COGNITIVE SCIENCE SERIES

Automatic Detection of Irony

Opinion Mining in Microblogs and Social Media

Jihen Karoui, Farah Benamara
and Véronique Moriceau

iSTE

WILEY

Automatic Detection of Irony

Series Editor
Patrick Paroubek

Automatic Detection of Irony

Opinion Mining in Microblogs and Social Media

Jihen Karoui
Farah Benamara
Véronique Moriceau

iSTE

WILEY

First published 2019 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George’s Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Control Number: 2019944081

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-78630-399-8

Contents

Preface

Sentiment analysis is an extremely active area of research in natural language processing (NLP). A wide range of textual resources containing opinions are currently available online, including Internet user opinions, forums, social networks and consumer surveys. Given this abundance of data, automating the synthesis of multiple opinions is crucial to obtain an overview of opinions on any given subject. These types of summaries are of considerable interest to companies in assessing customer reactions to their products, and for consumers themselves when making decisions about a future purchase, trip, etc.

A significant number of papers have been published in this field since the early 2000s, making sentiment analysis one of the most attractive applications in NLP. Overall, the existing systems have achieved good results in terms of automatic classification of documents by subjective or objective type. However, the results for polarity analysis (which consists of classifying a document on a scale of subjectivity from most positive to most negative) remain inconclusive. The main reason for this failure is that current algorithms are unable to understand all of the subtleties of human language. These include figurative language, in which figures of speech are used to convey non-literal meaning, i.e. a meaning that is not strictly the conventional or intended sense of the individual words in an 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 book focuses on irony and sarcasm detection in social media content.

To this end, we propose a supervised learning approach to predict whether or not a tweet is ironic. For this purpose, we followed a three-step approach. In the first step, drawing on linguistic studies, we investigated the pragmatic phenomena used to express irony to define a multilevel annotation scheme for irony. This annotation scheme was applied as part of an annotation campaign for a corpus of 2000 French tweets. In a second step, exploiting all of the observations made from the annotated corpus, we developed an automatic detection model for French tweets, using the internal context of the tweet (lexical and semantic features) and external context (information available online). Finally, in the third step, we studied the portability of the model for irony detection tasks 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 Arabic. The results that we obtained for this extremely complex task are very encouraging, and merit further investigation to improve polarity detection in sentiment analysis.

Jihen KAROUI
Farah BENAMARA
Véronique MORICEAU
July 2019

Introduction

I.1. Context and purpose

In recent years, the Internet has become a truly essential information source due to the quantity and variety of textual content available, notably, for our purposes, expressing user's opinions. This content takes a range of forms, from blogs to comments, forums, social networks, reactions or reviews, increasingly centralized by search engines. Given the wealth of data and range of sources, there is a clear need for tools to extract, synthesize and compare extracted opinions. Tools of this type present a considerable interest for companies looking for client's feedback on their products or brand image, and for consumers seeking guidance concerning a planned purchase, outing or trip. These tools are also of value for survey groups in evaluating market reactions to a product, for predicting the results of future elections, etc.

This is the context in which sentiment analysis, or opinion mining, emerged as a new area of research. The first work on automatic opinion extraction dates from the late 1990s, notably with Hatzivassiloglou and McKeown's (1997) work on determining adjective polarity, and the work on document classification according to polarity (positive or negative) presented in (Pang *et al*. 2002) and (Littman and Turney 2002). The number of publications in this subject has increased considerably since the early 2000s, and opinion extraction is one of the most active areas in natural language processing (NLP) (Liu 2015, Benamara *et al*. 2017). Several evaluation campaigns have also been carried out in this area, including the TREC (Text Retrieval Conference) (Ounis *et al*. 2008); the DEFT (*Défi fouille de textes*) data mining challenge in French, run for the first time in 2005 (Azé and Roche 2005); and the SemEval (Semantic Evaluation) campaign, started in 1998[1].

Overall, existing systems have produced good results in terms of subjectivity analysis, which consists of determining whether a portion of text carries an opinion (i.e. is subjective) or simply presents facts (i.e. is objective) (Turney 2002). For example, subjectivity lexicons, used alongside classification techniques where applicable, can be used to detect the fact that the author expresses a positive opinion of the Prime Minister in phrase (I.1) via the use of the positive-polarity adjective *excellent*:

(I.1) The Prime Minister gave an excellent speech.

However, for polarity analysis tasks, which consist of determining the overall polarity and/or opinion score actually contained in a portion of text that is known to be subjective, the results of existing systems remain inconclusive. The three examples given below, adapted from Benamara (2017), highlight the complexity of the task:

(I.2) [I bought a second-hand iPhone 5s three months ago.]$_{P1}$ [The image quality is *exceptional*.]$_{P2}$ [However, the safety glass protector is *of poor quality*]$_{P3}$ [and the battery *gave out after 2 weeks* !!]$_{P4}$

Phrase (I.2) contains four propositions, separated by square brackets. Only the first three propositions express opinions (shown in blue). Of these, the first two are explicit, i.e. can be identified based on words, symbols or subjective language, such as the adjective *exceptional*. The third is implicit, based on words or groups of words that describe a situation (fact or state), which is judged as desirable or undesirable based on cultural and/or pragmatic knowledge shared by the writer and readers.

In phrases (I.3) and (I.4), the author uses figurative language to express an opinion, a further element that complicates polarity analysis. The phrases in question express negative opinions, but the authors use positive language (love, thank you, fantastic):

[1] www.senseval.org/.

(I.3) I love the way your product breaks just when I need to use it.

(I.4) Thanks again, SNCF. What a fantastic start to the day, yet again.

Implicit opinions can sometimes be expressed using irony, presenting a further issue for polarity analysis. In tweet (I.5), translated from an entry in the French Irony Corpus (FrIC) (Karoui 2016), the user makes a false assertion (underlined text), which makes the message extremely negative with regard to Manuel Valls. Note the use of the #ironie (irony) hashtag, which helps readers to realize that the message is intended ironically:

(I.5) #Valls learned that #Sarkozy had his phones tapped from the papers. Good job he isn't Minister of the Interior #ironie

It would be extremely straightforward for a human reader to extract opinions from these examples, but for a computer program, the task is extremely complex. Current systems do not have the capacity to account for the context in which opinions are expressed, making it very difficult to distinguish between explicit/implicit opinions or identify figurative language, above and beyond the basic determination of subjective expressions.

Our aim in this book is to contribute to the development of automatic detection of figurative language, a linguistic phenomenon that is particularly widespread in social media content. This phenomenon has attracted significant attention within the field of NLP in recent years, essentially due to its importance for improving the performance of opinion analysis systems (Maynard and Greenwood 2014, Ghosh *et al*. 2015).

I.2. Figurative language: the basics

Unlike literal language, figurative language "twists" the inherent meaning of words to create a figurative or illustrative meaning, using metaphor, irony, sarcasm, satire, humor, etc. Irony is a complex phenomenon that has been studied at length by philosophers and linguists (Grice *et al*. 1975, Sperber and Wilson 1981, Utsumi 1996). Generally speaking, irony is defined as a rhetorical figure in which one expresses the opposite of what the listener or reader is meant to understand (see phrases (I.3) and (I.4)). In computational linguistics, irony is a generic term used to denote a group of figurative phenomena including sarcasm, although the latter features increased bitterness and aggression (Clift 1999).

Each type of figurative language possesses its own linguistic mechanisms that enable us to understand the figurative meaning. These include the inversion of reality/truth to express irony (Grice *et al*. 1975), the use of funny effects to express humor (Van de Gejuchte 1993, Nadaud and Zagaroli 2008), etc. In most cases, the context of the utterance is essential to correctly interpret the intended figurative meaning. As such, it is essential to find a means of inferring information above and beyond the lexical, syntactic and even semantic aspects of a text. These inferences may vary depending on speaker profile (e.g. gender) and the cultural context.

The majority of work on irony detection in NLP is based on corpora of tweets, as authors have the possibility to indicate the ironic character of their messages using specific hashtags: #sarcasm, #irony, #humor, etc. These hashtags can then be used to collect a manually annotated corpus, an essential resource for supervised classification of tweets into ironic or non-ironic groups. The majority of current publications concern tweets in English, but work has also been carried out on detecting irony and/or sarcasm in Italian, Chinese and Dutch (Farias *et al*. 2015, Jie Tang and Chen 2014, Liebrecht *et al*. 2013).

Overall, the vast majority of proposed approaches rely exclusively on the linguistic content of tweets. Two main types of cues have been used:

– Lexical cues (n-grams, word count, presence of opinion words or expressions of emotion) and/or stylistic cues (emoticons, interjections, quotations, use of slang, word repetition) (Kreuz and Caucci 2007, Burfoot and Baldwin 2009, Tsur *et al*. 2010, Gonzalez-Ibanez *et al*. 2011, Gianti *et al*. 2012, Liebrecht *et al*. 2013, Reyes *et al*. 2013, Barbieri and Saggion 2014b).

– Pragmatic cues, used to capture the context required to infer irony. These cues are also extracted from the linguistic content of messages, such as brusque changes in verb tense, the use of semantically distant terms or the use of common versus rare words (Burfoot and Baldwin 2009, Reyes *et al*. 2013, Barbieri and Saggion 2014b).

The results obtained using these approaches are promising[2]. However, we feel that although approaches of this type are essential, they are only the first step in the process; we need to go further, taking more pragmatic approaches in order to infer the extra-linguistic context required to understand this complex phenomenon.

I.3. Contributions

In this context, we have chosen to focus on French-language tweets for the first time, proposing a supervised learning approach in order to predict whether a tweet is ironic. Our contributions concern three main elements:

1) The development of a conceptual model used to identify pragmatic phenomena employed to express irony in Twitter messages. Drawing on work on irony in the field of linguistics, we developed a first multilevel annotation schema for irony. This schema, introduced at the ColTal@TALN2016 workshop, was applied as part of an annotation campaign for a corpus of 2,000 tweets (Karoui 2016). An expanded version of this corpus was used as training data for the first DEFT@TALN 2017 evaluation campaign for opinion analysis and figurative language[3] (Benamara *et al*. 2017). The annotation schema, along with the quantitative and qualitative results of the annotation campaign, are described in Chapter 3.

2) The development of a computational model to infer the pragmatic context required in order to detect irony. Using all of the observations obtained from the annotated corpus, we developed a model for automatic irony detection in tweets in French, using both the internal context of the tweet, via lexical and semantic features, and the external context, using information obtained from reliable external sources. Our model is notably able to detect irony expressed through the use of false assertions (see phrase (I.5)). This model, presented at TALN 2015 (Karoui *et al*. 2015) and ACL 2015, is presented in Chapter 4.

3) An investigation of the portability of both the conceptual and computational model for irony detection in a multilinugal context. We began by testing the portability of our annotation schema for tweets in Italian and

[2] For example, (Reyes *et al*. 2013) obtained a precision of 79% for tweets in English. See Chapter 2 for a detailed state of the art and a presentation of the results of existing approaches.

[3] https//deft.limsi.fr/2017/.

English, two Indo-European languages which are culturally similar to French. Our results, presented at EACL 2017, show that our scheme is perfectly applicable to these languages (Karoui *et al.* 2017). We then tested the portability of our computational model for Arabic, with tweets written both in standard Arabic and dialectal Arabic. Once again, the results showed that our model continues to perform well even with a language from a different family. The portability of our models is discussed in detail in Chapter 5.

In Chapters 1 and 2, we shall provide a full review of the state of the art of linguistic and computational approaches to irony detection, before going into detail concerning our specific contributions in Chapters 3 to 5. We conclude with a summary of the results obtained, and with a presentation of several directions for future research.

1

From Opinion Analysis to Figurative Language Treatment

1.1. Introduction

The first work on automatic opinion extraction (or opinion mining) dates back to the late 1990s, notably to (Hatzivassiloglou and McKeown 1997) seminal work on determining adjectival polarity in documents, i.e. identifying the positive or negative character of opinions expressed by these adjectives, and to (Pang *et al*. 2002) and (Littman and Turney 2002) work on classifying documents according to polarity.

Work on this subject has been in progress since the 2000s, and opinion extraction is one of the most active areas in both NLP and data mining, with over 26,000 publications identified by Google Scholar. Notable examples include (Wiebe *et al*. 2005) work on annotating the multi-perspective question answering (MPQA) opinion corpus, (Taboada *et al*. 2011) work on the effects of opinion operators, such as intensifiers, modalities and negations, and (Asher *et al*. 2009) and (Chardon *et al*. 2013) work on the use of the discursive structure in calculating the overall opinion expressed in a document. Finally, we note the emergence of a number of evaluation campaigns, such as the Text Retrieval Conference (TREC) (Ounis *et al*. 2008), the DEFT (*Défi fouille de textes*, data mining challenge) in French run

Automatic Detection of Irony: Opinion Mining in Microblogs and Social Media,
First Edition. Jihen Karoui; Farah Benamara and Véronique Moriceau.

for the first time in 2005 (Azé and Roche 2005), and the SemEval (Semantic Evaluation) campaign, started in 1998[1].

It is important to note that opinion analysis was already a subject of study in other domains, such as linguistics (Hunston and Thompson 2000), psychology (Davidson *et al*. 2009), sociology (Voas 2014) and economics (Rick and Loewenstein 2008) before it attracted the attention of computer scientists. Opinion analysis is a multidisciplinary domain that draws on a wide range of tools and techniques, as we shall see throughout this chapter.

The development of opinion analysis systems is no simple matter, and there are several different challenges that must be met: identifying portions of text that provide the opinions a user is looking for; evaluating the quality of opinions obtained in this way – positive, negative, etc.; presenting results to users in a relevant way; etc.

Most existing approaches are based on word-level lexical analysis, sometimes combined with phrase-level syntactic analysis to identify operators and calculate their effects on opinion words (Liu 2012). Evidently, this type of analysis is far from sufficient to take account of the full linguistic complexity of opinion expressions. Fine, or pragmatic, semantic analysis is therefore crucial, particularly when treating complex phenomena such as figurative language, the focus of our study.

The aim of this chapter is to provide a brief introduction to the field of opinion analysis and to establish key definitions relating to the notion of figurative language. Our overview makes no claim to be exhaustive, given the extent of the field of research in question. Readers interested in going further may wish to consult the excellent summaries found in (Liu 2015) and (Benamara *et al*. 2017).

This chapter begins with a presentation of the notion of opinion and of the main approaches used in the literature (section 1.2). In section 1.3, we present the main limitations of existing systems, focusing on the use of figurative language. Section 1.4 deals with this type of language, looking at four figurative phenomena: irony, sarcasm, satire and humor. Finally, we shall

[1] www.senseval.org/.

discuss the main challenges encountered in NLP in terms of automatic detection of figurative language.

1.2. Defining the notion of opinion

1.2.1. *The many faces of opinion*

In NLP, the word opinion is used as a generic term to denote a range of subjective expressions such as sentiments, attitudes, points of view, judgments and desires. The most widely used definition is as follows (Benamara 2017):

> "An *opinion* is a **subjective expression** of language which an **emitter** (a person, institution, etc.) uses to judge or evaluate a **subject** (an object, person, action, event, etc.), positioning it on a **polarized scale** in relation to a social norm (such as an aesthetic judgment) or a moral norm (such as the distinction between good and bad)".

Phrase (1.1) is a good illustration of this definition. The author expresses a positive opinion of the dishes served in the restaurant using a positive-polarity verb (to love). In opinion analysis, the ability to distinguish between subjective or objective expressions is key. Phrase (1.2) does not express an opinion, but a purely factual event.

(1.1) I loved the dishes served in this restaurant.

(1.2) The Prime Minister opened the new hospital.

The most important element of this definition is the notion of a polarized scale (positive vs. negative, good vs. bad, desirable vs. undesirable, agreement vs. disagreement, etc.). Thus, the sentiment of jealousy in phrase (1.3) expresses an emotion and may appear in isolation from an evaluative opinion of an entity. Similarly, certain predictive expressions, which relay opinions in everyday language, do not constitute evaluations. In the second section of phrase (1.4), for example, the author expresses a hypothesis regarding that evening's weather, but this does not constitute an evaluation of the weather in question.

(1.3) I'm jealous of my brother.

(1.4) I won't be able to go this evening, I think it's going to rain.

In what follows, we shall focus exclusively on the automatic detection of opinions expressed on a **polarized scale** or **evaluative opinions**.

1.2.2. *Opinion as a structured model*

Within the context of automatic extraction, (Liu 2012) proposed a structured model Ω made up of five elements:

– s is **the subject** of the opinion;

– a is **an aspect** of s;

– e is **the emitter**;

– $senti$ is **the sentiment** expressed by e toward s (and potentially a). $senti$ is generally represented by a triplet $(type,\ p,\ v)$ such that:

- $type$ is **the semantic type** of the sentiment being expressed. This type is defined in relation to predefined linguistic or psycho-linguistic categories. For example, in *This film bored me*, the author expresses a sentiment of boredom, while in (1.1) the author expresses an evaluative judgment;

- p is **the polarity**, which may be positive or negative;

- v is **the valency** (or **strength**), indicating the degree of positivity or negativity. Valency is often combined with polarity to obtain an opinion score. Thus, the score associated with the adjective *excellent* ($+2$, for example) will be higher than that for the adjective *good* (e.g. $+1$);

– d is **the date** on which the opinion was posted online.

The aim of automatic extraction is to identify each element of this quintuplet within a text. The model is designed to respond to the specific needs of **feature-based opinion mining systems**. These systems are very popular in the field of product reviews (movies, books, restaurants or any other product which can be broken down into parts) and aim to associate each extracted opinion $senti$ with a feature or element a. Liu stipulates that the presence of these five elements depends on the target application and that certain elements may be ignored, for example d or even s.

It is important to note that while the instantiation of the quintuplet $\Omega = (s,\ a,\ senti,\ e,\ d)$ appears simple for a human, automatic extraction is extremely complex for a computer program, mainly due to the incapacity of existing systems to grasp the context in which opinions are expressed.

One solution to this problem is to define opinion as a dynamic, rather than a static, model, in which each element of Ω depends on a variety of linguistic and extra-linguistic factors, such as domain dependency, operators or the discourse surrounding the phrase. Interested readers may wish to consult the new model proposed by Benamara *et al*. (2017), which builds on the model presented in (Liu 2012) to take account of the notion of context.

1.2.3. *Opinion extraction: principal approaches*

Existing opinion analysis systems generally focus on the extraction of one or more elements at the phrase or document level. This process involves three main tasks (Benamara 2017):

1) extraction of a subject and its features;

2) extraction of the emitter;

3) extraction of the sentiment. This task may be broken down into two steps:

- subjectivity analysis: used to determine whether a portion of text carries an opinion (is subjective) or simply presents facts (is objective);

- polarity analysis: used to determine the opinion that is effectively carried by a portion of text known to be subjective.

These subtasks may be carried out independently from each other, or simultaneously. When subtasks 1 and 2 are carried out together, each opinion is associated with a subject-feature couple. In this case, we speak of a feature-based opinion extraction system.

In most systems, the methods and techniques used to carry out these tasks are built on four major hypotheses:

– opinions concern a single subject s;

– there is a single emitter e;

– the phrases or documents being analyzed are independent of one another;

– a proposition (or phrase) contains one opinion at most.

Working from these hypotheses, existing systems focus exclusively on extracting explicit opinions and/or explicit features, using a bottom-up approach in which the calculation of an overall opinion expressed in a text is seen as a process of aggregation of opinions identified locally in propositions or phrases.

In what follows, we shall present a brief overview of these methods (for a detailed presentation, see Liu (2015), the reference in the domain), notably the work published in the context of the two **DEFT** (*Défi fouille de textes*, data mining challenge) campaigns: **DEFT'09** and **DEFT'15**.

The **DEFT'09** campaign included a subjectivity recognition task at document level for texts in French, English and Italian. The chosen corpus was made up of press articles. The best results for French and English were obtained by Bestgen and Lories (2009) system, which proposes a standard support vector machine (SVM) classification based on lemmatized unigrams, bigrams and trigrams, filtered using frequency thresholds. Note that different attempts to optimize the parameters of the system failed to produce better results than those obtained using system defaults. Other participants proposed approaches based on the k nearest neighbors algorithm (Létourneau and Bélanger 2009) and on the use of specialized lexicons as learning features for SVM (Toprak and Gurevych 2009), although their results fell short of those obtained by Bestgen and Lories Bestgen and Lories (2009).

The **DEFT'15** campaign was focused on opinion mining and sentiment/emotion analysis for Twitter messages on the subject of climate change. Three tasks were proposed: (1) determine the global polarity of tweets, (2) identify generic classes (opinion, sentiment, emotion, information) and specific classes (from a list of 18) for these tweets and (3) identify the source, the target and the expression carrying an opinion, sentiment or emotion. Twelve teams took part. The best results, in macro-precision, were 0.736 (polarity), obtained using the system proposed by Rouvier *et al*. (2015); 0.613 (generic classes), obtained using the system proposed by Abdaoui *et al*. (2015); and 0.347 (specific classes) obtained by the system proposed by

Rouvier *et al*. (2015). None of the participants submitted data for the final task. The methods used were mostly based on supervised statistical learning (SVM, naive Bayes, neurone networks, Performance in terms of Perplexity of MultiClass model (PPMC)), and used a range of opinion lexicons (ANEW, Casoar, Emotaix, Feel, Lidilem) and polarity lexicons (Polarimots) as characteristics.

Opinion analysis of social media publications has played an important role in a variety of domains, including monitoring the results of the 2009 US elections on Twitter, Facebook, Google+, Youtube and Instagram (Figure 1.1); real-time monitoring of political debates (Figure 1.2) and predicting the results of the November 2016 US presidential elections using Google, Twitter and Facebook (Figures 1.3 and 1.4); predicting the psychological state of users on social networks Losada and Crestani (2016), etc.

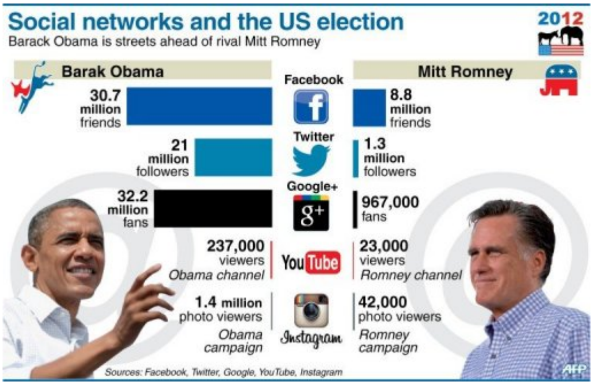


Figure 1.1. *Results of the 2009 US elections by social network. For a color version of the figures in this chapter see, www.iste.co.uk/karoui/irony.zip*

1.3. Limitations of opinion analysis systems

Globally, existing systems have produced good results in terms of automatic classification of the subjective or objective character of documents containing one or more phrases (section 1.2). However, the results obtained for polarity analysis (which consists of positioning a document on a subjectivity scale, from the most positive to the most negative) remain inconclusive. The main reason for this failing is that existing algorithms are unable to understand all of the subtleties of human language, as we shall see in the following sections.

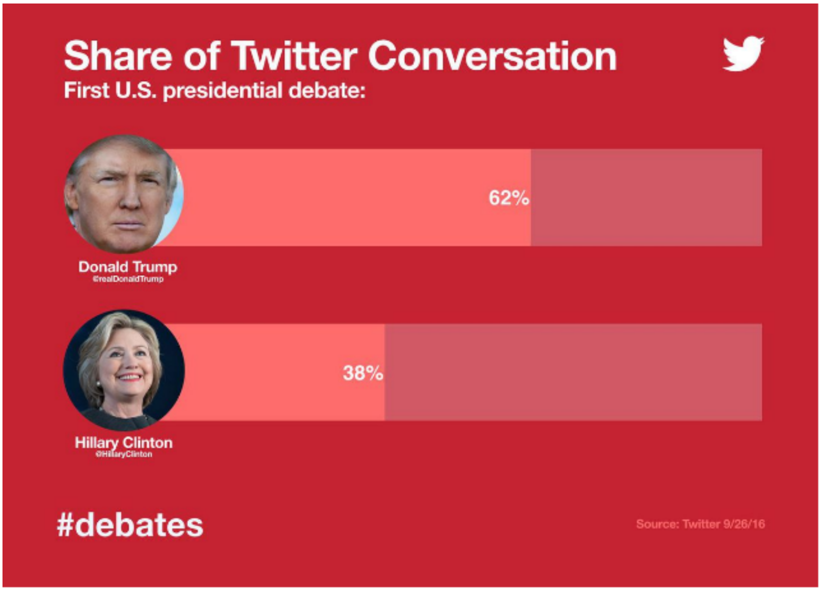


Figure 1.2. *Monitoring of US election debates on Twitter*

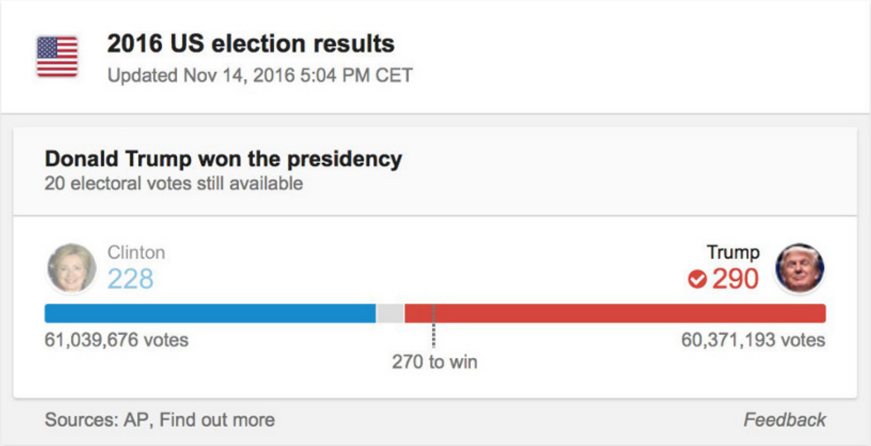
1.3.1. *Opinion operators*

Polarity (p) and/or valency (v) values, encoded out of context in lexicons or dictionaries, can be altered in the context of use by the presence of other elements in a phrase or text. These elements are known as **operators**. There are three main types of operator:

– negations, such as *doesn't*, *never*, *nothing*, *nobody*, etc. These operators reverse the value of p. However, in certain cases, the effect may also concern v. For example, in *This student isn't brilliant*, the opinion expressed is not negative, but rather a less intense positive;

– intensifiers, such as *very*, *less* and *quite*, which increase or decrease the value of v. Most intensifiers are adverbs. In some cases, punctuation, breaks or the repetition of characters can have the same effect;

– modalities, such as *maybe*, *believe* and *must*, which act on the strength of an expression and on its degree of certainty. For example, the phrase *This restaurant should be good* does not express an established opinion. In *You* *should go see this film*, on the other hand, the same modality reinforces a recommendation.



2016 US election results

Updated Nov 14, 2016 5:04 PM CET

Donald Trump won the presidency

20 electoral votes still available

Clinton 228

Trump 290

61,039,676 votes

270 to win

60,371,193 votes

Sources: AP, Find out more

Feedback

Figure 1.3. *Results of US elections on Google from the end of voting*

Most systems currently treat intensifiers and negations as polarity reversal phenomena (Polanyi and Zaenen 2006, Shaikh *et al*. 2007, Choi and Cardie 2008). Despite the evident importance of modalities in opinion analysis, they are almost never taken into account due to the difficulty of automatic treatment.

1.3.2. *Domain dependency*

Another factor affecting the values of p and v is the **domain**. An expression that is subjective in one domain may be factual in another: for example, the adjective *long* is factual in *A long skirt*, but subjective in *My battery life is long*. Even within the same domain, the polarity of an expression may not be fixed. The opinion expressed in *A hilarious movie* may be positive for a comedy, but negative for a drama. Expressions of surprise, such as *The movie was surprising*, also demonstrate contextual polarity. Finally, a remark such as *A little hotel* found on a booking site will have variable polarity depending on the reader.

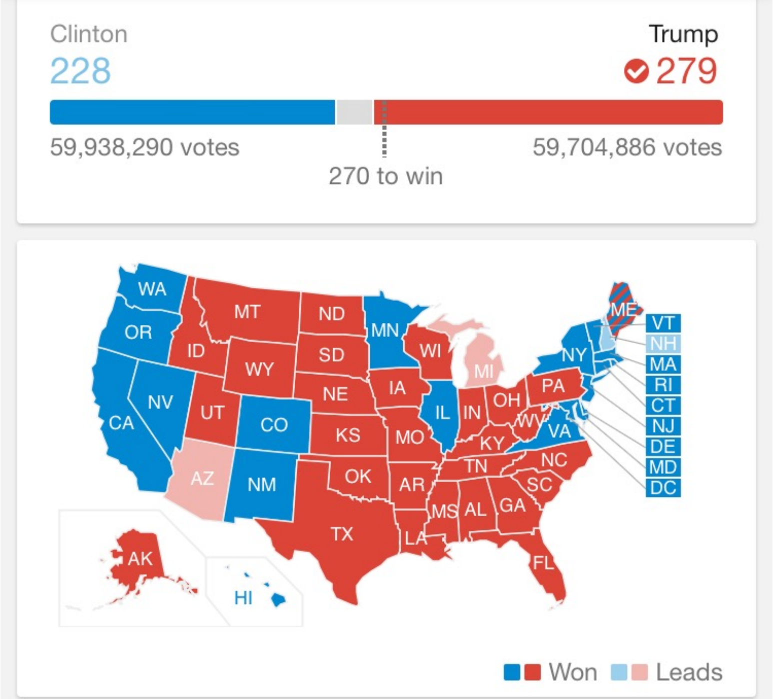


Figure 1.4. *Results of US elections on Google from the end of voting*

1.3.3. *Implicit opinions*

Opinions may be explicit or implicit. In the first case, opinions may be identified from words, symbols or subjective expressions in language, such as adjectives, adverbs, nouns, verbs, interjections or emoticons. Implicit opinions are found in words or groups of words that describe a situation (fact or state) judged to be desirable or undesirable on the basis of cultural and/or pragmatic knowledge common to the emitter and readers. For example, phrase 1.5 contains three opinions. The first two (underlined) are explicit and positive, whereas the third (in caps) is implicit and positive. Phrase (1.6) gives a further illustration of an implicit opinion, negative in this case, expressed on a review site.

(1.5) What an amazing movie. I was so drawn in that I DIDN’T MOVE AN INCH FROM MY SEAT.

(1.6) We bought this mattress in March. After trying it for a few days, we had the surprise of waking up in a ditch which had formed during the night.

Compared to explicit opinions, implicit opinions have received relatively little attention (see (Benamara *et al*. 2017) for a detailed state of the art of detection techniques for implicit opinions). However, their presence in texts is far from negligible. (Benamara *et al*. 2016) noted that around 25% of opinions in a corpus of TV series reviews were of this type, a figure which rose to 47% for a corpus of reactions to press articles.

1.3.4. *Opinions and discursive context above phrase level*

Discourse is an essential element in correctly understanding opinion texts, making it possible to analyze opinions at a higher level using the **rhetorical relations** between phrases (such as contrast, conditionality or elaboration). For example, consider the TV series review in phrase (1.7). The first three of the four opinions found in this text are, *a priori*, very negative. However, the final phrase, which contrasts with the three previous phrases, shows the true – positive – polarity of the document. Simply taking an average of the opinions expressed would have produced the wrong result; the discursive structure is essential to resolve the ambiguity concerning the overall polarity of the document.

(1.7) The characters are downright unpleasant. The scenario is utterly absurd. The sets are clearly of poor quality. But that’s what makes the series so unexpectedly great.

Similarly, conditionality can alter the positive or negative character of a subjective segment. For example, the negative phrase *If you have nothing better to do, go see this film* would be classified as positive by the majority of existing systems.

Each discursive relation has a specific effect on opinions. For example, contrast relations usually link phrases that are both subjective and of opposing polarities. Similarly, in an elaboration relation, the second phrase specifies or adds information to that which is introduced in the first, and the polarity is generally the same (an utterance such as *The movie is excellent. The actors are terrible* would not be discursively coherent). A statistical study of the effects of these relations is presented in (Benamara *et al.* 2016).

1.3.5. *Presence of figurative expressions*

Among all of the linguistic subtleties described in this section, we chose to focus on the detection of figurative language, particularly irony and sarcasm. The presence of one of these two phenomena, for example in a tweet, can result in erroneous predictions of overall opinion. For example, an opinion analysis system might classify tweet (1.8) as expressing a positive opinion, given the presence of the segments “better and better” and “this is progress”. To a human reader, however, this tweet is obviously critical of French president François Hollande’s policy on unemployment.

(1.8) **Better and better**. **This is progress** #irony #France @LeFigaroEmploi: Unemployment: Hollande’s magic trick.

Similarly, tweet (1.9) might be classed as positive based on the terms “I love” and “Is amazing” and on the positive emoticon “ :)”, which actually constitutes a criticism of working conditions – a negative situation.

(1.9) I **love** my job, 5 minutes for lunch and working until 8 pm **is amazing** :) #irony.

In phrases (1.8) and (1.9), the ironic nature of the tweet is indicated by use of the #irony hashtag, and can also be detected by readers on the basis of cultural knowledge.

The specificities of figurative language and the different forms which it may take are described in the following section. We have chosen to focus on irony, sarcasm, satire and humor, due to their frequent use in messages posted on social media.

1.4. Definition of figurative language

Unlike literal language, figurative language moves away from the original meaning of words or phrases to create a new, figurative or "illustrative" meaning. Figurative language is a way of using descriptions to create a specific image, often with an associated emotional component. It may also be used to humorous ends. It often consists of making comparisons, repeating sounds, exaggerating or creating sensory effects[2].

Analyzing figurative language represents a particularly difficult challenge for NLP. Unlike literal language, figurative language uses elements such as irony, sarcasm and humor to communicate more complex meanings, and can be challenging to interpret, for human listeners as well as for computers.

In this work, we shall focus on two main types of figurative language, irony and sarcasm; we shall also consider satire and humor, typically considered to be close to irony. These different types of figurative language have been defined in several different ways. In the following paragraphs, we shall cite some of the most important definitions put forward by philosophers and linguists.

1.4.1. *Irony*

Irony denotes a discrepancy between discourse and reality, between two realities or, more generally, between two perspectives to incongruous effect. The Oxford English dictionary defines irony as "the expression of meaning through the use of language signifying the opposite, typically for humorous effect". The French dictionary *Le Petit Robert* offers a similar definition, adding that an element of mockery is generally present.

Building on a similar definition found in the French dictionary *Le Petit Robert*, (Raeber 2011) identified two key aspects that characterize irony. The first takes account of *the illocutory effect of irony*, i.e. irony or teasing. The second indicates that irony includes an opposition between that which is said and that which is meant, i.e. an *antiphrase*. These two aspects are of very different natures. The first is of a pragmatic type, while the second is rhetorical.

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

According to (Mercier-Leca 2003), definitions of irony may take a very limited or far broader view of the subject. According to the narrowest definitions, irony is simply stating the opposite to what is meant; however, this view fails to take account of all existing forms. Taking a broader view, ironic discourse is considered as the communication of a meaning different to that associated with the words themselves (not necessarily the opposite).

Irony covers a range of distinct phenomena, the main forms being **verbal irony** and **situational irony**. According to (Niogret 2004), verbal irony expresses a contradiction between the speaker's thought and expression, and is created through language. Situational irony denotes any instance in which a situation contradicts the utterances or pretentions of a speaker (Niogret 2004). Other types of irony have been identified in the literature, but have not featured in detailed linguistic studies or automatic detection campaigns. These include Socratic irony, romantic irony and dramatic irony.

These different types of irony may be expressed in writing or orally. Philosophers and linguists have drawn a distinction between two main genres Tayot (1984), Didio (2007): **conversational (or interactive) irony** and **textual irony**.

Conversational irony is expressed orally in conversations or linguistic exchanges between at least two individuals. As such, it is identified through intonation, which may be the only way of detecting the speaker's ironic intention, through imitation, through gestures and through facial expressions (Didio 2007). Conversational irony is spontaneous, instinctive and not preprogramed.

Textual irony, on the other hand, is expressed in writing in both literary and non-literary texts. Didio (2007) notes that *textual irony* is very different to *conversational irony* in that it is planned and executed with great care prior to presentation. Furthermore, textual irony is intrinsically linked to the literary communication in which the author is writing and which the reader experiences. Irony is a popular tool for writers attempting to make a point. According to (Tayot 1984), these writers may take one of two approaches:

> "either calling a given order into question in order to impose their own point of view, or taking a two-level approach: first,

> questioning the order of things to create doubt, then second, holding up the world as it is against the world as it could be, without imposing any ideology".

Below, we shall present an overview of the first type of irony defined by Socrates and known as *Socratic irony*. We shall then go into greater detail concerning the two main types of irony, *verbal irony* and *situational irony*, presenting different theories in chronological order.

1.4.1.1. *Verbal irony*

Socratic irony, one of the primary triggers for the study of irony as a whole, is also one of the least-explored forms. Socratic irony is a form of irony in which the speaker feigns ignorance in order to highlight gaps in the listener's knowledge. Kierkegaard, whose work in this area is explored in *Le vocabulaire Kierkegaard* (Politis 2002), specified that the term *irony* is a rhetorical concept of Greek origin, meaning "feigned ignorance", often used by the philosopher Socrates:

> "Irony has an inventor, **Socrates**, and an apparent function, **refutation**: it appears as a rhetorical weapon used to refuse the rhetoric of another. Socrates insists on a step-by-step examination of his adversary's assurances. Starting from an affirmation of ignorance and progressing through a series of questions, the philosopher compels his adversary to confirm or reject successive affirmations, finally identifying the limits of his actual knowledge.
>
> Contrary to the common interpretation according to which Socrates has perfect knowledge of that which he pretends to ignore (from which we derive the modern notion of irony: stating the opposite of what one believes), irony involves a veritable suspension of opinion. By probing the depths of Socrates' own utterances along with those of his adversary, irony aims to undermine certainties and established knowledge. It forces the underlying discourse to the surface". (Encyclopedia *Larousse*)

Kerbrat-Orecchioni (1976) describes cues that may be used to construct and comprehend irony in a verbal sequence, and considers irony to be a rhetorical process based on antiphrases. For (Raeber 2011), this theory is problematic in the context of concrete case studies where no reversal of the encoded sense is present. Thus, according to (Raeber 2011), ignorance of the existence of other types of irony (Kerbrat-Orecchioni 1976) is hugely problematic. Kerbrat-Orecchioni considered these to be examples of situational, rather than verbal, irony, as in his view, an utterance should be considered ironic if, and only if, it describes a contradiction or paradox.

Writing in the 1970s, (Grice *et al.* 1970, 1975) supported the idea put forward in Kerbrat-Orecchioni (1976), considering that verbal irony should be treated as a negation (or an antiphrase). Conversely, (Sperber and Wilson 1981) treated irony as an interpretation or echoing act (the utterances of a speaker are echoed by another speaker, generally to mock or criticize the former). Comparing work by different linguists, Grice's theory and Sperber and Wilson's theory may be seen to present two broad views of the nature of verbal irony; other approaches are generally judged in terms of theoretical proximity to, or distance from, Grice or Sperber and Wilson.

According to (Grice *et al.* 1975), irony consists of using an utterance with a usual meaning "p" to transmit "non-p". Following the development of his conversational implicature theory (inference in meaning), Grice came to describe irony as a violation of the most important conversational maxim, the truthfulness maxim[3]. This idea is based on the fact that irony implies the expression of something that the speaker knows to be false, and was strongly criticized by supporters of Sperber and Wilson's theory.

Sperber and Wilson (1981) felt that Grice's theory fell short in its focus on *violation of the truthfulness maxim*, and ceased to apply in cases where irony is manifest through the violation of other maxims. Despite this shortcoming, the contributions of Gricean theory are evident, essentially in identifying irony as a linguistic phenomenon that may only be adequately interpreted with reference to the context of utterance.

In the 1980s, Sperber and Wilson (1981) put forward a theory based on the *use-mention distinction*. They defined irony as a special form of *mention* in which the speaker repeats a proposition or thought attributed to another, thus communicating their own critical attitude toward the content. This was encapsulated in Sperber and Wilson's **Mention Theory**, which attracted

[3] The truthfulness maxim prohibits stating something that one knows to be false.

considerable criticism on the grounds that it does not distinguish between ironic echoes and simple citations or reported speech. The authors went on to develop the notion of echo interpretation in the form of **Echo Theory**. If an utterance communicates agreement with an idea, then it cannot be considered ironic. However, if cues indicating the presence of mockery are found, then the utterance is ironic.

Take the following example:

(1.10) Speaker 1: The weather's nice today.
Speaker 2: The weather's super nice today!

To determine whether or not this example is ironic, we need to refer to reality. If the weather is actually nice, then the utterance made by speaker 2 is non-ironic. However, in the case of inclement weather conditions, the same utterance would be an ironical echo of the utterance made by speaker 1. The listener is required to perceive both the echoic aspect of the utterance (source of the echo) and their partner's opinion of that utterance in order to understand the ironic intent of their statement.

Over the course of the 1980s, a number of linguists proposed other visions of irony based on Grice's or Sperber and Wilson's works. (Clark and Gerrig 1984) notably proposed **Pretense Theory**, an extension of Grice's theory. According to this approach, while an echo is not a compulsory characteristic of irony, all speakers communicating in an ironic manner are pretending to adhere to a discourse that they do not, in fact, support.

The purpose of the speaker is to criticize and ridicule the content of a sincere discourse. Thus, to understand irony, a listener must be able to recognize the different roles being played by the speaker. This theory was extended further by Kumon-Nakamura *et al*. (1995) in the form of *the Allusional Pretense Theory of irony*.

