une polarité contextuelle. Enfin, un commentaire comme *Un petit hôtel*, posté sur un site de réservation, n'aura pas la même polarité selon le lecteur.

1.3.3 Opinions implicites

Une opinion peut être explicite ou implicite. Dans le premier cas, l'opinion est repérable par des mots, symboles ou expressions subjectives du langage, comme des adjectifs, adverbes, verbes, noms, interjections ou des émoticônes. Les opinions implicites sont des mots ou groupes de mots qui décrivent une situation (fait ou état) jugée désirable ou indésirable sur la base de connaissances culturelles et/ou pragmatiques communes à l'émetteur et aux lecteurs. Par exemple dans (1.5), il y a trois opinions : les deux premières (soulignées) sont explicites et positives alors que la dernière (en majuscule) est implicite et positive. (1.6) montre également un exemple d'opinion implicite, négative cette fois, extraite d'un site d'avis.

(1.5) *Quel film magnifique. J'étais tellement captivé que* JE N'AI PAS BOUGÉ UNE SECONDE DE MON SIÈGE.

(1.6) *Nous avons acheté en mars ce matelas. Après l'avoir essayé plusieurs jours, surprise :* RÉVEIL DANS UNE CUVETTE CREUSÉE AU COURS DE LA NUIT.

Comparée aux opinions explicites, l'identification d'opinions implicites n'a pas été beaucoup étudiée (voir (Benamara *et al.*, 2017b) pour un état de l'art détaillé des techniques de détection d'opinions implicites). Leur présence dans les textes n'est cependant pas négligeable. Benamara et al. (2016) rapporte que la proportion de ce type d'opinion est d'environ 25% dans un corpus de commentaires de séries télévisées et de 47% lorsqu'il s'agit de réactions à des articles de presse.

1.3.4 Opinion et contexte discursif au delà de la phrase

Le discours est un élément essentiel à la bonne compréhension d'un texte d'opinion car il permet l'analyse des opinions au-delà de la phrase en exploitant les **relations rhétoriques** qui relient les phrases entres elles (comme le contraste, le conditionnel ou l'élaboration). Considérons par exemple le commentaire de série télévisée dans (1.7). Sur les quatre opinions de ce texte, les trois premières sont a priori très négatives. Néanmoins, la dernière phrase, en relation de contraste avec les trois précédentes, nous permet de déterminer la

véritable polarité du document, qui est positive. Une simple moyenne des opinions aurait ici conduit à un contre-sens, et seule la prise en compte de la structure discursive permet de désambiguïser la polarité globale du document.

(1.7) *Les personnages sont antipathiques au possible. Le scénario est complètement absurde. Le décor est visiblement en carton-pâte. Mais c'est tous ces éléments qui font le charme improbable de cette série.*

De même, le conditionnel pour altérer la positivité ou la négativité d'un segment subjectif. Par exemple, dans l'extrait *Si vous n'avez rien à faire de mieux, allez voir le film*, l'opinion négative sera classée en positive par la plupart des systèmes actuels.

Chaque relation de discours à un effet spécifique sur l'opinion. Par exemple, les relations de contraste relient le plus souvent des phrases toutes deux subjectives et de polarités opposées. De même, la relation d'élaboration qui relie deux phrases où la seconde vient préciser ou ajouter de l'information introduite dans la première, préserve en général la polarité (annoncer *Le film est excellent. Les acteurs sont mauvais* n'est pas cohérent discursivement parlant). Le lecteur pourra se référer aux travaux de Benamara et al. (2016) pour une étude statistique sur l'effet de ces relations.

1.3.5 Présence d'expressions figuratives

Parmi toutes les subtilités de langage décrites dans cette section, notre travail porte sur la détection du langage figuratif, et plus particulièrement l'ironie et le sarcasme. La présence de l'un de ces deux phénomènes, dans les tweets par exemple, peut engendrer une mauvaise prédiction de l'opinion globale. Par exemple, le tweet (1.8) peut être classé par un système d'analyse d'opinion comme étant un tweet contenant une opinion positive vu la présence des segments « De mieux en mieux » et « on avance ». Cependant, on voit bien que ce tweet critique la politique du président français François Hollande vis-à-vis du chômage.

(1.8) ***De mieux en mieux***. ***On avance*** *#ironie #France @LeFigaroEmploi : Chômage : le tour de passe-passe de Hollande"*

De même, le tweet (1.9) peut être classé comme étant positif en se référant aux termes « J'adore » et « c'est top » et à l'émoticône positive « :) ». alors qu'il critique en fait les conditions de travail ce qui représente une situation négative.

(1.9) *J'**adore** le taff , manger en 5 minutes et travailler jusqu'à 20h **c'est top** :) #ironie*

Dans les deux exemples de tweets ci-dessus (1.8 et 1.9), on remarque que le recours au hashtag #ironie ainsi qu'aux connaissances culturelles permet aux lecteurs de comprendre que le tweet est ironique.

Nous détaillons dans la section suivante les spécificités du langage figuratif ainsi que ses différentes formes, en nous focalisant sur l'ironie, le sarcasme, la satire et l'humour, en raison de leur emploi fréquent dans les messages postés dans les réseaux sociaux.

1.4 Qu'est-ce que le langage figuratif ?

Contrairement au langage littéral, le langage figuratif détourne le sens propre pour lui conférer un sens dit figuré ou imagé. Le langage figuratif se réfère à une façon d'utiliser la description pour créer une image spéciale et faire ressortir les émotions. Il peut aussi servir d'outil humoristique. Il consiste souvent à faire des comparaisons, en répétant les sons, en exagérant ou en créant un appel aux sens [2].

L'analyse du langage figuratif est un des sujets difficiles auquel le TAL doit faire face. Contrairement au langage littéral, le langage figuratif profite des dispositifs linguistiques, tels que l'ironie, le sarcasme, la satire, l'humour, etc. afin de communiquer des significations plus complexes qui représentent un véritable défi, non seulement pour les ordinateurs, mais aussi pour l'être humain.

Dans ce travail, nous nous focalisons sur plusieurs types de langage figuratif à savoir l'ironie et le sarcasme auxquels nous associons la satire et l'humour qui sont considérés proches de l'ironie. Plusieurs définitions pour ces différents types de langage figuratif ont été avancées. Nous citons, dans ce qui suit, les définitions les plus significatives proposées par les philosophes et les linguistes.

1.4.1 Ironie

L'ironie désigne un décalage entre le discours et la réalité, entre deux réalités ou plus généralement entre deux perspectives, qui produit de l'incongruité. Commençons par reprendre la définition de l'ironie donnée par *le Petit Robert :*

> « Ironie : Manière de se moquer (de quelqu'un ou de quelque chose) en disant le contraire de ce qu'on veut faire entendre. »

En tenant compte de cette définition, Raeber (2011) a distingué deux aspects qui caractérisent l'ironie. Le premier prend en compte *l'effet illocutoire de l'ironie*, c'est-à-dire la

[2] http://www.sp-mc.com/la-definition-du-langage-figuratif/

moquerie/raillerie. Le second considère l'ironie comme une inversion entre ce qui est dit et ce qui est communiqué, autrement dit *une antiphrase*. Ces deux aspects sont de nature très différente. Le premier est d'ordre pragmatique, alors que le second est rhétorique.

Selon Mercier-Leca (2003), les définitions de l'ironie oscillent entre un point de vue restreint et un point de vue élargi. Du point de vue restreint, l'ironie se limite à dire l'inverse de ce que l'on pense, mais cette perspective ne rend pas compte de toutes les formes d'ironie existantes. D'un point de vue élargi, le discours ironique est un discours dans lequel on fait entendre autre chose que ce que disent les mots (et non pas spécifiquement l'inverse).

L'ironie recouvre un ensemble de phénomènes distincts dont les principaux sont **l'ironie verbale** et **l'ironie de situation**. Selon Niogret (2004), l'ironie verbale exprime une contradiction entre la pensée du locuteur et son expression. Elle est créée par le langage. Alors que l'ironie de situation (ou ironie du sort) désigne toute situation qui vient contredire les propos ou les prétentions d'une personne (Niogret, 2004). Plusieurs autres types d'ironie ont été cités dans la littérature sans faire l'objet d'une étude linguistique approfondie ou d'une étude de détection automatique. Parmi ces types d'ironie, citons l'ironie socratique, l'ironie romantique et l'ironie dramatique.

Ces différents types d'ironie peuvent être exprimés à l'écrit ou à l'oral. Par conséquent, les philosophes ainsi que les linguistes ont distingué deux grands genres d'ironie (Tayot, 1984; Didio, 2007) : **l'ironie conversationnelle (ou ironie interactive)** et **l'ironie textuelle**.

L'ironie conversationnelle se manifeste à l'oral dans les conversations ou entretiens langagiers entre au moins deux personnes. Par conséquent, l'ironie conversationnelle se distingue par l'intonation qui est parfois le seul moyen de percevoir l'intention ironique du locuteur, par les mimiques, les gestes et les grimaces (Didio, 2007). Par conséquent, l'ironie conversationnelle est spontanée, instinctive et non programmée à l'avance.

En revanche, **l'ironie textuelle** se manifeste à l'écrit dans des textes littéraires ainsi que dans des textes non littéraires. Didio (2007) précise que *l'ironie textuelle* a un aspect contradictoire à *l'ironie conversationnelle* en se reposant sur le fait que *l'ironie textuelle* est une ironie programmée et travaillée dans ses moindres détails et soigneusement préparée à l'avance. En plus, l'ironie textuelle met en valeur le problème de la communication littéraire dans laquelle l'auteur agit et le lecteur subit. Le style ironique est ainsi une arme appréciée des écrivains engagés. Selon Tayot (1984), ces écrivains ont deux façons de procéder : « soit en mettant en cause un ordre donné afin d'imposer leur point de vue, soit ils opèrent à deux niveaux : (1) ils ébranlent l'ordre des choses pour lui substituer le doute et (2) ils montrent le monde tel qu'il est et tel qu'il pourrait être sans imposer une quelconque idéologie ».

Dans ce qui suit, nous présentons un aperçu du premier type d'ironie défini par Socrate et connu sous le nom « ironie socratique ». Ensuite, nous détaillons les deux principaux types d'ironie : *l'ironie verbale* et *l'ironie de situation* en présentant les différentes théories dans l'ordre chronologique.

A Ironie verbale

Parmi les types d'ironie les moins exploités bien qu'il soit l'un des éléments déclencheur de l'étude de l'ironie, nous citons **l'ironie socratique**. *L'ironie socratique* est une forme d'ironie où l'on feint l'ignorance afin de faire ressortir les lacunes dans le savoir de son interlocuteur. Les études du philosophe Kierkegaard détaillées dans le livre *Le vocabulaire Kierkegaard* (Politis, 2002) précisent que le terme « ironie » est un concept de rhétorique qui provient du grec et qui signifie « ignorance feinte », une technique souvent employée par le philosophe *Socrate*.

> « *L'ironie a un inventeur,* ***Socrate****, et une fonction apparente,* ***la réfutation*** *: elle semble en effet être l'arme rhétorique de celui qui refuse la rhétorique. Face à l'assurance de son interlocuteur, Socrate réclame de pouvoir examiner pas à pas la thèse de son adversaire. Partant toujours d'une affirmation d'ignorance et avançant par une série de questions, le philosophe amène l'adversaire à confirmer ou infirmer des assertions successives, pour identifier finalement le peu qu'il sait réellement.*
> *Contrairement à l'interprétation courante selon laquelle Socrate sait parfaitement ce qu'il feint d'ignorer (et dont dérive le sens moderne de l'ironie : dire le contraire de ce que l'on pense), l'ironie mobilise une véritable suspension de l'opinion. En examinent minutieusement le propos de Socrate lui-même comme celui de son adversaire, l'ironie cherche à renverser les certitudes et les savoirs constituées. Elle force tout discours à s'exposer.* » (Encyclopédie LAROUSSE)

Kerbrat-Orecchioni (1976) a décrit les indices exploitables pour construire et saisir l'ironie d'une séquence verbale et considère l'ironie comme un procédé rhétorique basé sur l'antiphrase. D'après Raeber (2011), cette théorie pose un problème lors de l'étude des cas concrets qui ne présentent aucun renversement du sens encodé. Par conséquent, la critique de la théorie de Kerbrat-Orecchioni par quelques linguistes ne justifie pas l'ignorance de l'existence d'autres types d'ironie par Kerbrat-Orecchioni. En effet, ce dernier les a considérés comme étant une ironie de situation et non pas une ironie verbale car selon lui, un énoncé est considéré comme étant ironique si et seulement si il décrit une contradiction ou un paradoxe.

En parallèle dans les années 1970, Grice (1970) et Grice et al. (1975) ont partagé l'idée de Kerbrat-Orecchioni (1976) et ont considéré que l'ironie verbale devait être traitée comme étant une négation (ou comme une antiphrase). En revanche, Sperber et Wilson (1981) ont traité l'ironie comme étant une interprétation ou une mention échoïque (on fait écho à la parole de quelqu'un en reprenant cette parole, en général pour la moquer ou la critiquer). En comparant les différents travaux des linguistes, on peut dire que la théorie de Grice que celle de Sperber et Wilson représentent deux grandes visions sur la nature de l'ironie verbale,

alors que les autres approches sont jugées, en général, du point de vue de leur proximité ou éloignement théorique par rapport aux théories de Grice ou Sperber et Wilson.

Selon Grice et al. (1975), l'ironie consiste dans l'emploi d'un énoncé qui normalement signifie « p » pour transmettre « non-p ». Dès que la théorie des implicatures conversationnelles (une inférence sur le signifié) a été développée, l'attention de Grice s'est tournée vers la description de l'ironie comme une violation de la plus importante des maximes conversationnelles à savoir la maxime de qualité (ou de vérité) [3]. Cette idée repose sur le fait que l'ironie implique l'expression de quelque chose que le locuteur sait être faux. Ceci a été fortement critiqué par les adeptes de la théorie proposée par *Sperber et Wilson*. Ces derniers trouvent que la théorie de Grice repose uniquement sur *la violation de la maxime de qualité* alors que si l'ironie s'est manifestée avec la violation d'autres maximes que celle de qualité, la théorie de Grice présentera un échec. Malgré cet échec, personne ne peut nier les apports de *la théorie gricéenne* dont le principal est d'avoir fait de l'ironie un phénomène linguistique dont l'interprétation adéquate ne peut se faire qu'avec une prise en compte du contexte d'énonciation.

Dans les années 1980, Sperber et Wilson (1981) ont proposé une théorie qui consiste en l'exploitation du couple *emploi vs. mention*. Ils ont défini l'ironie comme étant une forme spéciale de *mention* par laquelle un locuteur répète une proposition ou une pensée attribuée à quelqu'un d'autre pour ainsi faire connaître à l'interlocuteur son attitude critique vis-à-vis du contenu. D'où la proposition d'une théorie nommée **la théorie de la mention**. Cette théorie a été beaucoup critiquée car elle ne permet pas de distinguer les énoncés échoïques ironiques des simples citations ou discours rapportés. Sperber et Wilson ont alors introduit la notion d'interprétation échoïque : **la théorie échoïque**. Si l'énoncé représente juste l'approbation d'une idée alors on ne peut pas parler d'ironie. En revanche, en prononçant un énoncé, si on attribue des indices qui prouvent la présence de moquerie alors l'énoncé devient ironique. Prenons l'exemple suivant :

(1.10) Personne 1 : Il fait beau aujourd'hui.
Personne 2 : Il fait vraiment beau aujourd'hui !

Afin de juger le sens ironique/non ironique de l'exemple, il faut se référer à la réalité. Par conséquent, si le temps était vraiment beau alors l'énoncé de *personne 2* est jugé comme non ironique. En revanche, si le temps était en réalité mauvais alors le même énoncé devient ironique échoïque à propos de l'énoncé de la *personne 1*. Par conséquent, l'interlocuteur est sensé percevoir la dimension échoïque (source de l'écho) ainsi que le regard que porte le locuteur sur l'énoncé, afin de comprendre l'intention ironique du locuteur.

En parallèle, dans les années 1980, quelques linguistes ont proposé d'autres visions de l'ironie en se référant à Grice, Sperber et Wilson. Parmi eux, Clark et Gerrig (1984) ont

[3] La maxime de qualité interdit de dire ce qu'on croit être faux

proposé **la théorie des faux semblants** qui représente une extension du travail de Grice. Selon eux, si l'écho n'est pas un trait obligatoire de l'ironie alors il faut retrouver toujours la même attitude chez le locuteur : celui-ci feint de tenir un discours auquel en fait il n'adhère pas. Par conséquent, le but du locuteur est de critiquer et de ridiculiser le contenu d'un discours sincère. Donc, la compréhension de l'ironie reviendrait pour un interlocuteur à reconnaitre les différents rôles joués par le locuteur. Cette théorie a été étendue par Kumon-Nakamura et al. (1995) qui ont proposé « la théorie du faux semblant allusif ». Afin de défendre leur théorie, Kumon-Nakamura et al. (1995) se sont référés aux travaux de Kreuz et Glucksberg (1989), qui affirment que l'allusion ne représente pas uniquement une référence à un propos ou un événement passé mais qu'elle exprime une divergence entre « ce qui est dit » et « ce qui aurait dû être dit au regard du contexte ».

Attardo (2000a) définit un énoncé ironique comme étant un énoncé inapproprié au regard du contexte, qui reste néanmoins pertinent dans l'interaction : le sens littéral n'aurait comme fonction que de signaler à l'interlocuteur que le locuteur est ironique, alors que le contexte permet d'inférer entièrement le sens ironique d'un énoncé. Par conséquent, Attardo admet la théorie de Grice en considérant que la violation des différentes maximes conversationnelles provoque l'ironie mais pas seulement. Selon lui, un énoncé est ironique s'il remplit les quatre conditions suivantes :

1. L'énoncé est contextuellement inapproprié.
2. Malgré tout, l'énoncé est pertinent dans la conversation.
3. Le locuteur de l'énoncé a conscience de l'impropriété et l'a produite intentionnellement.
4. Le locuteur suppose qu'une partie au moins de son public reconnaitra les points 2 et 3.

B Ironie de situation

L'ironie de situation, appelée aussi ironie du sort, représente un contraste entre ce que l'on espérait et la réalité observée à l'œil. Elle provoque la surprise chez l'observateur lorsqu'il se trouve avec une situation non prévue. Les images 1.5 et 1.6 illustrent l'aspect contradictoire entre la réalité et l'apparence.

Niogret (2004) a défini l'ironie de situation comme étant une ironie désignant toute situation qui vient contredire les propos ou les prétentions d'une personne. Alors que Lucariello (1994) ainsi que Shelly (2001) ont indiqué que l'ironie de situation n'implique pas l'existence d'une personne qui ironise, mais l'existence d'un observateur se trouvant à l'extérieur d'une situation ou un évènement perçu comme ironique.

FIGURE 1.5 : Exemple d'ironie de situation illustrée par une contradiction dans le texte accompagné par une image. *Et pour bien faire comprendre qu'il n'y a pas de neige, un palmier dessiné sur la piste.*

FIGURE 1.6 : Exemple d'ironie de situation illustré par une contradiction dans une image.

FIGURE 1.7 : Exemple de caricature sarcastique du blogueur « Nawak » sur le site web d'actualités « Yagg.com » [5]

1.4.2 Sarcasme

Selon le dictionnaire Le Grand Robert, le sarcasme est un « énoncé d'une ironie mordante et dédaigneuse ». Le locuteur s'exprime avec aigreur dans le but de blesser en présence de la personne visée (Simédoh, 2012). Par conséquent, le sarcasme souligne une agressivité. Cette agressivité n'empêche pas que le sarcasme puisse comporter des scènes de raillerie, de moquerie. Par conséquent, le sarcasme est considéré comme étant une alliance entre les procédés de l'humour et de l'ironie, mais une ironie blessante, ouvertement moqueuse (voir par exemple la caricature de la figure 1.7).

Didio (2007) ajoute « que le sarcasme est, dans sa première acception, ironie, raillerie acerbe, insultante ; dans la deuxième, trait d'ironie mordante et dans la troisième, figure de rhétorique, ironie cruelle ». De plus, afin de mieux justifier la liste des synonymes donnés pour le sarcasme, Didio (2007) se réfère à la définition du sarcasme donnée par Angenot (1982) : « Le sarcasme consiste à agresser l'adversaire en se montrant en apparence bienveillant, débonnaire, favorable à son égard. La figure apparaît selon l'opposition métalogique élémentaire : bienveillance apparente vs agression dissimulée. Le sarcasme peut consister à compenser un reproche par un éloge fallacieux, qui n'aboutit en fait qu'à aggraver le reproche même. »

Ainsi, le sarcasme est lié à l'agressivité, l'insulte et la méchanceté, des caractéristiques qui ne sont pas attribuées à l'ironie.

[5] http://mensongepourtous.yagg.com

1.4.3 Satire

Selon le dictionnaire Larousse, « la satire est présente dans des écrits, propos, œuvre par lesquels on raille ou on critique vivement quelqu'un ou quelque chose ». La satire ridiculise les travers des personnes et est moralisatrice. Dans son fonctionnement, la satire emploie l'ironie dans son aspect de jugement et de critique, mais elle emploie aussi l'humour pour divertir.

Selon Bautain (1816), « la satire frappe à l'endroit le plus sensible de l'âme ; elle atteint l'amour-propre. La satire représente une matière inépuisable et légitime pour blesser ». Les premières productions satiriques ont vu le jour au XVIIème siècle avec par exemple Les Fables de la Fontaine, *Le Malade imaginaire* de Molière , *Les Satires* de Boileau, etc.

Au XIXème siècle, la presse satirique a vu le jour en Europe dans le cadre de la critique politique dans le but de faire rire le lecteur en donnant une image volontairement déformée de la réalité. Parmi les journaux satiriques existants en France, nous citons : *Le Canard enchaîné* [6], *Charlie Hebdo* [7] et *Le Gorafi* [8]. La figure 1.8 illustre un article satirique publié par Le Gorafi.

1.4.4 Métaphore

La métaphore est une figure de style fondée sur l'analogie. Elle désigne une chose par une autre qui lui ressemble ou partage avec elle une qualité essentielle (Reboul, 1991). La métaphore peut être définie comme une comparaison sans utilisation de mot de comparaison (*comme, ainsi que, ressembler à, semblable à, tel que*, etc.). Par conséquent, le contexte est nécessaire à la compréhension de la métaphore car il permet de décider s'il faut prendre le mot dans son sens ordinaire ou pas. Les linguistes ont défini plusieurs types de métaphore dont *la métaphore annoncée, la métaphore directe*, et *la métaphore filée*.

La métaphore annoncée indique un rapport entre un comparant et un comparé en rapprochant les expressions qui les signifient. Ce type de métaphore est nommé également « métaphore explicite » ou « métaphore par comparaison » (par exemple, *son collègue est une tortue* pour signifier que le collègue est lent). Par contre, *la métaphore directe* compare deux entités ou réalités mais le comparé est absent et est sous-entendu (par exemple, *il travaille avec une tortue* pour signifier qu'il travaille avec un collègue qui est lent).

La *métaphore filée* est constituée d'un enchaînement de comparaisons implicites. Selon Riffaterre (1969), la *métaphore filée* est « une série de métaphores reliées les unes aux autres par la syntaxe - elles font partie de la même phrase ou de la même structure narrative - et par

[6] http://www.lecanardenchaine.fr/
[7] https://charliehebdo.fr/
[8] http://www.legorafi.fr/

FIGURE 1.8 : Exemple d'article de presse satirique publié par Le Gorafi.

le sens : chacune exprime un aspect particulier d'un tout, chose ou concept, que représente la première métaphore de la série ».

1.4.5 Humour

L'humour est considéré par les linguistes comme étant l'un des concepts les plus complexes à comprendre (van de Gejuchte, 1993; Nadaud & Zagaroli, 2008). Ce concept peut être défini par la présence d'effets amusants, tels que le rire ou les sensations de bien-être. L'humour, au sens large, est une forme d'esprit railleuse « qui s'attache à souligner le caractère comique, ridicule, absurde ou insolite de certains aspects de la réalité » (Larousse). Dans son sens strict, l'humour est une nuance du registre comique qui vise « à attirer l'attention, avec détachement, sur les aspects plaisants ou insolites de la réalité ». Toutefois, dans le langage courant, le sens du terme s'est élargi pour désigner le comique, c'est-à-dire l'ensemble des procédés visant à susciter le rire ou le sourire. Il existe principalement 6 formes

FIGURE 1.9 : Exemple de caricature humoristique publiée sur le site web evasion-online.com [10]

de comique : situation, mots, gestes, caractère, mœurs, répétition. L'humour utilise nécessairement une forme de comique, mais toute manifestation comique n'est pas forcément humoristique (exemple figure 1.9).

L'humour a fait l'objet d'études dans des disciplines telles que la philosophie, la linguistique, la psychologie et la sociologie qui ont tenté de définir un ensemble de caractéristique à ce type du langage figuratif. Des études linguistiques ont présenté l'humour à l'aide de modèles sémantiques et pragmatiques. Dans ses travaux, Attardo a défini l'humour comme un phénomène qui suppose la présence de certaines ressources du savoir, telles que le langage, les stratégies narratives, la cible, la situation, les mécanismes logiques pour produire un effet drôle (Attardo, 1994; Attardo, 2001). Du point de vue des sociologues (Hertzler, 1970), l'étude du contexte culturel est primordiale dans le compréhension de l'humour.

1.5 Traitement automatique du langage figuratif : un défi pour le TAL

En analysant finement les différentes définitions proposées par les linguistes, philosophes, psychologues et sociologues pour caractériser l'ironie, le sarcasme ou l'humour (cf. section précédente), il en ressort clairement que l'interprétation de ces phénomènes requiert une connaissance du contexte de l'énonciation. Ce contexte est relativement facile à retrouver par un humain dans le cadre d'un poème ou d'un texte long extrait d'un roman ou d'un livre. Par contre, ce contexte est plus difficile à identifier si le texte est court.

[10] http://evasion-online.com/tag/humour

Notre objectif étant l'identification du langage figuratif dans des textes courts postés sur Twitter, les questions suivantes se posent alors :

- *Les formes figuratives identifiées dans les textes littéraires sont-elles aussi employées dans des textes courts ?*
- *Existe-il des indices linguistiques qui permettent d'inférer l'ironie dans les textes courts ?*
- *Si oui, sont-ils suffisants ? sont-ils indépendants de la langue ?*
- *Si non, comment peut-on inférer le contexte nécessaire à la compréhension d'une forme non littérale d'un texte court ?*
- *Comment ces différents indices (linguistiques et contextuels) peuvent-ils être modélisés dans un système automatique ?*

Dans cette thèse, nous proposons d'apporter les réponses à chacune de ces questions en se focalisant sur l'ironie verbale exprimée dans des tweets. La frontière entre les différentes formes de langage figuratif présentées dans la section précédentes étant floues, nous considérons dans la suite de ce manuscrit le terme *ironie* comme un terme générique englobant l'ironie et le sarcasme. Nos contributions seront présentées dans les chapitres 3, 4 et 5.

1.6 Conclusion

Notre objectif dans le cadre de cette thèse est de proposer une approche pour la détection automatique de l'ironie dans les contenus générés par les utilisateurs sur le web et plus précisément les tweets en français avec des perspectives multilingues. Pour cela, nous avons présenté dans ce chapitre le domaine de l'analyse d'opinion d'une manière générale ainsi que les limites des systèmes d'analyse d'opinion. Nous avons également présenté les définitions données par les philosophes et les linguistes pour quelques formes de langage figuratif à savoir : l'ironie, le sarcasme, la satire, la métaphore et l'humour. Nous nous sommes focalisés sur l'ironie verbale car elle représente notre sujet d'intérêt dans cette thèse.

Nous présentons dans le chapitre 2 un état de l'art sur les différents travaux computationnels traitant le langage figuratif et l'ironie en particulier ainsi que les différents schémas d'annotation proposés pour l'annotation de ce phénomène.

Chapitre 2

Vers la détection automatique du langage figuratif

2.1 Introduction

Comme nous l'avons vu dans le chapitre précédent, l'ironie est un phénomène linguistique complexe largement étudié en philosophie et en linguistique (Grice *et al.*, 1975; Sperber & Wilson, 1981; Utsumi, 1996). Même si les théories différent au niveau de la définition de l'ironie, elles s'accordent sur le fait que l'ironie implique une incongruité entre ce qui est dit et la réalité. En regardant les différences entre les approches, l'ironie peut être définie comme une incongruité entre le sens littéral d'un énoncé et son sens voulu. La recherche d'un sens non-littéral commence lorsque l'auditeur se rend compte que l'énoncé du locuteur ne parvient pas à donner un sens par rapport au contexte (Grice *et al.*, 1975; Searle, 1979; Attardo, 2000a). Dans la plupart des travaux, l'étude de l'ironie se chevauche avec d'autres formes du langage figuratif tels que l'humour, la satire, la parodie, et le sarcasme (Clark & Gerrig, 1984; Gibbs, 2000). La distinction entre ces différentes formes du langage figuratif et en particulier la distinction entre ironie et sarcasme reste très compliquée et délicate. Cette difficulté s'explique par une frontière floue entre ces notions au niveau linguistique ainsi que par la complexité de la tâche de différenciation entre ces notions dans un texte au niveau computationnel.

Les théories discutées dans le chapitre 1 ont inspiré la plupart des indices ou traits utilisés pour la détection automatique. L'étude de l'état de l'art dans ce domaine montre que contrairement à la métaphore et à l'humour, l'ironie et le sarcasme sont les formes d'expressions figuratives les plus étudiées. La raison principale est bien évidemment l'importance de ces formes pour une analyse efficace des opinions et sentiments (cf. chapitre 1, section 1.3).

La majorité de travaux en TAL se sont focalisés sur les textes d'opinions, comme des

avis de consommateurs ou des textes courts issus de réseaux sociaux comme Twitter. En général, les avis de consommateurs négatifs sont supposés avoir plus de chance de contenir des expressions ironiques (Tsur *et al.*, 2010) ce qui est évidemment discutable. Dans les tweets, les messages associés aux hashtags *#sarcasme, #ironie,* ou *#satire* sont considérés comme ironiques ou sarcastiques. Ainsi grâce aux hashtags, il est relativement facile de recueillir des ensembles de données ironiques et/ou sarcastiques. Parfois, la pré-annotation binaire (ironique/non ironique) est augmentée par une annotation manuelle au niveau de l'opinion voire même au niveau de phénomènes plus pragmatiques.

Les hashtags figuratifs présents dans les tweets sont alors utilisés comme une étiquette de référence pour la détection automatique dans un cadre d'apprentissage supervisé. L'apprentissage s'appuie sur trois groupes de traits :

1. Traits surfaciques (ponctuations, émoticônes, etc.) et lexicaux (polarité de l'opinion véhiculée, le type d'émotion exprimée, etc.)
2. Traits pragmatiques qui capturent le contexte interne du message en utilisant exclusivement son contenu linguistique, comme l'usage de mots sémantiquement opposés.
3. Traits pragmatiques qui capturent le contexte externe du message en utilisant des connaissances extra-linguistiques, comme les fils de discussions ou encore le profil utilisateur.

Dans ce chapitre, nous dressons un panorama des principaux travaux de l'état de l'art sur la détection du langage figuratif en se focalisant d'une part sur les corpus utilisés et les schémas d'annotation proposés pour annoter ces corpus (cf. section 2.2), et d'autre part sur les méthodes mises en œuvre pour la détection automatique. Nous présentons non seulement les travaux sur la détection de l'ironie, du sarcasme et de la satire (cf. section 4.5), mais aussi les travaux sur la détection d'autres formes d'expressions figuratives, comme la métaphore (cf. section 2.4), la comparaison (cf. section 2.5) et l'humour (cf. section 2.6). Pour chacune de ces formes, nous détaillons les approches selon trois axes suivant qu'elles se basent sur l'un ou l'autre des trois ensembles de traits présentés plus haut. Nous terminons ce chapitre par un bilan qui permet de positionner nos travaux et apprécier nos contributions.

2.2 Principaux corpus existants pour le langage figuratif

La plupart des travaux exploitent les hashtags sans nécessairement recourir à des annotations manuelles[1]. Par exemple, Gonzalez-Ibanez et al., (2011) présentent un corpus en

[1]Lorsqu'elles sont effectuées, les annotations manuelles concernent un petit échantillon du corpus afin de juger de la fiabilité des hashtags.

anglais composé de 900 tweets et divisé en 3 catégories en fonction de leurs hashtags : sarcasme (*#sarcasm, #sarcastic*), un sentiment positif direct (*#happy, #joy, #lucky*), ou un sentiment négatif direct (*#sadness, #angry, #frustrated*). Reyes et al. (2013) ont construit un corpus formé de 40 000 tweets en anglais contenant *#irony*, *#education, #humor*, et *#politics*. Le corpus a été divisé en quatre parties qui contiennent chacune 10 000 tweets. La première partie est ironique (tweets contenant *#irony*) alors que les trois autres parties sont considérées comme étant non ironiques (tweets contenant *#education, #humor, #politics*).

Une approche similaire a été utilisée par Liebrecht et al. (2013) pour la collecte d'un corpus de tweets ironiques en néerlandais. Le corpus collecté est formé de deux sous-corpus. Le premier contient 77 948 tweets collectés à partir d'une base de données fournies par le centre d'e-Science néerlandais et publiés à partir de décembre 2010, la collecte a été effectuée en utilisant le hashtag *#sarcasme*. Le deuxième sous-corpus est formé de 3,3 millions de tweets publiés le 1er février 2013. Ce dernier contient 135 tweets avec *#sarcasme*.

En plus de l'annotation en ironique/non ironique basée sur les hashtags, d'autres travaux ont proposé d'annoter des informations différentes. Parmi eux, nous citons le schéma d'annotation du corpus de tweets en italien Senti-TUT (Gianti *et al.*, 2012) qui propose d'analyser l'impact de l'ironie dans l'expression des sentiments et des émotions. Les annotateurs, au nombre de trois, avait pour tâche de classer chaque tweet en cinq catégories mutuellement exclusives : *POS (positive), NEG (négatif), HUM (ironique), MIXTES (POS et NEG), et NONE (objectif)*. Citons également les travaux de Van Hee et al. (2015) qui s'intéressent à des formes spécifiques d'ironies dans les tweets en anglais et en néerlandais, à savoir : *ironique par opposition, ironique à travers l'hyperbole, ironique par euphémisme, peut-être ironique*, et *non ironique*.

Notre objectif étant une analyse fine des expressions ironiques en corpus, nous avons choisi dans cette section de présenter les principaux schémas d'annotation existants en nous focalisant exclusivement sur les approches qui vont au-delà de la simple pré-annotation binaire ironique vs. non ironique. Nous nous intéressons également aux corpus construits pour l'annotation d'expressions métaphoriques, car ces dernières sont considérées dans certains cas comme étant un des marqueurs de l'ironie (section A.1.1 du chapitre 3). Pour chacun de ces corpus, nous détaillons les phases de collecte des données, d'annotation manuelle ainsi que les résultats de la campagne d'annotation.

2.2.1 Corpus annotés en ironie/sarcasme

A Senti-TUT : un corpus de tweets en italien

Gianti et al. (2012) ont mené la première campagne d'annotation pour l'ironie dans le cadre du projet *Senti-TUT* [2] qui avait pour objectif le développement d'une ressource italienne et

[2] www.di.unito.it/ tutreeb/sentiTUT.html

l'étude de l'expression de l'ironie dans les réseaux sociaux. Dans ce qui suit, nous détaillons le processus d'annotation suivi dans le cadre de ce projet.

Collecte du corpus. Le corpus Senti-TUT est composé de deux sous-corpus de tweets politiques nommés *TWNews et TWSpino*. Les auteurs justifient le choix du domaine politique par le fait qu'il est considéré comme le domaine où l'ironie est fréquemment utilisée par les humains.

Le corpus *TWNews* a été collecté en appliquant des filtres basés sur le temps et les métadonnées afin de sélectionner des messages qui représentent une variété d'opinion politique. Ils ont utilisé *Blogometer* [3] qui exploite l'API de Twitter pour collecter les tweets publiés pendant les élections italiennes pour la période allant du 6 octobre 2011 au 3 février 2012, période pendant laquelle Mario Monti a remplacé Silvio Berluscconi comme premier ministre. Pour la collecte de ces tweets, les auteurs ont exploité la liste de mots-clés/hashtags suivants : *mario monti / #monti, governo monti /#monti, et professor monti / #monti* (en minuscule et majuscule). Ceci a permis la collecte de 19 000 tweets. 8 000 re-tweets ont ensuite été supprimés. L'ensemble des tweets restants ont été filtrés de nouveau par des annotateurs humains qui ont jugé 70% des tweets comme étant mal écrits, en double ou incompréhensibles vu l'absence de contexte. Après ces deux étapes de filtrage, le corpus obtenu est finalement formé de 3 288 tweets.

Le corpus *TWSpino* est formé de 1 159 tweets collectés à partir de la section *Twitter* de *Spinoza* [4], un blog italien très populaire contenant des messages politiques satiriques. Ces tweets ont été sélectionnés à partir des tweets publiés dans la période allant de juillet 2009 à février 2012. Les tweets contenant de la publicité, qui représentent 1,5% de l'ensemble collecté, ont été supprimés.

Annotation des tweets. Un schéma d'annotation à deux nivaux a été proposé qui permet d'annoter un tweet à la fois au niveau : (1) de la polarité globale et (2) de la morphologie et la syntaxe (Gianti *et al.*, 2012). Cinq catégories d'annotation ont été utilisées, comme détaillé ci-dessous. Pour chaque catégorie, nous donnons un exemple en italien extrait de Senti-TUT ainsi que sa traduction en anglais :

- Pos (positive) : l'opinion globale exprimée dans le tweet est positive, cf. exemple (2.1).

  (2.1) Marc Lazar : "Napolitano ? L'Europa lo ammira. Mario Monti ? Può salvare l'Italia."

[3] www.blogmeter.eu
[4] www.spinoza.it.

(Marc Lazar : "Napolitano ? Europe admires him. Mario Monti ? He can save Italy.")

- Neg (negative) : l'opinion globale exprimée dans le tweet est négative cf. exemple (2.2).

  (2.2) Monti è un uomo dei poteri che stanno affondando il nostro paese.
  (Monti is a man of the powers that are sinking our country.)

- Hum (ironic) : le tweet est ironique, cf. exemple (2.3).

  (2.3) Siamo sull 'orlo del precipizio, ma con me faremo un passo avanti (Mario Monti).
  (We're on the cliff's edge, but with me we will make a great leap forward (Mario Monti).)

- Mixed (Pos and Neg both) : le tweet est à la fois positif et négatif, cf. exemple (2.4).

  (2.4) Brindo alle dimissioni di Berlusconi ma sul governo Monti non mi faccio illusioni
  (I drink a toast to Berlusconi's resignation, but I have no illusion about Monti's government)

- None : le tweet n'est ni positif, ni négatif et ni ironique, cf. exemple (2.5).

  (2.5) Mar io Monti premier ? Tutte le indiscrezioni.
  (Mario Monti premier ? All the gossip.)

L'annotation a été réalisée par 5 annotateurs humains. Une première campagne d'annotation pour un sous-ensemble du corpus formé de 200 tweets a permis de valider les étiquettes, un accord inter-annotateurs de $k = 0,65$ (kappa de Cohen) a été observé. Une deuxième étape d'annotation a été faite pour 25% des tweets où les annotateurs ne sont pas d'accord. Après cette étape, 2% des tweets ont été jugés comme étant ambiguës et ont été par conséquent écartés du corpus. Le corpus final est donc formé de 3 288 tweets du corpus *TWNews*.

Analyse des résultats de la phase d'annotation. Après avoir achevé l'étape d'annotation, une tâche d'analyse de la campagne d'annotation manuelle a été effectuée. A ce niveau, deux hypothèses ont été testées : (H1) l'inversion de polarité est un indice d'ironie, et (H2) les expressions d'émotions sont très présentes dans les tweets ironiques.

Les annotations montrent que différents types d'émotions sont présents dans les corpus. Dans le corpus *TWNews-Hum*, les émotions les plus courantes sont la joie et la tristesse conceptualisée en termes de polarité inversée. Une plus grande variété de typologies d'ironie a été observée, à savoir : les tweets sarcastiques visant à blesser leur cible, et les tweets humoristiques (plutôt que d'invoquer une attitude négative, ceux-ci ont tendance à produire un effet comique ou parodique). En revanche, dans le corpus *TWSpino*, les émotions détectées sont la plupart du temps négatives, et les typologies d'ironie exprimées sont plus homogènes et sont principalement limitées au sarcasme et la satire politique. Cela pourrait être lié au fait que les messages de *TWSpino* sont sélectionnés et révisés par une équipe de rédaction. En outre, les éditeurs de *TWSpino* caractérisent explicitement le blog comme satirique.

Les différentes analyses effectuées dans ce projet montrent que l'ironie est souvent utilisée en conjonction avec une déclaration apparemment positive pour refléter une valeur négative, mais rarement l'inverse. Ceci est en accord avec les études théoriques qui ont noté que l'expression d'une attitude positive dans un mode négatif est rare et plus difficile pour l'être humain à traiter, par rapport à l'expression d'une attitude négative dans un mode positif.

Nous montrons au chapitre 5, comment une partie du corpus Senti-TUT a été utilisée dans le cadre de l'étude de la portabilité de notre approche à d'autres langues indo-européennes.

B Corpus de tweets en anglais et néerlandais

Un schéma d'annotation a été proposé par Hee et al. (2016) pour l'annotation d'un corpus de tweets en anglais et néerlandais. Toutes les annotations ont été effectuées à l'aide de l'outil d'annotation "Brat rapid annotation tool" [5] (Stenetorp *et al.*, 2012).

Collecte du corpus. Hee et al. (2016) ont collecté un corpus de 3 000 tweets en anglais et 3 179 tweets en néerlandais en utilisant l'API de Twitter. Les deux corpus ironiques ont été recueillis avec les hashtags *#irony, #sarcasm* (*#ironie* et *#sarcasme* en néerlandais) et *#not*.

Annotation des tweets. Le schéma d'annotation proposé permet : i) d'identifier les tweets ironiques dans lesquels se produit un changement de polarité et ii) d'indiquer les segments du texte en contradiction qui ont permis de repérer l'ironie. Le schéma proposé comporte trois étapes d'annotation :

1. Indiquer si un tweet est :

[5] http://brat.nlplab.org/

- *Ironique par opposition :* le texte du tweet exprime une polarité littérale (polarité exprimée explicitement dans le texte) qui est l'opposée de la polarité attendue (polarité qui représente la réalité selon le contexte). Par exemple le tweet exprime une opinion positive alors que d'après le contexte, la polarité doit être négative, comme le montrent les exemples (2.6) et (2.7).

  (2.6) Exams start tomorrow. Yay, can't wait !

  (2.7) My little brother is absolutely awesome ! #not.

- *Ironique par hyperbole :* le texte du tweet exprime une polarité littérale plus forte que la polarité attendue, cf. exemple (2.8).

  (2.8) 58 degrees and a few sunbeams breaking through the clouds. Now could the weather be any better for a picnic ?

- *Ironique par euphémisme :* le texte du tweet exprime une polarité littérale moins forte que la polarité attendue, cf. exemple (2.9).

  (2.9) A+ ? So you did quite well..

- *Peut-être ironique :* il n'y a pas de différence entre la polarité littérale et attendue. Cependant, le texte contient une autre forme d'ironie (comme l'ironie de situation dans l'exemple (2.10))[6].

  (2.10) Just saw a non-smoking sign in the lobby of a tobacco company #irony

- *Non ironique :* le tweet est non ironique, cf. exemple (2.11).

  (2.11) Drinking a cup of tea in the morning sun, lovely !

2. Si le tweet est ironique :

   - Indiquer si un hashtag ironique (par exemple, *#not, #sarcasm, #irony*) est nécessaire pour comprendre l'ironie.
   - Indiquer le degré de difficulté de la compréhension de l'ironie sur une échelle de 0 à 1.

3. Annoter les segments en contradiction.

La figure 2.1 illustre un exemple de tweet annoté comme étant *ironique par opposition* selon le schéma d'annotation détaillé ci-dessus. Dans cet exemple, les annotateurs ont considéré que l'expression « cannot wait » comme une évaluation littéralement positive (intensifiée par un point d'exclamation) qui est contrastée par l'acte d'aller chez le dentiste, qui est stéréo-typiquement perçu comme désagréable et implique donc implicitement un sentiment négatif.

[6] Voir chapitre 1, section B pour une définition de l'ironie de situation.

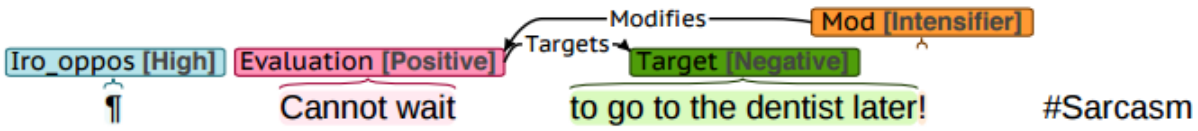

FIGURE 2.1 : Exemple de tweet ironique par opposition (Hee *et al.*, 2016).

Résultats de la procédure d'annotation. La procédure d'annotation a montré que 57% des tweets anglais ont été annotés comme étant ironiques, 24% comme étant peut-être ironiques et 19% comme étant non ironiques. Pour le corpus néerlandais, 74% des tweets ont été annotés comme étant ironiques, 20% comme étant peut-être ironiques et 6% comme étant non ironiques. La nécessité de la présence des hashtags ironiques afin de comprendre le sens ironique d'un tweet a été presque la même pour les deux corpus anglais (52%) et néerlandais (53%). En outre, la distribution des tweets avec une polarité positive ou négative est la même pour les deux langues avec une forte présence de tweets positifs. En regardant les déclencheurs d'ironie (opposition, hyperbole ou euphémisme), l'annotation des deux corpus a prouvé que dans 99% des cas, l'ironie a été exprimée par une opposition et seulement 1% des tweets ironiques sont exprimés par une hyperbole ou un euphémisme.

Nous montrons dans le chapitre 3 que l'ironie peut être exprimée dans les réseaux sociaux par d'autres mécanismes pragmatiques, autres que l'hyperbole ou l'euphémisme.

2.2.2 Corpus annotés en métaphore

Shutova et al. (2013) ont proposé un schéma d'annotation pour la métaphore exprimée par des verbes qui consiste à identifier des concepts métaphoriques et à mettre en relation ces concepts selon une relation source-cible.

Collecte du corpus. La campagne d'annotation a été réalisée sur un ensemble de textes tirés du "British National Corpus" (BNC). Le BNC est un corpus de 100 millions de mots contenant des échantillons de l'anglais britannique de la seconde moitié du XXe siècle (90% de l'écrit) et (10% de l'oral). Le corpus collecté comporte des échantillons de différents genres présents dans le BNC à savoir : de la fiction (5 293 mots), des journaux (2 086 mots) et articles de revues (1 485 mots), des essais sur la politique, relations internationales et sociologie (2 950 mots) ou encore des émissions de radio (1 828 mots, issus de la retranscription de parole). Le corpus d'étude comporte un nombre total de 13 642 mots.

Schéma d'annotation. Pour chaque genre de corpus du BNC, les annotateurs ont été invités à : (1) classer chaque verbe selon qu'il véhicule un sens métaphorique ou littéral, et (2) identifier les correspondances de domaine source-cible pour les verbes qui ont été marqués comme métaphoriques. Deux listes de catégories décrivant la source et la cible des concepts ont été mises à la disposition des annotateurs (Tableau 2.1). Les annotateurs devaient choisir une paire de catégories à partir de ces listes qui décrivent le mieux la correspondance métaphorique. Parallèlement à cela, les annotateurs ont été autorisés à introduire de nouvelles catégories s'ils n'arrivaient pas à choisir une catégorie parmi la liste prédéfinie.

Source concepts	**Target concepts**
PHYSICAL IBJECT	LIFE
LIVING BEING	DEATH
ADVERSARY/ENEMY	TIME/MOMENT IN TIME
LOCATION	FUTURE
DISTANCE	PAST
CONTAINER	CHANGE
PATH	PROGRESS/EVOLUTION/DEVELOPMENT
PHYSICAL OBSTACLE (e.g. BARRIER)	SUCCESS/ACCOMPLISHMENT
DIRECTIONALITY : (e.g. UP/DOWN)	CAREER
BASIS/PLATFORM	FEELINGS/EMOTIONS
DEPTH	ATTITUDES/VIEWS
GROWTH/RISE	MIND
SIZE	IDEAS
MOTION	KNOWLEDGE
JOURNEY	PROBLEM
VEHICULE	TASK/DUTY/RESPONSIBILITY
MACHINE/MECHANISM	VALUE
STORY	WELL-BEING
LIQUID	SOCIAL/ECONOMIC/POLITICAL SYSTEM
POSSESSIONS	RELATIONSHIP
INFECTION	-
VISION	-

TABLE 2.1 : Concepts sources et cibles suggérés pour l'annotation de la métaphore selon (Shutova *et al.*, 2013).

L'exemple (2.12) illustre la procédure d'annotation :

(2.12) If he **asked** her to **post** a letter or **buy** some razor blades from the chemist, she was **transported** with pleasure.

