Linguistic-based Patterns for Figurative Language Processing: The Case of Humor Recognition and Irony Detection

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Antonio Reyes Pérez

Departamento de Sistemas Informáticos y Computación

Universitat Politècnica de València

A thesis submitted for the degree of

PhilosophiæDoctor (PhD)

Under the supervision of

Dr. Paolo Rosso

July 2012

Para mi familia, que nunca paran de apoyarme y quererme.

Por mi propia familia, que me complementa y motiva.

Acknowledgements

I want to thank to all the people involved in this project. In particular to Paolo for his guidance, advices, and engagement. Alberto for his unconscious tutoring and help. David, Davide and Yassine for providing some nice ideas concerning my unusual research topic. I am also really grateful to Toni Martí, Carlo Strapparava, and Tony Veale for sharing their priceless knowledge. Finally, my thankfulness to the rest of the doctoral committee for being part of this thesis.

I would also like to thank to my family and friends for their unconditional love, confidence, and courage. Particularly to my mom, siblings, aunts, uncles, as well as my grandma and nephews. I cannot forget my dad: although you are not with us anymore, this achievement is also for you.

Last but not least, all my gratitude and love to the woman that has always been with me: Leticia. Thanks for everything and, especially, for the great new that is changing our lives: our baby.

All of you, with your faithful and continuous support, made this project could come true.

Funding

This work has been funded by the National Council for Science and Technology (CONACyT - Mexico); as well as partially supported by the Text-Enterprise 2.0 (TIN2009-13391-C04-03) within the Plan I+D+i of MICINN; the WIQ-EI IRSES project (grant no. 269180) within the EU FP7 Marie Curies People research projects; and the TeLMoSis (UPV PAID083294) research project. The work was partially done in the framework of the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems.

Dissertation Committee

1. Antónia Martí Antonín

2. Carlo Strapparava

3. Richard Anthony Veale

4. Walter Daelemans

5. José Antonio Troyano Jiménez

Day of the defense: July 2nd, 2012

Signature from head of PhD committee:

Abstract

Figurative language represents one of the most difficult tasks regarding natural language processing. Unlike literal language, figurative language takes advantage of linguistic devices such as irony, humor, sarcasm, metaphor, analogy, and so on, in order to communicate indirect meanings which, usually, are not interpretable by simply decoding syntactic or semantic information. Rather, figurative language reflects patterns of thought within a communicative and social framework that turns quite challenging its linguistic representation, as well as its computational processing.

In this Ph. D. thesis we address the issue of developing a linguistic-based framework for figurative language processing. In particular, our efforts are focused on creating some models capable of automatically detecting instances of two independent figurative devices in social media texts: humor and irony. Our main hypothesis relies on the fact that language reflects patterns of thought; i.e. to study language is to study patterns of conceptualization. Thus, by analyzing two specific domains of figurative language, we aim to provide arguments concerning how people mentally conceive humor and irony, and how they verbalize each device in social media platforms. In this context, we focus on showing how fine-grained knowledge, which relies on shallow and deep linguistic layers, can be translated into valuable patterns to automatically identify figurative uses of language. Contrary to most researches that deal with figurative language, we do not support our arguments on prototypical examples neither of humor nor of irony. Rather, we try to find patterns in texts such as blogs, web comments,

tweets, etc., whose intrinsic characteristics are quite different to the characteristics described in the specialized literature.

Apart from providing a linguistic inventory for detecting humor and irony at textual level, in this investigation we stress out the importance of considering user-generated tags in order to automatically build resources for figurative language processing, such as ad hoc corpora in which human annotation is not necessary.

Finally, each model is evaluated in terms of its relevance to properly identify instances of humor and irony, respectively. To this end, several experiments are carried out taking into consideration different data sets and applicability scenarios. Our findings point out that figurative language processing (especially humor and irony) can provide fine-grained knowledge in tasks as diverse as sentiment analysis, opinion mining, information retrieval, or trend discovery.

Resumen

El lenguaje figurado representa una de las tareas más difíciles del procesamiento del lenguaje natural. A diferencia del lenguaje literal, el lenguaje figurado hace uso de recursos lingüísticos tales como la ironía, el humor, el sarcasmo, la metáfora, la analogía, entre otros, para comunicar significados indirectos que la mayoría de las veces no son interpretables sólo en términos de información sintáctica o semántica. Por el contrario, el lenguaje figurado refleja patrones del pensamiento que adquieren significado pleno en contextos comunicativos y sociales, lo cual hace que tanto su representación lingüística, como su procesamiento computacional, se vuelvan tareas por demás complejas.