To support their theory, (Kumon-Nakamura *et al*. 1995) drew on work by Kreuz and Glucksberg (1989) affirming that allusion does not simply represent a reference to a past utterance or event, but that it also expresses a divergence between that which is said and that which should have been said in relation to the context. (Attardo 2000a) defines iconic utterances as essentially inappropriate while remaining relevant to the context: the literal meaning only serves to indicate the speaker's ironic intent, while the full ironic meaning can be inferred from the context. Attardo thus accepts Grice's theory, considering that the violation of conversational maxims may provoke irony, among other things.

In Attardo's view, an utterance is ironic if it respects the following four conditions:

1) the utterance is contextually inappropriate;

2) nevertheless, the utterance remains relevant to the conversation;

3) the speaker produces the utterance intentionally with full awareness of the contextual inappropriateness;

4) the speaker presumes that at least part of the audience will recognize points 2 and 3.

1.4.1.2. *Situational irony*

Situational irony is a contrast between what is hoped for and the observed reality. Observers may feel surprise in relation to an unexpected situation. Figures 1.5 and 1.6 highlight contradictions between reality and appearances.

Niogret (2004) defines situational irony as relating to any situation in contradiction with the utterances or pretentions of an individual. Lucariello (1994) and Shelley (2001) indicate that situational irony does not imply the existence of a person expressing irony, but requires the presence of an observer external to a situation or event that is perceived as ironic.

1.4.2. *Sarcasm*

According to the *Oxford English Dictionary*, sarcasm is "the use of irony to mock or convey contempt". The utterance is bitter in nature and is intended to hurt the target (Simédoh 2012). Sarcasm is thus characterized by aggression, although not to the exclusion of mockery or teasing. Sarcasm is considered as a combination of the processes involved in both humor and irony, but is hurtful and overtly mocks the target.

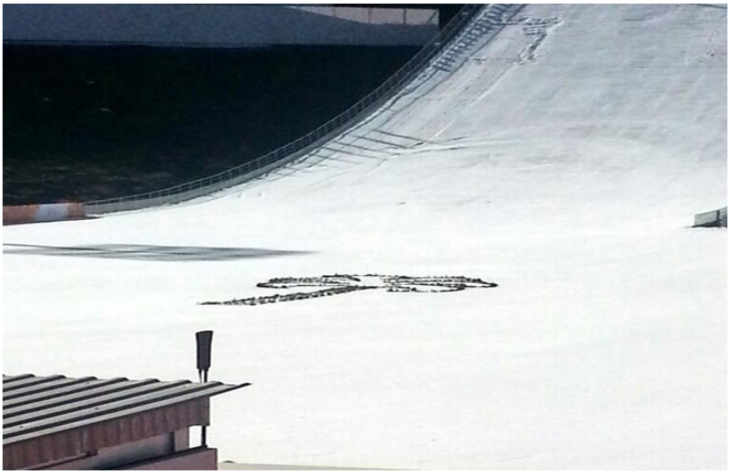


Figure 1.5. *Example of situational irony illustrated by a contradiction in the text found alongside an image. To make it clear that there was no snow, a palm tree was drawn on the slope (source: Twitter)*

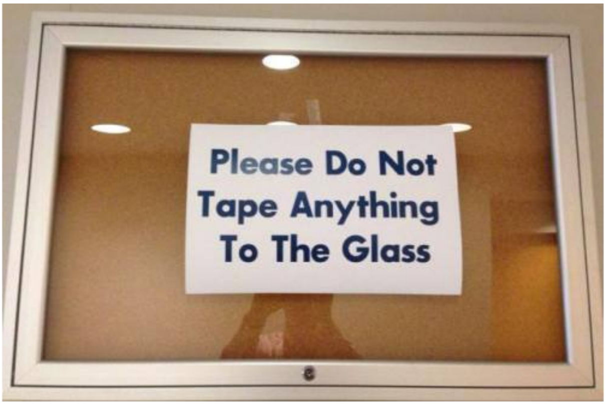


Figure 1.6. *Example of situational irony illustrated by a contradiction in an image (source: Twitter)*

Didio (2007) adds that "in its first sense, sarcasm is a form of irony, mockery or ridicule, which is acerbic and insulting; in the second sense, the property of biting irony, and in the third, a rhetorical figure, cruel irony". Furthermore, in justifying a list of synonyms for sarcasm, (Didio 2007) refers back to the definition of sarcasm put forward by Angenot (1982):

> "Sarcasm consists of attacking an adversary whilst maintaining an appearance of casual goodwill and favor toward them. It manifests as an elementary metalogical opposition between apparent goodwill and dissimulated aggression. Sarcasm may consist of fallacious praise made following a reproach, which in fact only serves to increase the force of the reproach itself".

Sarcasm is thus associated with aggression, insult and nastiness, traits that are not present in irony.

1.4.3. *Satire*

According to the *Oxford English Dictionary*, satire is "the use of humour, irony, exaggeration or ridicule to expose and criticize people's stupidity or vices". It uses irony as an expression of judgment and criticism, with the addition of humor for entertainment purposes.

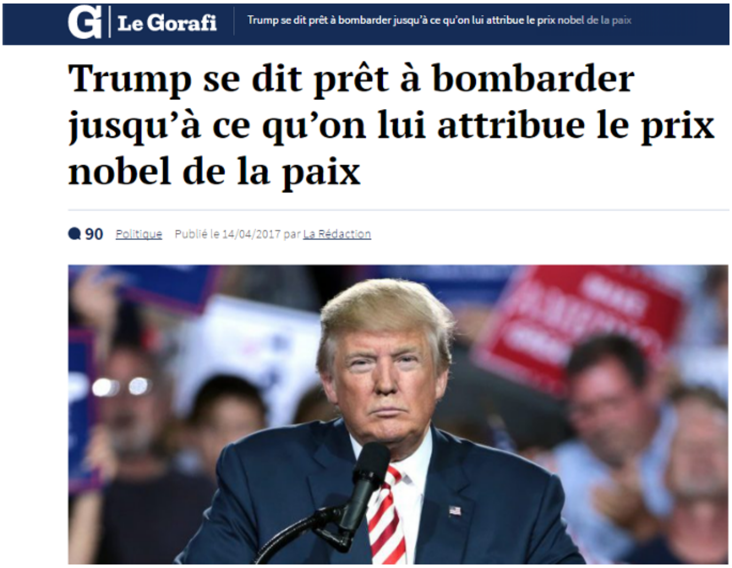
Bautain (1816) held that "satire strikes at the most tender part of the soul; it touches pride. Satire represents an inexhaustible and legitimate manner of causing hurt". Satire as a genre took off over the course of the 17th Century, with works such as La Fontaine's *Fables* and Molière's *Le Malade Imaginaire* in France, and Jonathan Swift's *Gulliver's Travels* and Alexander Pope's *The Rape of the Lock* in England.

The satirical press emerged in Europe in the 19th Century as a vehicle for political criticism, designed to provoke amusement by presenting a voluntarily distorted view of reality. The genre remains well represented in print and online. French examples include *Le Canard enchaîné*[4], *Charlie Hebdo*[5] and *Le Gorafi*, a site similar to *The Onion* and *The Daily Mash*[6]. Figure 1.7 shows a satirical article published in *Le Gorafi*.

[4] www.lecanardenchaine.fr/.

[5] www.charliehebdo.fr/.

[6] www.legorafi.fr/.



Le Gorafi | Trump se dit prêt à bombarder jusqu'à ce qu'on lui attribue le prix nobel de la paix

Trump se dit prêt à bombarder jusqu'à ce qu'on lui attribue le prix nobel de la paix

● 90 Politique Publié le 14/04/2017 par La Rédaction

Figure 1.7. *Example of a satirical press article published by Le Gorafi: Trump prepared to launch bombing attacks until he is awarded the Nobel Peace Prize*

1.4.4. *Metaphor*

A metaphor is a figure of speech based on analogy, in which a word or phrase is used to denote something to which it is not literally applicable, but which it resembles or with which it shares some essential quality Reboul (1991). A metaphor may be defined as a comparison made without using a comparison word (like, such as, similar to, etc.). Context is therefore essential to understanding metaphor, as it permits the listener or reader to determine whether or not the word is to be understood according to the usual sense. Linguists have defined several types of metaphor, including explicit metaphor, direct metaphor, and sustained metaphor.

Explicit metaphor indicates a relation between an object and that to which it is compared through the use of expressions. This type of metaphor is also referred to as a metaphor *in praesentia* or comparison metaphor: for example, "his colleague is a snail" implies that the colleague in question is slow. Direct metaphor, on the other hand, compares two entities or realities without explicitly including the second element. The connection must be made by the listener or reader: for example, “he works with a snail” indicates that the person works with a slow colleague.

A sustained metaphor, or literary conceit, is made up of a string of implicit comparisons – as in Shakespeare’s famous lines from *As You Like It*: *All the world’s a stage, and all the men and women merely players*. *They have their exits and their entrances, and one man in his time plays many parts...*. According to (Riffaterre 1969), a sustained metaphor is:

> “a series of metaphors connected to each other by syntax – they belong to the same phrase or narrative structure – and by meaning: each expresses a specific aspect of a whole thing or concept represented by the first metaphor in the series”.

1.4.5. *Humor*

Linguists consider humor to be one of the hardest concepts to understand (Van de Gejuchte 1993, Nadaud and Zagaroli 2008). It may be defined by the presence of amusing effects, such as laughter, or by a sensation of well-being. In the broadest sense, humor is a mocking approach used to highlight the comical, ridiculous, absurd or unexpected character of certain aspects of reality. In a stricter sense, humor is a nuance of the comic register that aims to draw attention to pleasing or unexpected aspects of reality, with a certain level of detachment. However, in common parlance, the term is generally used to describe any form of jest, i.e. processes intended to provoke laughter or amusement. There are six main forms of jest, based on situations, words, gestures, characters, mores and repetition. Humor necessarily makes use of one of these forms of jest, but jesting is not necessarily humorous.

Research on humor has been carried out in a range of disciplines, including philosophy, linguistics, psychology and sociology. Researchers have attempted to define a set of characteristics for this type of language. Linguistic studies have represented humor using semantic and pragmatic models. Attardo defined humor as a phenomenon that relies on the presence of certain knowledge resources, such as language, narrative strategies, a target, a situation and logical mechanisms to produce an amusing effect (Attardo 1994, 2001). From a sociological standpoint (Hertzler 1970), humor is generally approached within the framework of a cultural context.

1.5. Figurative language: a challenge for NLP

Looking closely at the different definitions of irony, sarcasm or humor put forward by linguists, philosophers, psychologists and sociologists (see above), it is evident that knowledge of the context is essential to correctly understand these phenomena. This context is relatively easy to identify for human readers within the framework of a poem or an extract from a literary work. However, it is harder to identify in the case of short texts.

Our objective here is to identify figurative language in short texts posted on Twitter. This may be divided into a number of questions:

– Are the figurative forms identified in literary texts also used in short texts?

– Are there linguistic cues which enable us to infer irony in short texts?

– If so, are these cues sufficient? Are they independent of the language used?

– If not, how can we infer the context required to understand non-literal forms in short texts?

– How can different cues (linguistic and contextual) be modeled in an automatic system?

We aim to respond to all of these questions in this book, focusing on verbal irony expressed in tweets. As the borders between the different forms of figurative language presented above are not clear-cut, we shall treat the word *irony* as a generic term encompassing both irony in the strictest sense and sarcasm. Our contributions are presented in Chapters 3–5.

1.6. Conclusion

Our objective in this work is to propose an approach for automatic detection of irony in content generated by Internet users. We shall specifically consider tweets written in French before considering multilingual situations. In this chapter, we provided a general overview of opinion analysis and of the limitations of existing analysis systems. We also presented definitions of certain forms of figurative language, established by philosophers and linguists: irony, sarcasm, satire, metaphor and humor. We focused on verbal irony, which is at the heart of our investigation.

In Chapter 2, we give the state of the art of different computational projects in the area of figurative language, particularly irony, and of the different annotation schemas used for this phenomenon.

2

Toward Automatic Detection of Figurative Language

2.1. Introduction

As we saw in Chapter 1, irony is a complex linguistic phenomenon that has been studied in detail in both philosophy and linguistics (Grice *et al*. 1975, Sperber and Wilson 1981, Utsumi 1996). Although authors differ in their definition of irony, all agree that it implies a mismatch between what is said and the reality. Taking account of the differences between approaches, irony can be defined as a mismatch between the literal and intended senses of an utterance. The search for non-literal meaning begins when a listener becomes aware that the utterance in question, when taken literally, makes no sense in the context (Grice *et al*. 1975, Searle 1979, Attardo 2000a). In most studies, irony is considered in conjunction with other forms of figurative language, such as humor, satire, parody and sarcasm (Clark and Gerrig 1984, Gibbs 2000). The distinction between these different forms of figurative language, particularly that between irony and sarcasm, is highly complex. This is the result of a blurred distinction between the notions in question at a linguistic level, and of the complexity involved in differentiating between notions within a text at a computational level.

The theories described in Chapter 1 form the basis for most of the cues used for automatic detection. A study of the state of the art in this area shows irony and sarcasm to be among the most widely studied forms of figurative language, unlike metaphor and humor, which have received considerably less attention.

Automatic Detection of Irony: Opinion Mining in Microblogs and Social Media,
First Edition. Jihen Karoui; Farah Benamara and Véronique Moriceau.

The main reason for this preference lies in the importance of these forms for effective opinion and sentiment analysis (see Chapter 1, section 1.3).

The majority of the work in the field of NLP has focused on opinion texts, such as consumer reviews, or short texts from social networks such as Twitter. Generally speaking, negative consumer reviews are presumed to be more likely to contain ironic expressions (Tsur *et al*. 2010), a debatable assumption. Tweets accompanied by the hashtags *#sarcasm, #irony* or *#satire* are considered to be ironic or sarcastic.

The use of these hashtags makes it relatively easy to collect ironic and/or sarcastic datasets. In some cases, binary pre-annotation (ironic/non-ironic) is supplemented by manual annotation for opinions or more pragmatic phenomena.

Figurative hashtags in tweets are used as reference labels for automatic learning within a supervised machine learning framework. Learning is carried out using three groups of features:

1) surface features (punctuation, emoticons, etc.) and lexical features (opinion polarity, type of emotion expressed, etc.);

2) pragmatic features relaying the internal context of the message based on linguistic content alone, such as the use of semantically opposed words;

3) pragmatic features relaying the external context of the message using non-linguistic knowledge, such as discussion threads or user profiles.

In this chapter, we shall provide the state of the art of figurative language detection, focusing first on the corpora used and the annotation schemas proposed for these corpora (see section 2.2), and then on the methods used for automatic detection.

In addition to work on detecting irony, sarcasm and satire (section 4.5), we shall present research concerning the detection of other forms of figurative expression, such as metaphor (section 2.4), comparison (section 2.5) and humor (section 2.6).

Approaches to each of these forms will be described from three different perspectives based on the three sets of above-mentioned features. In conclusion, we shall present a summary establishing the precise position of our own work and identifying our specific contributions.

2.2. The main corpora used for figurative language

Most existing work is based on the use of hashtags and does not necessarily use manual annotation[1]. For example, (Gonzalez-Ibanez *et al*. 2011) presents a corpus in English made up of 900 tweets, broken down into three categories by hashtag: sarcasm (*#sarcasm, #sarcastic*), a direct positive sentiment (*#happy, #joy, #lucky*) or a direct negative sentiment (*#sadness, #angry, #frustrated*). (Reyes *et al*. 2013) constructed a corpus of 40,000 tweets in English containing *#irony*, *#education, #humor* and *#politics*. The corpus is divided into four sections of 10,000 tweets. The first part is ironic (*#irony*), while the three remaining sections are considered to be non-ironic (*#education, #humor, #politics*).

A similar approach was used by (Liebrecht *et al*. 2013) in creating a corpus of ironic tweets in Dutch. The collected corpus is made up of two subcorpora. The first contains 77,948 tweets collected from a database provided by the Dutch e-Science Center, and was published in December 2010. The collection was created using the *#sarcasme* hashtag. The second subcorpus, made up of 3.3 million tweets, was published on February 1, 2013. This subcorpus contains 135 tweets with the *#sarcasme* hashtag.

Other authors have suggested annotating additional information alongside ironic/ironic labels based on hashtags. The Italian tweet annotation schema Senti-TUT (Gianti *et al*. 2012) aims to analyze the impact of irony in expressions of sentiment and emotions. The three annotators involved with the project were asked to classify tweets into five mutually exclusive categories: *POS (positive), NEG (negative), HUM (ironic), MIXTES (POS and NEG)* and *NONE (objective)*. (Van Hee *et al*. 2015) looked at different specific forms of irony in tweets in English and Dutch: ironic by clash, ironic by hyperbole, ironic by euphemism, potentially ironic, and non-ironic.

Given our objective – a fine-grained analysis of ironic expressions in a corpus – we shall present a number of notable annotation schemes in this section, focusing on those which go beyond a simple binary pre-annotation (ironic/non-ironic). We shall also consider corpora constructed for the purposes of annotating metaphorical expressions, as in certain cases, these

[1] In cases where manual annotation is used, it is only applied to a small portion of the corpus in order to assess the reliability of the selected hashtags.

expressions are considered to be a characteristic of irony (see section A.5). For each corpus, we shall describe the data collection and manual annotation phases along with the results of the annotation campaign.

2.2.1. *Corpora annotated for irony/sarcasm*

2.2.1.1. *Senti-TUT: a corpus of tweets in Italian*

Gianti *et al*. (2012) carried out the first irony annotation campaign as part of the Senti-TUT project[2], with the objective of developing a resource in Italian and of studying expressions of irony in social networks. The annotation process used in this project is described below.

2.2.1.1.1. Corpus collection

The Senti-TUT corpus is made up of two subcorpora of political tweets: TWNews and TWSpino. The authors explained their decision to focus on politics by the fact that irony is considered to be particularly widespread in the domain.

For the TWNews corpus, collection was carried out using time and metadata filters to select messages representing a variety of political opinions. The authors used *Blogometer*[3], using the Twitter API to collect tweets published over the period from October 6, 2011 to February 3, 2012, the election period in which Mario Monti took over from Silivio Berlusconi as Prime Minister. The authors established a list of keywords and hashtags to use in collecting tweets: *mario monti/#monti, governo monti/#monti and professor monti #monti* (in both lower and upper case).

This approach resulted in a collection of 19,000 tweets, including 8,000 re-tweets that were later deleted. The remaining tweets were filtered by human annotators who removed 70% of the data, including badly written tweets, duplicates and texts that were incomprehensible without contextual information. After this second stage of filtration, the final corpus contained 3,288 tweets.

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

[3] www.blogmeter.eu.

The TWSpino corpus is made up of 1,159 tweets collected from the Twitter section of Spinoza[4], an immensely popular Italian blog containing satirical political content. These tweets were selected from tweets published between July 2009 and February 2012. Tweets containing advertising, which accounted for 1.5% of the data, were removed.

2.2.1.1.2. Tweet annotation

A two-level annotation scheme was proposed, covering both the global polarity level and the morphology and syntax level of tweets (Gianti *et al*. 2012). The five annotation categories are listed below. For each example, we have provided an example in Italian, taken from Senti-TUT, and an English translation.

– *pos (positive)*: the overall opinion expressed in the tweet is positive – see phrase (2.1):

(2.1) Marc Lazar: “Napolitano? L’Europa lo ammira. Mario Monti? Può salvare l’Italia”.
(Marc Lazar: “Napolitano? Europe admires him. Mario Monti? He can save Italy”.)

– *neg (negative)*: the overall opinion expressed in the tweet is negative – see phrase (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)*: the tweet is ironic – see phrase (2.3):

(2.3) Siamo sull âorlo 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).)

[4] www.spinoza.it.

– *mixed (pos and neg)*: the tweet is both positive and negative – see phrase (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 illusions about Monti's government)

– *none*: the tweet is neither positive nor negative, nor is it ironic – see phrase (2.5)

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

Annotation was carried out by five human annotators. An initial campaign for a subcorpus of 200 tweets was used to validate labels, and inter-annotator agreement was calculated at $k = 0.65$ (Cohen's kappa). A second annotation step was carried out for 25% of the tweets for which the annotators did not agree. Following this stage, 2% of the tweets were rejected as being overly ambiguous and removed for the corpus. The final corpus thus consisted of 3,288 tweets from the TWNews corpus.

2.2.1.1.3. Results analysis for the annotation phase

Once the annotation process was completed, the results of the manual annotation campaign were analyzed. At this stage, two different hypotheses were tested: (H1) polarity inversion is an indicator of irony and (H2) expressions of emotions are frequent in ironic tweets.

The annotations highlighted the presence of different types of emotions in the corpora. The emotions most frequently expressed in the TWNews-Hum corpora were joy and sadness, conceptualized in terms of inverted polarity. A greater variety of irony typologies was observed, including sarcastic tweets, intended to hurt their target, and humorous tweets, which tend to produce a comic or parody effect instead of raising negative emotions. In the TWSpino corpus, however, the majority of the detected emotions were negative, and the irony typology was less varied, essentially limited to sarcasm and political satire. This may be due to the fact that the messages in TWSpino are all selected and reviewed by an editorial team; furthermore, the TWSpino editors explicitly state that the blog is intended to be satirical.

The different analyses carried out as part of this project showed that irony is often used in conjunction with a seemingly positive declaration to communicate a negative value; the reverse is rarely true. This corresponds to the results of theoretical studies, stating that it is rare for positive attitudes to be expressed in a negative mode, and that these expressions are harder for human listeners or readers to process than expressions of negative attitudes in a positive mode.

In the work presented in Chapter 5, part of the Senti-TUT corpus was used in studying the portability of our approach to other Indo-European languages.

2.2.1.2. *Corpus of English and Dutch tweets*

Hee *et al*. (2016) proposed an annotation scheme for a corpus of tweets in English and Dutch. All annotations were carried out using the BRAT rapid annotation tool[5] (Stenetorp *et al*. 2012).

2.2.1.2.1. Corpus collection

Hee *et al*. (2016) collected a corpus of 3,000 tweets in English and 3,179 tweets in Dutch using Twitter's API. The two ironic corpora were collected using the hashtags *#irony, #sarcasm* (*#ironie* and *#sarcasme* in Dutch) and *#not*.

2.2.1.2.2. Tweet annotation

The proposed annotation scheme enables (1) identification of ironic tweets in which a polarity shift is observed and (2) detection of segments of contradictory text showing the presence of irony. The scheme involves a three-step annotation process:

1) indicate whether a tweet is:

– ironic by clash: the text in the tweet expresses a literal polarity (expressed explicitly in the text), which is in contradiction with the expected polarity (representing the reality in the context). For example, a tweet may express a positive opinion in a situation where, based on context, the polarity should be negative, as shown in phases (2.6) and (2.7):

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

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

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

– ironic by hyperbole: the text in the tweet expresses a literal polarity, which is stronger than the expected polarity, as shown in phrase (2.8):

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

– ironic by euphemism: the text in the tweet expresses a literal polarity, which is weaker than the expected polarity, as shown in phrase (2.9):

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

– potentially ironic: there is no difference between the literal and expected polarities; however, the text contains another form of irony (such as situational irony, as shown in phrase (2.10))[6]:

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

– non ironic: the tweet is not ironic, as shown in phrase (2.11):

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

2) if the tweet is ironic:

– indicate whether an ironic hashtag (e.g. *#not, #sarcasm, #irony*) is necessary in order to understand the irony;

– indicate the degree of difficulty involved in understanding the irony, on a scale from 0 to 1;

3) annotate contradictory segments. enumerate

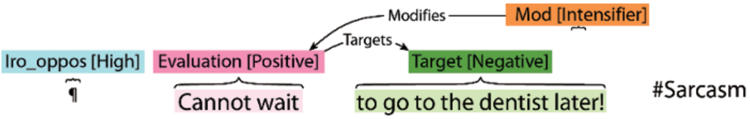


Figure 2.1. *Example of a tweet annotated as ironic by clash (Hee et al. 2016). For a color version of the figures in this chapter see, www.iste.co.uk/karoui/irony.zip*

[6] See Chapter 1, section 1.4.1.2 for a definition of situational irony.

Figure 2.1 shows an example of a tweet annotated as ironic by clash using the scheme described above. In this example, the annotators considered the expression *cannot wait* as a literally positive evaluation (intensified by an exclamation mark), which is in contradiction with the act of going to the dentist's office, typically perceived as an unpleasant experience and thus implying a negative sentiment.

2.2.1.2.3. Results of the annotation procedure

The annotation procedure showed that 57% of the English tweets in the corpus were ironic, 24% were potentially ironic and 19% were non-ironic. In the Dutch corpus, 74% of tweets were annotated as ironic, 20% as potentially ironic and 6% as non-ironic. The proportion of tweets in which an ironic hashtag was required in order to understand the ironic meaning was similar for the two languages: 52% for English and 53% for Dutch. Furthermore, the distribution of tweets with a positive or negative polarity was the same for both languages, with a majority of positive tweets. In terms of irony triggers (clash, hyperbole or euphemism), the annotation process showed that irony was expressed through a clash in 99% of cases; irony was expressed by hyperbole or euphemism in only 1% of cases.

As we demonstrate in Chapter 3, irony may be expressed on social media using other pragmatic mechanisms in place of hyperbole or euphemism.

2.2.2. *Corpus annotated for metaphors*

Shutova *et al.* (2013) proposed an annotation scheme for metaphors expressed using verbs. The scheme consisted of identifying metaphorical concepts and linking these concepts through source–target relationships.

2.2.2.1. *Corpus collection*

The annotation campaign was carried out on a set of texts taken from the British Corpus (BNC). The BNC is a corpus of 100 million words containing samples of British English from the second half of the 20th Century (90% written and 10% oral). The collected corpus includes samples from various genres included in the BNC: fiction (5,293 words), newspapers (2,086 mots) and magazine articles (1,485 mots), essays from the fields of politics, international relations and sociology (2,950 mots) and radio shows (1,828 transcribed words). The study corpus is made up of 13,642 words in total.

2.2.2.2. *Annotation scheme*

For each genre of corpus in the BNC, annotators were asked to (1) class each verb according to whether the meaning was metaphorical or literal and (2) identify source-target domain correspondences for those verbs marked as metaphorical. Two lists of categories describing concept sources and targets were provided (Table 2.1). Annotators were asked to choose the pair of categories from these lists, which best represented each metaphorical correspondence. Additionally, annotators were permitted to add new categories if they were unable to find a suitable option in the predefined list.

Phrase (2.12) gives an illustration of the annotation process:

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

Source concepts	**Target concepts**
Physical Object	Life
Living Being	Death
Adversary/Enemy	Time/Moment in time
Location	Future
Distance	Past
Container	Change
Path	Progress/Evolution/Development
Physical Obstacle (example: Barrier)	Success/Accomplishment
Directionality (example: Up/Down)	Career
Basis/Platform	Feelings/Emotions
Depth	Attitudes/Views
Growth/Rise	Mind
Size	Ideas
Motion	Knowledge
Journey	Problem
Vehicle	Task/Duty/Responsibility
Machine/Mechanism	Value
Story	Well-Being
Liquid	Social/Economic/Political System
Possessions	Relationship
Infection	–
Vision	–

Table 2.1. *Suggested source and target concepts for metaphor annotation (Shutove et al. 2013)*

According to the authors, the first three verbs (ask, post and buy) are used in their literal sense, whereas the fourth verb (transport) is used in a figurative sense (meaning "transported by a feeling" rather than "transported by a vehicle" in this case). The use of "transport" in this example is therefore metaphorical, hence the conceptual mapping of emotions to vehicles.

2.2.2.3. *Results of the annotation campaign*

Table 2.2 shows a Cohen's kappa of $\kappa = 0.64$ for the identification of metaphorical verbs. This is considered to represent a substantial level of agreement. Measuring agreement for the second task was more complex. The overall agreement level for metaphorical concept assignment was $\kappa = 0.57$, but the level was higher for the choice of target categories ($\kappa = 0.60$) than for source categories ($\kappa = 0.54$).

Tasks	**Kappa (κ)**	**Number of categories (n)**	**Number of annotated instances (N)**	**Number of annotators (k)**
Verb identification	0.64	2	142	3
Metaphorical concept assignment	0.57	26	60	2
Choice of target category	0.60	14	30	2
Choice of source category	0.54	12	30	2

Table 2.2. *Interannotator agreement for metaphor annotation (Shutova et al. 2013)*

A study of cases of annotator disagreement revealed that a partial overlap between the target concepts in the list was the main source of error. For example, *progress* and *success*, or *views*, *ideas* and *methods*, were often confused. These categories were finally combined in order to make the annotation coherent. This fusion increased the interannotator agreement level ($\kappa = 0.61$ instead of $\kappa = 0.57$).

Finally, (Shutova *et al*. 2013) showed metaphor to be a widespread phenomenon, and found that 68% of metaphors are expressed by verbs.

2.3. Automatic detection of irony, sarcasm and satire

At the same time as these annotation schemes for figurative language were being developed, in the 2000s, work also began on the automatic detection of figurative language. This subject has become particularly important within the domain of NLP due to progress in the field of sentiment analysis and the high levels of figurative language found online, on websites and in social media.

Broadly speaking, work on the automatic detection of figurative language has been based on three main approaches: (1) surface and semantic approaches, (2) pragmatic approaches using the internal context of utterances and (3) pragmatic approaches using context external to the utterance. The first approach (particularly the form using surface cues) has often been used as a baseline for work using the second or third approaches. These approaches have been used in work on irony, sarcasm and metaphor, but not for comparison or humor.

In this section, we shall describe the work that has been carried out on automatic detection of figurative language. We shall begin by presenting work on irony, sarcasm and satire (section 4.5), followed by work on the detection of metaphors (section 2.4), comparison (section 2.5) and, finally, humor (section 2.6). We will present the approaches proposed for each type of figurative language and the corpora used in these approaches.

2.3.1. *Surface and semantic approaches*

A brief overview of the lexical and semantic features that are most widely used for automatic detection of irony and sarcasm is provided in Table 2.3.

Burfoot and Baldwin (2009) collected a corpus of 4,233 press articles, of which 4,000 were non-satirical and 233 were satirical, as part of work on automatic detection of satire. An SVM-light classifier was used with default parameters for three sets of features: bag of words type features, lexical features and semantic features. The bag of words model takes account of either binary features, used to identify whether certain words are or are not present in the corpus, or of word weights using bi-normal separation feature scaling (BNS). In this approach, weights are assigned to features that are strongly correlated with positive or negative classes. In the case of lexical features, the selected elements included article titles, profanity and slang. The latter two features enable the detection of familiar and informal vocabulary, which is widely used in satirical articles. The results of the learning process showed that the combination of all of these features produces an F-measure score of 79.5%.

Research group	**Corpus**	**Features**	**Results**
(Burfoot and Baldwin 2009)	Press articles (4,233)	Bag of words, title, profanity, slang, semantic validity	Precision = 79.8%
(Carvalho *et al*. 2009)	Press articles (8,211) and comments (250,000)	Punctuation, quotation marks, emoticons, quotations, slang, interjections	Precision = 85.4%
(Veale and Hao 2010)	Similes collected online (20,299)	F-measure = 73 %	–
(Liebrecht *et al*. 2013)	Twitter (77,948)	Unigrams, bigrams and trigrams	AUC = 79%

Table 2.3. *Summary of the main surface and semantic approaches used to detect irony/sarcasm*

Carvalho *et al*. (2009) also based their study on press articles, building their corpus from a collection of 8,211 articles from a Portuguese newspaper, along with the comments associated with each article (250,000 comments). This corpus was used to study a set of simple linguistic cues associated with the expression of irony in Portuguese. Within this context, (Carvalho *et al*. 2009) used a pattern-based approach to identify ironic and non-ironic phrases (for example: $P_{laugh} = (LOL|AH|Emoticon)$, $P_{punct} = 4\text{-}GRAM^{+}$ $(!!|!?|?!)$).

The results showed the most productive patterns to be those containing *punctuation signs*, *quotation marks* and *emoticons*. The *quotation* and *slang* patterns also performed very well for irony detection, with a precision of 85.4% and 68.3%, respectively. Using these patterns, 45% of ironic phrases were detected. However, these results are not particularly representative as the coverage of these patterns is extremely low (around 0.18%), essentially due to the choice of starting phrases, which all contained opinion words and a named entity. A decision could not be reached for 41% of comments collected using the *interjection* pattern and for 27% of comments collected using *punctuation*. This is due to the lack of context and highlights the need for a more fine-grained analysis (e.g. including analysis of the previous phrases) in order to understand the ironic or non-ironic meaning of a comment.

In the same context, Veale and Hao (2010) analyzed a corpus of similes (comparisons expressing opposition). They began by harvesting data online using the pattern – *about as ADJ as ADJ* – in order to detect ironic intentions in creative comparisons (for example *he looked about as inconspicuous as a tarantula on a slice of angel food cake*). The collected extracts were filtered manually in order to separate similes from comparisons, resulting in a total collection of 20,299 similes. Manual annotation of the corpus gave a result of 76% ironic similes and 24% non-ironic similes. These similes were then grouped into three categories by opinion: positive, negative or hard to define. The results showed that the majority of ironic similes are used to relay negative sentiments using positive terms (71%). Only a small minority of similes (8%) aim to communicate a positive message using a negative ironic utterance. The simile classification process was automated using a 9-step model; evaluation of this model gave an F-measure of 73% for the ironic class and 93% for the non-ironic class.

Liebrecht *et al*. (2013) worked on the classification of sarcastic/ non-sarcastic tweets using two different corpora. The first is made up of 77,948 tweets in Dutch containing the *#sarcasme* hashtag; the second is made up of all tweets published in February 1, 2013. This second corpus contains 3.3 million tweets, only 135 of which include the *#sarcasme* hashtag. The authors used the *Balanced Winnow* supervised machine learning tool, taking unigrams, bigrams and trigrams as features. In this experiment, the *#sarcasme* hashtag was considered to be the reference annotation. Evaluation of the proposed model gave an area under the curve (AUC) score of 0.79 for the detection of sarcastic tweets.

From this overview, we see that all of the authors cited above agree that the use of surface and semantic features alone is not sufficient for irony and sarcasm detection, and that other, more pragmatic features need to be taken into account. These will be discussed below.

2.3.2. *Pragmatic approaches*

2.3.2.1. *Pragmatic approaches using the internal context of utterances*

Two main methods have been put forward, one using psycholinguistic protocols and one using learning techniques. The former, which will be presented first, can be used to test certain linguistic hypotheses relating to irony by comparing them with the judgments made by human annotators, for example via Mechanical Turk type platforms. Annotators are presented with a set of texts or expressions and asked to judge whether or not they are ironic on the basis of a set of linguistic features or cues. The latter, presented later, are based on supervised or semisupervised machine learning techniques.

2.3.2.1.1. Psycholinguistic approaches

One of the first attempts at automatic irony detection was described by Utsumi (1996). However, the model in question was designed to treat a specific type of irony observed in interactions between a public speaker and audience members. Later, (Utsumi 2004) defined irony as “a pragmatic phenomenon whose processing involves complex interaction between linguistic style and contextual information”. Building on this definition, the author developed a psycholinguistic method for the detection of irony, sarcasm and humor. An empirical study was carried out in order to examine the capacity of humans to detect ironic, sarcastic or humorous utterances based on the style and context of given examples. Annotators were also asked to specify the polarity of each of the studied utterances.

Note that the main purpose of this experimental study was to validate (Utsumi 2000) own implicit display theory, which comprises three main aspects:

1) Irony requires an ironic environment, a proper situational setting in the context of discourse. This environment presupposes (1) the speaker’s expectation, (2) incongruity between expectation and reality and (3) that the speaker has a negative attitude toward the incongruity. Consequently, an utterance should be interpreted ironically in cases where the discourse situation is identified as being an ironic environment.

2) Ironic utterances are utterances that implicitly display an ironic environment. This is achieved by an utterance that (4) alludes to the speaker’s expectation, (5) includes pragmatic insincerity by violating one of

the pragmatic principles and (6) indirectly expresses the speaker's negative attitude, being accompanied by ironic cues.

3) Irony is a prototype-based category characterized by the notion of implicit display. The prototype of irony is an abstract exemplar that completely meets all the three conditions for implicit display. The degree of irony can be assessed by the similarity between the prototype and a given utterance with respect to three conditions (opposition, rhetorical questions, circumlocution).

The implicit display theory is thus based on three hypotheses (see Figure 2.2):

1) The degree of irony is affected by linguistic choice, not by contextual setting, and it is high to the extent that the properties of implicit display are satisfied.

2) The degree of sarcasm of an ironic utterance is affected only by linguistic style and it is high to the extent that the properties of implicit display are satisfied.

3) The degree of humor of an ironic utterance is affected by both linguistic style and context, and it is high to the extent that a discourse context is incongruous to the ironic environment or that the utterance is dissimilar to the irony prototype.

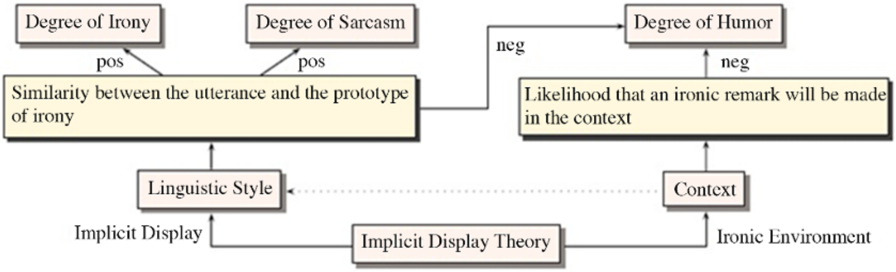


Figure 2.2. *General hypotheses for irony processing according to the implicit display theory (Utsumi 2004)*

This theory was validated by two experiments carried out on a study corpus, made up of 12 stories written in Japanese. The first experiment aimed to test the validity of the theory, examining the way in which linguistic style affects the degree of irony, sarcasm and humor. The linguistic style of irony was defined by two factors:

– **Sentence type**: three different types were identified: (1) opposition: an utterance in which the positive literal meaning is the opposite of the negative situation; (2) rhetorical question: an interrogative utterance in which the speaker rhetorically asks the addressee for an obvious fact; (3) circumlocution: a form of understatement, weakly linked to the speaker's expectation by a certain number of coherence relations (see section A.4).

– **Politeness level**: use or non-use of Japanese honorific titles, considered alongside the type of relation between the speaker and addressee (good or bad).

Note that 120 Japanese students were asked to study the 12-story corpus. Each participant was asked to read the stories one by one and to evaluate them, assigning two values to each story: a sarcasm value on a scale from 1 to 7, where 1 = not at all sarcastic and 7 = extremely sarcastic, and a humor value on a scale from 1 to 7, with 1 = not at all humorous and 7 = extremely humorous. After this stage, participants were asked to re-read the final sentence of each of the 12 stories and assign a third value from 1 to 7, where 1 = not at all ironic and 7 = extremely ironic). Finally, the annotators were asked to assign three values to each story, evaluating the absence or presence of sarcasm, humor and irony. The results of this first experiment showed that:

– oppositions were significantly more ironic and more sarcastic than circumlocutions, and more sarcastic than rhetorical questions;

– rhetorical questions were found to be significantly more sarcastic than circumlocutions;

– when the speaker was on good terms with the addressee, honorific utterances were rated as significantly more ironic and sarcastic than non-honorific utterances, but this difference disappeared when the speaker was on bad terms with the addressee;

– when the speaker and the addressee had a good relationship, circumlocutions without honorifics were rated as more humorous than those with honorifics but this difference was not observed when the relationship was bad.

The second experiment was designed to test the implicit display theory in terms of contextual effect on the degree of irony, sarcasm and humor. Two independent variables were examined: (1) situational negativity (whether the situation is weakly or strongly negative) and (2) the ordinariness of the negative situation (whether the negative situation is usual or unusual).

The author selected eight of the 12 stories used in Experiment 1, and recruited 48 further Japanese students for the experiment. The results showed that:

– context may have an indirect influence on the degree of irony when the speaker communicates in an implicit manner;

– the addressee is less likely to notice a speaker's beliefs, and thus interprets utterances as less ironic, when the negative behavior is usual than when this behavior is unusual;

– for the degree of sarcasm, no significant effects or interactions were observed in the new analysis. This suggests that the speaker's expectation may be an important property in distinguishing between irony and sarcasm, given that sarcasm does not require a speaker expectation;

– ironic utterances in contexts in which the addressee's negative behavior is usual were perceived as more humorous than the same utterances made in contexts where the negative behavior is not usual.

Kreuz and Caucci (2007) also took a psycholinguistic approach. The authors collected a corpus of 100 historical, romance and science-fiction novels, chosen at random from a list of works containing the expression *said sarcastically*, in order to study the influence of lexical cues on the perception of sarcasm in English. A group of 101 students was asked to classify extracts into sarcastic/non-sarcastic groups. Each participant was assigned 35 extracts (of which 20 featured sarcasm) and asked to assess the probability that an author is writing sarcastically, on a scale from 1 to 7. The authors calculated an average value across all of the extracts for each participant. As expected, the scores for the sarcastic extracts were higher (mean = 4.85, standard deviation = 0.67) than those for the control set (mean = 2.89, standard deviation = 0.86). In parallel, a second group, made up of two experts, was asked to determine the importance of lexical factors for the perception of sarcasm. This was achieved through regression analysis using the following five cues: (1) the number of words in each excerpt, (2) the number of words in bold type in each excerpt, (3) the presence of interjections, (4) the presence of adjectives and adverbs, and (5) the use of exclamation and question marks.

The results showed that the first three cues are relevant in detecting sarcasm, but that the two final cues have no visible effect.

2.3.2.1.2. Machine learning approaches

A summary of the pragmatic features that are most widely used for automatic detection is shown in Table 2.4.

This body of work may be split into two main groups: research based on online reviews, such as product or movie reviews, and research making use of tweets.