Selon les auteurs, les trois premiers verbes sont utilisés dans leur sens littéral (*demander (ask)*, *envoyer (post)* et *acheter (buy)*), alors que le quatrième verbe *transporter (trans-*

ported)" est employé dans un sens figuré (signifiant ici *"une personne est transportée par un sentiment"* et non pas *"transportée par un véhicule"*). Par conséquent, l'utilisation de *transporter* dans cet exemple est métaphorique d'où le mapping conceptuel EMOTIONS-VEHICULES.

Tâches	Kappa (κ)	Nombre de catégories (n)	Nombre d'instances annotées (N)	Nombre d'annotateurs (k)
Identification des verbes	0,64	2	142	3
Attribution des concepts métaphoriques	0,57	26	60	2
Choix des catégories cibles	0,60	14	30	2
Choix des catégories sources	0,54	12	30	2

TABLE 2.2 : Accord inter-annotateurs pour l'annotation du métaphore. (Shutova *et al.*, 2013)

Résultats de la campagne d'annotation. Comme le montre le tableau 2.2, l'identification des verbes métaphoriques a donné un Kappa de Cohen de $\kappa = 0,64$. Ce niveau d'accord est considéré comme substantiel. La mesure de l'accord dans la seconde tâche est apparue moins évidente. L'accord global résultant de l'attribution des concepts métaphoriques est de $\kappa = 0,57$ alors que l'accord a été plus fort sur le choix des catégories cibles ($\kappa = 0,60$) que les catégories sources ($\kappa = 0,54$).

L'étude des cas de désaccord entre les annotateurs a montré que la présence de catégories qui se chevauchent partiellement dans la liste des concepts cibles est la principale cause d'erreur. Par exemple, les catégories PROGRESS et SUCCESS, ou VIEWS, IDEAS et METHODS ont été souvent confondues. Par conséquent, une fusion de ces catégories a été effectuée afin d'avoir une annotation cohérente. Suite à cette fusion, l'accord inter-annotateurs s'est amélioré ($\kappa = 0,61$ au lieu de $\kappa = 0,57$).

Enfin, Shutova et al. (2013) ont montré que la métaphore est un phénomène très fréquent et que 68% des métaphores sont exprimées par des verbes.

2.3 Détection automatique de l'ironie, du sarcasme et de la satire

En parallèle des travaux qui ont proposé des schémas d'annotation pour le langage figuratif, dans les années 2000 un deuxième volet de recherche qui vise la détection automatique du langage figuratif a vu le jour. Ce sujet est devenu un sujet d'actualité en TAL en raison du

progrès des travaux relatifs à l'analyse des sentiments ainsi que la forte présence du langage figuratif dans les textes publiés sur le Web et les réseaux sociaux.

Globalement, les travaux sur la détection automatique du langage figuratif s'appuient sur trois approches : (1) approches surfaciques et sémantiques, (2) approches pragmatiques qui exploitent le contexte interne d'un énoncé et (3) approches pragmatiques qui exploitent un contexte supplémentaire externe à l'énoncé. La première approche (surtout celle qui se base sur les indices surfaciques) a souvent été considérée comme "baseline" dans la plupart des travaux utilisant la deuxième ou la troisième approche. Ces approches ont été exploitées dans les travaux traitant l'ironie, le sarcasme et la métaphore, contrairement aux travaux qui ont étudié la comparaison et l'humour.

Nous détaillons dans cette section les différents travaux de l'état de l'art sur la détection automatique du langage figuratif. Nous commençons par présenter les travaux traitant l'ironie, le sarcasme et la satire (cf. section 4.5). Nous poursuivons par ceux traitant la détection de la métaphore (cf. section 2.4), la comparaison (cf. section 2.5) et enfin l'humour (cf. section 2.6). Pour chaque type de langage figuratif, nous présentons les approches proposées ainsi que les corpus exploités.

2.3.1 Approches surfaciques et sémantiques

Un bref aperçu des traits lexicaux et sémantiques les plus utilisés pour la détection automatique de l'ironie et du sarcasme est présenté dans le tableau 2.3.

Groupes de recherche	Corpus	Traits	Résultats
(Burfoot & Baldwin, 2009)	Article de presse (4 233)	sac de mots, titre, profanation, argot, validité sémantique	Exactitude = 79,8%
(Carvalho *et al.*, 2009)	Article de presse (8 211) et commentaires (250 000)	ponctuation, guillemets, émoticône, citation, argot, interjection	Précision = 85,4%
(Veale & Hao, 2010)	Similitudes collectées à partir du web (20 299)	f-mesure = 73%	
(Liebrecht *et al.*, 2013)	Twitter (77 948)	unigrammes, bigrammes et trigrammes	AUC = 79%

TABLE 2.3 : Synthèse des principales approches surfaciques et sémantiques pour la détection de l'ironie/sarcasme.

Burfoot et Baldwin (2009) ont collecté un corpus de 4 233 articles de presse dont 4 000 non satiriques et 233 satiriques dans le cadre de la détection automatique de la satire. Un classifieur SVM-light avec des paramètres par défaut a été utilisé avec trois ensembles de traits : des traits de type sac de mots, des traits lexicaux et des traits sémantiques. Le modèle à base de sac de mots prend en compte soit des traits binaires permettant de capturer la présence ou l'absence des mots dans le corpus, soit une pondération des mots sur la base de la "Bi-normal separation Feature Scaling" (BNS). Cette dernière pondération permet d'affecter des poids aux traits qui sont fortement corrélés avec la classe négative ou positive. Pour les traits lexicaux, sont utilisés : le titre de l'article, la profanation, et l'argot. Les deux derniers traits permettent de capturer l'utilisation de vocabulaires familiers et informels, très utilisés dans des articles satiriques. Les résultats de l'apprentissage montrent que la combinaison de l'ensemble des traits donne un score de f-mesure de 79,5%.

Les articles de presse ont également fait l'objet des travaux de Carvalho et al. (2009) qui ont formé un corpus à partir d'une collection de 8 211 articles de presse collectés à partir d'un journal portugais accompagné des commentaires associés à chaque article (250 000 commentaires). Ce corpus a été exploité pour l'étude d'un ensemble d'indices linguistiques simples associés à l'expression de l'ironie en portugais. Dans ce cadre, Carvalho et al. (2009) utilisent une approche par patron pour identifier les phrases ironiques et non ironiques (e.g. $P_{laugh} = (LOL|AH|Emoticône)$, $P_{punct} = 4\text{-}GRAM^{+}$(!!|!?|?!)).

Les résultats montrent que les patrons les plus productifs sont ceux qui contiennent *des signes de ponctuation*, *des guillemets* et *des émoticônes*. D'autre part, les patrons *citations* et *argot* sont très performants pour la détection d'ironie, avec une précision de 85,4% et 68,3%, respectivement. L'utilisation de ces patrons a permis l'identification de 45% des phrases ironiques. Cependant, ces résultats ne sont pas très représentatifs car la couverture des patrons est extrêmement faible (environ 0,18%) principalement à cause du choix des phrases de départ qui correspondent à des phrases contenant des mots d'opinion et une entité nommée. Les cas d'indécision représentent 41% des commentaires collectés avec le patron *Interjection* et 27% des commentaires avec le patron *Ponctuation*. Ceci est dû à l'absence de contexte et à la nécessité d'une analyse plus fine (par exemple, analyser les phrases précédentes) afin de comprendre le sens ironique ou non d'un commentaire.

Dans le même cadre, Veale et Hao (2010) ont procédé à l'analyse d'un corpus de similitudes (comparaisons qui expriment une opposition). Ils ont commencé par une étape de collecte sur le web à l'aide du patron *« about as ADJ as ADJ »* dans le but de détecter les intentions ironiques dans les comparaisons créatives (par exemple : *He looked about as inconspicuous as a tarantula on a slice of angel food*). Les extraits collectés ont été filtrés à la main afin de séparer les similitudes des comparaisons, pour finalement obtenir 20 299 instances de similitudes. Une annotation manuelle du corpus a permis d'obtenir 76% de similitudes ironiques et 24% de similitudes non ironiques. Ces similitudes ont ensuite été classées en trois catégories selon l'opinion qu'elles véhiculent (*positive, négative* et *difficile à classer*). Les résultats montrent que les similitudes ironiques ont une forte tendance

à déguiser les sentiments négatifs en utilisant des termes positifs (71%). Alors qu'une petite minorité de similitudes (8%) tentent de transmettre un message positif dans un énoncé négatif ironique. L'automatisation de la classification des similitudes a été réalisée à travers un modèle qui comporte 9 étapes dont l'évaluation a obtenu une f-mesure de 73% pour la classe ironique et une f-mesure de 93% pour la classe non ironique.

Pour la classification de tweets en sarcastiques/non-sarcastiques, Liebrecht et al. (2013) ont collecté deux corpus : le premier est composé de 77 948 tweets en néerlandais contenant le hashtag *#sarcasme* ; le second comporte tous les tweets publiés le 1er février 2013 uniquement. Ce dernier corpus contient 3,3 million de tweets dont 135 tweets seulement contiennent le hashtag *#sarcasme*. Les auteurs ont exploité l'algorithme d'apprentissage supervisé *Balanced Winnow* avec les unigrammes, bigrammes et trigrammes comme traits. Dans cette expérimentation, le hashtag *#sarcasme* a été considéré comme étant l'annotation de référence. L'évaluation du modèle proposé a obtenu un score de *AUC (Area Under the Curve)* de 0,79 pour la détection des tweets sarcastiques.

Pour conclure, nous pouvons dire que les travaux présentés dans cette section s'accordent tous sur le fait que l'utilisation de traits surfaciques et sémantiques uniquement reste insuffisante pour la détection de l'ironie et du sarcasme et qu'il est nécessaire de recourir à d'autres traits plus pragmatiques, que nous détaillons dans les deux prochaines sections.

2.3.2 Approches pragmatiques exploitant le contexte interne de l'énoncé

Deux principales méthodes ont été proposées : celles qui utilisent des protocoles psycholinguistiques et celles qui utilisent des techniques d'apprentissage. Les premières, présentées en début de cette section, permettent de tester certaines hypothèses linguistiques sur l'ironie en les confrontant aux jugements d'annotateurs humains (via par exemple des plateformes de type Mechanical Turk). L'objectif est de présenter aux annotateurs un ensemble de textes ou expressions, et ces derniers doivent juger de leurs caractères ironiques ou non ironiques selon un ensemble de traits ou indices linguistiques. Les secondes, présentées plus loin dans cette section, se basent quant à elles sur de l'apprentissage supervisé ou semi-supervisé.

A Approches psycholinguistiques : travaux fondateurs d'Utsumi et de Kreuz et al

L'une des premières tentatives pour traiter automatiquement l'ironie a été décrite par Utsumi (1996). Cependant ce modèle était destiné à traiter un type particulier d'ironie qui se manifeste dans les interactions orateur-auditeur. Quelques années plus tard, Utsumi (2004) a défini l'ironie comme étant un phénomène pragmatique dont le traitement implique une interaction complexe entre le style linguistique et l'information contextuelle. En partant de cette définition, il a exploité une méthode psycholinguistique dans le but de détecter l'ironie,

le sarcasme et l'humour. Dans ce cadre, une étude empirique a été élaborée afin d'examiner la capacité des êtres humains à détecter des énoncés ironiques, sarcastiques et humoristiques en se basant sur le style et le contexte d'un énoncé donné. De plus, les annotateurs ont été invités à préciser la polarité de chaque énoncé étudié.

Il est à noter que cette étude expérimentale vise en grande partie à valider la théorie de *l'affichage implicite* proposée par Utsumi (2000). Cette dernière comporte trois volets :

1. L'ironie doit avoir un environnement ironique, un cadre approprié de la situation dans le contexte du discours. Cet environnement nécessite : (a) une attente de l'orateur, (b) une incongruité entre les attentes et la réalité, et (c) une attitude négative de l'orateur vers l'incongruité. Par conséquent, un énoncé doit être interprété comme étant ironique dans le cas où la situation du discours a été identifiée comme étant un environnement ironique.

2. L'ironie est un énoncé qui affiche implicitement un environnement ironique. Ceci est assuré par la présence d'un énoncé qui : (d) fait allusion à l'attente de l'orateur, (e) comprend une insincérité pragmatique par la violation de l'un des principes pragmatiques, et (f) exprime indirectement l'attitude négative de l'orateur en étant accompagné par des indices ironiques.

3. L'ironie est une catégorie basée sur un prototype caractérisé par la notion de l'affichage implicite. Le prototype de l'ironie est un modèle abstrait qui répond aux trois conditions de l'affichage implicite. Le degré d'ironie peut être évalué par la similitude entre le prototype et un énoncé donné par rapport aux trois conditions (opposition, question rhétorique, circonlocutions).

Par conséquent, la théorie de l'affichage implicite repose sur trois hypothèses à savoir (cf. figure 2.2) :

1. Le degré d'ironie est influencé par le choix linguistique, pas par le cadre contextuel. Le degré de l'ironie est élevé dans le cas où les propriétés de l'affichage implicite sont satisfaites.

2. Le degré de sarcasme d'un énoncé ironique est influencé seulement par le style linguistique. Le degré de sarcasme est élevé dans le cas où les propriétés de l'affichage implicite sont satisfaites.

3. Le degré de l'humour d'un énoncé ironique est influencé à la fois par le style linguistique et le contexte. Le degré de l'humour est élevé dans le cas où un contexte de discours est incongru avec l'environnement ironique ou l'énoncé est différent du prototype de l'ironie.

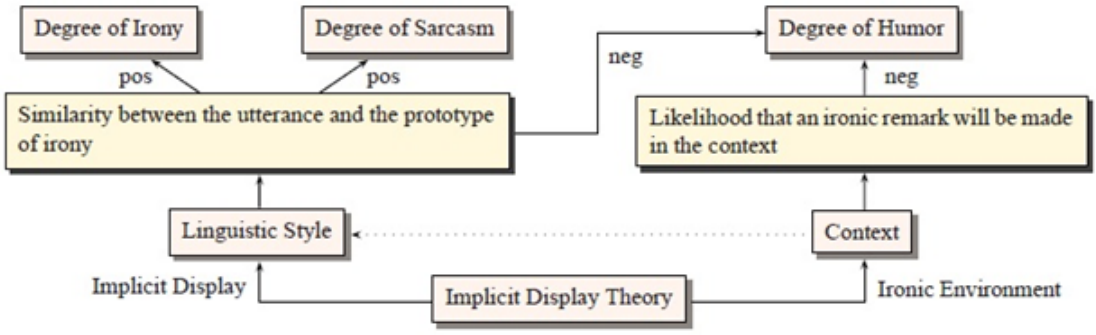

FIGURE 2.2 : Les hypothèses générales du traitement de l'ironie selon la théorie de l'affichage implicite (Utsumi, 2004).

Cette théorie a été validée par deux expérimentations sur un corpus d'étude formé de 12 histoires écrites en japonais. La première expérimentation a été menée dans le but de tester la validité de la théorie en examinant comment le style linguistique affecte le degré d'ironie, de sarcasme et d'humour. Le style linguistique de l'ironie a été défini par deux facteurs :

- **Le type de la phrase :** Trois types ont été identifiés, à savoir *(1) l'opposition* : déclaration dont le sens littéral positif est à l'opposé de la situation négative ; *(2) la question rhétorique* : déclaration interrogative par laquelle l'orateur demande rhétoriquement un fait qui est évident pour le destinataire ; *(3) la circonlocution :* sorte de litote qui est faiblement liée à l'attente de l'orateur par un certain nombre de relations de cohérence (cf. section D du chapitre 3).
- **Le niveau de politesse :** l'utilisation ou non des titres honorifiques japonais avec le type de la relation (bonne ou mauvaise) entre le locuteur et destinataire.

120 étudiants japonais ont été recrutés afin d'étudier le corpus formé de 12 histoires. Chaque participant devait lire les histoires une par une et les évaluer en attribuant deux valeurs pour chaque histoire, une valeur entre 1 et 7 (1= non sarcastique, 7= extrêmement sarcastique) et une deuxième valeur entre 1 et 7 (1= non humoristique, 7= extrêmement humoristique). Après cette étape, chaque participant devait relire les 12 histoires et attribuer une valeur entre 1 et 7 (1= non ironique, 7= extrêmement ironique). Enfin, les annotateurs devaient attribuer trois valeurs pour chaque histoire qui évaluent l'absence ou la présence de sarcasme, humour et ironie. Les résultats de cette première expérimentation ont prouvé que :

- *les oppositions* étaient beaucoup plus ironiques et plus sarcastiques que les circonlocutions, et plus sarcastiques que les questions rhétoriques.

- *les questions rhétoriques* ont été jugées beaucoup plus sarcastiques que les circonlocutions.
- lorsque l'orateur a une bonne relation avec le destinataire, *les énoncés honorifiques* ont été jugés beaucoup plus ironiques et sarcastiques que les énoncés non honorifiques, mais cette différence a disparu quand la relation entre l'orateur et le destinataire est mauvaise.
- lorsque l'orateur et l'interlocuteur avait une bonne relation, *les circonlocutions* avec absence de termes honorifiques ont été jugées plus humoristiques que celles accompagnées de termes honorifiques mais cette différence n'a pas été observée lorsque la relation était mauvaise.

Une deuxième expérimentation a été menée dans le but de tester la théorie de l'affichage implicite à l'égard de l'effet contextuel sur le degré d'ironie, de sarcasme et de l'humour. Dans la deuxième expérimentation, deux variables indépendantes ont été examinées : (1) la négativité de la situation : la situation est faiblement ou fortement négative ; (2) la banalité de la situation négative : la situation négative est habituelle ou non.

L'auteur a ainsi sélectionné 8 histoires parmi les 12 précédentes et 48 étudiants japonais pour une seconde expérimentation. Les résultats montrent que :

- Le contexte a une influence indirecte sur le degré d'ironie quand l'orateur s'exprime d'une manière implicite.
- Le destinataire est moins susceptible de détecter la croyance de l'orateur, et donc il perçoit un énoncé comme moins ironique quand son comportement négatif est habituel que dans le cas où il n'est pas habituel.
- Pour le degré de sarcasme, il n'y a pas d'effet et d'interaction significative dans la nouvelle analyse. Ce résultat suggère que les attentes de l'orateur peuvent être une propriété importante qui distingue l'ironie du sarcasme vu que le sarcasme ne nécessite pas d'attente de l'orateur.
- Les énoncés ironiques dans les contextes dans lequel le comportement négatif du destinataire est habituel, ont été considérés comme plus humoristique que les mêmes phrases dans le contexte dans lequel le comportement négatif était inhabituel.

Tout comme Utsumi, Kreuz et Caucci (2007) s'inscrivent dans les approches psycholinguistiques. Les auteurs ont collecté un corpus de 100 ouvrages composés de romans historiques, d'amour et de science-fiction afin d'étudier l'influence des indices lexicaux sur la perception du sarcasme dans la langue anglaise. Ces ouvrages ont été collectés au hasard

parmi les ouvrages contenant l'expression *"said sarcastically"*. Un groupe de 101 étudiants a été chargé de la classification des extraits d'ouvrages en sarcastiques/non sarcastiques. Chaque participant devait répondre pour 35 extraits (20 extraits contenant du sarcasme) à la question suivante : *Quelle est la probabilité que l'auteur soit sarcastique ?* (sur une échelle de 7 points). Pour chaque participant, les auteurs ont calculé une valeur moyenne pour l'ensemble des extraits. Comme prévu, les extraits sarcastiques ont des scores plus élevés (moyenne$= 4,85$, écart-type$= 0,67$) que les extraits de contrôle (moyenne$= 2,89$, écart-type$= 0,86$).

En parallèle, un deuxième groupe formé de deux experts a été chargé de déterminer l'importance des facteurs lexicaux sur la perception du sarcasme. Ceci a été effectué à travers une analyse de régression en exploitant les cinq indices suivants : (1) le nombre de mots dans chaque extrait, (2) le nombre de mots en caractères gras, (3) la présence d'interjections, (4) la présence d'adjectifs et d'adverbes, et (5) l'utilisation des points d'exclamation et d'interrogation. Les résultats montrent que les trois premiers indices sont pertinents pour la détection du sarcasme alors que les deux derniers le sont moins.

B Approches par apprentissage automatique

Un bref aperçu des traits pragmatiques les plus utilisés pour la détection automatique est présenté dans le tableau 2.4.

Deux grands groupes de travaux peuvent être distingués : ceux qui exploitent les commentaires en ligne tels que des commentaires de produits ou de films, et ceux qui exploitent des tweets.

Ironie dans des commentaires en ligne. (Tsur *et al.*, 2010) ont collecté un corpus de 66 000 commentaires en anglais sur les produits d'Amazon dans le but de présenter un algorithme nommé *SASI (Semi-supervised Algorithm for Sarcasm Identification)* pour la classification des commentaires en sarcastique/non sarcastique. Cet algorithme utilise deux types de traits. Le premier regroupe des traits qui s'appuient sur des patrons construits automatiquement en utilisant l'algorithme de Davidov et Rappoport (2006) et qui reflètent le sujet principal discuté (principalement le nom de produits ou de compagnies), ce qui permet de séparer les mots fréquents des mots de contenus. Le deuxième groupe de traits regroupe des traits lexico-syntaxiques comme la longueur de la phrase en mots, le nombre de points d'exclamation, d'interrogation et de citation dans la phrase, ou encore le nombre de mots en majuscules.

La combinaison de tous les traits a donné une valeur de f-mesure de 82% (chaque phrase a été annotée par 3 annotateurs). Les signes de ponctuation sont les prédicateurs les plus faibles (f-mesure$= 28,1\%$). Par contre, l'exploitation des traits pragmatiques à base de

Groupes de recherche	Corpus	Traits	Résultats
(Tsur *et al.*, 2010)	Commentaires Amazon (66 000)	fréquence du commentaire, type du produit, compagnie, titre, auteur, longueur, ponctuation, citation, majuscule	f-mesure = 82%
(Reyes & Rosso, 2011)	Commentaires Amazon et Slashdot.com (8 861)	n-grams, n-grams POS, profilage drôle, profilage négatif/postif, profilage affective, profilage d'agréabilité	f-mesure = 75,75%
(Reyes & Rosso, 2014)	Commentaires films (3 400), commentaires livres (1 500) et articles de presse (4 233)	Signature, Scénarios émotionnels, L'inattendu	-
(Buschmeier *et al.*, 2014)	Commentaires Amazon (1 254)	déséquilibre, hyperbole, citation, ponctuation, polarité pos/neg, interjection, émoticône, sac de mots	f-mesure = 74,4%
(Gonzalez-Ibanez *et al.*, 2011)	Twitter (2 700)	unigrams, dictionnaire, wordNet, interjection, ponctuation, émotion positive/négative, réponse à un autre utilisateur	Exactitude = 71%
(Reyes *et al.*, 2013)	Twitter (40 000)	signatures, l'inattendu, style et scénarios émotionnels	f-mesure = 76%
(Barbieri & Saggion, 2014b)	Twitter (40 000)	fréquence mots rares, synonyme, écart synonyme et ponctuation	f-mesure = 76%
(Barbieri *et al.*, 2014)	Twitter (60 000)	fréquence mots rares, synonyme, écart synonyme et ponctuation	f-mesure = 62%
(Joshi *et al.*, 2015)	Twitter (12 162) et discussion forum (1 502)	unigrams, majuscule, émoticône, ponctuation, incongruité implicite,incongruité explicite	f-mesure = 61%

TABLE 2.4 : Synthèse des principales approches pragmatiques exploitant le contexte interne de l'énoncé pour la détection de l'ironie/sarcasme.

patrons pour le type du produit, le nom de la société de production du produit, l'auteur du commentaire, etc. permet l'amélioration de la classification des commentaires avec une valeur de f-mesure de 76,9%. Par conséquent, la combinaison des traits surfaciques avec les traits pragmatiques maximise les performances de classification. D'où, l'importance de l'utilisation des traits pragmatiques pour inférer le langage figuratif.

Les commentaires postés en ligne ont aussi été utilisés par Reyes et Rosso (2011) qui ont collecté un corpus de 8 861 commentaires ironiques à partir de *Amazon.com (AMA)* et *Slashdot.com (SLA)* [7]. Un modèle formé de 6 traits a été proposé dont les traits pragmatiques suivants :

1. **profilage drôle (*funny profiling*) :** vise à caractériser les documents en termes de propriétés humoristiques identifiées par :
   - **des caractéristiques stylistiques :** selon les expérimentations de (Mihalcea & Strapparava, 2006), ces caractéristiques sont obtenues en collectant tous le mots annotés avec l'étiquette *"sexuality"* à partir de *WordNet Domains* (Bentivogli *et al.*, 2004).
   - **la centralité :** qui permet de capturer les relations sociales (mots récupérés dans la base lexicale WordNet (Miller, 1995)).
   - **des mot-clés spécifiques :** cette valeur est calculée en comparant la fréquence du mot dans des documents ironiques par rapport à sa fréquence dans un corpus de référence (ici Google N-grams (Reyes *et al.*, 2009)).
2. **profilage positif/négatif :** c'est-à-dire le fait d'exprimer une opinion négative par une opinion positive.
3. **profilage affectif :** WordNet-Affect est utilisé pour obtenir des termes affectifs.
4. **profilage d'agréabilité (*pleasantness profiling*) :** le dictionnaire des "Affect" anglais Whissell (Whissell, 1989) a été utilisé : chaque entrée comporte un score d'agréabilité fixé par annotation manuelle allant de 1 (non agréable) à 3 (agréable).

Trois classifieurs ont été entrainés avec ces traits (*Naïve Bayes, Support Vector Machine (SVM)* et *Arbres de décision*) dans une configuration en validation croisée (10-fold cross validation). La plupart des classifieurs ont obtenu une valeur d'exactitude supérieure à 70% avec une valeur d'exactitude maximale égale à 75,75% pour le sous-corpus *AMA* et un classifieur SVM. Les traits les plus performants pour le sous-corpus *AMA* sont les traits sac de mots (trigrammes), d'agréabilité et profilage drôle, alors que ceux pour le sous-corpus *SLA* sont l'agréabilité et les 5-grammes.

Le modèle de Reyes et Rosso (2011) a été ensuite étendu dans (Reyes & Rosso, 2014). Trois corpus ont été utilisés dont deux pour l'ironie, un pour la satire : des commentaires de films (*movies2*) issus des travaux de Pang et Lee (2004) et de Pang et al. (2002), des commentaires de livres construits par Zagibalov et al. (2010) et enfin des articles satiriques, collectés par Burfoot et Baldwin (2009). Le modèle proposé a pour but de fournir pour chaque document et phrase du corpus, la probabilité qu'il (elle) soit ironique ou non. Le modèle est organisé selon trois couches conceptuelles :

[7] http ://users.dsic.upv.es/grupos/nle

1. **Signature :** regroupe trois traits :
   (a) *Pointillisme :* s'intéresse à la détection des marqueurs explicites à savoir la ponctuation, les émotions, les citations et la majuscule.
   (b) *Contre-factualité :* s'intéresse aux marqueurs implicites, c'est-à- dire les termes discursifs qui font allusion à l'opposition ou contradiction dans un texte anglais, comme *"nevertheless", "nonetheless" ou "yet"*.
   (c) *Compression temporelle :* s'intéresse à l'identification des éléments liés à l'opposition dans le temps, c'est-à-dire des termes qui suggèrent un changement brusque de la séquence narrative. Ces éléments sont représentés par un ensemble d'adverbes temporels tels que *soudainement, maintenant, brusquement*.
2. **Scénarios émotionnels :** regroupe trois traits :
   (a) *Activation :* désigne le degré de réponse, soit passive ou active, que les humains présentent dans un état émotionnel.
   (b) *Imagerie :* tente de quantifier la facilité ou la difficulté de former une image mentale pour un mot donné.
   (c) *Plaisir (agrément) :* mesure le degré de plaisir suggéré par un mot.
3. **L'inattendu :** regroupe deux traits :
   (a) *Déséquilibre temporel :* est utilisé pour refléter le degré d'opposition temporelle dans un texte en ce qui concerne les temps présents et passés.
   (b) *Déséquilibre contextuel :* est utilisé pour capturer les incohérences dans un contexte.

Les résultats indiquent que la probabilité qu'un document soit ironique est plus élevée dans les commentaires de films. De plus, les documents avec une polarité positive sont ceux qui ont une grande probabilité de contenir un contenu figuratif (ironie, sarcasme, satire ou humour).

(Buschmeier *et al.*, 2014) ont exploité le corpus de (Filatova, 2012) qui comporte 437 commentaires sarcastiques et 817 non-sarcastiques sur les produits d'Amazon afin de proposer un ensemble de traits pour la détection automatique de l'ironie à savoir : *le déséquilibre entre le nombre d'étoile et la polarité du texte, l'hyperbole, les citations, séquence de mots positifs ou négatifs suivie d'au minimum deux points d'exclamation ou d'interrogation, séquence de mots positifs ou négatifs suivie de "...", ponctuation, interjection, émoticône et sac de mots*.

Les résultats de cette expérience indiquent que la combinaison de l'ensemble des traits avec le nombre d'étoiles relatifs à chaque commentaire a donné la meilleure valeur de f-mesure avec une valeur de 74,4% en utilisant le classifieur *Logistic Regression*.

Ironie dans les tweets. Une nouvelle ressource de données disponibles en ligne a vu le jour avec l'apparition des réseaux sociaux dont Twitter[8]. La forte utilisation de Twitter partout dans le monde et le grand nombre de tweets publiés par jour à propos de différents sujets a poussé les chercheurs en TAL a utilisé Twitter comme source de données pour étudier le différents phénomènes du langage figuratif.

Dans ce cadre, Gonzalez-Ibanez et al. (2011) ont collecté un corpus de 2 700 tweets en anglais dont 900 tweets sarcastiques (contenant les hashtags *#sarcasm, #sarcastic*), 900 tweets de polarité positive (*#happy, #joy, #lucky*) et 900 tweets de polarité négative (*#sadness, #angry, #frustrated*). La classification de 90 tweets de chaque classe (sarcastique, positif et négatif) pris aléatoirement a obtenu une valeur d'exactitude de 62,59% par une classification manuelle (3 annotateurs) et de 57,41% en utilisant le classifieur SMO (voir les traits dans le tableau 2.4). Quand il s'agit de classer les tweets en sarcastique vs. non sarcastique, la classification manuelle de 180 tweets a donné une exactitude de 66,85% et de 68,33% en utilisant SMO avec le trait unigramme. Les résultats obtenus montrent que les émoticônes jouent un rôle important en aidant les humains à distinguer les tweets sarcastiques des non sarcastiques. En revanche, un des annotateurs a mentionné qu'il était parfois nécessaire de faire appel à des connaissances du monde pour détecter le sarcasme. Ceci suggère que l'identification automatique du sarcasme sur Twitter nécessite des informations sur l'interaction entre les utilisateurs et la connaissance du monde.

Dans le même cadre, Reyes et al. (2013) ont proposé un modèle capable de représenter les attributs les plus marquants de l'ironie verbale dans un texte afin d'être capable de la détecter automatiquement. A cet effet, ils ont collecté un corpus de 40 000 tweets en anglais dont 10 000 ironiques et 30 000 non ironiques partagés d'une manière équitable entre les thèmes suivants : éducation, humour et politique. Un deuxième corpus de 500 tweets contenant le hashtag *#Toyota* et les émoticônes " :)" (250 tweets) et " :(" (250 tweets) a été collecté afin d'appliquer la méthode proposée sur un cas réel (ces tweets ne sont pas explicitement marqués comme ironique par leur auteur. Ils ont été annotés par 80 annotateurs). Ils ont ensuite défini un modèle pour l'extraction d'un ensemble de traits pour la détection de l'ironie. Le modèle proposé comporte quatre traits conceptuels à savoir : *les signatures, l'inattendu, le style et les scénarios émotionnels*. Selon les auteurs, ces traits permettent de capturer les propriétés de bas niveau et de haut niveau de l'ironie textuelle basée sur des descriptions conceptuelles dans la littérature.

Dans la phase d'évaluation du modèle proposé sur le premier corpus, deux classifieurs ont été exploités : *Naïve Bayes* et les *arbres de décision* avec 10-cross validation. Les meilleures résultats ont été obtenus avec les arbres de décision avec une répartition équilibrée entre les instances ironiques/non ironiques. Les résultats en terme de f-mesure vont de 70% à 76% selon les thèmes des corpus. Les résultats de l'apprentissage ont prouvé qu'il existe des traits qui semblent être inutiles dans la détection automatique des tweets ironiques

[8]https ://fr.wikipedia.org/wiki/Twitter

(par exemple, les traits de compression temporelle des signatures, le déséquilibre contextuel de l'inattendu, le style, et l'imagerie de scénarios émotionnels). L'application de l'approche proposée sur le second corpus de Toyota a prouvé que le système a détecté 123 tweets ironiques alors que les humains en ont détecté 147. L'approche proposée est donc fiable car les résultats obtenus automatiquement sont très proches de ceux des humains.

Le corpus construit par Reyes et al. (2013) a été réutilisé par Barbieri et Saggion (2014b) dans le cadre de la proposition d'un modèle qui regroupe sept traits lexicaux : la fréquence des mots, la moyenne des fréquences des mots du vocabulaire, la structure (nombre de mots, nombre de caractères, etc.), les intensifieurs, les sentiments, les synonymes et l'ambiguïté.

Afin de tester le modèle proposé, Barbieri et Saggion (2014b) ont utilisé une méthode d'apprentissage supervisé en utilisant trois corpus différents. Le premier corpus est constitué de 10 000 tweets contenant le hashtag *#irony* et 10 000 tweets contenant le hashtag *#education*. Le deuxième corpus compte 10 000 tweets contenant le hashtag *#irony* et 10 000 contenant le hashtag *#humor*. Et enfin, un troisième corpus de 10 000 tweets contenant le hashtag *#irony* et 10 000 contenant le hashtag *#politics*. Les résultats en terme de f-mesure obtenus pour les trois corpus sont les suivants : 72% pour le corpus ironique vs éducation, 75% pour le corpus ironique vs humour et 76% pour le corpus ironique vs politique. Une étude de pertinence des traits a montré que les traits *fréquence des mots rares, synonyme et ponctuation* sont les traits les plus discriminants pour la détection de l'ironie. En revanche, les traits discriminants ne sont pas les mêmes pour les différents corpus. ce qui prouve qu'il est très difficile de définir un ensemble de traits discriminants pour les différents thèmes.

Le modèle proposé par Barbieri et Saggion (2014b) a été repris par Barbieri et Saggion (2014) dans le but de la détection automatique du sarcasme. Dans ce cadre, ils ont exploité un corpus de 60 000 tweets en anglais répartis équitablement sur six thèmes à savoir : *sarcasme, éducation, ironie, politique, journaux*. L'ensemble des hashtags ont été enlevés pour la tâche de la classification automatique. La classification binaire des tweets a été effectuée à travers une approche d'apprentissage supervisé en utilisant les arbres de décision et l'ensemble des traits proposés par Barbieri et Saggion (2014b). Afin d'évaluer l'efficacité du modèle proposé pour la détection automatique du sarcasme, le corpus a été réparti en cinq sous-corpus équilibrés en termes d'ironiques/non ironiques : *sarcasme vs éducation, sarcasme vs humour, sarcasme vs ironie, sarcasme vs journaux* et *sarcasme vs politique*. Les résultats en termes de f-mesure sont de : 88% pour les thèmes éducation et humour, 62% pour le thème ironie, 97% pour le thème journaux et 90% pour le thème politique. Ainsi, les résultats obtenus ont prouvé que le modèle proposé obtient de bons résultats pour la distinction entre les tweets sarcastiques et non sarcastiques. En revanche, le modèle a obtenu de mauvais résultats pour la distinction entre les tweets ironiques et sarcastiques. Les auteurs expliquent ceci par le fait que l'ironie et le sarcasme ont des structures similaires dans le modèle proposé d'où la nécessité de proposer de nouveaux traits afin d'avoir une bonne distinction entre ces deux phénomènes. En parallèle, ils ont montré que les tweets sarcastiques contiennent moins d'adverbes que les tweets ironiques mais que ces adverbes

sont plus intenses et que les tweets sarcastiques contiennent des sentiments positifs plus que les tweets ironiques. La distinction entre ironie et sarcasme a également été traité par Sulis et al. (2016a) qui ont obtenu une f-mesure de $69,8\%$.

Enfin, dans le cadre d'une étude menée sur un corpus mixte formé à partir de Twitter et de forums de discussions, Joshi et al. (2015) ont proposé une approche qui utilise deux types d'incongruité : *explicite* et *implicite*. *L'incongruité explicite* est exprimée par des mots de sentiment de polarités différentes alors que *l'incongruité implicite* est exprimée par des phrases exprimant un sentiment implicite qui s'oppose à un mot de polarité positive ou négative. Afin de résoudre cette problématique, Joshi et al. (2015) ont proposé quatre groupes de traits à savoir : (1) lexical : unigrammes, (2) pragmatique : majuscule, émoticône, ponctuation, (3) incongruité implicite et (4) incongruité explicite : nombre d'incongruités entre sentiment positif et négatif, plus grande séquence positive/négative, nombre de mots positifs, nombre de mots négatifs et polarité globale du texte. Un apprentissage supervisé avec LibSVM a été appliqué. Les résultats obtenus en termes de f-mesure sont meilleurs que ceux obtenus par Riloff et al. (2013) et Maynard et Greenwood (2014).

2.3.3 Approches pragmatiques exploitant le contexte externe de l'énoncé

Un bref aperçu des traits pragmatiques les plus utilisés est présenté dans le tableau 2.5.

Groupes de recherche	Corpus	Traits	Résultats
(Wallace, 2015)	Commentaires articles politique (2 821) et discussion forum (1 502)	Sentiment, Subreddit(topic), phrase nominale	-
(Bamman & Smith, 2015)	Twitter (19 534)	n-grams, POS, sentiment, intensifieur, profil auteur, historique de discussion, etc.	Exactitude = 85,1%
(Joshi *et al.*, 2016)	Extraits de livres de GoodReads (3 629)	traits de (Liebrecht *et al.*, 2013), (Gonzalez-Ibanez *et al.*, 2011), (Buschmeier *et al.*, 2014) et (Joshi *et al.*, 2015)	f-mesure = 81,19%

TABLE 2.5 : Synthèse des principales approches pragmatiques exploitant le contexte externe de l'énoncé pour la détection de l'ironie/sarcasme.

Dans le cadre de la proposition d'une stratégie de classification de l'ironie verbale, Wallace et al. (2015) ont exploité un corpus de commentaires sur les articles politiques collecté à partir du site web *reddit.com* et utilisé dans le cadre d'une campagne d'annotation par

Wallace et al. (2014). Ce corpus d'étude est formé de trois sous-ensembles de commentaires. Le premier sous-corpus est formé de 1 825 commentaires dont 286 ont été annotés comme étant ironiques. Le second sous-corpus compte 996 commentaires politiques dont 154 ont été annotés comme étant ironiques. Le troisième sous-corpus compte 1 682 commentaires religieux dont 313 ont été annotés comme étant ironiques. Le méthode proposée utilise quatre traits :

- **Sentiment :** représente le sentiment inféré pour un commentaire donné (négatif, neutre ou positif).
- **Subreddit :** spécifie le subreddit (thème) sur lequel le commentaire a été publié (par exemple, progressiste ou conservateur, athéisme ou christianisme).
- **NNP :** relatif aux phrases nominales présentes dans un commentaire donné.
- **NNP+ :** relatif aux phrases nominales présentes dans un commentaire donné et au fil auquel le commentaire est relié (par exemple, le titre d'une image qui accompagne le commentaire).

La méthode proposée a été testée sur les trois corpus. Les résultats montrent une augmentation de la valeur moyenne du rappel (entre 2 et 12% selon le corpus) par rapport à la baseline (sac de mots).

Bamman et Smith (2015) ont collecté un corpus de 19 534 tweets dont la moitié est sarcastique (*#sarcasm, #sarcastic*). Ce corpus a été exploité dans une approche automatique pour la détection du sarcasme dans laquelle les auteurs ont utilisé quatre groupes de traits :

- **Traits de tweets :** regroupe un ensemble de neuf traits : (1) unigrammes et bigrammes de mots, (2) groupe d'unigrammes et bigrammes, (3) bigrammes de dépendance, (4) catégories grammaticales, (5) prononciation, (6) majuscule, (7) sentiment global du tweet, (8) sentiment des mots du tweet et (9) intensifieur.
- **Traits de l'auteur :** r (1) termes saillants de l'historique de l'auteur, (2) historique des mots-clés les plus utilisés par l'auteur, (3) information sur le profil de l'auteur, (4) historique des sentiments exprimés par l'auteur et (5) unigrammes du profil de l'auteur.
- **Traits de l'audience :** (1) combinaison de l'ensemble des traits du groupe "Traits de l'auteur", (2) sujet d'interaction Auteur/Destinataire et historique de la communication entre l'auteur et le destinataire.
- **Traits de l'environnement :** (1) interaction entre un tweet cible et le tweet auquel il répond en terme de couple de mots dans les deux tweets et (2) unigramme du tweet d'origine pour capturer le contexte linguistique original auquel un tweet répond.

Pour la tâche de la classification automatique des tweets en sarcastique/non sarcastique, Bamman et Smith (2015) ont exploité une régression logistique binaire avec validation croisée. Le premier groupe de traits *Traits de tweets* fournit une valeur moyenne d'exactitude de 75,4% alors que l'ajout des traits pragmatiques *historique de discussion* augmente la valeur d'exactitude à 77,3%. L'ajout des *traits de l'audience* repousse la valeur d'exactitude à 79% et les *traits de l'auteur* à 84,9%. En revanche, la combinaison de tous les traits atteint la meilleure valeur d'exactitude à 85,1%. Par conséquent, l'utilisation des traits de surface uniquement est insuffisante pour inférer les messages sarcastiques. D'où, l'importance des traits pragmatiques qui exploitent le contexte externe du tweet afin de maximiser la performance du système de détection automatique.

Enfin, Joshi et al. (2016) ont collecté un corpus de 3 629 extraits de livres sarcastiques et non sarcastiques à partir du site web *GoodReads* [9] Ce corpus a été exploité dans une approche automatique pour la détection du sarcasme dans laquelle Joshi et al. (2016) ont réutilisé l'ensemble des traits précédemment utilisés par Liebrecht et al. (2013), Gonzalez-Ibanez et al. (2011), Buschmeier et al. (2014) et Joshi et al. (2015) auxquels ils ont ajouté de nouveaux traits basés sur les « word embeddings » (par exemple, le score maximal de la paire de mots les plus dissemblables). La meilleure f-mesure (81,19%) a été obtenue en fusionnant les traits de Liebrecht et al. (2013) et les nouveaux traits proposés obtenus avec l'approche *Dependency Weights*[10].

2.4 Détection automatique de la métaphore

Bien que la plupart des travaux sur la détection automatique du langage figuratif se soient focalisés sur l'ironie et le sarcasme, d'autres travaux ont étudié la métaphore, la comparaison et l'humour mais leur nombre reste limité. De même que l'ironie, les travaux étudiant la métaphore se sont focalisés sur les traits surfaciques et sémantiques (Kintsch, 2000; Bestgen & Cabiaux, 2002) et sur le contexte interne de l'énoncé (Gedigian *et al.*, 2006; Oliveira & Ploux, 2009; Macwhinney & Fromm, 2014; Tsvetkov *et al.*, 2014; Huang, 2014). En revanche, comme pour l'ironie, l'exploitation du contexte externe à l'énoncé a vu le jour en 2015 avec les travaux de (Jang *et al.*, 2015b; Do Dinh & Gurevych, 2016; Su *et al.*, 2016; Goode *et al.*, 2017).

2.4.1 Approches surfaciques et sémantiques

Selon les définitions présentées dans le chapitre 1, la métaphore peut être considérée comme une comparaison. Les études menées sur la détection automatique de ce phénomène prouvent

[9] https://www.goodreads.com/

[10] https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/

que la détection automatique n'est pas simple et qu'il y a beaucoup de facteurs qui doivent être pris en considération.

Dans les année 2000, l'une des premières tentatives de proposition d'un modèle pour le traitement automatique de la métaphore a été décrite par Kintsch (2000). Kintsch (2000) a précisé dans ces études que la compréhension de la métaphore implique une interaction entre le sens du sujet et les termes véhiculant la métaphore. En partant de cette hypothèse, il a proposé un modèle qui s'appuie sur la conception interactive de l'interprétation des métaphores. Par exemple dans *"Mon avocat est un requin"*, le modèle proposé consiste à identifier *"avocat"* comme étant le *topique* et *"requin"* comme étant le *véhicule* et dans ce cas il faut sélectionner les traits relatifs à un requin (par exemple : sanguinaire ou vicieux) qui peuvent être attribuer à un avocat.

Afin d'implémenter ce modèle, Kintsch (2000) a commencé par l'identification des traits sémantiques qui participent au sens de la métaphore et a proposé un algorithme capable d'effectuer la sélection en suivant les étapes suivantes :

- construire un espace sémantique de grandes dimensions à partir de l'analyse statistique des co-occurrences dans un corpus de textes en exploitant l'analyse sémantique latente.
- représenter le sens de chaque mot par un vecteur.
- mesurer la similarité entre les mots en calculant le cosinus entre les vecteurs représentatifs de ces mots (la valeur du cosinus tend vers 1 en présence d'une similitude entre deux mots).

Cet algorithme employé pour déterminer le sens d'une prédication vise à sélectionner parmi les "traits" du prédicat ceux qui sont proches de l'argument en recherchant parmi les n plus proches voisins du prédicat et les k plus proches voisins de l'argument. Selon ce modèle, le seul facteur qui change lors de l'analyse d'un énoncé métaphorique au lieu d'un énoncé littéral est le paramètre n. Selon Kintsch (2000), pour un énoncé littéral, les 20 plus proches voisins sont suffisant, alors que pour un énoncé métaphorique, il faut aller jusqu'à 200 voire 500.

L'approche proposée par Kintsch (2000) a été critiquée par Bestgen et Cabiaux (2002) qui trouvent que les arguments avancés dans cette approche sont limités car l'approche ne couvre que quelques exemples de métaphores. Bestgen et Cabiaux (2002) proposent donc un autre modèle utilisant l'analyse sémantique latente qui peut être appliqué à des métaphores littéraires de différents types afin de vérifier son efficacité sur les expressions jugées par des lecteurs comme très métaphoriques ou peu métaphoriques et de définir un indice de l'intensité figurative. Pour cela, ils ont collecté un corpus de vingt phrases contenant des expressions métaphoriques qui ont été sélectionnées dans neuf contes de Maupassant. Dix

phrases exprimaient une métaphore *vive* et dix phrases exprimaient une métaphore *morte*. Les métaphores *mortes* (*la voix s'envole sans écho*) sont les métaphores employant des mots dans un sens que le dictionnaire *Petit Robert* qualifie de figuratif alors que le sens des mots employés dans les métaphores considérées comme *vives* (*des fusées de gaieté*) n'était pas mentionné dans ce même dictionnaire. Ils concluent que le modèle de Kintsch (2000) permet d'approximer le sens de métaphores littéraires de différents types et qu'il est possible d'en dériver un indice qui distingue les énoncés métaphoriques des énoncés littéraux.

Cette étude est encore loin de la proposition d'une procédure automatique pour l'identification et l'interprétation de métaphores étant donnés le type du modèle proposé et la taille du corpus d'étude.

2.4.2 Approches pragmatiques exploitant le contexte interne de l'énoncé

En 2006, Gedigian et al. (2006) ont proposé une approche pour la détection automatique de la métaphore en exploitant le contexte interne de l'énoncé. A cet effet, ils ont collecté un corpus d'articles de presse publiés dans le *Wall Street Journal* (WSJ). Les cibles verbales associées à trois sujets – le mouvement spatial, la manipulation et la santé – ont été annotées manuellement. Ont également été annotés : les cibles métaphoriques, le sens littéral des cibles et les cibles pour lesquelles les annotateurs n'arrivent pas à décider. Cette phase d'annotation a montré que plus que 90% des cibles ont été utilisées métaphoriquement. Les résultats du système proposé ont donné une exactitude de 95,12%.

Oliveira et Ploux (2009) ont proposé une méthode pour la détection automatique de la métaphore dans un corpus parallèle ou comparable de textes français et portugais. Le corpus d'étude est divisé en trois sous-corpus. Le premier sous-corpus littéraire est constitué d'environ 200 romans du XIXème siècle français ou portugais et leur traduction dans l'autre langue. Le deuxième sous-corpus journalistique est composé d'articles publiés entre 1997 et 2001. Le troisième sous-corpus regroupe les traités européens Les auteurs ont exploité le modèle *ACOM (Automatic Contexonym Organizing Model)* proposé par Hyungsuk et al. (2003) pour le calcul de la distance entre les contextes d'emploi les plus génériques des termes relatifs à une expression (métaphorique ou non). Selon les auteurs, les résultats de ce modèle seront exploités ultérieurement comme étant un critère de détection automatique de la métaphore.

Macwhinney et Fromm (2014) ont exploité un corpus multilingue (anglais, farsi, russe et espagnol) nommé *TenTen* qui contient environ 10 billions de mots par langue, déjà lemmatisés et étiquetés (catégories grammaticales et relations de dépendance entre les mots du domaine source et cible). Macwhinney et Fromm (2014) se sont focalisés sur le domaine de l'inégalité économique dans le but d'avoir un système qui permet la détection automatique du domaine source et cible. Les auteurs ont utilisé l'outil *SketchEngine* afin de construire des collections d'exemples métaphoriques pour chaque langue. L'évaluation a été effectuée

pour l'anglais uniquement et le système proposé *"WordSketch"* a obtenu des valeurs de précision de 0,98 et de rappel de 0,86, surpassant les résultats obtenus avec les méthodes CSF (Tsvetkov *et al.*, 2014), TRIPS (Wilks, 1978), VerbNet (Baker *et al.*, 2003) et l'ontologie construite dans le cadre du projet *Scone*[11].