Dentro de este contexto, en esta tesis de doctorado se aborda una problemática relacionada con el procesamiento del lenguaje figurado a partir de patrones lingüísticos. En particular, nuestros esfuerzos se centran en la creación de modelos capaces de detectar de manera automática instancias de humor e ironía en textos extraídos de medios sociales. Nuestra hipótesis principal se basa en la premisa de que el lenguaje refleja patrones de conceptualización; es decir, al estudiar el lenguaje, estudiamos tales patrones. Por tanto, al analizar estos dos dominios del lenguaje figurado, pretendemos dar argumentos respecto a cómo la gente los concibe, y sobre todo, a cómo esa concepción hace que tanto humor como ironía sean verbalizados de manera particular. En este sentido, uno de nuestros mayores intereses es demostrar cómo el conocimiento que proviene del análisis de diferentes niveles de estudio lingüístico puede representar un conjunto de patrones relevantes para identificar automáticamente usos figurados del lenguaje.

Cabe destacar que contrario a la mayoría de aproximaciones que se han enfocado en el estudio del lenguaje figurado, en nuestra investigación no buscamos dar argumentos basados únicamente en ejemplos prototípicos, sino en textos cuyas características intrínsecas son muy distintas de las descritas en la literatura especializada; por ejemplo, en blogs, comentarios web, tweets, etc.

Además de aportar un repertorio de patrones lingüísticos para detectar humor e ironía a nivel textual, en esta investigación hacemos énfasis en el hecho de considerar las etiquetas generadas por los mismos usuarios con el fin de crear recursos destinados al procesamiento del lenguaje figurado; por ejemplo, en la construcción automática de corpora especializados, en los cuales ya no es necesaria la intervención de anotadores humanos para etiquetar los datos.

Finalmente, describimos cómo evaluamos los modelos propuestos en términos de su capacidad y relevancia para identificar de manera correcta instancias de humor e ironía, respectivamente. Para ello, ejecutamos varios experimentos tomando en cuenta diferentes conjuntos de datos, así como diferentes escenarios de aplicabilidad. Los resultados obtenidos muestran que el procesamiento del lenguaje figurado (en especial, el humor y la ironía) puede aportar conocimiento relevante para tareas tan diversas como el análisis de sentimientos, el minado de opiniones, la recuperación de información o el descubrimiento de las tendencias de los usuarios.

Resum

El llenguatge figurat constitueix una de les tasques més difícils del processament del llenguatge natural. A diferència del llenguatge literal, el llenguatge figurat fa ús de recursos lingüístics com la ironia, l'humor, el sarcasme, la metàfora, l'analogia, entre d'altres, per comunicar significats indirectes que la majoria de vegades no es poden interpretar només amb informació sintàctica o semàntica. Contràriament, el llenguatge figurat reflexa patrons de pensament que adquireixen ple significat en contextos comunicatius i socials, cosa que fa que tant la seva representació lingüística com el seu processament computacional siguin tasques difícils.

En aquest context, en aquesta tesi de doctorat s'aborda una problemàtica relacionada amb el processament del llenguatge figurat a partir de patrons lingüístics. En particular, els nostres esforos se centren en la creació de models capaços de detectar automàticament instàncies d'humor i d'ironia en textos extrets de mitjans socials. La nostra hipòtesi principal es basa en la premissa que el llenguatge reflecteix patrons de conceptualització; és a dir, quan estudiem el llenguatge, estudiem aquests patrons. Per tant, en analitzar aquests dos dominis del llenguatge figurat, pretenem donar arguments respecte de com la gent els concep i, sobretot, com aquesta concepció fa que tant l'humor com la ironia siguin verbalitzats de manera particular. En aquest context, un dels nostres majors interessos és demostrar com el coneixement que prové de l'anàlisi de diferents nivells d'estudi lingüístics pot representar un conjunt de patrons rellevants per identificar automàticament usos figurats del llenguatge. Cal destacar que, al contrari de la majoria d'aproximacions que han estudiat el llenguatge

figurat, en la nostra investigació, no pretenem donar arguments basats únicament en exemples prototípics, sinó en exemples provinents de textos les característiques intrínseques dels quals són molt diferents de les descrites en la literatura especialitzada; per exemple, blogs, comentaris web, tweets, etc.

A més d'aportar un repertori de patrons lingüístics per detectar humor i ironia a nivell textual, en aquesta investigació, fem èmfasi en el fet de considerar les etiquetes generades pels mateixos usuaris amb la finalitat de crear recursos destinats al processament del llenguatge figurat, per exemple, la construcció automàtica de corpora especialitzats, en els quals ja no és necessària la intervenció d'anotadors humans per etiquetar les dades.