Irony in online reviews. Tsur *et al*. (2010) harvested a corpus of 66,000 Amazon product reviews in English, which they used to present their SASI algorithm *(Semi-Supervised Algorithm for Sarcasm Identification)* for comment classification. This algorithm uses two feature types. The first type of features are based on patterns constructed automatically using an algorithm designed by Davidov and Rappoport (2006), reflecting the main subject of discussion (generally product or company names), enabling the separation of frequent words and content words. The second type of features are lexico-syntactic, relating to aspects such as phrase length (in words), the number of exclamation, question and quotation marks in the sentence and the number of capitalized words.

The combination of all of these features resulted in an F-measure value of 82% (three annotators worked on each sentence). Punctuation signs were the weakest predictor, with an F-measure of 28.1%. Pattern-based pragmatic features related to product type, manufacturer name, comment author, etc., resulted in an improvement in review classification results, with an F-measure value of 76.9%. The combination of surface and pragmatic features maximizes performance in classification, highlighting the importance of pragmatic features in inferring figurative language.

Online reviews were also used by Reyes and Rosso (2011), who harvested a corpus of 8,861 ironic reviews from Amazon.com (AMA) and Slashdot.com (SLA)[7]. The authors proposed a six-feature model, including the following pragmatic features:

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

Research groups	Corpus	Features	Results
(Tsur *et al*. 2010)	Amazon reviews (66,000)	Comment frequency, product type, company, title, author, length, punctuation, quotes, capitalization	F-measure = 82%
(Reyes and Rosso 2011)	Amazon and Slashdot.com reviews (8,861)	n-grams, POS n-grams, funny profiling, positive/negative profiling, affective profiling, pleasantness profiling	F-measure = 75.75%
(Reyes and Rosso 2014)	Movie reviews (3,400), book reviews (1,500) and press articles (4,233)	Signature, emotional scenarios, unexpectedness	–
(Buschmeier *et al*. 2014)	Amazon reviews (1,254)	Imbalance, hyperbole, citation, punctuation, pos/neg polarity, interjection, emoticons, bag of words	F-measure = 74.4%
(Gonzalez-Ibanez *et al*. 2011)	Twitter (2,700)	Unigrams, dictionary, wordNet, interjection, punctuation, positive/negative emotion, response to another user	Accuracy = 71%
(Reyes *et al*. 2013)	Twitter (40,000)	Signatures, unexpectedness, style and emotional scenarios	f-measure = 76 %
(Barbieri and Saggion 2014b)	Twitter (40,000)	Frequency of rare words, synonyms, gap between synonyms and punctuation	F-measure = 76%
(Barbieri *et al*. 2014)	Twitter (60,000)	Frequency of rare words, synonyms, gap between synonyms, punctuation	F-measure = 62%
(Joshi *et al*. 2015)	Twitter (12,162) and forum discussions (1,502)	Unigrams, capitalization, emoticons, punctuations, implicit incongruity, explicit incongruity	F-measure = 61%

Table 2.4. *Summary of the main pragmatic approaches using the internal context of utterances for irony/sarcasm detection*

1) **funny profiling**: used to characterize documents in terms of humorous properties, identified using:

– *stylistic characteristics*: according to the experiments reported in (Mihalcea and Strapparava 2006), these characteristics were obtained by collecting all the words labeled with the tag “sexuality” in WordNet Domains (Bentivogli *et al*. 2004);

– *centeredness*: used to consider social relationships (words retrieved from the WordNet lexicon Miller (1995));

– *specific keywords*: this value is calculated by comparing word frequencies in the ironic documents against their frequencies in a reference corpus (in this case, Google N-grams (Reyes *et al*. 2009));

2) **positive/negative profiling**: used to indicate the communication of negative opinions using literally positive elements;

3) **affective profiling**: *WordNet-Affect* was used to obtain affective terms;

4) **pleasantness profiling**: the English affect dictionary (Whissell 1989) was used. Each entry includes a manually assigned pleasantness score from 1 (unpleasant) to 3 (pleasant).

These features were used for learning purposes in three different classifiers (*naive Bayes, support vector machine* (SVM) and decision trees) using a (*10-fold cross-validation*) configuration. Most classifiers obtained an accuracy value greater than 70% with a maximum accuracy score of 75.75% attained by an SVM classifier using the AMA subcorpus. For this subcorpus, the best performance was obtained using bag of words (trigram) features, pleasantness and funny profiling; for the SLA subcorpus, pleasantness profiling and 5-grams produced the best results.

The model proposed in Reyes and Rosso (2011) was extended further in Reyes and Rosso (2014). Three corpora were used in this case, including two for irony and one for satire: a corpus of movie reviews (*movies 2*) developed by Pang and Lee (2004) and Pang *et al*. (2002); a corpus of book reviews, collected by Zagibalov *et al*. (2010); and a corpus of satirical articles, collected by Burfoot and Baldwin (2009). The aim of the proposed model was to establish the probability of irony for each document and sentence in the corpus. It comprises three conceptual layers:

1) **signature**: this includes three features:

- pointedness: focused on the detection of explicit markers, specifically punctuation marks, emoticons, quotes and capitalized words;

- counterfactuality: detection of implicit marks, i.e. discursive terms that hint at opposition or contradiction in a text, such as *nevertheless*, *nonetheless* or *yet*;

- temporal compression: identification of elements related to opposition in time, i.e. terms that indicate an abrupt change in a narrative. These elements are represented by a set of temporal adverbs such as *suddenly, now* and *abruptly*;

2) **emotional scenarios**: again, this includes three features:

- activation : refers to the degree of response, either passive or active, that humans exhibit in an emotional state;

- imagery: quantifies how easy or difficult is to form a mental picture for a given word;

- pleasantness: measures the degree of pleasure suggested by a word;

3) **unexpectedness**: this covers two features:

- temporal imbalance: used to reflect the degree of opposition in a text with respect to the information profiled in the present and past tenses;

- contextual imbalance: used to capture inconsistencies within a context.

The results show that the probability of a document being ironic is higher in movie reviews. Furthermore, documents with positive polarity have a higher probability of including figurative content (irony, sarcasm, satire or humor).

Buschmeier *et al*. (2014) used a corpus developed by Filatova (2012), which comprised 437 sarcastic and 817 non-sarcastic reviews of products on Amazon, to propose a set of features for automatic irony detection. These include imbalance between overall text polarity and star rating, hyperbole, quotes, positive/negative word sequences followed by at least two exclamation or question marks, positive/negative word sequences followed by "...", interjections, emoticons and bag of words.

Irony in tweets. The development of social networks, notably Twitter[8], has provided researchers with a new online data source. The popularity of

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

Twitter worldwide and the large numbers of tweets published on different subjects every day make it an ideal source of data for studying figurative language phenomena.

Gonzalez-Ibanez *et al*. (2011), for example, collected a corpus of 2,700 tweets in English, of which 900 were sarcastic (containing the hashtags *#sarcasm, #sarcastic*), 900 tweets with positive polarity (*#happy, #joy, #lucky*) and 900 tweets with negative polarity (*#sadness, #angry, #frustrated*). Classification of 90 tweets taken at random from each class (sarcastic, positive and negative) gave an accuracy value of 62.59% using manual classification (three annotators) and a value of 57.41% using the SMO classifier (features listed in Table 2.4). For the binary sarcastic/non-sarcastic distinction, manual classification of 180 tweets resulted in an accuracy score of 66.85%, while SMO with the unigram feature achieved a score of 68.33%. The results of this study show that emoticons play an important role in helping humans to distinguish between sarcastic and non-sarcastic tweets. However, one annotator noted that contextual world knowledge was sometimes necessary in order to detect sarcasm, implying that information on user interactions and world knowledge are important in enabling automatic identification of sarcasm on Twitter.

In a similar framework, (Reyes *et al*. 2013) proposed a model for representing the most striking attributes of verbal irony in a text, thus enabling automatic detection. They collected a corpus of 40,000 tweets in English, of which 10,000 were ironic and 30,000 were non-ironic, spread evenly across the themes of education, humor and politics. A second corpus of 500 tweets containing the hashtag *#Toyota* and the emoticons ":)" (250 tweets) and ":(" (250 tweets) was collected in order to apply the proposed method to a real-world case. The tweets in this second corpus were not explicitly labeled as ironic by their authors, but were annotated by a group of 80 participants. The authors then defined a model to extract a set of features for irony detection. The proposed model included four conceptual features: signatures, unexpectedness, style and emotional scenarios. According to the authors, these features make it possible to capture both low- and high-level properties of textual irony based on conceptual definitions found in the literature.

Applied to the second Toyota corpus, the proposed approach succeeded in identifying 123 ironic tweets, compared to 147 detected by human annotators.

This proximity between the automatic and manual results shows the approach to be reliable.

The corpus constructed by Reyes *et al*. (2013) was reused by Barbieri and Saggion (2014b) in proposing a model made up of seven lexical features: word frequency, average frequency of words in vocabulary, structure (word count, character count, etc.), intensifiers, sentiments, synonyms and ambiguity.

Barbieri and Saggion (2014b) tested their proposed model using a supervised learning method on three different corpora. The first corpus was made up of 10,000 tweets containing the hashtag *#irony* and 10,000 tweets with the hashtag *#education*. The second corpus included 10,000 tweets containing the hashtag *#irony* and 10,000 with the hashtag *#humor*. Finally, the third corpus contained 10,000 tweets containing the hashtag *#irony* and 10,000 with the hashtag *#politics*. The F-measures obtained for the three corpora were as follows: 72% for the ironic versus education corpus, 75% for the ironic versus humor corpus, and 76% for the ironic versus politics corpus. A study of feature relevancy showed that the rare word frequency, synonym and punctuation features were most valuable in detecting irony. However, not all features performed equally well across the three corpora. This highlights the difficulty of defining sets of discriminating features for different themes.

Barbieri and Saggion (2014b) reused their own model in (Barbieri *et al*. 2014) in the context of sarcasm detection. In this case, they used a corpus of 60,000 tweets in English, split equally across six themes: *sarcasm, education, humor, irony, politics and newspapers*. All hashtags were removed prior to the automatic classification task. Binary classification was carried out via a supervised machine learning approach, using decision trees and the set of features proposed in Barbieri and Saggion (2014b). To evaluate the effectiveness of their proposed model in automatically detecting sarcasm, the authors split their corpus into five subcorpora with equal numbers of ironic/non-ironic tweets: sarcasm versus education, sarcasm versus humor, sarcasm versus irony, sarcasm versus newspapers and sarcasm versus politics. The results, expressed as F-measures, were 88% for the education and humor themes, 62% for the irony theme, 97% for the newspaper theme and 90% for the politics theme. These results show that the proposed model performs well in distinguishing between sarcastic and non-sarcastic tweets, but poorly in distinguishing between ironic and sarcastic tweets. The authors explain this by the fact that irony and sarcasm have similar structures in the proposed model; new features would need to be added in order to distinguish between the phenomena. They noted that sarcastic tweets tend to contain fewer adverbs than ironic tweets, but that these adverbs are more intense; additionally, sarcastic tweets contain more positive sentiments than ironic tweets. The distinction between irony and sarcasm was also explored by Sulis *et al.* (2016), who obtained an F-measure of 69.8%.

Finally, in a study of a mixed corpus containing data from Twitter and discussion forums, (Joshi *et al.* 2015) proposed an approach using two types of incongruity: explicit and implicit. Explicit incongruity is expressed by sentiment words with differing polarities, whereas implicit incongruity is expressed by sentences that express an implicit sentiment opposed to a word with positive or negative polarity. The authors proposed to resolve this problem using four groups of features: (1) lexical: unigrams; (2) pragmatic: capitalization, emoticons, punctuation; (3) implicit incongruity, and (4) explicit incongruity: the number of incongruities between positive and negative sentiments, longest positive/negative sequence, number of positive words, number of negative words and overall polarity of the text. A supervised machine learning method was used based on LibSVM. The results obtained (F-measures) were better than those presented by Riloff *et al.* (2013) and Maynard and Greenwood (2014).

2.3.2.2. *Pragmatic approaches using the external context of utterances*

A brief overview of the most widely used pragmatic features is provided in Table 2.5.

Proposing a classification strategy for verbal irony, (Wallace *et al.* 2015) used a corpus made up of comments on political articles, harvested from reddit.com and used by Wallace *et al.* (2014) in the context of an annotation campaign. This study corpus is made up three subsets of comments. The first comprises 1,825 comments, of which 286 are annotated as ironic. The second contains 996 political comments, 154 of which are ironic. The third subcorpus is made up of 1,682 comments on the theme of religion, of which 313 were annotated as ironic. The proposed method uses four feature types:

– **sentiment**: the inferred sentiment (negative/ neutral or positive) for a given comment;

– **subreddit**: the subreddit (e.g. progressive or conservative; atheism or Christianity) to which a comment was posted;

– **NNP**: noun phrases (e.g. proper nouns) extracted from comment texts;

– **NNP+**: noun phrases extracted from comment texts and the thread to which they belong (e.g. the title of an image accompanying the comment).

Research group	Corpus	Features	Results
(Wallace 2015)	Comments on political articles (2,821) and forum discussions (1,502)	Sentiment, subreddit (topic), noun phrase	–
(Bamman and Smith 2015)	Twitter (19,534)	n-grams, POS, sentiment, intensifier, author profile, discussion history, etc.	Accuracy = 85.1%
(Joshi *et al.* 2016)	Book extracts from GoodReads (3 629)	Features taken from (Liebrecht *et al.* 2013) Gonzalez-Ibanez *et al.* 2011, Buschmeier *et al.* 2014 and Joshi *et al.* 2015)	F-measure = 81.19%

Table 2.5. *Overview of the main pragmatic approaches using the external context of utterances for irony/sarcasm detection*

The proposed method was tested on all three corpora. The results showed an increase in the mean recall value (between 2% and 12% depending on the corpus) compared to the baseline (bag of words approach).

Bamman and Smith (2015) collected a corpus of 19,534 tweets of which half were sarcastic (*#sarcasm, #sarcastic*). This corpus was used in the context of automatic sarcasm detection using four feature types:

– **tweet features**, including a set of nine different features: (1) word unigrams and bigrams, (2) cluster unigrams and bigrams, (3) dependency bigrams, (4) parts of speech, (5) pronunciation, (6) capitalization, (7) whole tweet sentiment, (8) tweet word sentiment and (9) intensifiers;

– **author features**: (1) author historical salient terms, (2) author historical topics, (3) profile information, (4) author historical sentiment and (5) profile unigrams;

– **audience features**: (1) combination of the features from the "author features" type for the addressee, (2) author/addressee interaction topics and historical communications between author and addressee;

– **environment features**: (1) interaction between a target tweet and the tweet to which it responds, in terms of word pairings in the two tweets, and (2) unigram features of the original message, to capture the original linguistic context to which a tweet is responding.

Bamman and Smith (2015) used binary logistic regression with cross-validation in order to automatically classify tweets as sarcastic/non-sarcastic. The first feature type, tweet features, gave an average accuracy value of 75.4%; the addition of pragmatic features (discussion history) increased this value to 77.3%. The combination of tweet features and audience features produced a score of 79%. Even better results were obtained with a combination of tweet features and author features: 84.9%. The combination of all of these feature types resulted in an additional improvement, producing an accuracy value of 85.1%. These results prove that surface features alone are not sufficient to infer sarcasm in messages, and that the external context of tweets is relevant in maximizing the performance of automatic sarcasm detection systems.

Joshi *et al*. (2016) collected a corpus of 3,629 sarcastic and non-sarcastic tweets from the GoodReads website[9]. This corpus was used as part of an automatic sarcasm detection approach, in which (Joshi *et al*. 2016) used all of the features proposed by Liebrecht *et al*. (2013), Gonzalez-Ibanez *et al*. (2011), Buschmeier *et al*. (2014) and Joshi *et al*. (2015), adding new features based on word embeddings (e.g. the maximum score of the most dissimilar word pair). The highest F-measure (81.19%) was obtained by combining the features from (Liebrecht *et al*. 2013) with the new proposed features, obtained using the dependency weights approach[10].

2.4. Automatic detection of metaphor

Most work on the automatic detection of figurative language has focused on irony and sarcasm. However, some authors have studied the case of

[9] www.goodreads.com/.

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

metaphor, comparison and humor. As in the case of irony, work on metaphor has concentrated on surface and semantic features (Kintsch 2000, Bestgen and Cabiaux 2002) and on the internal context of utterances (Gedigian *et al*. 2006, Oliveira and Ploux 2009, Huang 2014, Macwhinney and Fromm 2014, Tsvetkov *et al*. 2014). The use of external context is a much more recent development, beginning with the work of (Jang *et al*. 2015, Do Dinh and Gurevych 2016, Su *et al*. 2017) and (Goode *et al*. 2017).

2.4.1. *Surface and semantic approaches*

According to the definitions presented in Chapter 1, metaphor may be considered as a comparison. Work on the automatic detection of this phenomenon has shown that automatic detection is not an easy matter, and that many different factors need to be taken into consideration.

One of the first attempts to propose a model for the automatic processing of metaphors was described by Kintsch (2000). The author noted that the understanding of metaphor implies an interaction between the meaning of a subject and the terms used to communicate the metaphor. Based on this hypothesis, he proposed a model using the interactive understanding of metaphor interpretation. For example, in My lawyer is a shark, the model consists of identifying lawyer as the topic and shark as the vehicle; in this case, the properties a shark (e.g. bloodthirsty or vicious), which might be assigned to a lawyer, must be selected.

To implement this model, (Kintsch 2000) began by identifying semantic features involved in communicating the meaning of a metaphor, proposing a selection algorithm with the following steps:

– construct a high-dimensional semantic space from the analysis of co-occurrences in a corpus of text using latent semantic analysis (LSA);

– represent the meaning of each word as a vector;

– measure the similarity between words by calculating the cosine of the vectors representing the words in question (the value of the cosine tends toward 1 with increasing similarity).

To determine the meaning of a predication, the algorithm aims to select those properties of a predicate which relate to the argument of the predication, selecting terms that are among the *n* closest neighbors of the predicate and the *k* closest neighbors of the argument. In this model, the only factor that changes when analyzing a metaphorical, rather than a literal, utterance is the parameter *n*. According to Kintsch (2000), while the 20 nearest neighbors are sufficient for a literal utterance, a value of 200 or even 500 may be necessary to understand a metaphorical utterance.

The approach proposed by Kintsch (2000) was criticized by Bestgen and Cabiaux (2002), who found that the arguments put forward in this approach were limited due to the fact that it only covered a few examples of metaphors. (Bestgen and Cabiaux 2002) proposed an alternative model based on LSA, applied to different types of literary metaphors in order to verify its effectiveness for expressions judged to be highly or weakly metaphorical by human readers, and to define a figurative intensity index. This was achieved using a corpus of 20 sentences containing metaphorical expressions, harvested from nine of Maupassant's short stories. Ten of these sentences expressed *live* metaphors, and 10 expressed *dead* metaphors. A dead metaphor (e.g. the hands of a clock) uses words, which have an accepted figurative meaning, while the words in live metaphors (about as much use as a chocolate teapot) are not generally used in this sense. In the case of the Maupassant extracts, (Bestgen and Cabiaux 2002) used the definitions in the Petit Robert dictionary to determine the live/dead status of metaphors.

The authors concluded that the model put forward in (Kintsch 2000) is effective in approximating the meaning of different types of literary metaphors, and that it may be used to derive an index for use in distinguishing between metaphorical and literal utterances.

Neither group went so far as to propose an automatic procedure for the automatic identification and interpretation of metaphors, as the model type and study corpus were not sufficient.

2.4.2. *Pragmatic approaches*

2.4.2.1. *Pragmatic approaches using the internal context of utterances*

In 2006, (Gedigian *et al.* 2006) proposed an automatic approach to metaphor detection using the internal context of utterances. They collected a corpus of articles published in the *Wall Street Journal*. Manual annotation was carried out for verbal targets associated with three subjects – motion in space, handling and health. The authors also labeled metaphorical targets, the literal meaning of targets and targets for which no decision could be reached. The annotation phase revealed that 90% of targets were used metaphorically. The proposed system achieved an accuracy score of 95.12%.

Oliveira and Ploux (2009) proposed a method for automatic metaphor detection in a parallel or comparable corpus of texts in French and Portuguese. The study corpus was split into three subcorpora. The first literary subcorpus was made up of around 200 20th Century novels in French or Portuguese, with their translations into the other language. The second subcorpus was made up of newspaper or magazine articles published between 1997 and 2001. The third contained European treaties. The authors used the ACOM (*Automatic Contexonym Organizing Model*) proposed by Hyungsuk *et al*. (2003) to calculate the distance between the contexts of use of the most generic terms relating to an expression (metaphorical or otherwise). The aim was to use the results from this model as a criterion for automatic metaphor detection.

Macwhinney and Fromm (2014) used the multilingual TenTen corpus (in English, Farsi, Russian and Spanish), which contains around 10 billion words for each language. These are all lemmatized and labeled (POS and dependency relations between words in source and target domains). The authors focused on the subject of economic inequality. Their aim was to obtain a system with the capacity for automatic source and target domain detection. They used the SketchEngine tool to construct collections of metaphorical examples for each language. Evaluation was carried out for English alone. The proposed system, *WordSketch*, obtained a precision score of 0.98 and a recall of 0.86, an improvement on the results obtained using CSF (Tsvetkov *et al*. 2014), TRIPS (Wilks 1978), VerbNet (Baker *et al*. 2003) and the ontology constructed as part of the Scone project[11].

Within the field of metaphor treatment in a multilingual context, (Tsvetkov *et al*. 2014) used a new corpus featuring the same languages as (Macwhinney and Fromm 2014) to propose an approach for automatic metaphor detection based on two syntactic structures: subject-verb-object (SVO) and adjective-noun (AN). The proposed approach uses three feature types:

[11] www.cs.cmu.edu/∼sef/scone/.

1) **abstractness and imageability**: most abstract things are hard to visualize. These features have been shown to be useful in detecting metaphors;

2) **supersenses**: coarse semantic classes originating in WordNet (15 classes for verbs and 26 classes for nouns);

3) **vector space word representations**: used to represent words in vector form using unsupervised algorithms.

Applying this approach to the English corpus gave an accuracy result of 82% for metaphor detection with SVO and 86% for metaphor detection with AN.

Huang (2014) addressed a specific type of metaphor in a social network context: *non-conventionalized* (non-stylized) metaphors. He collected a corpus of messages from an online breast cancer support page, Breastcancer.org, along with public user profiles. The corpus was used to implement a model based on JGibbLDA[12]. No data are available concerning the performance of this model.

Jang *et al*. (2015) used Huang's corpus to detect metaphors using the global context of discourse. They proposed an approach based on global contextual features (semantic category, topic distribution, lexical chain, context tokens) and local contextual features (semantic category, semantic relatedness, lexical abstractness, grammatical dependencies). Logistic regression was used for classification. The results showed that local contextual features perform better than global contextual features, with an accuracy score of 86.3%.

2.4.2.2. *Pragmatic approaches using the external context of utterances*

Jang *et al*. (2015) used the same corpus as (Jang *et al*. 2015) and (Huang 2014) to study the influence of situational factors (events linked to cancer: diagnosis, chemotherapy, etc.) on metaphor detection. They used the approach proposed by Wen *et al*. (2013) to extract the dates of cancer-related events for each user based on their public message history. In this way, they were able to compile a lot of terms used either metaphorically or literally in the study corpus. An SVM classifier was used, with the following features:

[12] A Java implementation of *Latent Dirichlet Allocation* (LDA), using Gibbs sampling to estimate parameters and for inference: http://jgibblda.sourceforge.net/.

(1) a binary feature, indicating whether a message was published during the critical period of each event; (2) a feature indicating the number of months separating the message date from the date of the event concerned by the message and (3) a binary feature indicating whether or not a message originated during a critical period for one of the events associated with a given method. The highest level of accuracy (83.36%) was obtained by combining features (1) and (2) with unigrams.

Do Dinh and Gurevych (2016) proposed a metaphor detection approach using neural networks and vector representations of words. They used a Multi-Layer Perceptron (MLP) of the feedforward type. They treated the metaphor detection issue as a tagging problem, adapting and extending the named entity recognition model constructed using the *Python deep learning library Theano* library, which was created by Bastien *et al.* (2012) as part of the Theano project. For network learning, they used the *stochastic gradient descent* (SGD) algorithm with log-likelihood. Experiments were carried out using pre-trained 300-dimensional word embeddings, created using word2vec[13] over the whole of Google News. Learning and test corpora were selected from the *VU Amsterdam Metaphor Corpus* (VUAMC) [14], in which each word is labeled with both a literal and metaphorical meaning. An F-measure of 56.18% was obtained using this approach.

Su *et al.* (2017) propose an approach for the automatic detection of nominal metaphor and for metaphor interpretation based on semantic relatedness. They make use of the fact that nominal metaphors consist of source and target domains, and that these domains are less related in the case of metaphor than in the literal case. The proposed metaphor detection and interpretation process therefore involves localizing concepts and calculating the semantic relatedness of these concepts. Each word/concept is represented by a vector, and semantic relatedness is calculated by comparing concept vectors with the cosine similarity value. After comparing the semantic relatedness of two concepts, the system consults WordNet to check for the existence of a hyponymy or hypernymy relation between the concepts in question. If a relationship of this type exists, then the system considers that these two concepts in the same sentence have a literal, non-metaphorical

[13] https://code.google.com/p/word2vec/.

[14] www.vismet.org/metcor/search/showPage.php?page=start.

meaning. The proposed approach was tested on two different corpora: the *Reader Corpus*[15], in Chinese, and an English corpus extracted from the *BNC Corpus*. The best accuracy scores for automatic detection were 0.850 for Chinese and 0.852 for English.

The second problem discussed in (Su *et al*. 2017) is the automatic interpretation of metaphor. Based on the hypothesis that the interpretation of metaphors is dependent on the abstract translation of an expression, the authors surmised that the source and target domains of a metaphor must come from two domains that are different but present similarities. In other terms, a metaphorical interpretation is a threefold cooperation between source and target domains: (1) the source and target share common properties; (2) the properties of the source and the target present certain similarities; (3) the target corresponds to one of the properties of the source domain. All properties of the source domain were extracted from the *Property Database*[16] and *Sardonicus*[17]. The test corpus included 100 metaphorical usages in Chinese and 100 metaphorical usages in English harvested from the Internet, newspapers, blogs and books. The interpretation was evaluated by five human annotators, who assigned values from 1 (highly unacceptable) to 5 (highly acceptable). Given an interannotator agreement of $kappa = 0.39$, all evaluations with an acceptability value of less than 3 were considered to be wrong and were eliminated. This resulted in an accuracy value of 87% for Chinese and 85% for English.

Current work extends beyond the detection of metaphors to consider ways in which metaphor detection may be used in more complex tasks, such as event detection. (Goode *et al*. 2017) studied blog behavior alongside metaphors in order to generate signals for event detection. They used a corpus of 589,089 documents collected from political blogs in Latin America. Metaphors in the corpus were identified using a metaphor detection system developed as part of the IARPA project[18]. Event detection was carried out using three feature types: (1) word count; (2) publication frequency and (3) the usage frequency of a given political metaphor. Blogs with strong grouping behaviors were more likely to coincide with events of interest than those with constant publication

[15] www.duzhe.com.

[16] A database developed by the NLP Lab at Xiamen University.

[17] http://afflatus.ucd.ie/sardonicus/tree.jsp.

[18] www.iarpa.gov/index.php/research-programs/metaphor.

rates. In other words, high levels of publication on a blog on a given date may indicate the existence of an important event.

2.5. Automatic detection of comparison

Mpouli and Ganascia (2015) studied another form of figurative language, comparison, which is similar to metaphor; the difference lies in the fact that comparison makes explicit use of comparative words (see Chapter 1). The authors proposed an algorithm using a surface parser (chunker) and manual rules in order to extract and analyze pseudo-comparisons in texts.

Figurative comparisons in texts were identified using a three-step process: (1) extraction of comparative and pseudo-comparative structures from a text; (2) identification of the components of these structures and (3) disambiguation of the structures in question.

Two types of figurative comparisons were considered. *Type I* concerned those introduced by comparative words (like, such as, just as, etc.), while *Type II* concerned comparisons based on adjectives (better than, worse than, similar to, etc.), verbs (look like, seem to, makes one think of, etc.), suffixes or prepositional locutions (in the manner of, in the image of, etc.). Only structures of the form marker + noun syntagm or marker, etc. noun syntagm, in which the comparer is not a subject, were extracted.

The proposed algorithm was tested using a manually annotated corpus of prose poems. The results obtained using this approach were better than those produced by the *Berkeley Parser* in terms of verb and comparer detection (with precision values of 52.8% and 96.7%, respectively), but poorer in terms of compared element and adjective detection.

2.6. Automatic detection of humor

Merriam-Webster defines humor as, among other things, *that quality which appeals to a sense of the ludicrous or absurdly incongruous: a funny or amusing quality*. The detection of humor has been addressed by a number of authors, including (Mihalcea and Strapparava 2006, Purandare and Litman 2006, Sjöbergh and Araki 2007, Taylor 2009, Raz 2012, Radev *et al*. 2015, Yang *et al*. 2015, Bertero *et al*. 2016, Bertero and Fung 2016), whose work is presented in this section.

Purandare and Litman (2006) analyzed conversations from the TV show *Friends*[19], examining acoustic, prosodic and linguistic characteristics and studying their utility in automatic humor recognition. They used a simple annotation scheme to automatically label passages followed by laughter as being humorous (43.8% of passages were found to be humorous).

The authors defined a set of acoustic and prosodic features (pitch, energy, timing) and other features (lexical, word count, speaker).

Automatic classification was carried out using supervised machine learning with a decision tree. An accuracy value of 64% was obtained using all features.

Bertero *et al.* (2016) presented a comparison of different supervised learning methods for humor detection in a corpus made up of audio recordings from the TV show *The Big Bang Theory*[20]. Two feature sets were defined: acoustical features and linguistic features (lexical, syntax, structure, sentiment, antonyms and speaker).

These features were used with three classifiers: *conditional random field* (CFR), *recurrent neural network* (RNN) and *convolutional neural network* (CNN). The best results were obtained using the CNN classifier, with an F-measure of 68.5% and an accuracy score of 73.8%.

Mihalcea and Strapparava (2006) based their approach on characteristics of humor identified in the field of linguistics. A corpus of 16,000 humorous phrases and a corpus of non-humorous phrases was collected online. The authors obtained an accuracy score of 96.95% using a *naive Bayes* classifier with stylistic features often found in humor (alliteration, antonyms, slang) and content-based features.

Certain studies have shown that understanding is not necessary in order to recognize humor. A set of surface features is therefore sufficient for automatic detection. For example, Sjöbergh and Araki (2007) used their own classification algorithm, in which a threshold value is calculated for each feature in order to separate training examples into two groups (humorous and non-humorous). This threshold should result in the lowest possible mean

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

[20] bigbangtrans.wordpress.com.

entropy. To classify each new example, a feature check is carried out on the new input and on both example groups. If the level of correspondence is highest between the new example and the humorous group, then the new example is considered to be humorous and vice versa. The proposed features were grouped into five types: text similarity, joke words (words commonly found in humorous texts), ambiguity average number of word meanings), style (negation, repetition, pronouns, antonyms etc.) and idiomatic expressions. An accuracy score of 85.4% was obtained.

As in the case of work on figurative language, humor detection has also been addressed based on social media content. Raz (2012) proposed an approach for the automatic detection of humor in a tweet corpus. A corpus of funny tweets was collected from a websites[21]. The author of the study proposed a set of features of varying types: syntactic, lexical, morphological (verb tense, etc.), phonological (homophony, in order to recognize puns), pragmatic (number of results returned by a search engine for the verbs present in the tweet) and stylistic (emoticons, punctuation). Unfortunately, this approach has not been evaluated.

Radev *et al*. (2015) chose to study a different type of corpus, made up of the captions from 298,224 cartoons published in *The New Yorker*.

The authors developed more than a dozen unsupervised approaches to classify these captions. The first group of methods was based on originality: for example, the *LexRank* algorithm [22] was used to identify the most central caption, and the *Louvain* graph-based classifier proposed by was used to group captions by theme. The second group was content based: for example, *Freebase*[23] was used to tag noun phrases in captions, and polarity was annotated using Stanford CoreNLP[24] . The third and final group was made up of generic methods, such as the use of syntactic complexity as proposed by Charniak and Johnson (2005).

The three methods were evaluated using *Amazon Mechanical Turk* (AMT). Each AMT microtask consisted of a cartoon along with two captions, A and B.

[21] www.funny-tweets.com.

[22] 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/.

Annotators were asked to identify the funniest caption. The results showed that methods based on negative sentiments and lexical centrality performed best in detecting the funniest captions.

Yang *et al.* (2015) used the corpus created by Mihalcea and Strapparava (2005) and proposed four feature types, respecting the following latent semantic structure: incongruity (disconnection, repetition), ambiguity (possible meanings according to WordNet), interpersonal effect (positive/negative polarity, strong/weak subjectivity) and phonetic style (alliteration, rhyme).

Using these features, the authors applied the random forest classification algorithm, obtaining an accuracy value of 85.4%. They concluded that the detection of humor and associated markers is dependent on understanding the meaning of the phrase and on external knowledge.

2.7. Conclusion

In this chapter, we presented the state of the art concerning automatic detection of figurative language, focusing first on proposed annotation schemes, and second on psycholinguistic or automatic approaches to detecting irony, sarcasm and satire (section 4.5), metaphor (section 2.4), comparison (section 2.5) and humor (section 2.6).

This work has shown that automatic analysis of irony and sarcasm is one of the main challenges encountered in natural language processing. Most recent studies have focused on detecting the phenomenon in corpora harvested from social networks such as Twitter; in tweets, authors may use specific hashtags (e.g. *#irony, #sarcasm*) to guide readers to an understanding of the image which they wish to convey (Gonzalez-Ibanez *et al.* 2011, Reyes *et al.* 2013, Barbieri and Saggion 2014a, Barbieri *et al.* 2014, Joshi *et al.* 2015, Bamman and Smith 2015). These hashtags are extremely valuable to researchers, as they make it easy to obtain annotated corpora for machine learning systems, classifying tweets as ironic or non-ironic.

Methods used to detect irony essentially draw on the linguistic content of texts, including the presence of punctuation, emoticons, positive or negative opinion words, etc. (Burfoot and Baldwin 2009, Tsur *et al.* 2010, Gonzalez-Ibanez *et al.* 2011, Reyes *et al.* 2013, Barbieri *et al.* 2014).

However, these methods rapidly cease to be effective when pragmatic or extra-linguistic knowledge is required to understand an ironic message. Pragmatic approaches, using external context, have recently been proposed in response to this problem (Bamman and Smith 2015, Wallace 2015, Joshi *et al*. 2016).

In this context, we propose a supervised learning approach in order to predict whether or not a tweet is ironic. To this end, we follow a three-step method:

1) analysis of pragmatic phenomena used to express irony, drawing on work in linguistics in order to define a multi-level annotation scheme for irony (Chapter 3);

2) application of observations from the annotated corpus to develop an automatic detection model for tweets in French, using both the internal context of a tweet, via lexical and semantic features, and external context, in the form of information available online (Chapter 4);

3) studying model portability for irony detection in a multilingual context (Italian, English and Arabic). We test the portability of our annotation scheme for Italian and Arabic, and test the performance of the feature-based automatic detection model for Arabic (Chapter 5).

3

A Multilevel Scheme for Irony Annotation in Social Network Content

3.1. Introduction

The aim of this chapter is to propose an annotation scheme for irony in a specific type of text, i.e. tweets. In Chapter 2 (section 2.2), we provided an overview of the different schemes that have been put forward for annotating tweets in Italian and English (Gianti *et al*. 2012, Shutova *et al*. 2013, Van Hee *et al*. 2016). These schemes are similar in that they all take a global approach to characterize irony, without considering linguistic or extra-linguistic cues at message level. The majority only include one level of annotation, characterizing tweets by figurative type (ironic/non-ironic), polarity (positive, negative or neutral) or, more rarely, the pragmatic device used to create irony (polarity reversal, hyperbole or euphemism).

Considering the work carried out in the field of linguistics on verbal irony markers in poems, novels, etc. (see section 3.3), we see that work on irony in social networks from a computational perspective has barely scratched the surface of the problem, without going into specifics. Our objective here is to discuss in detail, providing a fine-grained study of different markers and responding to the following questions:

– *can the different types of irony identified in the field of linguistics be found in a specific corpus harvested from social networks, such as Twitter?*

– *if so, which types are encountered most frequently?*

– *are these types marked explicitly?*

Automatic Detection of Irony: Opinion Mining in Microblogs and Social Media,
First Edition. Jihen Karoui; Farah Benamara and Véronique Moriceau.

– *what correlations exist between types of irony and these markers?*

– *how might these correlations be used for automatic detection purposes?*

We began our work by analyzing the different categories of irony proposed in linguistic studies (Attardo 2000b, Ritchie 2005, Didio 2007, Burgers 2010), retaining only those most appropriate for analyzing irony in tweets. To characterize and quantify the relevance of these categories, we propose a first multilevel annotation scheme and a corpus of tweets annotated using this scheme. Our approach, drawing on work in the field of linguistics, is intended to permit an in-depth study of expressions of irony in social networks at different levels of granularity, in terms of:

– **global messages**: ironic versus non-ironic;

– **irony type**: explicit versus implicit, showing the importance of context in understanding figurative language;

– **irony category**: each type of irony is associated with one or more categories, taking account of the pragmatic phenomena involved in the production of verbal irony;

– **linguistic cues**: each category of irony may be triggered by a set of specific linguistic markers.

This chapter is organized in the following manner. First, in section 3.2, we shall present our study corpus, FrIC (*French Irony Corpus*), followed by our annotation scheme in section 3.3. We shall describe those categories of irony used in our annotation approach alongside those proposed by linguists. In section 3.4, we describe our annotation campaign, followed by a presentation of the quantitative and qualitative results of this campaign. Particular attention will be given to the interactions between (1) types of irony triggers and markers, (2) irony categories and markers, and (3) the impact of external knowledge on irony detection. Our results show that implicit irony triggers represent a major challenge to be addressed in future systems.

Part of the FrIC was used in the context of the first opinion analysis and figurative language evaluation campaign, DEFT@TALN 2017, which we organized in collaboration with the LIMSI[1] (Benamara *et al*. 2017).

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

3.2. The FrIC

In the absence of a corpus of ironic tweets in French, we began by constructing our own, containing both ironic and non-ironic tweets. Initially, tweets containing the hashtags *#ironie* or *#sarcasme* were considered to be ironic, while messages without these markers were taken to be non-ironic.

We began the collection process by selecting a set of themes discussed in the media from Spring 2014 to Autumn 2016. The hypothesis underpinning our choice of themes was that the pragmatic context required in order to infer irony is more likely to be understood by annotators if it related to known current affairs, rather than to the specific context of personal tweets.

We selected 186 themes spread across nine categories (politics, sport, music, etc.). For each theme, we chose a set of keywords with and without hashtags, for example for politics (Sarkozy, #Hollande, UMP, etc.), health (cancer, flu), sport (#Zlatan, #FIFAworldcup, etc.), social media (#Facebook, Skype, MSN), artists (Rihanna, Beyoncé, etc.), television (TheVoice, XFactor), countries or cities (North Korea, Brazil, Paris, etc.), the Arab Spring (Marzouki, Ben Ali, etc.) and other more generic themes (pollution, racism, etc.). We then selected ironic tweets containing our keywords along with the *#ironie* or *#sarcasme* hashtag. Non-ironic tweets were selected in the same way (i.e. based on the absence of *#ironie* or *#sarcasme*).

Twitter's API was used to collect the corpus. After harvesting, we removed replications, retweets and tweets containing images, as we felt that the latter were more likely to contain situational irony (illustrated by the image), which is harder to detect automatically. After filtering, we were left with a corpus of 18,252 tweets, of which 2,073 were ironic and 16,179 were non-ironic (Table 3.1). For the experiments described below, the *#ironie* and *#sarcasme* hashtags were removed.

An initial annotation step was carried out on a subset of the corpus in order to verify the reliability of the *#ironie* and *#sarcasme* hashtags. This task was carried out by two human annotators, who manually labeled 100 tweets (a mixture of 50 ironic and 50 non-ironic tweets) from which the *#ironie* and *#sarcasme* hashtag had been removed. The annotation phase resulted in a Cohen's kappa of $\kappa = 0.78$ when comparing these annotations with the reference hashtags. This figure indicates that the hashtags are relatively reliable.

Themes	**Ironic**	**Non-ironic**
TV shows	81	3,060
Economy	85	273
Generic	189	777
Cities or countries	245	805
Artists	232	192
Politics	1,035	10,629
Social networks	19	0
Health	3	32
Sport	178	411
Total	2,073	16,179

Table 3.1. *Distribution of tweets in the FrIC*

Most disagreements between the two annotators were due to the presence of negation (phrase (3.1)) or to the need for external knowledge not contained within the tweet in order to understand the ironic meaning (phrase (3.2)).

(3.1) C'est chez Hollande qu'il y a du Berlusconi vous ne trouvez pas. Un côté bounga-bounga non ?
(*President François Hollande has a touch of the Berlusconi about him, don't you think? A bit of a bunga-bunga side?*

(3.2) Qu'est-ce qui pourrait détruire notre monde ? — La Corée du Nord
(*What could destroy the world? - North Korea*)

Note that while many of the corpora of ironic/sarcastic tweets presented in the literature were gathered using the *#ironie* or *#sarcasme* hashtags, very few authors have verified the reliability of these hashtags (Hee *et al*. 2016).

3.3. Multilevel annotation scheme

3.3.1. *Methodology*

The first stage in defining our scheme was to study the different irony markers found in literature on linguistics. More than 126 markers have been identified, such as counter-factuality, exaggeration and exclamation (Tayot 1984, Attardo 2000b, Mercier-Leca 2003, Ritchie 2005, Didio 2007, Burgers 2010).

Table 3.2 provides an overview of these markers, focusing specifically on textual irony. In this table, we present a list of references for each category, along with one or two of the definitions provided in the literature. All of the definitions cited by the linguists in question are given in Appendix section A.1.