Dans le même cadre du traitement de la métaphore dans un corpus multilingue, Tsvetkov et al. (2014) ont exploité un nouveau corpus pour les même langues que Macwhinney et Fromm (2014) afin de proposer une approche pour la détection automatique de la métaphore qui se manifeste par les deux structures syntaxiques suivantes : sujet-verbe-objet (SVO) et adjectif-nom (AN). L'approche proposée repose sur trois catégories de traits :

1. **Abstraction et imagibilité :** la plupart des choses abstraites sont difficiles à visualiser. Ces traits se sont révélés être utiles dans la détection des métaphores.

2. **Supersenses :** ce sont des catégories sémantiques grossières venant de*WordNet* (15 classes pour les verbes et 26 classes pour les noms).

3. **Vecteurs représentant les mots :** ils représentent les mots sous forme de vecteur en utilisant des algorithmes non supervisé.

L'exploitation de cette approche sur le corpus anglais a donné une exactitude pour la détection de la métaphore avec SVO de 82% et avec AN de 86%.

Huang (2014) a traité un type particulier de métaphore dans les réseaux sociaux à savoir *la métaphore non-conventionnalisée (non stylisée)*. Il a collecté un corpus de messages de soutien de malades du cancer du sein (et leurss profils utilisateurs publics) publiés sur le site web *Breastcancer.org*. Ce corpus a été utilisé afin d'implémenter un modèle qui utilise l'outil *JGibbLDA* [12]. Les performances de ce modèle ne sont pas connues.

Jang et al. (2015a) ont réutilisé le corpus collecté par (Huang, 2014) dans le but de détecter la métaphore en exploitant le contexte global du discours. Ils ont proposé une approche à base de traits contextuels globaux (catégorie sémantique, distribution des topiques, chaines lexicales, présence de mots de contexte) et locaux (catégorie sémantique, proximité sémantique, abstraction lexicale, relations de dépendance). Pour la classification, ils ont utilisé la régression logistique. Les résultats ont montré que les traits contextuels locaux sont plus performants que les traits contextuels globaux avec une exactitude de 86,3%.

[11] http://www.cs.cmu.edu/~sef/scone/

[12] Une implémentation Java de Latent Dirichlet Allocation (LDA) en utilisant Gibbs Échantillonnage pour l'estimation des paramètres et l'inférence : http://jgibblda.sourceforge.net/

2.4.3 Approches pragmatiques exploitant le contexte externe de l'énoncé

Jang et al. (2015b) ont réutilisé le même corpus utilisé dans (Jang *et al.*, 2015a) et (Huang, 2014) afin d'étudier l'influence des facteurs situationnels (les évènements sur le cancer : diagnostic, chimiothérapie, etc.) sur la détection de la métaphore. Ils ont appliqué l'approche proposée par (Wen *et al.*, 2013) afin d'extraire les dates des évènements relatifs au cancer pour chacun des utilisateurs à partir de leurs historiques des messages publiés. Ils ont ainsi défini une liste de termes qui ont été utilisés soit métaphoriquement soit littéralement dans le corpus d'étude. Pour la tâche de classification, un classifieur SVM avec les traits suivants a été utilisé : (1) un trait binaire qui indique si un message a été publié pendant la période critique de chaque évènement, (2) un trait qui indique le nombre de mois qui séparent la date d'un message et la date relative à l'évènement sujet du message et (3) un trait binaire qui indique si un message appartient ou pas à la période critique de l'un des évènements associés à une métaphore donnée. La meilleure exactitude (83,36%) a été obtenue en combinant les traits (1) et (2) avec les unigrammes.

Do Dinh and Gurevych (2016) ont proposé une approche pour la détection de la métaphore en exploitant des réseaux de neurones et des représentations vectorielles de mots. A cet effet, ils ont exploité le perceptron multicouche (MLP) de type *feedforward*. Ils ont traité la problématique de la détection de la métaphore comme étant un problème d'étiquetage. Par conséquent, ils ont repris et étendu le modèle de la reconnaissance des entités nommées qui a été construit en utilisant la bibliothèque *Python deep learning library Theano* élaboré par Bastien et al. (2012) dans le cadre du projet *Theano*. Pour l'apprentissage du réseau, ils ont exploité l'algorithme *Stochastic gradient descent* (SGD) avec la vraisemblance logarithmique (log-likelihood). Les expériences ont été réalisées en utilisant des mots pré-entraînés à 300 dimensions créés avec *word2vec* [13] sur l'ensemble de données de *Google News*. Les corpus d'apprentissage et de test ont été sélectionnés à partir du corpus *VU Amsterdam Metaphor Corpus* (VUAMC) [14] dans lequel chaque mot est annoté avec le sens littéral et le sens métaphorique. Ils ont obtenu une f-mesure de 56,18%.

Su et al. (2016) ont proposé une approche pour la détection automatique des références métaphoriques nominales et pour l'interprétation des métaphores en exploitant la connexité sémantique. Ils s'appuient sur le fait que la métaphore nominale se compose d'un domaine source et un domaine cible et que ces deux domaines sont moins liés sémantiquement dans le cas métaphorique que dans le cas littéral. Une étape de localisation des concepts ainsi que de calcul de la connexité sémantique entre les concepts est nécessaire pour la détection et l'interprétation de la métaphore. Chaque mot/concept est représenté par un vecteur et la connexité sémantique est calculée en comparant les vecteurs relatifs aux concepts avec la valeur de *similarité cosinus*. Après avoir comparé la connexité sémantique des deux concepts, le système interroge *WordNet* afin de vérifier l'existence ou pas d'une relation d'hyponymie

[13] https://code.google.com/p/word2vec/
[14] http://www.vismet.org/metcor/search/showPage.php?page=start

ou hyperonymie entre les deux concepts. Si une telle relation existe alors le système considère que l'utilisation de ces deux concepts dans la même phrase a un sens littéral et donc est non métaphorique. L'approche proposée a été testée sur deux corpus différents, le premier est un corpus en chinois nommé *Reader Corpus*[15] alors que le second est un corpus en anglais collecté à partir de *BNC Corpus*. Les meilleures résultats obtenus pour la détection automatique de la métaphore en termes d'exactitude sont de 0,850 pour le chinois et 0,852 pour l'anglais.

Le deuxième défi évoqué par (Su *et al.*, 2016) est l'automatisation de l'interprétation de la métaphore. En partant de l'hypothèse suivante : l'interprétation de la métaphore repose sur la traduction abstraite d'une expression (paraphrase), (Su *et al.*, 2016) supposent que les domaines source et cible d'une métaphore proviennent de deux domaines différents mais qui contiennent des similitudes. Autrement dit, une interprétation métaphorique est une coopération entre les domaines source et cible de trois manières : (1) la source et la cible partagent des propriétés communes ; (2) les propriétés de la source et de la cible présentent certaines similitudes ; (3) la cible correspond à une des propriétés du domaine source. L'ensemble des propriétés du domaine source ont été extraites à partir de *Property Database*[16] et *Sardonicus*[17]. Le corpus de test comportent 100 usages métaphoriques en chinois et 100 usages métaphoriques en anglais collectés à partir du web, de journaux, blogs, et livres. L'évaluation de l'interprétation a été effectuée par cinq annotateurs humains en attribuant une valeur de 1 (fortement non acceptable) à 5 (fortement acceptable). Etant donné que l'accord inter-annotateur est de $kappa = 0,39$, toutes les évaluations ayant une valeur d'acceptabilité au dessous de 3 ont été considérées comme fausses et éliminées. Ceci a permis d'avoir une valeur d'exactitude égale à 87% pour le chinois et de 85% pour l'anglais.

Les travaux actuels ne se limitent pas à la détection de la métaphore uniquement mais à l'exploitation de la détection de la métaphore pour la réalisation de tâches plus complexes comme la détection des événements par exemple. Ceci a fait l'objet des travaux de (Goode *et al.*, 2017) qui ont étudié le comportement des blogs ainsi que des métaphores afin de générer des signaux pour la détection des événements. Ils ont ainsi exploité un corpus de 589 089 documents collectés à partir des blogs politiques d'Amérique Latine. Les métaphores présentes dans le corpus ont été identifiées à l'aide du système de détection de métaphores développé dans le cadre du projet *IARPA*[18]. En revanche, la détection des événements a été effectuée en exploitant trois traits à savoir : (1) le nombre de mots ; (2) la fréquence de publication et (3) la fréquence de l'utilisation d'une métaphore politique donnée. Par conséquent, les blogs ayant un comportement de regroupement élevé sont plus susceptibles de coïncider avec des événements d'intérêt que ceux ayant un taux de publication constant. Autrement dit, le taux de publication élevé dans un blog à une date précise peut être une indice de la

[15] www.duzhe.com
[16] Une base développée par le NLP Lab de l'université de Xiamen
[17] http://afflatus.ucd.ie/sardonicus/tree.jsp
[18] http://www.iarpa.gov/index.php/research-programs/metaphor

présence d'un événement important.

2.5 Détection automatique de la comparaison

Proche de la métaphore, Mpouli et Ganascia (2015) ont étudié un autre type de langage figuratif qui se manifeste par les comparaisons. La différence entre une métaphore et une comparaison est le fait que la comparaison utilise explicitement des mots qui expriment une comparaison (voir chapitre 1). Mpouli and Ganascia (2015) ont ainsi proposé un algorithme qui utilise un analyseur syntaxique de surface (chunker) et des règles manuelles afin d'extraire et d'analyser les pseudo-comparaisons présentes dans un texte. La reconnaissance des comparaisons figuratives dans un texte a été effectuée en suivant trois étapes à savoir : (1) l'extraction des structures comparatives et pseudo-comparatives contenues dans un texte, (2) l'identification des constituants de ces structures, et (3) la désambiguïsation de ces structures. Deux types de comparaisons figuratives ont été étudiés : (type I) les comparaisons qui sont introduites par des comparatifs (*comme, tel, ainsi que, de même que, ...*) et (type II) les comparaisons qui reposent sur des adjectifs (*semblable à, pareil à, similaire, ...*), des verbes (*ressembler à, sembler, faire l'effet de, faire penser à, ...*), des suffixes ou des locutions prépositionnelles (*à la manière de, à l'image de, ...*). Seules les structures de la forme « marqueur + syntagme nominal » ou « marqueur, ... , syntagme nominal » dans lesquelles le comparant n'est pas un sujet, sont extraites. Afin de tester l'algorithme proposé, un corpus de poèmes en prose a été exploité. Ce corpus a été annoté manuellement. L'algorithme proposé a permis d'obtenir des résultats meilleurs par rapport à ceux obtenus par *Berkeley Parser* au niveau de la détection des verbes et des comparants (respectivement $Précision = 52,8\%$ et $Précision = 96,7\%$) mais des résultats moins bons pour la détection des comparés et des adjectifs.

2.6 Détection automatique de l'humour

Selon la définition du dictionnaire *Larousse*, l'humour est : *« une forme d'esprit railleuse qui s'attache à souligner le caractère comique, ridicule, absurde ou insolite de certains aspects de la réalité »*. La détection de l'humour a fait l'objet de nombreux travaux que nous présentons ici (Purandare & Litman, 2006; Mihalcea & Strapparava, 2006; Sjöbergh & Araki, 2007; Taylor, 2009; Raz, 2012; Radev *et al.*, 2015; Yang *et al.*, 2015; Bertero *et al.*, 2016; Bertero & Fung, 2016).

Purandare et Litman (2006) ont analysé les conversations de la série « FRIENDS »[19], en examinant les caractéristiques acoustiques, prosodiques et linguistiques, et leur utilité

[19] http://www.friendscafe.org/scripts.shtml

dans la reconnaissance automatique de l'humour. Ils ont exploité un schéma d'annotation simple qui permet d'annoter automatiquement les passages suivis pas des rires comme étant humoristiques (43,8% des passages sont humoristiques). Ils ont défini un ensemble de traits acoustiques et prosodiques (pitch, énergie, temporel) et d'autres traits (lexicaux, nombre de mots, intervenant). L'automatisation de la classification des données a été effectuée en utilisant un apprentissage supervisé un arbre de décision. En exploitant l'ensemble des traits, la valeur d'exactitude obtenue est de 64%.

Bertero et al. (2016) ont présenté une comparaison entre différents méthodes d'apprentissage supervisé pour la détection de l'humour dans un corpus humoristique composé des enregistrements audio de la série « The Big Bang Theory »[20]. Deux ensembles de traits ont été définis : des traits acoustiques et des traits linguistiques (lexique, syntaxe, structure, sentiment, antonymes, intervenant). Ces traits ont été exploités à travers trois classifieurs : *Conditional Random Field* (CFR), *Recurrent Neural Network* (RNN) et *Convolutional Neural Network* (CNN). Les meilleures résultats ont été obtenus par le classifieur CNN avec une f-mesure de 68,5% et une exactitude de 73,8%.

Mihalcea et Strapparava (2006) se sont inspirés des caractéristiques de l'humour étudié par les linguistes. Un corpus de 16 000 phrases courtes humoristiques a été collecté sur le Web ainsi qu'un corpus non humoristique. Ils ont obtenu une exactitude de 96,95% en utilisant un classifieur Naïve Bayes avec des traits stylistiques propres à l'humour (allitération, antonymie, argot) et des traits basés sur le contenu.

Certains travaux ont prouvé que la reconnaissance de l'humour ne nécessite pas une compréhension du sens. L'exploitation d'un ensemble de traits de surface est suffisant pour la détection automatique de l'humour. Parmi ces travaux, Sjöbergh et Araki (2007) ont exploité leur propre algorithme de classification qui consiste à : pour chaque trait, une valeur seuil est calculée dans le but de séparer les exemples d'apprentissage en deux groupes (humoristique et non humoristique). Le seuil doit rendre l'entropie moyenne le plus faible possible. Pour classer un nouvel exemple, une vérification de la présence des traits dans l'exemple est faite ainsi que les deux groupes d'exemples. Si la correspondance des traits est plus grande entre l'exemple et le groupe humoristique alors l'exemple est considéré comme humoristique et vice versa. Les traits proposés ont été regroupés en cinq groupes à savoir : *similitude de texte, mots de plaisanterie* (mots communs dans les documents humoristiques), *ambiguïté des mots* (nombre moyen de sens des mots), *style* (négation, répétition, pronoms, antonymes...) et *expressions idiomatiques*. Une exactitude de 85,4% a été obtenue.

En parallèle de tous les travaux qui s'intéressent au langage figuratif et avec l'apparition des réseaux sociaux, Raz (2012) a proposé une approche pour la détection automatique de l'humour dans un corpus de tweets. Un corpus de tweets humoristiques a été collecté à partir d'un site web contenant des tweets humoristiques[21]. L'auteur a proposé un ensemble de

[20] https://bigbangtrans.wordpress.com
[21] http://www.funny-tweets.com

traits syntaxiques, lexicaux, morphologiques (temps des verbes, ...), phonologiques (homophones pour reconnaître les jeux de mots), pragmatiques (nombre de résultats renvoyés par un moteur de recherche pour une requête sur les verbes présents dans le tweet) et stylistiques (émoticônes, ponctuation). Cette approche n'a malheureusement pas été évaluée.

Radev et al. (2015) ont quant à eux choisi un type particulier de corpus : le corpus choisi est composé de 298 224 légendes de bandes dessinées publiées dans *The New Yorker Cartoon*. Les auteurs ont développé plus d'une douzaine de méthodes non supervisées afin de classer les légendes. Le premier groupe de méthodes représente les méthodes à base d'originalité : par exemple, l'algorithme *LexRank* [22] est utilisé afin d'identifier la légende la plus centrale ; la méthode de classification *Louvain* à base de graphe proposé par (CAMPIGOTTO *et al.*, 2014) qui permet le regroupement des légendes par thématique. Le deuxième groupe est formé des méthodes à bases de contenu : par exemple, *Freebase* [23] est utilisé afin d'annoter les phrases nominales dans les légendes ; la polarité est annotée avec *Stanford CoreNLP* [24] . Enfin, le troisième groupe représente les méthodes génériques : utilisation de la complexité syntaxique en utilisant la méthode proposé par (Charniak & Johnson, 2005). L'évaluation de ces trois méthodes a été réalisée en utilisant *Amazon Mechanical Turk* (AMT). Chaque micro-tâche AMT consistait en un dessin animé ainsi que deux légendes (A et B). Les annotateurs ont été chargés de déterminer la légende la plus drôle. Les résultats ont montré que les méthodes qui reposent sur les sentiments négatifs et la centralité lexicale sont les plus performantes pour la détection des légendes les plus drôles.

Yang et al. (2015) ont utilisé le corpus de (Mihalcea & Strapparava, 2005) et ont proposé quatre ensembles de traits en respectant la structure sémantique latente suivante : *incongruité* (déconnexion, répétition), *ambiguïté* (sens possibles à travers WordNet), *effet interpersonnel* (polarité positive/négative, subjectivité faible/forte), *style phonétique* (allitération, rime). En exploitant les traits ci-dessus, les auteurs ont utilisé l'algorithme de classification *Random Forest*. Une exactiture de 85,4% a été obtenue. Les auteurs en déduisent que la détection de l'humour et des marqueurs nécessite une bonne compréhension du sens de la phrase ainsi que des connaissance externes.

2.7 Bilan et positionnement de nos travaux

Dans ce chapitre, nous avons présenté un état de l'art sur la détection automatique du langage figuratif en se focalisant d'une part sur les travaux proposant des schémas d'annotation et d'autre part sur les travaux qui ont proposé des approches psycholinguistiques ou automa-

[22] http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume22/erkan04a-html/erkan04a.html

[23] https://en.wikipedia.org/wiki/Freebase

[24] http://stanfordnlp.github.io/CoreNLP/

tiques pour la détection de l'ironie, du sarcasme, de la satire (cf. section 4.5), la métaphore (cf. section 2.4), la comparaison (cf. section 2.5) et l'humour (cf. section 2.6).

L'ensemble de ces travaux ont prouvé que l'analyse automatique de l'ironie et du sarcasme est l'un des défis majeurs du Traitement Automatique des Langues. La plupart des travaux récents se focalisent sur la détection de ce phénomène dans des corpus issus de réseaux sociaux comme Twitter car dans un tweet, l'auteur peut utiliser des hashtags spécifiques (comme *#ironie, #sarcasme*) afin de guider le lecteur dans la compréhension de la forme imagée qu'il veut exprimer (Gonzalez-Ibanez *et al.*, 2011; Reyes *et al.*, 2013; Barbieri & Saggion, 2014a; Barbieri *et al.*, 2014; Joshi *et al.*, 2015; Bamman & Smith, 2015). Ces hashtags sont très précieux car ils permettent aux chercheurs d'avoir des corpus annotés qui seront utilisés par des systèmes d'apprentissage automatique pour classer un tweet dans la classe ironique ou non ironique.

Les méthodes utilisées dans l'état de l'art pour détecter l'ironie s'appuient essentiellement sur le contenu linguistique des textes, comme la présence de ponctuations, d'émoticônes, de mots d'opinions positifs ou négatifs, etc. (Burfoot & Baldwin, 2009; Tsur *et al.*, 2010; Gonzalez-Ibanez *et al.*, 2011; Reyes *et al.*, 2013; Barbieri *et al.*, 2014). Cependant, ces méthodes atteignent vite leurs limites lorsque la compréhension du message ironique nécessite des connaissances pragmatiques ou extra-linguistiques. Par conséquent, des approches pragmatiques exploitant le contexte externe à l'énoncé ont vu le jour à partir de 2015 afin de résoudre cette problématique (Wallace, 2015; Bamman & Smith, 2015; Joshi *et al.*, 2016).

Dans ce cadre, nous proposons une approche par apprentissage supervisé afin de prédire si un tweet est ironique ou pas. Pour ce faire, nous avons suivi une démarche en trois étapes :

1. Analyse des phénomènes pragmatiques utilisés pour exprimer l'ironie en nous inspirant des travaux en linguistique afin de définir un schéma d'annotation multi-niveaux pour l'ironie (cf. chapitre 3).

2. En exploitant l'ensemble des observations faites sur le corpus annoté, développement d'un modèle de détection automatique pour les tweets en français qui exploite à la fois le contexte interne du tweet à travers des traits lexicaux et sémantiques et le contexte externe en recherchant des informations disponibles sur le Web (cf. chapitre 4).

3. Enfin, étude de la portabilité du modèle pour la détection de l'ironie dans un cadre multilingue (italien, anglais et arabe). Nous avons ainsi testé la portabilité du schéma d'annotation proposé sur l'italien et l'anglais et testé la performance du modèle de détection automatique à base de traits sur la langue arabe (cf. chapitre 5).

Chapitre 3

Un schéma multi-niveaux pour l'annotation de l'ironie dans les réseaux sociaux

3.1 Introduction

L'objectif de ce chapitre est de proposer un schéma d'annotation pour l'ironie pour un type particulier de textes à savoir les tweets. Dans la section 2.2 du chapitre 2, nous avons fait un état des lieux des différents schémas d'annotation proposés pour l'annotation des tweets ironiques en italien et en anglais (Gianti *et al.*, 2012; Shutova *et al.*, 2013; Van Hee *et al.*, 2016). Nous avons vu que ces schémas ont tous un point commun : ils s'intéressent à caractériser l'ironie dans les tweets d'une façon globale sans se préoccuper ni des indices linguistiques au niveau du message ni même des indices extra-linguistiques. Ainsi, ces schémas comportent pour la plupart un seul niveau d'annotation qui permet de caractériser un tweet selon son type figuratif (ironique vs. non ironique), sa polarité (positif, négatif, neutre), ou encore, dans une moindre mesure, le dispositif pragmatique déclencheur de l'ironie (inversement de polarité, hyperbole, euphémisme).

Comparés aux travaux menés dans la littérature linguistique sur les différents marqueurs de l'ironie verbale dans les textes de type poème, roman, etc. (cf. section 3.3), il est évident que les travaux computationnels traitant l'ironie dans les réseaux sociaux ont étudié la problématique d'une manière superficielle sans entrer dans les spécificités de l'ironie étudiées par les linguistes. Notre objectif dans cette thèse est d'aller plus loin afin étudier finement ces différents marqueurs et répondre aux questions suivantes : *Est-ce que les différents types d'ironie étudiés dans les travaux en linguistique peuvent être présents dans un corpus particulier collecté à partir des réseaux sociaux comme Twitter ? Si oui, quels sont les types les plus fréquents ? Ces types sont-ils explicitement marqués ? Quelles sont les corrélations*

entre les types d'ironie et ces marqueurs ? Comment pouvons-nous exploiter ces corrélations dans un but de détection automatique ?

Pour atteindre notre objectif, nous avons d'abord analysé les différentes catégories d'ironie proposées dans la littérature linguistique (Attardo, 2000b; Ritchie, 2005; Didio, 2007; Burgers, 2010) afin de retenir uniquement celles qui sont les plus appropriées pour l'analyse de l'ironie dans les tweets. Pour les caractériser et quantifier leur pertinence pour une détection automatique, nous proposons pour la première fois un nouveau schéma d'annotation multi-niveaux et un corpus de tweets annotés selon ce schéma. Ce schéma, inspiré des travaux linguistiques sur l'ironie, vise à une étude approfondie de l'expression de l'ironie dans les réseaux sociaux à différents grains :

- ***Au niveau du message dans sa globalité*** : ironique vs. non ironique.
- ***Au niveau du type de l'ironie*** : explicite vs. implicite. Ce qui permet d'appréhender l'importance du contexte dans la compréhension du langage figuratif.
- ***Au niveau des catégories de l'ironie*** : chaque type d'ironie est ainsi associé à une ou plusieurs catégories qui rendent compte des phénomènes pragmatiques mis en jeu lors de la production de l'ironie verbale.
- ***Au niveau des indices linguistiques*** : chaque catégorie d'ironie peut être déclenchée par un ensemble de marqueurs linguistiques spécifiques.

Ce chapitre est organisé comme suit. Nous commençons par présenter en section 3.2 notre corpus d'étude, nommé FrIC (*French Irony Corpus*), puis en section 3.3, notre schéma d'annotation. Nous détaillons non seulement les catégories d'ironie retenues pour l'annotation mais également celles proposées par les linguistes. En section 3.4, nous décrivons la campagne d'annotation, puis nous enchaînons par la présentation des résultats quantitatifs et qualitatifs de cette campagne. Nous nous intéressons en particulier à étudier les interactions entre : (1) les types d'activation d'ironie et les marqueurs, (2) les catégories d'ironie et les marqueurs, et (3) l'impact des connaissances externes sur la détection d'ironie. Nos résultats démontrent que l'activation implicite de l'ironie est un défi majeur pour les systèmes futurs.

Une partie du corpus FrIC a été utilisée dans le cadre de la première campagne d'évaluation sur l'analyse d'opinion et le langage figuratif DEFT@TALN 2017 que nous avons co-organisée en collaboration avec le LIMSI[1] (Benamara *et al.*, 2017a).

3.2 Le corpus FrIC

Étant donnée l'absence d'un corpus de tweets ironiques en français, nous avons commencé par la construction d'un corpus de tweets ironiques et non ironiques. Dans un premier temps,

[1] https://deft.limsi.fr/2017/

nous considérons comme ironiques les tweets contenant les hashtags #*ironie* ou #*sarcasme*, les autres sont considérés comme non ironiques.

Pour la collecte des tweets, nous avons d'abord sélectionné un ensemble de thèmes discutés dans les médias du printemps 2014 jusqu'à l'automne 2016. Notre intuition derrière le choix de ces thèmes est que le contexte pragmatique nécessaire pour inférer l'ironie est plus susceptible d'être compris par les annotateurs s'il fait référence à des faits d'actualité connus par rapport aux contextes de tweets personnels.

Nous avons choisi 186 thèmes répartis en 9 catégories (politique, sport, musique, etc.). Pour chaque thème, nous avons sélectionné un ensemble de mots-clés avec et sans hashtag, par exemple : politique (*Sarkozy, #Hollande, UMP, ...*), santé (*cancer, grippe*), sport (*#Zlatan, #FIFAworldcup, ...*), médias sociaux (*#Facebook, Skype, MSN*), artistes (*Rihanna, Beyoncé, ...*), télévision (*TheVoice, XFactor*), pays ou villes (*Corée du Nord, Brésil, ...*), Printemps Arabe (*Marzouki, Ben Ali, ...*) et d'autres thèmes plus génériques (*pollution, racisme*). Nous avons ensuite sélectionné des tweets ironiques contenant les mots-clés, le hashtag #*ironie* ou #*sarcasme*. De la même manière, nous avons aussi sélectionné des tweets non ironiques (*i.e.* ne contenant pas #*ironie* or #*sarcasme*).

Pour la collecte du corpus, nous avons utilisé l'API de Twitter. Une fois les tweets collectés, nous avons supprimé les doublons, les retweets et les tweets contenant des images car nous estimons que ces derniers sont plus susceptibles de contenir de l'ironie de situation (illustrée par les images), plus difficile à détecter automatiquement. Après cette étape de filtrage, nous avons obtenu un corpus de 18 252 tweets répartis comme suit : 2 073 tweets ironiques et 16 179 tweets non ironiques (Tableau 3.1). Pour les expériences décrites par la suite, les hashtags #*ironie* et #*sarcasme* ont été supprimés des tweets.

Thèmes	*Ironique*	*Non ironique*
Émissions de télé	81	3 060
Économie	85	273
Générique	189	777
Villes ou Pays	245	805
Artistes	232	192
Politique	1 035	10 629
Réseaux sociaux	19	0
Santé	3	32
Sport	178	411
Total	**2 073**	**16 179**

TABLE 3.1 : Répartition des tweets dans le corpus FrIC.

Afin de vérifier la fiabilité des hashtags *#ironie* et *#sarcasme*, une première étape d'an-

notation d'un sous-ensemble du corpus a été nécessaire. Cette tâche a été réalisée par deux annotateurs humains qui ont annoté manuellement 100 tweets (50 tweets ironiques et 50 tweets non ironiques mélangés) après avoir supprimé les hashtags *#ironie* et *#sarcasme*. La phase d'annotation a résulté en un kappa de Cohen de $\kappa = 0,78$ en comparant les annotations par rapport aux hashtags de référence. Ce résultat montre que ces hashtags sont relativement fiables.

La plupart des désaccords entre les deux annotateurs provient de la présence de négation (Exemple 3.1) ou de la nécessité de connaissances externes au texte du tweet afin de comprendre le sens ironique (Exemple 3.2).

(3.1) C'est chez Hollande qu'il y a du Berlusconi vous ne trouvez pas. Un côté bounga-bounga non ?

(3.2) Qu'est-ce qui pourrait détruire notre monde ? — La Corée du Nord

Notons que la plupart des travaux de la littérature ont collecté des corpus de tweets ironiques/sarcastiques en se basant sur les hashtags *#ironie* ou *#sarcasme* sans vérification de la fiabilité de ces hashtags (Hee *et al.*, 2016) .

3.3 Schéma d'annotation multi-niveaux

3.3.1 Méthodologie

Pour définir notre schéma, la première étape a été d'étudier les différents marqueurs de l'ironie étudiés dans la littérature linguistique. Plus de 16 marqueurs ont été proposés comme la contre-vérité, l'exagération, l'exclamation, etc. (Tayot, 1984; Attardo, 2000b; Mercier-Leca, 2003; Ritchie, 2005; Didio, 2007; Burgers, 2010). Le tableau 3.2 présente une synthèse de ces marqueurs en se focalisant sur un type particulier d'ironie à savoir l'ironie textuelle. Dans ce tableau, nous donnons pour chaque catégorie la liste des références bibliographiques associées ainsi qu'une ou deux définitions citées dans la littérature linguistique. Nous détaillons l'ensemble des définitions citées par les linguistes dans l'annexe A.1.

Il est important de noter que les catégories de l'ironie présentées dans le tableau 3.2 ont principalement été identifiées dans des textes littéraires (livres, poèmes, etc.). Une première étape a donc été de vérifier leur présence dans un échantillon de 300 tweets de notre corpus. Quatre principales observations résultent de cette analyse :

1. Selon la définition générale, l'ironie exprime une contradiction entre ce qui est dit et ce qui est signifié. Dans les tweets, nous avons constaté que les utilisateurs utilisent

Marqueurs de l'ironie	**Références**	**Définitions dans un usage ironique**
Métaphore	(Grice, 1970; Kittay, 1990; Song, 1998)	Selon (Kittay, 1990), l'ironie peut être exprimée par la métaphore qui est un sens du second ordre qui est obtenu lorsque les caractéristiques de l'énoncé et de son contexte indiquent à l'auditeur ou le lecteur que le sens du premier ordre de l'expression est indisponible ou ne convient pas.
Hyperbole	(Kreuz & Roberts, 1993; Pougeoise, 2001; Mercier-Leca, 2003; Didio, 2007)	Selon (Didio, 2007), l'ironie peut être exprimée par l'hyperbole, une figure qui augmente les choses avec excès, avec exagération.
Exagération	(Didio, 2007)	L'exagération est une figure qui amplifie la réalité ou bien elle la présente en lui donnant plus d'importance qu'elle n'en a réellement.
Euphémisme	(Muecke, 1978; Fromilhague, 1995; Seto, 1998; Yamanashi, 1998; Mercier-Leca, 2003)	Selon (Muecke, 1978; Seto, 1998), l'euphémisme est une figure de style qui consiste à atténuer l'expression de faits ou d'idées considérés comme désagréables dans le but d'adoucir la réalité.
Question rhétorique	(Muecke, 1978; Barbe, 1995; Burgers, 2010)	Selon (Burgers, 2010), une question rhétorique n'est pas une question réelle : c'est une question pour laquelle le locuteur ne prévoit pas de recevoir une réponse, parce que la réponse est déjà connue. Cela signifie qu'une question rhétorique représente un point de vue et non pas une question.
Changement de registre	(Attardo, 2000b; Haiman, 2001; Burgers, 2010)	Selon (Burgers, 2010), un changement de registre est un changement soudain dans le style. Dans un énoncé, le changement de registre est exprimé par l'utilisation de mots inattendus d'un autre registre (dans un texte formel on utilise soudain des mots informels ou vice versa). Il se manifeste aussi par le changement brusque de sujet de la phrase ou l'utilisation exagérée de la politesse dans une situation où ceci est inapproprié.
Fausse logique / contradiction	(Tayot, 1984; Barbe, 1995; Didio, 2007)	Selon (Didio, 2007), les contradictions dans un discours permettent à l'énonciataire de comprendre le sens ironique d'un texte en partant de l'idée que la contradiction unit deux énoncés qui affirment et nient un même objet de connaissance.
Oxymore	(Gibbs, 1994; Song, 1998; Mercier-Leca, 2003)	Selon (Gibbs, 1994; Song, 1998; Mercier-Leca, 2003), l'oxymore est une figure de construction qui repose sur une apparente contradiction logique. C'est une figure d'opposition. Elle se repère au niveau de l'énoncé par le rapprochement syntaxique de deux éléments qui forment une contradiction sémantique.
Paradoxe	(Tayot, 1984; Barbe, 1995; Mercier-Leca, 2003)	Selon (Mercier-Leca, 2003), l'ironie repose sur un paradoxe dont le caractère frappant est accentué par la syntaxe asyndétique (économisant les liens logiques), qui, par contraste, met en valeur la seule conjonction de coordination présente dans un énoncé par exemple le « mais ».
Absurdité	(Didio, 2007)	Selon (Didio, 2007), l'absurdité est exprimée par un raisonnement illogique. L'absurde peut être lié à une réaction comique ou tragique. Il signifie ce qui n'est pas en harmonie avec quelqu'un ou quelque chose.
Effet de surprise	(Colston & Keller, 1998; Didio, 2007)	Selon (Colston & Keller, 1998), la surprise est une réaction courante lorsque les événements ne se passent pas comme prévu. Les gens peuvent exprimer cette surprise en notant verbalement le contraste entre ce qui était attendu et ce qui est réellement arrivé.
Répétition	(Muecke, 1978; Berntsen & Kennedy, 1996; Burgers, 2010)	Selon (Burgers, 2010), un écrivain peut ironiquement répéter quelque chose prononcé par une autre personne plus tôt dans le texte ou en cas d'interaction orale - dans le dialogue. Ce type de répétition est appelé une répétition sur la base de co-texte. Dans ce cas, un énoncé ou une partie d'un énoncé est ironiquement répété dans le même texte (qui n'a pas été utilisé ironiquement dans son premier usage).
Guillemets	(Tayot, 1984; Gibbs, 1994; Attardo, 2001; Burgers, 2010)	Selon (Gibbs, 1994), l'utilisation des guillemets est un geste non verbal dans le vocabulaire de nombreux orateurs américains dans le but d'exprimer l'ironie. L'utilisation des guillemets annonce que l'orateur va imiter le discours ou l'état d'esprit de la personne citée, souvent pour obtenir un effet sarcastique.
Emoticônes	(Tayot, 1984; Kreuz, 1996; Burgers, 2010)	Selon (Tayot, 1984), une émoticône indique que l'intonation, la mimo-gestualité (par exemple le "tongue in cheek" britannique, ou encore le clin d'oeil français) marquent parfois l'ironie orale.
Exclamation	(Attardo, 2001; Didio, 2007; Burgers, 2010)	Selon (Attardo, 2001; Didio, 2007; Burgers, 2010), une ironie peut être marquée par l'exclamation à l'oral et par le point d'exclamation à l'écrit.
Majuscule	(Haiman, 1998; Burgers, 2010)	-
Texte barré et caractères spéciaux	(Burgers, 2010)	-

TABLE 3.2 : Les différents marqueurs de l'ironie étudiés dans la littérature linguistique.

deux mécanismes pour exprimer cette contradiction : (a) recourir exclusivement aux indices lexicaux présents dans le tweet ou (b) combiner ces indices avec un contexte pragmatique externe à l'énoncé. Nous avons alors défini *deux types de contradiction* à savoir *explicite* pour le cas (a) et *implicite* pour le cas (b). Chaque type de contradiction peut s'exprimer par différentes catégories d'ironie.

2. Plusieurs catégories peuvent être regroupées car il est très difficile de les distinguer dans des messages courts, comme par exemple, l'hyperbole et l'exagération, ou la métaphore et la comparaison.

3. Certaines catégories doivent être écartées car trop spécifiques aux textes littéraires, par exemple l'absurdité.

4. Les catégories de l'ironie telles que définies dans la littérature ne peuvent pas être toutes mises au même niveau. Par exemple, les guillemets peuvent être présents dans un tweet ironique de type hyperbole ou de type euphémisme. Nous avons alors décidé de différencier entre *catégories de l'ironie* (hyperbole, euphémisme, question rhétorique, etc.) et *indices de l'ironie* (ponctuation, majuscule, etc.).

Ces observations ont permis de mettre en place trois niveaux d'analyse : type de l'ironie (explicite vs. implicite), catégorie de l'ironie pour chaque type, et indices linguistiques présents dans chaque catégorie. Au final, huit catégories et dix-huit indices ont été retenus, que nous allons détailler dans la section suivante.

3.3.2 Le schéma d'annotation

Le schéma proposé contient quatre niveaux, comme le montre la figure 5.1.

A Niveau 1 : Classe du tweet

Dans ce schéma, on s'intéresse à la classification des tweets en trois classes à savoir :

- **Ironique :** Un tweet est ironique s'il exprime une ironie verbale, ironie situationnelle, sarcasme, satire ou humour (e.g. *Un truc avec DSK, mais quoi ? Aucun site internet n'en parle. Surement parce qu'on ne sait rien de ce qu'il s'est réellement passé ?*).

- **Non ironique :** Un tweet est dit non ironique s'il ne correspond à aucune forme d'ironie citée ci-dessus (e.g. *L'ecotaxe, c'est pour sauver la planète pas pour redresser la France et c'est une idée de Sarko. #idiotie*).

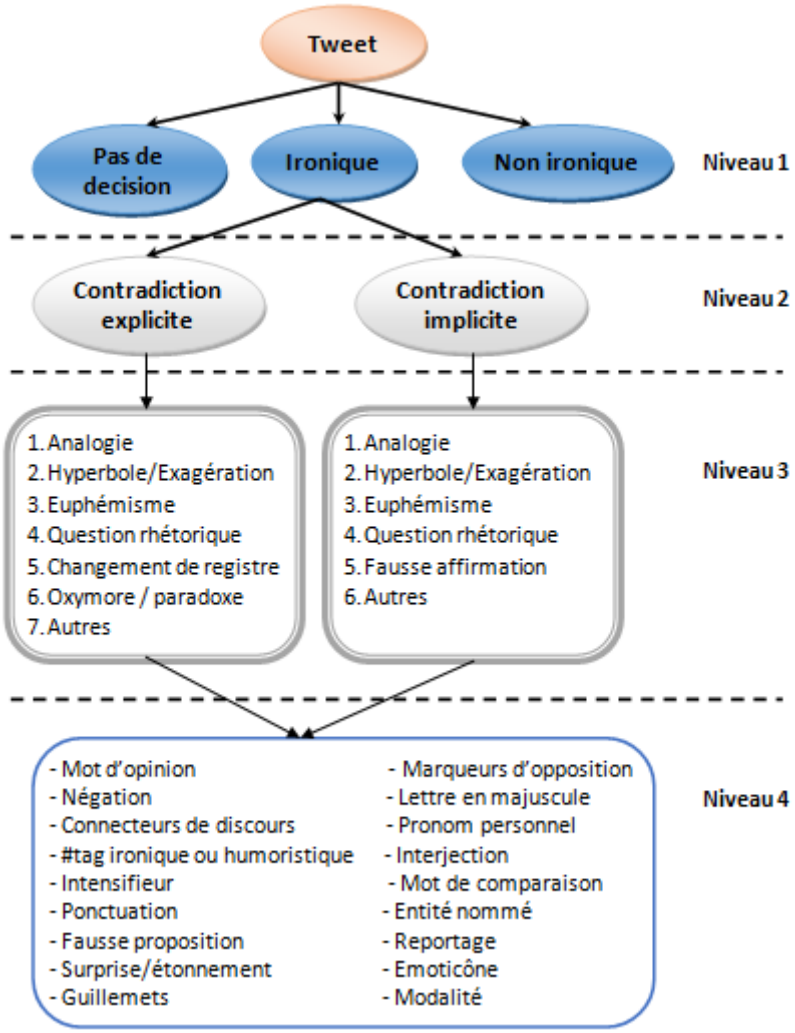

FIGURE 3.1 : Schéma d'annotation.

- **Pas de décision :** Un tweet est classé sous la classe "Pas de décision" dans le cas où on ne parvient pas à décider si un tweet est ironique ou non (e.g. *Si cela ne vous donne pas envie de voter PS en 2012, je ne comprends plus rien à rien*).

B Niveau 2 : Types de l'ironie

L'incongruité dans les énoncés ironiques, et dans les tweets en particulier, consiste souvent en une opposition entre au moins deux propositions (ou deux mots) P_1 et P_2. Les propositions P_1 et P_2 peuvent être toutes deux une partie du contexte interne d'un énoncé (explicitement lexicalisé), ou bien l'une est présente et l'autre implicite. Il existe donc deux

façons de déduire l'ironie des tweets : se fier exclusivement aux indices lexicaux internes à l'énoncé ou bien combiner ces indices avec un contexte pragmatique supplémentaire extérieur à l'énoncé. Ainsi, nous définissons deux types de contradiction : *explicite* et *implicite*.

B.1 Contradiction explicite. La contradiction explicite peut impliquer une contradiction entre les mots d'une proposition P_1 et les mots d'une proposition P_2 qui, soit ont des polarités opposées, comme dans l'exemple 3.3, soit sont sémantiquement non liés, comme dans l'exemple 3.4. Une opposition explicite peut également résulter d'un contraste positif/négatif explicite entre une proposition subjective P_1 et une situation P_2 qui décrit une activité ou un état indésirable. L'ironie est déduite de l'hypothèse suivante : l'auteur et le lecteur partagent une connaissance commune d'une situation qui est jugée négative par les normes culturelles ou sociales. Par exemple, le tweet 3.5 suppose que tout le monde s'attend à ce que son téléphone portable sonne assez fort pour être entendu.

Lors de l'annotation, si l'annotateur n'a pas besoin de connaissances externes pour comprendre la contradiction, le tweet est annoté comme ironique avec une contradiction explicite.

(3.3) [J'adore]$_{P1}$ quand mon téléphone [tombe en panne]$_{P2}$ quand j'en ai besoin.

(3.4) [The Voice]$_{P1}$ est plus important que [Fukushima]$_{P2}$ ce soir.

(3.5) [J'aime]$_{P1}$ quand mon téléphone[baisse le volume automatiquement]$_{P2}$.

B.2 Contradiction implicite. L'activation *implicite* de l'ironie provient d'une contradiction entre une proposition lexicalisée P_1 décrivant un événement ou un état et un contexte pragmatique P_2 externe à l'énoncé dans lequel P_1 est fausse, pas probable ou contraire à l'intention de l'auteur. L'ironie se produit parce que l'auteur croit que son public peut détecter la disparité entre P_1 et P_2 sur la base de connaissances contextuelles ou d'antécédents communs partagés avec lui. Par exemple, dans 3.6, le fait nié dans P_1 permet d'inférer l'ironie dans le tweet.

(3.6) La #NSA a mis sur écoutes un pays entier. Pas d'inquiétude pour la #Belgique : [ce n'est pas un pays entier]$_{P1}$.
$\longrightarrow P2$: Belgique est un pays.

C Niveau 3 : Catégories de l'ironie

Les contradictions *explicites* et *implicites* peuvent être exprimées de différentes manières que nous appelons catégories d'ironie. La littérature linguistique définit plusieurs catégories sur lesquelles nous nous sommes appuyés pour définir les 8 catégories de notre schéma d'annotation : trois peuvent se produire exclusivement dans un type spécifique (marqué *Exp* pour explicite ou *Imp* pour implicite), tandis que cinq sont plus susceptibles d'être trouvées dans les deux types de contradiction (marqués *Exp/Imp*). Ces catégories ne sont pas mutuellement exclusives. Un tweet ironique peut donc être associé à une ou plusieurs catégories.

Le tableau 3.3 résume les principales catégories de la littérature linguistique ainsi que les huit catégories retenues pour l'annotation de tweets.

Nous présentons ici chacune de ces catégories et nous les illustrons par des exemples issus de notre corpus.

C.1 Analogie$^{Exp/Imp}$

L'analogie est un processus de pensée par lequel on remarque une similitude entre deux éléments de différentes natures ou classes. Dans le discours, une analogie explicite est une comparaison (cf. tweet 3.7), tandis qu'une analogie implicite est une métaphore (cf. tweets 3.8 et 3.9).

Dans notre schéma d'annotation, nous utilisons l'*analogie* comme un terme plus large qui couvre l'analogie, la comparaison et la métaphore qui sont des outils qui impliquent une similitude entre deux entités qui ont des concepts, des domaines ou des classes ontologiques différents, sur lesquels une comparaison peut être fondée.

(3.7) *(Exp)* Le dimanche **c'est comme** Benzema en équipe de France : il sert à rien... :D

(3.8) *(Imp)* Pour une fois que je regarde la télé, c'est pour voir **Depardieu en député communiste.** #Savoureux.

(3.9) *(Imp)* On n'avait qu'à écouter ses déclarations des dernières années pour savoir que **Depardieu était en fait très belge**.

C.2 Hyperbole/Exagération$^{Exp/Imp}$

L'hyperbole/exagération est une figure de style qui consiste à exprimer une idée ou un sentiment d'une manière exagérée. Il est souvent utilisé pour faire une impression forte ou pour souligner un point, comme dans les exemples 3.10 et 3.11.

Catégories de l'état de l'art	**Nos catégories**	**Utilisation**
Métaphore (Grice, 1970; Kittay, 1990; Song, 1998)	Analogie$^{Exp/Imp}$ (métaphore et comparaison)	Couvre l'analogie, la comparaison et la métaphore. Implique la similitude entre deux concepts ou deux entités de domaines ontologiques différents, sur lesquels une comparaison peut être fondée.
Hyperbole (Berntsen & Kennedy, 1996; Mercier-Leca, 2003; Didio, 2007)	Hyperbole/ Exagération$^{Exp/Imp}$	Exprime une forte impression ou met l'accent sur un point particulier.
Exagération (Didio, 2007)		
Euphémisme (Muecke, 1978; Seto, 1998)	Euphémisme$^{Exp/Imp}$	Réduit l'effet d'une expression ou une idée considérée comme désagréable pour adoucir la réalité.
Question rhétorique (Barbe, 1995; Berntsen & Kennedy, 1996)	Question Rhétorique$^{Exp/Imp}$	Pose une question afin de mettre en avant un point plutôt que d'obtenir une réponse (P_1 : poser une question avec l'intention d'avoir une réponse, P_2 : aucune intention d'avoir une réponse parce qu'elle est déjà connue).
Changement de registre (Haiman, 2001; Leech, 2016)	Changement de registre^{Exp}	Un changement soudain du sujet / cadre, l'utilisation d'une politesse exagérée dans une situation où cela est inapproprié, etc.
Fausse logique (Didio, 2007)	Fausse affirmation^{Imp}	Une affirmation, un fait ou un événement faux dans la réalité.
Oxymore (Gibbs, 1994; Mercier-Leca, 2003)	Oxymore/ paradoxe^{Exp}	Opposition explicite de deux mots. Le paradoxe est équivalent à la "fausse affirmation" sauf que la contradiction est explicite.
Paradoxe (Tayot, 1984; Barbe, 1995)		
Ironie de situation (Shelley, 2001; Niogret, 2004)	Autres$^{Exp/Imp}$	Ironie humoristique ou situationnelle (ironie où l'incongruité n'est pas dû à l'utilisation de mots mais à une contradiction non intentionnelle entre deux faits ou événements).

TABLE 3.3 : Catégories d'ironie dans notre schéma d'annotation.

(3.10) *(Exp)* Le PS a **tellement bien** réussi que tt va moins bien : pollution, logement, sécurité #PARISledebat #Paris2014

(3.11) *(Imp)* @morandiniblog C'est vrai que **c'est un saint** #Berlusconi, il ne mérite vraiment pas tout cet acharnement...

C.3 Euphémisme$^{Exp/Imp}$

L'euphémisme est une figure de style qui sert à réduire l'effet d'une expression ou une idée considérée comme désagréable pour adoucir la réalité (comme l'utilisation de *moins bien* au lieu de *pire* dans l'exemple 3.10).

C.4 Question rhétorique$^{Exp/Imp}$

La question rhétorique est une figure de style sous la forme d'une question posée afin de présenter un point de vue plutôt que d'obtenir une réponse, comme dans l'exemple 3.12.

(3.12) "Miss France c'est une compétition" **Non sérieux ?** parce que je ne savais pas !

C.5 Changement de registre^{Exp}

Cela se produit par le changement soudain du sujet/cadre dans le tweet, comme dans l'exemple 3.13 où la première phrase concerne le départ de Duflot du gouvernement alors que la seconde concerne le carême.

Le décalage contextuel peut également se produire en utilisant une politesse exagérée dans une situation où cela est inapproprié, comme dans l'exemple 3.14 où l'auteur est trop poli pour une conversation normale entre amis (c'est ce qu'on appelle l'hyperformalité).