Finalment, descrivim com avaluem els models proposats en termes de la seva capacitat i rellevància per identificar de manera correcta instàncies d'humor i ironia, respectivament. Amb aquest fi, duem a terme diversos experiments tenint en compte diferents conjunts de dades, així com diferents escenaris d'aplicabilitat. Els resultats obtinguts mostren que el processament del llenguatge figurat (especialment, l'humor i la ironia) pot aportar coneixement rellevant per a tasques tan diverses com l'anàlisi de sentiments, la mineria d'opinions, la recuperació d'informació o el descobriment de les tendències dels usuaris.

Contents

List of Figures

LIST OF FIGURES

List of Tables

1

Introduction

Dios te libre, lector, de prólogos largos y de malos epítetos.

Francisco de Quevedo [42]

Figurative language is one of the most arduous topics that natural language processing (NLP) has to face. Unlike literal language, the former takes advantage of linguistic devices such as metaphor, analogy, ambiguity, irony, and so on, in order to project more complex meanings which, usually, represent a real challenge, not only for computers, but for humans as well. This is the case of humor and irony. Each device exploits different linguistic mechanisms in order to produce its effect (e.g. ambiguity and alliteration regarding humor (Mihalcea and Strapparava [102], Sjöbergh and Araki [154]); similes regarding irony (Veale and Hao [173])). Sometimes such mechanisms are similar (e.g. use of satirical or sarcastic expressions to communicate a negative attitude (Kumon-Nakamura et al. [83], Attardo [7])). Both figurative devices, moreover, entail cognitive capabilities to make abstractions as well as to interpret the meaning beyond literal words; i.e. figurative language reflects patterns of thought within a communicative, and thus, social framework.

In this respect, communication is more than sharing a common code, but being capable of inferring information beyond syntax or semantics; i.e. figurative language implies information not grammatically expressed. If such information is not correctly unveiled, then the real meaning is not achieved and accordingly,

the figurative effect is lost. Let us consider a joke. The amusing effect sometimes relies on not given information. If such information is not filled, the result is a bad, or better said, a misunderstood joke.

In addition, the necessary processes to properly interpret such figurative information entail a great challenge because they point to social and cognitive layers that are quite difficult to be computationally represented. However, regardless of the inconveniences that figurative language entails, the approaches to automatically process figurative devices, such as humor, irony or sarcasm, seem to be quite encouraging. For instance, the research works concerning automatic humor generation (Binsted and Ritchie [18], Stock and Strapparava [159]) and automatic humor recognition (Mihalcea and Strapparava [102], Mihalcea and Pulman [98]), as well as the investigations concerning irony detection (Utsumi [168], Veale and Hao [173], Carvalho et al. [27]), satire detection (Burfoot and Baldwin [24]), and sarcasm detection (Tsur et al. [166]), have shown the feasibility of computationally approaching figurative language. Moreover, figurative language might also unveil valuable knowledge for tasks such as edutainment, advertising, sentiment analysis, trend discovery, computer assisted translation, and so on.

This investigation, thus, aims at showing how two specific domains of figurative language: humor and irony, can be automatically handled by means of considering linguistic-based patterns. We are especially focused on discussing how underlying knowledge, which relies on shallow and deep linguistic layers, can represent relevant information to automatically identify figurative uses of language. In particular, and contrary to most researches on figurative language, we aim to identify figurative uses of language in social media. This means that we are not focused on analyzing prototypical jokes or literary examples of irony; rather, we try to find patterns in social media texts of informal register whose intrinsic characteristics are quite different to the characteristics described in the specialized literature. For instance, a joke which exploits phonetic devices to produce a funny effect, or a tweet in which irony is self-contained in the situation. In this context, we propose a set of features which work together as a system: no single feature is particularly humorous or ironic, but all together provide a useful linguistic inventory for detecting humor and irony at textual level.

1.1 The Problem: Humor and Irony

This investigation is focused on analyzing two playful domains of figurative language: humor and irony. In particular, we are focused on **textual instances of verbal humor and verbal irony**, respectively.