Irony markers	References	Definitions in ironic use
Metaphor	(Grice 1970, Kittay 1990, Song 1998)	According to (Kittay 1990), irony may be expressed through metaphor, a second order of meaning obtained when the characteristics and context of an utterance indicate to the addressee or reader that the first-order sense of the expression is unavailable or inappropriate.
Hyperbole	(Kreuz and Roberts 1993, Pougeoise 2001, Mercier-Leca 2003, Didio 2007)	According to (Didio 2007), irony may be expressed by hyperbole, a way of magnifying something to excess, through exaggeration.
Exaggeration	(Didio 2007)	Exaggeration is a figure which amplifies reality, or presents it in a way which assigns more importance to it than it really has.
Euphemism	(Muecke 1978, Fromilhague 1995, Seto 1998, Yamanashi 1998, Mercier-Leca 2003)	According to (Muecke 1978, Seto 1998), euphemism is a figure of style that consists of attenuating the expression of facts or ideas considered to be disagreeable in order to “soften” the reality.
Rhetorical question	(Muecke 1978, Barbe 1995, Burgers 2010)	According to (Burgers 2010), a rhetorical question is not a real question: the speaker does not expect to receive a response, as the answer is already know. Thus, a rhetorical question represents a point of view rather than a question.
Register shift	(Attardo 2000b, Haiman 2001, Burgers 2010)	According ton (Burgers 2010), a register shift is a sudden change in style. In utterances, register changes are expressed through the use of unexpected words belonging to a different register (e.g. the presence of informal words in formal text, or vice-versa). It may also take the form of a sudden change in the subject of a sentence, or of exaggerated politeness in a situation where this is not appropriate.
False logic/ontradiction	(Tayot 1984, Barbe 1995, Didio 2007)	According to (Didio 2007), contradictions in a discourse enable the addressee to understand the ironic meaning of text, based on the notion that a contradiction combines two utterances, which confirm and deny the same element of knowledge.

Oxymoron	(Gibbs 1994, Song 1998, Mercier-Leca 2003)	According to (Gibbs 1994, Song 1998, Mercier-Leca 2003), an oxymoron is a figure of construction based on an apparent logical contradiction. It is an opposing figure, identified at utterance level through the syntactic combination of two elements that form a semantic contradiction.
Paradox	(Tayot 1984, Barbe 1995, Mercier-Leca 2003)	According to (Mercier-Leca 2003), irony is based on a paradox, the striking nature of which is accentuated by asyndetic syntax (sparse use of logical connectors) which, through contrast effects, highlights the only coordination conjunction present in an utterance, for example "but".
Absurdity	(Didio 2007)	According to (Didio 2007), absurdity is expressed through illogical reasoning. It may be associated with a cominc or tragic reaction. Absurdity indicates something which is not in harmony with another thing or person.
Surprise effect	(Colston and Keller 1998, Didio 2007)	According to (Colston and Keller 1998), surprise is a frequent reaction when things do not go as expected. This surprise may be expressed as a verbal note of the contrast between expected and actual events.
Repetition	(Muecke 1978, Berntsen and Kennedy 1996, Burgers 2010)	According to (Burgers 2010), a writer may ironically repeat something said by another person earlier in a text, or, in the case of verbal interactions, in dialog. This type of repetition is known as co-text based repetition. In this case, an utterance or part of an utterance is repeated, ironically, in the same text (where the first usage was non-ironic).
Quotation marks	(Tayot 1984, Gibbs 1994, Attardo 2001, Burgers 2010)	According to (Gibbs 1994), quotation marks are used as a non-verbal gesture by many American speakers to express irony. The use of quotation marks indicates that the speaker is about to imitate the discourse or views of a cited individual, often to sarcastic effect.
Emoticons	(Tayot 1984, Kreuz 1996, Burgers 2010)	According to (Tayot 1984), an emoticon indicates intonation or mimo-gestuality (e.g. the British "tongue in cheek", or winking) used to mark irony in oral communication.
Exclamation	(Attardo 2001, Didio 2007, Burgers 2010)	According to (Attardo 2001), (Didio 2007) and (Burgers 2010), irony may be marked by exclamation, in oral communications, and by an exclamation mark in writing.
Capitalization	(Haiman 1998, Burgers 2010)	–
Strikethrough text and special characters	(Burgers 2010)	–

Table 3.2. *Different irony markers studied in the field of linguistics*

It is important to note that the irony categories presented in Table 3.2 were, for the most part, identified in literary texts (books, poems, etc.). Our first step was to verify their presence in a sample of 300 tweets from our corpus. Four main observations were made:

1) According to the general definition, irony expresses a contradiction between what is said and what is meant. We noted that tweet authors used two mechanisms to express this contradiction: (a) lexical cues in the text of the tweet and (b) lexical cues plus external pragmatic context. We thus defined two forms of contradiction: explicit, for case (a), and implicit, for case (b). Each type of contradiction may be expressed by different categories of irony.

2) Many categories can be grouped together as it is hard to distinguish between them in short messages – for example hyperbole and exaggeration, or metaphor and comparison.

3) Some categories are specific to literary texts and are not truly applicable to tweets, for example absurdity.

4) The categories of irony defined in the literature cannot all be considered on the same level. For example, quotation marks may be found in ironic tweets of the hyperbole or euphemism type. We thus decided to distinguish between categories of irony (hyperbole, euphemism, rhetorical question, etc.) and irony cues (punctuation, capitalization, etc.).

Based on these observations, we propose three levels of analysis: irony type (implicit/explicit), category of irony for each type, and the linguistic cues present in each category. Finally, eight categories and 10 cues were adopted. These are described below.

3.3.2. *Annotation scheme*

The proposed scheme comprises four levels, as shown in Figure 3.1.

3.3.2.1. *Level 1: tweet classes*

In this scheme, tweets are grouped into three classes:

– **ironic**: a tweet is ironic if it expresses verbal irony, situational irony, sarcasm, satire or humor (for example: *un truc avec DSK, mais quoi? Aucun site internet n'en parle. Sûrement parce que l'on ne sait rien de ce qu'il s'est réellement passé?* (Dominique Strauss-Kahn did something, but what? There's nothing about it anywhere online. Maybe because we don't actually know what really happened));

– **non-ironic**: a tweet is considered non-ironic if it does not correspond to any of the forms of irony cited above (for example: *l'écotaxe, c'est pour sauver la planète pas pour redresser la France et c'est une idée de Sarko. #idiotie* (the ecotax is intended to help the planet, not France, and it was Sarkozy's idea));

– **no decision**: tweets are placed into this class if it is not possible to decide whether or not the message is ironic (for example: *si cela ne vous donne pas envie de voter PS en 2012, je ne comprends plus rien à rien* (if that doesn't make you want to vote for the socialists in 2012, I don't know what will)).

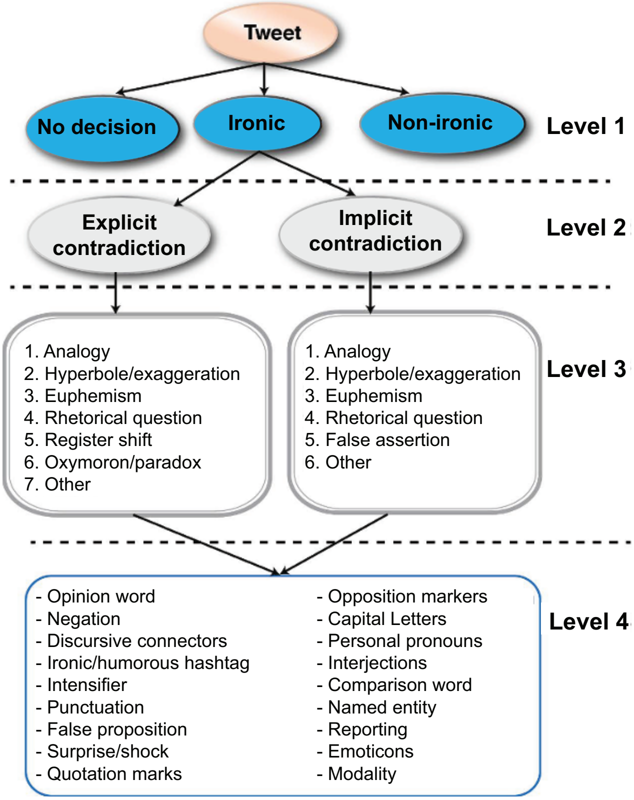


Figure 3.1. *Annotation scheme. For a color version of the figures in this chapter see, www.iste.co.uk/karoui/irony.zip*

3.3.2.2. *Level 2: types of irony*

Incongruity in ironic utterances, and particularly in tweets, often consists of an opposition between at least two propositions (or two words) P_1 and P_2. Propositions P_1 and P_2 may both form part of the internal context of an utterance (explicitly lexicalized), or one may be present with the other being implied. There are thus two means of deducing irony in tweets: first, based exclusively on the lexical cues found in an utterance, or second, by using these cues in conjunction with supplementary, external pragmatic context. This relates to the two forms of contradiction mentioned previously: explicit and implicit.

3.3.2.2.1. Explicit contradiction

Explicit contradiction may imply a contradiction between the words in a proposition P_1 and those in a proposition P_2, which either have opposing polarities, as in phrase (3.3), or are semantically unrelated, as in phrase (3.4). Explicit opposition may also result from an explicit positive/negative contrast between a subjective proposition P_1 and a situation P_2 describing an undesirable activity or state. Irony is deduced based on the hypothesis that the author and the reader have shared knowledge of a situation, which is judged to be negative according to cultural or social norms. For example, tweet (3.5) presumes that everybody expects their cell phone to ring loudly enough to be heard.

Tweets for which annotators do not require any external knowledge to understand a contradiction are labeled as ironic with explicit contradiction.

(3.3) [I love it]$_{P1}$ when my cell phone [stops working]$_{P2}$ just when I need it.

(3.4) [The Voice]$_{P1}$ is more important than [Fukushima]$_{P2}$ this evening.

(3.5) [I love it]$_{P1}$ when my cell phone [automatically turns the volume down]$_{P2}$.

3.3.2.2.2. Implicit contradiction

Implicit irony results from a contradiction between a lexicalized proposition P_1, describing an event or a state, and a pragmatic context P_2 external to the utterance in which P_1 is false, improbable or contrary to the author's intentions. Irony occurs because the author believes that their audience will detect the disparity between P_1 and P_2 based on contextual knowledge or shared common antecedents. For example, in phrase (3.6), the fact that is denied by P_1 allows us to infer that the tweet is ironic.

(3.6) The #NSA wiretapped a whole country. Shouldn't be a problem for #Belgium: [it's not a whole country]$_{P1}$.
$\longrightarrow$ $P2$: Belgium is a country.

3.3.2.3. *Level 3: categories of irony*

Explicit and implicit contradictions may be expressed in different ways, which we refer to as categories of irony. Many different categories have been defined in literature from the field of linguistics, and we used these as a basis to define the eight categories used in our annotation scheme. Three of these categories can only occur with a specific type of irony (noted *Exp* for explicit or *Imp* for implicit), while the remaining five categories may be found with either type of contradiction (marked *Exp/Imp*). These categories are not mutually exclusive: an ironic tweet may be associated with one or more categories.

Table 3.3 gives a summary of the main categories found in the literature, along with the eight categories selected for tweet annotation.

Each of our categories is presented below, illustrated with examples taken from our corpus.

3.3.2.3.1. Analogy$^{Exp/Imp}$

Analogy is a thought process in which a similarity between two elements of different types or classes is noted. In discourse, a comparison is an explicit analogy (see tweet (3.7)), while a metaphor is an implicit analogy (see tweets (3.8) and (3.9)).

In our annotation scheme, analogy is used in a broader sense to cover analogy, comparison and metaphor, three tools which imply a similarity between two entities relating to different concepts, domains or ontological classes, which may form the basis for a comparison.

(3.7) *(Exp)* Le dimanche **c'est comme** Benzema en équipe de France : il sert à rien... :D
*(Sunday **is like** Benzema in the French national team: pointless... :D).*

(3.8) *(Imp)* Pour une fois que je regarde la télé, c'est pour voir **Depardieu en député communiste.** #Savoureux.
*(The one time I watch TV, I see **Depardieu playing a communist congressman.** #Wowee.)*

(3.9) *(Imp)* On n'avait qu'à écouter ses déclarations des dernières années pour savoir que **Depardieu était en fait très belge**.
*(If you'd listened to anything he's said over the last few years, you'd know that **Depardieu is actually very Belgian).***

3.3.2.3.2. Hyperbole/Exaggeration$^{Exp/Imp}$

Hyperbole/exaggeration is a figure of style which consists of expressing an idea or sentiment in an exaggerated manner. It is often used to make a strong impression or make a point, as in phrases (3.10) and (3.11).

(3.10) *(Exp)* Le PS a **tellement bien** réussi que tt va moins bien : pollution, logement, sécurité #PARISledebat #Paris2014
*(The socialist party have **done such a great job** that everything's gone downhill: pollution, accommodation, security #PARISledebat #Paris2014)*

(3.11) *(Imp)* @morandiniblog C'est vrai que **c'est un saint** #Berlusconi, il ne mérite vraiment pas tout cet acharnement...
*(@morandiniblog Obviously #Berlusconi **is a saint**, he doesn't deserve all of this negative attention...)*

Categories: state of the art	**Our categories**	**Use**
Metaphor (Grice 1970, Kittay 1990, Song 1998)	Analogy$^{Exp/Imp}$ (metaphor and comparison)	Covers analogy, comparison and metaphor. Implies a similarity between two concepts or two entities from different ontological domains, on which a comparison may be based.
Hyperbole (Berntsen and Kennedy 1996, Mercier-Leca 2003, Didio 2007) Exaggeration (Didio 2007)	Hyperbole/ exaggeration$^{Exp/Imp}$	Expresses a strong impression or highlights a particular point.
Euphemism (Muecke 1978, Seto 1998)	Euphemism$^{Exp/Imp}$	Attenuates the effect of an expression or idea considered to be unpleasant in order to soften the reality.
Rhetorical question (Barbe 1995, Berntsen and Kennedy 1996)	Rhetorical question$^{Exp/Imp}$	A question asked in order to emphasize a point rather than to obtain a response (P_1: ask a question with the intention of obtaining a response, P_2: there is no intention of obtaining a response as the answer is already known).
Register shift (Haiman 2001, Leech 2016)	Register shift^{Exp}	A sudden change of subject/framework, use of exaggerated politeness in a situation where this is not appropriate, etc.
False logic (Didio 2007)	False affirmation^{Imp}	An affirmation, fact or event, which is not true in reality.
Oxymoron (Gibbs 1994, Mercier-Leca 2003) Paradox (Tayot 1984, Barbe 1995)	Oxymoron/paradox^{Exp}	Explicit opposition between two words. A paradox differs from a false affirmation in that the contradiction is explicit.
Situational irony (Shelley 2001, Niogret 2004)	Other$^{Exp/Imp}$	Humorous or situational irony (irony in which the incongruity is not due to the use of words, but to a non-intentional contradiction between two facts or events).

Table 3.3. *Categories of irony in our annotation scheme*

3.3.2.3.3. Euphemism$^{Exp/Imp}$

Euphemism is a figure of style used to reduce the effect of an expression or idea considered to be unpleasant in order to soften the reality (as in the case of *moins bien* (less good, translated as "downhill" in the example) instead of *worse* in tweet (3.10)).

3.3.2.3.4. Rhetorical question$^{Exp/Imp}$

The rhetorical question is a figure of style that takes the form of a question, but is asked in order to emphasize a point rather than to obtain a response, as in phrase (3.12).

(3.12) "Miss France c'est une compétition" **Non sérieux?** parce que je ne savais pas!
("Miss France is a competition". **Seriously?** I had no idea!)

3.3.2.3.5. Register shift^{Exp}

A register shift may take the form of a sudden change of subject/framework in a tweet, as in phrase (3.13), where the first phrase concerns the resignation of Duflot, a government minister, while the second concerns the period of Lent.

A contextual shift may also be observed through the use of exaggerated politeness in a situation where this is not appropriate, as in phrase (3.14), where the author is too polite for a normal conversation between friends (this is known as hyperformality).

Changes of context may also result from the use of polysemic words, where irony is triggered by the contrast between meanings in different context. For example, in French, "se rencontrer" implies interrogation in the case of suspects in a police investigation, but in a different context, "rencontrer" might mean spending time with an attractive woman.

(3.13) Duflot quitterait le gouvernement. **En plein carême, on ne peut même pas le fêter.** Décidément, elle embête jusqu'au bout... *soupire*
(I hear Duflot is leaving the government. **In the middle of Lent, so we can't even celebrate.** Annoying to the end... *sigh*)

(3.14) Vous pouvez m'accorder l'honneur d'écouter un à l'autre de vos prédictions fines
(Would you please do me the honor of listening to your respective predictions?)

3.3.2.3.6. False affirmation^{Imp}

This indicates that a fact or affirmation does not make sense in reality; the author is expressing the opposite of what they think, or saying something which is false in the context. External knowledge is thus required in order to understand the irony. For example, tweets (3.15)–(3.17) are ironic as the highlighted situations are absurd, false or impossible in reality. Note that tweet (3.17) is also an example of a rhetorical question.

(3.15) La #NSA a mis sur écoute un pays entier. Pas d'inquiétude pour la #Belgique: **ce n'est pas un pays entier**.
*(The #NSA wiretapped a whole country. Shouldn't be a problem for #Belgium: **it's not a whole country**)*

(3.16) @Vince75015 Les agences de notation **ne font pas de politique**.
*(@Vince75015 Credit ratings agencies **don't do politics**.)*

(3.17) @infos140 @mediapart Serge Dassault? Corruption? Non! Il doit y avoir une erreur. **C'est l'image même de la probité en politique**.
*(@infos140 @mediapart Serge Dassault? Corruption? No! That can't be true. **He's a paragon of political virtue**.)*

3.3.2.3.7. Oxymoron/paradox^{Exp}

This category differs from the false affirmation category only in that the contradiction is explicit, as seen in the use of two antonyms in the first sentence of phrase (3.10) (bien réussi vs moins bien) and the use of two opposing facts in phrase (3.18).

(3.18) Ben non! **Matraquer et crever des yeux**, ce **n'est pas violent** et ça respecte les droits!!! #assnat #polqc #ggi
*(No way! **Clubbing people and taking their eyes out isn't violent behavior** and it respects human rights!!! #assnat #polqc #ggi)*

3.3.2.3.8. Other$^{Exp/Imp}$

This category contains tweets that are judged to be ironic, containing explicit/implicit contradiction, but which do not fit into any of the categories above as they relate more closely to humor, satire or situational irony. A number of examples are shown below:

(3.19) Palme d'Or pour un film sur l'homosexualité le jour de la #manifpourtous #Cannes2013
(A movie about homosexuality was awarded the Palme d'Or on the same day as the #manifpourtous [mass demonstration against the legalization of same-sex marriage]#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!
(Air quality alert: citizens are advised to avoid biking to work at 9am... but a diesel car is fine!)

(3.21) Merci Hollande d'avoir sauvé le monde! Sans toi, la terre serait actuellement entrée en 3^{e} guerre mondiale
(Thank you for saving the world, [President] Hollande! Without you, the world would have plunged into WWIII)

3.3.2.4. *Level 4: Linguistic markers of irony*

As Table 3.2 shows, other forms of irony category have been identified in literature on linguistics, such as surprise and repetition. From a computational perspective, we chose to make a clear distinction between categories of irony – the pragmatic irony mechanisms defined in the previous section – and irony cues, a set of segments (words, symbols and propositions) that may trigger irony based on the linguistic content of a tweet alone.

This distinction also reflects the fact that cues may be found in specific categories of irony and may or may not be found in non-ironic tweets.

Table 3.4 shows the cues that we selected for our tweet annotation scheme. Note that 19 different markers were used in our study. Some have already been shown to be helpful as surface characteristics in irony detection: punctuation signs, capitalization, emoticons, interjections, negations, opinions and emotion words (Davidov *et al*. 2010, Gonzalez-Ibanez *et al*. 2011, Reyes *et al*. 2013).

Irony cues	**References**
Discursive connectors	–
Punctuation	(Tayot 1984, Wilson and Sperber 1992, Seto 1998, Attardo 2001, Didio 2007, Burgers 2010)
Opinion words	(Reyes and Rosso 2011, 2012)
Emoticons	(Tayot 1984, Kreuz 1996, Burgers 2010, Gonzalez-Ibanez *et al*. 2011, Buschmeier *et al*. 2014)
Contradiction markers	(Attardo 2000b, Didio 2007)
Capitalization	(Haiman 2001, Burgers 2010, Tsur *et al*. 2010, Reyes *et al*. 2013)
Intensifiers	(Liebrecht *et al*. 2013, Barbieri and Saggion 2014b)
Comparison words	(Veale and Hao 2010)
Modality	–
Negation	–
Citation	(Tayot 1984, Gibbs 1994, Attardo 2001, Burgers 2010, Tsur *et al*. 2010, Reyes *et al*. 2013)
Interjection	(Gonzalez-Ibanez *et al*. 2011, Kreuz and Caucci 2007)
Personal pronouns	–
Reporting verbs	–
Surprise/shock	(Didio 2007, Colston and Keller 1998)
Named entities	–
False proposition	(Tayot 1984, Attardo 2000b, Didio 2007, Barbe 1995)
Ironic or humorous hashtag	–
URL	–

Table 3.4. *Irony cues in our annotation scheme. Cues that were not discussed in the state of the art are highlighted*

We have added a number of new markers (in bold font in Table 3.4): **discursive connectors** (which may mark opposition, chains of argument and consequences); **modality**; **reporting verbs**, **negation**, **ironic or humorous hashtags**; **named entities** and **personal pronouns** (both of which may indicate that a tweet is personal or relates to a mediatized subject); **URLs** (the linked webpages provide contextual information that may help readers to detect irony); and, finally, **false propositions** that mention facts or events that do not correspond to reality (and which may contain negations). The final four markers may be valuable for automatic detection of implicit irony, for example by highlighting the need for external context.

For example, in tweet (3.22), the markers are negations (n', pas, non (French translation of "no")), punctuation (! or !!!), and the opinion word (violent), whereas in tweet (3.23), the markers are named entities (NSA, Belgium), negations (pas, n'...pas (French translation of "no")) and a false proposition (isn't a whole country).

(3.22) Ben **non!** Matraquer et crever des yeux, ce **n'**est **pas violent** et ça respecte les droits **!!!** #ironie
*(**No way!**Clubbing people and taking their eyes out **isn't violent** behavior and it respects human rights **!!!** #irony)*

(3.23) La #**NSA** a mis sur écoute un pays entier. **Pas** d'inquiétude pour la #**Belgique**: ce **n'**est **pas** un pays entier. #ironie
*(The #**NSA** wiretapped a whole country. **Shouldn't** be a problem for #**Belgium**: it's **not** a whole country.) #ironie*

3.4. The annotation campaign

In this section, we shall describe the procedure used to annotate tweets using our proposed scheme via the Glozz annotation tool[2].

3.4.1. *Glozz*

Glozz, developed as part of the Annodis project (Péry-Woodley *et al.* 2009), is a tool that provides a dedicated interface for annotation. Annotation

[2] www.glozz.org.

is carried out using a Glozz scheme, developed using the elements presented in the previous section.

Each tweet was annotated using Glozz for units and relations between units (where applicable). Relations provide the means of linking textual units within a tweet. We identified three relation types:

– **comparison relation**: consists of linking two units or parts of a text that are in comparison;

– **explicit contradiction relation**: links parts of the text that are in explicit contradiction to one another;

– **cause/consequence relation**: consists of connecting two parts of text where one is a case and the other is the consequence of the first.

Glozz requires multiple input files, notably a version of the proposed annotation scheme in Glozz input format, and produces an output file containing the different annotations created by the user.

3.4.2. *Data preparation*

A preliminary data processing step was required before beginning the annotation process. This consisted of preannotating tweets and generating the input files required by Glozz.

In the preannotation stage, we automatically annotated a set of cues in order to make the annotation process easier and faster. This concerned specific cues, namely punctuation, intensifiers, emoticons, opposition words, comparison words, personal pronouns and negation words.

Automatic preannotation for these linguistic markers was carried out using two lexicons: CASOAR[3] and EMOTAIX[4] for opinion and emotion words, intensifiers and interjections; a syntax analyzer, MEIT[5] was used for automatic annotation of named entities. The automatic annotations were corrected manually to add missing markers or rectify incorrect annotations.

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

[4] www.tropes.fr/download/EMOTAIX_2012_FR_V1_0.zip.

[5] http://alpage.inria.fr/statgram/frdep/fr_stat_dep_malt.html.

Additionally, the preannotation step included automatic assignment of a number of elements:

– a unique identifier for each tweet harvested using Twitter's API;

– an incremental identifier, making it easy for annotators to detect tweets requiring annotation;

– tweet publication date;

– the keyword used to collect the tweet.

This preannotation is possible if the data can be retrieved using the Twitter API, else a default null value is used for these attributes. The text of each tweet was saved to a Glozz input file, with the hashtags *#ironie* and *#sarcasme* removed.

3.4.3. *Annotation procedure*

For each tweet t, the annotation works as follows[6]:

a) Classify t as *ironic*/*non-ironic*. If annotators do not understand the tweet because of cultural references or lack of background knowledge, t can be classified into the *No decision* class.

b) If t is ironic, define its trigger type, i.e. explicit contradiction or implicit contradiction. To do this, annotators must look for two contradicting propositions P_1 and P_2 in the tweet. If these are found, then the trigger is explicit; otherwise, it is implicit.

c) Once irony type has been identified, specify the pragmatic devices used to express irony by selecting one or more categories (see section 3.3.2.3).

d) Identify text spans within the tweet that correspond to a predefined list of linguistic markers.

Annotators were asked to respect the following rules during the annotation process:

[6] The annotation manual is available at: github.com/IronyAndTweets/Scheme.

1) all cues should be annotated, whatever the type of irony (ironic with explicit contradiction, implicit contradiction, non-ironic);

2) ironic tweets with explicit or implicit contradiction must be placed into one of the categories belonging to the relevant irony type;

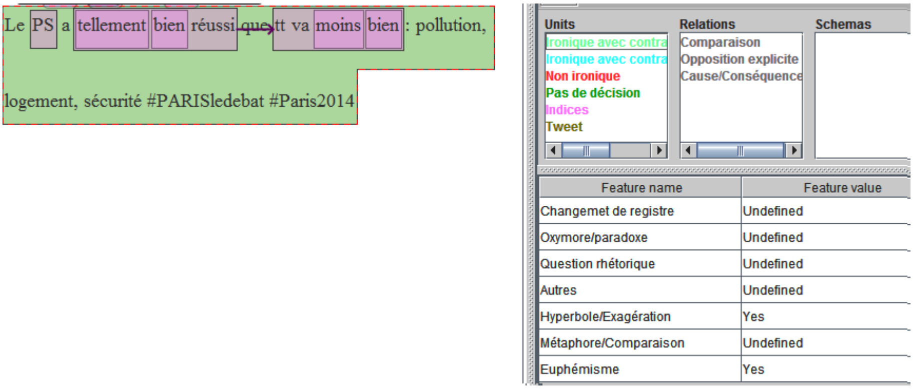
3) for tweets containing negation, only one negation word needs to be labeled;

4) for ironic tweets containing an explicit contradiction, the text segments in question should be linked by an "explicit contradiction" relation;

5) for ironic tweets containing an implicit contradiction, no connection can be made using the defined relation;

6) furthermore, if a tweet contains a comparison marker, the two concepts or text segments being compared must be linked by a "comparison" relation.

Additionally, automatic detection was used to ensure that annotations conformed to the constraints listed above, detecting the most common error types: ironic tweets with no trigger type or irony category, missing markers, etc. In cases where errors were detected, the annotator was asked to correct their mistake before moving on to other tweets.



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

Feature name	Feature value
Changemet de registre	Undefined
Oxymore/paradoxe	Undefined
Question rhétorique	Undefined
Autres	Undefined
Hyperbole/Exagération	Yes
Métaphore/Comparaison	Undefined
Euphémisme	Yes

Figure 3.2. *Example of a tweet annotated using Glozz*

Figure 3.2 is an illustration of the annotation procedure for an ironic tweet with explicit opposition between two segments: (1) "tellement bien réussi" (done such a good job) and (2) "tout va moins bien" (everything's going downhill). The first segment expresses hyperbole, whereas the second contains euphemism. There are thus two categories of irony in this tweet: hyperbole and euphemism. The terms "PS" (socialist party: named entity), "tellement" (such: intensifier), "bien" (well/good: positive opinion word) and "moins" (less: intensifier) were annotated as cues.

3.5. Results of the annotation campaign

In this section, we shall present the different statistics obtained from the annotated corpus. This work was carried out in collaboration with Farah Benamara and Véronique Moriceau.

The annotated corpus was made up of 2,000 tweets, of which 80% were ironic and 20% were non-ironic, based on the presence or absence of the *#ironie* and *#sarcasme* hashtags. The annotation campaign was carried out by three French-speaking human annotators.

The corpus was divided into three parts. The first part comprised 100 tweets (50 ironic and 50 non-ironic) and was used for annotator training. The second part of the corpus was made up of 300 tweets (250 ironic, 50 non-ironic) and was annotated by two annotators in order to calculate interannotator agreement. The third and final part of the corpus comprised 1,700 tweets, of which 80% were ironic, and was used in the final annotation campaign.

In what follows, we shall present the interannotator agreement results for the 300 tweets processed by two annotators, along with statistical results from the annotation campaign for the full 2,000 tweet corpus.

3.5.1. *Qualitative results*

Working on a total of 300 tweets, the human annotators were in agreement in 255 cases (174 ironic an d 63 non-ironic tweets), of which 18 were classed as No decision. We obtained a Cohen's kappa of 0.69 for ironic/non-ironic classification: this is a very good score. The kappa obtained in comparison with the reference tags (*#ironie* and *#sarcasme*) was also good (0.63), showing these hashtags to be sufficiently reliable. Furthermore, we noted that over 90% of the tweets labeled No decision due to a lack of external context were, in fact, ironic, a fact revealed by the reference hashtags. However, we decided to keep them in the no decision group for the purposes of our experiment.

For explicit versus implicit irony classification, we obtained a kappa of 0.65. It is interesting to note that the majority of the tweets in question contained implicit triggers (76.42%). This result is important, demonstrating the annotators' capacity to identify text fields which trigger incongruity in ironic tweets, whether explicit or implicit. Our hope was that the automatic system would perform as well as the human annotators (see Chapter 4 for the results obtained using our computational models).

In the case of category identification, the calculation was more complex as a single ironic tweet may belong to several different categories. We calculated agreement by counting the number of categories identified by both annotators, then dividing this score by the total number of annotated categories. We obtained a score of 0.56, which is moderately acceptable. This score reflects the complexity of the identification process for pragmatic devices. By grouping similar devices together (notably combining hyperbole/exaggeration and euphemism, all used to increase or weaken the strength of meaning), we were able to obtain a higher score of 0.60.

3.5.2. *Quantitative results*

The main objective of our corpus study was to verify whether different linguistic theories and definitions of irony may be applied to social media content, notably tweets. In addition to standard frequencies, we calculated correlations between types of irony triggers and markers, and between categories and markers, in order to highlight those characteristics that would be most useful for the purposes of automatic irony detection. Note that in all of these studies, the obtained frequencies are statistically significant based on the χ^2 test ($P < 0.05$).

3.5.2.1. *Tweet frequency by class*

Figure 3.3 shows the frequencies of annotated tweets for our three classes: ironic, non-ironic and no decision. This first step corresponds to level 1 of our annotation scheme (see section 3.3.2).

Based on the reference hashtags *#ironie* and *#sarcasme*, our corpus contained 1,600 ironic tweets and 400 non-ironic tweets. The human annotators considered 1,460 tweets (73%) to be ironic and 380 tweets (19%) to be non-ironic. The remaining 160 tweets (8%) were sorted into the no decision category.

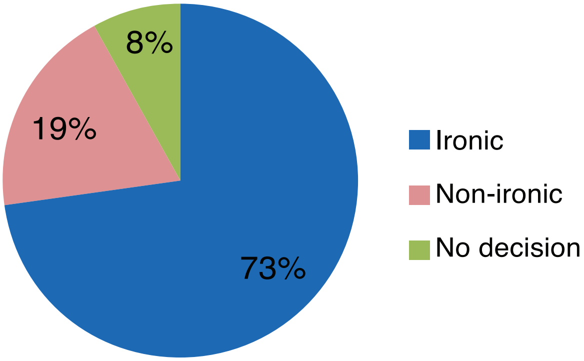


Figure 3.3. *Class distribution of annotated tweets (level 1)*

These results indicate that tweets containing the *#ironie* or *#sarcasme* hashtag are not necessarily ironic, and that irony may occur without these hashtags (for example: *@MelvinLeroux Quels autres problèmes? La France va bien, le taux de chômage n'est pas du tout élevé, tout le monde est heureux...*) (*@MelvinLeroux What other problems? France is doing great, unemployment isn't high at all, everybody's happy)...*

3.5.2.2. *Tweet frequency by irony type*

Figure 3.4 shows our results for the second level of the annotation scheme, concerning the type of irony trigger (explicit or implicit).

From a total of 1,460 tweets labeled as ironic, 1,066 tweets (73%) were judged to be ironic with implicit contradiction; only 394 tweets (27%) featured an explicit contradiction. This indicates that irony is a phenomenon generally expressed in an implicit manner.

3.5.2.3. *Tweet frequency by irony category*

Table 3.5 shows quantitative results for the third level of the annotation scheme: irony categories.

These results show significant differences in terms of the presence of different categories for each type of irony. The euphemism and hyperbole/exaggeration categories, for example, appeared with similar frequencies in both irony types (explicit and implicit contradiction). The oxymoron/paradox category, on the other hand, only occurred in cases of explicit contradiction. Similarly, false affirmations only appeared in ironic tweets with implicit opposition. This is explained by the definitions of the categories oxymoron/paradox and false affirmation themselves (section 3.3.2.3). We also note that most of the tweets in the *other* category are ironic with implicit contradiction. This indicates that the decision task is harder in cases where irony is expressed through implicit contradiction.

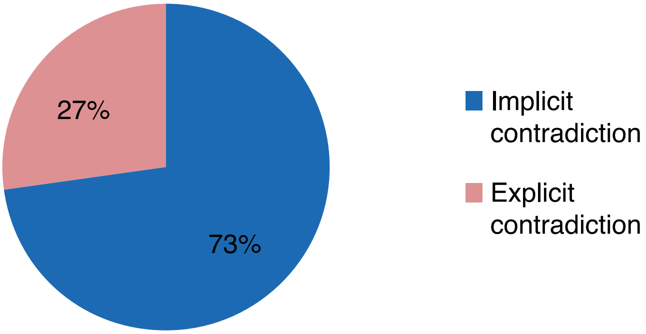


Figure 3.4. *Distribution of annotated tweets by irony type (level 2)*

	Ironic with explicit contradiction	**Ironic with implicit contradiction**
Analogy	12	2
Register shift	1	–
Euphemism	1	1
Rhetorical question	10	14
Oxymoron/paradox	**66**	–
Hyperbole/exaggeration	8	10
False affirmation	–	**56**
Other	21	**32**

Table 3.5. *Percentage of tweets in each category for each type of irony (level 3)*[7]

[7] The bold values represent the best results or notable values.

As classes are not mutually exclusive, 64 tweets of the explicit contradiction type belong to more than one category, and 134 of the implicit contradiction type belong to more than one category. The most common combination for explicit contradictions is oxymoron/paradox + hyperbole/exaggeration, while for implicit contradiction cases, the most frequent combination is false affirmation + hyperbole/exaggeration. The tweet in phrase (3.24) is annotated for two categories of irony: the first is hyperbole, expressed by *tellement bien (so well)*, whereas the second is euphemism, in the form of *moins bien (less well)*.

(3.24) Le PS a **tellement bien** réussi que tt va **moins bien** : pollution, logement, sécurité #PARISledebat #Paris2014
*(The socialist party have done **such a great** job that everything's **gone downhill**: pollution, accommodation, security #PARISledebat #Paris2014)*

3.5.2.4. *Tweet frequency by linguistic cue*

Three statistical studies were carried out at this level. The first is a quantitative study concerning the first and fourth levels of the annotation scheme, considering the presence of different cues in ironic and non-ironic tweets (see Table 3.6). The second study addresses the second and fourth levels of the annotation scheme, studying the presence of different cues in ironic tweets with explicit or implicit contradictions (Table 3.6). Our third and final study relates to the third and fourth levels of the annotation scheme, covering the presence of cues in each irony category (Table 3.7).

Table 3.6 indicates that most of our cues occur more frequently in ironic tweets than in non-ironic tweets. Additionally, negation words, intensifiers, oppositions, interjections, comparisons and opinion words occur most frequently in tweets containing an explicit contradiction.

URLs and false propositions, on the other hand, occur most in tweets with an implicit contradiction. The strong presence of URLs in ironic tweets with implicit opposition reflects the fact that readers take account of the external context provided by the linked site in order to understand the irony of the tweet (see Figures 3.5 and 3.6).

	Ironic with explicit contradiction	Ironic with implicit contradiction	Non-ironic
Emoticon	7	6	5
Negation	**37**	**34**	**58**
Discursive connectors	6	4	4
#tag humorous	2	4	0
Intensifier	22	19	11
Punctuation	**51**	**51**	28
False proposition	8	**54**	0
Surprise	3	3	2
Modality	0	0	1
Quotation	6	6	1
Opposition word	9	3	4
Capitalized word	3	2	3
Personal pronoun	**31**	**31**	**30**
Interjection	14	12	2
Comparison	8	2	4
Named entity	**97**	**91**	**82**
Reporting	1	1	3
Opinion	**48**	**41**	**35**
URL	21	26	**36**

Table 3.6. *Marker distribution across ironic (explicit/implicit) and non-ironic tweets as a percentage of total tweets. The most frequent cues are shown in bold*

Table 3.7 illustrates the percentage of tweets in each category with markers. We see that, for each category, there are at least two marker types that occur frequently. For example, negations are most frequent in the analogy, register shift, euphemism, rhetorical question and oxymoron/ paradox categories, whereas false propositions are common in the euphemism, hyperbole/exaggeration and false affirmation categories.

3.5.2.5. *Relation frequency in ironic tweets with explicit contradictions*

Figure 3.7 shows the high frequency of opposition relations in ironic tweets with explicit contradiction. This is explained by the significant number of opposing text segments in individual tweets. The remainder of relations were almost equally split between comparison relations and cause/ consequence relations.

	Analogy	**Register shift**	**Euphemism**	**Hyperbole/ exaggeration**	**Rhetorical questions**	**Oxymoron/ paradox**	**False affirmation**	**Other**
Emoticon	6	0	0	5	6	6	5	8
Negation	**46**	**40**	**50**	25	**43**	**35**	18	26
Discursive connector	6	0	6	5	2	4	4	5
#tag humorous	6	0	0	3	2	0	3	5
Intensifier	21	0	**50**	**57**	17	21	10	15
Punctuation	**49**	**60**	**72**	**56**	**93**	**49**	29	**45**
False proposition	13	0	**44**	**53**	9	11	**95**	11
Surprise	0	0	0	3	4	4	3	1
Modality	0	0	0	0	0	0	0	0
Quotation	0	0	0	8	7	5	4	8
Opposition word	6	0	0	2	3	12	3	2
Capitalization	5	0	0	2	5	4	2	3
Personal pronoun	**38**	**40**	22	29	31	32	31	29
Interjection	6	20	6	18	13	15	13	10
Comparison	**43**	20	0	0	2	2	2	1
Named entity	**100**	**80**	**94**	**88**	**90**	**99**	**90**	**91**
Reporting	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. *Marker distribution across different categories as a percentage of total tweets. The most frequent cues are shown in bold*

3.5.3. *Correlation between different levels of the annotation scheme*

We also carried out a statistical study on the strength of links between different levels in our proposed annotation scheme. We began by considering the link between the first level (ironic/non-ironic) and the fourth level (cues), before looking at the relationship between levels 2 (explicit/implicit contradictions) and 4. Finally, we examined the relationship between levels 3 (irony categories) and 4.

The conventional means of verifying whether a link exists between two crossed qualitative variables in a contingency table is to use the χ^2 test. The strength of the link between two variables may be measured using the Phi (ϕ) test or Cramer's V (Cohen 1988). The Phi (ϕ) test can only be used for 22, tables, whereas Cramer's V is suitable for use with larger tables. Given the size of the tables in our study, we used the latter test to verify the strength of the link between elements at different levels of our annotation scheme.