Le changement de contexte peut également se produire par l'utilisation de mots polysémiques, comme dans (23) où l'ironie est activée par le contraste entre un contexte où "se rencontrer" a la signification d'interroger une personne impliquée dans une enquête et une autre où "rencontrer" a la signification : "Passer du temps avec une belle femme" (puisque Boschi est connue pour sa beauté).

(3.13) Duflot quitterait le gouvernement. **En plein carême, on ne peut même pas le fêter.** Décidément, elle embête jusqu'au bout... *soupire*

(3.14) Vous pouvez m'accorder l'honneur d'écouter un à l'autre de vos prédictions fines

C.6 Fausse affirmation^{Imp}

Cela indique qu'un fait ou une affirmation ne parvient pas à donner un sens à la réalité. L'auteur exprime le contraire de ce qu'il pense ou quelque chose d'erroné par rapport au contexte. Par conséquent, des connaissances externes à l'énoncé sont nécessaires pour comprendre l'ironie. Par exemple, les tweets 3.15, 3.16 et 3.17 sont ironiques car les situations en caractères gras sont absurdes, fausses ou impossibles en réalité. Notons que le tweet 3.17 est également un exemple de la catégorie *question rhétorique*.

(3.15) La #NSA a mis sur écoutes un pays entier. Pas d'inquiétude pour la #Belgique : **ce n'est pas un pays entier**.

(3.16) @Vince75015 Les agences de notation **ne font pas de politique**.

(3.17) @infos140 @mediapart Serge Dassault ? Corruption ? Non ! Il doit y avoir une erreur. **C'est l'image même de la probité en politique**.

C.7 Oxymore/paradoxe^{Exp}

Cette catégorie est équivalente à la catégorie « Fausse affirmation », sauf que la contradiction est explicite, comme l'utilisation de deux antonymes dans la première phrase de l'exemple 3.10 (*bien réussi* vs *moins bien*) et l'utilisation de deux faits opposés dans l'exemple 3.18.

(3.18) Ben non ! **Matraquer et crever des yeux**, ce **n'est pas violent** et ça respecte les droits ! ! ! #assnat #polqc #ggi

C.8 Autres$^{Exp/Imp}$

Cette catégorie représente les tweets qui sont jugés ironiques avec contradiction explicite et/ou implicite mais qui ne peuvent pas être classés sous une des catégories précédentes parce qu'ils expriment plus de l'humour, de la satire ou une ironie situationnelle. Voici quelques exemples :

(3.19) Palme d'Or pour un film sur l'homosexualité le jour de la #manifpourtous #Cannes2013

(3.20) Alerte à la pollution de l'air : il est déconseillé de prendre son vélo pour aller au travail à 9h... mais pas sa voiture diesel !

(3.21) Merci Hollande d avoir sauvé le monde ! Sans toi, la terre serait actuellement entrée en 3ème guerre mondiale

D Niveau 4 : Marqueurs linguistiques de l'ironie

Comme le montre le tableau 3.2, la littérature linguistique considère d'autres formes de catégories d'ironie, telles que l'effet de surprise, la répétition, etc. Dans une perspective computationnelle, nous avons préféré distinguer clairement entre *catégories d'ironie* qui sont des dispositifs pragmatiques d'ironie tels que définis dans la section précédente, et *indices d'ironie* qui sont un ensemble de segments (mots, symboles, propositions) qui peuvent activer l'ironie sur la base du contenu linguistique du tweet seulement. Cette distinction a également été motivée par le fait que les indices peuvent être présents dans des catégories d'ironie distinctes, pas du tout présents ou présents dans des tweets non ironiques.

Indices de l'ironie	**Références**
Connecteurs de discours	-
Ponctuation	(Tayot, 1984; Wilson & Sperber, 1992; Seto, 1998; Attardo, 2001; Didio, 2007; Burgers, 2010)
Mot d'opinion	(Reyes & Rosso, 2011; Reyes & Rosso, 2012)
Emoticône	(Tayot, 1984; Kreuz, 1996; Burgers, 2010; Gonzalez-Ibanez *et al.*, 2011; Buschmeier *et al.*, 2014)
Marqueur de contradiction	(Attardo, 2000b) (Didio, 2007)
Majuscule	(Haiman, 2001; Burgers, 2010; Tsur *et al.*, 2010; Reyes *et al.*, 2013)
Intensifieur	(Liebrecht *et al.*, 2013; Barbieri & Saggion, 2014b)
Mot de comparaison	(Veale & Hao, 2010)
Modalité	-
Négation	-
Citation	(Tayot, 1984; Gibbs, 1994; Attardo, 2001; Burgers, 2010; Tsur *et al.*, 2010; Reyes *et al.*, 2013)
Interjection	(Gonzalez-Ibanez *et al.*, 2011) (Kreuz & Caucci, 2007)
Pronom personnel	-
Verbe de reportage	-
Surprise/étonnement	(Didio, 2007; Colston & Keller, 1998)
Entité nommée	-
Fausse proposition	(Tayot, 1984; Attardo, 2000b; Didio, 2007; Barbe, 1995)
Hashtag ironique ou humoristique	-
URL	-

TABLE 3.4 : Indices de l'ironie dans notre schéma d'annotation. Les indices en gras sont nouveaux par rapport à l'état de l'art.

Le tableau 3.4 résume les indices retenus pour l'annotation de tweets. Dix-neuf marqueurs ont été sélectionnés pour notre étude. Certains d'entre eux ont montré leur efficacité lorsqu'ils sont utilisés comme caractéristiques de surface dans la détection d'ironie :

les signes de ponctuation, les lettres majuscules, les émoticônes, les interjections, les négations, les opinions et les mots d'émotion (Davidov *et al.*, 2010; Gonzalez-Ibanez *et al.*, 2011; Reyes *et al.*, 2013). Nous étudions en outre de nouveaux marqueurs (voir les indices en gras dans le tableau 3.4) : **connecteurs de discours** (ils peuvent marquer des oppositions, des chaînes d'argumentation et des conséquences) ; **modalité** ; **verbes de reportage**, **négation**, **hashtag ironique ou humoristique** ; **entités nommées** et **pronoms personnels** (pour ces deux derniers, nous supposons qu'ils peuvent être indicateurs de tweets personnels ou de tweets traitant de sujets médiatisés) ; **URL** (les pages Web pointées par les URL donnent des informations contextuelles qui peuvent aider le lecteur à détecter l'ironie) ; et enfin les **fausses propositions** qui mentionnent des faits ou événements faux dans la réalité (et dans lesquelles on peut trouver des négations). Ces quatre derniers marqueurs pourraient être de bons indicateurs pour une détection automatique de l'ironie implicite, par exemple en détectant qu'un contexte externe est nécessaire.

Par exemple, dans le tweet (3.22) les marqueurs sont des négations (*n', pas, non*), la ponctuation (*! ou !!!*), le mot d'opinion (*violent*) alors que dans le tweet (3.23) les marqueurs sont des entités nommées (*NSA, Belgique*), des négations (*Pas, n'...pas*) et une fausse proposition (*ce n'est pas un pays entier*).

(3.22) Ben **non !** Matraquer et crever des yeux, ce **n'**est **pas violent** et ça respecte les droits **!!!** #ironie

(3.23) La #**NSA** a mis sur écoutes un pays entier. **Pas** d'inquiétude pour la #**Belgique** : ce **n'**est **pas** un pays entier. #ironie

3.4 Campagne d'annotation

Dans cette section, nous présentons la procédure suivie pour l'annotation des tweets en appliquant le schéma d'annotation proposé et en utilisant l'outil d'annotation Glozz[2].

3.4.1 Présentation de l'outil Glozz

Glozz est un outil qui propose une interface dédiée à l'annotation, développé dans le cadre du projet ANNODIS (Péry-Woodley *et al.*, 2009). L'annotation est effectuée selon un schéma d'annotation Glozz qui suit soigneusement les éléments présentés dans la section précédente.

Chaque tweet doit être annoté en utilisant l'outil Glozz, en termes d'unités et relations entre les unités (si la relation existe). Les relations permettent de relier des unités textuelles figurant dans un tweet. Nous distinguons trois types de relation :

[2] www.glozz.org

- **Relation de comparaison :** consiste à relier les deux unités ou les deux parties du texte qui sont en comparaison (cf. figure 3.2).

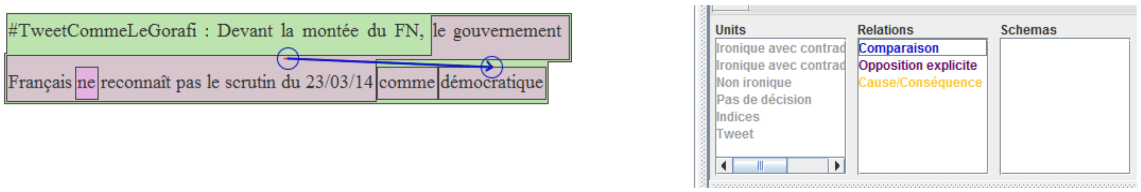

FIGURE 3.2 : Relation de comparaison entre deux unités.

- **Relation de contradiction explicite :** consiste à relier les parties du texte en contradiction explicite (figure 3.3).

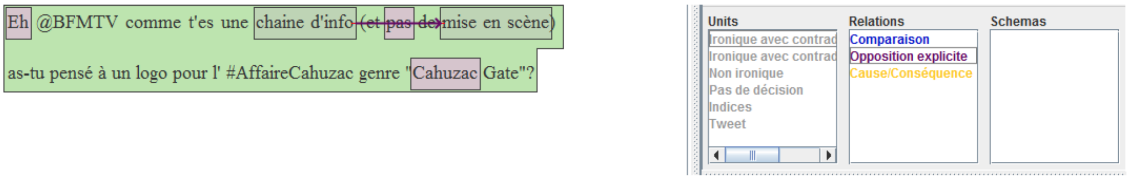

FIGURE 3.3 : Relation d'opposition explicite entre deux unités.

- **Relation de cause/conséquence :** consiste à relier les parties du texte où l'une est la cause et la deuxième est la conséquence de la première (figure 3.4).

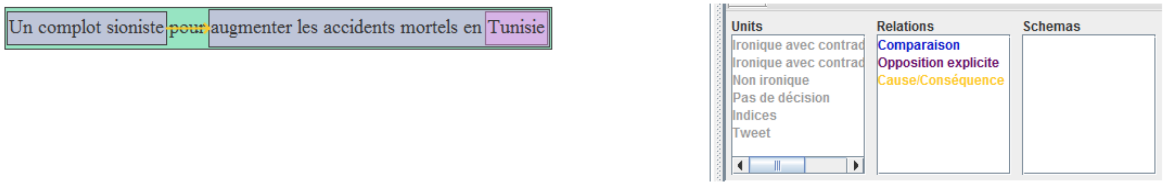

FIGURE 3.4 : Relation de cause/conséquence entre deux unités.

Glozz nécessite plusieurs fichiers en entrée notamment un fichier qui représente le schéma d'annotation proposé au format d'entrée de Glozz et un fichier de sortie qui contient les différentes annotations réalisées par l'annotateur.

3.4.2 Préparation des données

Une étape de préparation des données a été nécessaire avant d'entamer la procédure d'annotation. Cette étape a consisté à pré-annoter les tweets et à générer les fichiers d'entrée nécessaires pour Glozz.

L'étape de pré-annotation consiste à annoter automatiquement un ensemble d'indices afin de faciliter la tâche d'annotation et de la rendre plus rapide. Les indices automatiquement annotés sont : la ponctuation, les intensifieurs, les émoticônes, les mots exprimant une opposition, les mots de comparaison, les pronoms personnels et les mots de négation.

La pré-annotation automatique de ces marqueurs linguistiques a été effectuée en s'appuyant sur deux lexiques : **CASOAR**[3] et **EMOTAIX**[4] pour l'annotation automatique des mots d'opinion et d'émotion, les intensifieurs, les interjections ; et l'analyseur syntaxique **MElt**[5] pour l'annotation automatique des entités nommées. En cas de marqueurs manquants ou d'annotations erronées, les annotations automatiques ont été corrigées manuellement.

L'étape de pré-annotation comprend également l'attribution automatique :

- d'un identifiant unique pour chaque tweet récupéré via l'API de Twitter
- d'un identifiant incrémental qui permet aux annotateurs de repérer facilement les tweets à annoter.
- de la date de publication du tweet.
- du mot-clé utilisé pour la collecte du tweet.

Ces pré-annotations sont possibles si elles peuvent être récupérables en utilisant l'API de Twitter sinon ces attributs prennent une valeur "*null*" par défaut. Le texte du tweet est enregistré dans un fichier d'entrée Glozz après avoir enlevé les hashtags *#ironie* et *#sarcasme*.

3.4.3 Procédure d'annotation

Pour chaque tweet t, l'annotation fonctionne comme suit [6] :

(a) Classifier t dans *Ironique/Non ironique*. Dans le cas où les annotateurs ne comprennent pas le tweet en raison de références culturelles ou de connaissances insuffisantes, t peut être classé dans la classe *Pas de décision*.

[3] https://projetcasoar.wordpress.com/
[4] http://www.tropes.fr/download/EMOTAIX_2012_FR_V1_0.zip
[5] http://alpage.inria.fr/statgram/frdep/fr_stat_dep_malt.html
[6] Le manuel d'annotation est disponible à l'adresse : github.com/IronyAndTweets/Scheme

(b) Si t est ironique, il faut définir son type d'activation, à savoir *contradiction explicite* ou *contradiction implicite*. Pour ce faire, les annotateurs doivent essayer de retrouver dans le texte du tweet deux propositions P_1 et P_2 qui sont en contradiction. Si c'est le cas, alors l'activation est explicite, sinon, elle est implicite.

(c) Une fois le type d'ironie identifié, il faut spécifier les dispositifs pragmatiques utilisés pour exprimer l'ironie en sélectionnant une ou plusieurs catégories d'ironie (cf. section C.)

(d) Identifier les segments de texte dans le tweet qui correspondent à une liste prédéfinie de marqueurs linguistiques.

Au cours de l'annotation, les annotateurs sont invités à respecter les contraintes ci-dessous :

1. Quel que soit le type de l'ironie (ironique avec contradiction explicite, implicite ou non ironique), il faut annoter tous les indices.
2. Pour les tweets ironiques avec contradiction explicite ou implicite, il faut obligatoirement classer le tweet dans une des catégories appartenant à chaque type d'ironie.
3. Pour les indices de négation, il suffit d'annoter un seul mot de négation par tweet.
4. Pour les tweets ironiques avec contradiction explicite, il faut obligatoirement relier les parties du texte en contradiction explicite avec la relation "contradiction explicite".
5. Dans le cas des tweets ironiques avec contradiction implicite, il est impossible de relier deux parties du texte par les relations que nous avons définies.
6. Si le tweet contient un marqueur linguistique de comparaison, il faut obligatoirement relier les deux concepts ou les deux parties du texte comparés par la relation "comparaison".

De plus, pour s'assurer que les annotations sont conformes aux contraintes listées ci-dessus, les erreurs les plus courantes sont automatiquement détectées : tweets ironiques sans type d'activation ou catégorie d'ironie, absence de marqueurs, etc. En cas d'erreurs, les annotateurs sont invités à corriger leurs erreurs avant de continuer à annoter de nouveaux tweets.

La figure 3.5 illustre un exemple de la procédure d'annotation d'un tweet ironique avec *opposition explicite* entre deux segments : (1) « tellement bien réussi » et (2) « tt va moins bien ». Le premier segment exprime une *hyperbole* alors que le deuxième segment exprime un *euphémisme*. L'ironie dans ce tweet a donc deux catégories : hyperbole et euphémisme. Les termes : "PS" (entité nommée), "tellement" (intensifieur), "bien" (mot d'opinion positif) et "moins" (intensifieur) ont été annotés comme étant des indices.

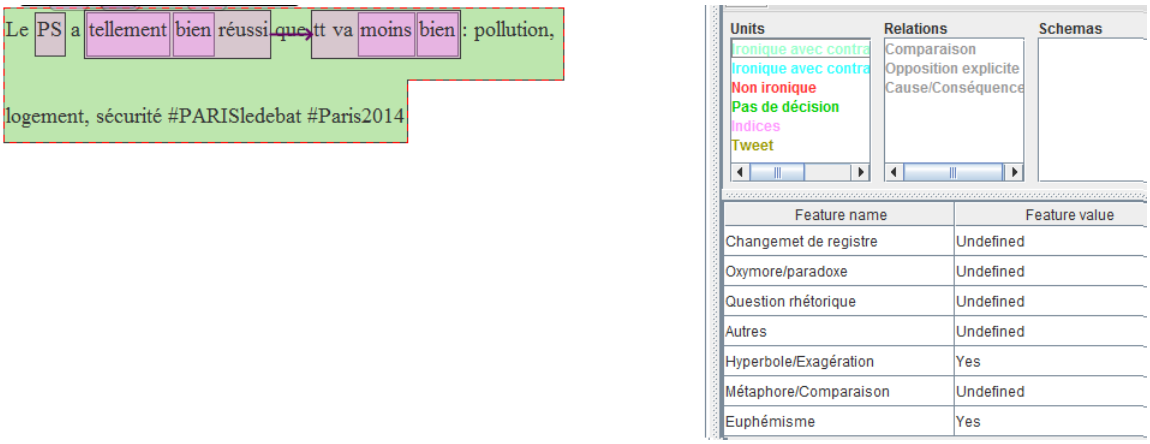

FIGURE 3.5 : Exemple de tweet annoté avec Glozz.

3.5 Résultats de la campagne d'annotation

Nous présentons dans cette section les différentes statistiques élaborées sur le corpus annoté. Ce travail a été réalisé en collaboration avec Farah Benamara et Véronique Moriceau.

Le corpus annoté est formé de 2 000 tweets dont 80% des tweets sont ironiques et 20% non ironiques en prenant comme référence la présence ou absence des hashtags *#ironie* et *#sarcasme*. La campagne d'annotation a été réalisée par trois annotateurs humains francophones.

Le corpus a été divisé en trois parties. Une première partie du corpus comporte 100 tweets (50 tweets ironiques et 50 tweets non ironiques) et a été exploitée dans la phase d'entrainement des trois annotateurs. Une deuxième partie du corpus comporte 300 tweets (250 tweets ironiques et 50 tweets non ironiques) et a été annotée par deux annotateurs afin de pouvoir calculer un accord inter-annotateur. Une troisième partie du corpus formé de 1 700 tweets dont 80% ironiques a été consacrée à la campagne d'annotation finale.

Dans ce qui suit, nous présentons les accords inter-annotateurs sur les 300 tweets annotés par deux annotateurs ainsi que les différentes statistiques quantitatives résultant de la campagne d'annotation sur les 2 000 tweets.

3.5.1 Résultats qualitatifs

Parmi les 300 tweets, les annotateurs sont d'accord sur 255 tweets (174 ironiques et 63 non ironiques), parmi lesquels 18 ont été classés comme *Pas de décision*. Nous obtenons un Kappa de Cohen de 0,69 pour la classification *Ironique/Non ironique*, ce qui est un très bon score. Par rapport aux étiquettes de référence (*#ironie* et *#sarcasme*), nous avons également

obtenu une bonne mesure Kappa (0,62), ce qui montre que ces hashtags sont assez fiables. Nous avons également noté que plus de 90% des tweets annotés comme *Pas de décision* en raison du manque de contexte externe, sont ironiques selon les étiquettes de référence. Nous avons cependant décidé de les garder pour les expériences.

Pour l'activation de l'ironie *explicite* vs. *implicite*, l'accord sur le type d'activation du tweet ironique a obtenu un Kappa de 0,65. Il est intéressant de noter que l'activation implicite représente la majorité (76,42%). Il s'agit d'un résultat important qui montre que les annotateurs sont en mesure d'identifier quels sont les champs textuels qui activent l'incongruité des tweets ironiques, qu'ils soient explicites ou implicites, et nous nous attendons à ce que les systèmes automatiques soient aussi efficaces que les humains (voir chapitre 4 pour les résultats obtenus par nos modèles computationnels).

Enfin, pour l'identification de la catégorie d'ironie, puisque un tweet ironique peut appartenir à plusieurs catégories d'ironie, nous avons calculé les accords en comptant, pour chaque tweet, le nombre de catégories communes, puis en divisant par le nombre total de catégories annotées. Nous avons obtenu un score de 0,56 qui est modéré. Ce score reflète la complexité de l'identification des dispositifs pragmatiques. Lorsque des dispositifs similaires sont regroupés (principalement *hyperbole/exagération* et *euphémisme*, car ils sont utilisés pour rendre le sens voulu soit plus fort soit plus faible), le score augmente à 0,60.

3.5.2 Résultats quantitatifs

L'objectif principal de notre étude de corpus est de vérifier si les différentes théories linguistiques et les définitions de l'ironie peuvent être appliquées aux médias sociaux, en particulier aux tweets. En plus des fréquences standards, nous fournissons les corrélations entre les types d'activation de l'ironie et les marqueurs et entre les catégories et les marqueurs afin de mettre en évidence des caractéristiques qui pourraient être utilisées dans une perspective de détection automatique de l'ironie. Notons que dans toutes ces études, les fréquences présentées sont statistiquement significatives de ce qui serait attendu par hasard en utilisant le test du χ^2 ($p < 0,05$).

A Fréquence des tweets selon les classes

La figure 3.6 montre les fréquences des tweets annotés selon les trois classes : ironique, non ironique et pas de décision. Nous rappelons que cette première étape correspond au niveau 1 de notre schéma d'annotation (cf. section 3.3.2).

En partant des hashtags de référence *#ironie* et *#sarcamse*, nous avons 1 600 tweets ironiques et 400 tweets non ironiques. En revanche, les annotateurs humains ont jugé 1 460 tweets (73%) comme étant *ironiques*, 380 tweets (19%) *non ironiques* et 160 tweets (8%) ont été classés dans la classe *Pas de décision*.

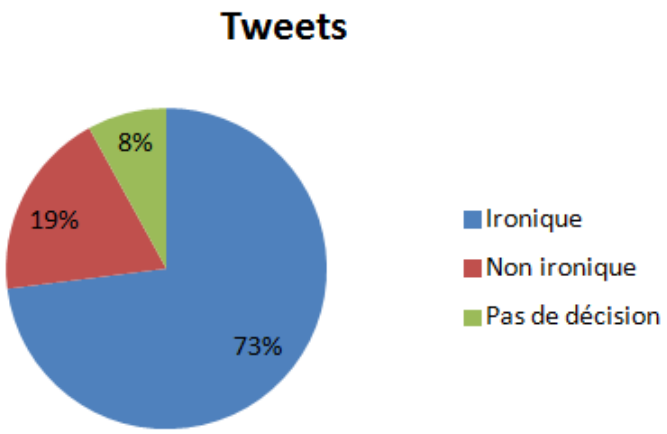

FIGURE 3.6 : Répartition des tweets annotés par classe (niveau 1).

Ces résultats prouvent qu'un tweet accompagné de *#ironie* ou *#sarcasme* n'est pas forcément ironique, alors qu'un tweet sans ces hashtags peut être ironique (par exemple : *@MelvinLeroux Quels autres problèmes ? La France va bien, le taux de chômage n'est pas du tout élevé, tout le monde est heureux ...*).

B Fréquence des tweets selon le type de l'ironie

La figure 3.7 montre les statistiques pour le deuxième niveau du schéma d'annotation, c'est-à-dire le type d'activation de l'ironie (explicite ou implicite).

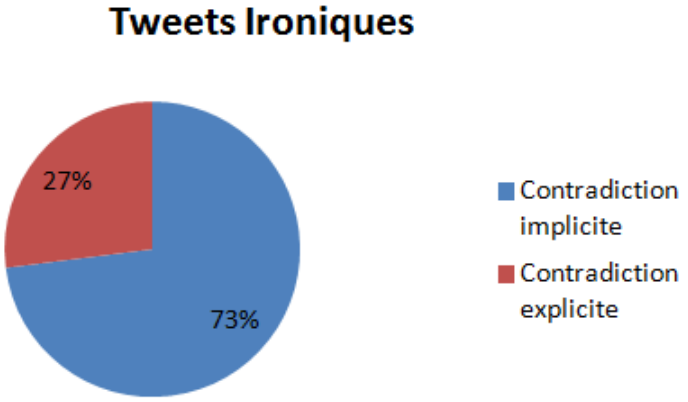

FIGURE 3.7 : Répartition des tweets annotés comme ironiques selon le type d'activation (niveau 2).

Parmi les 1 460 tweets annotés comme ironiques, on a 1 066 tweets (73%) *ironiques*

avec contradiction implicite et seulement 394 tweets (27%) *ironiques avec contradiction explicite*. Ceci prouve que l'ironie est un phénomène qui s'exprime généralement d'une manière implicite.

C Fréquence des tweets selon les catégories de l'ironie

Le tableau 3.5 présente les résultats quantitatifs pour le troisième niveau du schéma d'annotation, c'est-à-dire les catégories d'ironie.

	Ironique avec contradiction explicite	*Ironique avec contradiction implicite*
Analogie	12	2
Changement de registre	1	-
Euphémisme	1	1
Question rhétorique	10	14
Oxymore / paradoxe	**66**	-
Hyperbole / Exagération	8	10
Fausse affirmation	-	**56**
Autres	21	**32**

TABLE 3.5 : Pourcentage de tweets dans chaque catégorie pour chaque type d'ironie (niveau 3).

Nous observons des différences importantes concernant la présence des différentes catégories pour chaque type d'ironie. Par exemple, les catégories *euphémisme* et *hyperbole/ exagération* sont présentes d'une manière équivalente dans les deux types d'ironie (avec contradiction explicite et implicite) alors que la catégorie *oxymore/paradoxe* est présente uniquement dans les tweets ironiques avec contradiction explicite. De même, la catégorie *fausse affirmation* n'est présente que dans les tweets ironiques avec opposition implicite. Ceci est justifié par les définitions relatives aux catégories *Oxymore/paradoxe* et *Fausse affirmation* (sous-section C de la section 3.3.2). Pour la catégorie *Autres*, nous constatons que la plupart des tweets sont ironiques avec *contradiction implicite*. Cela prouve que la tâche de décision est plus difficile dans le cas où l'ironie est exprimée avec une contradiction implicite.

Comme les classes ne sont pas mutuellement exclusives, il y a 64 tweets avec *contradiction explicite* appartenant à plus d'une catégorie et 134 tweets avec *contradiction implicite* appartenant à plus d'une catégorie. La combinaison la plus fréquente pour les contradictions explicites est *Oxymore/ Paradoxe* + *Hyperbole/ Exagération* tandis que pour les ironiques

avec *contradiction implicite* la combinaison la plus fréquente est *Fausse affirmation + Hyperbole/ Exagération*. L'exemple 3.24 illustre un tweet annoté sous deux catégories d'ironie : la première catégorie est l'hyperbole exprimé par *tellement bien* alors que la deuxième catégorie est l'euphémisme exprimé par *moins bien*.

(3.24) Le PS a **tellement bien** réussi que tt va **moins bien** : pollution, logement, sécurité #PARISledebat #Paris2014

D Fréquence des tweets selon les indices linguistiques

À ce niveau, nous avons élaboré trois études statistiques. La première est une étude quantitative entre le premier et le quatrième niveau du schéma d'annotation dans laquelle nous avons étudié la présence des différents indices dans les tweets ironiques et non ironiques (cf. tableau 3.6). La deuxième étude est consacrée aux deuxième et quatrième niveau du schéma d'annotation dans laquelle nous avons étudié la présence des différents indices dans les tweets ironiques avec contradictions explicites et implicites (cf. tableau 3.6). Enfin, la troisième étude est consacrée aux troisième et quatrième niveaux du schéma d'annotation dans laquelle nous avons étudié la présence des indices dans chaque catégorie de l'ironie (cf. tableau 3.7).

Le tableau 3.6 indique que la plupart des marqueurs sont plus fréquents dans les tweets ironiques que les tweets non ironiques. De plus, les mots de négation, les intensifieurs, les oppositions, les interjections, les mots de comparaison, et les mots d'opinion sont plus fréquents dans les tweets ironiques avec contradiction explicite. En revanche, les URL ainsi que les fausses propositions sont plus présentes dans les tweets avec contradiction implicite. La forte utilisation des URL dans les tweets ironiques avec opposition implicite aident les lecteurs à comprendre le sens ironique du tweet en prenant en compte le contexte externe figurant dans l'URL (voir les figures 3.8 et 3.9).

Le tableau 3.7 illustre le pourcentage de tweets appartenant à chaque catégorie et ayant des marqueurs. Nous observons que pour chaque catégorie, il y a au moins deux indices ayant une fréquence importante. Par exemple, les *négations* sont plus fréquentes dans les catégories *analogie, changement de registre, euphémisme, question rhétorique,* et *oxymore/ paradoxe*, alors que les *fausses propositions* sont très fréquentes dans les catégories *euphémisme, hyperbole/exagération* et *fausse affirmation*.

E Fréquences des relations dans les tweets ironiques avec contradiction explicite

La figure 3.10 montre la forte utilisation de la relation d'opposition dans les tweets ironiques avec contradiction explicite. Ceci est justifié par la présence importante des segments textuels qui s'opposent dans le même tweet. Le nombre de relations de comparaison ainsi que de cause/conséquence est presque le même.

	Ironique avec contradiction explicite	*Ironique avec contradiction implicite*	*Non ironique*
Emoticône	7	6	5
Négation	**37**	**34**	**58**
Connecteur de discours	6	4	4
#tag humoristique	2	4	0
Intensifieur	22	19	11
Ponctuation	**51**	**51**	28
Fausse proposition	8	**54**	0
Surprise	3	3	2
Modalité	0	0	1
Citation	6	6	1
Mot d'opposition	9	3	4
Mot en majuscule	3	2	3
Pronom personnel	**31**	**31**	**30**
Interjection	14	12	2
Comparaison	8	2	4
Entité Nommée	**97**	**91**	**82**
Reportage	1	1	3
Opinion	**48**	**41**	**35**
URL	21	26	**36**

TABLE 3.6 : Répartition des marqueurs entre les tweets ironiques (explicites et implicites) et non ironiques en terme de pourcentage de tweets.

3.5.3 Corrélation entre les différents niveaux du schéma d'annotation

Dans cette sous-section, nous présentons une étude statistique sur l'intensité du lien entre les différents niveaux du schéma d'annotation proposé. Nous commençons par l'étude de la force du lien entre le premier niveau (ironique/non ironique) et le quatrième niveau (les indices). Par la suite, nous présentons les résultats de la relation entre le niveau 2 (ironique avec contradiction explicite ou implicite) et le niveau 4. Et enfin, nous présentons les résultats de la relation entre le niveau 3 (les catégories de l'ironie) et le niveau 4.

Traditionnellement, pour vérifier s'il existe un lien entre deux variables qualitatives croisées dans un tableau de contingence, on utilise le test du χ^2, alors que la vérification de l'intensité du lien entre deux variables nécessite l'utilisation du test Phi (ϕ) ou le test V de Cramer (Cohen, 1988). L'utilisation de Phi (ϕ) est limitée aux tableaux de taille 2×2, tandis que le test V de Cramer peut être utilisé pour des tableaux plus grands. Vu que les tableaux de notre étude statistique sont de taille supérieure à 2×2, nous avons utilisé le

	Analogie	*Changement de registre*	*Euphémisme*	*Hyperbole/ exagération*	*Questions rhétoriques*	*Oxymore/ Paradoxe*	*Fausse affirmation*	*Autres*
Emoticône	6	0	0	5	6	6	5	8
Négation	**46**	**40**	**50**	25	**43**	**35**	18	26
Connecteur de discours	6	0	6	5	2	4	4	5
#tag humoristique	6	0	0	3	2	0	3	5
Intensifieur	21	0	**50**	**57**	17	21	10	15
Ponctuation	**49**	**60**	**72**	**56**	**93**	**49**	29	**45**
Fausse proposition	13	0	**44**	**53**	9	11	**95**	11
Surprise	0	0	0	3	4	4	3	1
Modalité	0	0	0	0	0	0	0	0
citation	0	0	0	8	7	5	4	8
Mot d'opposition	6	0	0	2	3	12	3	2
Majuscule	5	0	0	2	5	4	2	3
Pronom personnel	**38**	**40**	22	29	31	32	31	29
Interjection	6	20	6	18	13	15	13	10
Comparaison	**43**	20	0	0	2	2	2	1
Entité Nommées	**100**	**80**	**94**	**88**	**90**	**99**	**90**	**91**
Reportage	2	0	0	3	1	1	1	1
Opinion	**41**	**60**	**56**	**84**	**45**	**55**	**45**	32
URL	13	0	22	21	25	11	25	30

TABLE 3.7 : Répartition des marqueurs entre les différentes catégories en terme de pourcentage de tweets.

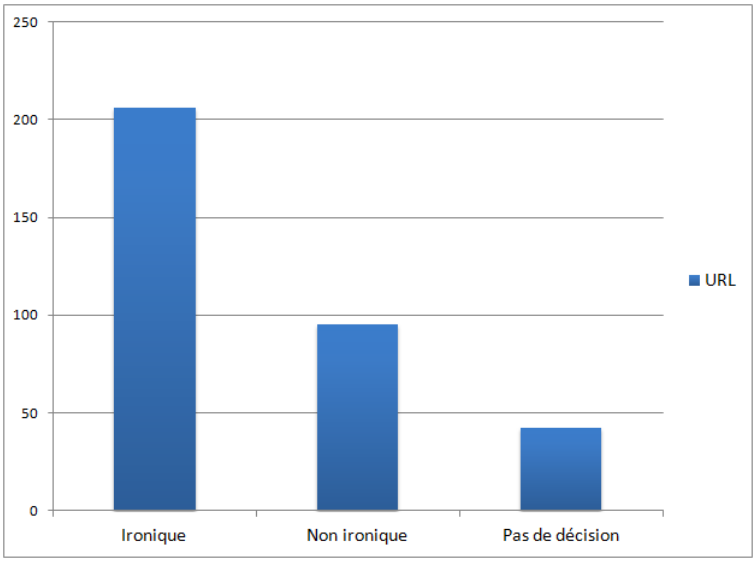

FIGURE 3.8 : Présence des URL dans les tweets.

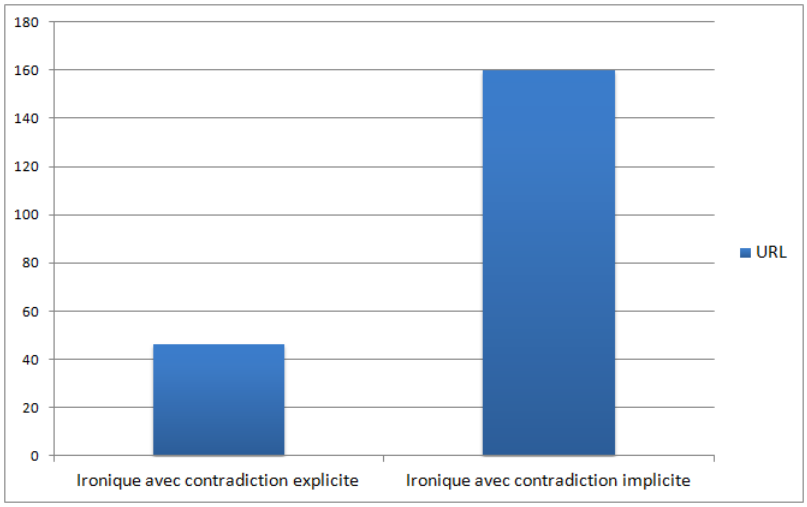

FIGURE 3.9 : Présence des URL dans les tweets ironiques.

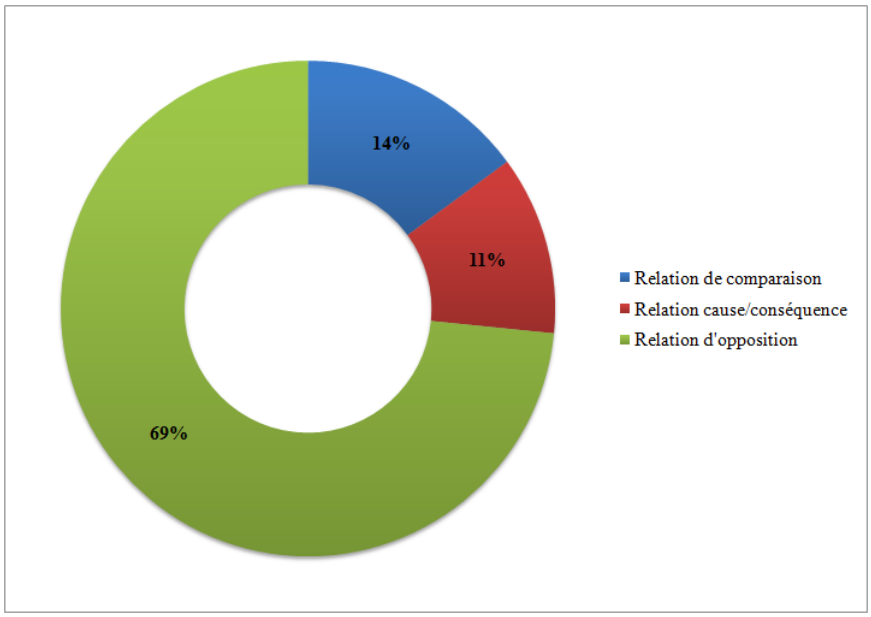

FIGURE 3.10 : Répartition des relations dans les tweets ironiques avec contradiction explicite en terme de pourcentage de tweets.

test V de Cramer pour la vérification de l'intensité du lien entre les éléments des différents niveaux de notre schéma d'annotation.

Le calcul du V de Cramer est donné par la formule suivante :

$$V = \sqrt{\frac{\chi^2}{n \times (k-1)}}$$

Le V de Cramer est la racine carrée du χ^2 divisée par $n \times (k-1)$ où n repésente l'effectif et k le plus petit côté du tableau (nombre de lignes ou de colonnes). Plus V est proche de zéro, moins les variables étudiées sont dépendantes. Au contraire, il vaudra 1 lorsque les deux variables sont complètement dépendantes, puisque le χ^2 est alors égal à $n \times (k-1)$ (dans un tableau 2×2, il prend une valeur comprise entre -1 et 1). Donc, plus V est proche de 1, plus la liaison entre les deux variables étudiées est forte.

Corrélation entre le type et les indices de l'ironie

Commençons par la vérification de l'intensité de la relation entre le niveau 1 du schéma d'annotation (classe de l'ironie : ironique/non ironique) et le niveau 4 (les indices de l'ironie) (cf. figure 5.1). Nous avons obtenu une valeur V de Cramer de $0,156$ avec $df = 14$ [7]. Selon la table des valeurs de V de Cramer définie par (Cohen, 1988), la valeur obtenue est statistiquement significative ($p < 0,05$) et indique une forte corrélation entre les classes ironiques/non ironiques et l'ensemble des indices annotés.

Corrélation entre l'activation et les indices de l'ironie

Une deuxième étude a été élaborée pour la vérification de l'intensité de la relation entre le niveau 2 du schéma d'annotation (activation de l'ironie : contradiction implicite ou explicite) et le niveau 4 (les indices de l'ironie). Cette étude indique une forte corrélation entre les types d'ironie (ironique avec contradiction explicite ou implicite) et les différents indices annotés avec une valeur de V de Cramer de $0,196$ avec $df = 16$. L'étude de l'intensité de la corrélation entre le type de l'ironie et les indices pris un par un ($df = 1$) a montré qu'il existe une liaison d'intensité moyenne à forte entre le type de l'ironie et quelques indices : les négations, les interjections, les entités nommées et les URL avec $0,140 < V < 0,410$.

Corrélation entre les différents indices de l'ironie

Une troisième étude a été consacrée à la vérification de la corrélation entre les différents indices de l'ironie. Cette étude a montré que les indices les plus corrélés avec les classes d'ironie (ironique/non ironique) sont : les négations, les interjections, les entités nommées

[7] df = min (r - 1, c - 1) où r = nombre de lignes et c = nombre de colonnes dans le tableau de contingence

et les URL ($0,140 < V < 0,410, df = 1$), alors que les indices les plus corrélés avec l'activation ironique avec contradiction explicite/implicite sont : les marqueurs d'opposition, les mots de comparaison et les fausses propositions ($0,140 < V < 0,190,\ df = 1$).

Corrélation entre les catégories et les indices de l'ironie

Une quatrième étude a été élaborée pour la vérification de la corrélation entre le niveau 3 du schéma d'annotation (catégorie de l'ironie) et le niveau 4 (les indices). Cette étude a prouvé que les indices les plus corrélés avec les catégories de l'ironie sont : les intensifieurs, la ponctuation, les fausses propositions et les mots d'opinion ($0,267 < V < 0,565,\ df = 4$). Enfin, nous précisons que malgré la forte fréquence des mots d'opinion dans les tweets ironiques, les corrélations ont prouvé que les mots d'opinion ne sont pas discriminants dans la distinction entre ironique/non ironique ni entre ironique avec contradiction explicite et ironique avec contradiction implicite.

Ces différentes études montrent que les indices ainsi que les catégories peuvent êtres utiles dans la tâche de classification des tweets en ironique/non ironique, ironique explicite/implicite et même pour distinguer les différentes catégories de l'ironie.

3.6 Conclusion

Dans ce chapitre, nous avons présenté un schéma d'annotation multi-niveaux pour l'ironie. Nous avons en particulier étudié les catégories de l'ironie proposées par les linguistes et présenté les catégories retenues pour notre schéma d'annotation. Le schéma d'annotation proposé a été validé par une campagne d'annotation réalisée par trois annotateurs francophones. Un sous-ensemble du corpus FrIC a été doublement annoté par deux annotateurs et les accords inter-annotateurs obtenus pour les différents niveaux du schéma sont satisfaisants. Enfin, nous avons présenté la procédure d'annotation avec l'outil d'annotation Glozz et les études statistiques élaborées pour chaque niveau du schéma d'annotation.

Les résultats obtenus, en particulier ceux relatifs à l'étude de la corrélation entre les catégories et les types d'ironie, sont cruciaux dans le but d'une détection automatique de l'ironie. Un sous-ensemble du corpus FrIC (voir figure 3.11) est utilisé dans ce but et nous présentons notre approche au chapitre 4. De plus, l'application de notre schéma d'annotation sur des corpus de tweets en anglais et en italien a montré que notre schéma est relativement portable sur des langues culturellement proches (cf. chapitre 5, section 5.2.2).

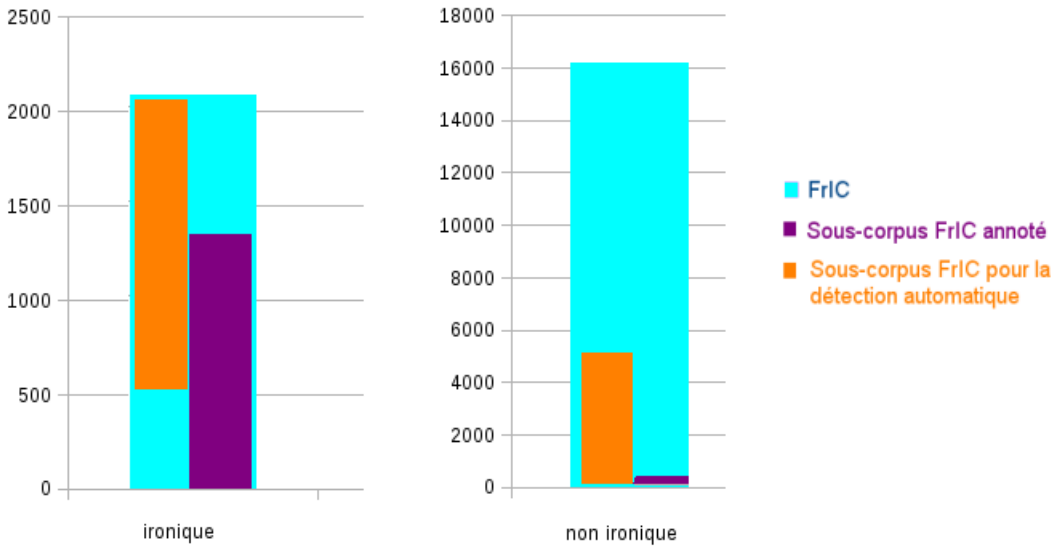

FIGURE 3.11 : Distribution des tweets dans le corpus FrIC et les sous-ensembles utilisés pour l'annotation manuelle et les expériences de détection automatique.

Chapitre 4

Détection automatique de l'ironie

4.1 Introduction

L'analyse d'un sous-ensemble (formé de 2 000 tweets) du corpus FrIC présenté dans le chapitre précédent a permis les observations importantes suivantes :

- L'ironie dans les tweets se manifeste par la présence de deux propositions P_1 et P_2 qui sont en contradiction. Nous avons identifié deux types de contradiction : *explicite* dans le cas où P_1 et P_2 sont toutes deux présentes lexicalement dans le tweet, et *implicite* dans le cas où P_1 est présente et P_2 doit être inférée du contexte externe au tweet. Nous avons montré que l'ironie dans 76,42% des tweets était dûe à une contradiction implicite contre 23,58% pour les contradictions explicites.
- Les catégories d'ironie les plus fréquentes sont : *oxymore / paradoxe* dans le cas des contradictions explicites et *fausse affirmation* dans le cas des contradictions implicites, avec des fréquences respectives de 66% et 56%.
- Les indices linguistiques les plus présents dans les tweets de notre corpus sont : les entités nommées, la ponctuation, les expressions d'opinion, les marqueurs de négation, les pronoms personnels et les URL. De plus, il apparait que les marqueurs de négation (comme *ne ...pas*, *jamais*, etc.) sont l'un des indices les plus fréquents à la fois dans les tweets ironiques et non ironiques avec respectivement 35% et 58%.

Au vu de ces observations, nous avons décidé de développer un modèle de détection automatique de l'ironie textuelle dans les tweets qui permet à la fois d'identifier l'ironie dans les contradictions explicites et implicites. Notre modèle permet, en particulier, de détecter l'ironie qui se manifeste par des fausses assertions. Compte tenu de l'importance des négations dans notre corpus, le modèle a pour objectif de tester les hypothèses suivantes :

- *Hypothèse (H1)* : la présence de négations, en tant que propriété interne d'un énoncé, peut aider à détecter la disparité entre le sens littéral et le sens voulu d'un énoncé.

- *Hypothèse (H2)* : un tweet contenant un fait affirmé de la forme $Not(P)$ est ironique si et seulement si on peut prouver P sur la base de certaines connaissances communes externes à l'énoncé et partagées par l'auteur et le lecteur.

Pour tester la validité des hypothèses ci-dessus, nous proposons une procédure en trois étapes impliquant les trois nouveaux modèles suivants :

1. En utilisant une méthode d'apprentissage supervisé, détecter si un tweet est ironique ou pas en s'appuyant exclusivement sur le contexte interne au tweet. Nous développons ici deux modèles :

   (a) Un premier modèle, nommé **SurfSystem**, basé sur les traits surfaciques.

   (b) Un second modèle, nommé **PragSystem** basé sur les traits pragmatiques extraits du contenu linguistique du tweet.

   Les deux modèles reposent sur les traits de l'état de l'art dont l'efficacité a été prouvée empiriquement et de nouveaux groupes de traits.

2. Un troisième modèle, nommé **QuerySystem**, valide le contexte interne de l'énoncé par rapport au contexte « extérieur ». Nous proposons un algorithme qui traite les sorties du classifieur construit selon le modèle **PragSystem** et corrige les instances ironiques mal classées de la forme $Not(P)$ en recherchant P dans des sources d'informations externes et fiables sur le Web, telles que Wikipedia ou des journaux en ligne. Nous avons effectué deux expériences, la première en se référant aux hashtags de références (*#ironie* et *#sarcasme*), et la seconde en se référant aux classes prédites par le classifieur. Si P est trouvée dans des sources d'information fiables, alors le tweet est susceptible de véhiculer un sens non littéral, c'est-à-dire ironique.

L'approche en trois étapes que nous proposons est nouvelle. En effet, à notre connaissance, il n'y a pas de travaux qui abordent la détection automatique de l'ironie en exploitant à la fois la présence de négations et des connaissances externes permettant de capturer l'ironie qui se manifeste par des contradictions implicites.

Avant de détailler ces trois modèles (SurfSystem, PragSystem et QuerySystem), nous présentons dans ce qui suit les données qui ont servi pour entraîner et tester nos modèles.

4.2 Le corpus FrIC^{Auto}

Le corpus FrIC^{Auto} utilisé pour les expériences de détection automatique est un sous-ensemble du corpus FrIC (cf. section 3.2 du chapitre 3). Il est composé de 1 545 tweets ironiques et 5 197 tweets non ironiques sur des sujets discutés dans les médias (voir tableau 4.1). Dans FrIC^{Auto}, notons que seulement 50% des tweets ont été annotés manuellement selon le schéma multi-niveaux présenté dans le chapitre précédent.

Catégories	Mots clés
Politique	Ayrault, Fillon, Hollande, Le Pen, FN, DSK, UMP, etc.
Santé	cancer, grippe, sida, dépression, angoisse, psychologie, etc.
Réseaux sociaux	Skype, Facebook, MSN, WhatsApp, etc.
Sport	Zlatan, PSG, football, Ribéry, Zidane, équipe de France, ligue des champions, jeux olympiques, etc.
Printemps arabe	Marzouki, Ben Ali, Bachar, Moubarak, Al-Assad, Morsi, Kadhafi, etc.
Pays/ville	Algérie, Égypte, Syrie, Tunisie, Iran, Washington, Mali, etc.
Artistes	Rihanna, Beyoncé, Carla Bruni, Madonna, Nabilla, Justin Bieber, Adèle, etc.
Télévision	Fast and Furious, Xfactor, The Voice, etc.

TABLE 4.1 : Ensemble des catégories utilisées pour la collecte du corpus ainsi quelques mots-clés correspondants.

Comme pour les annotations manuelles, les hashtags *#ironie* et *#sarcasme* ont été supprimés des tweets pour les expériences décrites par la suite.

Une étude approfondie de l'ironie dans les tweets français durant la campagne d'annotation détaillée dans le chapitre 3, a mis en lumière la forte présence des mots de négation tels que *ne, n', pas, non, ni, sans, plus, jamais, rien, aucun(e), personne*. Dans l'ensemble des tweets collectés, nous avons environ 62,75% des tweets contenant des mots de négation. La négation nous semble donc être un indice important dans les énoncés ironiques, notamment dans les fausses affirmations.

Pour mesurer l'effet de la négation sur la tâche de détection de l'ironie, nous avons constitué 3 corpus : les tweets avec négation (*NegOnly)*, les tweets sans négation (*NoNeg*), et un corpus regroupant l'ensemble des tweets (*All*). Le tableau 4.2 montre la répartition des tweets dans chaque corpus.

L'identification des négations a été effectuée automatiquement en exploitant deux analyseurs syntaxiques à savoir *XIP* [1] et *MElt* [2]. Une analyse manuelle des résultats des deux

[1] https://open.xerox.com/Services/XIPParser
[2] http://alpage.inria.fr/statgram/frdep/fr_stat_dep_malt.html

Corpus	**Ironique**	**Non ironique**	TOTAL
NegOnly	470	3 761	**4 231**
NoNeg	1 075	1 436	**2 511**
All	1 545	5 197	**6 742**

TABLE 4.2 : Répartition des tweets dans le corpus.

analyseurs pour la tâche de la détection des vraies négations a montré que *MElt* est plus performant que *XIP* malgré la présence de quelques erreurs. Par conséquent, nous avons décidé d'utiliser *MElt*. Ce dernier se trompe pour trois mots de négation à savoir : *personne, plus* et *pas*. Ces mots ont toujours été considérés par MElt comme étant des mots de négation alors que le mot *personne* doit être considéré comme une négation seulement s'il est pronom indéfini (exemple où *personne* est un nom : 4.1). De même, le mot *plus* est une négation s'il n'est pas utilisé comme comparatif ou superlatif (exemple 4.2). Enfin, le mot *pas* est une négation s'il n'est pas un nom (exemple 4.3). Nous avons alors développé un post-traitement de correction pour l'identification correcte des négations en exploitant ces règles de grammaire avec un script java. Cette correction est effectuée en sortie de l'analyseur syntaxique *MElt*.

(4.1) @EloiseLEspagnol va se coucher. J'suis seule avec **1 personne** dans ma TL. #GENIAL

(4.2) Je ne sais pas laquelle des deux entre #trierweiler et #Gayet est **la plus** folle de rage de voir @RoyalSegolene à côté de Flanby !

(4.3) @OMissud @RoyalSegolene je leur conseille le tango, **un pas** en avant, **un pas** en arrière ! ! !

Dans notre travail, nous avons considéré les hashtags *#ironie* et *#sarcasme* comme étant des hashtags fiables. En revanche, l'absence de l'un de ces hashtags dans un tweet ne confirme jamais qu'un tweet n'est pas ironique. Afin de s'assurer de la fiabilité des hashtags, deux annotateurs humains ont annoté manuellement 3 sous-ensembles composés de : 50 tweets ironiques et 50 tweets non ironiques pour chacun des corpus *All*, *NoNeg* et *NegOnly*. L'accord inter-annotateur (kappa de Cohen) par rapport aux hashtags de référence est $\kappa = 0,78$ pour le corpus *All*, $\kappa = 0,73$ pour le corpus *NoNeg*, et $\kappa = 0,43$ pour le corpus *NegOnly*. Les scores montrent que les hashtags *#ironie* et *#sarcasme* sont relativement fiables mais la présence d'un mot de négation provoque une ambiguïté pour la compréhension et l'identification de l'ironie par les humains.

4.3 Le modèle SurfSystem : Détection de l'ironie sur la base de traits surfaciques

Nous présentons ici le modèle que nous avons développé pour la détection de l'ironie et qui s'appuie uniquement sur des traits surfaciques. Nous présentons les traits utilisés ainsi que les expériences et résultats obtenus.

4.3.1 Traits utilisés

Dans le cadre de cette première expérience, nous avons réutilisé l'ensemble des traits de surface de l'état de l'art pour la classification des tweets français en ironique/non ironique , et proposé de nouveaux traits :

- Longueur du tweet en nombre de mots (Tsur *et al.*, 2010).
- Présence de la ponctuation (Kreuz & Caucci, 2007; Gonzalez-Ibanez *et al.*, 2011), cf. exemple 4.4 et 4.5

  (4.4) **Ah** oui exact' **!** #SuisJeBête **Mais** il y a rien de fait pour le PSG en championnat hein **;)** #ironie

  (4.5) Comment ce faire hair par sa #LT : L'algérie n'ira pas a la Coupe Du Monde moi je vous le dis =)**!!!!!!!!!!!! JE REPETE JE RIGOLE!!** #ironie

- Présence de mots en majuscules (Tsur *et al.*, 2010; Reyes *et al.*, 2013), cf. exemple 4.5
- Présence d'interjections (Gonzalez-Ibanez *et al.*, 2011; Buschmeier *et al.*, 2014), cf. exemple 4.4
- présence d'émoticônes (Gonzalez-Ibanez *et al.*, 2011; Buschmeier *et al.*, 2014), cf. exemple 4.4
- Présence de citations (Tsur *et al.*, 2010; Reyes *et al.*, 2013), cf. exemple 4.6

  (4.6) **"1 million de chômeurs, c'est 1 millions d'immigrés de trop"**...connaissais pas...sympa le slogan, et pas du tt simpliste #fn #lepen #ironie

- Présence d'argot (Burfoot & Baldwin, 2009), cf. exemple 4.7

  (4.7) On nous a expliqué que #Hollande se ferait **bouffer** à l'international. Effectivement, #Obama avait l'air de le mépriser ce matin...

- Présence de mots d'opposition (Utsumi, 2004), cf. exemple 4.4
- Présence de séquence de points d'exclamation ou d'interrogation (Carvalho *et al.*, 2009), cf. 4.5
- Présence d'une combinaison de points d'exclamation et d'interrogation (Buschmeier *et al.*, 2014), cf. exemple 4.8

  (4.8) Depuis quelques jours certains murmurent que le racisme se serait ... MAL**?!** **??!!!!!!** Hummmmmmmm #MoueDubitative #ironie

- Présence de connecteurs discursifs qui ne déclenchent pas d'opposition. Ceci est un trait nouveau par rapport à l'état de l'art, cf. exemple 4.9

  (4.9) Vous pourrez remercier Jérôme de votre service NC de Lyon **ainsi que** vos techniciens. On ne peut que vous recommander #ironie

Pour implémenter ces traits, nous avons utilisé quatre lexiques :

- Un lexique de connecteurs discursifs défini par Rose et al., (2012) composé de 328 mots.
- Un lexique d'argot formé de 389 mots et construit manuellement à partir de diverses sources trouvées sur le Web[3].
- Le lexique CASOAR (Benamara *et al.*, 2014) qui comporte 236 interjections.
- Un lexique de 595 émoticônes que nous avons collecté manuellement à partir de Twitter.

4.3.2 Expériences et résultats

Pour l'apprentissage, nous avons testé plusieurs classifieurs sous la plateforme Weka avec les paramètres par défaut, à savoir : Sequential Minimal Optimization (SMO), Decision Tree (DT) et Naïve Bayes (NB). Les meilleurs résultats ont été obtenus avec SMO. Par conséquent, nous ne présentons ici que les résultats obtenus avec SMO pour l'ensemble des expériences.

Comme nous avons 3 corpus (*NegOnly*, *NoNeg* et *All*), nous avons entraîné 3 classifieurs, un par corpus, notés $C_{NegOnly}$, C_{NoNeg}, et C_{All}. Comme le nombre d'instances ironiques

[3] http://www.linternaute.com/dictionnaire/fr/usage/argot/1/

dans *NegOnly* est relativement petit (470 tweets), le classifieur $C_{NegOnly}$ a été entraîné sur un sous-ensemble équilibré de 940 tweets avec une validation croisée sur 10 échantillons. Pour C_{NoNeg} et C_{All}, nous avons utilisé 80% du corpus pour l'apprentissage et 20% pour le test, avec une distribution égale entre les instances ironiques (notées `IR`) et non ironiques (notées `NIR`)[4].

Le Tableau 4.3 illustre les résultats des 3 classifieurs avec tous les traits de surface en termes de précision (P), rappel (R) et f-mesure (F). Le classifieur $C_{NegOnly}$ obtient la meilleure valeur en terme d'exactitude (72,23%) par rapport aux classifieurs C_{NoNeg} et C_{All}. Ce résultat prouve de nouveau l'importance de la présence de la négation pour la détection des tweets ironiques bien que la négation ait représenté un handicap pour les humains dans le cadre de la compréhension de l'ironie lors de l'annotation manuelle.

	Ironique (IR)			**Non ironique (NIR)**			
	P	R	F	P	R	F	Exactitude
$C_{NegOnly}$	**0,847**	0,543	0,661	0,664	**0,902**	**0,765**	**72,23%**
C_{NoNeg}	0,635	0,623	0,629	0,630	0,642	0,636	63,25%
C_{All}	0,531	**0,955**	0,682	**0,774**	0,155	0,259	55,50%

TABLE 4.3 : Résultats du modèle SurfSystem.

Afin d'étudier l'apport de chaque trait surfacique au processus d'apprentissage, nous avons effectué l'apprentissage en rajoutant les traits un par un afin d'interpréter l'influence de chaque trait sur les résultats du classifieur en terme d'exactitude. Nous détaillons les résultats obtenus pour les corpus All, $NegOnly$ et $NoNeg$ dans le tableau 4.4.

Pour le corpus All, la combinaison de l'ensemble des traits de surface a permis d'avoir une valeur d'exactitude égale à $55,50\%$. Par contre, l'utilisation des traits *présence de ponctuation* et *présence de mot en majuscule* uniquement permet d'avoir une meilleure valeur d'exactitude de $56,31\%$. De même, pour le corpus $NegOnly$, l'exploitation des deux traits *présence de ponctuation* et *présence de mot en majuscule* permet d'avoir une meilleure valeur d'exactitude égale à $73,08\%$. Alors que pour le corpus $NoNeg$, la combinaison de l'ensemble des traits de surface a permis d'avoir une meilleure valeur d'exactitude de $63,25\%$.

En comparant les résultats obtenus avec ceux de l'état de l'art, nous avons la même tendance de valeurs obtenues pour d'autres langues et avec d'autres types de corpus. Par exemple, parmi ces travaux, (Burfoot & Baldwin, 2009) ont obtenu une f-mesure de $79,5\%$ sur un corpus d'articles de presse en anglais, (Carvalho *et al.*, 2009) ont obtenu une valeur de précision de $85,4\%$ sur un corpus d'articles de presse en portugais et (Tsur *et al.*, 2010) une précision de 50% sur un corpus de commentaires en anglais sur les produits d'Amazon.

[4]Pour C_{NoNeg} et C_{All}, nous avons testé une validation croisée sur 10 échantillons avec une distribution équilibrée entre les instances ironiques et non ironiques mais les résultats sont beaucoup moins bons.

Corpus	Traits	N° de traits	Ironique (IR)			Non ironique (NIR)			Exactitude
			P	R	F	P	R	F	
C_{All}	1.Ponctuation	1	0,517	0,848	0,642	0,577	0,207	0,305	52,75%
	2.Majuscule	1+2	0,556	0,631	0,591	0,573	0,495	0,531	**56,31%**
	3.Interjection	1 à 3	0,556	0,631	0,591	0,573	0,495	0,531	56,31%
	4.Emoticône	1 à 4	0,556	0,631	0,591	0,573	0,495	0,531	56,31%
	5.Citation	1 à 5	0,556	0,631	0,591	0,573	0,495	0,531	56,31%
	6.Connecteur de discours-opposition	1 à 6	0,556	0,631	0,591	0,573	0,495	0,531	56,31%
	7.Argot	1 à 7	0,556	0,631	0,591	0,573	0,495	0,531	56,31%
	8.Opposition	1 à 8	0,531	0,955	0,682	0,774	0,155	0,259	55,50%
	9.Exclamation	1 à 9	0,531	0,955	0,682	0,774	0,155	0,259	55,50%
	10.Interrogation	1 à 10	0,531	0,955	0,682	0,774	0,155	0,259	55,50%
	11.Exclamation et interrogation	1 à 11	0,531	0,955	0,682	0,774	0,155	0,259	55,50%
	12.Nombre de mots	1 à 12	0,531	0,955	0,682	0,774	0,155	0,259	55,50%
$C_{NegOnly}$	1.Ponctuation	1	0,492	0,279	0,356	0,497	0,713	0,586	49,57%
	2.Majuscule	1+2	0,883	0,532	0,664	0,665	0,93	0,776	**73,08%**
	3.Interjection	1 à 3	0,883	0,532	0,664	0,665	0,93	0,776	73,08%
	4.Emoticône	1 à 4	0,883	0,532	0,664	0,665	0,93	0,776	73,08%
	5.Citation	1 à 5	0,883	0,532	0,664	0,665	0,93	0,776	73,08%
	6.Connecteur de discours-opposition	1 à 6	0,883	0,532	0,664	0,665	0,93	0,776	73,08%
	7.Argot	1 à 7	0,883	0,532	0,664	0,665	0,93	0,776	73,08%
	8.Opposition	1 à 8	0,883	0,532	0,664	0,665	0,93	0,776	73,08%
	9.Exclamation	1 à 9	0.861	0.538	0.662	0.664	0.913	0.769	72.55%
	10.Interrogation	1 à 10	0.861	0.538	0.662	0.664	0.913	0.769	72.55%
	11.Exclamation et interrogation	1 à 11	0.861	0.538	0.662	0.664	0.913	0.769	72.55%
	12.Nombre de mots	1 à 12	0.847	0.543	0.661	0.664	0.902	0.765	72.23%
C_{NoNeg}	1.Ponctuation	1	0,581	0,833	0,685	0,705	0,4	0,51	61,62%
	2.Majuscule	1+2	0,581	0,833	0,685	0,705	0,4	0,51	61,62%
	3.Interjection	1 à 3	0,593	0,298	0,396	0,531	0,795	0,637	54,65%
	4.Emoticône	1 à 4	0,593	0,298	0,396	0,531	0,795	0,637	54,65%
	5.Citation	1 à 5	0,593	0,298	0,396	0,531	0,795	0,637	54,65%
	6.Connecteur de discours-opposition	1 à 6	0,575	0,642	0,607	0,595	0,526	0,558	58,37%
	7.Argot	1 à 7	0,584	0,66	0,62	0,61	0,53	0,567	59,53%
	8.Opposition	1 à 8	0,591	0,651	0,619	0,611	0,549	0,578	60,00%
	9.Exclamation	1 à 9	0,591	0,651	0,619	0,611	0,549	0,578	60,00%
	10.Interrogation	1 à 10	0,591	0,651	0,619	0,611	0,549	0,578	60,00%
	11.Exclamation et interrogation	1 à 11	0,591	0,651	0,619	0,611	0,549	0,578	60,00%
	12.Nombre de mots	1 à 12	0,635	0,623	0,629	0,63	0,642	0,636	**63,25%**

TABLE 4.4 : Résultats d'apprentissage trait par trait du modèle SurfSystem obtenus pour les corpus All, $NegOnly$ et $NoNeg$.

4.4 Le modèle PragSystem : Détection de l'ironie sur la base de traits contextuels internes

Dans cette section, nous détaillons une deuxième expérience qui regroupe les traits de surfaces utilisés dans le modèle précédent et des traits pragmatiques issus du contexte interne d'un tweet.

Un tweet est représenté par un vecteur composé de 6 groupes de traits. Certains d'entre eux ont été déjà exploités avec succès pour la détection de l'ironie dans des langues différentes du français (dans ce cas, nous citons les références), d'autres sont nouveaux. Dans cette section, nous allons présenter l'ensemble des traits utilisés ainsi que les résultats obtenus pour chaque groupe de traits et pour chacun des sous-corpus : *All, NegOnly, NoNeg*.

4.4.1 Traits utilisés

Traits de surface : ce sont les traits de surface que nous avons exploités lors de la première expérience détaillée dans la section précédente (cf. Section 4.3).

Traits de sentiment : ce sont les traits qui indiquent la présence de mots ou d'expressions d'opinion positive ou négative (Reyes & Rosso, 2011; Reyes & Rosso, 2012) et leur nombre (Barbieri & Saggion, 2014b). Nous avons ajouté 3 nouveaux traits :

- la présence de mots ou expressions de **surprise** ou d'**étonnement**, cf exemple 4.10

  (4.10) **Quelle surprise** la victoire de Naouelle ! ! ! Heureusement qu'il n y as pas eu de fuite pour garder le suspens ! ! #ironie #topchef

- la présence et le nombre d'**opinions neutres**, c'est-à-dire des opinions à la fois positives et négatives ou bien sous-entendues, cf exemple 4.11

  (4.11) C'est vrai que le PSG c'est un club qui a une histoire on **sens** que les mecs sont investie et qu'ils ne font pas sa pour l'argent #ironie)

Pour obtenir ces traits, nous avons utilisé deux lexiques :

1. **CASOAR** [5] (Benamara *et al.*, 2014), un lexique pour le français de 2 830 mots ou expressions d'opinion classés en 4 catégories sémantiques (REPORTAGE, JUGEMENT, SENTIMENT-APPRÉCIATION et CONSEIL), répartis comme suit : 1 142 adjectifs, 605

[5] https://projetcasoar.wordpress.com/

adverbes, 415 noms, 308 verbes, 292 expressions, 62 interjections et 6 conjonctions/ prépositions/pronoms.

Ce lexique distingue clairement les entrées purement subjectives des entrées de type intensifieur ou négation qui affectent les expressions d'opinion au niveau de la phrase soit en inversant, en intensifiant ou en minimisant sa polarité et / ou sa force. Les entrées subjectives sont principalement des adjectifs, des noms, des verbes, des adverbes de manière, des interjections et des émoticônes. Ils sont représentés selon trois axes :

- Polarité : elle peut être positive, négative ou neutre.
- Force : notée sur une échelle de 3 points [1, 3] (où 1 indique une force faible).
- Catégorie sémantique : cela correspond aux 4 catégories suivantes : REPORTAGE (e.g. *savoir, voir, annoncer*), JUGEMENT (e.g. *espoir, clair, nul*), SENTIMENT-APPRÉCIATION (e.g. *aimer, dommage, inquiétude*) et CONSEIL (e.g. *proposer, conseiller, souhaiter*).

2. **EMOTAIX**[6], un lexique émotionnel et affectif de 4 921 entrées regroupées en 9 catégories : malveillance, mal-être, anxiété, bienveillance, bien-être, sang-froid, surprise, impassibilité, émotion non spécifique. Il contient 1 308 entrées positives, 3 078 négatives et 535 neutres. Chaque catégorie est composée de sous-catégories, elles-mêmes associées à des catégories de base, par exemple : la catégorie « Malveillance » regroupe deux sous-catégories « Haine » et « Agressivité ». La sous-catégorie « Haine » est reliée à 4 catégories de base à savoir : *Ressentiment, Dégoût, Mépris et Irritation* (cf. Figure 4.1)

Traits pour les modifieurs : ce sont des traits binaires qui vérifient si un tweet contient : un **intensifieur** (*très, assez, beaucoup,* etc.) (Liebrecht *et al.*, 2013; Barbieri & Saggion, 2014b) (cf. exemple 4.12), une **modalité** (*devoir, vouloir, permettre*, etc.) (cf. exemple 4.13), un **mot de négation** (*ne, pas, plus, jamais, etc.*) (cf. exemple 4.14) ou un **verbe de discours rapporté** (*annoncer, dire, penser,* etc.) (cf. exemple 4.15). L'identification de ces modifieurs dans le corpus a été effectuée en utilisant l'analyseur syntaxique MElt en exploitant le lexique CASOAR.

(4.12) je suis **très très** surpris ! Bourdin a voté Hollande ? J y crois pas MDR #ironie

(4.13) #qc2014 P.Marois **veut** que P.Couillard «dénonce» l'Arabie S. Et «dénoncer» aussi Obama ? #ironie

[6] http://www.tropes.fr/download/EMOTAIX_2012_FR_V1_0.zip

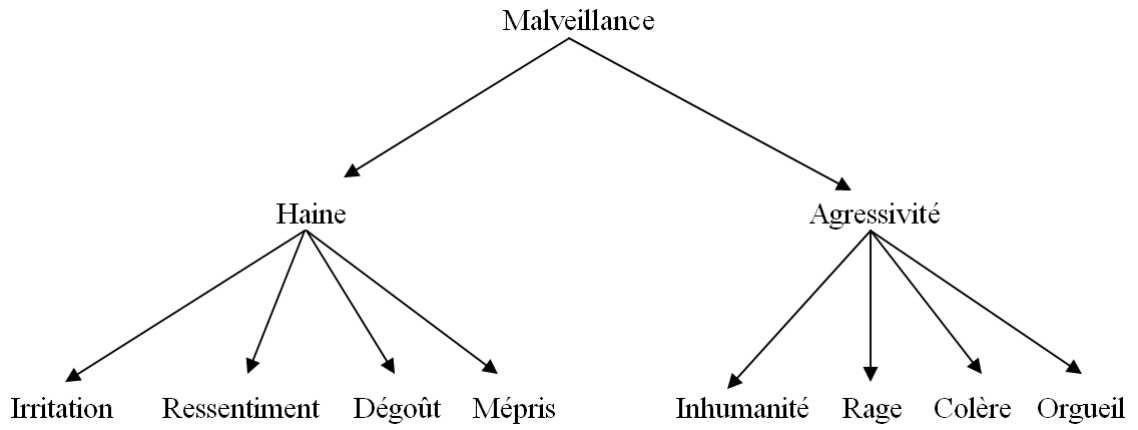

FIGURE 4.1 : Exemple de catégories et sous-catégories du lexique EMOTAIX.

(4.14) La bombe à fragmentation de DSK dans le Guardian C'est marrant mais **jamais** je n'ai pensé au complot #ironie

(4.15) Pas de ministère intouchable pour les économies 2014 **annonce** #Cahuzac, le ministre intouchable #ironie

Traits pour les modifieurs de sentiment : ils regroupent deux nouveaux traits qui indiquent si un tweet contient un mot d'opinion dans la portée d'une modalité ou d'un adverbe d'intensité. Par exemple, dans le tweet de l'exemple 4.16, le mot d'opinion positive *intelligent* est modifié par l'adverbe *très*.

(4.16) en même temps tu regarde les Tweets, tu vois bien que c'est des gens **très intelligents** qui on votez Hollande... #ironie

Traits de contexte : le contexte d'énonciation est important pour comprendre l'ironie d'un énoncé. Ces traits indiquent donc la présence/absence d'éléments de contexte tels que *les pronoms personnels* qui permettent d'identifier les tweets plus personnels, les *mots-clés* d'un thème donné et les *entités nommées* identifiées par l'analyseur syntaxique MElt.

Traits d'opposition : ils sont nouveaux par rapport à ceux traditionnellement utilisés. Ils indiquent la présence d'opposition explicite grâce à des patrons lexico-syntaxiques spécifiques. Ces traits ont été partiellement inspirés de (Riloff *et al.*, 2013) qui a proposé une méthode par « bootstrapping » pour détecter les tweets sarcastiques correspondant à une opposition entre un sentiment/opinion positif et une situation négative. Nous avons donc

étendu ce patron afin de traiter d'autres types d'opposition. Par exemple, nos patrons indiquent si un tweet contient (a) une opposition explicite de sentiment/opinion, ou (b) une opposition implicite entre une proposition subjective exprimant un sentiment/opinion et une proposition objective.

Soit P_+ (resp. P_-) une proposition subjective contenant une expression positive (resp. négative),
soit P_{obj} une proposition objective ne contenant pas d'expression d'opinion (P_{obj} peut contenir une négation ou non),
et soit Neg un opérateur qui change la polarité des mots subjectifs dans P_+ (resp. P_-),
les patrons pour **(a) les oppositions explicites** de sentiment/opinion sont de la forme :

- $[Neg(P_+)].[P'_+]$ ou $[P_+].[Neg(P'_+)]$:

  Vraiment, [je **comprend pas** pourquoi Jerome Safar s'est fait battre par les verts...]$_{Neg(P_+)}$ [**Super**]$_{P'_+}$ #Municipales2014 #Grenoble

- $[Neg(P_-)].[P'_-]$ ou $[P_-].[Neg(P'_-)]$:

  [**Las** des écoutes]$_{P_-}$, Sarkozy veut mener une vie de "citoyen normal". [**Pas idiot**]$_{Neg(P'_-)}$, c'est vrai que le "citoyen normal", lui, personne ne l'écoute.

- $[P_-].[P'_+]$ ou $[P_+].[P'_-]$:

  Émotion. Clap de fin pr le gouvernement Ayrault, probablement [le plus **mauvais** qu'ait jamais connu la 5ième république.]$_{P_+}$ [On s'est bien **marré**.]$_{P'_-}$

Les patrons pour **(b) les oppositions implicites** sont de la forme :

- $[Neg(P_+)].[P'_{obj}]$ ou $[P_{obj}].[Neg(P'_+)]$:

  Franchement, [je **ne comprends pas** pourquoi on critique Evra]$_{Neg(P_+)}$. Le côté gauche français est une forteresse ! ! ! [Rien ne passe]$_{P'_{obj}}$ #UKRFRA

- $[Neg(P_-)].[P'_{obj}]$ ou $[P_{obj}].[Neg(P'_-)]$:

  [Avoir pour seul chroniqueur politique nicodomenach...]$_{P_{obj}}$ [On ne peut cependant **pas douter** de la neutralité de Canal+.]$_{Neg(P'_-)}$ #foutagedegueule

- $[P_+].[P'_{obj}]$ ou $[P_{obj}].[P'_+]$:

  Le soccer aux Jeux olympiques [c'est une **bonne chose**]$_{P_+}$ parce qu'[on n'en voit nulle part ailleurs]$_{P'_{obj}}$

- $[P_{\_}].[P'_{obj}]$ ou $[P_{obj}].[P'_{\_}]$:
  [Kadhafi est mort tué d'une balle.]$_{P_{obj}}$ [C'est **moche** la guerre.]$_{P'_{\_}}$ Je suis contre la guerre. #bhl

Nous considérons qu'un mot d'opinion est dans la portée d'une négation si le mot de négation et le mot d'opinion sont séparés par au maximum deux tokens. Cette règle simple a montré son efficacité étant donné que les tweets sont des messages courts limités à 140 caractères.

4.4.2 Expériences et résultats

Pour ces expériences aussi, nous ne présentons que les meilleurs résultats, obtenus avec le classifieur SMO.

Comme nous avons 3 corpus (*NegOnly*, *NoNeg* et *All*), nous avons entraîné 3 classifieurs, un par corpus, notés $C_{NegOnly}$, C_{NoNeg}, et C_{All} dans les mêmes conditions que pour le modèle SurfSystem. Les résultats présentés ici ont été obtenus en entraînant C_{NoNeg} sur 1 720 tweets et en testant sur 430 tweets. C_{All} a été entraîné sur 2 472 tweets (1 432 contenant une négation dont 404 `IR` et 1 028 `NIR`) et testé sur 618 tweets (360 contenant une négation dont 66 `IR` et 294 `NIR`).

A Étude de la pertinence des traits au processus d'apprentissage

Pour chaque classifieur, nous avons mesuré l'impact de chaque groupe de traits (détaillés dans la Section 4.4.1) sur la tâche de détection de l'ironie. Pour toutes les expériences, nous avons utilisé les traits de surface comme approche de base. Pour C_{NoNeg} et $C_{NegOnly}$, le trait qui indique la présence d'une négation a été désactivé.

B Apport de chaque groupe de traits au processus d'apprentissage

Nous avons effectué l'apprentissage en rajoutant les traits un par un afin d'interpréter l'influence de chaque trait sur les résultats du classifieur en terme d'exactitude. Nous détaillons les résultats obtenus pour les corpus All, $NoNeg$ et $NegOnly$ dans les tableaux 4.5, 4.6 et 4.7.

En ajoutant les traits un par un, nous avons obtenu une meilleure valeur d'exactitude de $87,70\%$ pour le corpus All en combinant l'ensemble des traits (Tableau 4.5). Par contre, pour le corpus $NegOnly$, nous avons obtenu une meilleure valeur d'exactitude égale à $73,82\%$ en combinant les traits de surface avec les traits de sentiment. De même, pour le corpus $NoNeg$, nous avons obtenu la meilleure valeur d'exactitude à $68,83\%$ en combinant les traits de surface avec les traits de sentiment uniquement.

Catégorie des traits	Traits	N° de traits	Ironique (IR)			Non ironique (NIR)			Exactitude
			P	R	F	P	R	F	
Traits de surface (baseline)		1 à 12	0,531	0,955	0,682	0,774	0,155	0,259	55,50%
Traits de sentiment	13.Opinion positive	1 à 13	0,567	0,563	0,565	0,566	0,57	0,568	56,63%
	14.Opinion négative	1 à 14	0,567	0,563	0,565	0,566	0,57	0,568	56,63%
	15.Opinion neutre	1 à 15	0,567	0,563	0,565	0,566	0,57	0,568	56,63%
	16.Nombre d'opinion positive	1 à 12+ 16	0,565	0,647	0,603	0,587	0,502	0,541	57,44%
	17.Nombre de mot d'opinion négative	1 à 12 + 16 à 17	0,575	0,686	0,625	0,61	0,492	0,545	58,89%
	18.Nombre de mot d'opinion neutre	1 à 12+ 16 à 18	0,584	**0,699**	**0,636**	**0,625**	0,502	**0,557**	**60,03%**
	19.Surprise ou étonnement	1 à 12+ 16 à 19	0,584	0,699	0,636	0,625	0,502	0,557	60,03%
Trait pour les modifieurs de sentiment	20.Intensifieur suivi opinion	1 à 12+ 16 à 20	**0,588**	0,683	0,632	0,622	0,521	**0,567**	**60,19%**
	21.Modalité suivi opinion	1 à 12+ 16 à 21	0,586	**0,686**	0,632	0,621	0,515	0,563	60,03%
Traits pour les modifieurs	22.Intensifieur	1 à 12 + 16 à 22	0,582	0,676	0,626	0,614	0,515	0,56	59,54%
	23.Modalité	1 à 12 +16 à 23	0,584	0,676	0,627	0,615	0,518	0,562	59,70%
	24.Reportage	1 à 12 +16 à 24	0,596	0,696	0,642	0,634	0,528	0,576	61,16%
	25.Négation	1 à 12 +16 à 25	**0,942**	**0,786**	**0,857**	**0,817**	**0,951**	**0,879**	**86,89%**
Traits d'opposition	26.Opposition implicite	1 à 12 + 16 à 26	0,942	0,786	0,857	0,817	0,951	0,879	86,89%
	27.Opposition explicite	1 à 12 +16 à 27	0,942	0,786	0,857	0,817	0,951	0,879	86,89%
Traits de contexte	28.Pronom personnel	1 à 12 +16 à 28	0,942	0,786	0,857	0,817	0,951	0,879	86,89%
	29.Présence du topic dans le texte	1 à 12 +16 à 29	0,942	0,786	0,857	0,817	0,951	0,879	86,89%
	30.Entité nommée	1 à 12 +16 à 30	**0,93**	**0,816**	**0,869**	**0,836**	**0,939**	**0,884**	**87,70%**

TABLE 4.5 : Résultats d'apprentissage trait par trait obtenus pour le corpus *All*.

Catégorie des traits	Traits	N° de traits	Ironique (IR)			Non ironique (NIR)			Exactitude
			P	R	F	P	R	F	
Traits de surface (baseline)		1 à 12	**0,847**	0,543	0,661	0,664	**0,902**	**0,765**	**72,23%**
Traits de sentiment	13.Opinion positive	1 à 13	0,853	0,543	0,663	0,665	0,906	0,767	72,44%
	14.Opinion négative	1 à 14	**0,867**	0,54	0,666	0,666	**0,917**	**0,772**	**72,87**%
	15.Opinion neutre	1 à 15	0,861	0,543	0,666	0,666	0,913	0,77	72,76%
	16.Nombre d'opinion positive	1 à 12 + 16	0,856	0,543	0,664	0,665	0,909	0,768	72,55%
	17.Nombre de mot d'opinion négative	1 à 12 + 16 à 17	0,859	0,543	0,665	0,666	0,911	0,769	72.65%
	18.Nombre de mot d'opinion neutre	1 à 12 + 16 à 18	0,863	0,538	0,663	0,665	0,915	0,77	72,65%
	19.Surprise ou étonnement	1 à 12 + 16 à 19	**0,854**	0,574	0,687	0,679	**0,902**	**0,775**	**73,82%**
Trait pour les modifieurs de sentiment	20.Intensifieur suivi opinion	1 à 15 + 19 à 20	0,836	0,543	0,658	0,661	0,894	0,76	71,80%
	21.Modalité suivi opinion	1 à 15 + 19 à 21	0,844	0,543	0,661	0,663	0,9	0,764	72.12%
Traits pour les modifieurs	22.Intensifieur	1 à 15 + 19 à 22	0,842	0,543	0,66	0,662	0,898	0,762	72,02%
	23.Modalité	1 à 15 + 19 à 23	0,842	0,543	0,66	0,662	0,898	0,762	72,02%
	24.Reportage	1 à 15 + 19 à 24	0,833	0,543	0,657	0,661	0,891	0,759	71,70%
	25.Négation	-	-	-	-	-	-	-	-
Traits d'opposition	26.Opposition implicite	1 à 15 + 19 à 26	0,851	0,57	0,683	0,677	0,9	0,773	73,51%
	27.Opposition explicite	1 à 15 + 19 à 27	0,851	0,57	0,683	0,677	0,9	0,773	73,51%
Traits de contexte	28.Pronom personnel	1 à 15 + 19 à 28	0,851	0,57	0,683	0,677	0,9	0,773	73,51%
	29.Présence du topic dans le texte	1 à 15 + 19 à 29	0,851	0,57	0,683	0,677	0,9	0,773	73.51%
	30.Entité nommée	1 à 15 + 19 à 30	0,883	0,532	0,664	0,665	0,93	0,776	73,08%

TABLE 4.6 : Résultats d'apprentissage trait par trait obtenus pour le corpus $NegOnly$.

Catégorie des traits	Traits	N° de traits	Ironique (IR)			Non ironique (NIR)			Exactitude
			P	R	F	P	R	F	
Traits de surface (baseline)		1 à 12	0,635	0,623	0,629	0,63	0,642	0,636	**63,25%**
Traits de sentiment	13.Opinion positive	1 à 13	0,668	0,6	0,632	0,637	0,702	0,668	65,11%
	14.Opinion négative	1 à 14	0,695	0,647	0,67	0,67	0,716	0,692	68,13%
	15.Opinion neutre	1 à 15	**0,701**	0,656	0,678	0,677	0,**0,721**	0,698	**68,83%**
	16.Nombre d'opinion positive	1 à 12 + 16	0,677	0,586	0,628	0,635	0,721	0,675	65,34%
	17.Nombre de mot d'opinion négative	1 à 12 + 16 à 17	0,682	0,567	0,619	0,629	0,735	0,678	65,11%
	18.Nombre de mot d'opinion neutre	1 à 12 + 16 à 18	0,71	0,581	0,639	0,646	0,763	0,699	67,20%
	19.Surprise ou étonnement	1 à 12 + 16 à 19	0,656	0,665	0,661	0,66	0,651	0,656	65,81%
Trait pour les modifieurs de sentiment	20.Intensifieur suivi opinion	1 à 15 + 19 à 20	0,695	0,656	0,675	0,674	0,712	0,692	68,37%
	21.Modalité suivi opinion	1 à 15 + 19 à 21	0,695	0,656	0,675	0,674	0,712	0,692	68,37%
Traits pour les modifieurs	22.Intensifieur	1 à 15 + 19 à 22	0,695	0,656	0,675	0,674	0,712	0,692	68,37%
	23.Modalité	11 à 15 + 19 à 23	0,706	0,647	0,675	0,674	0,73	0,701	68,83%
	24.Reportage	1 à 15 + 19 à 24	0,698	0,656	0,676	0,675	0,716	0,695	68,60%
	25.Négation								
Traits d'opposition	26.Opposition implicite	1 à 15 + 19 à 26	0,654	**0,702**	0,677	0,678	0,628	0,652	66,51%
	27.Opposition explicite	1 à 15 + 19 à 27	0,652	0,637	0,647	0,645	0,66	0,653	64,88%
Traits de contexte	28.Pronom personnel	1 à 15 + 19 à 28	0,66	0,651	0,656	0,656	0,665	0,661	65,81%
	29.Présence du topic dans le texte	1 à 15 + 19 à 29	0,665	0,665	0,665	0,665	0,665	0,665	66,51%
	30.Entité nommée	1 à 15 + 19 à 30	0,67	0,651	0,66	0,661	0,679	0,67	66,51%

TABLE 4.7 : Résultats d'apprentissage trait par trait obtenus pour le corpus $NoNeg$.

C Résultats pour les meilleures combinaisons de traits

Nous avons étudié l'influence des traits pour chaque sous-corpus afin de sélectionner le sous-ensemble de traits qui forment la meilleure combinaison permettant d'atteindre la meilleure valeur d'exactitude. Nous détaillons les résultats obtenus ainsi que la meilleure combinaison de traits pour chaque sous-corpus dans les tableaux 4.8, 4.9 et 4.10.

Traits	Ironique (IR)			Non ironique (NIR)			Exactitude
	P	R	F	P	R	F	
Baseline	0,531	0,955	0,682	0,774	0,155	0,259	55,50%
Tous les traits	0,93	0,816	0,869	0,836	0,939	0,884	87,70%
Meilleure combinaison de traits : majuscule, mot d'opposition, nombre de mots, intensifieur suivi d'opinion, opposition explicite, opposition implicite, négation, surprise/étonnement	0,93	0,816	0,869	0,836	0,939	0,884	**87,70%**

TABLE 4.8 : Comparaison des résultats obtenus pour le corpus All.

Traits	Ironique (IR)			Non ironique (NIR)			Exactitude
	P	R	F	P	R	F	
Baseline	0,847	0,543	0,661	0,664	0,902	0,765	72,23%
Tous les traits	0,847	0,564	0,677	0,673	0,898	0,769	73.08%
Meilleure combinaison de traits : majuscule, citation, opposition implicite	0,889	0,56	0,687	0,679	0,93	0,785	**74,46%**

TABLE 4.9 : Comparaison des résultats obtenus pour le corpus $NegOnly$.

Dans une deuxième expérience, nous avons calculé l'exactitude pour chaque groupe de traits afin d'avoir une idée de l'influence de chaque groupe de traits sur la performance (cf. Tableau 4.11).

Comparée aux autres traits, la baseline obtient de bons résultats sur $NegOnly$ alors que les résultats sont beaucoup moins bons sur les 2 autres corpus. Pour $NoNeg$, les meilleurs résultats sont obtenus en utilisant les traits {longueur du tweet, interjection, connecteur discursif, ponctuation, citation} alors que pour All, la meilleure combinaison correspond à {présence de ponctuation, mot en lettres majuscules}.

Les principales conclusions que l'on peut tirer du tableau 4.11 sont :

- Dans *NegOnly*, les traits sémantiques pris séparément (sentiment, modifieurs, opposition, etc.) ne sont pas suffisants pour classer les tweets `NIR` et `IR`. On note en parti-

Traits	Ironique (IR)			Non ironique (NIR)			Exactitude
	P	R	F	P	R	F	
Baseline	0,591	0,651	0,619	0,611	0,549	0,578	63,25%
Tous les traits	0,673	0,642	0,657	0,658	0,688	0,673	66,51%
Meilleure combinaison de traits : ponctuation, majuscule, interjection, citation, connecteurs de discours et d'opposition, opposition, intensifieur, modalité, opinion positive, opinion négative, opinion neutre, nombre de mots, intensifieur suivi d'opinion, modalité suivi d'opinion	0.711	0.651	0.68	0.678	0.735	0.705	**69.30%**

TABLE 4.10 : Comparaison des résultats obtenus pour le corpus $NoNeg$.

	NegOnly	*NoNeg*	*All*
Baseline (traits de surface)	**73,08**	63,25	55.50
Meilleurs traits de surface	73,08	64,65	56,31
Meilleurs traits de sentiment	57,02	**67,90**	58,25
Modifieurs de sentiment	53,51	56,51	51,94
Modifieurs	53,72	55,81	**86,89**
Opposition	55,31	63,02	79,77
Contexte interne	55,53	53,25	53,55

TABLE 4.11 : Résultats des 3 expériences par groupe de traits en terme d'exactitude.

culier que les résultats de la baseline pour la classe `NIR` sont meilleurs que ceux pour `IR` (respectivement 77,60 et 66,40 en f-mesure).

- Les traits de sentiment sont les plus fiables pour C_{NoNeg} en utilisant le trait de surprise/étonnement associé aux traits de fréquence des mots d'opinion. Ici aussi, la classe `NIR` obtient 12,7 points de plus que la classe `IR` avec une f-mesure de 73,30.

- Les traits pour les modifieurs et oppositions sont les meilleurs pour C_{All}. Comme pour les autres corpus, on remarque que les prédictions du classifieur sont meilleures pour la classe `NIR` que pour `IR` mais avec un écart plus petit (2,2 en utilisant les modifieurs et 7,4 en utilisant les oppositions).

Le tableau 4.12 détaille les résultats globaux dans le cas où les classifieurs sont entraînés sur tous les traits pertinents de chaque groupe. Les résultats sont donnés en termes de précision (P), rappel (R), f-mesure (macro-moyenne) et exactitude. Les résultats sont

	Ironique (IR)			**Non ironique (NIR)**		
	P	R	F	P	R	F
$C_{NegOnly}$	0,889	0,56	0,687	0,679	0,933	0,785
C_{NoNeg}	0,711	0,651	0,68	0,678	0,735	0,705
C_{All}	0,93	0,816	0,869	0,836	0,939	0,884

	Résultats (meilleure combinaison)	
	f-mesure (macro-moyenne)	Exactitude
$C_{NegOnly}$	73,60	74,46
C_{NoNeg}	69,25	69,30
C_{All}	**87,65**	**87,70**

TABLE 4.12 : Résultats pour les meilleures combinaisons de traits.

meilleurs pour C_{All} que pour $C_{NegOnly}$ et C_{NoNeg}. Ces résultats sont obtenus en utilisant les **3 traits de surface** {mots en lettres majuscules, connecteurs d'opposition, longueur du tweet}, **les modifieurs** {présence d'intensifieurs et négations} et **les traits d'opposition** {présence d'opposition explicite et implicite}. La meilleure combinaison pour $C_{NegOnly}$ est composée de **2 traits de surface** {mots en lettres majuscules, citation} et des **traits d'opposition**. Finalement, si on ne considère pas les tweets contenant des négations (i.e. C_{NoNeg}), les performances tombent à 69,30% d'exactitude. La meilleure combinaison est la suivante : **traits de surface** {ponctuation, mots en lettres majuscules, interjection, citation, connecteurs discursifs, mots d'opposition, longueur du tweet}, **traits de sentiment** {présence de mots d'opinion positifs/négatifs/neutres} et **traits pour les modifieurs de sentiment** {mots d'opinion modifiés par un intensifieur ou une modalité}.

Ces résultats nous permettent de tirer les 4 conclusions suivantes :

- Les traits de surface sont essentiels pour la détection de l'ironie, surtout pour les tweets sans négation,
- La négation est un trait important pour cette tâche mais ne suffit pas : en effet, parmi les 76 tweets mal classés par C_{All}, 60% contiennent des négations (37 `IR` et 9 `NIR`),
- Pour les tweets contenant une négation, les traits d'opposition sont les plus efficaces,
- Les mots d'opinion sont plus susceptibles d'être utilisés dans les tweets sans négation.

4.4.3 Discussion

Les résultats obtenus lors des expériences avec les corpus All ($P = 93\%$, $R = 81,6\%$, f-mesure $= 86,9\%$), $NegOnly$ ($P = 88,9\%$, $R = 56\%$, f-mesure $= 68,7\%$) et $NoNeg$

($P = 71,1\%$, $R = 65,1\%$, f-mesure $= 68\%$) sont très encourageants en comparaison avec d'autres travaux qui ont traité la même tâche et avec le même type de corpus. Par exemple, (Liebrecht *et al.*, 2013) ont obtenu une valeur de précision de 30% pour le néerlandais. Pour l'anglais, des f-mesures de 76% (Reyes *et al.*, 2013), 76% (Barbieri & Saggion, 2014b) et $88,76\%$ (Joshi *et al.*, 2015) ont été obtenues.

Pour les 3 classifieurs, une analyse d'erreur montre que les erreurs de classification sont principalement dues à quatre facteurs :

- **Présence de comparaison** : Ceci est une forme d'ironie par laquelle on attribue des caractéristiques à un élément en le comparant à un élément complètement différent (*e.g. "Benzema en équipe de France c'est comme le dimanche. Il sert à rien"*). Ce type d'ironie utilise souvent des marqueurs de comparaison (cf. chapitre 1 et chapitre 2, section 2.5). Nous ne traitons pas ce phénomène pour le moment mais une approche par similarité sémantique pourrait être utilisée (Veale & Hao, 2010).

- **Absence de contexte** : Elle est responsable de la majorité des erreurs. En effet, l'interprétation des tweets mal classés nécessite des connaissances contextuelles extérieures aux tweets. Cette absence de contexte peut se manifester de plusieurs façons :

  1. Le thème du tweet n'est pas mentionné (*e.g. Elle nous avait manqué tiens ! #ironie*) ou bien l'ironie doit être inférée des hashtags (e.g. *#poissondavril*) ;
  2. L'ironie porte sur une situation spécifique (ironie situationnelle) (*e.g. #Sarkozy fut un des auteurs de la loi du 9 mars 2004 encadrant les écoutes téléphoniques #ironie* : l'ironie vient de la situation où Sarkozy a été victime d'écoutes téléphoniques alors qu'il avait lui-même proposé la loi le permettant) ;
  3. La présence de fausses assertions (*e.g. Ne vous inquietez pas. Le Sénégal sera champion du monde de Football*) ;
  4. Des oppositions qui impliquent une contradiction entre 2 mots qui ne sont pas sémantiquement reliés (*e.g. "ONU"* et *"organisation terroriste"*, *"Tchad"* et *"élection démocratique"*). Ce cas est plus fréquent dans les tweets sans négation alors que les cas (2) et (3) sont plus fréquents dans les tweets avec négation.

- **Humour** : L'humour dans les tweets accompagnés du hashtag *#ironie* peut provoquer une mauvaise classification automatique. Ceci est dû à l'utilisation d'un vocabulaire spécifique exprimant l'humour qui n'est pas couvert par les traits que nous avons utilisés pour la détection de l'ironie (*e.g. La bombe à fragmentation de DSK dans le Guardian C'est marrant mais jamais je n'ai pensé au complot #ironie*).

- **Mauvaise utilisation des hashtags** *#ironie* ou *#sarcasme* dans des tweets contenant des textes non ironiques (*e.g. #Seydou remercie #Thauvin pour "tout son professionalisme" et "son amour pour le club" #OM #Losc #ironie* où l'intention ironique de l'auteur n'est pas évidente puisqu'il ne fait que citer des propos).

L'analyse d'erreurs a également montré que les erreurs de classification étaient globalement les mêmes que les cas de désaccord entre annotateurs humains lors de la phase d'annotation manuelle décrite dans la section 4.2 (annotation pour vérifier la fiabilité des hashtags de référence). Cette phase d'annotation a aussi montré que la présence d'une négation est une cause d'ambiguïté pour la compréhension et l'identification de l'ironie par les humains. Cette observation est valable aussi pour la classification automatique des tweets contenant des négations puisque les résultats sont moins bons en termes d'exactitude sur $NegOnly$ $(74,46\%)$ que sur All $(87,70\%)$. En revanche, la valeur d'exactitude a atteint son minimum pour le corpus $NoNeg$ $(69,30\%)$. Ces résultats laissent penser que la négation est un indicateur primordial pour la détection automatique de l'ironie bien qu'il ait été considéré comme un obstacle pour la compréhension de l'ironie par les humains et pour le classifieur ne traitant que les tweets avec négation.

L'analyse d'erreur sur l'ensemble des expériences réalisées ayant montré que la présence de la négation ainsi que l'absence de contexte sont responsables de la majorité des erreurs, nous avons décidé de nous focaliser sur ces deux phénomènes afin de proposer une nouvelle méthode qui permet d'améliorer la classification des tweets mal classés par le classifieur. Nous détaillons cette méthode dans la prochaine section.

4.5 Le modèle QuerySystem : Vers un modèle pragmatique contextuel pour la détection automatique de l'ironie

L'ensemble des expériences réalisées dans les deux sections précédentes a confirmé la nécessité de recourir à des connaissances externes afin de comprendre le sens ironique d'un tweet. Cette problématique a été évoquée par les annotateurs humains durant la campagne d'annotation détaillée dans le chapitre 3 ainsi que dans quelques travaux récents (cf. les travaux de Wallace et al. (2015), Bamman & Smith (2015) et Joshi et al. (2016) présentés dans la section du chapitre 2).