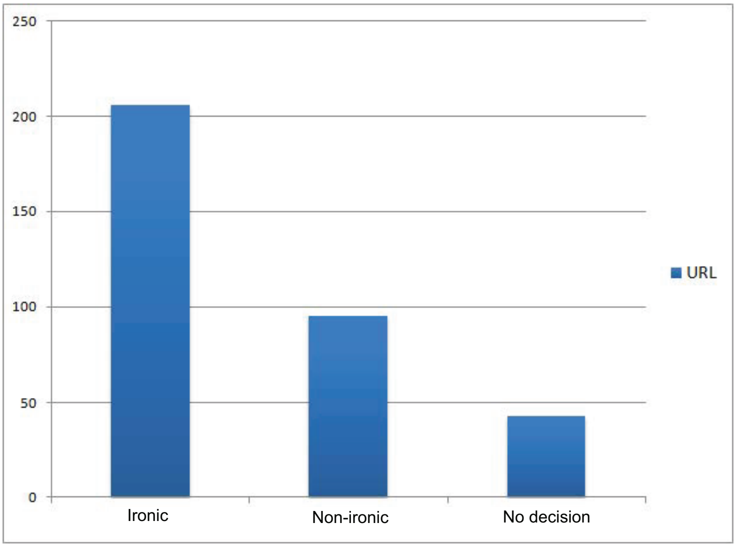


Figure 3.5. *Presence of URLs in tweets*

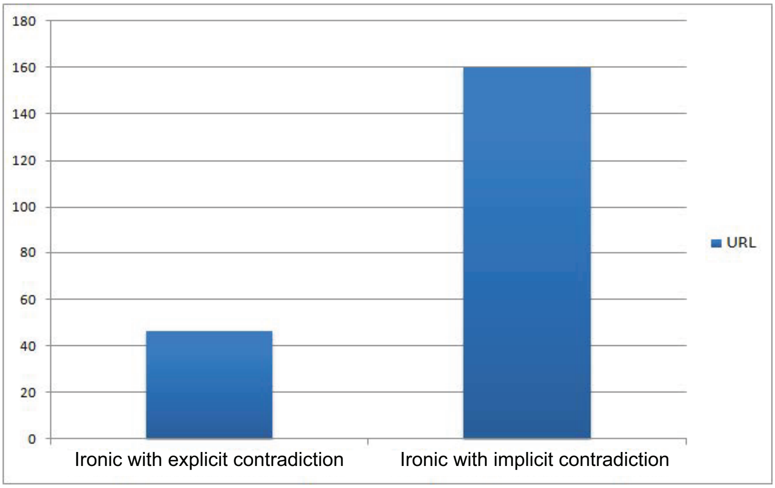


Figure 3.6. *Presence of URLs in ironic tweets*

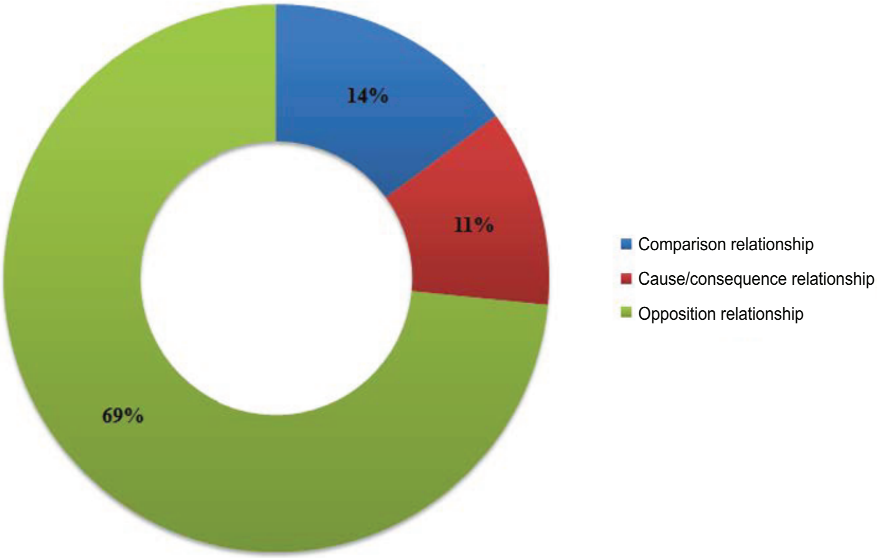


Figure 3.7. *Distribution of relations in ironic tweets with explicit contradictions, expressed as a percentage*

Cramer’s V is given by the following formula:

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

Cramer’s V is the square root of χ^2 divided by $n(k-1)$, where n represents the sample size and k the shortest side of the table (number of rows or columns). The value of V is located in the interval from 0 to 1 inclusive, where 0 implies no association between variables and 1 represents complete dependency, as χ^2 will be equal to $n(k-1)$ in this case (in a 22 table, the value can also be negative, within an interval from -1 and 1). The closer the V value is to 1, the stronger the link is between the two variables in question.

3.5.3.1. *Correlation between irony type and cues*

We began by verifying the strength of the association between level 1 of our annotation scheme (irony class: ironic/non-ironic) and level 4 (irony cues, see Figure 3.1). We obtained a value of 0.156 for Cramer's V, with $df = 14$[8]. According to the table of V values defined by Cohen (1988), this value is statistically significant ($P < 0.05$), indicating a strong correlation between ironic/non-ironic class and the set of annotated cues.

3.5.3.2. *Correlation between irony trigger and cues*

Our second study concerned the strength of the association between level 2 of the annotation scheme (irony trigger: implicit or explicit contradiction) and level 4 (irony cues).

This study showed a strong correlation between irony types (explicit or implicit triggers) and different annotation cues, with a Cramer's V of 0.196 and $df = 16$. A one-by-one study of the strength of the association between irony types and cues ($df = 1$) indicated a medium to strong association between irony type and particular cues, with negations, interjections, named entities and URLs falling into the interval $0.140 < V < 0.410$.

3.5.3.3. *Correlation between different irony cues*

The third step was to check for correlation between different irony cues. Our study showed that the cues which correlate most closely to irony classes (ironic/non-ironic) include negations, interjections, named entities and URLs ($0.140 < V < 0.410$, $df = 1$), while opposition markers, comparison words and false propositions are must closely correlated with explicit/implicit contradictions ($0.140 < V < 0.190$, $df = 1$).

3.5.3.4. *Correlation between irony categories and cues*

Our fourth study verified the correlation between level 3 of the annotation scheme (irony category) and level 4 (cues). The results indicated that the strongest correlations between categories and cues occurred in the case of intensifiers, punctuation, false propositions and opinion words ($0.267 < V < 0.565$, $df = 4$). Note that although opinion words occur frequently in ironic tweets, our results indicate that these words are not helpful in distinguishing between ironic/non-ironic tweets, or between explicit/implicit contradictions.

[8] df = min (r - 1, c - 1), where r = number of rows and c = number of columns in the contingency table.

These results show that both cues and categories are helpful in classifying tweets as ironic/non-ironic and explicit/implicit ironic, and even in distinguishing between different categories of irony.

3.6. Conclusion

In this chapter, we presented a multilevel annotation scheme for irony, focusing on categories of irony proposed by linguists and presenting those selected for use in our approach. The annotation scheme was validated by means of an annotation campaign carried out by three French speakers. A subset of the FrIC was labeled by two different annotators in order to calculate interannotator agreement for all levels of the scheme; the results of this test were satisfactory. Finally, we processed the corpus using the Glozz annotation tool and carried out statistical tests for each level of our scheme.

The results of this work, particularly in terms of the correlation between irony categories and types, are crucial to the development of an effective automatic detection system for irony. A subset of the FrIC (Figure 3.8) was used for this purpose, and our approach is presented in the following chapter. Our scheme is also relatively portable for culturally similar languages, as we shall see in Chapter 5, section 5.2.2), where we present our results for tweet corpora in English and Italian.

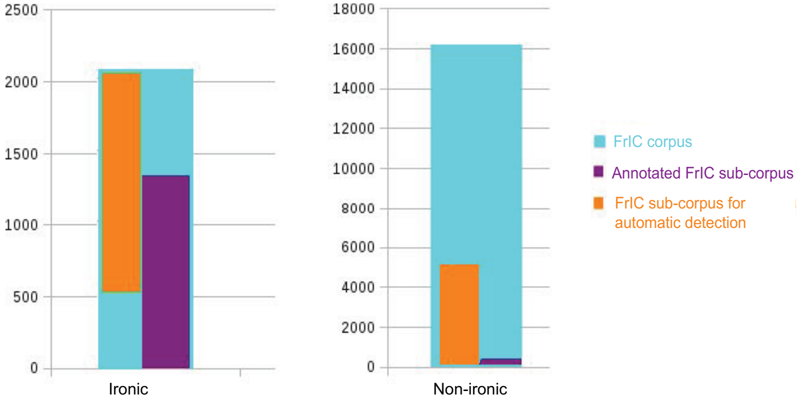


Figure 3.8. *Tweet distribution in the FrIC and subcorpora used for manual annotation and automatic detection experiments*

4

Three Models for Automatic Irony Detection

4.1. Introduction

In the previous chapter, we analyzed a 2,000 tweet subset of the FrIC, making the following observations:

– Irony in a tweet is indicated by the presence of two propositions P_1 and P_2 that contradict one another. We identified two types of contradiction: explicit, in cases where both P_1 and P_2 are present in lexical form in the tweet, and implicit, in cases where P_1 is present but P_2 must be inferred from external context. We showed that in 76.42% of cases in our corpus, irony took the form of an implicit contradiction, with only 23.58% of cases containing explicit contradictions.

– The most common irony categories were oxymoron/paradox in the case of explicit contradictions, and false assertion in the case of implicit contradictions, with frequencies of 66% and 56%, respectively.

– The most common linguistic cues found in tweets in our corpus included named entities, punctuation, opinion expressions, negation markers, personal pronouns and URLs. Furthermore, negation markers (no, never, won't, etc.) are one of the most common cues found in both ironic and non-ironic tweets, with respective frequencies of 35% and 58%.

Based on these observations, we decided to develop a model for automatic detection of textual irony in tweets in cases of both explicit and implicit contradictions. Our model notably has the capacity to detect irony expressed

Automatic Detection of Irony: Opinion Mining in Microblogs and Social Media,
First Edition. Jihen Karoui; Farah Benamara and Véronique Moriceau.

through false assertions. Given the high numbers of negations in our corpus, we wish to test the following hypotheses:

– hypothesis (H1): the presence of negations as an internal property of an utterance may help to detect disparity between the literal and intended meanings of the utterance;

– hypothesis (H2): a tweet containing an affirmed fact of the form $Not(P)$ is ironic if, and only if, P can be confirmed based on general knowledge, external to the utterance, shared by both the author and the reader.

These hypotheses will be tested using a three-step procedure using three new models:

1) a supervised machine learning method, used to detect whether a tweet is ironic on the basis of internal content alone. This method involves two models:

- the first, named **SurfSystem**, is based on surface features alone;

- the second, **PragSystem**, is based on pragmatic features extracted from the linguistic content of the tweet.

Both models use features described in the state of the art that have been proved empirically to be effective, alongside new groups of features;

2) a third model, known as **QuerySystem**, used to validate the internal context of an utterance in relation to "external" context. We propose an algorithm to treat classifier output constructed using the **PragSystem** model, and to correct wrongly classified ironic utterances of the form $Not(P)$ by searching for P in external, reliable online information sources, such as Wikipedia and online newspaper sites. We carried out two experiments, the first using reference hashtags (*#ironie* and *#sarcasme*) and the second using the classes predicted by the classifier. If P is found in a reliable information source, the tweet in question may be expressing a non-literal, i.e. ironic, meaning.

This three-step approach is new. To our knowledge, this is the first study to use both the presence of negations and external knowledge for the automatic detection of irony manifest through implicit contradiction.

Before going into detail concerning our three models (SurfSystem, PragSystem and QuerySystem), we shall present the data used to train and test these models.

4.2. The FrIC^{Auto} corpus

The FrIC^{Auto} corpus, used for automatic detection experiments, is a subset of the FrIC (see Chapter 3, section 3.2). It is made up of 1,545 ironic tweets and 5,197 non-ironic tweets on subjects discussed in the media (see Table 4.1). Note that only 50% of the tweets in FrIC^{Auto} were annotated manually following the multilevel scheme presented in the previous chapter.

Categories	**Mots clés**
Politics	Ayrault, Fillon, Hollande, Le Pen, FN, DSK, UMP, etc.
Health	cancer, grippe, sida, dépression, angoisse, psychologie, etc.
Social networks	Skype, Facebook, MSN, WhatsApp, etc.
Sport	Zlatan, PSG, football, Ribéry, Zidane, équipe de France, ligue des champions, jeux olympiques, etc.
Arab Spring	Marzouki, Ben Ali, Bachar, Moubarak, Al-Assad, Morsi, Kadhafi, etc.
Country/City	Algérie, Égypte, Syrie, Tunisie, Iran, Washington, Mali, etc.
Artists	Rihanna, Beyoncé, Carla Bruni, Madonna, Nabilla, Justin Bieber, Adèle, etc.
Television	Fast and Furious, Xfactor, The Voice, etc.

Table 4.1. *Categories used to collect the corpus with corresponding keywords (in French)*

As in the case of manual annotation, the hashtags *#ironie* and *#sarcasme* were removed from the tweets for the purposes of our experiments.

The detailed study of irony in French tweets conducted during the annotation campaign, described in Chapter 3, highlighted the strong presence of negation words such as *ne, n', pas, non, ni, sans, plus, jamais, rien, aucun(e) and personne (translation: no, never, nothing, nobody)*. Around 62.75% of all collected tweets contained negation words. For this reason, we feel that negation may be an important cue in ironic utterances, notably in the context of false assertions.

To measure the effect of negation on irony detection, we created three corpora: tweets with negation (NegOnly), tweets without negation (NoNeg) and a corpus containing all of the tweets (All). The distribution of tweets in each corpus is shown in Table 4.2.

Corpus	Ironic	Non-ironic	Total
NegOnly	470	3,761	**4,231**
NoNeg	1,075	1,436	**2,511**
All	1,545	5,197	**6,742**

Table 4.2. *Distribution of tweets in the corpus*

Negations were identified automatically using two syntax analyzers: XIP[1] and MElt[2]. Manual analysis of the analyzer output showed that MElt was more effective than XIP in detecting genuine negations, although a number of errors remained. For this reason, we chose to use MElt for the purposes of our study. Issues remain for three common French negation words, which also have other meanings: *personne* (no-one/a person), *plus* (no more, more, plus), and *pas* (not/step).

MElt systematically considers these words as negations. However, personne should only be considered to be a negation when used as an indefinite pronoun (i.e. where *personne* represents a name: phrase (4.1)). Similarly, *plus* is only a negation if it is not being used as a comparative or superlative (phrase (4.2)). Finally, the word *pas* is a negation if it is not a noun (phrase (4.3)). We developed a postprocessing correction approach to correctly identify negations based on these grammar rules using a Java script, applied to output from the MElt syntax analyzer.

(4.1) @EloiseLEspagnol va se coucher. J'suis seule avec **1 personne** dans ma TL. #GENIAL
*(@EloiseLEsoagnol is going to bed. I'm on my own with **1 person** in my TL. #GREAT)*

(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!
*(I don't know who's madd**est** at seeing @RoyalSegolene [Hollande's ex-wife] next to Flanby [French president François*

[1] https://open.xerox.com/Services/XIPParser.

[2] http://alpage.inria.fr/statgram/frdep/fr_stat_dep_malt.html.

Hollande], #trierweiler [Hollande's former partner] or #Gayet his new partner])

(4.3) @OMissud @RoyalSegolene je leur conseille le tango, **un pas** en avant, **un pas** en arrière !!!
(@OMissud @RoyalSegolene I suggest the tango, ***one step*** *forward,* ***one step*** *back !!!*

For the purposes of our study, we assumed that the hashtags #ironie and #sarcasme were reliable. However, the absence of these hashtags in a tweet in no way indicates that the tweet is not ironic. To check the reliability of the hashtags, two human annotators manually labeled three subsets containing 50 ironic and 50 non-ironic tweets for the All, NoNeg and NegOnly corpora. The interannotator agreement score (Cohen's kappa) for the reference hashtags was $\kappa = 0.78$ for the All corpus, $\kappa = 0.73$ for the NoNeg corpus and $\kappa = 0.43$ for the NegOnly corpus. These scores indicate that the #ironie and #sarcasme are relatively reliable, but also that the presence of a negation term can create ambiguity and make it harder for humans to identify irony.

4.3. The SurfSystem model: irony detection based on surface features

In this section, we shall present the first model we developed for irony detection based on surface features alone. We shall describe the selected features, our experiments and the results we obtained.

4.3.1. *Selected features*

For this first experiment, we reused the set of surface features identified in the state of the art in order to classify French tweets into ironic/non ironic group, adding a number of new features:

– tweet length in words (Tsur *et al*. (2010));

– presence of punctuation (Kreuz and Caucci (2007), Gonzalez-Ibanez *et al*. (2011)), see phrases (4.4) and (4.5):

(4.4) **Ah** oui exact'**!** #SuisJeBête **Mais** il y a rien de fait pour le PSG en championnat hein **:)** #ironie
__Oh__ yeah, you're right__!__ #SillyMe __But__ PSG [soccer team] hasn't done much in the championships, huh __:)__ #irony)

(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*
(How to get your #LT to hate you: Algeria won't be going to the World Cup, I'm telling you =) __!!!!!!!!!!! I REPEAT, I'M JOKING !!__ #irony)

– presence of words in all capitals (Tsur *et al*. (2010), Reyes *et al*. (2013)), see phrase (4.5);

– presence of interjections (Gonzalez-Ibanez *et al*. (2011), Buschmeier *et al*. (2014)), see phrase (4.4);

– presence of emoticons (Gonzalez-Ibanez *et al*. (2011), Buschmeier *et al*. (2014)), see phrase (4.4);

– presence of quotations (Tsur *et al*. (2010), Reyes *et al*. (2013)), see example (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
(__"1 million people unemployed means 1 million too many immigrants"__...hadn't heard that before...nice slogan, not at all oversimplified #fn #lepen #irony)

– use of slang (Burfoot and Baldwin (2009)), see phrase (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...
(They told us #Hollande was going to __get trampled__ on the international stage. #Obama didn't look too convinced by him this morning...)

– presence of opposition words (Utsumi (2004)), see phrase (4.4);

– series of multiple exclamation or question marks (Carvalho *et al.* (2009)), see phrase (4.5);

– presence of a combination of exclamation and question marks (Buschmeier *et al.* (2014)), see phrase (4.8):

(4.8) Depuis quelques jours certains murmurent que le racisme se serait ... MAL **?!??!!!!!!** Hummmmmmmm #MoueDubitative #ironie
(In recent days, there have been rumors that racism might be... BAD ***?!??!!!!!!*** *Hmmmmmmmm #SkepticalFace #irony)*

– presence of discursive connectors not associated with opposition. This is an entirely new feature that we have chosen to introduce here (see phrase (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
(Please thank Jerome from the NC service in Lyon ***along with*** *your technical staff. We won't hesitate to recommend you #irony)*

These features were implemented using four lexicons:

– a lexicon of 328 discursive connectors, defined by Roze *et al.* (2012);

– a lexicon of 389 slang words, constructed manually from a variety of sources found online[3];

– the CASOAR lexicon (Benamara *et al.* (2014)), made up of 236 interjections;

– a lexicon of 595 emoticons, collected manually from Twitter.

4.3.2. *Experiments and results*

Several classifiers were tested for training purposes using the Weka platform with the default parameters. These included sequential minimal optimization (SMO), decision tree (DT) and naive Bayes (NB). The best results were obtained using SMO; as such, we shall only present the results obtained using SMO for all of our experiments.

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

As three corpora were used (NegOnly, NoNeg and All), we trained three classifiers, one for each corpus, denoted as$C_{NegOnly}$, C_{NoNeg} and C_{All}. Given that the number of ironic instances (IR) in NegOnly was relatively small (470 tweets), the $C_{NegOnly}$ classifier was trained using a balanced subset of 940 tweets with cross-validation for 10 samples. In the case of C_{NoNeg} and C_{All}, 80% of the corpora was used for training and 20% for testing, with an equal distribution of IR and non-ironic instances (NIR)[4].

Table 4.3 shows the results of three classifiers using all surface features in terms of precision (P), recall (R) and f-measure (F). The $C_{NegOnly}$ classifier produced the best results in terms of accuracy (72.23%). Once again, this result highlights the importance of negation in detecting ironic tweets, in spite of the fact that this feature proved problematic for human understanding in the context of manual annotation.

	Ironic (IR)			**Non-ironic (NIR)**			
	P	**R**	**F**	**P**	**R**	**F**	**Accuracy**
$C_{NegOnly}$	**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. *Results of the SurfSystem model. The best results are shown in bold*

To assess the contribution of each surface feature to the training process, we added features one by one in order to observe their influence on the accuracy of the classifier results. The results obtained for the All, NegOnly and NoNeg corpora are shown in Table 4.4.

For the All corpus, an accuracy value of 55.50% was obtained using all features. However, using the presence of punctuation and capitalization features alone produced a more accurate result, with a value of 56.31%.

Similarly, for the NegOnly corpus, the use of the presence of punctuation and capitalization alone resulted in a better accuracy score of 73.08%. For the NoNeg corpus, on the other hand, the best accuracy value (63.25%) was obtained using all surface features together.

[4] For C_{NoNeg} and C_{All}, we tested cross-validation over 10 samples with equal distribution of ironic and non-ironic instances, but the results were considerably poorer.

Corpus	Features	No. of features	Ironic (IR)			Non-ironic (NIR)			Accuracy (%)
–	–	–	P	R	F	P	R	F	
C_{All}	1. Punctuation	1	0.517	0.848	0.642	0.577	0.207	0.305	52.75 %
	2. Capitalization	1+2	0.556	0.631	0.591	0.573	0.495	0.531	**56.31 %**
	3. Interjection	1 – 3	0.556	0.631	0.591	0.573	0.495	0.531	56.31 %
	4. Emoticon	1 – 4	0.556	0.631	0.591	0.573	0.495	0.531	56.31 %
	5. Quotation	1 – 5	0.556	0.631	0.591	0.573	0.495	0.531	56.31 %
	6. Disc. connect.-opposition	1 – 6	0.556	0.631	0.591	0.573	0.495	0.531	56.31 %
	7. Slang	1 – 7	0.556	0.631	0.591	0.573	0.495	0.531	56.31 %
	8. Opposition	1 – 8	0.531	0.955	0.682	0.774	0.155	0.259	55.50 %
	9. Exclamation	1 – 9	0.531	0.955	0.682	0.774	0.155	0.259	55.50 %
	10. Quest. mark	1 – 10	0.531	0.955	0.682	0.774	0.155	0.259	55.50 %
	11. Exclam. and quest. marks	1 – 11	0.531	0.955	0.682	0.774	0.155	0.259	55.50 %
	12. Number of words	1 – 12	0.531	0.955	0.682	0.774	0.155	0.259	55.50 %
$C_{NegOnly}$	1. Punctuation	1	0.492	0.279	0.356	0.497	0.713	0.586	49.57 %
	2. Capitalization	1+2	0.883	0.532	0.664	0.665	0.93	0.776	**73.08 %**
	3. Interjection	1 – 3	0.883	0.532	0.664	0.665	0.93	0.776	73.08 %
	4. Emoticon	1 – 4	0.883	0.532	0.664	0.665	0.93	0.776	73.08 %
	5. Quotation	1 – 5	0.883	0.532	0.664	0.665	0.93	0.776	73,08 %
	6. Disc. connect.-opposition	1 – 6	0.883	0.532	0.664	0.665	0.93	0.776	73.08 %
	7. Slang	1 – 7	0.883	0.532	0.664	0.665	0.93	0.776	73.08 %
	8. Opposition	1 – 8	0.883	0.532	0.664	0.665	0.93	0.776	73.08 %
	9. Exclamation	1 – 9	0.861	0.538	0.662	0.664	0.913	0.769	7.,55 %
	10. Quest. mark	1 – 10	0.861	0.538	0.662	0.664	0.913	0.769	72.55 %
	11. Exclam. and quest. mark	1 – 11	0.861	0.538	0.662	0.664	0.913	0.769	72.55 %
	12. Number of words	1 – 12	0.847	0.543	0.661	0.664	0.902	0.765	72.23 %
C_{NoNeg}	1. Punctuation	1	0.581	0.833	0.685	0.705	0.4	0.51	61.62 %
	2. Capitalization	1+2	0.581	0.833	0.685	0.705	0.4	0.51	61.62 %
	3. Interjection	1 – 3	0.593	0.298	0.396	0.531	0.795	0.637	54.65 %
	4. Emoticon	1 – 4	0.593	0.298	0.396	0.531	0.795	0.637	54.65 %
	5. Quotation	1 – 5	0.593	0.298	0.396	0.531	0.795	0.637	54.65 %
	6. Disc. connect.-opposition	1 – 6	0.575	0.642	0.607	0.595	0.526	0.558	58.37 %
	7. Slang	1 – 7	0.584	0.66	0.62	0.61	0.53	0.567	59.53 %
	8. Opposition	1 – 8	0.591	0.651	0.619	0.611	0.549	0.578	60.00 %
	9. Exclamation	1 – 9	0.591	0.651	0.619	0.611	0.549	0.578	60.00 %
	10. Quest. mark	1 – 10	0.591	0.651	0.619	0.611	0.549	0.578	60.00 %
	11. Exclam. and quest. mark	1 – 11	0.591	0.651	0.619	0.611	0.549	0.578	60.00 %
	12. Number of words	1 – 12	0.635	0.623	0.629	0.63	0.642	0.636	**63.25 %**

Table 4.4. *Feature-by-feature training results obtained using the SurfSystem model for the All, NegOnly and NoNeg corpora. The best results are shown in bold*

Comparing our results with existing literature, we note that the same trends have been observed in other languages and other corpus types.

For example, (Burfoot and Baldwin 2009) obtained an f-measure of 79.5% for a corpus of press articles in English, (Carvalho *et al*. 2009) obtained an accuracy value of 85.4% for a corpus of press articles in Portuguese and Tsur *et al*. (2010) noted a precision score of 50% for a corpus of Amazon product reviews in English.

4.4. The PragSystem model: irony detection based on internal contextual features

This section is devoted to our second experiment, in which the surface features used in the previous model were combined with pragmatic features found in the internal context of tweets.

Each tweet is represented by a vector made up of six groups of features. Some of these features have already been shown to be effective in detecting irony in other languages (references are provided below); others are entirely new. In this section, we shall present all of the features used in our study, along with the results obtained for each feature group and each subcorpus (All, NegOnly, and NoNeg).

4.4.1. *Selected features*

4.4.1.1. *Surface features*

The surface features are the same as those used in our previous experiment (see section 4.3).

4.4.1.2. *Sentiment features*

These features indicate the presence of words or expressions conveying a positive or negative opinion Reyes and Rosso (2011, 2012) and the number of these elements (Barbieri and Saggion 2014b). We added three new features:

– the presence of words or expressions conveying **surprise** or **eshock** – see phrase (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
*(**What a surprise**, Naouelle won!!! Good job there were no leaks to spoil the suspense!! #irony #topchef)*

– the presence and number of **neutral opinions**, i.e. opinions that are both positive and negative, or implied – see phrase (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
*(It's true that the PSG has history, you can **tell that** the guys are really committed and that they aren't just doing it for the money #irony)*

We used two lexicons to identify these features:

1) **CASOAR**[5] (Benamara *et al*. (2014)), a French lexicon of 2,830 words or opinion expression grouped into four semantic categories (*reporting, judgment, sentiment-assessment* and *advice*), comprising 1,142 adjectives, 605 adverbs, 415 nouns, 308 verbs, 292 expressions, 62 interjections and six conjunctions/prepositions/pronouns. The lexicon draws a clear distinction between purely subjective entries and intensifier/negation entries, which affect expressions of opinion at phrase level by inverting, intensifying or minimizing the polarity and/or strength. Most subjective entries are adverbs, nouns, verbs, manner adverbs, interjections and emoticons. They can be split into three groups:

- polarity: positive, negative or neutral;

- strength: assessed on a scale from 1 to 3, from weak to strong;

- semantic category: there are four of these categories, including reporting (e.g. know, see and announce), judgment (e.g. hope, clear and pathetic), sentiment-assessment (e.g. like, shame and worry) and advice (e.g. suggest, advise and recommend);

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

2) **EMOTAIX**[6], an emotional and affective lexicon containing 4,921 entries spread across nine categories: ill-will, unhappiness, anxiety, goodwill, well-being, sang-froid, surprise, imperturbability and non-specific emotion. It contains 1,308 positive entries, 3,078 negative entries and 535 neutral entries. Each category is made up of subcategories that are themselves connected to base categories. For example, the category "ill will" contains the subcategories "hate" and "aggression". The "hate" subcategory is connected to four base categories: resentment, disgust, contempt and irritation (see Figure 4.1).

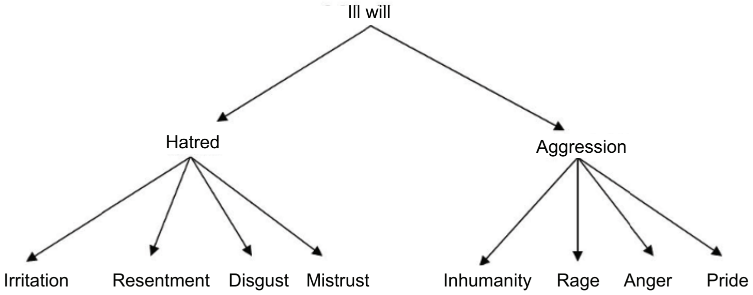


Figure 4.1. *Example of categories and subcategories in the EMOTAIX lexicon. For a color version of the figures in this chapter see, www.iste.co.uk/karoui/irony.zip*

4.4.1.3. *Modifier features*

These binary features indicate whether a tweet contains an **intensifier** (very, quite, a lot, etc.; (Liebrecht *et al*. 2013, Barbieri and Saggion 2014b)) (see phrase (4.12)), a **modality** (have to, want to, permit, etc.) (see phrase (4.13)), a **negation word** (no, not, never, etc.) (see phrase (4.14)) or a **reported discourse verb** (state, say, think, etc.) (see phrase (4.15)). Modifiers in the corpus were detected using the MElt syntax analyzer with the CASOAR lexicon.

(4.12) je suis **très très** surpris! Bourdin a voté Hollande? J y crois pas MDR #ironie
*(I'm **very very** surprised! Bourdin voted for Hollande? I don't believe it LOL #irony)*

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

(4.13) #qc2014 P.Marois **veut** que P.Couillard “énonce” l’Arabie S. Et “énoncer” aussi Obama? #ironie
*(P.Marois **wants** P.Couillard to “denounce” S. Arabia. And “denounce” Obama too? #irony)*

(4.14) La bombe à fragmentation de DSK dans le Guardian C’est marrant mais **jamais** je n’ai pensé au complot #ironie
*(DSK’s splinter bomb in the Guardian was funny, but I’d **never** have thought it was a conspiracy #irony)*

(4.15) Pas de ministère intouchable pour les économies 2014 **annonce** #Cahuzac, le ministre intouchable #ironie
*(2014 budget cuts: no ministry will be spared, **says** #Cahuzac, the untouchable minister #irony)*

4.4.1.4. *Sentiment modifier features*

These two new features indicate whether a tweet contains an opinion word within the range of a modality or intensity adverb. For example, in tweet (4.16), the positive opinion word intelligent is modified by the adverb très (very).

(4.16) en même temps tu regarde les Tweets, tu vois bien que c’est des gens **très intelligents** qui on votez Hollande... #ironie
*(If you look at the tweets, you clearly see that **very smart** people voted for Hollande... #irony)*[7]

4.4.1.5. *Contextual features*

The context of an utterance is important in order to understand irony. These features indicate the presence/absence of contextual elements such as personal pronouns, useful in the detection of personal tweets, keywords for given themes and named entities identified by the MElt syntax analyzer.

[7] Translator’s note: There is an additional, unintentional irony in that the original French version of this tweet contains a number of grammatical errors.

4.4.1.6. *Opposition features*

These features are a new addition to the traditional list. They indicate the presence of explicit opposition using specific lexico-syntactic patterns. Our opposition features are partially inspired by Riloff *et al*. (2013) bootstrapping method for detecting sarcastic tweets, corresponding to an opposition between a positive sentiment/opinion and a negative situation. We extended this pattern in order to treat other types of opposition. For example, our patterns indicate whether a tweet contains (a) an explicit opposition in terms of sentiment/opinion, or (b) an implicit opposition between a subjective proposition expressing a sentiment/opinion and an objective proposition.

Let P_+ (respectively, P_-) be a subjective proposition containing a positive (respectively, negative) expression, P_{obj} an objective proposition with no expressions of opinion (P_{obj} may contain a negation) and Neg an operator that changes the polarity of the subjective words in P_+ (respectively, P_-). The patterns for **(a) explicit oppositions** of sentiment/opinion take the form:

– $[Neg(P_+)].[P'_+]$ or $[P_+].[Neg(P'_+)]$: Vraiment, [je **comprend pas** pourquoi Jerome Safar s'est fait battre par les verts...]$_{Neg(P_+)}$ [**Super**]$_{P'_+}$ #Municipales2014 #Grenoble
*(I really [**don't understand** why Jerome Safar got beaten by the Green Party...]$_{Neg(P_+)}$ [**Super**]$_{P'_+}$ #Municipales2014 #Grenoble)*;

– $[Neg(P_-)].[P'_-]$ or $[P_-].[Neg(P'_-)]$: [**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
*([**Sick of** being wiretapped]$_{P_-}$, Sarkozy wants to live like "an ordinary citizen". [**Not a bad idea**]$_{Neg(P'_-)}$, since no-one listens to "ordinary citizens")*;

– $[P_-].[P'_+]$ 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**.]$_{P'_-}$.
*(Emotional. Round of applause for the Ayrault government, probably [the absolute **worst** in the history of the Fifth Republic.]$_{P_+}$ [Very **entertaining**.]$_{P'_-}$.)*

The patterns for **(b) implicit oppositions** take the form:

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

*(Frankly, [I **don't understand** why people are criticizing Evra]$_{Neg(P_+)}$. The French left is solid as a rock !!! [Nothing's getting through]$_{P'_{obj}}$ #UKRFRA*;

– $[Neg(P_-)].[P'_{obj}]$ or $[P_{obj}].[Neg(P'_-)]$: [Avoir pour seul chroniqueur politique nicodomenach..]$_{P_{obj}}$ [On ne peut cependant **pas douter** de la neutralité de Canal+.]$_{Neg(P'_-)}$ #foutagedegueule
*([nicodomenach is the only political correspondent...]$_{P_{obj}}$ [But there's absolutely **no doubt** that Canal+ is neutral.]$_{Neg(P'_-)}$ #WhatAJoke)*;

– $[P_+].[P'_{obj}]$ or $[P_{obj}].[P'_+]$: Le soccer aux Jeux olympiques [c'est une **bonne chose**]$_{P_+}$ parce qu'[on n'en voit nulle part ailleurs]$_{P'_{obj}}$
*(Soccer at the Olympic Games [is a **great idea**]$_{P_+}$ since [we never get to see it anywhere else]$_{P'_{obj}}$)*;

– $[P_-].[P'_{obj}]$ or $[P_{obj}].[P'_-]$: [Kadhafi est mort tué d'une balle.]$_{P_{obj}}$ [C'est **moche** la guerre.]$_{P'_-}$ Je suis contre la guerre. #bhl.
*([Gaddafi was shot dead.]$_{P_{obj}}$ [War is **horrible**.]$_{P'_-}$ I don't like war. #bhl*

An opinion word is considered to be within range of a negation if the two words are separated by a maximum of two tokens. This simple rule has been shown to be effective, given that tweets are short messages containing a maximum of 140 characters.

4.4.2. *Experiments and results*

As in our previous experiment, we shall only present the best results, which were obtained using the SMO classifier.

We trained three classifiers, one for each corpus (NegOnly, NoNeg and All), denoted as$C_{NegOnly}$, C_{NoNeg} and C_{All}, under the same conditions used for the SurfSystem model.

The results presented here were obtained by training C_{NoNeg} using 1,720 tweets and testing the system using 430 tweets. C_{All} was trained using 2,472 tweets (1,432 of which contained negation, including 404 IR and 1,028 NIR) and tested on 618 tweets (360 containing negation, of which 66 were IR and 294 NIR).

4.4.2.1. *Feature relevancy for the training process*

For each classifier, we measured the impact of each feature group (described in section 4.4.1) for irony detection. For all of these experiments, we used surface features as the basic approach. For C_{NoNeg} and $C_{NegOnly}$, the feature indicating the presence of negation was deactivated.

4.4.2.2. *Contribution of each feature group to the training process*

During the training process, features were added one by one in order to highlight the influence of each on the accuracy of the classifier output. The results for the All, NoNeg and NegOnly corpora are shown in Tables 4.5–4.7.

Adding features one-by-one, we obtained the highest accuracy score, 87.70%, for the All corpus by combining all of the features (Table 4.5). For the NegOnly corpus, however, the highest accuracy score, 73.82%, was obtained by combining surface and sentiment features. Similarly, for the NoNeg corpus, the best accuracy score, 68.83%, was obtained by combining surface features with sentiment features alone.

4.4.2.3. *Results for the best feature combinations*

We examined the influence of different features for each subcorpus in order to identify the best combination in terms of accuracy. The results obtained and the best combination of features for each corpus are presented in Tables 4.8–4.10.

In our second experiment, we calculated the accuracy for each group of features in order to assess the impact of each feature group on performance (Table 4.11).

For the NegOnly corpus, the baseline produced good results compared to the other features; however, the baseline results were considerably poorer for two other corpora. The best results for the NoNeg corpus were obtained using the {tweet length, interjection, discursive connector, punctuation and quotation} features; for the All corpus, the best combination included {presence of punctuation and (capitalized word}.