Dans cette section, nous présentons une nouvelle approche qui consiste à la correction de la sortie du système de détection automatique décrit dans la section précédente afin de bien classer les tweets en ironiques/non ironiques en faisant recours aux connaissances externes.

4.5.1 Approche proposée

L'hypothèse principale de notre approche suppose que les utilisateurs qui publient des tweets à propos de sujets médiatisés ont tendance à commenter ou critiquer une situation ou une personne qui a fait l'actualité. Il est donc possible de vérifier si un fait ou un événement

relaté dans un tweet s'est produit dans la réalité. Dans le cas contraire, il est probable que le tweet soit une fausse affirmation et donc susceptible d'exprimer de l'ironie.

Notre approche consiste à chercher le *contexte externe* d'un tweet en interrogeant Google via son API pour vérifier la véracité des tweets. En partant de la définition de l'ironie : « L'ironie est une manière de se moquer de quelqu'un ou quelque chose en disant le contraire de ce qu'on veut faire entendre (Le Petit Robert) », nous avons décidé de nous focaliser sur un type particulier de tweets ironiques où l'ironie se manifeste par la présence d'une négation. Nous avons fait ce choix pour trois raisons principales :

- Les négations peuvent être le moyen d'exprimer une fausse affirmation ironique (*e.g. Si Hollande est élu, il serait capable de donner des responsabilités à DSK. Pourquoi pas la direction du FMI tant qu'on y est.*).
- La forte présence des négations dans les tweets ($62,75\%$ des tweets contiennent une négation) et des fausses affirmations (elles représentent 56% des tweets ironiques avec contradiction implicite).
- Notre classifieur obtient de moins bons résultats quand il est appliqué sur les tweets contenant une négation (cf. section 4.5)

Par conséquent, l'approche proposée a pour but de vérifier la véracité des tweets avec négation qui ont été classés comme non ironiques par le classifieur alors qu'ils sont ironiques selon les hashtags de référence. Ainsi, si un tweet de la forme $Not(P)$ a été classé comme non ironique alors que P est vérifié sur le Web, alors la classe du tweet doit être corrigée en classant le tweet comme ironique au lieu de non ironique.

Le changement de la classe d'un tweet s'effectue selon l'algorithme suivant. Considérons $WordsT$ l'ensemble de mots excluant les mots vides qui appartiennent à un tweet t, et soit kw le mot-clé (topic) utilisé pour collecter le tweet t. Soit $N \subset WordsT$ l'ensemble des mots de négation de t. L'algorithme proposé est défini comme suit :

1. Segmenter t en un ensemble de phrases S.

2. Pour chaque phrase $s \in S$ telle que $\exists neg \in N$ et $neg \in s$:

   2.1. Supprimer les symboles # et @, les émoticônes, et neg, puis extraire l'ensemble des tokens $P \subset s$ qui sont dans la porté de neg (en respectant une distance maximale de 2 tokens).

   2.2. Générer une requête $Q_1 = P \cup kw$ et la soumettre à Google qui renverra 20 résultats au maximum formés d'un titre et d'un extrait (snippet).

2.3. A partir des résultats renvoyés par Google, ne garder que les plus fiables (Wikipedia, articles de journaux, sites Web ne contenant pas "blog" or "twitter" dans leur URL). Ensuite, pour chaque résultat, si les mots-clés de la requête Google ont été trouvés dans le titre ou dans l'extrait, alors t est considéré comme étant ironique. STOP.

3. Générer une deuxième requête $Q_2 = (WordsT - N) \cup kw$ et la soumettre à Google. Puis suivre la procédure décrite dans 2.3. Si Q_2 est trouvé, alors t est considéré comme ironique. Sinon, la classe prédite par le classifieur reste inchangée.

En lançant une requête, Google renvoie une page web contenant une liste de **snippets** qui désignent la façon dont Google présente les résultats de recherche. Le snippet comprend souvent un titre, une URL, une description et parfois des informations supplémentaires (image, votes, prix, ...). Google limite le nombre de caractères à 66 pour le titre, 156 pour la description et 65 pour l'URL (sauf exception). Par conséquent, la zone de description doit contenir un nombre de caractères inférieur à la limite pour éviter d'avoir un contenu tronqué. Le snippet peut être personnalisé par les webmasters en adaptant le contenu des pages web. Google a également mis en place des règles pour comprendre le contenu des pages web. Ainsi, un *rich snippet* désigne un système de balisage qui est invisible pour l'utilisateur mais lu par Google. Ce dernier comprend mieux certaines données (prix d'un produit, temps de réalisation d'une recette, ...). Le moteur de recherche peut alors adapter un snippet pour mieux présenter les résultats aux internautes. Profitant de l'existence d'une API Google qui permet la récupération de snippets ainsi que le paramétrage du nombre de snippets à récupérer par requête, nous avons décidé d'exploiter cette technologie dans notre expérience en récupérant au maximum 20 snippets par requête Google.

Illustrons l'algorithme proposé par un exemple de tweet de notre corpus et collecté par l'API de Twitter avec le mot-clé *Valls* :

(4.17) #Valls a appris la mise sur écoute de #Sarkozy en lisant le journal. Heureusement qu'il n'est pas Ministre de l'Intérieur.

L'étape 1 de l'algorithme permet la segmentation du tweet en deux phrases :

- $s1$ (#Valls a appris la mise sur écoute de #Sarkozy en lisant le journal.)
- $s2$ (Heureusement qu'il n'est pas Ministre de l'Intérieur.).

L'étape 2.1 de l'algorithme supprime les mots de négation "n'" et "pas" et isole le segment dans la portée de la négation. En résultat, nous avons $P = \{ministre, intérieur\}$.

L'étape 2.2 génère une première requête $Q_1 =$ *Valls ministre intérieur*.

L'étape 2.3 permet d'avoir un ensemble de 20 snippets. Les 2 premiers snippets sont les suivants :

<Résultat id="1">
<Titre>*Manuel Valls - Wikipedia*</Titre>
<Url>https://fr.wikipedia.org/wiki/Manuel_Valls</Url>
<Snippet>... Homme politique français. Pour le compositeur espagnol, voir Manuel **Valls** (compositeur). .... **Valls** a été nommé **ministre** de l'**Intérieur** dans le Cabinet d'Ayrault en mai 2012.</Snippet>.

<Résultat id="2">
<Titre>Pendant les jours de sang qu'a connus la France, le vrai président ...</Titre>
<Url>http://www.atlantico.fr/decryptage/pendant-jours-sang-qu-connus-france-vrai\-president-republique-est-appele-manuel-valls\-benoit-rayski-1950467.html</Url>
<Snippet>12 janv. 2015 ... Mais heureusement pour notre dignité qu'il était là. ... Valls a été rocardien et il en a gardé le meilleur : le parler vrai. ... Alors que son ministre de l'Intérieur, Bernard Cazeneuve, poussait des ronrons de satisfaction, saluant ...</Snippet>

L'ensemble des mots-clés sont présents dans le texte du premier snippet. Chaque mot-clé de la requête est marqué par les balises . Par conséquent, nous considérons que la requête est vérifiée, que la proposition *Heureusement qu'il n'est pas Ministre de l'Intérieur* est une fausse affirmation et nous pouvons conclure que le tweet est ironique.

4.5.2 Expériences et résultats

Nous avons effectué plusieurs expériences pour évaluer comment la méthode à base de requête améliore la classification des tweets. À cette fin, nous avons appliqué la méthode sur les deux corpus All et $NegOnly$ (le corpus $NoNeg$ ne contenant pas de négation, les requêtes ne peuvent pas être appliquées) :

- ① Une première expérience évalue la méthode sur les tweets avec négation classés comme non ironique (`NIR`) par le classifieur *PragSystem* mais qui sont ironiques selon les hashtags de référence.
- ② : Une deuxième expérience consiste à appliquer la méthode sur tous les tweets avec négation qui ont été classés comme `NIR` par le classifieur *PragSystem*, peu importe si la classe prédite est correcte ou non.

Le tableau 4.13 illustre les résultats des deux expériences.

	①		②	
Nombre de tweets `NIR` pour lesquels :	*All*	*NegOnly*	*All*	*NegOnly*
une requête est appliquée	37	207	327	644
des résultats sont trouvés sur Google	25	102	166	331
la classe est corrigée en `IR`	5	35	69	178
Exactitude du classifieur	87,7	74,46	**87,7**	**74,46**
Exactitude après les requêtes	**88,51**	**78,19**	78,15	62,98

TABLE 4.13 : Résultats de la méthode à base de requêtes Google (expériences 1 et 2).

Tous les scores pour la méthode à base de requête sont statistiquement significatifs par rapport aux scores du classifieur ($p\_value < 0,0001$ calculée avec le test de McNemar). Une analyse d'erreur montre que 65% des tweets qui sont encore mal classés avec cette méthode sont les tweets pour lesquels trouver leur contenu en ligne est presque impossible parce que ce sont des tweets personnels ou par manque de contexte interne. Une conclusion à tirer est que cette méthode ne doit pas être appliquée sur ce type de tweets.

C'est pourquoi, nous avons refait les expériences ① et ② uniquement sur les tweets qui ne sont pas susceptibles de contenir des contenus personnels. Pour sélectionner ces derniers, nous avons utilisé certains traits de contexte interne. Ainsi, nous avons considéré que ce sont des tweets qui ne devaient pas contenir de pronoms personnels à la premier personne (*je, moi, nous*) et qui contiennent des entités nommées : en effet, les tweets ne contenant pas de tels pronoms personnels ni d'entité nommée sont susceptibles de ne pas contenir de contenu personnel qui serait impossible à valider sur le Web (*e.g. Elle nous avait manqué tiens ! #ironie*).

Le tableau 4.14 illustre les résultats de ces expériences. Tous les scores pour la méthode à base de requêtes sont également statistiquement significatifs par rapport aux scores du classifieur.

Pour l'expérience ①, la méthode n'est pas appliquée sur le corpus All car tous les tweets mal classés contiennent un pronom personnel et aucune entité nommée. D'une façon globale, la méthode à base de requêtes améliore les résultats du classifieur dans tous les cas,

	①		②	
Nombre de tweets `NIR` *pour lesquels :*	*All*	*NegOnly*	*All*	*NegOnly*
une requête est appliquée	0	18	40	18
des résultats sont trouvés sur Google	-	12	17	12
la classe est corrigée en `IR`	-	4	7	4
Exactitude du classifieur	87,7	74,46	**87,7**	74,46
Exactitude après les requêtes	**87,7**	**74,89**	86,57	**74,89**

TABLE 4.14 : Résultats de la méthode à base de requêtes Google pour les tweets non personnels (Expérience 3).

sauf pour le corpus All où les résultats sur Google ont été trouvés pour seulement 42,5% des requêtes, alors que plus de 50% des requêtes ont trouvé des résultats dans toutes les autres expériences (le maximum est de 66,6% dans $NegOnly$). Les tweets pour lesquels aucun résultat n'est trouvé sont des tweets avec des entités nommées, mais qui ne parlent pas d'un événement (*e.g. AHAHAHAHAHA ! Pas de respect #Legorafi*, où 'Legorafi ' est un journal satirique).

4.5.3 Évaluation de la méthode à base de requêtes

Pour évaluer la difficulté de la tâche de classification, deux annotateurs ont également été invités à étiqueter comme ironiques ou non ironiques les 50 tweets (40 dans All et 18 dans $NegOnly$) pour lesquels la méthode des requêtes est appliquée. Le score inter-annotateur en terme de Kappa de Cohen est seulement $\kappa = 0,41$. Parmi les 12 tweets corrigés en ironique, les deux annotateurs ne sont pas d'accord entre eux pour 5 tweets. Même si cette expérience n'est pas assez poussée pour mener à une conclusion formelle en raison du petit nombre de tweets annotés, cela tend à montrer que les êtres humains ne feraient pas mieux pour les tweets où le classifieur se trompe.

Il est intéressant de noter que, même si les traits du contexte interne n'étaient pas pertinents pour la classification automatique des tweets (voir le modèle PragSystem), nos résultats montrent qu'ils sont tout de même utiles pour l'amélioration de la classification. Comme le montre l'expérience ①, la méthode à base de requêtes est plus efficace lorsqu'elle est appliquée sur des tweets mal classés. Nous pouvons alors considérer que l'utilisation de traits contextuels internes (présence de pronoms personnels et entités nommées) peut être un moyen de détecter automatiquement les tweets susceptibles d'être mal classés.

4.6 Conclusion

Dans ce chapitre, nous avons proposé une approche pour la détection automatique de l'ironie dans les tweets en français qui a pour but de valider deux hypothèses principales :

- *Hypothèse (H1)* : la présence de négations, en tant que propriété interne d'un énoncé, peut aider à détecter la disparité entre le sens littéral et le sens voulu d'un énoncé.
- *Hypothèse (H2)* : un tweet contenant un fait affirmé de la forme $Not(P)$ est ironique si et seulement si on peut prouver P sur la base de certaines connaissances communes externes à l'énoncé et partagées par l'auteur et le lecteur.

Pour tester la validité des hypothèses, notre corpus FrIC^{Auto} a été divisé en trois parties, selon que les tweets contiennent des négations (*NegOnly*), ne contiennent pas de négation (*NoNeg*) ou contiennent ou non des négations (*All*). Trois modèles ont été proposés pour détecter l'ironie dans ces trois corpus :

1. **SurfSystem**, un modèle sur la base de traits surfaciques traditionnellement utilisés dans l'état de l'art. Nos résultats ont montré que ces traits déjà utilisés pour cette tâche dans d'autres langues sont aussi efficaces pour le français (par exemple la présence de ponctuation ou de mots en majuscule) et qu'ils sont plus performants sur le corpus *NegOnly*.
2. **PragSystem**, un modèle qui utilise les traits pragmatiques extraits du contenu linguistique du tweet. Nous avons utilisé des traits de l'état de l'art et proposé 3 nouveaux traits : les modifieurs de sentiment, les traits de contexte et d'opposition. Ces derniers, obtenus grâce à des patrons d'opposition, ont été les plus productifs. Les meilleurs résultats sont obtenus sur le corpus All (exactitude de $87,7\%$) et l'analyse d'erreurs a montré que les tweets avec négation méritaient un traitement spécifique.
3. **QuerySystem**, une méthode à base de requêtes qui s'applique sur les tweets contenant des négations et qui permet de vérifier la véracité des propositions de la forme $Not(P)$ dans des sources fiables du Web. L'idée est de corriger les prédictions du classifieur *PragSystem* pour les tweets contenant des négations classés comme non ironiques. Les expériences ont montré que cette méthode permet d'améliorer la classification quand elle est appliquée sur des tweets non personnels.

Les résultats obtenus lors des évaluations de ces trois modèles ont donc permis de valider les hypothèse (H1) et (H2).

Dans le chapitre suivant, nous étudions la portabilité de nos modèles conceptuel et computationnel pour la détection de l'ironie dans un cadre multi-lingue. Nous vérifions d'abord la portabilité de notre schéma d'annotation sur des tweets en italien et en anglais, deux langues indo-européennes puis testons notre modèle computationnel sur l'arabe.

Chapitre 5

Vers un système multi-lingue pour la détection automatique de l'ironie

5.1 Introduction

Dans les deux chapitres précédents, nous avons proposé un schéma d'annotation multi-niveaux pour l'ironie ainsi qu'un système de détection automatique de l'ironie dans les tweets en français. Ce présent chapitre a pour but de montrer la portabilité à la fois du schéma d'annotation et du système de détection automatique à d'autres langues. Deux expériences ont ainsi été réalisées.

La première consiste à tester la portabilité du schéma d'annotation multi-niveaux pour l'ironie détaillé dans le chapitre 3 sur d'autres langues de la même famille que le français à savoir l'anglais et l'italien. Cette expérience consiste en une campagne d'annotation dans laquelle le même schéma d'annotation, initialement conçu pour analyser les phénomènes pragmatiques de l'ironie sur des tweets en français, a été utilisé. Cette première expérimentation a comme objectif non seulement de tester la performance du schéma d'annotation sur des langues indo-européennes culturellement proches du français mais aussi de mesurer l'impact d'un ensemble de phénomènes pragmatiques sur l'interprétation de l'ironie et d'étudier comment ces phénomènes interagissent avec le contexte local d'un tweet dans des langues de la même famille. Ce travail a été réalisé en collaboration avec Viviana Patti et Cristina Bosco, maîtres de conférences à l'université de Turin et Véronique Moriceau, maître de conférences à l'université Paris-Sud.

La deuxième expérience consiste à tester la performance du système de détection automatique de l'ironie à base de traits (cf. les modèles SurfSystem et PragSystem présentés dans le chapitre 4) sur des tweets écrits en arabe. Pour cette expérience, nous avons construit le premier corpus de tweets ironiques et non ironiques en arabe afin d'étudier la performance

des traits ainsi que les algorithmes utilisés pour la tâche de classification des tweets en ironique/non ironique.

Dans ce qui suit, nous détaillons les deux expériences proposées ainsi que les résultats obtenus. Nous commençons dans la section 5.2 par la présentation de la première expérience en décrivant les corpus utilisés pour l'anglais et l'italien ainsi que les résultats quantitatifs pour chaque niveau du schéma d'annotation et pour chaque langue. Dans la section 5.3, nous présentons la deuxième expérience en donnant un aperçu des spécificités de la langue arabe ainsi que le corpus de tweets utilisé. Ensuite, nous détaillons les résultats quantitatifs obtenus ainsi qu'une étude comparative avec les résultats obtenus pour le français présentés dans le chapitre 4.

5.2 L'ironie dans les langues indo-européennes

> « *En linguistique, les langues indo-européennes (appelées autrefois langues indo-germaniques, ou bien encore langues scythes) forment une famille de langues étroitement apparentées, ayant pour origine ce qu'il est consensuellement convenu d'appeler l'indo-européen commun* « ***dont les éléments lexicologiques, morphologiques et syntaxiques présentent, pour la plupart d'entre elles, des ressemblances de nature telle que ces langues peuvent se ramener à l'unité ; le présupposé est alors que chaque groupe d'éléments comparés procède d'évolutions divergentes à partir de formes originelles disparues*** *». Au nombre d'environ un millier, elles sont actuellement parlées par près de trois milliards de locuteur.* » (Wikipedia [1])

Partant de cette définition des langues indo-européennes, nous remarquons que les linguistes ont mis l'accent sur les ressemblances morphologiques et syntaxiques entre la plupart des langues indo-européennes. Ceci nous a encouragé à étudier le phénomène de l'ironie dans différentes langues appartenant à la famille des langues indo-européennes. Nous nous focalisons ici sur l'anglais et l'italien.

5.2.1 Corpus

A Collecte du corpus anglais

Nous avons décidé de construire notre propre corpus de tweets ironiques et non ironiques en anglais malgré l'existence de ce type de corpus (Reyes *et al.*, 2013). En effet, ces derniers sont essentiellement constitués de tweets personnels (*e.g. Don't worry about what people*

[1] https://fr.wikipedia.org/wiki/Langues_indo-europ%C3%A9ennes

think. They don't do it very often) et cela ne nous permet pas de nous placer dans une situation comparable à celle de l'étude pour le français. Nous avons donc construit notre propre corpus en suivant la même procédure de collecte exploitée pour la construction du corpus français *FrIC*. Nous avons sélectionné un ensemble de thèmes appartenant aux même catégories utilisées pour le français tout en respectant les actualités relatives à la langue, par exemple : pour la catégorie **politique**, nous avons sélectionné *Obama, Trump, Clinton,* etc., pour la catégorie **artistes**, nous avons sélectionné de nouveaux mots-clés comme *Justin Bieber, kardashian, Beyoncé*. Nous avons ensuite sélectionné les tweets ironiques contenant ces mots-clés et le hashtag #*ironic* ou #*sarcasm*. De la même façon, nous avons sélectionné des tweets non ironiques (*i.e.* ne contenant pas #*ironic* ou #*sarcasm*).

Une fois les tweets collectés, nous avons supprimé les doublons, les retweets et les tweets contenant des images. Après cette étape de filtrage, nous avons obtenu un corpus de 11 289 tweets répartis comme suit : 5 173 tweets ironiques et 6 116 tweets non ironiques. La répartition des tweets entre les catégories est détaillée dans le tableau 5.1.

Thèmes	*Ironique*	*Non ironique*
Économie	117	79
Générique	311	873
Villes ou Pays	1 014	891
Artistes	472	836
Politique	2 560	2 294
Santé	142	160
Sport	557	983
Total	**5 173**	**6 116**

TABLE 5.1 : Répartition des tweets anglais.

B Collecte du corpus italien

Pour l'italien, les tweets proviennent de deux ressources existantes annotées dans le cadre du projet Senti-TUT[2] :

- le corpus Sentipolc, utilisé pour la campagne d'évaluation *Evalita 2014*[3] sur l'analyse de sentiment et la détection de l'ironie sur Twitter (Basile *et al.*, 2014). Le corpus Sentipolc est une collection de tweets en italien dérivée de deux corpus existants : Senti-TUT (Bosco *et al.*, 2013) et TWITA (Basile & Nissim, 2013). Il contient des

[2] www.di.unito.it/~tutreeb/sentiTUT.html
[3] http://di.unito.it/sentipolc14

tweets utilisant des mots-clés et hashtags spécifiques au thème de la politique (noms de personnalités politiques, etc.). Dans Sentipolc, chaque tweet est annoté selon 5 catégories mutuellement exclusives : opinion positive, opinion négative, opinion à la fois positive et négative, ironie, et objectif.

- le corpus TW-SPINO qui contient des tweets provenant de Spinoza[4], un blog italien satirique sur la politique. Ces tweets sont sélectionnés et révisés par une équipe éditoriale qui les identifie comme ironiques ou satiriques.

Le corpus italien est donc constitué de 3 079 tweets ironiques (806 de Sentipolc et 2 273 de TW-SPINO) et 5 642 tweets non ironiques (de Sentipolc).

5.2.2 Résultats de la procédure d'annotation

Pour l'étude de la portabilité de notre schéma d'annotation, nous nous sommes concentrés sur l'annotation d'un sous-ensemble de tweets en anglais et en italien. Ceci est une première étape qui a pour principal but de tester la performance de notre schéma sur d'autres langues et de comparer les résultats statistiques obtenus pour les trois langues.

Nous présentons dans cette section les résultats quantitatifs obtenus sur les deux corpus annotés (anglais et italien). Pour chaque langue, deux annotateurs humains ont assuré la procédure d'annotation qui a été effectuée en deux étapes. Une première étape d'entrainement pendant laquelle 100 tweets (50 tweets ironiques et 50 tweets non ironiques) pour chaque langue ont été utilisés, puis une deuxième étape d'annotation pendant laquelle 550 tweets en anglais et 500 tweets en italien ont été doublement annotés (80% ironiques et 20% non ironiques). La première étape est nécessaire afin que les annotateurs se familiarisent avec le schéma d'annotation ainsi que le type de corpus.

Pour rappel, notre schéma d'annotation propose une annotation à quatre niveaux (cf. figure 5.1) : le niveau 1 identifie les tweets ironiques et non ironiques ; le niveau 2 les contradictions explicites ou implicites pour les tweets ironiques ; le niveau 3 la catégorie d'ironie (analogie, hyperbole, etc.) et le niveau 4 les marqueurs (négation, ponctuation, etc.).

Pour la procédure d'annotation du corpus de tweets anglais, les annotateurs ont utilisé le même outil déjà utilisé pour l'annotation du corpus français FrIC, à savoir l'outil *Glozz*, et annotés les tweets selon les 4 niveaux du schéma d'annotation. Nous avons fourni aux deux annotateurs, une version traduite en anglais du guide d'annotation ainsi que les fichiers nécessaires pour Glozz.

Pour le corpus italien, le niveau 1 (ironique/ non ironique) étant déjà annoté, les annotateurs ont annoté manuellement uniquement les niveaux 2 et 3 du schéma d'annotation

[4] http://www.spinoza.it/

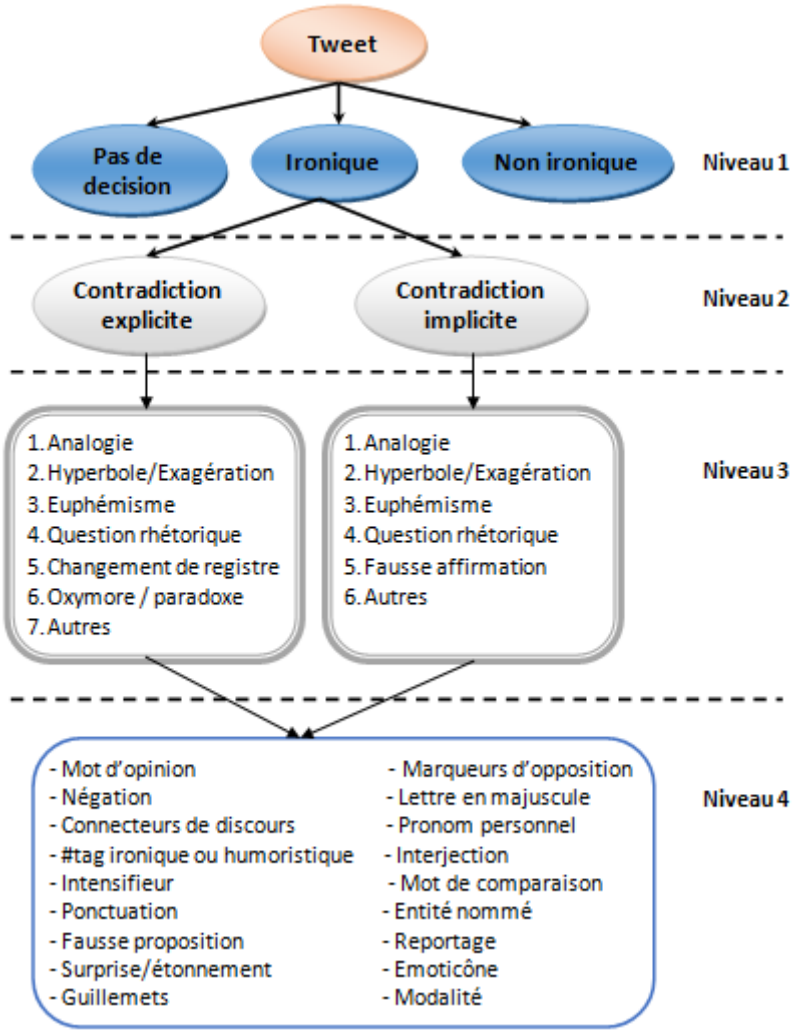

FIGURE 5.1 : Schéma d'annotation.

(types et catégories de l'ironie) en respectant le guide d'annotation utilisé pour le français et l'anglais [5]. En revanche, les indices (niveau 4 du schéma d'annotation) ont été annotés automatiquement mais seuls certains indices ont été vérifiés manuellement (négations, émoticones). Nous ne donnons donc ici que les résultats pour les indices qui ont été vérifiés manuellement.

Cette nouvelle campagne d'annotation a permis d'effectuer les analyses suivantes :

- présence des indices dans les tweets ironiques,

[5] https://github.com/IronyAndTweets/Scheme

- variation de la présence des indices pour chaque type d'ironie (ironique avec contradiction explicite ou implicite) et chaque catégorie d'ironie (hyperbole, analogie, paradoxe, etc.),
- fréquence des catégories de l'ironie pour chaque type d'ironie,

A Résultats quantitatifs de la procédure d'annotation en ironique/non ironique

En se fiant aux hashtags de référence *#ironic* et *#sarcasm* dans le corpus anglais, nous avons 440 (80%) tweets ironiques et 110 (20%) tweets non ironiques alors que les annotateurs humains ont jugé 427 (77,63%) tweets comme ironiques, 99 tweets non ironiques (18%) et 24 tweets (4,37%) ont été classés dans la classe « Pas de décision » (figure 5.2). Ces résultats prouvent que, indépendamment de la langue, un tweet accompagné d'un hashtag marquant l'ironie n'est pas forcément ironique et qu'un tweet ne contenant pas l'un de ces hashtags peut être ironique.

Contrairement, aux deux corpus français et anglais, la répartition des tweets en ironiques/non ironiques dans le corpus italien a été effectuée par des humains et non par les hashtags (cette annotation a été faite préalablement dans le cadre du projet *Senti-TUT*). Nous avons 400 tweets ironiques, 100 tweets non ironiques et aucun tweet dans la catégorie « Pas de décision » (figure 5.2).

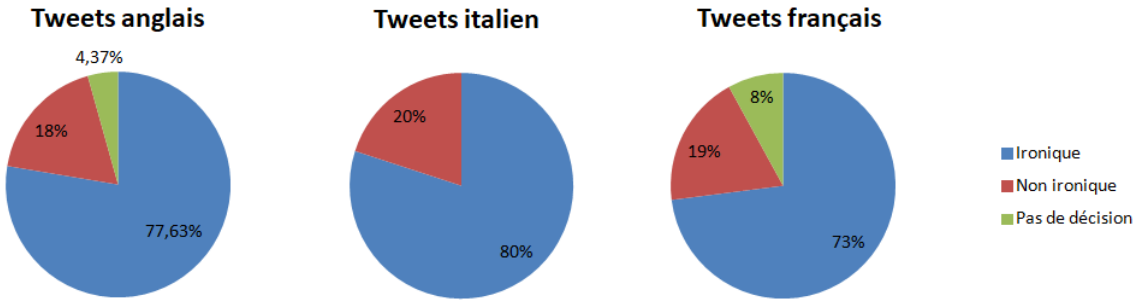

FIGURE 5.2 : Répartition des tweets anglais, italien et français.

B Résultats quantitatifs de la procédure d'annotation sur les types de l'ironie

Le tableau 5.2 montre le nombre total de tweets annotés et le type d'activation de l'ironie dans chaque corpus.

Pour le corpus anglais, parmi les 427 tweets annotés comme ironiques, nous avons 283 tweets ironiques avec contradiction implicite et 144 tweets ironiques avec contradiction

explicite. Ceci prouve que l'ironie est un phénomène qui s'exprime généralement d'une manière implicite pour l'anglais (66,28%), comme c'était le cas pour le français (73,01%).

Par contre, pour le corpus italien, l'ironie se manifeste par des contradictions explicites dans la plupart des cas (65%). Ceci peut s'expliquer par le fait que les utilisateurs italiens n'utilisent pas de hashtags spécifiques pour marquer l'ironie et sont donc peut-être plus explicites.

	Ironique		**Non Ironique**	**Pas de décision**	*Total*
	explicite	implicite			
Français	394 (19,7%)	1 066 (53,3%)	380 (19%)	160 (8%)	2 000
Anglais	144 (26,2%)	283 (51,45%)	99 (18%)	24 (4,35%)	550
Italien	260 (52%)	140 (28%)	100 (20%)	–	500

TABLE 5.2 : Nombre de tweets annotés dans les corpus français, anglais et italien.

C Résultats quantitatifs de la procédure d'annotation sur les catégories de l'ironie

Le tableau 5.3 présente le pourcentage de tweets appartenant à chaque catégorie d'ironie selon l'activation explicite/implicite de l'ironie [6]. Pour les trois corpus français, anglais et italien, nous observons une différence importante dans la présence de catégories. Les résultats montrent que :

- Pour l'ironie avec contradiction explicite : la catégorie *Oxymore/paradoxe* est la catégorie la plus fréquente pour les trois langues : français, anglais et italien,
- Pour l'ironie avec contradiction implicite : la *fausse affirmation* et *autres* sont les catégories les plus fréquentes en français et en anglais. Pour l'italien, les catégories *fausse affirmation, analogie* et *autres* sont les catégories les plus fréquentes.
- En observant les tweets annotés sous la catégorie *autres*, nous constatons que la plupart sont ironiques avec contradiction implicite. Cela prouve que la tâche de décision est plus difficile pour les humains dans le cas où l'ironie est exprimée avec une contradiction implicite indépendamment de la langue.

Comme les classes ne sont pas mutuellement exclusives, nous avons :

[6] Les fréquences les plus élevées sont en caractères gras.

- Pour le corpus anglais, 35 tweets avec contradiction explicite appartenant à plus d'une catégorie et 62 tweets avec contradiction implicites appartenant à plus d'une catégorie. La combinaison la plus fréquente pour les contradictions explicites est *Oxymore/Paradoxe* + *Question rhétorique* tandis que pour les ironiques avec contradiction implicite la combinaison la plus fréquente est *Métaphore/comparaison* + *Autres*.
- Pour le corpus français *FrIC*, la combinaison la plus fréquente pour les contradictions explicites est *Oxymore/Paradoxe* + *Hyperbole/Exagération* alors que la combinaison la plus fréquente pour les contradictions implicites est *Fausse affirmation* + *Hyperbole/Exagération*.
- Pour le corpus italien, les annotateurs ont choisi de n'attribuer qu'une seule catégorie, celle qui leur paraissait la plus porteuse d'ironie.

Ainsi, pour le français et l'anglais, la catégorie *Oxymore/Paradoxe* fait partie des combinaisons les plus fréquentes pour les ironiques avec contradiction explicite. Pour les ironiques avec contradiction implicite, ce sont des combinaisons différentes.

	Analogie			**Changement de registre**			**Euphémisme**			**Hyperbole**		
	F	A	I	F	A	I	F	A	I	F	A	I
Explicite	12%	17%	21%	1%	6%	19%	1%	1%	5%	8%	2%	9%
Implicite	2%	13%	**26%**	-	-	-	1%	1%	4%	10%	7%	5%

	Question rhétorique			**Oxymore/ paradoxe**			**Fausse affirmation**			**Autres**		
	F	A	I	F	A	I	F	A	I	F	A	I
Explicite	10%	15%	10%	**66%**	**81%**	**28%**	-	-	-	21%	6%	7%
Implicite	14%	1%	12%	-	-	-	**56%**	20%	**34%**	**32%**	**65%**	**19%**

TABLE 5.3 : Répartition des catégories selon les activations explicite ou implicite dans les corpus français (F), anglais (A) et italien (I).

D Résultats quantitatifs de la procédure d'annotation pour les indices de l'ironie

A ce niveau, nous avons élaboré trois études statistiques. La première étude est une étude quantitative entre le premier et le quatrième niveau du schéma d'annotation dans laquelle nous avons étudié la présence des différents indices dans les tweets ironiques et non ironiques (cf. Tableau 5.4). La deuxième étude est consacrée aux deuxième et quatrième niveaux dans laquelle nous avons étudié la présence des différents indices dans les tweets ironiques avec contradiction explicite ou implicite (cf. Tableau 5.4). Enfin, nous avons consacré

la troisième étude aux troisième et quatrième niveaux : nous avons étudié la présence des indices dans chaque catégorie d'ironie (cf. Tableau 5.5).

Le tableau 5.4 indique le pourcentage de tweets contenant des indices pour les tweets ironiques (explicites ou implicites) et non ironiques (NIR, lignes en gris).

En français, les indices *intensifieurs, ponctuation* et *interjections* sont plus fréquents dans les tweets ironiques alors que *les citations* sont plus fréquentes dans les tweets non ironiques.

En anglais, les indices *connecteurs de discours, citations, mots de comparaison* et *verbes rapporteurs* sont deux fois plus fréquents dans les tweets ironiques que dans les tweets non ironiques alors que c'est le contraire pour les *pronoms personnels*. Nous précisons que pour l'anglais, le corpus ne contient pas de tweets ironiques contenant des URL car ceux-ci ont tous été annotés comme « Pas de décision » en raison d'un manque de connaissances des annotateurs qui n'ont pas réussi à comprendre le tweet et le contenu de la page Web pointée par l'URL.

En italien, la plupart des indices sont plus fréquents dans les tweets ironiques, tandis que certains, comme les *citations* et les URL sont plus fréquents dans les tweets non ironiques.

	Emoticône			**Négation**			**Connecteurs de discours**			**# Humoris-* tique**			**Intensifieur**			**Ponctua-tion**		
	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I
Explicite	7	2	1	37	58	15	6	41	29	2	14	-	22	9	2	51	30	14
Implicite	6	4	7	34	61	9	4	29	16	4	15	-	19	12	0	51	28	5
NIR	5	10	0	58	75	9	4	13	18	0	0	-	11	9	0	28	30	17

	Fausse* proposition			**Surprise**			**Modalité**			**Citation**			**Opposition**			**Majuscule**		
	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I
Explicite	8	0	-	3	0	-	0	2	3	6	21	3	9	18	4	3	8	-
Implicite	54	18	-	3	3	-	0	2	6	6	21	6	3	11	6	2	6	-
NIR	0	0	-	2	0	-	1	6	3	1	10	26	4	14	4	3	3	-

	Pronom* personnel			**Interjection**			**Compa-* raison**			**Entités* Nommées**			**Verbe de reportage**			**Opinion**			**URL***		
	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I
Explicite	31	21	5	14	2	11	8	8	4	97	100	65	1	17	0	48	75	-	33	0	10
Implicite	31	24	3	12	0	13	2	12	3	91	97	43	1	14	0	41	74	-	29	0	2
NIR	30	40	1	2	2	12	4	6	1	82	88	98	3	7	1	35	68	-	42	0	44

TABLE 5.4 : Répartition des indices dans les tweets ironiques (explicites ou implicites) et les tweets non ironiques (NIR) en français (F), anglais (A) et italien (I) en terme de pourcentage. Les indices marqués par * n'ont pas été étudiés dans la littérature.

Le tableau 5.5 indique le pourcentage de tweets contenant des indices dans chaque catégorie d'ironie.

La *négation* est plus fréquente dans la catégorie *euphémisme* pour le français et dans les catégories *changement de registre, euphémisme et question rhétorique* pour l'anglais.

	Négation			Connecteurs de discours			# Humoris-* tique			Intensifieur			Ponctua-tion			Fausse* proposition		
	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I
Analogie	46	56	2	6	29	8	6	15	-	21	10	0	49	24	2	13	8	-
Changement de Registre	40	**100**	3	0	11	3	0	0	-	0	0	1	**60**	**44**	1	0	0	-
Euphémisme	**50**	**67**	1	6	0	2	0	0	-	**50**	**33**	0	**72**	0	1	44	0	-
Hyperbole	25	42	1	5	25	2	3	8	-	**57**	**38**	0	56	21	2	**53**	**46**	-
Question rhétorique	43	**70**	2	2	36	3	2	17	-	17	9	0	**93**	**86**	1	9	3	-
Oxymore/paradoxe	35	59	3	4	**43**	6	0	14	-	21	10	1	49	26	2	11	0	-
Fausse affirmation	18	57	1	4	25	3	3	7	-	10	16	0	29	14	2	**95**	**89**	-
Autre	26	62	2	5	31	3	5	18	-	15	11	0	45	20	2	11	3	-

	Modalité			Citation			Opposition			Pronom* personnels			Interjection			Compa-* raison		
	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I
Analogie	0	3	2	0	24	1	6	11	2	38	19	2	6	0	3	**43**	**42**	3
Changement de registre	0	11	0	0	**44**	0	0	11	1	**40**	**33**	1	20	0	2	20	6	0
Euphémisme	0	**33**	0	0	0	1	0	0	0	22	0	0	6	0	1	0	0	0
Hyperbole	0	0	0	8	4	0	2	4	0	29	**33**	1	18	0	2	0	8	0
Question rhétorique	0	3	0	7	23	1	3	15	1	31	27	0	13	2	1	2	5	0
Oxymore/paradoxe	0	2	1	5	20	0	**12**	**19**	1	32	21	0	15	3	2	2	6	0
Fausse affirmation	0	0	0	4	16	1	3	4	1	31	**36**	1	13	0	1	2	13	1
Autre	0	2	1	8	25	1	2	11	2	29	22	0	10	0	2	1	10	0

	Entités* nommées			Verbe de reportage			Opinion			URL*		
	F	A	I	F	A	I	F	A	I	F	A	I
Analogie	100	100	17	2	16	0	41	68	-	13	0	1
Changement de registre	80	100	8	0	22	0	60	68	-	0	0	1
Euphémisme	94	100	2	0	**33**	0	56	67	-	22	0	1
Hyperbole	88	88	6	3	13	0	**84**	**88**	-	21	0	1
Question rhétorique	90	97	9	1	17	0	45	73	-	25	0	1
Oxymore/paradoxe	99	100	10	1	19	0	55	75	-	11	0	2
Fausse affirmation	90	93	8	1	13	0	45	79	-	25	0	0
Autre	91	98	6	1	16	0	32	74	-	30	0	1

TABLE 5.5 : Répartition des tweets dans chaque catégorie d'ironie contenant des indices en terme de pourcentage pour le français (F), anglais (A) et italien (I).

Les *intensifieurs* sont plus fréquents dans les catégories *euphémisme* et *hyperbole* pour le français et l'anglais.

Les *ponctuations* sont plus fréquentes dans les catégories *changement de registre, euphémisme* et *question rhétorique* pour le français et les catégories *changement de registre et question rhétorique* pour l'anglais.

Les *fausses propositions* sont plus fréquentes dans les catégories *hyperbole* et *fausse affirmation* pour le français ainsi que pour l'anglais.

Les *mots d'opposition* sont plus fréquents dans la catégorie *oxymore/paradoxe* pour le français et l'anglais.

Les *pronoms personnels* sont plus fréquents dans la catégorie *changement de registre* pour le français et dans les catégories *changement de registre, hyperbole* et *fausse affirmation* pour l'anglais.

Les *mots de comparaison* sont plus fréquents dans la catégorie *analogie* pour le français et l'anglais.

Les *mots d'opinion* sont plus fréquents dans la catégorie *hyperbole* pour le français ainsi que pour l'anglais.

Pour l'italien, les pourcentages de tweets contenant des indices dans chaque catégorie d'ironie sont très faibles et les tweets sont répartis d'une manière presque équitable entre les différentes catégories (par exemple la négation est présente dans toutes les catégories avec un pourcentage de 1% à 3%, les mots d'oppositions sont présents avec des pourcentages qui ne dépassent pas 2%).

L'ensemble des études quantitatives a montré qu'indépendamment de la langue, les auteurs des tweets ironiques ont tendance à utiliser des indices tels que les mots d'opinion, les entités nommées, les mots de négation dans les catégories *analogie, question rhétorique, oxymore/paradoxe, fausse affirmation* et autres.

E Résultats quantitatifs de la procédure d'annotation des relations de contradiction

La figure 5.3 montre que la répartition de chaque relation est presque la même dans les corpus français et anglais. Par exemple, la relation *opposition* est la plus utilisée dans les tweets ironiques en français (69%) et en anglais (80%). En outre, le nombre de relations de *comparaison* est moins important pour l'anglais (20%) ainsi que pour le français (14%) et les relations *cause/conséquence* sont les moins utilisées pour l'anglais (6%) et pour le français (11%). Pour l'italien, l'annotation des indices a été effectué automatiquement. Par conséquent, les annotateurs n'ont pas annoté les relations de contradiction qui nécessitent une annotation manuelle.

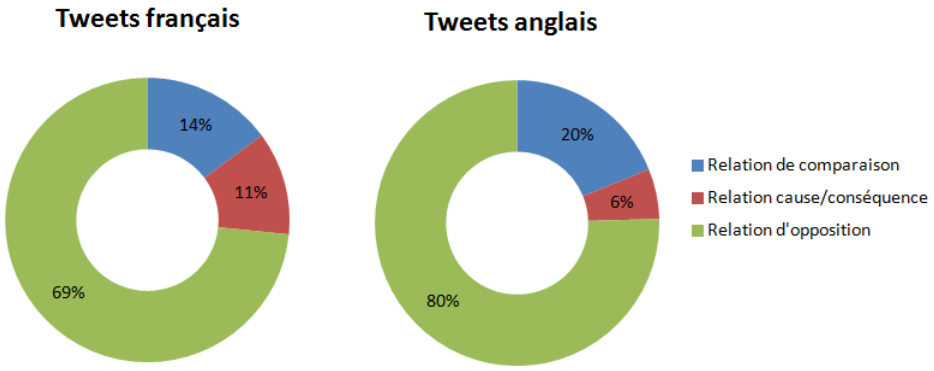

FIGURE 5.3 : Répartition des relations dans les tweets ironiques avec contradiction explicite en terme de pourcentage, pour le français et l'anglais.

F Corrélation entre les différents niveaux du schéma d'annotation

Nous présentons ici une étude comparative des corrélations entre les différents niveaux du schéma pour les trois langues : français, anglais et italien. La même démarche que pour le français (voir Chapitre 3) a été suivie pour l'anglais et l'italien .

Nous avons donc étudié la corrélation entre les indices de l'ironie et les types d'activation de l'ironie (contradiction explicite ou implicite) ainsi qu'entre les indices de l'ironie et les catégories de l'ironie. Notre but est d'analyser dans quelle mesure ces indices peuvent être des indicateurs pour la prédiction de l'ironie dans ces langues.

En appliquant le test *V de Cramer* (Cohen, 1988) sur le nombre d'occurrences de chaque indice, nous obtenons (toutes les corrélations sont statistiquement significatives) :

- Entre les indices et la classe ironique/non ironique :
  - une corrélation forte pour le français $(V = 0,156, df = 14)$ et l'italien $(V = 0,31, df = 6)$,
  - une corrélation entre moyenne et forte pour l'anglais $(V = 0,132, df = 9)$.
- Entre les indices et les types d'activation de l'ironie (contradiction explicite ou implicite) :
  - une corrélation forte pour le français $(V = 0,196, df = 16)$,
  - une corrélation entre moyenne et forte pour l'italien $(V = 0,138, df = 5)$,
  - une corrélation moyenne pour l'anglais $(V = 0,083, df = 12)$.

Nous avons également analysé les corrélations par indice $(df = 1)$. Les indices les plus corrélés à la classe *ironique/non ironique* sont les suivants :

- *négations, interjections, entités nommées* et *URL* pour le français $(0,14 < V < 0,41)$;
- *négations, connecteurs de discours* et *pronoms personnels* pour l'anglais $(0,12 < V < 0,17)$;
- *citations, entités nommées* et *URL* pour l'italien $(0,310 < V < 0,416)$.

Les indices les plus corrélés à l'activation *explicite/implicite* sont :

- *marqueurs d'opposition, mots de comparaison* et *fausse affirmation* pour le français $(0,140 < V < 0,190)$;

- *marqueurs d'opposition* et *connecteurs de discours* pour l'anglais $(0,110 < V < 0,120)$;

- *connecteurs de discours, ponctuation* et *entités nommées* pour l'italien $(0,136 < V < 0,213)$.

Nous avons remarqué que malgré la fréquence importante des mots d'opinion dans les tweets ironiques pour le français et l'anglais, l'indice *opinion* n'est pas corrélé à la classification ironique/non ironique ou à l'activation explicite/implicite $(V < 0,06)$, car de nombreux tweets non ironiques contiennent également des mots d'opinion.

Enfin, nous avons analysé la corrélation entre les indices et les catégories d'ironie. Les indices les plus discriminants selon le test *V de Cramer* sont :

- *intensifieur, ponctuation, fausse affirmation* et *mot d'opinion* pour le français avec une corrélation forte ;

- *négation, connecteur de discours* et *pronom personnel* pour l'anglais avec une corrélation moyenne ;

- *ponctuation, interjection* et *entité nommée* pour l'italien avec une corrélation moyenne.

5.2.3 Synthèse

L'ensemble des résultats obtenus sont très encourageants car ils montrent que notre schéma d'annotation défini pour le français peut être appliqué sur d'autres langues indo-européennes (anglais et italien). Les phénomènes pragmatiques propres à l'ironie, initialement identifiés dans les tweets en français, sont également présents lorsque l'ironie est exprimée dans des langues de la même famille. Nous observons ainsi les mêmes tendances en termes de catégories d'ironie et de marqueurs ; ainsi que des corrélations entre les marqueurs et la classe ironique/non ironique et entre les marqueurs et le type d'activation (explicite ou implicite).

Ce travail d'évaluation de la portabilité du schéma d'annotation permet d'envisager dans le futur le développement d'un système de détection automatique de l'ironie dans un cadre multi-lingue.

Nous avons testé une première approche vers un tel système en évaluant notre modèle de détection automatique sur l'arabe. Nous présentons cette expérimentation dans les sections suivantes.

5.3 L'ironie dans les langues sémitiques

> « *Les langues sémitiques sont un groupe de langues parlées dès l'antiquité au Proche-Orient, en Afrique du Nord et dans la corne de l'Afrique. Ces langues sont qualifiées de « sémitiques » depuis 1781, d'après de nom biblique de Sem, fils de Noé. Elles forment une des branches de la famille des langues chamito-sémitique (dites aussi afro-asiatiques ou afrasiennes), répandues de la moitié nord de l'Afrique jusqu'au Moyen-Orient. L'origine et la direction de l'expansion géographique de ces langues restent incertaines, de l'Asie vers l'Afrique ou de l'Afrique vers l'Asie. Plusieurs langues sémitiques sont des langues officielles ou administratives : arabe (environ 375 millions de locuteurs), amharique (plus de 90 millions), hébreu (8 millions), tigrigna (6,75 millions) et maltais (400 000 locuteurs). D'autres langues sémitiues sont utilisées en Éthiopie, en Érythrée, à Djibouti et en Somalie, et au Proche-Orient (langue néo-araméennes par exemple). L'arabe est marqué par une importante diglossie entre l'arabe littéral, langue véhiculaire surtout écrite, et l'arabe dialectal, langue vernaculaire surtout orale. L'arabe littéral comprend l'arabe classique et l'arabe standard moderne. L'arabe dialectal comprend de nombreuses variétés régionales, pas toutes intelligibles entre elles.* » (Wikipedia [7])

Parmi les langues appartenant à la famille des langues sémitiques, nous nous sommes intéressés dans ce chapitre à la langue arabe. Selon la définition précedente, nous remarquons que les linguistes ont mis l'accent sur la grande différence entre l'arabe littéral et l'arabe dialectal. Ces deux types de la langue arabe ont fait l'objet de beaucoup de travaux de traitement informatique dont la plupart traite l'arabe littéral et plus particulièrement l'arabe standard moderne.

L'arabe standard moderne (ASM) est une version modernisée et standardisée de l'arabe classique[8] utilisé dans l'écriture et les discours formels dans le domaine de l'éducation, les journaux et quelques programmes à la télévision. L'ASM possède une structure linguistique complexe avec une morphologie riche et une syntaxe complexe (Al-Sughaiyer & Al-Kharashi, 2004), (Ryding, 2005). Le traitement automatique de la langue arabe a fait l'objet d'une grande attention dans la littérature depuis plus de vingt ans (Habash, 2010). Plusieurs ressources et outils ont été construits pour traiter la morphologie et la syntaxe arabe allant de l'analyse superficielle à l'analyse profonde (Eskander *et al.*, 2013; Pasha *et al.*, 2014; Green & Manning, 2010; Marton *et al.*, 2013). Il existe également un nombre important d'applications pour le TAL arabe (ANLP) à savoir : les systèmes de questions-réponses (Bdour & Gharaibeh, 2013; Hammo *et al.*, 2002; Abouenour *et al.*, 2012), la traduction automatique (Sadat & Mohamed, 2013; Carpuat *et al.*, 2012), l'analyse des sentiments (Abdul-Mageed

[7] https://fr.wikipedia.org/wiki/Langues_s%C3%A9mitiques

[8] L'arabe classique ou arabe coranique est la langue utilisée dans les textes littéraires et religieux.

et al., 2014) et la reconnaissance des entités nommées(Darwish, 2013; Oudah & Shaalan, 2012). Quelques travaux se sont également intéressés au traitement automatique de l'arabe dialectal à savoir : la compréhension automatique de l'oral arabe spontané (Afify *et al.*, 2006; Biadsy *et al.*, 2009; Bahou *et al.*, 2010), la phonétisation du dialecte tunisien (Masmoudi *et al.*, 2014), la constructions d'ontologies de domaines pour le dialecte arabe (Graja *et al.*, 2011; Karoui *et al.*, 2013), l'analyse morphologique (Habash & Rambow, 2006; Habash & Rambow, 2007), l'identification automatique de l'arabe dialectal (Alorifi, 2008), etc. Cependant, à notre connaissance, aucun travail n'a abordé la problématique de la détection automatique de l'ironie pour la langue arabe et plus particulièrement dans les réseaux sociaux.

Dans ce qui suit, nous présentons un aperçu sur la spécificité de la langue arabe (Section 5.3.1) en distinguant l'arabe standard moderne et l'arabe dialectal. Nous présentons dans la section 5.3.2 le corpus ainsi que les ressources que nous avons utilisés pour la détection automatique. Dans la section 5.3.3, nous détaillons l'expérience réalisée ainsi que les résultats obtenus. Enfin, dans la section C, nous comparons les résultats obtenus avec ceux obtenus pour le français (détaillés dans le chapitre 4).

5.3.1 Les spécificités de la langue arabe

La langue arabe a une écriture de droite à gauche avec des formes ambiguës. En effet, les lettres arabes ont des formes différentes selon la position de la lettre dans le mot. La lettre peut être en position autonome (forme non reliée), initiale (la lettre au début du mot), forme médiane (la lettre au milieu du mot) ou finale (la lettre à la fin du mot) (voir Tableau 5.6).

autonome	م	ش	غ
initiale	مـ	شـ	غـ
médiane	ـمـ	ـشـ	ـغـ
finale	ـم	ـش	ـغ

TABLE 5.6 : Lettres en caractères arabes selon leur position dans le mot (Habash, 2010).

La langue arabe est principalement caractérisée par : l'absence de signes diacritiques (lettres dédiées pour représenter les voyelles courtes), l'agglutination complexe, et la structure de mot d'ordre libre. Ces caractéristiques rendent le traitement arabe plus difficile. Par exemple, Farghaly et al., (2003) estiment que le nombre d'ambiguïtés moyen pour un token en arabe standard peut atteindre 19,2 alors qu'il est de 2,3 dans la plupart des autres langues.

La langue arabe a 28 consonnes qui peuvent être entrelacées avec différentes voyelles

longues et courtes, comme le montre la figure 5.4 [9]. Les voyelles courtes ont été représentées par des signes diacritiques qui sont des marques au-dessus ou au-dessous des lettres, telles que Fatha (une petite diagonale placée au-dessus d'une lettre), Kasrah (une petite diagonale placée sous une lettre) et Dammah (un petit diacritique comme une boucle placée au-dessus d'une lettre). Les textes arabes peuvent être entièrement diacritisés, partiellement diacritisés ou non diacritisés.

Voyelles simples	Avec consonne	Nom	Trans.	Valeur
ـَ	بَ	fatḥa	a	[a]
ـِ	بِ	kasra	i	[i]
ـُ	بُ	ḍamma	u	[u]
ـَا	بَا	fatḥa ʾalif	ā	[aː]
ـَى	بَى	fatḥa ʾalif maqṣūra	ā / aỳ	[aː]
ـِي	بِي	kasra yāʾ	ī / iy	[iː]
ـُو	بُو	ḍamma wāw	ū / uw	[uː]

FIGURE 5.4 : Types de diacritiques arabes (Wikipédia)

.

Les voyelles courtes sont rarement explicitement marquées par écrit. En effet, les diacritiques ne sont ni écrits dans l'écriture arabe de l'usage quotidien, ni dans les publications générales. Les textes dépourvus de diacritiques sont très ambigus. Par exemple, le mot علم peut être diacritisé sous 9 formes différentes (Maamouri *et al.*, 2006) : عِلم (science), عَلَم (drapeau), عَلَّمَ (il a enseigné), etc. Un mot non diacritisé peut avoir des caractéristiques morphologiques différentes et, dans certains cas, une catégorie morpho-syntaxique différente, surtout s'il est interprété hors contexte.

En revanche, pour l'arabe dialectal, les textes non diacritisés ne représentent pas la seule source d'ambiguïté au niveau de la compréhension d'un mot. En effet, un mot peut avoir plusieurs sens selon le dialecte d'un pays particulier et de même un objet peut être désigné par plusieurs noms selon le pays ou la région. Par exemple, une valise est appelée فَاليجه ou فَالِيز en dialecte tunisien mais سفر شنطة en dialecte égyptien. C'est le cas des textes publiés sur les réseaux sociaux, les blogs, les commentaires sur les sites de vente en ligne, etc. On se trouve avec une variété de mots qui peuvent désigner la même chose.

[9] https://fr.wikipedia.org/wikiDiacritiques_de_l'alphabet_arabe

Dans notre travail, nous nous focalisons sur un type particulier de texte non diacritisé qui mélange arabe standard et arabe dialectal : les textes publiés sur les réseaux sociaux et particulièrement les tweets.

5.3.2 Corpus et ressources

Collecte du premier corpus arabe pour l'ironie

Étant donnée l'absence d'un corpus de tweets ironiques en arabe, nous avons suivi la même procédure appliquée pour la collecte des corpus français et anglais. Nous avons d'abord étudié la possibilité d'avoir les mêmes catégories utilisées pour la collecte du corpus français (politique, économie, santé, etc.) mais avec des thèmes spécifiques aux actualités du monde arabe. Malheureusement, nous avons remarqué que la plupart des tweets ironiques sont reliés à des sujets politiques. Par conséquent, nous n'avons collecté que des tweets relatifs à la catégorie *politique* avec un ensemble de six thèmes : هيلَاري (Hillary), ترَامب (Trump), السيسي (Al-sissi : président de l'Égypte), مبَارك (Moubarak : ancien président de l'Égypte) et مرسي (Morsi : ancien président de l'Égypte).

Pour la construction d'un corpus de tweets ironiques/non ironiques, nous avons considéré les tweets contenant les hashtags : #سخرية, #مسخرة, #تهكم , #استهزَاء (traduction de *#ironie* et *#sarcasme*). Le tweets 5.1 illustre un message ironique alors que l'exemple 5.2 ne l'est pas.

Le tweet (5.2) est rédigé en arabe standard. Le tweet (5.1) est rédigé en arabe standard contenant un seul mot en dialecte égyptien/tunisien مش (*non*).

(5.1) فعلَا وَاضح جدَاً أن الَانقلَاب العسكري ضد الرءيس المنتخب مرسي كَان لمصلحة مصر مش أي أطمَاع وَالآن صرَاعَات علي منصب الرآسة #سخرية

(En fait, c'est clair que le coup d'État militaire contre le président élu Morsi était dans l'intérêt de l'Égypte et non pas des ambitions et maintenant des conflits pour la présidence #ironie)

(5.2) تّغير الحيَاة ولَا تّغير النسَاء هكذَا فهمت من دموع هيلَاري كلينتون عند تلقيهَا لّهزيمة أمَام ترَامب

(La vie change mais les femmes ne changent jamais c'est comme ça que j'ai compris les larmes de Hillary Clinton quand elle a reçu la défaite face à Trump)

Une fois les tweets collectés, nous avons supprimé les doublons, les retweets et les tweets contenant des images. Après cette étape de filtrage, nous avons obtenu un corpus de **3 479** tweets répartis comme suit : **1 733** tweets ironiques et **1 746** tweets non ironiques. Le corpus collecté est formé de tweets rédigés en arabe standard et dialectal et dans la plupart du temps le même tweet contient un mélange d'arabe standard et dialectal. Etant donné que l'API de Twitter ne permet pas de distinguer l'arabe standard, l'arabe dialectal et les différents dialectes, nous avons obtenu un corpus de tweets qui mélange beaucoup de dialectes dont les dialectes égyptien, syrien et saoudien. D'autres dialectes ont été rarement utilisés comme le dialecte tunisien et algérien. Pour les expériences décrites par la suite, les hashtags #سخرية, #مسخرة, #تهكم , #استهزاء (*#ironie* et *#sarcasme*) ont été supprimés des tweets.

Ressources linguistiques

Notre approche de détection automatique détaillée dans le chapitre précédent s'appuie sur des lexiques dédiés pour identifier les mots d'opinion, les intensifieurs, négations, émotions, etc. Pour étudier la portabilité de notre système à l'arabe, nous avons recherché des lexiques existants pour la langue arabe (standard et dialectal). Certains d'entre eux se sont révélés performants pour la détection de l'ironie et d'autres non. Dans ce dernier cas, nous avons construit nos propres lexiques. Voici la liste de ressources linguistiques que nous avons utilisées :

- Lexique de connecteurs de discours arabes issu des travaux d'Iskandar Keskes (Keskes *et al.*, 2014). Ce lexique contient 416 connecteurs, comme يضَاف الَى ذلك (*en outre*), لذلك (*donc*), etc.

- Lexique de 4 501 entités nommées (Keskes *et al.*, 2014), auxquelles nous avons ajouté les entités utilisées pour la collecte des tweets, par exemple كلنتون (*Clinton*), etc.

- Lexique de 119 verbes de reportage qui ont été utilisés dans les travaux de Keskes et al. (Keskes *et al.*, 2014), comme قَال (*dire*), أعـلن (*annoncer*), etc.

- Un lexique de mots d'opinions constitués de 22 239 mots d'opinion négative et de 26 777 mots d'opinion positive que nous avons obtenus en fusionnant deux ressources : *Arabic Emoticon Lexicon* et *Arabic Hashtag Lexicon* (dialectal)[10] (Saif *et al.*,

[10]Ressources disponibles à http://saifmohammad.com/WebPages/ArabicSA.html

2016). Ces lexiques ont été exploités dans la tâche 7 de la campagne SemEval'2016[11] [12], par exemple الفشل (*échec*), نقمة (*indignation*).

- Un lexique de 681 émoticônes. Ce dernier est le même que celui déjà utilisé pour le français.
- Un lexique de pronoms personnels ainsi qu'un lexique de mots de négation que nous avons construit manuellement, comme أنَا (*je*), نحن (*nous*), لم ou ليس (*pas/non/n'est pas*), etc.
- Un lexique de 25 intensifieurs traduits d'un ensemble d'intensifieurs utilisés pour le français, par exemple كثير (*beaucoup*), جدًا (*très*), etc.

5.3.3 Détection automatique de l'ironie dans les tweets arabes

Dans cette section, nous présentons l'ensemble des traits utilisés lors de la procédure d'apprentissage, les différents algorithmes utilisés en se focalisant sur l'algorithme le plus performant pour la tâche de classification, et enfin, nous présentons les résultats obtenus.

A Traits utilisés pour la détection de l'ironie

Pour le français, notre système utilise 30 traits dont 8 sont obtenus par l'utilisation d'outils de pré-traitement morphosyntaxique, comme MElt. Ne disposant pas d'analyseur performant pour des textes dégradés écrits à la fois en arabe standard et dialectal, nous avons sélectionné un sous-ensemble de 22 traits qui peuvent être extraits sans analyseur. Nous les avons rassemblés en quatre groupes, comme le montre le tableau 5.7.

B Expériences et résultats

Pour la classification des tweets en ironiques/non ironiques, nous avons exploité l'ensemble des traits déjà défini, avec plusieurs classifieurs sous la plateforme Weka : SMO, Naïve Bayes, Régression Logistique Multinomiale, Régression Linéaire, Random Tree et Random Forest. Nous avons entrainé les classifieurs avec un corpus équilibré de 1 733 tweets ironiques et 1 733 tweets non ironiques avec 80% pour l'entrainement et 20% pour le test pour une première expérience et validation croisée à 10 replis pour une deuxième expérience.

[11] http://alt.qcri.org/semeval2016/task7/

[12] Nous avons testé d'autres lexiques, tels que *Arabic translation of Bing Liu's Lexicon* et *Arabic translation of MPQA Subjectivity Lexicon*. Cependant, les expériences que nous avons menées avec ces lexiques n'ont pas été concluantes.

Groupes de traits	**Traits**	**Types de traits**
Traits de surface	Ponctuation (... / ! / ?)	Binaire
	Emoticône	Binaire
	Nombre d'émoticônes	Numérique
	Citation (présence de texte entre « »)	Binaire
	Connecteurs de discours qui ne déclenchent pas d'opposition	Binaire
	Mots d'opposition	Binaire
	Exclamation (! ! / ! ! ! / ou plus)	Binaire
	Interrogation (? ? / ? ? ? / ou plus)	Binaire
	Exclamation+Interrogation (? ! / ! ?)	Binaire
	Nombre de mots	Numérique
	Interjection	Binaire
	Nombre d'interjections	Numérique
Traits de sentiments	Opinion négative	Binaire
	Opinion positive	Binaire
	Nombre d'opinions positives	Numérique
	Nombre d'opinions négatives	Numérique
Traits pour les modifieurs	Intensifieur	Binaire
	Verbe de reportage	Binaire
	Mot de négation	Binaire
Traits de contexte	Pronom personnel	Binaire
	Entité nommée	Binaire
	Nombre d'entités nommées	Numérique

TABLE 5.7 : Ensemble des traits exploités pour l'apprentissage pour l'arabe.

Nous avons opté pour ces deux expériences car la taille du corpus reste limitée (3 466 tweets). Le classifieur Random Forest avec les paramètres par défaut obtient les meilleurs résultats. Nous les détaillons dans le tableau 5.8.

	Train/test				**10-Cross validation**			
	Précision	Rappel	F-mesure	Exactitude	Précision	Rappel	F-mesure	Exactitude
IR	0,728	0,707	0,718	72,29	0,719	0,735	**0,727**	**72,36**
NIR	0,718	0,739	0,728		0,729	0,713	**0,721**	

TABLE 5.8 : Résultats de la classification des tweets en ironique (IR)/non ironique (NIR) obtenus avec Random Forest en exploitant tous les traits

Afin d'améliorer ces résultats, nous avons appliqué trois algorithmes de sélection de traits sous la plateforme Weka à savoir : *Chi2* et *GainRatio* (voir Figure 5.5) pour avoir l'ordre décroissant des traits les plus performants (du plus performant au moins performant) et *CfsSubsetEval* qui donne la combinaison des meilleurs traits en considérant la capacité prédictive individuelle de chaque trait avec le degré de redondance entre eux. Ce dernier algorithme a donné la combinaison de traits suivante : *nombre d'émoticônes, exclamations, négations, nombre d'interjections, entités nommées, nombre d'entités nommées*.

Malheureusement, l'utilisation du sous-ensemble de traits jugés les plus performants par les algorithmes de sélection n'a pas donné de meilleurs résultats que ceux obtenus en utilisant tous les traits. Par conséquent, nous avons testé une autre approche en effectuant l'apprentissage du classifieur Random Forest en ajoutant les traits un par un (en respectant l'ordre donné par les algorithmes de sélection) afin d'interpréter l'influence de chaque trait. Ceci devrait nous permettre d'avoir le sous-ensemble de traits qui permettent de maximiser les performances.

Cette étude a montré que l'utilisation de tous les traits sauf le trait *verbe de reportage* maximise les résultats avec des valeurs d'exactitude de $72,76\%$ au lieu de $72,36\%$ avec tous les traits, ainsi qu'une f-mesure de 73% au lieu de $72,70\%$ pour la classe ironique et une f-mesure de $72,50\%$ au lieu de $72,10\%$ pour la classe non ironique (voir Tableau 5.9).

	Train/test				**10-Cross validation**			
	Précision	Rappel	F-mesure	Exactitude	Précision	Rappel	F-mesure	Exactitude
IR	0,723	0,696	0,709	71,57	0,724	0,736	**0,730**	**72,76**
NIR	0,709	0,736	0,722		0,731	0,720	**0,725**	

TABLE 5.9 : Résultats de la classification des tweets en ironique (IR)/non ironique (NIR) obtenus avec Random Forest en exploitant la meilleure combinaison de traits.

Kchi²		GainRatio	
0.10578219	Nb_Entités_Nommées	0.1091403	Entiés_Nommées
0.08806588	Entiés_Nommées	0.069659	Nb_Entités_Nommées
0.01807329	Nb_émoticônes	0.0501787	Nb_Interjection
0.01788328	Emoticône	0.03029	Nb_Mots
0.01240071	Nb_Interjection	0.0245372	Nb_émoticônes
0.00526265	Nb_Mots	0.0240979	Emoticônes
0.00506479	Interjection	0.020977	Interjection
0.00289741	Ponctuation	0.0127604	Interrogation
0.00289665	Exclamation	0.0122677	Exclamation
0.00288611	Nb_OpinionPositive	0.0060024	OpinionPositive
0.00209972	Négation	0.0046953	OpinionNégative
0.00199026	OpinionNégative	0.0043288	Nb_OpinionPositive
0.00197259	OpinionPositive	0.0032733	Négation
0.00118275	Interrogation	0.0029279	Ponctuation
0.00099095	Pronom_personnel	0.0016047	Opposition
0.00032808	Oposition	0.0010676	Pronom_personnel
0.00015437	Intensifieur	0.0007447	Intensifieur
0.00003629	ConnecteurDiscours-Opposition	0.0000376	ConnecteurDiscours-Opposition
0.00000962	Citation	0.0000307	Exclamation_Interrogation
0.00000371	Exclamation_Interrogation	0.0000295	Citation
0	Nb_OpinionNégative	0	Nb_OpinionNégative

FIGURE 5.5 : Résultats donnés par les algorithmes de sélection de traits.

C Discussion

Bien que nous ayons exploité un ensemble de traits dont la plupart sont des traits de surface, nous avons obtenu des résultats très encourageants. En comparant les résultats obtenus pour la classifications des tweets arabes en ironique/non ironique avec ceux pour les tweets français en exploitant les même traits (voir 4), nous remarquons que le comportement des algorithmes de classification n'est pas le même pour les deux langues. Pour le français, l'algorithme de classification SMO était le plus performant avec une f-mesure de $85,70\%$ pour la classe ironique alors que pour l'arabe la f-mesure ne dépasse pas $62,50\%$. Par contre, l'algorithme de classification *Random Forest* a été plus performant pour la tâche de classification des tweets ironiques arabes avec une f-mesure de 73% et moins performant pour la classification des tweets français avec une f-mesure de $75,40\%$.

5.4 Conclusion

Dans ce chapitre, nous avons proposé deux expériences. La première a été consacrée à l'étude de la portabilité de notre schéma d'annotation défini pour le français sur un corpus multi-lingue (anglais et italien). Cette expérience nous a permis de mesurer l'impact des phénomènes pragmatiques dans l'interprétation de l'ironie. Les résultats obtenus ont prouvé que notre schéma est fiable pour le français ainsi que pour l'anglais et l'italien en observant relativement les mêmes tendances en termes de catégories et indices d'ironie. En particulier, nous avons observé des corrélations entre les indices et les classes ironiques/non ironiques, entre les indices et les types d'activations de l'ironie (explicites ou implicites) et entre les indices et les catégories de l'ironie pour les trois langues étudiées. Ces observations sont intéressantes dans une perspective de détection automatique multi-lingue de l'ironie.

La deuxième expérience a été consacrée à la détection automatique de l'ironie dans un corpus de tweets arabes. Nous avons entrainé un modèle qui utilise certains traits de surface définis pour le français. Les résultats obtenus sont très encourageants étant donné (1) la difficulté de traitement des textes en arabe qui mélangent arabe standard et arabe dialectal, et (2) les résultats comparables obtenus sur d'autres langues. Par exemple, dans le cadre de notre travail pour le français, nous avons obtenu une précision de 93% pour la classe ironique alors que pour l'arabe, la précision est de 72,4%. D'autres travaux qui se sont intéressés à cette même tâche ont atteint des scores de précision de 30% pour le néerlandais (Liebrecht *et al.*, 2013) et 79% pour l'anglais (Reyes *et al.*, 2013).

Conclusion et perspectives

Nos travaux avaient un double objectif : (a) proposer le premier système de détection de l'ironie sur les réseaux sociaux pour la langue française, et (2) étudier la portabilité de ce système à d'autres langues. Ce travail se situe dans un champ de recherche extrêmement actif en traitement du langage, principalement en raison de l'importance de la détection de l'ironie et du sarcasme pour améliorer les performances des systèmes d'analyse d'opinions.

Pour atteindre nos objectifs, nous avons d'abord effectué un état de l'art complet sur les approches linguistiques et computationnelles de détection du langage figuratif. Bien que nos travaux concernent l'ironie et le sarcasme, notre état de l'art a couvert divers phénomènes figuratifs tels que l'humour, la satire, la métaphore ou encore la comparaison, car la frontières entre ces différents phénomènes est parfois floue. L'étude de l'existant a mis en évidence deux principales constatations :

1. Les travaux en linguistique étudient le langage figuratif d'un point de vue sémantique et pragmatique en se focalisant sur les mécanismes mis en œuvre dans l'expression linguistique de ce type de langage comme l'hyperbole, la question rhétorique, ou encore la fausse assertion. Ces travaux concernent principalement des textes littéraires comme des romans ou des poèmes.

2. Les travaux computationnels se focalisent sur la détection de l'ironie en la considérant comme un terme générique englobant le sarcasme et parfois la satire. Ces travaux concernent principalement les réseaux sociaux comme Twitter en raison de la présence de hashtags spécifiques qui permettent de préciser la pensée ironique ou sarcastique de l'auteur et en font des données de référence. Les approches proposées utilisent de l'apprentissage supervisé à base de divers traits lexicaux, syntaxiques voire plus rarement pragmatiques.

Nous avons alors adopté une approche mixte, à la fois linguistique et computationnelle. En effet, il a paru difficile d'aborder le traitement automatique de phénomènes complexes tels que le langage figuratif sans une étude approfondie de ces phénomènes en corpus. Nous avons alors suivi une démarche en trois étapes.

Premièrement, nous nous sommes intéressés à l'analyse des phénomènes pragmatiques utilisés pour exprimer l'ironie. La principale préoccupation a été de vérifier si les différents types d'ironie étudiés dans les travaux en linguistique peuvent être présents dans un corpus particulier collecté à partir des réseaux sociaux comme Twitter. Pour ce faire, nous avons proposé un schéma d'annotation multi-niveaux permettant d'identifier pour chaque tweet s'il est ironique ou non, son type d'ironie (explicite vs. implicite), les catégories d'ironie utilisées et enfin, les indices linguistiques déclencheurs de l'ironie (comme les émoticônes, la ponctuation, les mots d'opinion, etc.). Ce schéma d'annotation a été exploité dans le cadre d'une campagne d'annotation d'un corpus de 2 000 tweets en français. Les résultats quantitatifs ainsi que les analyses de corrélations entre les différents niveaux du schéma ont montré que l'activation de l'ironie dans la majorité des tweets ironiques est dûe soit à des contradictions implicites mettant en œuvre des fausses assertions, soit à des contradictions explicites sous forme d'oxymores ou de paradoxes. De plus, parmi les indices, la négation semblent être un marqueur extrêmement présent à la fois dans les tweets ironiques et non ironiques.

Ensuite, en exploitant l'ensemble des observations faites sur le corpus annoté, nous avons développé un système de détection automatique pour les tweets en français. Trois modèles ont été proposés : (1) *SurfSystem*, un modèle sur la base de traits surfaciques traditionnellement utilisés dans l'état de l'art, (2) *PragSystem*, un modèle qui utilise les traits pragmatiques extraits du contenu linguistique du tweet et de nouveaux traits, notamment les patrons d'opposition, qui ont été les plus productifs avec une exactitude de 87,7%, et enfin (3) *QuerySystem*, une méthode à base de requêtes qui s'applique sur les tweets contenant des fausses assertions contenant des négations et qui ont été mal classés par le classifieur *PragSystem*. Les expériences ont montré que cette dernière méthode permet d'améliorer la classification quand elle est appliquée sur des tweets non personnels (augmentation de l'exactitude à 88,51%).

Enfin, nous avons étudié la portabilité à la fois du schéma d'annotation et des modèles computationnels pour la détection de l'ironie dans un cadre multilingue (italien, anglais et arabe). Nous avons ainsi testé la performance du schéma d'annotation proposé sur l'italien et l'anglais et nous avons testé la performance du modèle de détection automatique à base de traits sur la langue arabe. Les résultats obtenus montrent que notre schéma s'applique parfaitement sur l'italien et l'anglais où les mêmes tendances que le français ont été observées. De plus, l'application d'un sous-ensemble de traits du modèle *PragSystem* sur un corpus de tweets arabes a montré que ces derniers sont portables à l'arabe où une exactitude de 72,76% a été atteinte. Ce dernier résultat, bien que faible par rapport à celui obtenu pour le français, montre cependant des perspectives encourageantes pour la détection de l'ironie dans des tweets arabes mélangeant arabe classique et arabe dialectal.

Les travaux de cette thèse ouvre des perspectives intéressantes pour le futur. Nous envisageons 3 pistes. La première piste est d'étudier la possibilité de l'amélioration de la détection automatique de la polarité des tweets ironiques/sarcastiques lors de l'analyse de

sentiments. Dans ce cadre, nous avons proposé trois tâches d'analyse des tweets lors de la de la première campagne d'évaluation sur l'analyse d'opinion et le langage figuratif DEFT@TALN 2017 (Benamara *et al.*, 2017a) que nous avons co-organisée en collaboration avec le LIMSI. Pour cette nouvelle édition du défi, nous avons proposé trois tâches : (1) classification des tweets non figuratifs selon leur polarité (objectif, positif, négatif ou mixte) ; (2) identification du langage figuratif (ironie, sarcasme ou humour) ; (3) classification des tweets figuratifs et non figuratifs selon leur polarité (objectif, positif, négatif ou mixte). Pour ce défi, le corpus *FrIC* a été étendu à 7 724 tweets en français qui portent sur des sujets d'actualité (politique, sport, cinéma, émissions TV, artistes, etc.) collectés entre 2014 et 2016, en fonction de la présence de mots-clés (Hollande, Valls, #DSK, #FIFA,...) et/ou de hashtags spécifiques, indicateurs du langage figuratif (*#ironie, #sarcasme, #humour, #joke*). Douze équipes ont participé à ce défi. Les meilleurs résultats, en macro f-mesure, sont de 0,650 pour la tâche (1), 0,783 pour la tâche (2) et 0,594 pour la tâche (3). Ces résultats montrent clairement que l'usage du langage figuratif complique considérablement l'analyse d'opinions.

La second piste est d'étudier comment l'application de notre schéma peut contribuer à éclairer la question de la distinction entre l'ironie et le sarcasme. Cette distinction commence à faire l'objet de travaux récents qui proposent de distinguer automatiquement les tweets ironiques des tweets sarcastiques (Sulis *et al.*, 2016b). Il pourrait être intéressant de voir la relation entre les phénomènes pragmatiques de granularité fine liés à l'ironie que nous avons proposés dans la présente étude et la distinction de niveau supérieur entre l'ironie et le sarcasme.

Enfin, nous envisageons la proposition d'un système automatique pour la détection de l'ironie avec un corpus multilingue. Dans ce cadre, nous voulons évaluer la performance d'un classifieur entrainé sur un corpus et testé sur un autre corpus avec une langue différente dans le but de trouver une combinaison de traits performants pour la tâche de détection de l'ironie et indépendants de la langue. De plus, nous envisageons l'amélioration de la détection automatique de l'ironie/sarcasme en exploitant un modèle d'apprentissage profond à base de réseaux de neurones. Ce travail est actuellement en cours dans le cadre d'une collaboration avec l'université de Turin et l'université de Valence en Espagne.

Appendices

Annexe A

Annexes

A.1 Catégories de l'ironie

A.1.1 Catégories étudiées dans la littérature linguistiques

Le tableau A.1 présente une synthèse des principales catégories et marqueurs de l'ironie étudiés par les linguistes en se focalisant sur un type particulier de l'ironie à savoir l'ironie textuelle. Dans cette section, nous commençons par la présentation de ces différents marqueurs puis nous présentons ceux qui ont été retenus pour l'étude de l'ironie dans les tweets.

A Contradiction / Fausse logique

Partant de la définition *« l'ironie verbale exprime une contradiction entre la pensée du locuteur et son expression »* (Niogret, 2004), la contradiction devient l'un des principaux marqueurs d'ironie.

Dans un énoncé ironique, on se trouve avec deux segments textuels. Dans le premier, on trouve une affirmation et dans le deuxième on trouve l'information qui contredit la première. Autrement dit, l'ironiste dit le contraire de ce qu'il pense en laissant une trace dans le texte qui prouve que sa déclaration est ironique. Ceci va permettre au lecteur de comprendre le sens ironique/non ironique d'un texte.

Dans ce cadre, Attardo (2000b) admet l'idée de Kerbrat-Orecchioni (1976) et de Muecke (1978) qui considèrent que l'ironie est marqué par une contradiction ou un contraste entre ce qui est dit et ce qui est attendu. Ceci a été plus détaillé par Didio (2007), qui considèrent que les contradictions dans un discours permettent à l'énonciataire de comprendre le sens ironique d'un texte en partant de l'idée que la contradiction unit deux énoncés qui affirment et nient un même objet de connaissance. Prenant l'exemple suivant cité par Didio (2007) :

Marqueurs	**Références**
Contradiction / Fausse logique	(Kerbrat-Orecchioni, 1976) (Muecke, 1978) (Tayot, 1984) (Attardo, 2000b) (Barbe, 1995) (Didio, 2007)
Métaphore	(Grice, 1970) (Boyd, 1979) (Wilson & Sperber, 1986) (Wilson & Sperber, 1988) (Wilson & Sperber, 1992) (Kittay, 1990) (Kreuz & Roberts, 1993) (Barbe, 1995) (Song, 1998) (Ritchie, 2005) (Burgers, 2010) (Bres, 2010)
Hyperbole / Exagération	(Kreuz & Roberts, 1993) (Kreuz & Roberts, 1995) (Mercier-Leca, 2003) (Didio, 2007) (Burgers, 2010)
Euphémisme	(Muecke, 1978) (Fromilhague, 1995) (Seto, 1998) (Yamanashi, 1998) (Mercier-Leca, 2003) (Didio, 2007) (Burgers, 2010)
Absurdité	(Didio, 2007)
Effet de surprises	(Colston & Keller, 1998) (Didio, 2007)
Répétition	(Muecke, 1978) (Berntsen & Kennedy, 1996) (Wilson & Sperber, 2004) (Burgers, 2010)
Question rhétorique	(Muecke, 1978) (Berntsen & Kennedy, 1996) (Haiman, 1998) (Attardo, 2000b) (Burgers, 2010)
Changement de registre	(Haiman, 1998) (Attardo, 2000b) (Burgers, 2010)
Oxymore	(Gibbs, 1994) (Song, 1998) (Mercier-Leca, 2003)
Paradoxe	(Tayot, 1984) (Barbe, 1995) (Mercier-Leca, 2003)
Guillemets	(Tayot, 1984) (Gibbs, 1994) (Attardo, 2001) (Burgers, 2010)
Emoticônes	(Tayot, 1984) (Kreuz, 1996) (Burgers, 2010)
Exclamation	(Tayot, 1984) (Wilson & Sperber, 1992) (Seto, 1998) (Attardo, 2000b) (Attardo, 2001) (Didio, 2007) (Burgers, 2010)
Majuscule	(Haiman, 1998) (Burgers, 2010)
Texte barré et caractères spéciaux	(Burgers, 2010)

TABLE A.1 : Les marqueurs de l'ironie étudiés par dans la littérature linguistiques.

(A.1) Mademoiselle de Kerkabon, qui n'avait jamais été mariée, quoiqu'elle eût grande envie de l'être, conservait de la fraîcheur à l'âge de quarante-cinq ans ; son caractère était bon et sensible ; elle aimait le plaisir et était dévote (Didio, 2007).

Didio (2007) considère que cet exemple représente deux contradictions. La première est exprimé au niveau de la phrase *"conservait de la fraîcheur à l'âge de quarante-cinq ans"*. Malgré la multitude des points de vue à propos l'âge de la fraîcheur des femmes,

les auteurs jugent qu'une femme peut conserver la fraîcheur à l'âge de trente ans, non pas à l'âge de quarante-cinq ans. A ce niveau, la contradiction a été exprimé implicitement vu que le lecteur doit faire recours à ces connaissances pour comprendre le sens de la contradiction.

Une deuxième contradiction exprimé au niveau de cet exemple *"elle aimait le plaisir et était dévote"*. Une contradiction ironique a été exprimé explicitement par la présence de *"aimer le plaisir"* et *"était dévote"* vu qu'une personne ne peut pas aimer le plaisir et être dévote à la fois.

La contradiction a été particulièrement considéré par Didio (2007) comme étant *une fausse logique ou contre-sens*. Ceci est exprimé par le faite qu'on exprime volontairement le contraire de ce que l'on pense ou bien quelque chose qui est faux par rapport à un contexte (Exemple A.2). Barbe (1995) a appelé cette catégorie *"mensonges"*.

(A.2) Voilà qu'un corsaire de Salé fond sur nous et nous aborde ; nos soldats se défendirent comme des soldats du pape : ils se mirent tous à genoux en jetant leurs armes, et en demandant au corsaire une absolution in articulo mortis (Didio, 2007).

Dans cet exemple, la fausse logique est exprimé par le faite que les soldats se défendent en jetant leurs armes alors qu'ils ne peuvent pas se défendre sans armes.

Barbe (1995) a considéré que *la fausse proposition* exprime *un mensonge*. Il a défini le mensonge comme un phénomène contenant une *opposition vérité-mensonge*. Le menteur veut cacher la vérité. Il le fait en imitant les caractéristiques de la parole de vérité et en évitant les signaux qui peuvent mettre en péril ses déclarations.

Dans un sens superficiel, l'ironie et le mensonge ont la même définition : une phrase prononcée qui cache une phrase non prononcé. Même si l'ironie pourrait être classé comme un type de mensonge, la vérité et le mensonge ne forment pas une opposition dans l'ironie (Barbe, 1995).

L'idée de rapprocher les deux notions mensonge/contre vérité avec l'ironie a été étudié aussi par Tayot (1984) qui a trouvé que l'ironiste se plait par le faite de ne pas laisser des marqueurs physiques à la disposition du récepteur. Ce dernier se voit donc contraint d'explorer le contexte linguiste ou extra-linguistique de la sequence afin de détecter l'impertinence qui dévoilera la supercherie du discours offert à son attention. Sa perspicacité lui permettra, dans certaine cas, de démasquer une *"contre vérité"*, arme dont les menteurs et les ironistes se partagent l'usage.

B Métaphore

Partant d'une définition du dictionnaire *Larousse* : « la métaphore est une figure de style qui consiste à établir une comparaison entre deux réalités, comparaison qui est fondée sur

une analogie que l'on instaure entre les deux référents ». Contrairement à la comparaison proprement dite, la métaphore ne comporte aucun élément grammatical, par exemple les marqueurs *comme, ainsi que, tel, semblable à*, explicitant le rapport comparatif.

La métaphore est non seulement le trope le plus courant, mais il a également reçu la plus grande quantité d'attention des psychologues, des philosophes et des théoriciens de la littérature (Grice, 1970; Kittay, 1990; Kreuz & Roberts, 1993; Barbe, 1995; Ritchie, 2005).

Afin d'avoir une définition approfondie de la métaphore, Song (1998) a reprit les différentes définitions cités par les linguistes. Kittay (1990) a interprété la métaphore comme un sens du second ordre qui est obtenu lorsque les caractéristiques de l'énoncé et de son contexte indiquent à l'auditeur ou le lecteur que le sens du premier ordre de l'expression est indisponible ou ne convient pas. (Wilson & Sperber, 1986; Wilson & Sperber, 1988; Wilson & Sperber, 1992; Grice, 1970; Kittay, 1990) affirment qu'un lecteur arrive à interpréter un énoncé comme étant métaphorique sauf s'il arrive, en faisant recours à ses connaissances, à trouver le sens littéral de l'énoncé. Par exemple :

(A.3) Across the ice, the snow is sweeping
lonely the wind, the snow, the heart
are playing together. (Berntsen & Kennedy, 1996)

Selon (Barbe, 1995), la métaphore se présente au cœur de la langue. Un terme est représenté avec un sens littéral et un sens figuré. En reliant la métaphore à l'ironie, Barbe (1995) indique que bien que ces deux phénomènes pousse le lecteur à lire entre les lignes, leurs applications diffèrent. Parmi les différences, la métaphore est une figure de style alors que l'ironie est une attitude. Cela ne nie pas que la métaphore peut être utilisée à des fins ironiques. Une autre différence et que la métaphore est utilisé pour clarifier, éclairer, ou expliquer afin de construire un type de description, alors que l'ironie constitue un commentaire critique ou une évaluation et il permet de transmettre une attitude à l'égard d'une situation. Le point commun entre la métaphore et l'ironie est la nécessité des connaissances partagées afin de comprendre le sens ironique ou métaphorique d'un énoncé. Par exemple :

(A.4) I once had a girlfriend who had a child. I tell you she was a real beast. She was an Aquarius just like you. (Barbe, 1995)

Une décennie après ces études, les linguistes ont commencé à ce concentrer à la forte relation entre la métaphore et l'ironie.

En partant des différentes études menés sur la métaphore, l'ironie et l'humour, Ritchie (2005) a conclue que la métaphore, l'ironie et l'humour peuvent générer des changements importants à des environnements cognitifs et que l'humour et l'ironie servent souvent des fins de communication de base de façon subtile et multiformes qui les placent, aux côtés

de la métaphore et de la métonymie. Cette analyse a été reprise dans les travaux de Burgers (2010) qui considère la métaphore comme étant un marqueur de l'ironie. Selon Bres (2010), l'ironie fait partie, à l'instar de la métaphore, de ces plus vieux objets linguistiques du monde qui stimulent la réflexion sans jamais l'épuiser. Bres (2010) trouve que la relation entre l'ironie et la métaphore : « comme le cocktail ou le vin d'assemblage, procède d'une association délicate d'autres breuvages. Que l'un d'entre eux fasse défaut, et le cocktail voit ses arômes se volatiliser, s'affadit ou devient de mauvais goût » .

C Hyperbole / Exagération

L'hyperbole est une figure de style qui consiste à exprimer de façon exagérée une idée ou un sentiment. Elle est souvent utilisée pour produire une forte impression ou pour insister sur un point. Kreuz et Roberts (1993) considèrent que l'hyperbole est une forme du langage figuratif très utilisée, mais elle a été relativement négligé par les linguistiques malgré que l'hyperbole a été présent dans plus que 27% des histoires courtes américaines. Cette négligence à pousser Kreuz et Roberts (1995) à mener une étude approfondie sur l'hyperbole. Cette étude fructueuse a prouvé que la présence de l'hyperbole dans un texte suggère l'intention ironique dans certain cas.

En particulier, Kreuz et Roberts (1995) trouvent que l'hyperbole se produit très fréquemment dans l'ironie verbale et qu'il joue un rôle important dans la perception des déclarations ironiques. À un niveau intuitif, la relation entre l'hyperbole et l'ironie semble être importante. (Kreuz & Roberts, 1995) ont constaté que l'hyperbole et l'ironie partagent un certain nombre d'objectifs de discours importants, tels que *"l'humour", "mettre l'accent sur quelque chose" et "clarifier quelque chose"*.

Pour conclure Kreuz et Roberts (1995) et Burgers (2010), précisent que l'hyperbole peut jouer un rôle important dans la perception de l'ironie et que l'hyperbole fonctionne probablement comme un indice fiable pour la reconnaissance de l'intention ironique. En outre, la présence de l'hyperbole augmente la probabilité d'une interprétation ironique, même en l'absence d'une remarque non véridique. Autrement dit, même si ces déclarations ne sont pas contraires à la réalité des choses, l'exagération, par elle-même, peut suggérer une intention ironique (ExempleA.5).

(A.5) We till and sell and pile our money
and the hedge is ten feet high
we dead the future, what it will bring
vexation, bad luck and troubles.
I trudge my round with the dog and the gun
and if anyone enters, they'll get shot
for oh-so-envious people are
just beacause we are doing so well (Berntsen & Kennedy, 1996)

Dans le même cadre, Mercier-Leca (2003) et Didio (2007) ont défini l'hyperbole comme une exagération du propos afin de produire une plus grande impression. Ils ont suggéré que l'hyperbole est l'un des signaux les plus voyants de l'ironie mais toutes les exagérations ne sont pas forcément ironiques. Didio (2007) s'est référé à l'exemple A.6 pour montrer que l'hyperbole peut être employée à des fins comiques, voire ironiques.

(A.6) « L'autre jour, Mme de la Villemenue, vieille coquette qui désire encore plaire, a voulu essayer ses charmes surannés sur le philosophe [Voltaire] : elle s'est présentée à lui dans tout son étalage et, prenant occasion de quelque phrase galante qu'il lui disait et de quelques regards qu'il jetait en même temps sur sa gorge fort découverte : – Comment, s'écria-t-elle, Monsieur de Voltaire, est-ce que vous songeriez encore à ces petits coquins-là ? Petits coquins, reprend avec vivacité le malin vieillard, petits coquins, Madame ! ce sont de bien grands pendards ! (Mémoires de Bachaumont, 30 mars 1778) »

Didio (2007) précisent que cet exemple est très particulier puisque l'hyperbole (apparente) est, en fait, tout à fait dépréciative en raison du jeu de mots dû à la polysémie du terme "pendards".

Didio (2007) a étudié l'exagération comme étant un phénomène séparé de l'hyperbole (Didio, 2007) sans préciser une différence entre l'hyperbole et l'exagération. Elle a défini l'exagération comme étant un moyen qui permet l'amplification de la réalité ou en lui donnant plus d'importance qu'elle n'en a réellement. La plupart des linguistes ne distinguent pas entre l'hyperbole et l'exagération. Ils considèrent que l'hyperbole est une exagération (Kreuz & Roberts, 1993; Kreuz & Roberts, 1995; Pougeoise, 2001; Mercier-Leca, 2003; Burgers, 2010).

D Euphémisme

L'euphémisme est une figure de style qui consiste à atténuer l'expression de faits ou d'idées considérés comme désagréables dans le but d'adoucir la réalité (Muecke, 1978; Seto, 1998; Burgers, 2010). Par conséquent, l'euphémisme est l'antonyme de l'hyperbole. A travers l'hyperbole, un orateur exagèrent dans l'évaluation littérale d'un message alors que dans un euphémisme, quelqu'un dit moins dans l'évaluation littérale que ce qui l'est en réalité. Par conséquent, un euphémisme est vraiment l'antithèse d'une exagération. Cela implique qu'une personne affaiblit une émotion forte ou une expression. Un euphémisme peut également être utilisé ironiquement.

Yamanashi (1998) se rejoint à Burgers (2010) et Seto (1998) pour confirmer que l'euphémisme est utilisé pour représenter quelque chose comme étant "moins importante qu'elle ne l'est vraiment" afin d'atténuer la réalité (exemple A.7). Ainsi, l'utilisation ironique ne peut

pas être correctement prédite par la définition traditionnelle de l'ironie verbale qui consiste à comprendre le contraire de ce qui est dit.

L'euphémisme a été considéré comme étant un marqueur de l'ironie par Mercier-Leca (2003) sous le nom **"Litote"**. Mercier-Leca (2003) a reprit la définition de Fromilhague (1995) qui a défini la litote comme suit : "On feint d'atténuer une vérité que l'on affirme implicitement avec force : on dit le moins pour le plus". Selon Fromilhague (1995) la litote peut être exprimer par *une négation* ou *une assertion restrictive* (il s'agit d'une assertion qui s'accompagne d'adverbes à portée restrictive, comme "peu", "pas beaucoup", "difficilement", etc.). Par exemple l'utilisation de :

(A.7) "moins bien" au lieu de "très mal".
"rejoindre les étoiles" ou "ne plus être" Pour signifier "le fait de mourir".
"Non-voyant" pour désigner "un aveugle"

E Absurdité

L'absurdité est exprimée par un raisonnement illogique. L'absurde peut être lié à une réaction comique ou tragique. Il signifie ce qui n'est pas en harmonie avec quelqu'un ou quelque chose (Didio, 2007) (Exemple A.8).

(A.8) Après le tremblement de terre qui avait détruit les trois quarts de Lisbonne, les sages du pays n'avaient pas trouvé un moyen plus efficace pour prévenir une ruine totale que de donner au peuple un bel auto-da-fé ; il était décidé par l'université de Coïmbre que le spectacle de quelques personnes brûlées à petit feu, en grande cérémonie, est un secret infaillible pour empêcher la terre de trembler (Didio, 2007).

La relation entre l'absurdité et l'ironie n'a pas été traité suffisamment par les linguistes. Didio (2007) est la seule linguiste qui a évoqué la présence d'une relation entre l'absurdité et l'ironie.

F Effet de surprise

La surprise est un état émotionnel provoqué par un événement inattendu ou par une révélation allant à l'encontre de l'image qu'on se faisait d'une situation. Elle est généralement brève, puis s'estompe ou laisse place à une autre émotion (wikipédia [1]).

La relation entre l'effet de surprise et l'ironie a fait partie d'un nombre restreint de travaux de linguistes. Didio (2007) considère que l'effet de surprise est un marqueur de

[1] https://fr.wikipedia.org/wiki/Surprise

l'ironie sans donné une définition particulière alors que Colston et Keller (1998) ont étudié le relation entre la surprise et l'ironie d'une manière inversé c'est à dire que l'ironie est un mécanisme d'expression de surprise. Dans ces travaux, ils ont défini la surprise comme suit : "la surprise est une réaction courante lorsque les événements ne se passent pas comme prévu. Les gens peuvent exprimer cette surprise en notant verbalement le contraste entre ce qui était attendu et ce qui est réellement arrivé. L'hyperbole verbale et l'ironie sont utiles dans l'expression de surprise car ils utilisent ce contraste d'une manière concise".

G Répétition

Partant des travaux de Wilson et Sperber (2004) qui ont démontré que l'écho (répétition) peut être une forte indication qu'un texte est ironique, Burgers (2010) a considéré la répétition ou l'écho comme étant un marqueur de l'ironie. Ce phénomène a été traité par d'autres linguistes à savoir Muecke (1978) et Berntsen (1996) sous le nom de **"Parodie"**.

Selon Burgers (2010) :

- Un écrivain peut ironiquement répéter quelque chose prononcé par une autre personne plutôt dans le texte ou en cas d'interaction orale - dans le dialogue. Ce type de répétition est appelé une répétition sur la base de co-texte. Dans ce cas, un énoncé ou une partie d'un énoncé est ironiquement répétée dans le même texte (qui n'a pas été utilisé ironiquement dans sa première usage) (Exemple A.9).

  (A.9) (1) This movie was fantastic.
  (2) No, really fantastic.
  (3) FAN-TAS-TIC. (Burgers, 2010)

- Un écrivain peut ironiquement répéter quelque chose qui n'a pas été mentionné plus haut dans le texte en discussion ou dans le même dialogue, mais il a été mentionné dans un autre texte. Ce type de répétition est appelé une répétition en fonction du contexte.

Afin d'éviter toute confusion entre les deux, (Burgers, 2010) à attribuer à une répétition ironique dans le co-texte l'étiquète "répétition de co-textuel" et pour une répétition ironique dans le contexte, l'étiquète "écho".