Feature category	**Features**	**No of features**	**Ironic (IR)**			**Non ironic (NIR)**			**Accuracy**
–	–	–	P	R	F	P	R	F	–
Surface features (*baseline*)	–	1 – 12	0.531	0.955	0.682	0.774	0.155	0.259	55.50 %
Sentiment features	13. Positive opinion	1 – 13	0.567	0.563	0.565	0.566	0.57	0.568	56.63 %
	14. Negative opinion	1 – 14	0.567	0.563	0.565	0.566	0.57	0.568	56.63 %
	15. Neutral opinion	1 – 15	0.567	0.563	0.565	0.566	0.57	0.568	56.63 %
	16. Number of positive opinion words	1 – 12 + 16	0.565	0.647	0.603	0.587	0.502	0.541	57.44 %
	17. Number of negative opinion words	1 – 12 + 16– 17	0.575	0.686	0.625	0.61	0.492	0.545	58.89 %
	18. Number of neutral opinion words	1 – 12 + 16 – 18	0.584	**0.699**	**0.636**	**0.625**	0.502	**0.557**	**60.03 %**
	19. Surprise or shock	1 – 12 + 16 – 19	0.584	0.699	0.636	0.625	0.502	0.557	60.03 %
Sentiment modifier features	20. Intensifier followed by opinion	1 – 12 + 16 – 20	**0.588**	0.683	0.632	0.622	0.521	**0.567**	**60.19 %**
	21. Modality followed by opinion	1 – 12 + 16 – 21	0.586	**0.686**	0.632	0.621	0.515	0.563	60.03 %
Modifier features	22. Intensifier	1 – 12 + 16 – 22	0.582	0.676	0.626	0.614	0.515	0.56	59.54 %
	23. Modality	1 – 12 + 16 – 23	0.584	0.676	0.627	0.615	0.518	0.562	59.70 %
	24. Reporting	1 – 12 + 16 – 24	0.596	0.696	0.642	0.634	0.528	0.576	61.16 %
	25. Negation	1 – 12 + 16 – 25	**0.942**	**0.786**	**0.857**	**0.817**	**0.951**	**0.879**	**86.89 %**
Opposition features	26. Implicit opposition	1 – 12 + 16 – 26	0.942	0.786	0.857	0.817	0.951	0.879	86.89 %
	27. Explicit opposition	1 – 12 + 16 – 27	0.942	0.786	0.857	0.817	0.951	0.879	86.89 %
Context features	28. Personal pronoun	1 – 12 + 16 – 28	0.942	0.786	0.857	0.817	0.951	0.879	86.89 %
	29. Presence of topic in the text	1 – 12 + 16 – 29	0.942	0.786	0.857	0.817	0.951	0.879	86.89 %
	30. Named entity	1 – 12 + 16 – 30	**0.93**	**0.816**	**0.869**	**0.836**	**0.939**	**0.884**	**87.70** %

Table 4.5. *Results of the training process for the All corpus, feature-by-feature. The best results are shown in bold*

Feature category	**Features**	**No of features**	**Ironic (IR)**			**Non ironic (NIR)**			**Accuracy**
–	–	–	P	R	F	P	R	F	–
Surface features (baseline)	–	1 – 12	**0.847**	0.543	0.661	0.664	**0.902**	**0.765**	**72.23 %**
Sentiment features	13. Positive opinion	1 – 13	0.853	0.543	0.663	0.665	0.906	0.767	72.44 %
	14. Negative opinion	1 – 14	**0.867**	0.54	0.666	0.666	**0.917**	**0.772**	**72.87** %
	15. Neutral opinion	1 – 15	0.861	0.543	0.666	0.666	0.913	0.77	72.76 %
	16. Number of positive opinion words	1 – 12 + 16	0.856	0.543	0.664	0.665	0.909	0.768	72.55 %
	17. Number of negative opinion words	1 – 12 + 16 – 17	0.859	0.543	0.665	0.666	0.911	0.769	72.65 %
	18. Number of neutral opinion words	1 – 12 + 16 – 18	0.863	0.538	0.663	0.665	0.915	0.77	72.65 %
	19. Surprise or shock	1 – 12 + 16 – 19	**0.854**	0.574	0.687	0.679	**0.902**	**0.775**	**73.82 %**
Sentiment modifier features	20. Intensifier followed by opinion	1 – 15 + 19– 20	0.836	0.543	0.658	0.661	0.894	0.76	71.80 %
	21. Modality followed by opinion	1 – 15 + 19 – 21	0.844	0.543	0.661	0.663	0.9	0.764	72.12 %
Modifier features	22. Intensifier	1 – 15 + 19 – 22	0.842	0.543	0.66	0.662	0.898	0.762	72.02 %
	23. Modality	1 – 15 + 19 – 23	0.842	0.543	0.66	0.662	0.898	0.762	72.02 %
	24. Reporting	1 – 15 + 19 – 24	0.833	0.543	0.657	0.661	0.891	0.759	71.70 %
	25. Negation	–	–	–	–	–	–	–	–
Opposition features	26. Implicit opposition	1 – 15 + 19 – 26	0.851	0.57	0.683	0.677	0.9	0.773	73.51 %
	27. Explicit opposition	1 – 15 + 19 – 27	0.851	0.57	0.683	0.677	0.9	0.773	73.51 %
Contextual features	28. Personal pronoun	1 – 15 + 19 – 28	0.851	0.57	0.683	0.677	0.9	0.773	73.51 %
	29. Presence of topic in the text	1 – 15 + 19 – 29	0.851	0.57	0.683	0.677	0.9	0.773	73.51 %
	30. Named entity	1 – 15 + 19 – 30	0.883	0.532	0.664	0.665	0.93	0.776	73.08 %

Table 4.6. *Results of the training process for the NegOnly corpus, feature-by-feature. The best results are shown in bold*

Feature category	Features	No. of features	Ironic (IR)			Non-ironic (NIR)			Accuracy
–	–	–	P	R	F	P	R	F	–
Surface features (*baseline*)	–	1 – 12	0.635	0.623	0.629	0.63	0.642	0.636	**63.25 %**
Sentiment features	13. Positive opinion	1 – 13	0.668	0.6	0.632	0.637	0.702	0.668	65.11 %
	14. Negative opinion	1 – 14	0.695	0.647	0.67	0.67	0.716	0.692	68.13 %
	15. Neutral opinion	1 – 15	**0.701**	0.656	0.678	0.677	**0.721**	0.698	**68.83 %**
	16. Number of positive opinion words	1 – 12 + 16	0.677	0.586	0.628	0.635	0.721	0.675	65.34 %
	17. Number of negative opinion words	1 – 12 + 16 – 17	0.682	0.567	0.619	0.629	0.735	0.678	65.11 %
	18. Number of neutral opinion words	1 – 12 + 16 – 18	0.71	0.581	0.639	0.646	0.763	0.699	67.20 %
	19. Surprise or shock	1 – 12 + 16 – 19	0.656	0.665	0.661	0.66	0.651	0.656	65.81 %
Sentiment modifier features	20. Intensifier followed by opinion	1 – 15 + 19 – 20	0.695	0.656	0.675	0.674	0.712	0.692	68.37 %
	21. Modality followed by opinion	1 – 15 + 19 – 21	0.695	0.656	0.675	0.674	0.712	0.692	68.37 %
Modifier features	22. Intensifier	1 – 15 + 19 – 22	0.695	0.656	0.675	0.674	0.712	0.692	68.37 %
	23. Modality	11 – 15 + 19 – 23	0.706	0.647	0.675	0.674	0.73	0.701	68.83 %
	24. Reporting	1 – 15 + 19 – 24	0.698	0.656	0.676	0.675	0.716	0.695	68.60 %
	25. Negation	–	–	–	–	–	–	–	–
Opposition features	26. Implicit opposition	1 – 15 + 19 – 26	0.654	**0.702**	0.677	0.678	0.628	0.652	66.51 %
	27. Explicit oposition	1 – 15 + 19 – 27	0.652	0.637	0.647	0.645	0.66	0.653	64.88 %
Contextual features	28. Personal pronoun	1 – 15 + 19 – 28	0.66	0.651	0.656	0.656	0.665	0.661	65.81 %
	29. Presence of topic in the text	1 – 15 + 19 – 29	0.665	0.665	0.665	0.665	0.665	0.665	66.51 %
	30. Named entity	1 – 15 + 19 – 30	0.67	0.651	0.66	0.661	0.679	0.67	66.51 %

Table 4.7. *Results of the training process for the NoNeg corpus, feature-by-feature. The best results are shown in bold*

Features	Ironic (IR)			Non-ironic (NIR)			Accuracy (%)
–	P	R	F	P	R	F	–
Baseline	0.531	0.955	0.682	0.774	0.155	0.259	55.50 %
All features	0.93	0.816	0.869	0.836	0.939	0.884	87.70 %
Best combination of features: capitalization, opposition word, number of words, intensifier followed by opinion, explicit opposition, implicit opposition, negation, surprise/shock	0.93	0.816	0.869	0.836	0.939	0.884	**87.70 %**

Table 4.8. *Comparison of results obtained for the All corpus. The best results are shown in bold*

Features	Ironic (IR)			Non ironic (NIR)			Accuracy (%)
–	P	R	F	P	R	F	–
Baseline	0.847	0.543	0.661	0.664	0.902	0.765	72.23 %
All features	0.847	0.564	0.677	0.673	0.898	0.769	73.08 %
Best combination of features: capitalization, quotation, implicit opposition	0.889	0.56	0.687	0.679	0.93	0.785	**74.46 %**

Table 4.9. *Comparison of results obtained for the NegOnly corpus. The best results are shown in bold*

Features	Ironic (IR)			Non-ironic (NIR)			Accuracy (%)
–	P	R	F	P	R	F	–
Baseline	0.591	0.651	0.619	0.611	0.549	0.578	63.25 %
All features	0.673	0.642	0.657	0.658	0.688	0.673	66.51 %
Best combination of features: punctuation, capitalization, interjection, quotation, discursive and opposition connectors, opposition, intensifier, modality, positive opinion, negative opinion, neutral opinion, number of words, intensifier followed by opinion, modality followed by opinion	0.711	0.651	0.68	0.678	0.735	0.705	**69.30 %**

Table 4.10. *Comparison of results obtained for the NoNeg corpus. The best results are shown in bold*

	NegOnly	*NoNeg*	*All*
***Baseline* (Surface features)**	**73.08**	63.25	55.50
Best surface features	73.08	64.65	56.31
Best sentiment features	57.02	**67.90**	58.25
Sentiment modifiers	53.51	56.51	51.94
Modifiers	53.72	55.81	**86.89**
Opposition	55.31	63.02	79.77
Internal context	55.53	53.25	53.55

Table 4.11. *Accuracy results of all three experiments by feature group. The best results are shown in bold*

From Table 4.11, we draw the following conclusions:

– In NegOnly, semantic features alone (sentiment, modifiers, opposition, etc.) are not sufficient to classify NIR and IR tweets. Note that the baseline results for the NIR class are better than those for IR (f-measures of 77.60 and 66.40, respectively).

– Sentiment features are most reliable for C_{NoNeg}, using the surprise/shock feature alongside opinion word frequency features. Once again, the results for the NIR class are significantly better than for IR, with an f-measure of 73.30, 12.7 points higher than IR.

– Modifier and opposition features perform best for C_{All}. As for the other corpora, we see that the classifier predictions are better for the NIR class than for the IR, but the difference here is smaller (2.2 points for modifiers, 7.4 points for oppositions).

Table 4.12 shows the overall results obtained by a classifier trained using all relevant features for each group. These results are expressed in terms of precision (P), recall (R), f-measure (macro-mean) and accuracy. The results are better for C_{All} than for $C_{NegOnly}$ and C_{NoNeg}. These figures were obtained using **three surface features** {capitalized words, opposition connectors, tweet length}, **modifiers** {presence of intensifiers and negations} and **opposition features** {presence of explicit and implicit opposition}. The best combination for $C_{NegOnly}$ consists of **two surface features** {capitalization, quotation} and **opposition features**. Finally, in the case of tweets without negation (C_{NoNeg}), the highest accuracy score of 69.30% is obtained using the following combination: **surface features** {punctuation, capitalization, interjection, quotation, discursive connectors, opposition words, tweet length}, **sentiment features** {presence of positive/negative/ neutral opinion words} and **sentiment modifier features** {opinion words modified by an intensifier or modality}.

	Ironic (IR)			**Non-ironic (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
	Results (best combination)					
	f-measure (macro-mean)			Accuracy		
$C_{NegOnly}$	73.60			74.46		
C_{NoNeg}	69.25			69.30		
C_{All}	**87.65**			**87,70**		

Table 4.12. *Results for the best feature combinations. The best results are shown in bold*

Based on these results, the following four conclusions are obtained:

– surface features are essential in detecting irony, especially for tweets not containing negation;

– negation is an important feature for this task, but is not sufficient: of the 76 tweets that were wrongly classified by C_{All}, 60% contained negations (37 IR and 9 NIR);

– for tweets containing negation, opposition features are most effective;

– opinion words are more likely to be used in tweets without negation.

4.4.3. *Discussion*

The results of our experiments using All ($P = 93\%$, $R = 81.6\%$, f-measure $= 86.9\%$), NegOnly($P = 88.9\%$, $R = 56\%$, f-measure $= 68.7\%$) and NoNeg corpora ($P = 71.1\%$, $R = 65.1\%$, f-measure $= 68\%$) are very encouraging when compared to other work carried out on the same subject and using the same type of corpus. For example, (Liebrecht *et al*. 2013) obtained a precision value of 30% for Dutch. For English, other researchers have obtained F-measures of 76% (Reyes *et al*. 2013), 76% (Barbieri and Saggion 2014b) and 88.76% (Joshi *et al*. 2015).

Analysis of the errors made by the three classifiers shows that these mistakes mostly result from four factors:

– **Presence of comparison**: a form of irony in which characteristics are attributed to an element by comparing it to a completely different element (for example, this employee is about as much use as a chocolate teapot). This type of irony often includes comparison markers (see Chapters 1 and 2, section 2.5). We shall not go into detail regarding this phenomenon, but a semantic similarity approach might be helpful (Veale and Hao 2010).

– **Lack of context**: this accounts for the majority of errors. Many wrongly classified tweets cannot be interpreted without contextual knowledge, which is not contained within the tweet. This lack of context may take a variety of forms, for example:

- the theme of the tweet is not mentioned (Hey, we missed her! #irony), or the irony can only be inferred from hashtags (e.g. *#AprilFools*);

- the irony relates to a specific situation (situational irony) (e.g. #Sarkozy was one of the authors responsible for the law of March 9, 2004 establishing rules for wiretapping #irony: the irony arises from the fact that Sarkozy himself had his phone tapped, and had contributed to the law authorizing the practice);

- the presence of false assertions (e.g. Don't worry. Senegal will be the world soccer champions);

- oppositions implying a contradiction between two words that are not semantically linked (e.g. "UN" and "terrorist organization" or "Chad" and "democratic elections"). This situation is most common in tweets without negation, whereas cases (2) and (3) are more frequently encountered in tweets where negation is present.

– **Humor**: humor in tweets with the *#irony* hashtag may lead to automatic classification errors. This is due to the use of specific vocabulary expressing humor, which is not covered by the feature used to detect irony (e.g.: DSK's splinter bomb in the Guardian was funny, but I'd never have thought it was a conspiracy #irony).

– **Misuse of hashtags**, where the *#ironie* or *#sarcasme* labels were used in tweets containing non-ironic text (e.g. #Seydou thanks #Thauvin for "his professionalism" and "his love for the club" #OM #Losc #irony: in this case, the author's ironic intent is not clear as he is simply quoting statements).

This error analysis also showed that, globally, the tweets concerned with classification errors and the tweets concerned with interannotator disagreements during the manual annotation phase (see section 4.2, for the purposes of hashtag verification) were one and the same. Additionally, the annotation phase indicated that the presence of negations may give rise to ambiguity in terms of the understanding and identification of irony by human readers. This observation also holds true for the automatic classification process, as we obtained lower accuracy scores for the NegOnly corpus (74.46%) than for the All corpus (87.70%). However, the lowest accuracy score was obtained for the NoNeg corpus (69.30%). These results seem to suggest that negation is a key indicator for automatic irony detection, despite being seen as an obstacle to understand for human readers and for the automatic classifier used to detect tweets containing negation.

Error analysis across all of our experiments shows that the presence of negation and the absence of context are responsible for the majority of errors. For this reason, we chose to focus on these phenomena in developing a new method to improve the classification of those tweets, which were wrongly labeled by the classifier. This method is described below.

4.5. The QuerySystem model: developing a pragmatic contextual approach for automatic irony detection

The results of the experiments described above confirm the need for external knowledge in understanding the ironic meaning of tweets. This issue was brought up by the human annotators working on the campaign described in Chapter 3, and in a number of recent works by other authors, including (Wallace 2015, Bamman and Smith 2015) and (Joshi *et al*. 2016), presented in Chapter 2.

In this section, we present a new approach, which corrects the output from the automatic detection system presented above in order to correctly classify tweets as ironic/non-ironic using external knowledge.

4.5.1. *Proposed approach*

The main hypothesis guiding our approach is that users who publish tweets on popular topics tend to be commenting on or criticizing a situation or person featured in the news. This means that the facts or events related in tweets can be verified. If not, the tweet is likely to contain a false assertion, and as such may well express irony.

Our approach consists of searching for the external context of a tweet, carrying out a Google search via the API to check the veracity of tweets. Based on the definition of irony as a means of mocking someone or something by saying the opposite of what one intends to convey, we decided to focus specifically on tweets in which irony manifests through the presence of negation. There were three main reasons for this choice:

– negations may be the means of expressing and ironic false assertion (for example, Si Hollande est élu, il serait capable de donner des responsabilités à DSK. Pourquoi pas la direction du FMI tant que l'on y est. (If Hollande is elected he might give DSK an important role. Why not make him director of the IMF while we are at it.);

– the strong presence of negations in tweets (62.75% of the tweets in our corpus contain a negation) and the role of false assertions (accounting for 56% of ironic tweets with implicit contradiction);

– the results of our classifier are poorer for tweets containing negations (see section 4.5).

The aim of the proposed approach is to verify the truth of tweets containing negation that have been judged to be non-ironic by the classifier, but which include the reference hashtag(s). Thus, if a tweet of the form $Not(P)$ has been classed as non-ironic but P is verified by online sources, then the tweet class should be corrected to “ironic”.

The class of a tweet may be changed using the following algorithm. Let $WordsT$ be the set of words, excluding empty words, in a tweet t, and let kw be the keyword (*topic*) used to collect tweet t. Let $N \subset WordsT$ be the set of negation words in t. The proposed algorithm is defined as follows:

1) segment t into a set of sentences S;

2) for each sentence $s \in S$ such that $\exists neg \in N$ and $neg \in s$:

a) 2.1) remove symbols # and @, emoticons and neg, then extract all tokens $P \subset s$ that are within the range of neg (within a maximum distance of two tokens);

b) 2.2) generate a query $Q_1 = P \cup kw$ and submit it to Google, which will provide a maximum of 20 results composed of a title and a snippet;

c) 2.3) select the most reliable of the results provided by Google (Wikipedia, press articles and websites without the terms "Blog" or "Twitter" in their URL). For each result, if the keywords from the Google query are found in the title or snippet, then t is considered to be ironic STOP;

3) generate a second query $Q_2 = (WordsT - N) \cup kw$ and submit it to Google, then follow the procedure described in step 2.3. If Q_2 is found, then t is considered to be ironic. Otherwise, the class predicted by the system will remain unchanged.

Google responds to queries by showing a web page containing a list of **snippets** representing the search results. A snippet generally contains a title, a URL, a description and, in some cases, additional information (image, votes, price, etc.). Google limits the number of characters to 66 for the title, 156 for the description and 65 for the URL (with some exceptions). The number of characters within the description zone must be below this threshold in order to avoid content being cut off.

Webmasters have the ability to personalize snippets by adapting the content of their web pages. Google also uses rules to understand the content of websites. The term *rich snippet* denotes a tagging system, which is read by Google, but is not visible to the user. This makes it easier to identify certain data (e.g. product price and recipe preparation time). The search engine may then adapt a snippet to improve its presentation to users. Google's API permits snippet collection and allows users to parameterize the number of snippets to show per query. We were able to make use of these possibilities in our experiments, limiting our collection to a maximum of 20 snippets per query.

The example given below shows the application of our algorithm to a tweet from the corpus, collected using the Twitter API with the keyword *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.
(#Valls learned that #Sarkozy had been wiretapped from the newspaper. Good job he's not Minister of the Interior).

In step 1 of the algorithm, the tweet is segmented into two 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).

In step 2.1, the negation words "n" and "pas" are removed, and the segment within range of the negation is extracted. This gives us $P = \{ministre, int\tilde{\mathrm{A}}rieur\}$.

Step 2.2 generates a first request $Q_1 =$ *Valls ministre intérieur*.

Step 2.3 results in a collection of 20 snippets, of which the first two are shown below:

> <Résultat id="1">
> <Titre>*Manuel <b>Valls</b> - Wikipedia*</Titre>
> <Url>https://fr.wikipedia.org/wiki/ManuelValls</Url>
> <Snippet>... Homme politique français. Pour le compositeur espagnol, voir Manuel <b>**Valls**</b> (compositeur). .... <b>**Valls**</b> a été nommé <b>**ministre**</b> de l'<b>**Intérieur**</b> 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. ... <b>Valls</b> a été rocardien et il en a gardé le meilleur : le parler vrai. ... Alors que son <b>ministre</b> de l'<b>Intérieur</b>, Bernard Cazeneuve, poussait des ronrons de satisfaction, saluant ...</Snippet>
> <Résultat id="3">
> <Titre>Le gouvernement tente de minimiser le différend <b>Valls</b>-Taubira</Titre>
> <Url>http://www.lemonde.fr/societe/article/2013/08/14/apres-sa-missive-contre-taubira-manuel-valls-fait-profil-bas34612033224.html</Url>
> <Snippet>14 août 2013 ... Le <b>ministre</b> de l'<b>intérieur</b> et le premier <b>ministre</b> se sont exprimés après la ... "Qu'il y ait débat, c'est normal, heureusement qu'il y a des débats.</Snippet>

All of our keywords are found in the text of the first snippet, a Wikipedia page on Manuel Valls. Each keyword is labeled using the tags <b> </b>. The query is therefore verified, and the proposition *Heureusement qu'il n'est pas ministre de l'Intérieur* is a false assertion (Manuel Valls was, in fact, Minister of the Interior). We thus conclude that the tweet is ironic.

4.5.2. *Experiments and results*

Several experiments were carried out to assess the impact of our query-based system on tweet classification. The method was applied to the All and NegOnly corpora (as the NoNeg corpus contains no negation, the query system would not be relevant):

– ① the first experiment evaluated our method for tweets with negation classed as non-ironic (NIR) by the PragSystem classifier, but classed as ironic according to the reference hashtags;

– ② our second experiment consisted of applying the method to all tweets with negation that were classified as NIR by PragSystem, whether the prediction was correct.

Table 4.13 shows the results of these experiments.

	①		②	
Number of NIR tweets for which:	**All**	**NegOnly**	**All**	**NegOnly**
Query applied	37	207	327	644
Results obtained from Google	25	102	166	331
Class corrected to IR	5	35	69	178
Classifier accuracy	87.7	74.46	**87.7**	**74.46**
Accuracy after queries	**88.51**	**78.19**	78.15	62.98

Table 4.13. *Results of the query method using Google (experiments 1 and 2). The best results are shown in bold*

All of the scores obtained using the query method are statistically significant in comparison with the classifier scores (P_value <0.0001 calculated using McNemar's test). Error analysis showed that 65% of tweets which were still wrongly classified after applying this method were almost impossible to verify online, as they were either of a personal nature or lacking internal context. From this, we conclude that our method is not suitable for application to tweets of this type.

We repeated experiments ① and ② using only tweets free from personal content, selected based on internal context features, containing neither named entities nor first-person personal pronouns (I, me, you). Tweets that contain no personal pronouns or named entities may contain personal content that would be impossible to validate online (for example, She missed us! #irony).

The results of these experiments are shown in Table 4.14. Again, all of the scores for the query-based method are statistically significant compared to the classifier scores.

	①		②	
Number of NIR tweets for which:	**All**	**NegOnly**	**All**	**NegOnly**
Query applied	0	18	40	18
Results obtained from Google	–	12	17	12
Class corrected to IR	–	4	7	4
Classifier accuracy	87.7	74.46	**87.7**	74.46
Accuracy after queries	**87.7**	**74.89**	86.57	**74.89**

Table 4.14. *Results of the query-based method using Google for non-personal tweets (experiment 3). The best results are shown in bold*

For experiment ①, our method was not applied to the All corpus, as all of the wrongly classified tweets contained a personal pronoun and no named entity. Overall, the query-based method resulted in an improvement in classifier results in all cases, except for the All corpus, where Google results were obtained for only 42.5% of queries. Results were obtained for over 50% of queries in all other experiments (with a maximum of 66.6% in NegOnly). Tweets for which no results were found are those that featured named entities, but did not mention an event (e.g., AHAHAHAHAHA! No respect #Legorafi: "Legorafi" is a satirical paper).

4.5.3. *Evaluation of the query-based method*

To assess the difficulty of the classification task, two human annotator were also asked to label the 50 tweets (40 from All and 10 from NegOnly) to which the query system was applied as ironic or non-ironic. The Cohen's kappa between the annotators was low, at only κ =0.41. Of the 12 tweets, which were corrected to "Ironic", the annotators failed to reach an agreement on five tweets. Although this experiment does not provide sufficient basis for a formal conclusion, due to the small number of tweets which were annotated, it seems to indicate that human operators would not have a higher success rate than the automatic system in cases where the latter made errors.

It is interesting to note that although internal contextual features were not relevant for automatic tweet classification (see PragSystem model), our results shows that they remain useful for improving classification. As experiment ① demonstrated, the query-based method is more effective when applied to wrongly classed tweets. We thus consider that the use of internal contextual features (presence of personal pronouns and named entities) may offer a means of automatically detecting tweets, which may be wrongly classified.

4.6. Conclusion

In this chapter, we proposed an approach for automatically detecting irony in French tweets, aiming to verify two hypotheses:

– Hypothesis (H1): the presence of negations as an internal property of an utterance may help to detect disparity between the literal and intended meanings of the utterance.

– Hypothesis (H2): a tweet containing an affirmed fact of the form $Not(P)$ is ironic if, and only if, P can be confirmed based on general knowledge, external to the utterance, shared by both the author and the reader.

To test the validity of these hypotheses, we split our FrIC^{Auto} into three parts, for tweets containing negations (NegOnly), tweets without negations (NoNeg) and tweets of both types (All). Three models were proposed for detecting irony in three corpora:

1) **SurfSystem**, a model based on the surface features presented in existing literature. Our results showed that the features used for this task in other languages are also effective for French (for example, the presence of punctuation or of words in capitals), and that they work best on the NegOnly corpus.

2) **PragSystem**, a model using pragmatic features extracted from the linguistic content of tweets. We used features previously described in literature and added three new features: sentiment modifiers, context features and opposition features. The latter, obtained using opposition patterns, were the most effective. The best results were obtained for the All corpus (87.7% accuracy). Error analysis showed that specific treatment of tweets with negation would be beneficial.

3) **QuerySystem**, a query-based method for tweets containing negations, with the capacity to verify the truth of propositions of the form $Not(P)$ using reliable online sources. The idea is to correct the predictions made by the PragSystem classifier for tweets containing negation and classed as non-ironic. Our experiments showed that this method improves classification when applied to non-personal tweets.

Our evaluation of these three models shows hypotheses (H1) and (H2) to be valid.

In the following chapter, we shall examine the portability of our conceptual and computational models for irony detection in a multilingual context. We shall begin by examining the portability of our annotation scheme for tweets in two indo-European languages, Italian and English, before testing our computational model on Arabic.

5

Towards a Multilingual System for Automatic Irony Detection

5.1. Introduction

In Chapters 3 and 4, we proposed a multilevel annotation scheme for irony and an automatic irony detection system for tweets in French. In this chapter, we shall assess the portability of both the annotation scheme and the automatic detection system for other languages. Two experiments will be presented.

The first experiment consists of testing the portability of the multilevel annotation scheme for irony, described in Chapter 3, for other languages in the same family as French: English and Italian. An annotation campaign is carried out using our scheme, designed to analyze pragmatic phenomena relating to irony in French tweets. The aim of this first experiment is not only to test the performance of the annotation scheme on other Indo-European languages that are culturally similar to French, but also to measure the impact of a set of pragmatic phenomena on the interpretation of irony. Additionally, we shall consider the way in which these phenomena interact with the local context of a tweet in languages belonging to the same family.

The second experiment consists of testing the performance of the feature-based automatic irony detection system (using the SurfSystem and PragSystem models – see Chapter 4) on tweets written in Arabic. For the purposes of this experiment, we constructed the first corpus of ironic and non-ironic tweets in Arabic, studying the performance of features and assessing the algorithms used to classify tweets as ironic/non-ironic.

Automatic Detection of Irony: Opinion Mining in Microblogs and Social Media,
First Edition. Jihen Karoui; Farah Benamara and Véronique Moriceau.

In what follows, we shall describe our two experiments and the obtained results. Section 5.2 is devoted to the first experiment and includes a description of the corpora used for English and Italian, along with the quantitative results for each level of the annotation scheme in each language. Section 5.3 provides a description of the second experiment, with an overview of the specificities of Arabic and a presentation of the tweet corpus used in this context. We shall then provide the quantitative results obtained through our experiment, comparing them with the results obtained for French as presented in Chapter 4.

5.2. Irony in Indo-European languages

> “In linguistics, the Indo-European languages (formerly known as Indo-Germanic or Scythian languages) form a family of closely related languages with shared roots in what is commonly referred to as proto-Indo-European. They possess strong lexical, morphological and syntactic similarities; it is thus supposed that each group of comparable elements evolved from the same original form, now extinct. There are around one thousand languages in this family, currently spoken by approximately three billion people”.[1]

From this definition, we see that linguists have noted considerable morphological and syntactic similarities between most Indo-European languages. This is encouraging for the purposes of our research, considering the irony phenomenon in different languages within this family. Within this framework, we shall focus on English and Italian.

5.2.1. *Corpora*

5.2.1.1. *Collection of the English corpus*

Although other corpora had already been developed (Reyes *et al*. 2013), we decided to construct our own corpus of ironic and non-ironic tweets in English. The existing corpora were essentially composed of personal tweets (e.g. *Don’t worry about what people think. They don’t do it very often*), meaning that they were not comparable to the corpus used for research in

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

French. To create our corpus, we followed the same collection procedure used for the French corpus FrIC. We selected a set of themes belonging to the same categories used previously, adapted to language and culture-specific news. For example, for the **politics** category, we selected Obama, Trump, Clinton, etc.; for **artists**, we selected keywords including Justin Bieber, Kardashian, Beyoncé, etc. Next, we selected ironic tweets containing these keywords along with the hashtag *#ironic* or *#sarcasm*. Non-ironic tweets (e.g. tweets without the *#ironic* or *#sarcasm* hashtags) were collected in the same way.

Following collection, we removed all duplicates, retweets and tweets containing images. At the end of the filtering process, we had a corpus of 11,289 tweets, of which 5,173 were ironic and 6,116 were non-ironic. The distribution of these tweets by category is shown in Table 5.1.

Themes	**Ironic**	**Non-ironic**
Economy	117	79
Generic	311	873
Cities or countries	1,014	891
Artists	472	836
Politics	2,560	2,294
Health	142	160
Sports	557	983
Total	**5,173**	**6,116**

Table 5.1. *Distribution of tweets in the English corpus*

5.2.1.2. *Collection of the Italian corpus*

For Italian, we made use of two existing resources, annotated as part of the Senti-TUT project[2]:

– The Sentipolc corpus used in the *Evalita 2014* evaluation campaign[3] for sentiment analysis and irony detection in tweets (Basile *et al*. 2014). The Sentipolc corpus is a collection of tweets in Italian derived from two existing corpora: Senti-TUT (Bosco *et al*. 2013) and TWITA (Basile and Nissim 2013). It contains tweets using keywords and hashtags on the theme of politics (names of politicians, etc.). Each tweet in Sentipolc is annotated using five mutually

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

[3] http://di.unito.it/sentipolc14.

exclusive categories: positive opinion, negative opinion, both positive and negative opinion, ironic and objective.

– The TW-SPINO corpus, containing tweets from Spinoza[4], a satirical political blog in Italian. These tweets were selected and reviewed by a team of editors who identified them as being ironic or satirical.

The Italian corpus is made up of 3,079 ironic tweets (806 from Sentipolc and 2,273 from TW-SPINO) and 5,642 non-ironic tweets (from Sentipolc).

5.2.2. *Results of the annotation process*

To study the portability of our annotation scheme, we focused on annotating a subset of tweets in English and Italian. The aim of this first step was to test the performance of our scheme for other languages, and to compare the statistical results obtained for English, Italian and French.

In this section, we will present the quantitative results obtained from the annotated corpora in English and Italian. Two human annotators worked on each corpus, following a two-step process. In the first step, 100 tweets in each language (50 ironic, 50 non-ironic) were used for training. In the second step, 550 tweets in English and 500 tweets in Italian (80% ironic, 20% non-ironic) were annotated. The first step was essential in order for annotators to become familiar with the annotation scheme and the corpora in question.

Our annotation scheme, described in Chapter 3, includes four levels (see Figure 5.1). In level 1, tweets are identified as ironic or non-ironic. Level 2 identifies explicit or implicit contradictions in ironic tweets. Level 3 concerns the category of irony (analogy, hyperbole, etc.), while level 4 relates to markers (negation, punctuation, etc.).

For the English corpus, the annotators used Glozz, the same tool as was used for the French corpus, and annotated tweets using four levels of our scheme. We provided the annotators with an English translation of our annotation guide and the files needed to operate Glozz.

Level 1 of the annotation process (ironic/non-ironic) was already included in our Italian corpus. Manual annotation was only carried out for levels 2 and 3

[4] www.spinoza.it/.

of the scheme (irony types and categories), following the same guide used for English and French[5]. Cue annotation (level 4) was carried out automatically, and certain markers were verified manually (negations and emoticons). For this reason, we only provide results for those markers that were manually verified.

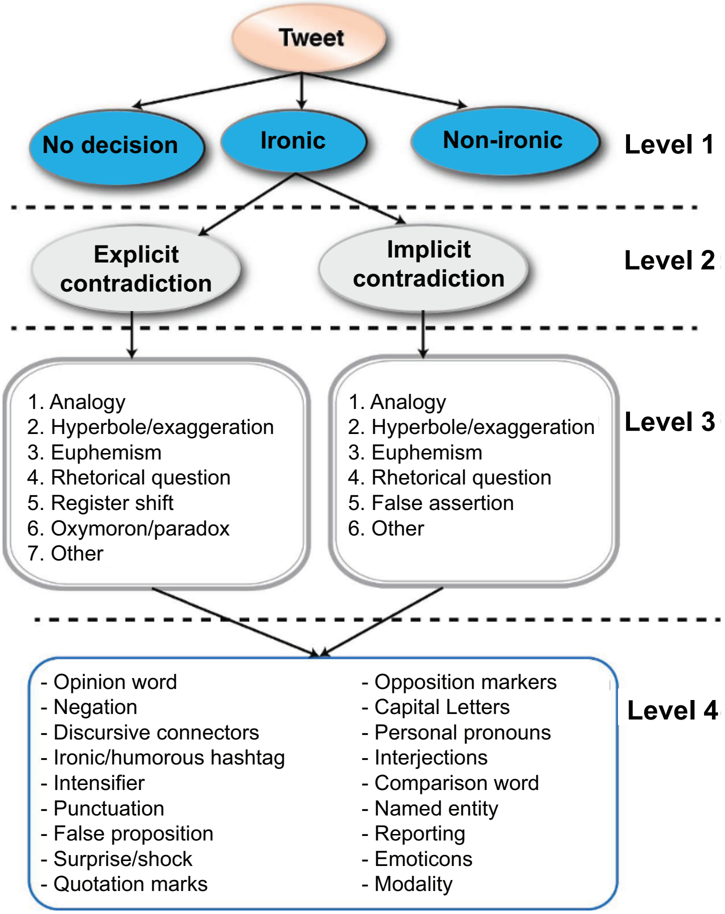


Figure 5.1. *Annotation scheme. For a color version of the figures in this chapter see, www.iste.co.uk/karoui/irony.zip*

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

Through this annotation campaign, we were able to analyze the following:

– presence of markers in ironic tweets;

– variation in the presence of markers for each irony type (explicit/implicit contradiction) and each category (hyperbole, analogy, paradox, etc.);

– frequency of irony categories for each type of irony.

5.2.2.1. *Quantitative results for ironic/non-ironic annotation*

Based on the reference hashtags *#ironic* and *#sarcasm* in the English corpus, 440 (80%) of the tweets were ironic and 110 (20%) were non-ironic. The human annotators assessed 427 (77.63%) of the tweets as ironic and 99 (18%) as non-ironic, with the remaining 24 tweets (4.37%) being placed into the "no decision" class (Figure 5.2). These results show that, whatever the language used, a tweet including an irony hashtag is not necessarily ironic, nor is the hashtag a prerequisite for irony.

Unlike the French and English corpora, the Italian corpus was annotated as ironic/non-ironic by humans instead of using hashtags (as part of the Senti-TUT project). This corpus contains 400 ironic tweets and 100 non-ironic tweets, with no tweets in the "no decision" class (Figure 5.2).

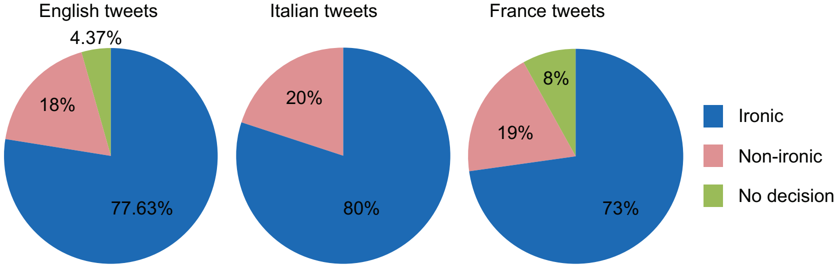


Figure 5.2. *Distribution of English, Italian and French tweets*

5.2.2.2. *Quantitative results for irony type annotation*

Table 5.2 shows the total number of annotated tweets and the type of irony trigger in each corpus.

In the English corpus, out of 427 tweets annotated as ironic, 283 featured implicit contradiction and 144 contained explicit contradiction. This indicates that irony is generally expressed implicitly in English (66.28% of cases), as it is in French (73.01%).

In the Italian corpus, however, the majority of ironic tweets featured explicit contradictions (65%). This may stem from the fact that Italian users do not use specific hashtags to indicate irony, perhaps causing them to be more explicit in their expressions.

	Ironic		**Non-ironic**	**No decision**	**Total**
	Explicit	**Implicit**	**–**	**–**	**–**
French	394 (19.7%)	1,066 (53.3%)	380 (19%)	160 (8%)	2,000
English	144 (26.2%)	283 (51.45%)	99 (18%)	24 (4.35%)	550
Italian	260 (52%)	140 (28%)	100 (20%)	–	500

Table 5.2. *Number of tweets annotated by irony type in the French, English and Italian corpora*

5.2.2.3. *Quantitative results for irony category annotation*

Table 5.3 shows the percentage of tweets belonging to each category of irony, split into explicit/implicit trigger groups[6]. Significant differences can be seen in terms of the categories of irony found in the French, English and Italian corpora. The results show that:

– for irony including explicit contradiction, oxymoron/paradox is the most common category for all three languages (French, English and Italian);

– for irony with implicit contradiction, false assertion and other are the most common categories in French and English. In Italian, the most frequent categories are false assertion, analogy and other;

– considering the tweets in the other category, we see that the majority are ironic with implicit contradiction. This indicates that the decision task is harder for humans in cases where irony is expressed through an implicit contradiction, whichever language is used.

As the classes are not mutually exclusive:

– for the English corpus, 35 tweets with explicit contradiction belong to more than one category, and 62 tweets with implicit contradiction

[6] The bold values show the highest frequencies.

belong to more than one category. For explicit contradictions, the most frequent combination is oxymoron/paradox + rhetorical question, while for ironic tweets with implicit contradiction, the most frequent contradiction is metaphor/comparison + other;

– for the French corpus FrIC, the most frequent combination for explicit contradictions is oxymoron/paradox + hyperbole/exag-geration; for implicit contradictions, the most frequent combination is false assertion + hyperbole/exaggeration;

– for the Italian corpus, the annotators chose to assign a single category to each tweet, selecting the option which seemed to express irony most strongly.

Thus, for French and English, the oxymoron/paradox category is one of the most frequent combinations for irony with explicit contradiction. For cases featuring implicit contradiction, the combinations are different for the two languages.

	Analogy			**Register shift**			**Euphemism**			**Hyperbole**		
	F	A	I	F	A	I	F	A	I	F	A	I
Explicit	12%	17%	21%	1%	6%	19%	1%	1%	5%	8%	2%	9%
Implicit	2%	13%	**26%**	–	–	–	1%	1%	4%	10%	7%	5%

	Rhetorical question			**Oxymoron/paradox**			**False assertion**			**Other**		
	F	A	I	F	A	I	F	A	I	F	A	I
Explicit	10%	15%	10%	**66%**	**81%**	**28%**	–	–	–	21%	6%	7%
Implicit	14%	1%	12%	–	–	–	**56%**	20%	**34%**	**32%**	**65%**	**19%**

Table 5.3. *Distribution of categories for explicit/implicit trigger types in the French (F), English (A) and Italian (I)corpora. The best results are shown in bold*

5.2.2.4. *Quantitative results of the annotation procedure for irony markers*

Three statistical studies were carried out for these levels. The first is a quantitative study looking at the first and fourth levels of the annotation scheme, concerning the presence of different markers in ironic and non-ironic tweets (Table 5.4). The second study concerns the second and fourth levels, focusing on the presence of different markers in tweets with explicit versus implicit contradiction (Table 5.4). Finally, our third study relates to the third and fourth levels, specifically the presence of markers in each irony category (Table 5.5).

	Emoticon			**Negation**			**Discursive connectors**			**# Humorous***			**Intensifier**			**Punctuation**		
	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I
Explicit	7	2	1	37	58	15	6	41	29	2	14	-	22	9	2	51	30	14
Implicit	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

	False* proposition			**Surprise**			**Modality**			**Quotation**			**Opposition**			**Capital letters**		
	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I
Explicit	8	0	-	3	0	-	0	2	3	6	21	3	9	18	4	3	8	-
Implicit	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	-

	Personal* pronoun			**Interjection**			**Comparison***			**Named* entities**			**Reporting verb**			**Opinion**			**URL***		
	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I
Explicit	31	21	5	14	2	11	8	8	4	97	100	65	1	17	0	48	75	-	33	0	10
Implicit	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. *Marker distribution in ironic tweets (explicit or implicit) and non-ironic tweets (NIR) in French (F), English (A) and Italian (I), expressed as a percentage. The markers with an asterisk * are those not covered by existing literature. The most frequent cues in each category are shown in bold*

Table 5.4 indicates the percentage of tweets containing markers in the ironic category (making a distinction between explicit/implicit) and the non-ironic category (NIR, in gray).

In French, the intensifier, punctuation and interjection markers were most common in ironic tweets, while quotations were most frequent in non-ironic tweets.

In English, the discursive connectors, quotation, comparison words and reporting verbs were twice as common in ironic tweets than in non-ironic tweets; the reverse is true for personal pronouns. Note that the English corpus does not contain any ironic tweets including URLs, as all tweets of this type were annotated as "no decision": the annotators were unable to understand the tweet and the content of the web page linked to the URL.

In Italian, most markers were more common in ironic tweets, although some, such as quotations and URLs, occurred more frequently in non-ironic tweets.

	Negation			Discursive connectors			# Humor-* ous			Intensifier			Punctua-tion			False* assertion		
	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I
Analogy	46	56	2	6	29	8	6	15	-	21	10	0	49	24	2	13	8	-
Register shift	40	**100**	3	0	11	3	0	0	-	0	0	1	**60**	**44**	1	0	0	-
Euphemism	**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**	-
Rhetorical question	43	**70**	2	2	36	3	2	17	-	17	9	0	**93**	**86**	1	9	3	-
Oxymoron/Paradox	35	59	3	4	**43**	6	0	14	-	21	10	1	49	26	2	11	0	-
False assertion	18	57	1	4	25	3	3	7	-	10	16	0	29	14	2	**95**	**89**	-
Other	26	62	2	5	31	3	5	18	-	15	11	0	45	20	2	11	3	-

	Modality			Quotation			Opposition			Personal* pronoun			Interjection			Comparison*		
	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I	F	A	I
Analogy	0	3	2	0	24	1	6	11	2	38	19	2	6	0	3	**43**	**42**	3
Register shift	0	11	0	0	**44**	0	0	11	1	**40**	**33**	1	20	0	2	20	6	0
Euphemism	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
Rhetorical question	0	3	0	7	23	1	3	15	1	31	27	0	13	2	1	2	5	0
Oxymoron/Paradox	0	2	1	5	20	0	**12**	**19**	1	32	21	0	15	3	2	2	6	0
False assertion	0	0	0	4	16	1	3	4	1	31	**36**	1	13	0	1	2	13	1
Other	0	2	1	8	25	1	2	11	2	29	22	0	10	0	2	1	10	0

	Named* entities			Reporting verb			Opinion			URL*		
	F	A	I	F	A	I	F	A	I	F	A	I
Analogy	100	100	17	2	16	0	41	68	-	13	0	1
Register shift	80	100	8	0	22	0	60	68	-	0	0	1
Euphemism	94	100	2	0	**33**	0	56	67	-	22	0	1
Hyperbole	88	88	6	3	13	0	**84**	**88**	-	21	0	1
Rhetorical question	90	97	9	1	17	0	45	73	-	25	0	1
Oxymoron/Paradox	99	100	10	1	19	0	55	75	-	11	0	2
False assertion	90	93	8	1	13	0	45	79	-	25	0	0
Other	91	98	6	1	16	0	32	74	-	30	0	1

Table 5.5. *Distribution of tweets containing markers by irony category, expressed as percentages for French (F), English (A) and Italian (I)*

Table 5.5 indicates the percentage of tweets containing markers in each irony category.

Negation is most common in the *euphemism* category for French, and in the register shift, euphemism and rhetorical question categories in English.

Intensifiers were most common in the *euphemism* and hyperbole categories in both French and English.

Punctuation occurred most frequently in the *register shift, euphemism* and *rhetorical question* categories in French, and in the *register shift and rhetorical question* categories in English.

False propositions were most common in the *hyperbole* and *false assertion* categories in both French and English. *Opposition words* occurred most often in the *oxymoron/paradox* category in French and English.

Personal pronouns were most frequent in the *register shift* category for French, and in the *register shift, hyperbole* and *false assertion* categories for English.

Comparison words occurred most frequently in the *analogy* category for French and English.

Opinion words were most common in the *hyperbole* category for both French and English.