Dans les phrase de (1) à (3) de l'exemple A.9, l'orateur répète sa déclaration que le film a été fantastique. Selon (Burgers, 2010), dans le cas où l'orateur n'a pas aimé le film, les énoncés (1) et (3) sont considérés comme étant ironiques et que dans l'énoncé (2), la répétition du mot "fantastique" représente un marqueur d'ironie.

H Question rhétorique

Une question rhétorique n'est pas une question réelle ; c'est une question sur laquelle le locuteur ne prévoit pas de recevoir une réponse, parce que la réponse est déjà claire. Cela signifie qu'une question rhétorique représente un point de vue et non pas une question (Exemple A.10) (Burgers, 2010). Plusieurs linguistes comme Muecke (1978) et Barbe (1995) ont considéré les questions rhétoriques comme étant un marqueur de l'ironie sans avoir donnée de définitions précises à ce phénomène.

(A.10) Could the weather be any better for a picnic ? (Burgers, 2010)

I Changement de registre

Un changement de registre est un changement soudain dans le style. Dans un énoncé, le changement de registre est exprimé par l'utilisation des mots inattendus d'un autre registre (dans un texte formel on utilise soudain des mots informels ou vice versa). Il se manifeste aussi par le changement brusque de sujet de la phrase ou l'utilisation exagéré de la politesse dans une situation où ceci est inapproprié (Burgers, 2010).

Attardo (2000b) et Haiman (1998) confirment la relation entre l'ironie et la politesse en considérant qu'une remarque ironique est plus polie qu'une critique directe. Dans ce cas, l'ironie est le but essentiel de l'ironiste mais l'utilisation de la politesse permet d'atténuer l'agressivité (Exemple A.11).

(A.11) En s'adressant à un ami : "Vous pouvez m'accorder l'honneur d'écouter un à l'autre de vos prédictions fines."

J Oxymore

L'oxymore est une figure de construction qui repose sur une apparente contradiction logique. C'est une figure d'opposition. Elle se repère au niveau de l'énoncé par le rapprochement syntaxique de deux éléments qui forment une contradiction sémantique (Gibbs, 1994; Song, 1998; Mercier-Leca, 2003).

Mercier-Leca (2003) considère que l'oxymore présente une affinités avec l'ironie dans la mesure où elle repose sur une feinte : on fait semblant d'opposer des éléments qui, en réalité, sont compatibles (Exemple A.12).

(A.12) "Je suis la plaie et le couteau ! [...] Et la victime et le bourreau !"
"une obscure clarté", "un silence éloquent" (Mercier-Leca, 2003).

(A.13) "Le règne de Louis VIII débutait triomphalement par un bain de sang qui étendit le domaine royal jusqu'à la Méditerranée. Malheureusement, ce règne prometteur tourna court".

Mercier-Leca (2003) indique que l'ironie repose dans l'Exemple A.13 sur l'oxymore opposant "triomphalement" et "bain de sang".

K Paradoxe

Selon Mercier-Leca (2003), l'ironie repose sur un paradoxe dont le caractère frappant est accentué par la syntaxe asyndétique (économisant les liens logique), qui, par contraste, met en valeur la seule conjonction de coordination présente à savoir le « mais » (Exemple A.14).

(A.14) "On pense à moi pour une place, mais par malheur j'y étais propre : il fallait un calculateur, ce fut un danseur qui l'obtint" (Mercier-Leca, 2003).

Tayot (1984) considère que le paradoxe est un instrument du sarcasme alors que Barbe (1995) considèrent que l'ironie exprimée en termes de paradoxe ressemble à l'opposition ironique (Barbe, 1995).

L Guillemets

Le mot mis entre guillemets "s'emploie pour exprimer une réserve sur un terme que l'on ne prend pas à son compte", selon le Dictionnaire de l'Académie Française, et "se dit pour indiquer qu'on ne prend pas à son compte le mot ou la locution qu'on emploie" selon le Petit Robert. De ce fait, on peut constater que si les termes mis entre guillemets veulent juste signifier leurs contraires, il y aura de l'ironie. Autrement dit, la mise entre guillemets peut représenter, selon le cas, un type d'ironie (Tayot, 1984; Gibbs, 1994; Attardo, 2001; Burgers, 2010).

La transcription écrite de la langue parlée est faite selon des conventions typographiques. Ceci peut être utile pour exprimer une intonation ironique. Les guillemets sont utilisés pour transmettre un certain détachement d'un énoncé écrit et donc de l'ironie (Attardo, 2001).

Selon Gibbs (1994), l'utilisation des guillemets est un geste non verbal dans le vocabulaire de nombreux orateurs américains dans le but d'exprimer l'ironie. L'utilisation des guillemets annonce que l'orateur va imiter le discours ou l'état d'esprit de la personne citée, souvent pour obtenir un effet sarcastique.

M Emoticônes

Depuis les années 80, vu l'absence total de l'utilisation des émoticônes dans les romans, les poèmes, etc. quelques linguistes ont mis l'accent que les expression du visage peuvent être un indice très fort de l'ironie (Tayot, 1984; Kreuz, 1996). Malgré que la première trace d'une émoticône daterait de 1648 qui a été utilisé dans la poème *To Fortune* du poète anglais *Robert Herrick* (wikipédia [2]), la faible utilisation des émoticônes a permis la négligence de l'importance de l'utilisation des émoticônes qui permet d'exprimer les expression du visage.

Commençant par la définition de ce phénomène, une émoticône est une courte figuration symbolique d'une émotion, d'un état d'esprit, d'un ressenti, d'une ambiance ou d'une intensité, utilisée dans un discours écrit (Burgers, 2010).

Tayot (1984) indique que l'intonation, la mimo-gestualité (par exemple le "tongue in cheek" britannique, infime déformation de la joue obtenue par un subtil déplacement de la langue ou encore le clin d'œil français) marquent parfois l'ironie orale.

Sous le nom **expression du visage**, Kreuz (1996) explique la façon avec laquelle l'expression du visage peut être un indice de l'ironie comme suit : « Tout en parlant à une personne, un orateur peut exprimer des attitudes au sujet de ses déclarations à travers une variété de signaux physiques. Par exemple, un orateur peut involontairement secouer la tête pour indiquer merveille, ou mouiller ses lèvres pour indiquer la nervosité. D'autres mouvements de la tête, les yeux et les paupières peuvent être tout à fait volontaire, et beaucoup peuvent être utilisés pour indiquer l'intention ironique. Le clin d'œil peut être utilisé pour suggérer que l'orateur n'a pas l'intention que son énoncé soit pris au sérieux. L'orateur peut également choisir de hocher lentement la tête, ou à rouler ses yeux ».

Symbolisant directement les émotions, les émoticônes se prêtent facilement à certaines figures de style telles que l'ironie. On pourra ainsi voir une personne feindre la détresse psychologique à l'annonce d'un départ de quelques jours ou manifester la joie lors d'un événement malheureux. Au contraire, l'émoticône est chargée de donner à la phrase une signification, une nuance, qu'elle n'avait pas.

Burgers (2010) a reprit les travaux des linguistes en attribuant le nouveaux nom "émoticône" attribué au "expression du visage" utilisé généralement dans les réseaux sociaux. Il considère que les émoticônes comme :-) ou ;-) peuvent marquer l'ironie.

N Exclamation

L'utilisation de la ponctuation comme "!", "?" et même la combinaison des deux "!?", "?!" a été considéré comme marqueur d'ironie (Attardo, 2000b). La plupart des travaux linguistiques se sont focalisés sur l'utilisation du point d'exclamation.

[2] https://fr.wikipedia.org/wiki/%C3%89motic%C3%B4ne

Le point d'exclamation est utilisé dans plusieurs énoncés pour marquer la valeur inverse du propos. Le plus intéressant c'est que le point d'exclamation, signe de ponctuation employé dans la langue écrite, correspond dans la langue orale à une exclamation, c'est-à-dire à une intonation ascendante.

Beaucoup d'énoncés ironiques, tirés du dictionnaire, ont un point d'exclamation. Voyons-en quelques exemples : "C'est du beau travail !" ; "C'est un beau gâchis !". On peut donc dire que *l'exclamation* peut être un marqueur de *l'ironie conversationnelle* ; tandis que *le point d'exclamation* peut être aussi un marqueur de *l'ironie textuelle*. Mais il faut souligner que toute proposition exclamative n'est pas nécessairement ironique. Cependant, une expréssion d'ironie peut être marquée par l'exclamation à l'oral et par le point d'exclamation à l'écrit. On peut donc dire que l'exclamation ou le point d'exclamation seront aussi, selon le cas, des marqueurs d'ironie (Tayot, 1984; Wilson & Sperber, 1992; Seto, 1998; Attardo, 2001; Didio, 2007; Burgers, 2010).

Attardo (2001) considèrent que le point d'exclamation est utilisé pour mettre en évidence l'ironie. Didio (2007) indiquent que faute d'un point d'ironie (figure A.1 conçu par Alcanter de Brahm) qui n'a pas rencontré de succès, le point d'exclamation sert dans plusieurs énoncés pour marquer la valeur inverse du propos.

FIGURE A.1 : Point d'ironie

O Majuscule, texte barré et caractères spéciaux

L'utilisation de l'écriture en majuscule a été considéré par quelques linguistes (Haiman, 1998; Burgers, 2010) comme étant un marqueur d'ironie (Exemple A.15). En revanche, l'utilisation des textes barrés (Exemple A.16) et des caractères spéciaux (par exemple : Your Weather Report^{TM} is great) ont été particulièrement étudiés par (Burgers, 2010).

(A.15) It is GREAT weather. (Burgers, 2010)

(A.16) It is horribly ~~great~~ weather. (Burgers, 2010)

Bibliographie

[Abdaoui *et al.*, 2015] ABDAOUI A., TAPI NZALI M. D., AZÉ J., BRINGAY S., LAVERGNE C., MOLLEVI C. & PONCELET P. (2015). Advanse : Analyse du sentiment, de l'opinion et de l'émotion sur des tweets français. In *Actes de la 11e Défi Fouille de Texte*, p. 78–87, Caen, France : Association pour le Traitement Automatique des Langues.

[Abdul-Mageed *et al.*, 2014] ABDUL-MAGEED M., DIAB M. & KÜBLER S. (2014). Samar : Subjectivity and sentiment analysis for arabic social media. *Computer Speech & Language*, **28**(1), 20–37.

[Abouenour *et al.*, 2012] ABOUENOUR L., BOUZOUBAA K. & ROSSO P. (2012). Idraaq : New arabic question answering system based on query expansion and passage retrieval.

[Afify *et al.*, 2006] AFIFY M., SARIKAYA R., KUO H.-K. J., BESACIER L. & GAO Y. (2006). On the use of morphological analysis for dialectal arabic speech recognition. In *INTERSPEECH*.

[Al-Sughaiyer & Al-Kharashi, 2004] AL-SUGHAIYER I. A. & AL-KHARASHI I. A. (2004). Arabic morphological analysis techniques : A comprehensive survey. *Journal of the American Society for Information Science and Technology*, **55**(3), 189–213.

[Alorifi, 2008] ALORIFI F. S. (2008). *Automatic identification of arabic dialects using hidden markov models*. PhD thesis, University of Pittsburgh.

[Angenot, 1982] ANGENOT M. (1982). *La parole pamphlétaire : contribution à la typologie des discours modernes*, volume 17. Payot.

[Asher *et al.*, 2009] ASHER N., BENAMARA F. & MATHIEU Y. Y. (2009). Appraisal of opinion expressions in discourse. *Lingvisticæ Investigationes*, **32**(2), 279–292.

[Attardo, 1994] ATTARDO S. (1994). *Linguistic theories of humor*, volume 1. Walter de Gruyter.

[Attardo, 2000a] ATTARDO S. (2000a). Irony as relevant inappropriateness. *Journal of pragmatics*, **32**(6), 793–826.

[Attardo, 2000b] ATTARDO S. (2000b). Irony markers and functions : Towards a goal-oriented theory of irony and its processing. *Rask*, **12**(1), 3–20.

[Attardo, 2001] ATTARDO S. (2001). *Humorous texts : A semantic and pragmatic analysis*, volume 6. Walter de Gruyter.

[Azé & Roche, 2005] AZÉ J. & ROCHE M. (2005). Présentation de l'atelier deft'05. In *Proc. of TALN*, p. 99–111.

[Bahou *et al.*, 2010] BAHOU Y., MASMOUDI A. & HADRICH BELGUITH L. (2010). Traitement des disfluences dans le cadre de la compréhension automatique de l'oral arabe spontané. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles*, Montréal, Canada : Association pour le Traitement Automatique des Langues.

[Baker *et al.*, 2003] BAKER C. F., FILLMORE C. J. & CRONIN B. (2003). The structure of the framenet database. *International Journal of Lexicography*, **16**(3), 281–296.

[Bamman & Smith, 2015] BAMMAN D. & SMITH N. A. (2015). Contextualized sarcasm detection on twitter. In *Proceedings of the International Conference on Web and Social Media*, ICWSM, p. 574–577.

[Barbe, 1995] BARBE K. (1995). *Irony in context*, volume 34. John Benjamins Publishing.

[Barbieri & Saggion, 2014a] BARBIERI F. & SAGGION H. (2014a). Modelling irony in twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, p. 56–64.

[Barbieri & Saggion, 2014b] BARBIERI F. & SAGGION H. (2014b). Modelling irony in twitter : Feature analysis and evaluation. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, p. 4258–4264.

[Barbieri *et al.*, 2014] BARBIERI F., SAGGION H. & RONZANO F. (2014). Modelling sarcasm in twitter, a novel approach. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, p. 50–58.

[Basile *et al.*, 2014] BASILE V., BOLIOLI A., NISSIM M., PATTI V. & ROSSO P. (2014). Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proc. of EVALITA 2014*, p. 50–57, Pisa, Italy : Pisa University Press.

[Basile & Nissim, 2013] BASILE V. & NISSIM M. (2013). Sentiment analysis on italian tweets. In *Proceedings of of WASSA 2013*, p. 100–107.

[Bastien *et al.*, 2012] BASTIEN F., LAMBLIN P., PASCANU R., BERGSTRA J., GOODFELLOW I., BERGERON A., BOUCHARD N., WARDE-FARLEY D. & BENGIO Y. (2012). Theano : new features and speed improvements. *arXiv preprint arXiv :1211.5590*.

[Bautain, 1816] BAUTAIN L. (1816). *De la satire*. de l'Imprimerie de C.-F. Patris.

[Bdour & Gharaibeh, 2013] BDOUR W. N. & GHARAIBEH N. K. (2013). Development of yes/no arabic question answering system. *arXiv preprint arXiv :1302.5675*.

[Benamara, 2017] BENAMARA F. (2017). Analyse automatique d'opinions États des lieux et perspectives. *Techniques de l'ingénieur Représentation et traitement des documents numériques*, **base documentaire : TIB312DUO.**(ref. article : h7270). fre.

[Benamara *et al.*, 2016] BENAMARA F., ASHER N., MATHIEU Y., POPESCU V. & CHARDON B. (2016). Evaluation in Discourse : a Corpus-Based Study. *Dialogue and Discourse*, **7**(1).

[Benamara *et al.*, 2017a] BENAMARA F., GROUIN C., KAROUI J., MORICEAU V. & ROBBA I. (2017a). Défi fouille de textes@taln/recital 2017. In *Actes de la 13e Défi Fouille de Texte*, p. à paraître, Orléans, France : Association pour le Traitement Automatique des Langues.

[Benamara *et al.*, 2014] BENAMARA F., MORICEAU V. & MATHIEU Y. Y. (2014). *TALN-RECITAL 2014 Workshop DEFT 2014 : DÉfi Fouille de Textes (DEFT 2014 Workshop : Text Mining Challenge)*, chapter Catégorisation sémantique fine des expressions d'opinion pour la détection de consensus, p. 36–44. Association pour le Traitement Automatique des Langues.

[Benamara *et al.*, 2017b] BENAMARA F., TABOADA M. & MATHIEU Y. Y. (2017b). Evaluative language beyond bags of words : Linguistic insights and computational applications. *Computational Linguistics*, **43**(1), 201–264.

[Bentivogli *et al.*, 2004] BENTIVOGLI L., FORNER P., MAGNINI B. & PIANTA E. (2004). Revising the wordnet domains hierarchy : semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Ressources*, p. 101–108 : Association for Computational Linguistics.

[Berntsen & Kennedy, 1996] BERNTSEN D. & KENNEDY J. M. (1996). Unresolved contradictions specifying attitudes—in metaphor, irony, understatement and tautology. *Poetics*, **24**(1), 13–29.

[Bertero & Fung, 2016] BERTERO D. & FUNG P. (2016). A long short-term memory framework for predicting humor in dialogues. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, San Diego, USA*.

[Bertero *et al.*, 2016] BERTERO D., FUNG P., LI X., WU L., LIU Z., HUSSAIN B., CHONG W. C., LAU K. M., YUE P. C., ZHANG W. *et al.* (2016). Deep learning of audio and language features for humor prediction. *Journal of Lightwave Technology*, **34**(2016), 1.

[Bestgen & Cabiaux, 2002] BESTGEN Y. & CABIAUX A.-F. (2002). L'analyse sémantique latente et l'identification des métaphores. In *Rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues*, p. 331—-337 : TALN.

[Bestgen & Lories, 2009] BESTGEN Y. & LORIES G. (2009). Un niveau de base pour la tâche 1 (corpus français et anglais) de deft'09. *Actes du cinquième DÉfi Fouille de Textes*, p.6̃5.

[Biadsy *et al.*, 2009] BIADSY F., HIRSCHBERG J. & HABASH N. (2009). Spoken arabic dialect identification using phonotactic modeling. In *Proceedings of the eacl 2009 workshop on computational approaches to semitic languages*, p. 53–61 : Association for Computational Linguistics.

[Bosco *et al.*, 2013] BOSCO C., PATTI V. & BOLIOLI A. (2013). Developing corpora for sentiment analysis : The case of irony and senti-tut. *IEEE Intelligent Systems*, **28**(2), 55–63.

[Boyd, 1979] BOYD R. (1979). *Metaphor and theory change : what is" metaphor" a metaphor for ¿. In A. Ortony (Ed.), Metaphor and Thought*. Cambridge : Cambridge University Press.

[Bres, 2010] BRES J. (2010). L'ironie, un cocktail dialogique ? In *2ème Congrès Mondial de Linguistique Française*, p. 046 : EDP Sciences.

[Burfoot & Baldwin, 2009] BURFOOT C. & BALDWIN C. (2009). Automatic satire detection : Are you having a laugh ? In *Proceedings of the ACL-IJCNLP 2009 conference short papers*, p. 161–164 : Association for Computational Linguistics.

[Burgers, 2010] BURGERS C. (2010). *Verbal irony : Use and effects in written discourse*. PhD thesis, Radboud Universiteit Nijmegen.

[Buschmeier *et al.*, 2014] BUSCHMEIER K., CIMIANO P. & KLINGER R. (2014). An impact analysis of features in a classification approach to irony detection in product reviews. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, p. 42–49.

[CAMPIGOTTO *et al.*, 2014] CAMPIGOTTO R., CONDE CÉSPEDES P. & GUILLAUME J.-L. (2014). La méthode de Louvain générique : un algorithme adaptatif pour la détection de communautés sur de très grands graphes. In *ROADEF* -

15ème congrès annuel de la Société française de recherche opérationnelle et d'aide à la décision, Bordeaux, France : Société française de recherche opérationnelle et d'aide à la décision.

[Carpuat *et al.*, 2012] CARPUAT M., MARTON Y. & HABASH N. (2012). Improved arabic-to-english statistical machine translation by reordering post-verbal subjects for word alignment. *Machine Translation*, **26**(1-2), 105–120.

[Carvalho *et al.*, 2009] CARVALHO P., SARMENTO L., SILVA M. J. & OLIVEIRA E. D. (2009). Clues for detecting irony in user-generated contents : oh... ! ! it's so easy ;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, p. 53–56 : ACM.

[Chardon *et al.*, 2013] CHARDON B., BENAMARA F., MATHIEU Y. Y., POPESCU V. & ASHER N. (2013). Measuring the effect of discourse structure on sentiment analysis. In *CICLing*, p. 25–37.

[Charniak & Johnson, 2005] CHARNIAK E. & JOHNSON M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, p. 173–180 : Association for Computational Linguistics.

[Choi & Cardie, 2008] CHOI Y. & CARDIE C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of EMNLP'08*, p. 793–801 : Association for Computational Linguistics.

[Clark & Gerrig, 1984] CLARK H. H. & GERRIG R. J. (1984). On the pretense theory of irony. *Journal of Experimental Psychology : General*, **113**(1), 121–126.

[Clift, 1999] CLIFT R. (1999). Irony in conversation. *Language in Society*, **28**, 523–553.

[Cohen, 1988] COHEN J. (1988). *Statistical power analysis for the behavior science (Second Edition)*. LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS.

[Colston & Keller, 1998] COLSTON H. L. & KELLER S. B. (1998). You'll never believe this : Irony and hyperbole in expressing surprise. *Journal of Psycholinguistic Research*, **27**(4), 499–513.

[Darwish, 2013] DARWISH K. (2013). Named entity recognition using cross-lingual resources : Arabic as an example. In *ACL (1)*, p. 1558–1567.

[Davidov & Rappoport, 2006] DAVIDOV D. & RAPPOPORT A. (2006). Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *Proceedings of the 21st International Conference on Computational Linguistics and*

the 44th annual meeting of the Association for Computational Linguistics, p. 297–304 : Association for Computational Linguistics.

[Davidov *et al.*, 2010] DAVIDOV D., TSUR O. & RAPPOPORT A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, p. 107–116.

[Davidson *et al.*, 2003] DAVIDSON R. J., SCHERER K. R. & GOLDSMITH H. H. (2003). *Handbook of Affective Sciences*. Oxford : Oxford University Press.

[Didio, 2007] DIDIO L. (2007). *Une approche sémantico-sémiotique de l'ironie*. PhD thesis, Université de Limoges.

[Do Dinh & Gurevych, 2016] DO DINH E.-L. & GUREVYCH I. (2016). Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, p. 28–33, San Diego, California : Association for Computational Linguistics.

[Eskander *et al.*, 2013] ESKANDER R., HABASH N. & RAMBOW O. (2013). Automatic extraction of morphological lexicons from morphologically annotated corpora. In *EMNLP*, p. 1032–1043.

[Farghaly *et al.*, 2003] FARGHALY A., SENELLART J. *et al.* (2003). Intuitive coding of the arabic lexicon. In *SYSTRAN, MT, Summit IX Workshop, Machine Translation for Semitic Languages : Issues and Approaches, Tuesday September*, volume 23 : Citeseer.

[Farias *et al.*, 2015] FARIAS D. I. H., SULIS E., PATTI V., RUFFO G. & BOSCO C. (2015). Valento : Sentiment analysis of figurative language tweets with irony and sarcasm. *SemEval-2015*, p. 694.

[Filatova, 2012] FILATOVA E. (2012). Irony and sarcasm : Corpus generation and analysis using crowdsourcing. In *LREC*, p. 392–398.

[Fromilhague, 1995] FROMILHAGUE C. (1995). Les figures de style. *Paris, Nathan, coll*, **128**, 1894–1899.

[Gedigian *et al.*, 2006] GEDIGIAN M., BRYANT J., NARAYANAN S. & CIRIC B. (2006). Catching metaphors. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding*, p. 41–48 : Association for Computational Linguistics.

[Ghosh *et al.*, 2015] GHOSH A., LI G., VEALE T., ROSSO P., SHUTOVA E., BARNDEN J. & REYES A. (2015). Semeval-2015 task 11 : Sentiment Analysis of Figurative Language in Twitter. In *Proceedings of SemEval 2015, Co-located with NAACL*, p. 470–478 : ACL.

[Gianti *et al.*, 2012] GIANTI A., BOSCO C., PATTI V., BOLIOLI A. & CARO L. D. (2012). Annotating irony in a novel italian corpus for sentiment analysis. In *Proceedings of the 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals, Istanbul, Turkey*, p. 1–7.

[Gibbs, 1994] GIBBS R. W. (1994). *The poetics of mind : Figurative thought, language, and understanding*. Cambridge University Press.

[Gibbs, 2000] GIBBS R. W. (2000). Irony in talk among friends. *Metaphor and symbol*, **15**(1-2), 5–27.

[Gonzalez-Ibanez *et al.*, 2011] GONZALEZ-IBANEZ R., MURESAN S. & WACHOLDE N. (2011). Identifying sarcasm in twitter : a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies : short papers-Volume 2*, p. 581–586 : Association for Computational Linguistics.

[Goode *et al.*, 2017] GOODE B. J., REYES J. I. M., PARDO-YEPEZ D. R., CANALE G. L., TONG R. M., MARES D., ROAN M. & RAMAKRISHNAN N. (2017). Time-series analysis of blog and metaphor dynamics for event detection. In *To appear in Advances in Cross-Cultural Decision Making*, p. 17–27. Springer.

[Graja *et al.*, 2011] GRAJA M., JAOUA M. & BELGUITH L. H. (2011). Building ontologies to understand spoken tunisian dialect. *CoRR*, **abs/1109.0624**.

[Green & Manning, 2010] GREEN S. & MANNING C. D. (2010). Better arabic parsing : Baselines, evaluations, and analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics*, p. 394–402 : Association for Computational Linguistics.

[Grice, 1970] GRICE H. P. (1970). *Logic and conversation*. na.

[Grice *et al.*, 1975] GRICE H. P., COLE P. & MORGAN J. L. (1975). Syntax and semantics. *Logic and conversation*, **3**, 41–58.

[Habash & Rambow, 2006] HABASH N. & RAMBOW O. (2006). Magead : a morphological analyzer and generator for the arabic dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, p. 681–688 : Association for Computational Linguistics.

[Habash & Rambow, 2007] HABASH N. & RAMBOW O. (2007). Morphophonemic and orthographic rules in a multi-dialectal morphological analyzer and generator for arabic verbs. In *International symposium on computer and arabic language (iscal), riyadh, saudi arabia*, volume 2006.

[Habash, 2010] HABASH N. Y. (2010). *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

[Haiman, 1998] HAIMAN J. (1998). *Talk is cheap : Sarcasm, alienation, and the evolution of language*. Oxford University Press on Demand.

[Haiman, 2001] HAIMAN J. (2001). *Talk is cheap : Sarcasm, alienation, and the evolution of language*. Oxford University Press, USA.

[Hammo *et al.*, 2002] HAMMO B., ABU-SALEM H. & LYTINEN S. (2002). Qarab : A question answering system to support the arabic language. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, p. 1–11 : Association for Computational Linguistics.

[Hatzivassiloglou & McKeown, 1997] HATZIVASSILOGLOU V. & MCKEOWN K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, p. 174–181 : Association for Computational Linguistics.

[Hee *et al.*, 2016] HEE C. V., LEFEVER E. & HOSTE V. (2016). Exploring the realization of irony in twitter data. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, S. GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France : European Language Resources Association (ELRA).

[Hertzler, 1970] HERTZLER J. O. (1970). *Laughter : A socio-scientific analysis*. Exposition Press.

[Huang, 2014] HUANG T.-H. K. (2014). Social metaphor detection via topical analysis. In *Sixth International Joint Conference on Natural Language Processing*, p.1̃4.

[Hunston & Thompson, 2000] S. HUNSTON & G. THOMPSON, Eds. (2000). *Evaluation in Text : Authorial Distance and the Construction of Discourse*. Oxford University Press.

[Hyungsuk *et al.*, 2003] HYUNGSUK J., PLOUX S. & WEHRLI E. (2003). Lexical knowledge representation with contexonyms. In *9th MT summit Machine Translation*, p. 194–201.

[Jang *et al.*, 2015a] JANG H., MOON S., JO Y. & ROSÉ C. P. (2015a). Metaphor detection in discourse. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, p. 384.

[Jang *et al.*, 2015b] JANG H., WEN M. & ROSÉ C. P. (2015b). Effects of situational factors on metaphor detection in an online discussion forum. *NAACL HLT 2015*, p.1̃.

[jie Tang & Chen, 2014] JIE TANG Y. & CHEN H.-H. (2014). Chinese Irony Corpus Construction and Ironic Structure Analysis. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*, p. 1269—-1278.

[Joshi *et al.*, 2015] JOSHI A., SHARMA V. & BHATTACHARYYA P. (2015). Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 2, p. 757–762.

[Joshi *et al.*, 2016] JOSHI A., TRIPATHI V., PATEL K., BHATTACHARYYA P. & CARMAN M. (2016). Are word embedding-based features useful for sarcasm detection ? *arXiv preprint arXiv :1610.00883*.

[Karoui, 2016] KAROUI J. (2016). Fric : Un corpus et un schéma d'annotation multi-niveaux pour l'ironie dans les tweets. *Traitement Automatique des Langues Naturelles (TALN)*.

[Karoui *et al.*, 2015a] KAROUI J., BENAMARA F., MORICEAU V., AUSSENAC-GILLES N. & BELGUITH L. H. (2015a). Towards a contextual pragmatic model to detect irony in tweets. In *Proceedings of ACL-IJCNLP 2015, Volume 2 : Short Papers*, p. 644–650 : The Association for Computer Linguistics.

[Karoui *et al.*, 2017] KAROUI J., BENAMARA F., MORICEAU V., PATTI V., BOSCO C. & AUSSENAC-GILLES N. (2017). Exploring the impact of pragmatic phenomena on irony detection in tweets : A multilingual corpus study. In *Proceedings of the 15th edition of the European Chapter of the Association for Computational Linguistics Conference (EACL)*.

[Karoui *et al.*, 2015b] KAROUI J., BENAMARA ZITOUNE F., MORICEAU V., AUSSENAC-GILLES N. & HADRICH BELGUITH L. (2015b). Détection automatique de l'ironie dans les tweets en français. In *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles*, p. 460–465, Caen, France : Association pour le Traitement Automatique des Langues.

[Karoui *et al.*, 2013] KAROUI J., GRAJA M., BOUDABOUS M. & BELGUITH L. H. (2013). Domain ontology construction from a tunisian spoken dialogue corpus. *Proc. ICWIT*.

[Kerbrat-Orecchioni, 1976] KERBRAT-ORECCHIONI C. (1976). Problèmes de l'ironie. *Linguistique et sémiologie*, **2**, 10–46.

[Keskes *et al.*, 2014] KESKES I., ZITOUNE F. B. & BELGUITH L. H. (2014). Learning explicit and implicit arabic discourse relations. *Journal of King Saud University-Computer and Information Sciences*, **26**(4), 398–416.

[Kintsch, 2000] KINTSCH W. (2000). Metaphor comprehension : A computational theory. *Psychonomic bulletin & review*, **7**(2), 257–266.

[Kittay, 1990] KITTAY E. F. (1990). *Metaphor : Its cognitive force and linguistic structure*. Oxford University Press.

[Kreuz, 1996] KREUZ R. J. (1996). The use of verbal irony : Cues and constraints. *Metaphor : Implications and applications*, p. 23–38.

[Kreuz & Caucci, 2007] KREUZ R. J. & CAUCCI G. M. (2007). Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on computational approaches to Figurative Language*, p. 1–4 : Association for Computational Linguistics.

[Kreuz & Glucksberg, 1989] KREUZ R. J. & GLUCKSBERG S. (1989). How to be sarcastic : The echoic reminder theory of verbal irony. *Journal of Experimental Psychology : General*, **118**(4), 374.

[Kreuz & Roberts, 1993] KREUZ R. J. & ROBERTS R. M. (1993). The empirical study of figurative language in literature. *Poetics*, **22**(1), 151–169.

[Kreuz & Roberts, 1995] KREUZ R. J. & ROBERTS R. M. (1995). Two cues for verbal irony : Hyperbole and the ironic tone of voice. *Metaphor and symbol*, **10**(1), 21–31.

[Kumon-Nakamura *et al.*, 1995] KUMON-NAKAMURA S., GLUCKSBERG S. & BROWN M. (1995). How about another piece of pie : The allusional pretense theory of discourse irony. *Journal of Experimental Psychology : General*, **124**(1), 3.

[Leech, 2016] LEECH G. N. (2016). *Principles of pragmatics*. Routledge.

[Létourneau & Bélanger, 2009] LÉTOURNEAU D. & BÉLANGER M. (2009). Impacts de la variation du nombre de traits discriminants sur la catégorisation des documents. *Actes du cinquième DÉfi Fouille de Textes*, p.7̃7.

[Liebrecht *et al.*, 2013] LIEBRECHT C., KUNNEMAN F. & VAN DEN B. A. (2013). The perfect solution for detecting sarcasm in tweets# not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, p. 29–37 : New Brunswick, NJ : ACL.

[Littman & Turney, 2002] LITTMAN T. & TURNEY P. (2002). *Unsupervised learning of semantic orientation from a hundred-billion-word corpus*. Rapport interne, Technical Report ERB-1094, National Research Council Canada.

[Liu, 2012] LIU B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, **5**(1), 1–167.

[Liu, 2015] LIU B. (2015). *Sentiment Analysis : Mining Opinions, Sentiments, and Emotions*. Cambridge : Cambridge University Press.

[Losada & Crestani, 2016] LOSADA D. E. & CRESTANI F. (2016). A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings*, p. 28–39.

[Lucariello, 1994] LUCARIELLO J. (1994). Situational irony : A concept of events gone awry. *Journal of Experimental Psychology : General*, **123**(2), 129.

[Maamouri *et al.*, 2006] MAAMOURI M., BIES A. & KULICK S. (2006). Diacritization : A challenge to arabic treebank annotation and parsing. In *Proceedings of the Conference of the Machine Translation SIG of the British Computer Society*.

[Macwhinney & Fromm, 2014] MACWHINNEY B. & FROMM D. (2014). Two approaches to metaphor detection. In N. C. C. CHAIR), K. CHOUKRI, T. DECLERCK, H. LOFTSSON, B. MAEGAARD, J. MARIANI, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland : European Language Resources Association (ELRA).

[Marton *et al.*, 2013] MARTON Y., HABASH N. & RAMBOW O. (2013). Dependency parsing of modern standard arabic with lexical and inflectional features. *Computational Linguistics*, **39**(1), 161–194.

[Masmoudi *et al.*, 2014] MASMOUDI A., KHEMAKHEM M. E., ESTÈVE Y., BOUGARES F., DABBAR S. & BELGUITH L. H. (2014). Phonétisation automatique du dialecte tunisien. *30ème Journée d'études sur la parole, Le Mans-France*.

[Maynard & Greenwood, 2014] MAYNARD D. & GREENWOOD M. A. (2014). Who cares about sarcastic tweets ? investigating the impact of sarcasm on sentiment analysis. In *LREC*, p. 4238–4243.

[Mercier-Leca, 2003] MERCIER-LECA F. (2003). *L'ironie*. Hachette supérieur.

[Mihalcea & Strapparava, 2005] MIHALCEA R. & STRAPPARAVA C. (2005). Making computers laugh : Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, p. 531–538 : Association for Computational Linguistics.

[Mihalcea & Strapparava, 2006] MIHALCEA R. & STRAPPARAVA C. (2006). Learning to laugh (automatically) : Computational models for humor recognition. *Computational Intelligence*, **22**(2), 126–142.

[Miller, 1995] MILLER G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, **38**(11), 39–41.

[Mpouli & Ganascia, 2015] MPOULI S. & GANASCIA J.-G. (2015). Extraction et analyse automatique des comparaisons et des pseudo-comparaisons pour la détection des comparaisons figuratives. In *22e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2015)*.

[Muecke, 1978] MUECKE D. C. (1978). Irony markers. *Poetics*, **7**(4), 363–375.

[Nadaud & Zagaroli, 2008] NADAUD B. & ZAGAROLI K. (2008). *Surmonter ses complexes : les comprendre pour les assumer*. Editions Eyrolles.

[Niogret, 2004] NIOGRET P. (2004). *Les figures de l'ironie dans A la recherche du temps perdu de Marcel Proust*. Editions L'Harmattan.

[Oliveira & Ploux, 2009] OLIVEIRA I. & PLOUX S. (2009). Vers une méthode de détection et de traitement automatique de la métaphore. In *Passeurs de mots, passeurs d'espoir. Actes des 8èmes Journées scientifiques du Réseau LTT*, p. 1–11.

[Oudah & Shaalan, 2012] OUDAH M. & SHAALAN K. F. (2012). A pipeline arabic named entity recognition using a hybrid approach. In *Coling*, p. 2159–2176.

[Ounis *et al.*, 2008] OUNIS I., MACDONALD C. & SOBOROFF I. (2008). *Overview of the TREC-2008 blog track*. Rapport interne, DTIC Document.

[Pang & Lee, 2004] PANG B. & LEE L. (2004). A sentimental education : Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, p. 271 : Association for Computational Linguistics.

[Pang *et al.*, 2002] PANG B., LEE L. & VAITHYANATHAN S. (2002). Thumbs up ? : sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, p. 79–86 : Association for Computational Linguistics.

[Pasha *et al.*, 2014] PASHA A., AL-BADRASHINY M., DIAB M. T., EL KHOLY A., ESKANDER R., HABASH N., POOLEERY M., RAMBOW O. & ROTH R. (2014). Madamira : A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *LREC*, volume 14, p. 1094–1101.

[Péry-Woodley *et al.*, 2009] PÉRY-WOODLEY M.-P., ASHER N., ENJALBERT P., BENAMARA F., BRAS M., FABRE C., FERRARI S., HO-DAC L.-M., LE DRAOULEC A., MATHET Y. *et al.* (2009). Annodis : une approche outillée de l'annotation de structures discursives. In *TALN 2009 (Conférence sur le Traitement Automatique des Langues Naturelles)*, p. paper_TALN_52.

[Polanyi & Zaenen, 2006] POLANYI L. & ZAENEN A. (2006). Contextual valence shifters. In *Computing Attitude and Affect in Text : Theory and Applications*, The Information Retrieval Series, p. 1–10 : Springer-Verlag.

[Politis, 2002] POLITIS H. (2002). *Kierkegaard*. Ellipses.

[Pougeoise, 2001] POUGEOISE M. (2001). *Dictionnaire de rhétorique*. Armand Colin.

[Purandare & Litman, 2006] PURANDARE A. & LITMAN D. (2006). Humor : prosody analysis and automatic recognition for f* r* i* e* n* d* s. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, p. 208–215 : Association for Computational Linguistics.

[Radev *et al.*, 2015] RADEV D., STENT A., TETREAULT J., PAPPU A., ILIAKOPOULOU A., CHANFREAU A., DE JUAN P., VALLMITJANA J., JAIMES A., JHA R. *et al.* (2015). Humor in collective discourse : Unsupervised funniness detection in the new yorker cartoon caption contest. *arXiv preprint arXiv :1506.08126*.

[Raeber, 2011] RAEBER T. (2011). L'ironie ; réactualisation de pensée et contenus non posés : une approche pragmatique. Master's thesis, Université de Neuchâtel.

[Raz, 2012] RAZ Y. (2012). Automatic humor classification on twitter. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies : Student Research Workshop*, p. 66–70 : Association for Computational Linguistics.

[Reboul, 1991] REBOUL O. (1991). *Introduction à la rhétorique : théorie et pratique*. Presses universitaires de France.

[Reyes & Rosso, 2011] REYES A. & ROSSO P. (2011). Mining subjective knowledge from customer reviews : a specific case of irony detection. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, p. 118–124 : Association for Computational Linguistics.

[Reyes & Rosso, 2012] REYES A. & ROSSO P. (2012). Making objective decisions from subjective data : Detecting irony in customer reviews. *Decision Support Systems*, **53**(4), 754–760.

[Reyes & Rosso, 2014] REYES A. & ROSSO P. (2014). On the difficulty of automatically detecting irony : beyond a simple case of negation. *Knowledge and Information Systems*, **40**(3), 595–614.

[Reyes *et al.*, 2009] REYES A., ROSSO P. & BUSCALDI D. (2009). Humor in the blogosphere : First clues for a verbal humor taxonomy. *Journal of Intelligent Systems*, **18**(4), 311–332.

[Reyes *et al.*, 2013] REYES A., ROSSO P. & VEALE T. (2013). A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, **47**(1), 239–268.

[Rick & Loewenstein, 2008] RICK S. & LOEWENSTEIN G. (2008). The role of emotion in economic behavior. In M. LEWIS, J. M. HAVILAND-JONES & L. F. BARRETT, Eds., *Handbook of Emotions*, p. 138–156. Guilford.

[Riffaterre, 1969] RIFFATERRE M. (1969). La métaphore filée dans la poésie surréaliste. *Langue française*, **3**(1), 46–60.

[Riloff *et al.*, 2013] RILOFF E., QADIR A., SURVE P., SILVA L. D., GILBERT N. & HUANG R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, p. 704–714.

[Ritchie, 2005] RITCHIE D. (2005). Frame-shifting in humor and irony. *Metaphor and Symbol*, **20**(4), 275–294.

[Rouvier *et al.*, 2015] ROUVIER M., FAVRE B. & ANDIYAKKAL RAJENDRAN B. (2015). Talep @ deft'15 : Le plus cooool des systèmes d'analyse de sentiment. In *Actes de la 11e Défi Fouille de Texte*, p. 97–103, Caen, France : Association pour le Traitement Automatique des Langues.

[Roze *et al.*, 2012] ROZE C., DANLOS L. & MULLER P. (2012). Lexconn : A french lexicon of discourse connectives. *Discours, Multidisciplinary Perspectives on Signalling Text Organisation*, **10**, (on line).

[Ryding, 2005] RYDING K. C. (2005). *A reference grammar of modern standard Arabic*. Cambridge university press.

[Sadat & Mohamed, 2013] SADAT F. & MOHAMED E. (2013). Improved arabic-french machine translation through preprocessing schemes and language analysis. In *Canadian Conference on Artificial Intelligence*, p. 308–314 : Springer.

[Saif *et al.*, 2016] SAIF M., MOHAMMAD S. & SVETLANA K. (2016). Sentiment lexicons for arabic social media. In N. CALZOLARI, K. CHOUKRI, T. DECLERCK, S.

GOGGI, M. GROBELNIK, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Eds., *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.* : European Language Resources Association (ELRA).

[Searle, 1979] SEARLE J. (1979). *Expression and meaning : Studies in the theory of speech acts*. Cambridge University.

[Seto, 1998] SETO K.-I. (1998). On non-echoic irony. *Relevance Theory : Applications and Implications*, **37**, 239.

[Shaikh *et al.*, 2007] SHAIKH M. A., PRENDINGER H. & MITSURU I. (2007). Assessing sentiment of text by semantic dependency and contextual valence analysis. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, ACII, p. 191–202 : Springer-Verlag.

[Shelley, 2001] SHELLEY C. (2001). The bicoherence theory of situational irony. *Cognitive Science*, **25**(5), 775–818.

[Shutova *et al.*, 2013] SHUTOVA E., DEVEREUX B. J. & KORHONEN A. (2013). Conceptual metaphor theory meets the data : a corpus-based human annotation study. *Language resources and evaluation*, **47**(4), 1261–1284.

[Simédoh, 2012] SIMÉDOH V. (2012). *L'humour et l'ironie en littérature francophone subsaharienne : des enjeux critiques à une poétique du rire*. Peter Lang.

[Sjöbergh & Araki, 2007] SJÖBERGH J. & ARAKI K. (2007). Recognizing humor without recognizing meaning. In *International Workshop on Fuzzy Logic and Applications*, p. 469–476 : Springer.

[Song, 1998] SONG N. S. (1998). Metaphor and metonymy. *Relevance theory : Applications and implications*, p. 87–104.

[Sperber & Wilson, 1981] SPERBER D. & WILSON D. (1981). Irony and the use-mention distinction. *Radical pragmatics*, **49**, 295–318.

[Stenetorp *et al.*, 2012] STENETORP P., PYYSALO S., TOPIĆ G., OHTA T., ANANIADOU S. & TSUJII J. (2012). Brat : a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 102–107 : Association for Computational Linguistics.

[Su *et al.*, 2016] SU C., HUANG S. & CHEN Y. (2016). Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*.

[Sulis *et al.*, 2016a] SULIS E., FARÍAS D. I. H., ROSSO P., PATTI V. & RUFFO G. (2016a). Figurative messages and affect in twitter : Differences between #irony, #sarcasm and #not. *Knowl.-Based Syst.*, **108**, 132–143.

[Sulis *et al.*, 2016b] SULIS E., HERNÁNDEZ FARÍAS D. I., ROSSO P., PATTI V. & RUFFO G. (2016b). Figurative messages and affect in twitter : Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*. Available on line, In press.

[Taboada *et al.*, 2011] TABOADA M., BROOKE J., TOFILOSKI M., VOLL K. & STEDE M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, **37**, 267–307.

[Taylor, 2009] TAYLOR J. M. (2009). Computational detection of humor : A dream or a nightmare ? the ontological semantics approach. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 03*, p. 429–432 : IEEE Computer Society.

[Tayot, 1984] TAYOT C. (1984). *L'ironie*. PhD thesis, Claude Bernard University (Lyon).

[Toprak & Gurevych, 2009] TOPRAK C. & GUREVYCH I. (2009). Document level subjectivity classification experiments in deft'09 challenge. *Actes du cinquième DÉfi Fouille de Textes*, p.$\tilde{9}$1.

[Tsur *et al.*, 2010] TSUR O., DAVIDOV D. & RAPPOPORT A. (2010). Icwsm-a great catchy name : Semi-supervised recognition of sarcastic sentences in online product reviews. In *ICWSM*.

[Tsvetkov *et al.*, 2014] TSVETKOV Y., BOYTSOV L., GERSHMAN A., NYBERG E. & DYER C. (2014). Metaphor detection with cross-lingual model transfer. *Proceedings of ACL, Baltimore, MD*.

[Turney, 2002] TURNEY P. D. (2002). Thumbs up or thumbs down ? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of Meeting of the Association for Computational Linguistics*, ACL, p. 417–424 : Association for Computational Linguistics.

[Utsumi, 1996] UTSUMI A. (1996). A unified theory of irony and its computational formalization. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, p. 962–967 : Association for Computational Linguistics.

[Utsumi, 2000] UTSUMI A. (2000). Verbal irony as implicit display of ironic environment : Distinguishing ironic utterances from nonirony. *Journal of Pragmatics*, **32**(12), 1777–1806.

[Utsumi, 2004] UTSUMI A. (2004). Stylistic and contextual effects in irony processing. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, p. 1369–1374.

[van de Gejuchte, 1993] VAN DE GEJUCHTE I. (1993). L'humour comme discours. *Revue de l'Institut de sociologie*, (1-4), 399–411.

[Van Hee *et al.*, 2015] VAN HEE C., LEFEVER E. & HOSTE V. (2015). *Guidelines for Annotating Irony in Social Media Text*. Rapport interne, Language and Translation Technology Team.

[Van Hee *et al.*, 2016] VAN HEE C., LEFEVER E. & HOSTE V. (2016). Exploring the realization of irony in twitter data. In *Language Resources and Evaluation Conference* : European Language Resources Association (ELRA).

[Veale & Hao, 2010] VEALE T. & HAO Y. (2010). Detecting ironic intent in creative comparisons. In *ECAI*, volume 215, p. 765–770.

[Voas, 2014] VOAS D. (2014). Towards a sociology of attitudes. *Sociological Research Online*, **19**(1), 12.

[Wallace, 2015] WALLACE B. C. (2015). Computational irony : A survey and new perspectives. *Artificial Intelleligence Review*, **43**(4), 467–483.

[Wallace *et al.*, 2015] WALLACE B. C., CHOE D. K. & CHARNIAK E. (2015). Sparse, contextually informed models for irony detection : Exploiting user communities, entities and sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, ACL-IJCNLP 2015, p. 1035–1044 : Association for Computational Linguistics.

[Wallace *et al.*, 2014] WALLACE B. C., DO KOOK CHOE L. K., KERTZ L. & CHARNIAK E. (2014). Humans require context to infer ironic intent (so computers probably do, too). In *ACL (2)*, p. 512–516.

[Wen *et al.*, 2013] WEN M., ZHENG Z., JANG H., XIANG G. & ROSÉ C. P. (2013). Extracting events with informal temporal references in personal histories in online communities. In *ACL (2)*, p. 836–842.

[Whissell, 1989] WHISSELL C. (1989). The dictionary of affect in language. *Emotion : Theory, research, and experience*, **4**(113-131), 94.

[Wiebe *et al.*, 2005] WIEBE J., WILSON T. & CARDIE C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, **39**(2-3), 165–210.

[Wilks, 1978] WILKS Y. (1978). Making preferences more active. *Artificial Intelligence*, **11**(3), 197–223.

[Wilson & Sperber, 1986] WILSON D. & SPERBER D. (1986). *Relevance : Communication and cognition*. Mass.

[Wilson & Sperber, 1988] WILSON D. & SPERBER D. (1988). Representation and relevance. *Mental representations : The interface between language and reality*, p. 133–153.

[Wilson & Sperber, 1992] WILSON D. & SPERBER D. (1992). On verbal irony. *Lingua*, **87**(1), 53–76.

[Wilson & Sperber, 2004] WILSON D. & SPERBER D. (2004). Relevance theory. In *Handbook of Pragmatics*, p. 607–632.

[Yamanashi, 1998] YAMANASHI M.-A. (1998). Some issues in the treatment of irony and related tropes. *Relevance Theory : Applications and implications*, **37**, 271.

[Yang *et al.*, 2015] YANG D., LAVIE A., DYER C. & HOVY E. (2015). Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 2367–2376 : Citeseer.

[Zagibalov *et al.*, 2010] ZAGIBALOV T., BELYATSKAYA K. & CARROLL J. (2010). Comparable english-russian book review corpora for sentiment analysis. In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, p. 67–72.