For Italian, the percentage of tweets containing markers in each irony category was very low, and tweets were distributed almost equally across different categories (for example, negation is found in all categories at between 1% and 3%, and the frequency of opposition words does not exceed 2%).

This set of quantitative studies show that, whatever the language, the authors of ironic tweets tend to use markers such as opinion words, named entities and negation words in categories including analogy, rhetorical question, oxymoron/paradox, false assertion and other.

5.2.2.5. *Quantitative results of the annotation procedure for contradiction relationships*

Figure 5.3 shows that the distribution of relationships is almost identical in the French and English corpora. For example, the opposition relationship is most common in ironic tweets in French (69%) and in English (80%). Furthermore, the number of comparison relationships is smaller in both English (20%) and French (14%), while cause/consequence relationships are the least frequent in the two languages (6% for English, and 11% for French). Marker annotation was carried out automatically for the Italian corpus, meaning that contradiction relationships were not noted, given that this must be done manually.

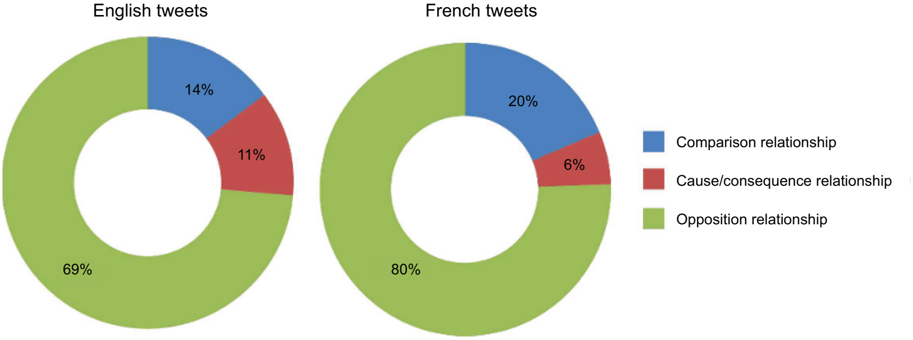


Figure 5.3. *Distribution by percentage of relationships in ironic tweets with explicit contradiction in French and English*

5.2.2.6. *Correlation between different levels of the annotation scheme*

We also carried out a comparative study of correlations between different levels of the scheme in all three languages: French, English and Italian. The approach previously used in French (see Chapter 3) was applied to the English and Italian results.

We looked at the correlation between irony markers and trigger types (explicit or implicit contradiction) and between irony markers and category. The aim of this study was to analyze the extent to which markers may be used to predict irony in the languages in question.

Applying Cramer's V (Cohen 1988) to the number of occurrences of each marker, we obtained the following results (all of the correlations are statistically significant):

– between markers and the ironic/non-ironic class:

- a strong correlation for French ($V = 0.156, df = 14$) and Italian ($V = 0.31, df = 6$);

- a medium to strong correlation for English ($V = 0.132, df = 9$);

– between markers and irony trigger type (explicit or implicit contradiction):

- a strong correlation for French ($V = 0.196, df = 16$);

- a medium to strong correlation for Italian ($V = 0.138, df = 5$);

- a medium correlation for English ($V = 0.083, df = 12$).

We also analyzed correlations by marker $(df = 1)$. The markers with the strongest class correlation (ironic/non-ironic) were:

– negations, interjections, named entities and URLs for French ($0.14 < V < 0.41$);

– negations, discursive connectors and personal pronouns for English ($0.12 < V < 0.17$);

– quotations, named entities and URLs for Italian ($0.310 < V < 0.416$).

The markers most strongly correlated with triggers (explicit/implicit) are:

– opposition markers, comparison words and false assertions for French ($0.140 < V < 0.190$);

– opposition markers and discursive connectors for English ($0.110 < V < 0.120$);

– discursive connectors, punctuation and named entities for Italian ($0.136 < V < 0.213$).

We noted that, in spite of the high frequency of opinion words in ironic tweets in both French and English, the opinion marker is not correlated with the ironic/non-ironic classification or with explicit/implicit triggers ($V < 0.06$), as many non-ironic tweets also include opinion words.

Finally, we analyzed the correlation between markers and irony categories. According to the results of Cramer's V test, the most decisive markers were:

– intensifiers, punctuation, false assertion and opinion word for French, with strong correlation;

– negation, discursive connectors and personal pronouns for English, with medium correlation;

– punctuation, interjections and named entities for Italian, with medium correlation.

5.2.3. *Summary*

These results are encouraging, as they show that our annotation scheme, defined for French, can be applied to other Indo-European languages (English and Italian). The pragmatic phenomena that we identified as being specific to ironic contexts in French are also present when irony is expressed in other languages belonging to the same family. The same trends are present in terms of irony categories and markers, in correlations between markers and the ironic/non-ironic class, and in correlations between markers and trigger type (explicit/implicit).

The results of this portability evaluation for our annotation scheme indicate that, in future, it may be possible to develop an automatic irony detection system in a multilingual context.

We made the first step in this direction by evaluating our automatic detection model for text in Arabic. This investigation is presented below.

5.3. Irony in Semitic languages

> “The Semitic languages are a group of languages spoken from ancient times in the Near East, North Africa and the Horn of Africa. The term ’Semitic’ was coined in 1781, derived from Shem, one of the sons of Noah in the Old Testament. They form a branch of the Afroasiatic language family, present across the northern half of Africa and into the Middle East. The origin and direction of the geographic expansion of these languages is uncertain; they may have expanded from Asia into Africa, or from Africa into Asia. The most widely spoken Semitic languages today are Arabic (approx. 375 million speakers), Amharic (over 90 million), Hebrew (8 million), Tigrinya (6.75 million) and Maltese (400,000 speakers). Other Semitic languages are used in Ethiopia, Eritrea, Djibouti and Somalia, and in the Near East (e.g. neo-Aramaic languages). Arabic is notable for the distinction made between literary Arabic, the lingua franca generally found in writing, and spoken dialects. Literary Arabic includes both Classical Arabic and Modern Standard Arabic. There are many regional variations in spoken (dialectal) Arabic, and not all are mutually comprehensible”. (Adapted from French Wikipedia[7])

[7] https://fr.wikipedia.org/wiki/Langues_s%C3%A9mitiques.

In the context of our study, we have chosen to focus on Arabic. According to the previous definition, we note that linguists have emphasized the great difference between literal Arabic and dialectal Arabic. A considerable volume of work has been carried out in the field of natural language processing for Arabic, mostly for literary Arabic, and more specifically for Modern Standard Arabic (MSA).

MSA is a modernized and standardized derivative of Classical Arabic[8] used in writing and in formal speech in the domains of education, newspapers, and to some extent, TV shows. MSA has a complex linguistic structure with a rich morphology and complex syntax (Al-Sughaiyer and Al-Kharashi 2004, Ryding 2005). Work on automatic processing for Arabic has been ongoing for over 20 years (Habash 2010), and several resources and tools have been developed to handle Arabic morphology and syntax, from superficial to deep analysis (Eskander *et al*. 2013, Pasha *et al*. 2014, Green and Manning 2010, Marton *et al*. 2013). Additionally, many applications have been developed for Arabic NLP (ANLP), including question–answer systems (Bdour and Gharaibeh 2013, Hammo *et al*. 2002, Abouenour *et al*. 2012), automatic translation (Sadat and Mohamed 2013, Carpuat *et al*. 2012), sentiment analysis (Abdul-Mageed *et al*. 2014) and named entity recognition (Darwish 2013, Oudah and Shaalan 2012). Work has also been carried out in the field of NLP for colloquial Arabic, notably concerning automatic understanding of spontaneous speech (Afify *et al*. 2006, Biadsy *et al*. 2009, Bahou *et al*. 2010), phonetization of the Tunisian dialect (Masmoudi *et al*. 2014), the construction of domain ontologies for Arabic dialects (Graja *et al*. 2011, Karoui *et al*. 2013), morphological analysis (Habash and Rambow 2006, 2007), automatic identification of colloquial Arabic (Alorifi 2008), etc. However, to the best of our knowledge, the problem of automatic irony detection for Arabic, and more specifically in the context of social media, has yet to be addressed.

In what follows, we shall present an overview of the specificities of Arabic (section 5.3.1), distinguishing between MSA and colloquial forms. In section 5.3.2, we shall present the corpus and resources used for automatic detection. Section 5.3.3 describes our experiment and results. Finally, in section 5.3.3.3, we compare these results with those obtained for French (presented in Chapter 4).

[8] Classical, Quranic, Arabic is used in literary and religious texts.

5.3.1. *Specificities of Arabic*

Arabic is written from right to left, with ambiguous letter forms that change depending on the position of the letter in a word. Letters may take different forms according to whether they are autonomous (unconnected), initial (at the start of a word), median (within a word) or final (at the end of a word) (see Table 5.6).

Autonomous	م	ش	غ
Initial	مـ	شـ	غـ
Median	ـمـ	ـشـ	ـغـ
Final	ـم	ـش	ـغ

Table 5.6. *Letters in Arabic according to their position in a word (Habash 2010)*

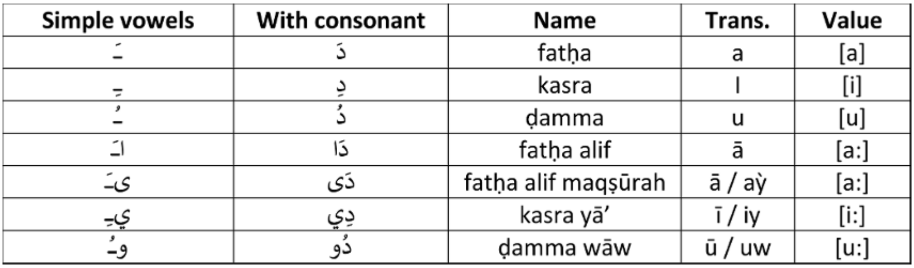
Arabic is characterized by the absence of diacritical signs (dedicated letters for short sounds), complex agglutination, and free word order structure. These characteristics make Arabic particularly difficult to process. For example, (Farghaly *et al*. 2003) estimate that the average number of ambiguities for a token in standard Arabic may be as high as 19.2, compared to an average of 2.3 in most other languages.

Arabic has 28 consonants, which may be combined with different ling and short vowels, as shown in Table 5.7[9]. Short vowels are represented by diacritical signs in the form of marks above or below letters, such as the *fathah* (a short diagonal line placed above a letter), the *kasrah* (a short diagonal line placed below a letter) and the *dammah* (a small, curl-like sign above a letter). Arabic texts may be full, partial or non-diacritized.

Short vowels are rarely marked explicitly in writing: the associated diacritics are not used in everyday written Arabic, or in general publications. However, texts without diacritics are extremely ambiguous. For example, the word علم may be diacritized in nine different ways (Maamouri *et al*. 2006): علم (science), علم (flag), علّم (he taught), etc. A non-diacritized word may have different morphological characteristics and, in certain cases, may belong to a

[9] https://fr.wikipedia.org/wiki/Diacritiques_de_l%27alphabet_arabe.

different morpho-syntactic category, particularly if it is interpreted out of context.



Simple vowels	With consonant	Name	Trans.	Value
ـَ	دَ	fatḥa	a	[a]
ـِ	دِ	kasra	I	[i]
ـُ	دُ	ḍamma	u	[u]
ـَا	دَا	fatḥa alif	ā	[a:]
ـَى	دَى	fatḥa alif maqṣūrah	ā / aỳ	[a:]
ـِي	دِي	kasra yā’	ī / iy	[i:]
ـُو	دُو	ḍamma wāw	ū / uw	[u:]

Table 5.7. *Types of Arabic diacritics (Wikipedia)*

Furthermore, in colloquial Arabic, there are additional sources of ambiguity in terms of understanding. One word may have several meanings according to the dialect being spoken; similarly, a single object may have several names, according to the country or region. For example, the word “suitcase” is written فاليجه or فاليز in the Tunisian dialect, but شنطة سفر in the Egyptian dialect. This is an issue in texts published on social networks, blogs, the review sections on online shopping sites, etc., where a variety of words may be used to signify the same thing.

In our work, we have chosen to focus on social media texts, specifically tweets, as a form of non-diacritized text combining standard and colloquial Arabic.

5.3.2. *Corpus and resources*

5.3.2.1. *Collection of the first Arabic corpus for irony*

Given the absence of an existing corpus of ironic tweets in Arabic, we followed the same procedure as for French and English. Initially, we planned to use the same categories as for the French corpus (politics, economics, health, etc.) with themes specific to the Arab world. However, it rapidly became apparent that the vast majority of ironic tweets concerned political subjects. For this reason, we only collected tweets in the politics category, using a set of five themes: هيلاري (Hillary), ترامب (Trump), السيسي (Al-Sissi: the Egyptian president), مبارك (Mubarak: former Egyptian president) and مرسي (Morsi: former Egyptian president).

To construct a corpus of ironic/non-ironic tweets, we harvested examples containing the hashtags #سهرية, #مسهرة, #تهكم and #استهزا (translations of *#irony* and *#sarcasm*). Tweet (5.1) is an example of an ironic message, while (5.2) is non-ironic.

Tweet (5.2) is written in standard Arabic. Tweet (5.1) combines standard Arabic with a single Egyptian/Tunisian dialect word, مش (not).

(5.1) فعلا واضح جداا ءن الانقلاب العسكري ضد الرئيس المنتخب مرسي كان لمصلحة مصر

مش إي ءطماع والآن صراعات علي منصب الرآسة #سخرية

(Actually, it is obvious that the military coup against President-Elect Morsi was in the best interests of Egypt, and not of ambitions, and now conflicts, regarding the presidency #irony)

(5.2) تّغير الحياة ولا تّغير النساء هكذا فهمت من دموع هيلاري كلينتون عند تلقيها لّهزيمة

ءمام ترامب

(Life changes, but women never change. That’s what I learned from Hillary Clinton’s tears when she was defeated by Trump)

After the collection stage, we removed duplicates, retweets and tweets containing images. This filtering process left us with a corpus of **3,479** tweets, of which **1,733** were ironic and **1,746** were non-ironic. The corpus included tweets in both standard and colloquial Arabic, and in the majority of cases, a combination of the two forms. Given that Twitter’s API does not distinguish between standard Arabic and colloquial Arabic or between different dialects, several dialects are present in the corpus, including Egyptian, Syrian and Saudi. Other dialects, such as Tunisian and Algerian, occur less frequently. For the purposes of our study, the hashtags #سهرية, #مسخرة, #تهكم and #استهزاء (*#irony* and *#sarcasm*) were removed.

5.3.2.2. *Linguistic resources*

Our automatic detection approach, described in the previous chapter, makes use of dedicated lexicons to identify opinion words, intensifiers, negations, emotions, etc. To study the portability of our system for Arabic, we looked for existing lexicons (for both standard and dialect forms of the language). Some of the lexicons we found performed well in the context of irony detection; others, less so. We constructed our own lexicons to replace those in the latter category.

The following linguistic resources were used:

– a lexicon of Arabic discursive connectors based on work by Iskandar (Keskes *et al*. 2014). This lexicon includes 416 connectors, such as يضاف الى ذلك (furthermore) and لذلك (thus);

– a lexicon of 4,501 named entities Keskes *et al*. (2014), to which we added the entities used for tweet collection, e.g. كلنتون (Clinton);

– a lexicon of 119 reporting verbs used by Keskes *et al*. (2014), such as قال (to say) and ءعلن (to announce);

– a lexicon of opinion words, made up of 22,239 negative opinion words and 26,777 positive words. This was obtained by combining two resources: the Arabic Emoticon Lexicon and the Arabic Hashtag Lexicon (dialectal)[10] (Saif *et al*. 2016). These lexicons were used in Task 7 of the SemEval'2016 campaign[11,12], and include entries such as الفشل (failure) and نقمة (indignation);

– a lexicon of 681 emoticons, used previously for French;

[10] Available at http://saifmohammad.com/WebPages/ArabicSA.html.

[11] http://alt.qcri.org/semeval2016/task7/.

[12] We tested other lexicons, such as the Arabic translation of Bing Liu's Lexicon and the Arabic translation of MPQA's Subjectivity Lexicon, but the results of experiments using these lexicons were inconclusive.

– a lexicon of personal pronouns and a lexicon of negation words, which we constructed manually, including terms such as ءانا (*I*), نحن (we), لم or ليس (not/no/isn’t);

– a lexicon of 25 intensifiers, translated from a set of intensifiers used for French, including كثير (lots) and جدا (very).

5.3.3. *Automatic detection of irony in Arabic tweets*

In this section, we shall present the features used in the training process and the different algorithms used, focusing on the algorithm that performed best for classification purposes. Finally, we shall present our results.

5.3.3.1. *Features used for irony detection*

For the French corpus, our system used 30 features, eight of which were obtained using morphosyntactic pre-processing tools such as MElt. In the absence of a satisfactory analyzer for texts in standard and colloquial Arabic, we manually selected a subset of 22 features, which can be extracted without these tools. These were divided into four groups, as shown in Table 5.8.

5.3.3.2. *Experiments and results*

To classify tweets as ironic/non-ironic, we used all of the features defined above, testing several classifiers using the Weka platform: SMO, Naive Bayes, multinomial logistic regression, linear regression, random tree and random forest. We trained the classifiers using a balanced corpus of 1,733 ironic tweets and 1,733 non-ironic tweets. For the first experiment, 80% of the corpus was used for training and 20% for testing. For the second experiment, we used 10-fold cross-validation. We chose to focus on these two experiments due to the limited size of the corpus (3,466 tweets). The best results were obtained using the random forest classifier with default parameters. These are shown in Table 5.9.

Three feature selection algorithms were applied using Weka with the aim of improving our results: *Chi2* and *GainRatio* (see Table 5.10) to obtain a list of features in decreasing performance order (best to worst), and *CfsSubsetEval option* to obtain the best feature combination, considering the individual predictive capacity of each feature and the degree of redundancy between features. The last algorithm indicated that the combination of *number of emoticons, exclamations, negations, number of interjections, named entities* and *number of named entities* would produce the best results.

Feature groups	**Features**	**Feature type**
Surface features	Punctuation (.../!/?)	Binary
	Emoticon	Binary
	Number of emoticons	Numerical
	Quotation (text in “”)	Binary
	Discursive connectors that do not trigger opposition	Binary
	Opposition words	Binary
	Exclamation (!!/!!!/or more)	Binary
	Question (??/???/or more)	Binary
	Exclamation + Question (?!/!?)	Binary
	Number of words	Numerical
	Interjection	Binary
Sentiment features	Number of interjections	Numerical
	Negative opinion	Binary
	Positive opinion	Binary
	Number of positive opinions	Numerical
	Number of negative opinions	Numerical
Modifier features	Intensifier	Binary
	Reporting verb	Binary
	Negation word	Binary
Contextual features	Personal pronoun	Binary
	Named entity	Binary
	Number of named entities	Numerical

Table 5.8. *Feature set used for training in Arabic*

	Train/test				**10-Fold cross-validation**			
	Precision	**Recall**	**F-measure**	**Accuracy**	**Precision**	**Recall**	**F-measure**	**Accuracy**
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 5.9. *Results of ironic/non-ironic classification obtained using random forest and all features. The best results are shown in bold*

Unfortunately, this subset of features identified by the selection algorithms failed to produce better results than those obtained using all features. We thus tested a different approach, adding features one by one in the training process for the random forest classifier (following the order given by the selection algorithm) in order to assess the influence of each feature. In this, we aimed to identify a subset of features, which would maximize performance.

Kchi^2		**GainRatio**	
0.10578219	Nb_Named_Entities	0.1091403	Named_Entities
0.08806588	Named entities	0.069659	Nb_Named_Entities
0.01807329	Nb_emoticons	0.0501787	Nb_Interjections
0.01788328	Emoticon	0.03029	Nb_Words
0.01240071	Nb_Interjections	0.0245372	Nb_emoticons
0.00526265	Nb_Words	0.0240979	Emoticons
0.00506479	Interjection	0.020977	Interjection
0.00289741	Punctuation	0.0127604	Question
0.00289665	Exclamation	0.0122677	Exclamation
0.00288611	Nb_PositiveOpinion	0.0060024	PositiveOpinion
0.00209972	Negation	0.0046953	NegativeOpinion
0.00199026	NegativeOpinion	0.0043288	Nb_PositiveOpinion
0.00197259	PositiveOpinion	0.0032733	Negation
0.00118275	Question	0.0029279	Punctuation
0.00099095	Personal_pronoun	0.0016047	Opposition
0.00032808	Opposition	0.0010676	Personal_pronoun
0.00015437	Intensifier	0.0007447	Intensifier
0.00003629	DiscursiveConnector-Opposition	0.0000376	DiscursiveConnector-Opposition
0.00000962	Quotation	0.0000307	Exclamation_Question
0.00000371	Exclamation_Question	0.0000295	Quotation
0	Nb_NegativeOpinion	0	Nb_NegativeOpinion

Table 5.10. *Results produced by the feature selection algorithms*

This approach showed that the use of all features with the exception of reporting verb produced the best results, with an accuracy value of 72.76% compared to 72.36% for all features, an F-measure of 73% instead of 72.70% for the ironic class and an F-measure of 72.50% instead of 72.10% for the non-ironic class (see Table 5.11).

5.3.3.3. *Discussion*

Although most of the features used here are surface features, the results obtained are very encouraging. Comparing the results obtained for the classification of Arabic tweets into ironic/non-ironic sets with those obtained using the same features in French, we see that the classification algorithms behave differently in the two languages. In French, the SMO classification algorithm was most effective, with an F-measure of 85.70% for the ironic class; in Arabic, the F-measure did not exceed 62.50% using this approach. The random forest classification algorithm performed better for classifying ironic tweets in Arabic, with an F-measure of 73%, but its performance in French was lower than that of the SMO algorithm, with an F-measure of 75.40%.

	Train/test				**10-Fold cross-validation**			
	Precision	**Recall**	**F-measure**	**Accuracy**	**Precision**	**Recall**	**F-measure**	**Accuracy**
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 5.11. *Results of tweet classification into ironic (IR)/ non-ironic (NIR) obtained using random forest and the best feature combination. The best results are shown in bold*

5.4. Conclusion

In this chapter, we described two experiments. In the first experiment, we studied the portability of our annotation scheme—designed for French—to a multilingual context (English and Italian). We were able to measure the impact of pragmatic phenomena in irony interpretation. The results indicated that our scheme is reliable for French, English and Italian, and the same trends can be seen in terms of irony categories and markers. We notably found correlations between markers and ironic/non-ironic classes, between markers and irony trigger types (explicit/implicit), and between markers and irony categories in all three of the languages considered. These observations are valuable in the context of developing a multilingual automatic irony detection system.

Our second experiment concerned automatic irony detection in a corpus of Arabic tweets. We trained a model using some of the surface features defined for the French corpus. The results of our experiment were encouraging, particularly given (1) the difficulty of processing texts that combine standard Arabic with dialect forms, and (2) the comparable results obtained for other languages. In our work on French, we obtained a precision of 93% for the ironic class; for Arabic, the precision was 72.4%. Other authors working on this problem have obtained precision scores of 30% for Dutch (Liebrecht *et al*. 2013) and 79% for English (Reyes *et al*. 2013).

Conclusion

Our aims in this study were twofold: (1) to propose the first system for irony detection in social media content in French, and (2) to assess the portability of this system for other languages. The field of language processing to which our work is intended to contribute is a particularly active one, notably due to the importance of irony and sarcasm detection for improving the performance of opinion analysis systems.

Our first task was to establish a full state of the art concerning linguistic and computational approaches for the detection of figurative language. While our work focused specifically on irony and sarcasm, we also described other authors' contributions in areas such as humor, satire, metaphor and comparison, as the borders between these phenomena are somewhat permeable. Based on our literature review, we made two main observations:

1) Research in the field of linguistics has approached figurative language from a semantic and pragmatic perspective, concentrating on the mechanisms involved in linguistic expressions of this type of language. These include hyperbole, rhetorical questions, false assertions, etc. Work in this area tends to focus on literary works, such as novels or poetry.

2) In computational work, irony has mostly been considered as a generic term, extended to cover sarcasm and, in some cases, satire. Studies in this area have made extensive use of social networks, such as Twitter; the presence of specific hashtags indicating the use of irony or sarcasm makes these data extremely valuable. Proposed approaches use feature-based supervised learning, using lexical, syntactic and, more rarely, pragmatic features.

Automatic Detection of Irony: Opinion Mining in Microblogs and Social Media,
First Edition. Jihen Karoui; Farah Benamara and Véronique Moriceau.

We adopted a mixed approach, combining elements of existing linguistic and computational methods; it would be difficult to treat complex phenomena such as figurative language using an automatic approach without building on a detailed study of these phenomena in a corpus setting. Our chosen approach consisted of three steps.

First, we analyzed the pragmatic phenomena used to express irony. Our main aim was to verify whether the different types of irony identified in linguistics are present in specific corpora collected from social networks such as Twitter. To do this, we proposed a multilevel annotation scheme to determine whether or not individual tweets are ironic, the type of irony involved (explicit/implicit), the category of irony used, and the linguistic cues revealing the existence of this irony (such as emoticons, punctuation and opinion words). This annotation scheme was used for a campaign covering a corpus of 2,000 tweets in French. The quantitative results, along with analysis of the correlations between different levels of the scheme, showed that in most ironic tweets, irony is triggered either by implicit contradictions involving false assertions or by explicit contradictions in the form of an oxymoron or paradox. In the case of cues, negation was seen to be a particularly common marker in both ironic and non-ironic tweets.

Next, using our observations from the annotated corpus, we developed an automatic detection system for tweets in French. Three models were proposed: (1) **SurfSystem**, a model based on surface features found in the state of the art; (2) **PragSystem**, a model using pragmatic features extracted from the linguistic content of tweets alongside new features, notably opposition patterns, which proved to be most successful with an accuracy score of 87.7%; and (3) **QuerySystem**, a query-based method applied to tweets containing false assertions with negations, which were wrongly classified by **PragSystem**. Testing showed that this final method improves classification when applied to non-personal tweets, increasing accuracy to 88.51%.

Finally, we studied the portability of both the annotation scheme and the computational models used to detect irony in a multilingual context (for Italian, English and Arabic). We tested the performance of our proposed annotation scheme for Italian and English, and tested the performance of our feature-based automatic detection model for Arabic. The results of these experiments showed our scheme to be entirely relevant for Italian and

English, languages which present the same tendencies as French. Applying a subset of features from the **PragSystem** model to a corpus of tweets in Arabic, we were also able to demonstrate the portability of these features, obtaining an accuracy value of 72.76%. Although this result is lower than that obtained for French, it is encouraging with regard to the development of irony detection approaches for Arabic tweets, combining both standard and colloquial forms of the language.

Our work opens up a number of interesting pathways for future research. The first of these relates to improving automatic polarity detection for ironic/sarcastic tweets within the context of sentiment analysis. To this end, we proposed three tweet analysis tasks as part of the DEFT@TALN 2017 evaluation campaign for opinion analysis and figurative language (Benamara *et al*. 2017), which we co-organized in collaboration with the LIMSI. In this latest edition of the challenge, we proposed three tasks: (1) classification of non-figurative tweets by polarity (objective, positive, negative or mixed); (2) identification of figurative language (irony, sarcasm or humor); (3) classification of figurative and non-figurative tweets by polarity (objective, positive, negative or mixed). For the challenge, the **FrIC** was expanded to include 7,724 tweets in French concerning news topics (politics, sports, movies, TV shows, artists, etc.) for the period from 2014 to 2016, selected on the basis of keywords (Hollande, Valls, #DSK, #FIFA, etc.) and/or specific hashtags, indicating the presence of figurative language (*#ironie, #sarcasme, #humor, #joke*). Twelve teams participated in the challenge. The best results, in terms of macro F-measures, were 0.650 for task (1), 0.783 for task (2) and 0.594 for task (3). These results clearly show that the use of figurative language makes it considerably difficcult to analyze opinions.

The second pathway for future investigation relates to ways in which our scheme may contribute to a better definition of the border between irony and sarcasm. Work has recently been carried out in this area, with Sulis *et al.* (2016) proposing a means of automatically distinguishing irony and sarcasm in tweets. It may be interesting to examine the relationship between fine-granularity pragmatic phenomena linked to irony, as proposed in this book, and the higher level distinction between irony and sarcasm.

Our third and final pathway concerns the development of an automatic irony detection system for multilingual corpora. In this context, we wish to evaluate the performance of a classifier trained using one corpus and tested using a second corpus in a different language. This would enable us to identify the best combination of features for irony detection, independent of language. Furthermore, we believe that automatic detection methods for irony/sarcasm may be improved by the use of a deep learning model based on neurone networks. Work in this area is already under way, in collaboration with the University of Turin, Italy, and the University of Valencia, Spain.

Appendix

Categories of Irony Studied in Linguistic Literature

Table A.1 shows a summary of the main categories and markers of irony studied by linguists, focusing principally on textual irony. In this section, we shall begin by presenting these different markers before focusing more closely on those used for studying irony in tweets.

A.1. Contradiction/false logic

Based on the definition of verbal irony as "expressing a contradiction between the thoughts and speech of a speaker" Niogret (2004), contradiction may be considered to be one of the main markers of irony.

Ironic utterances contain two textual segments. The first contains an affirmation, and the second contains information which contradicts the first. In other words, the speaker says the opposite of what they think, but leaves a trace in the text to show that their declaration is ironic. In this way, the reader is able to identify a text as ironic/non-ironic.

In this respect, (Attardo 2000b) supports the idea expressed by Kerbrat-Orecchioni (1976) and Muecke (1978) that irony is marked by a contradiction or contrast between that which is said and that which is expected. The idea was studied in greater detail by Didio (2007), who considered that contradictions in discourse allow the listener to understand the ironic meaning of a text based on the idea that contradiction unites two utterances, which confirm and deny the same knowledge object. The author cited the following example (Didio 2007):

Automatic Detection of Irony: Opinion Mining in Microblogs and Social Media,
First Edition. Jihen Karoui; Farah Benamara and Véronique Moriceau.
© ISTE Ltd 2019. Published by ISTE Ltd and John Wiley & Sons, Inc.

Markers	References
Contradiction/false logic	(Kerbrat-Orecchioni 1976, Muecke 1978, Tayot 1984, Attardo 2000b, Barbe 1995, Didio 2007)
Metaphor	(Grice 1970, Boyd 1979, Wilson and Sperber 1986, 1988, 1992, Kittay 1990, Kreuz and Roberts 1993, Barbe 1995, Song 1998, Ritchie 2005, Burgers 2010, Bres 2010)
Hyperbole/exaggeration	(Kreuz and Roberts 1993, 1995, Mercier-Leca 2003, Didio 2007, Burgers 2010)
Euphemism	(Muecke 1978, Fromilhague 1995, Seto 1998, Yamanashi 1998, Mercier-Leca 2003, Didio 2007, Burgers 2010)
Absurdity	(Didio 2007)
Surprise effects	(Colston and Keller 1998, Didio 2007)
Repetition	(Muecke 1978, Berntsen and Kennedy 1996, Wilson and Sperber 2004, Burgers 2010)
Rhetorical questions	(Muecke 1978, Berntsen and Kennedy 1996, Haiman 1998, Attardo 2000b, Burgers 2010)
Register shift	(Haiman 1998, Attardo 2000b, Burgers 2010)
Oxymoron	(Gibbs 1994, Song 1998, Mercier-Leca 2003)
Paradox	(Tayot 1984, Barbe 1995, Mercier-Leca 2003)
Quotation marks	(Tayot 1984, Gibbs 1994, Attardo 2001, Burgers 2010)
Emoticons	(Tayot 1984, Kreuz 1996, Burgers 2010)
Exclamations	(Tayot 1984, Wilson and Sperber 1992, Seto 1998, Attardo 2000b, 2001, Didio 2007, Burgers 2010)
Capital letters	(Haiman 1998, Burgers 2010)
Strikeout text and special characters	(Burgers 2010)

Table A.1. *Irony markers encountered in linguistic literature*

(A.1) *Mademoiselle de Kerkabon, who had never been married, although she would have greatly liked to be so, remained youthful at the age of forty-five; in character, she was good and sensitive; she enjoyed pleasure and was pious (Didio 2007).*

Didio (2007) identified two contradictions in this example, the first in the phrase “remained youthful at the age of forty-five”. Notwithstanding the variety of opinions regarding the youthfulness of the feminine sex, the authors judged that a woman could not seriously be referred to as youthful at 45. In this case, contradiction is expressed implicitly, as the reader must draw upon their own knowledge to understand the contradiction in the phrase.

Didio identified a second contradiction in the example: “she enjoyed pleasure, and was pious”. The author considered this to be an explicit contradiction, stating that “enjoyed pleasure” and “was pious” were contradictory and that a person could not do both at once.

Didio (2007) specifically considered contradiction as false logic or countersense. This can be seen in the fact that, in ironic utterances, speakers voluntarily express the opposite of what they think, or say something that they know to be false in a given context (see phrase (A.2)). Barbe (1995) referred to this category as “lies”.

(A.2) *At that time, a corsair from Salé came upon us and accosted us; our soldiers defended themselves like the Pope’s men, falling to their knees and throwing down their arms, begging the corsair for absolution in articulo mortis (Voltaire, Candide, cited in (Didio 2007)*).

In this example, false logic is expressed in the idea that the soldiers defend themselves by throwing down their arms: they cannot defend themselves without weapons.

Barbe (1995) considered that false propositions express a lie. He defined a lie as a phenomenon containing a truth–lie opposition. The liar wishes to hide the truth, and does so by imitating the characteristics of true speech while avoiding signals that might endanger his or her declarations. Irony might be said to be a type of lie, truth and lies are not opposed in irony (Barbe 1995).

The idea of assimilating the notion of irony to lies/counter-truth was also explored by Tayot (1984). In this case, the author found that ironic speakers take pleasure in leaving no physical clues for the audience; the listener or reader must explore the linguistic or extra-linguistic context of the utterance in order to detect the ironic meaning concealed beneath. The perspicacity of the audience is called into play to unmask a "counter-truth", a phenomenon used in both irony and lies.

A.2. Metaphor

According to the *Larousse* dictionary, "a metaphor is a figure of style which consists of establishing a comparison between two realities, based on an analogy created between the two referents". Unlike comparisons, metaphors do not include explicit comparison words such as *like, as if* and *similar to*.

The metaphor is not only the most common trick, but it has also garnered the greatest attention from psychologists, philosophers and literary theorists (Grice 1970, Kittay 1990, Kreuz and Roberts 1993, Barbe 1995, Ritchie 2005).

Song (1998) developed a more detailed definition of metaphor by building on the different definitions put forward by linguists. Kittay (1990) interpreted metaphor as a second order of meaning, obtained when the characteristics of an utterance and of its context indicate that the first-order meaning of the expression is not applicable or unsuitable. Wilson and Sperber (1986, 1988, 1992), Grice (1970), Kittay (1990) noted that readers will interpret an utterance as metaphorical unless they are able to find an acceptable literal meaning based on their own knowledge. For example:

(A.3) *Across the ice, the snow is sweeping; lonely the wind, the snow, the heart are playing together (Berntsen and Kennedy 1996).*

According to Barbe (1995), metaphor is at the heart of language. A term represents both literal and figurative meanings. Connecting metaphor and irony, Barbe (1995) indicated that while both phenomena require the audience to read between the lines, they differ in terms of application – notably in that metaphor is a figure of style, while irony is an attitude. This does not preclude the use of metaphor for ironic purposes. Another difference is that metaphor may be used to clarify, enlighten or explain something in order to create a type of description, whereas irony constitutes a critical commentary or evaluation, and is used to convey an attitude regarding a situation. The point which metaphor and irony have in common is that shared knowledge is required to understand the ironic or metaphorical meaning of an utterance. For example:

(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)*

Ten years after these studies were published, linguists began to focus on the strength of the connection between metaphor and irony.

Based on the findings of different studies concerning metaphor, irony and humor, (Ritchie 2005) concluded that all three phenomena may generate significant changes in cognitive environments, and that humor and irony are often subtle means of conveying basic messages in a variety of forms, serving similar purposes to metaphor and metonymy. This analysis was supported by Burgers (2010), who considered metaphor as a marker of irony. According to (Bres 2010), irony and metaphor are two of the oldest linguistic phenomena, used to stimulate reflection without exhausting it. For (Bres 2010), the relation between irony and metaphor is "like a cocktail or blended wine, proceeding from a delicate association of other drinks. If one element is missing, the cocktail loses its distinctive aromas, becomes dull or unpleasant to drink".

A.3. Hyperbole/exaggeration

Hyperbole is a figure of style which consists of expressing an idea or sentiment in an exaggerated manner. It is often used to produce a strong impression or to highlight a point. Kreuz and Roberts (1993) considered hyperbole to be a particularly common form of figurative language, but one which has been neglected by many linguists, despite its presence in over 27% of American short stories. This observation led (Kreuz and Roberts 1995) to carry out an in-depth study of hyperbole, which indicated that the presence of hyperbole in a text can, in certain cases, indicate ironic intent.

Kreuz and Roberts (1995) noted that hyperbole is a very common feature of verbal irony, and that it plays an important role in the perception of ironic declarations. The relationship between hyperbole and irony, on an intuitive level, appears to be important. Kreuz and Roberts (1995) also noted that hyperbole and irony share a significant number of discursive aims, such as humor, emphasis and clarification.

In conclusion, (Kreuz and Roberts 1995) and (Burgers 2010) noted that hyperbole may play an important role in the perception of irony, and that hyperbole is probably a reliable indicator for recognizing ironic intent. Furthermore, the presence of hyperbole increases the probability of an ironic interpretation, even if no untruthful remarks are made. In other words, even if these declarations are not contrary to reality, exaggeration may, in and of itself, suggest ironic intent (phrase (A.5)).

(A.5) We till and sell and pile our money
and the hedge is ten feet high
we dread 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 because we are doing so well (Berntsen and Kennedy 1996).

In the same context, (Mercier-Leca 2003) and (Didio 2007) defined hyperbole as an exaggeration of a proposition intended to create a stronger impression. They suggested that hyperbole is one of the most visible signals of irony, but that not all exaggerations are necessarily ironic. Didio (2007) used phrase (A.6) to illustrate the use of hyperbole for comic, or even ironic, purposes.

(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 : â 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)*
>
> *(The other day, Mme de la Villemenue, an old coquette still desirous to please, wished to essay her superannuated charms upon the philosopher [Voltaire]: presenting herself to him in her full finery and, profiting from some gallant utterance which he made, and which was accompanied by sidelong glances at her generous bosom, she declared: "How now, Monsieur de Voltaire, are you still concerned with impertinent duckies?" "Impertinent duckies?" replied the elderly gentleman with spirit, "impertinent duckies, Madame? Great hanging game-birds, perhaps!" (Mémoires de Bachaumont, 30 March 1778))*

Didio (2007) notes that this example is unusual as the apparent hyperbole is derogatory, given the word-play inherent in the term "pendards" – a criminal fit for hanging, or, in this case, the woman's pendulous breasts. The English translation carries much of the same meaning, although some of this final subtlety is lost.

Didio (2007) studied exaggeration as a phenomenon separate from hyperbole, but never provides a precise definition of the distinction between the two. The author defined exaggeration as a means of amplifying reality, giving it more importance than it actually has. Most linguists do not draw a distinction between hyperbole and exaggeration (Kreuz and Roberts 1993, 1995, Pougeoise 2001, Mercier-Leca 2003, Burgers 2010).

A.4. Euphemism

Euphemism is a figure of style that consists of attenuating the expression of facts or ideas considered unpleasant in order to "soften" the reality (Muecke 1978, Seto 1998, Burgers 2010). As such, euphemism is the opposite of hyperbole. In cases of hyperbole, speakers exaggerate the literal evaluation of a message, while in cases of euphemism, the speaker understates the reality, making it the very antithesis of exaggeration. Euphemism weakens a strong emotion or expression and may also be used ironically.

Yamanashi (1998), like (Burgers 2010) and (Seto 1998), confirms that euphemism serves to represent a thing as "less important than it really is" in order to attenuate the reality (phrase (A.7)). In these cases, ironic use cannot be correctly predicted based on the traditional definition of verbal irony, which entails understanding the opposite of that which is said.

Mercier-Leca (2003) considered euphemism as a marker for irony, under the name **litote**. The author adopted Fromilhague's (1995) definition: the speaker "appears to attenuate a truth which is, in fact, being forcibly affirmed: we say less to mean more". According to Fromilhague (1995), the litote may be expressed by negation or by restrictive assertion (an assertion accompanied by restrictive scope adverbs, e.g. "little", "not very much", "with difficulty", etc.). Examples include:

(A.7) "Not great" instead of "terrible"
"Passed on" or "fallen asleep" to signify "dead"
"Visually impaired" instead of "blind".

A.5. Absurdity

Absurdity is expressed by illogical reasoning and may be linked to a comic or tragic reaction. It highlights that which is not in harmony with a person or thing (Didio 2007) (phrase (A.8)).

(A.8) After the earthquake, which had destroyed three-quarters of the city of Lisbon, the wise men of that country could think of no means more effectual to preserve the kingdom from utter ruin than to entertain the people with an auto-da-fe, it having been decided by the University of Coimbra, that the sight of several persons being burned alive in great ceremony is an infallible secret for preventing earthquakes (Voltaire, *Candide*).

The relationship between absurdity and irony has yet to receive sufficient attention from linguists; at the time of writing, (Didio 2007) is the only linguist to have discussed the existence of this relationship.

A.6. Surprise

Surprise is an emotional state brought about by an unexpected event or revelation contrary to the perceived image of a situation. The effect is generally brief, dying away or giving way to another emotion[1].

The relationship between surprise and irony has received relatively little attention in linguistics. Didio (2007) considers surprise effects to be a marker of irony, without giving a precise definition; (Colston and Keller 1998) studied the relationship between surprise and irony from the opposite perspective, i.e. considering irony as a mechanism for expressing surprise. They provided the following definition:

> "Surprise is a common reaction when things do not turn out as expected. People can express this surprise by verbally noting the contrast between what was expected and what actually happened. Verbal hyperbole and irony are useful in expressing surprise because they concisely make use of this contrast".

A.7. Repetition

Following Wilson and Sperber (2004), who demonstrated that echoing (repetition) may provide a strong indication that a text is ironic, Burgers (2010) considered repetition or echoing to be a marker for irony. This phenomenon has also been addressed by other linguists, notably (Muecke 1978) and Berntsen and Kennedy (1996), under the name of **parody**.

According to (Burgers 2010):

– a writer may ironically repeat something that she (or a spokesperson) said earlier in the text or, in the case of spoken interaction, in the dialogue; a repetition based on co-text. In the case of a repetition based on co-text, an ironic utterance ironically repeats (part of) an earlier utterance from the same text that was not used ironically in its first usage (phrase (A.9)):

(A.9) (1) This movie was fantastic.
(2) No, really fantastic.
(3) FAN-TAS-TIC. (Burgers 2010)

[1] https://fr.wikipedia.org/wiki/Surprise.

– a writer may ironically repeat something that was not mentioned earlier in the text under discussion or in the same dialogue, but was mentioned somewhere else; a repetition based on context.

To avoid confusion, (Burgers 2010) used the label "co-textual repetition" for ironic repetition in co-text and the label "echo" for context-based ironic repetition.

In phrases (1) through (3) in phrase (A.9), the speaker repeats his or her declaration that the movie was fantastic. According to (Burgers 2010), if the speaker did not enjoy the movie, utterances (1) and (3) are considered ironic, and the repetition of the word "fantastic" in utterance (2) is an irony marker.

A.8. Rhetorical question

A rhetorical question is not a real question, and the speaker does not expect to receive an answer, as the response is already evident. A rhetorical question represents a point of view and not a question (phrase (A.10)) (Burgers 2010). Many linguists, such as (Muecke 1978) and (Barbe 1995), have considered rhetorical questions as a marker of irony without precisely defining the phenomenon.

(A.10) Could the weather be any better for a picnic? (Burgers 2010)

A.9. Register shift

A register shift is a sudden change in style. In utterances, a register shift is seen in the use of unexpected words from a different register (informal words in a formal register, or vice versa). It may also take the form of a sudden change in the subject of a phrase, or of exaggerated politeness in a situation where this is not appropriate (Burgers 2010).

Attardo (2000b) and Haiman (1998) also noted the relationship between irony and politeness, considering that ironic remarks are more polite than direct criticism. In this case, irony is the essential goal of the speaker, but the use of politeness attenuates the aggressive aspect (phrase (A.11)).

(A.11) Spoken to a friend: "You may grant me the honor of listening to another one of your fine predictions".

A.10. Oxymoron

An oxymoron is a figure of construction based on an apparent logical contradiction. It is an opposition figure. An oxymoron in an utterance takes the form of a syntactic rapprochement of two elements forming a semantic contradiction (Gibbs 1994, Song 1998, Mercier-Leca 2003).

Mercier-Leca (2003) considers that oxymorons have certain similarities with irony in that both phenomena are based on pretense: the speaker claims to oppose elements which are, in reality, compatible (phrase (A.12)).

(A.12) “I am the wound and the knife! [...] And the victim and the torturer!” “a dark lightness”, “a loud silence”.

(A.13) “The reign of Louis VII began in triumph, with a bloodbath extending the royal domain as far as the Mediterranean. Unfortunately, this promising reign was cut short”.

Mercier-Leca (2003) indicates that the irony in phrase (A.13) arises from the oxymoron opposing “triumph” and “bloodbath”.

A.11. Paradox

According to Mercier-Leca (2003), irony is based on paradox, accentuated by asyndetic syntax (featuring few logical connections): this emphasizes the only coordinating conjunction found in the phrase, in this case “but” (phrase (A.14)).

(A.14) “They thought of me for the job, but unfortunately I was suitable: they needed someone who could calculate, a dancer got the place”.

Tayot (1984) considers paradox to be a tool for sarcasm, whereas (Barbe 1995) considers irony expressed through paradox to be similar to ironic opposition (Barbe 1995).

A.12. Quotation marks

Quotation marks “are used to express reserve with regard to a term for which one does not accept responsibility”, according to the *Académie française* (French Academy dictionary) and “are used to indicate that one does not accept responsibility for the word or term being used”, according to *Le Petit Robert*. As such, we note that if terms placed in quotation marks are used to signify their opposite, then irony is present. In other terms, in certain cases, the use of quotation marks can represent a form of irony (Tayot 1984, Gibbs 1994, Attardo 2001, Burgers 2010).

Written transcription of speech follows certain typographical conventions. This may be useful in expressing ironic intonation. Quotation marks are used to convey a certain detachment with regard to a written utterance and, thus, irony (Attardo 2001).

According to (Gibbs 1994), many American speakers use quotation marks as a non-verbal gesture intended to convey irony. The use of quotation marks indicates that the speaker is imitating the discourse or state of mind of the cited individual, often with sarcastic intent.

A.13. Emoticons

An emoticon is a short, symbolic representation of an emotion, state of mind, feeling, ambiance or intensity, used in written discourse (Burgers 2010).

From the 1980s onwards, certain linguists drew attention to the fact that facial expressions may be a strong marker of irony (Tayot 1984, Kreuz 1996). Although there are suggestions that an emoticon was used as early as 1648, in the English poet Robert Herrick’s *To Fortune*[2], the general lack of these elements in formal written speech has led many to neglect the importance of facial expressions.

Tayot (1984) indicates that intonation, mimo-gestuality (e.g. the British “tongue in cheek”, or winking) may be used to indicate irony in speech.

[2] https://fr.wikipedia.org/wiki/%C3%89motic%C3%B4ne.

Kreuz (1996) explains how **facial expressions** may indicate irony: while speaking to someone, a person may communicate their attitudes regarding their declarations through a variety of physical signs. For example, a speaker may involuntarily shake their head slightly to indicate astonishment, or lick their lips as a sign of nervousness. Other movements of the head, eyes and eyelids may be made voluntarily, and there are several means of conveying an ironic intention. Winking, for example, may indicate that the speaker does not expect to be taken seriously. Nodding, or eye-rolling, may have a similar effect.

As direct representations of emotions, emoticons fit easily into certain figures of style, such as irony. A writer may feign distress at the news that someone is leaving for a short while, or show joy in response to an unfortunate event. In such cases, the emoticon adds a meaning or nuance to the phrase, which is not evident from the words alone.

Burgers (2010) applied existing findings on facial expressions to emoticons, generally found in social media content. He considered emoticons such as :-) or ;-) as potential irony markers.

A.14. Exclamation

The use of punctuation such as "!", "?" and the combination of the two, "!?" or "?!", has been considered as an irony marker (Attardo 2000b). Most work in this area has focused on the use of the exclamation mark.

In many utterances, exclamation marks are used to highlight a value opposite to that which is expressed by the words themselves. It is interesting to note that the written exclamation mark corresponds to oral exclamations in the form of rising intonation.

Many ironic utterances contain exclamation marks, such as "great work!" or "what a beautiful day!". Exclamation, in speech, may be a marker for conversational irony, while the exclamation mark may also be a marker for textual irony. Not all exclamatory propositions are necessarily ironic; as with many other markers, exclamationsor exclamation marks can indicate irony in certain cases (Tayot 1984, Wilson and Sperber 1992, Seto 1998, Attardo 2001, Didio 2007, Burgers 2010).

Attardo (2001) indicates that exclamation marks may be used to indicate irony. Didio (2007) states that, in the absence of an irony mark (such as that shown in Figure A.1, proposed by Alcanter de Brahm, never widely adopted), the exclamation mark often fulfills this role in indicating that text should be understood at a second level.

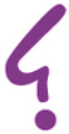


Figure A.1. *Irony mark*

A.15. Capital letters, barred text and special characters

Some linguists (Haiman 1998, Burgers 2010) have considered the use of capital letters as a potential marker for irony (phrase (A.15)). The use of barred (strikethrough) text (phrase (A.16)) and special characters (as in *Your Weather Report*^{TM} *is great*) were first considered by Burgers (2010).

(A.15) It is GREAT weather (Burgers 2010).

(A.16) It is horribly great weather (Burgers 2010).

References

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. *Actes du 11^{e} Défi Fouille de Texte*, Caen, France, 78–87. Available at: http://www.atala.org/taln_archives/DEFT/DEFT-2015/deft-2015-long-009.

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, L., Bouzoubaa, K., Rosso, P. (2012). Idraaq: new arabic question answering system based on query expansion and passage retrieval. CELCT 2012. Available at: https://pdfs.semanticscholar.org/650c/6525691136c50b312710662bb7c7b0da2bba.pdf.

Afify, M., Sarikaya, R., Kuo, H.-K.J., Besacier, L., Gao, Y. (2006). On the use of morphological analysis for dialectal arabic speech recognition. *INTERSPEECH 2006 – ICSLP*, Pittsburgh, PA.

Alorifi, F.S. (2008). Automatic identification of arabic dialects using hidden markov models. Doctorate Thesis, Université de Pittsburgh, Pittsburgh, PA.

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 (JASIST)*, 55(3), 189–213.

Angenot, M. (1982). *La parole pamphlétaire*. Payot, Paris.

Automatic Detection of Irony: Opinion Mining in Microblogs and Social Media,
First Edition. Jihen Karoui; Farah Benamara and Véronique Moriceau.
© ISTE Ltd 2019. Published by ISTE Ltd and John Wiley & Sons, Inc.

Asher, N., Benamara, F., Mathieu, Y.Y. (2009). Appraisal of opinion expressions in discourse. *Lingvisticæ Investigationes*, 32(2), 279–292.

Attardo, S. (1994). *Linguistic Theories of Humor*. Walter de Gruyter, Berlin.

Attardo, S. (2000a). Irony as relevant in appropriateness. *Journal of Pragmatics*, 32(6), 793–826.

Attardo, S. (2000b). Irony markers and functions: towards a goal-oriented theory of irony and its processing. *Rask – International journal of Language and Communication*, 12(1), 3–20.

Attardo, S. (2001). *Humorous texts: a semantic and pragmatic analysis*. Walter de Gruyter, Berlin.

Azé, J., Roche, M. (2005). Présentation de l'atelier DEFT'05. *Proceedings of TALN 2005 – Atelier DEFT'05*, Dourdan, France 2, 99–111.

Bahou, Y., Masmoudi, A., Hadrich Belguith, L. (2010). Traitement des disfluences dans le cadre de compréhension automatique de l'oral arabe spontané. *Actes de la 17*^e^ *conférence sur le traitement automatique des langues naturelles, association pour le traitement automatique des langues*, Montreal, Canada. Available at: http://www.atala.org/taln_archives/TALN/TALN-2010/taln-2010-long-021.

Baker, C.F., Fillmore, C.J., Cronin, B. (2003). The structure of the frame net database. *International Journal of Lexicography*, 16(3), 281–296.

Bamman, D., Smith, N.A. (2015). Contextualized sarcasm detection on Twitter. *Proceedings of the 9th International AAAI Conference on Web and Social Media*, Oxford, UK, 574–577.

Barbe, K. (1995). *Irony in Context*. John Benjamins Publishing, Amsterdam.

Barbieri, F., Saggion, H. (2014a). Modelling irony in Twitter. *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, Gothenburg, Sweden, 56–64.

Barbieri, F., Saggion, H. (2014b). Modelling irony in Twitter: feature analysis and evaluation. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, 4258–4264.

Barbieri, F., Saggion, H., Ronzano, F. (2014). Modelling sarcasm in Twitter, a novel approach. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Baltimore, MD, 50–58.

Basile, V., Nissim, M. (2013). Sentiment analysis on Italian tweets. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Atlanta, GA, 100–107.

Basile, V., Bolioli, A., Nissim, M., Patti, V., Rosso, P. (2014). Overview of the Evalita 2014 SENTIment POLarity Classification Task. *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, Pisa University Press, Pisa, Italy, 50–57.

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*. Cornell University, Ithaca.

Bautain, L. (1816). *De la satire*. De l'Imprimerie De C.-F. Patris, Paris.

Bdour, W.N., Gharaibeh, N.K. (2013). Development of yes/no arabic question answering system. *International Journal of Artificial Intelligence & Applications*, 4(1), 51–63.

Benamara, F. (2017). Analyse automatique d'opinions états des lieux et perspectives. *Techniques de l'ingénieur. Représentation et traitement des documents numériques*. Available at: http://www.techniques-ingenieur.fr/base-documentaire/technologies-de-l-information-th9/representation-et-traitement-des-documents-numeriques-42312210/analyse-automatique-d-opinions-h7270/.

Benamara, F., Moriceau, V., Mathieu, Y.Y. (2014). Catégorisation sémantique fine des expressions d'opinion pour la détection de consensus, *Actes du 10^{e} Défi Fouille de Textes*, Marseille, France, 36–44.

Benamara, F., Asher, N., Mathieu, Y., Popescu, V., Chardon, B. (2016). Evaluation in discourse: a corpus-based study. *Dialogue and Discourse*, 7(1), 1–49.

Benamara, F., Grouin, C., Karoui, J., Moriceau, V., Robba, I. (2017a). Analyse d'opinion et langage figuratif dans des tweets: présentation et résultats du Défi Fouille de Textes DEFT2017. *Actes du 13^e^ Défi Fouille de Textes*, Orléans, France. Available at: https://deft.limsi.fr/2017/actes_DEFT_2017.pdf.

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, L., Forner, P., Magnini, B., Pianta, E. (2004). Revising the Wordnet Domains Hierarchy: semantics, coverage and balancing. *Proceedings of the Workshop on Multilingual Linguistic Resources*, Geneva, Switzerland, 101–108.

Berntsen, D., Kennedy, J.M. (1996). Unresolved contradictions specifying attitude sinmetaphor, irony, understatement and tautology. *Poetics*, 24(1), 13–29.

Bertero, D., Fung, P. (2016). Along short-term memory framework for predicting humor in dialogues. *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA.

Bertero, D., Fung, P. (2016). Deep learning of audio and language features for humor prediction. *Journal of Lightwave Technology*, Portorož, Slovenia, 496–501.

Bestgen, Y., Cabiaux, A.-F. (2002). L'analyse sémantique latente et l'identification des métaphores. *6^e^ Rencontre des étudiants chercheurs en informatique pour le traitement automatique des langues*, Nancy, France, 331–337.

Bestgen, Y., Lories, G. (2009). Un niveau de base pour la tâche 1 (corpus français et anglais) de DEFT'09. *Actes du 5^e^ Défi Fouille de Textes*, Paris, France.

Biadsy, F., Hirschberg, J., Habash, N. (2009). Spoken arabic dialect identification using phonotactic modeing. *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Athens, Greece, 53–61.

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, R. (1979). Metaphor and theory change: what is "metaphor" a metaphor for? In *Metaphor and Thought*, Ortony, A. (ed.). Cambridge University Press, Cambridge.

Bres, J. (2010). L'ironie, un cocktail dialogique? *2e Congrès mondial de linguistique française*, New Orleans, LA.

Burfoot, C., Baldwin, C. (2009). Automatic satire detection: Are you having a laugh? *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Suntec, Singapore, 161–164.

Burgers, C. (2010). Verbal irony: use and effects in written discourse. Doctorate Thesis, Radboud Universiteit Nijmegen, Nijmegen, The Netherlands.

Buschmeier, K., Cimiano, P., Klinger, R. (2014). An impact analysis of features in a classification approach to irony detection in product reviews. *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Baltimore, MD, 42–49.

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. *ROADEF – 15e congrès annuel de la Société française de recherche opérationnelle et d'aide à la décision*. Bordeaux, France. Available at: https://hal.archives-ouvertes.fr/hal-00946481.

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, P., Sarmento, L., Silva, M.J., Oliveira, E.D. (2009). Clues for detecting irony in user-generated contents: Oh...!! it's "so easy" ;-). *Proceedings of the 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*. Hong Kong, China, 53–56.

Chardon, B., Benamara, F., Mathieu, Y.Y., Popescu, V., Asher, N. (2013). Measuring the effect of discourse structure on sentiment analysis. *14th International Conference on Intelligent Text Processing and Computational Linguistics*, Samos, Greece, 25–37.

Charniak, E., Johnson, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, MI, 173–180.

Choi, Y., Cardie, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, HI, 793–801.

Clark, H.H., Gerrig, R.J. (1984). On the pretense theory of irony. *Journal of Experimental Psychology: General*, 113(1), 121–126.

Clift, R. (1999). Irony in conversation. *Language in Society*, 28, 523–553.

Cohen, J. (1988). *Statistical Power Analysis for the Behavior Science*. Lawrence Erlbaum Associates, Mahwah, NJ.

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, K. (2013). Named entity recognition using cross-lingual resources: arabic as an example. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, 1558–1567.

Davidov, D., Rappoport, A. (2006). Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sydney, Australia, 297–304.

Davidov, D., Tsur, O., Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. *Proceedings of the 14^{th} Conference on Computational Natural Language Learning*, Uppsala, Sweden, 107–116.

Davidson, R.J., Scherer, K.R., Goldsmith, H.H. (2009). In *Handbook of Affective Sciences*, Oxford University Press, Oxford.

Didio, L. (2007). Une approche émantico-sémiotique de l'ironie. Doctorate Thesis, Université de Limoges, Limoges, France.

Do Dinh, E.-L., Gurevych, I. (2016). Token-level metaphor detection using neural networks. *Proceedings of the 4th Workshop on Metaphor in NLP*, San Diego, CA, 28–33. Available at: http://www.aclweb.org/ anthology/W16-1104

Eskander, R., Habash, N., Rambow, O. (2013). Automatic extraction of morphological lexicons from morphologically annotated corpora. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Seattle, WA, 1032–1043.

Farghaly, A., Senellart, J. *et al*. (2003). Intuitive coding of the arabic lexicon. *SYSTRAN, MT, Summit IX Workshop, Machine Translation for Semitic Languages: Issues and Approaches*, New Orleans, USA.

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*, Duisburg, Germany, 694–698.

Filatova, E. (2012). Irony and sarcasm: corpus generation and analysis using crowdsourcing. In *LREC*, Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds). European Language Resources Association, Istanbul, Turkey, 392–398.

Fromilhague, C. (1995). *Les figures de style*. Armand Colin, Paris, France.

Gedigian, M., Bryant, J., Narayanan, S., Ciric, B. (2006). Catching metaphors. *Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, New York, NY, 41–48.

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. *Proceedings of Sem Eval 2015*, Co-located with NAACL, ACL, Beijing, China, 470–478.

Gianti, A., Bosco, C., Patti, V., Bolioli, A., Caro, L.D. (2012). Annotating irony in a novel italian corpus for sentiment analysis. *Proceedings of the 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals*, Istanbul, Turkey.

Gibbs, R.W. (1994). *The Poetics of Mind: Figurative Thought, Language, and Understanding*. Cambridge University Press, Cambridge.

Gibbs, R.W. (2000). Irony in talk among friends. *Metaphor and Symbol*, 15(1–2), 5–27.

Gonzalez-Ibanez, R., Muresan, S., Wacholde, N. (2011). Identifying sarcasm in Twitter: a closer look. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-2*, Portland, OR, 581–586.

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 *Advances in Cross-Cultural Decision Making*, Schatz, S., Hoffman, M. (eds). Springer, 17–27.

Graja, M., Jaoua, M., Belguith, L.H. (2011). Building ontologies to understand spoken tunisian dialect. *CoRR*. Available at: http://arxiv.org/abs/1109.0624.

Green, S., Manning, C.D. (2010). Better arabic parsing: baselines, evaluations, and analysis. *Proceedings of the 23rd International Conference on Computational Linguistics*. Beijing, China, 394–402.

Grice, H.P. (1970). *Logic and Conversation*. Harvard University, Cambridge, MA.

Grice, H.P., Cole, P., Morgan, J.L. (1975). Syntax and semantics. In *Logic and Conversation*, Grice, H.P. (ed.)., Harvard University, Cambridge, MA, 3, 41–58.

Habash, N., Rambow, O. (2006). Magead: a morphological analyzer and generator for the arabic dialects. *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 681–688.

Habash, N., Rambow, O. (2007). Morphophonemic and orthographic rules in a multi-dialectal morphological analyzer and generator for arabic verbs. *International Symposium on Computer and Arabic Language*, Riyadh, Saudi Arabia.

Habash, N.Y. (2010). *Introduction to Arabic natural language processing*. Morgan & Claypool, San Rafael, CA.

Haiman, J. (1998). *Talk is cheap: sarcasm, alienation, and the evolution of language*. Oxford University Press, Oxford.

Haiman, J. (2001). *Talk is cheap: sarcasm, alienation, and the evolution of language*. Oxford University Press, Oxford.

Hammo, B., Abu-Salem, H., Lytinen, S. (2002). Qarab: a question answering system to support the Arabic language. *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*, Philadelphia, PA, 1–11.

Hatzivassiloglou, V., McKeown, K.R. (1997). Predicting the semantic orientation of adjectives. *Proceedings of the 8th Conference on European Chapter of the Association for Computational Linguistics*, Madrid, Spain, 174–181.

Hee, C.V., Lefever, E., Hoste, V. (2016). Exploring the realization of irony in Twitter data. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. In Chair, N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds). European Language Resources Association, Paris, France.

Hertzler, J.O. (1970). *Laughter: a socio-scientific analysis*. Exposition Press, New York, USA.

Huang, T.-H.K. (2014). Social metaphor detection via topical analysis. *6th International Joint Conference on Natural Language Processing*. Nagoya, Japan.

Hunston, S., Thompson, G. (eds). (2000). *Evaluation in Text: Authorial Distance and the Construction of Discourse*. Oxford University Press, Oxford.

Hyungsuk, J., Ploux, S., Wehrli, E. (2003). Lexical knowledge representation with contexonyms. *9th MT Summit Machine Translation*, New Orleans, LA, 194–201.

Jang, H., Moon, S., Jo, Y., Rosé, C.P. (2015). Metaphor detection in discourse. *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Prague, Czech Republic.

Jang, H., Wen, M., Rosé, C.P. (2015). Effects of situational factors on metaphor detection in an online discussion forum. *Proceedings of the 3rd Workshop on Metaphor in NLP*, Denver, CO, 1–10.

Jie Tang, Y., Chen, H.-H. (2014). Chinese irony corpus construction and ironic structure analysis. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland, 1269–1278.

Joshi, A., Sharma, V., Bhattacharyya, P. (2015). Harnessing context incongruity for sarcasm detection. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, 757–762.

Joshi, A., Tripathi, V., Patel, K., Bhattacharyya, P., Carman, M. (2016). Are word embedding-based features useful for sarcasm detection? *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, TX, 1006–1011.

Karoui, J. (2016). Fric: un corpus et un schéma d'annotation multiniveau pour l'ironie dans les tweets. *Atelier Communautés: outils et applications en TAL*, dans le cadre de la conférence *JEP-TALN-RECITAL 2016*, Paris, France.

Karoui, J., Graja, M., Boudabous, M., Belguith, L.H. (2013). Domain ontology construction from a tunisian spoken dialogue corpus. *Proc. ICWIT*, Hammamet, Tunisia.

Karoui, J., Benamara, F., Moriceau, V., Aussenac-Gilles, N., Belguith, L.H. (2015a). Towards a contextual pragmatic model to detect irony in tweets. *Proceedings of ACL-IJCNLP 2015, Volume 2: Short Papers*. The Association for Computer Linguistics, Beijing, China, 644–650. Available at: http://aclweb.org/ anthology/P/P15/P15-2106.pdf.

Karoui, J., Benamara Zitoune, F., Moriceau, V., Aussenac-Gilles, N., Hadrich Belguith, L. (2015b). Détection automatique de l'ironie dans les tweets en français. *Actes de la 22^{e} conférence sur le traitement automatique des langues naturelles*, Caen, France, 460–465. Available at: http://www.atala.org/taln_archives/TALN/TALN-2015/taln-2015-court-022.

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. *Proceedings of the 15th edition of the European Chapter of the Association for Computational Linguistics Conference*, Valencia, Spain.

Kerbrat-Orecchioni, C. (1976). Problèmes de l'ironie. *Linguistique et sémiologie*, 2, 10–46.

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, W. (2000). Metaphor comprehension: a computational theory. *Psychonomic Bulletin & Review*, 7(2), 257–266.

Kittay, E.F. (1990). *Metaphor: Its Cognitive Force and Linguistic Structure*, Oxford University Press, Oxford.

Kreuz, R.J. (1996). The use of verbal irony: cues and constraints. *Metaphor: Implications and Applications*, Part 1, Chapter 2, 23–38.

Kreuz, R.J., Caucci, G.M. (2007). Lexical influences on the perception of sarcasm. *Proceedings of the Workshop on Computational Approaches to Figurative Language*. New York, NY, 1–4.

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, R.J., Roberts, R.M. (1993). The empirical study of figurative language in literature. *Poetics*, 22(1), 151–169.

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, 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, G.N. (2016). *Principles of Pragmatics*. Routledge, London.

Létourneau, D., Bélanger, M. (2009). Impacts de la variation du nombre de traits discriminants sur la catégorisation des documents. *Actes du 5^e Défi Fouille de Textes*, Paris, France.

Liebrecht, C., Kunneman, F., Van Den, B.A. (2013). The perfect solution for detecting sarcasm in tweets# not. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis,* Atlanta, GA29–37.

Littman, T., Turney, P. (2002). Unsupervised learning of semantic orientation from a hundred-billion-word corpus, Technical Report ERB-1094, National Research Council Canada, Canada.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 1–167.

Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, Cambridge.

Losada, D.E., Crestani, F. (2016). A test collection for research on depression and language use. *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, 28–39. Available at: http://dx.doi.org/10.1007/978-3-319-44564-9_3.

Lucariello, J. (1994). Situational irony: a concept of events gone awry. *Journal of Experimental Psychology: General*, 123(2), 129.

Maamouri, M., Bies, A., Kulick, S. (2006). Diacritization: a challenge to Arabic tree bank annotation and parsing. *Proceedings of the Conference of the Machine Translation SIG of the British Computer Society*, London.

Macwhinney, B., Fromm, D. (2014). Two approaches to metaphor detection. *Proceedings of the 9th International Conference on Language Resources and Evaluation*. In Chair, N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijkand, J., Piperidis, S. (eds). European Language Resources Association, Reykjavik, Iceland.

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, A., Khemakhem, M.E., Estève, Y., Bougares, F., Dabbar, S., Belguith, L.H. (2014). Phonétisation automatique du dialecte tunisien. *30^{e} Journée d'études sur la parole*, Le Mans, France.

Maynard, D., Greenwood, M.A. (2014). Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. *LREC*, 4238–4243.

Mercier-Leca, F. (2003). *L'ironie*. Hachette, Paris.

Mihalcea, R., Strapparava, C. (2005). Making computers laugh : investigations in automatic humor recognition. *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Vancouver, Canada, 531–538.

Mihalcea, R., Strapparava, C. (2006). Learning to laugh (automatically): computational models for humor recognition. *Computational Intelligence*, 22(2), 126–142.

Miller, G.A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11), 39–41.

Mpouli, S., Ganascia, J.-G. (2015). Extraction et analyse automatique des comparaisons et des pseudo-comparaisons pour la détection des comparaisons figuratives. *22^{e} Conférence sur le traitement automatique des langues naturelles*, Caen, France.

Muecke, D.C. (1978). Irony markers. *Poetics*, 7(4), 363–375.

Nadaud, B., Zagaroli, K. (2008). *Surmonter ses complexes: les comprendre pour les assumer*. Eyrolles, Paris, France.

Niogret, P. (2004). *Les figures de l'ironie dans À la recherche du temps perdu de Marcel Proust*. L'Harmattan, Paris, France.

Oliveira, I., Ploux, S. (2009). Vers une méthode de détection et de traitement automatique de la métaphore. *Passeurs de mots, passeurs d'espoir. Actes des 8^{e} Journées scientifiques du Réseau LTT*, Lisbon, Portugal, 1–11.

Oudah, M., Shaalan, K.F. (2012). A pipeline Arabic named entity recognition using a hybrid approach. *Proceedings of COLING 2012: Technical Papers*, Mumbai, India, 2159–2176.

Ounis, I., Macdonald, C., Soboroff, I. (2008). Overview of the trec-2008 blog track. Technical Report, Glasgow.

Pang, B., Lee, L. (2004). A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Barcelona, Spain.

Pang, B., Lee, L., Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, 10, 79–86.

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. *LREC*, 14, 1094–1101.

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. *TALN 2009 Conférence sur le traitement automatique des langues naturelles*, Paris, France.

Polanyi, L., Zaenen, A. (2006). Contextual valence shifters. *Computing Attitude and Affect in Text: Theory and Applications*. Springer-Verlag, Berlin, 1–10.

Politis, H. (2002). *Kierkegaard*. Ellipses, New York, USA.

Pougeoise, M. (2001). *Dictionnaire de rhétorique*. Armand Colin, Paris, France.

Purandare, A., Litman, D. (2006). Humor: prosody analysis and automatic recognition for friends. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, 208–215.

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. *Computation and Language*, 1, 475–479.

Raeber, T. (2011). L'ironie: réactualisation de pensée et contenus non posés: une approche pragmatique. Master's Thesis, Université de Neuchâtel, Neuchâtel, Switzerland.

Raz, Y. (2012). Automatic humor classification on Twitter. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, Montreal, Canada, 66–70.

Reboul, O. (1991). *Introduction à la rhétorique*. *Théorie et pratique*. Presses universitaires de France, Paris.

Reyes, A., Rosso, P. (2011). Mining subjective knowledge from customer reviews: a specific case of irony detection. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Portland, OR, 118–124.

Reyes, A., Rosso, P. (2012). Making objective decisions from subjective data: detecting irony in customer reviews. *Decision Support Systems*, 53(4), 754–760.

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, 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, A., Rosso, P., Veale, T. (2013). A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, 47(1), 239–268.

Rick, S., Loewenstein, G. (2008). The role of emotion in economic behavior. In *Handbook of Emotions*, Lewis, M., Haviland-Jones, J.M., Barrett, L.F. (eds). The Guilford Press, New York, 138–156.

Riffaterre, M. (1969). La métaphore filée dans la poésie surréaliste. *Langue française*, 3(1), 46–60.

Riloff, E., Qadir, A., Surve, P., Silva, L.D., Gilbert, N., Huang, R. (2013). Sarcasm as contrast between a positive sentiment and negative situation. *EMNLP*, Sofia, Bulgaria, 704–714.

Ritchie, D. (2005). Frame-shifting in humor and irony. *Metaphor and Symbol*, 20(4), 275–294.

Rouvier, M., Favre, B., Andiyakkal Rajendran, B. (2015). Talep deft'15: le plus cooool des systèmes d'analyse de sentiment. *Actes du 11^{e} Défi Fouille de Texte*. Association pour le traitement automatique des langues, Caen, France, 97–103. Available at: http://www.atala.org/taln_archives/DEFT/DEFT-2015/deft-2015-long-011.

Roze, C., Danlos, L., Muller, P. (2012). Lexconn: a French lexicon of discourse connectives. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*. Presses universitaires de Caen, France.

Ryding, K.C. (2005). *A reference grammar of modern standard Arabic*. Cambridge University Press, Cambridge.

Sadat, F., Mohamed, E. (2013). Improved Arabic-French machine translation through preprocessing schemes and language analysis. In *Canadian Conference on Artificial Intelligence*, Springer-Verlag, Berlin, 308–314.

Saif, M., Mohammad, S., Svetlana, K. (2016). Sentiment lexicons for Arabic social media. *Proceedings of the 10th International Conference on Language Resources and Evaluation LREC 2016*. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds). Portorož, Slovenia. Available at: http://www.lrec-conf.org/proceedings/lrec2016/summaries/234.html.

Searle, J. (1979). *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press, Cambridge.

Seto, K.-I. (1998). On non-echoic irony. *Relevance Theory: Applications and Implications*, 37, 239.

Shaikh, M.A., Prendinger, H., Mitsuru, I. (2007). Assessing sentiment of text by semantic dependency and contextual valence analysis. *Proceedings of the International Conference on Affective Computing and Intelligent Interaction,* Springer-Verlag, Berlin, Germany, 191–202.

Shelley, C. (2001). The bicoherence theory of situational irony. *Cognitive Science*, 25(5), 775–818.

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, V. (2012). *L'humour et l'ironie en littérature francophone subsaharienne: des enjeux critiques à une poétique du rire*. Peter Lang, Berlin, Germany.

Sjöbergh, J., Araki, K. (2007). Recognizing humor without recognizing meaning. *International Workshop on Fuzzy Logic and Applications*, Springer-Verlag, Berlin, Germany, 469–476.

Song, N.S. (1998). Metaphor and metonymy. *Relevance Theory: Applications and Implications*, John Benjamins Publishing, Amsterdam, The Netherlands, 37, 87–104.

Sperber, D., Wilson, D. (1981). Irony and the use-mention distinction. *Radical pragmatics*, 49, 295–318.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France, 102–107.

Su, C., Huang, S., Chen, Y. (2017). Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, 219, 300–311.

Sulis, E., Hernández Farías, D.I., Rosso, P., Patti, V., Ruffo, G. (2016). Figurative messages and affect in Twitter: differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*. Available at: https://doi.org/10.1016/j.knosys.2016.05.035.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37, 267–307.

Taylor, J.M. (2009). Computational detection of humor: a dream or a nightmare? The ontological semantics approach. *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Washington, DC, 429–432.

Tayot, C. (1984). L'ironie. Doctorate Thesis, Université Claude Bernard, Lyon, France.

Toprak, C., Gurevych, I. (2009). Document level subjectivity classification experiments in DEFT 2009 challenge. *Actes du 5^{e} Défi Fouille de Textes*, Senlis, France.

Tsur, O., Davidov, D., Rappoport, A. (2010). ICWSM a great catchy name: semi-supervised recognition of sarcastic sentences in online product reviews. *ICWSM*, George Washington University, Washington.

Tsvetkov, Y., Boytsov, L., Gershman, A., Nyberg, E., Dyer, C. (2014). Metaphor detection with cross-lingual model transfer. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, MD, 248–258.

Turney, P.D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Meeting of the Association for Computational Linguistics*, Philadelphia, PA, 417–424.

Utsumi, A. (1996). A unified theory of irony and its computational formalization. *Proceedings of the 16th Conference on Computational Linguistics*, Copenhague, Denmark, 962–967.

Utsumi, A. (2000). Verbal irony as implicit display of ironic environment: distinguishing ironic utterances from non irony. *Journal of Pragmatics*, 32(12), 1777–1806.

Utsumi, A. (2004). Stylistic and contextual effects in irony processing. *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*. Lawrence Erlbaum Associates, Mahwah, New Jersey, 1369–1374.

Van de Gejuchte, I. (1993). L'humour comme discours. *Revue de l'Institut de sociologie*, 12(1-4), 399–411.

Van Hee, C., Lefever, E., Hoste, V. (2015). Guidelines for annotating irony in social media text. Technical Report, Department of Translation, Interpreting and Communication, Ghent University, Belgium.

Van Hee, C., Lefever, E., Hoste, V. (2016). Exploring the realization of irony in Twitter data. *Proceedings of the 10th International Conference on Language Resources and Evaluation*, Portorož, Slovenia, 1795–1799.

Veale, T., Hao, Y. (2010). Detecting ironic intent in creative comparisons. *ECAI*, 215, 765–770.

Voas, D. (2014). Towards a sociology of attitudes. *Sociological Research Online*, 19(1), 12.

Wallace, B.C. (2015). Computational irony: a survey and new perspectives. *Artificial Intelleligence Review*, 43(4), 467–483.

Wallace, B.C., Choe, D.K., Kertz, L., Charniak, E. (2014). Humans require context to infer ironic intent (so computers probably do, too). *ACL,* (2), 512–516.

Wallace, B.C., Choe, D.K., Charniak, E. (2015). Sparse, contextually informed models for irony detection: exploiting user communities, entities and sentiment. *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*, Beijing, China, 1035–1044.

Wen, M., Zheng, Z., Jang, H., Xiang, G., Rosé, C.P. (2013). Extracting events with informal temporal references in personal histories in online communities. *ACL*, (2), 836–842.

Whissell, C. (1989). The dictionary of affect in language. *Emotion: Theory, Research, and Experience*, 4(113-131), 94.

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, Y. (1978). Making preferences more active. *Artificial Intelligence*, 11(3), 197–223.

Wilson, D., Sperber, D. (1986). *Relevance: Communication and Cognition*, Harvard University Press, Cambridge.

Wilson, D., Sperber, D. (1988). Representation and relevance. *Mental Representations: The Interface Between Language and Reality*, Cambridge University Press, Cambridge, 133–153.

Wilson, D., Sperber, D. (1992). On verbal irony. *Lingua*, 87(1), 53–76.

Wilson, D., Sperber, D. (2004). Relevance theory. In *Handbook of Pragmatics*. Blackwell Publishing, Oxford, 607–632.

Yamanashi, M.-A. (1998). Some issues in the treatment of irony and related tropes. *Relevance Theory: Applications and Implications*, 37, 271.

Yang, D., Lavie, A., Dyer, C., Hovy, E. (2015). Humor recognition and humor anchor extraction. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2367–2376.

Zagibalov, T., Belyatskaya, K., Carroll, J. (2010). Comparable english-russian book review corpora for sentiment analysis. *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, Lisbon, Portugal, 67–72.

Index

Automatic Detection of Irony: Opinion Mining in Microblogs and Social Media,
First Edition. Jihen Karoui; Farah Benamara and Véronique Moriceau.

Other titles from

ISTE

in

Cognitive Science and Knowledge Management

2018

BONFANTE Guillaume, GUILLAUME Bruno, PERRIER Guy
Application of Graph Rewriting to Natural Language Processing
(Logic, Linguistics and Computer Science Set – Volume 1)

CLAVEL Chloé
Opinion Analysis in Interactions: From Data Mining to Human-Agent Interaction

MARTINOT Claire, BOŠNJAK BOTICA Tomislava, GEROLIMICH Sonia, PAPROCKA-PIOTROWSKA Urszula
Reformulation and Acquisition of Linguistic Complexity
(Interaction of Syntax and Semantics in Discourse Set – Volume 2)

PENNEC Blandine
Discourse Readjustment(s) in Contemporary English
(Interaction of Syntax and Semantics in Discourse Set – Volume 1)

2017

KURDI Mohamed Zakaria
Natural Language Processing and Computational Linguistics 2: Semantics, Discourse and Applications

MAESSCHALCK Marc
Reflexive Governance for Research and Innovative Knowledge
(Responsible Research and Innovation Set – Volume 6)

PELLÉ Sophie
Business, Innovation and Responsibility
(Responsible Research and Innovation Set - Volume 7)

2016

BOUVARD Patricia, SUZANNE Hervé
Collective Intelligence Development in Business

CLERC Maureen, BOUGRAIN Laurent, LOTTE Fabien
Brain–Computer Interfaces 1: Foundations and Methods
Brain–Computer Interfaces 2: Technology and Applications

FORT Karën
Collaborative Annotation for Reliable Natural Language Processing

GIANNI Robert
Responsibility and Freedom
(Responsible Research and Innovation Set – Volume 2)

GRUNWALD Armin
The Hermeneutic Side of Responsible Research and Innovation
(Responsible Research and Innovation Set – Volume 5)

KURDI Mohamed Zakaria
Natural Language Processing and Computational Linguistics 1: Speech, Morphology and Syntax

LENOIR Virgil Cristian
Ethical Efficiency: Responsibility and Contingency
(Responsible Research and Innovation Set – Volume 1)

MATTA Nada, ATIFI Hassan, DUCELLIER Guillaume
Daily Knowledge Valuation in Organizations

NOUVEL Damien, EHRMANN Maud, ROSSET Sophie
Named Entities for Computational Linguistics

PELLÉ Sophie, REBER Bernard
From Ethical Review to Responsible Research and Innovation
(Responsible Research and Innovation Set - Volume 3)

REBER Bernard
Precautionary Principle, Pluralism and Deliberation
(Responsible Research and Innovation Set – Volume 4)

SILBERZTEIN Max
Formalizing Natural Languages: The NooJ Approach

2015

LAFOURCADE Mathieu, JOUBERT Alain, LE BRUN Nathalie
Games with a Purpose (GWAPs)

SAAD Inès, ROSENTHAL-SABROUX Camille, GARGOURI Faïez
Information Systems for Knowledge Management

2014

DELPECH Estelle Maryline
Comparable Corpora and Computer-assisted Translation

FARINAS DEL CERRO Luis, INOUE Katsumi
Logical Modeling of Biological Systems

MACHADO Carolina, DAVIM J. Paulo
Transfer and Management of Knowledge

TORRES-MORENO Juan-Manuel
Automatic Text Summarization

2013

TURENNE Nicolas
Knowledge Needs and Information Extraction: Towards an Artificial Consciousness

ZARATÉ Pascale
Tools for Collaborative Decision-Making

2011

DAVID Amos
Competitive Intelligence and Decision Problems

LÉVY Pierre
The Semantic Sphere: Computation, Cognition and Information Economy

LIGOZAT Gérard
Qualitative Spatial and Temporal Reasoning

PELACHAUD Catherine
Emotion-oriented Systems

QUONIAM Luc
Competitive Intelligence 2.0: Organization, Innovation and Territory

2010

ALBALATE Amparo, MINKER Wolfgang
Semi-Supervised and Unsupervised Machine Learning: Novel Strategies

BROSSAUD Claire, REBER Bernard
Digital Cognitive Technologies

2009

BOUYSSOU Denis, DUBOIS Didier, PIRLOT Marc, PRADE Henri
Decision-making Process

MARCHAL Alain
From Speech Physiology to Linguistic Phonetics

PRALET Cédric, SCHIEX Thomas, VERFAILLIE Gérard
Sequential Decision-Making Problems / Representation and Solution

SZÜCS Andras, TAIT Alan, VIDAL Martine, BERNATH Ulrich
Distance and E-learning in Transition

2008

MARIANI Joseph
Spoken Language Processing