Verbal humor explicitly refers to the type of verbally expressed humor. Non-verbal forms of humor (e.g. visual or situational) are beyond the scope of this thesis. Verbal irony, in contrast, is a linguistic phenomenon in which there is opposition between what it is literally communicated and what it is figuratively implicated[1]. Putting into context these concepts, we could simply define humor by the presence of amusing effects, such as laughter or well-being sensations, whose main function is to release emotions, sentiments or feelings. In a social context, humor's cathartic properties make most people react to a humorous stimulus regardless of their beliefs, social status or cultural differences, thereby providing valuable information related to linguistic, psychological, neurological and sociological phenomena. However, given its complexity, humor is still an undefined phenomenon. Partly, because the stimuli that make people laugh can hardly be generalized or formalized. For instance, cognitive aspects as well as cultural knowledge, are some of the multi-factorial variables that should be analyzed in order to understand humor's properties. Despite such inconveniences, different disciplines such as philosophy (Halliwell [65]), linguistics (Attardo [5]), psychology (Ruch [143]), or sociology (Hertzler [69]), have attempted to study humor in order to provide formal insights to explain better its basic characteristics.

With respect to irony, most studies have a linguistic approach. Such studies define irony basically as a communicative act that expresses the opposite of what it is literally said (Wilson and Sperber [181]). However, experts can distinguish among: situational irony (Lucariello [91]), where some confluence of objects or events upends our common-sense expectations of the world (e.g. finding out that the Dalai Lama is a meat-eater, or that Adolf Hitler was a vegetarian); poetic irony (Colston [35]), where a protagonist suffers an apt yet unexpected setback

[1]Unlike verbal humor, we do not reject the possibility that our findings can be applied to situational irony, not least because much of the irony in online texts exhibits precisely this type of irony.

(e.g. when the head of a major fast food corporation has a heart-attack); cosmic irony, where nature seems to mock man's efforts to control events (e.g. a tornado tearing through a drive-in movie theater while it is showing the movie "Twister"); dramatic irony (Attardo [7]), where the reader of a novel or the viewer of a film knows more about a fictional character than the character himself (e.g. as in Shakespearean tragedies such as MacBeth); and verbal irony (Colston and Gibbs [36]), where a speaker uses a form of speech that is superficially more appropriate to a very different context or meaning. As previously noted, it is the latter type, verbal irony, that mainly interest us, and especially, its use in social media. However, once one actually views the data itself, it becomes clear that casual speakers rarely recognize the pragmatic boundaries concerning the types of irony outlined above; i.e. texts by non-experts who use an intuitive and unspoken definition of irony rather than one sanctioned by a dictionary or a text-book.

It is worth noting that both figurative devices: verbal humor and verbal irony were selected due to i) verbal humor is the most tangible, and perhaps, the most widely type of humor (Mihalcea [97]); ii) whereas verbal irony, unlike situational or dramatic irony, is intrinsically intentional; therefore, it is more tangible (Valitutti [169]). In addition, both devices are suitable to be computationally represented by means of linguistic patterns. Finally, we think that there are various patterns given in situations where humor and irony are implied that are worth analyzing for practical tasks such as sentiment analysis, e-commerce, or information retrieval.

1.2 The Core of the Problem: (Figurative) Language in Social Media

Web-based technologies have become a significant source of data in a variety of scientific and humanistic fields. Such technologies provide a rich vein of information that is easily mined. User-generated content (such as text, audio and images) provides knowledge that is topical, task-specific, and dynamically updated to broadly reflect changing trends, behavior patterns and social preferences. Consider, for instance, the research work described by Pang et al. [113] which shows

the role of implicit knowledge in automatically determining the subjectivity and polarity of movie reviews; or the findings reported by Balog et al. [12] regarding the role of user-generated tags for analyzing mood patterns among bloggers.

In this context, figurative language can be found on almost every web site in a variety of guises and with varying degrees of obviousness. For instance, when analyzing instances of irony, one of the most important micro-blogging sites, Twitter, allows its users to self annotate their posts with user-generated tags (or hashtags according to Twitter's terminology). Thus, the hashtag #irony is used by people in order to self-annotate all varieties of irony, whether they are chiefly the results of deliberate word-play or merely observations of the humor inherent in everyday situations. Similar situation concerning social (non-prototypical) examples of verbal humor: people self-annotate their posts (web comments, tweets, user reviews, etc.) by highlighting certain characteristics that throw focus onto certain aspects of a funny text. For instance, when analyzing such examples, one realizes that they are often observations of life's little ironies (e.g. "Sitting in the eye-doctor's office, waiting for the doctor to see me"), or simply sarcastic expressions (e.g. "I thank God that you are unique!"). Such behavior makes quite complicated establishing accurate boundaries to differentiate specific (prototypical) examples of figurative language in social media texts. The safest generalization that one can draw is that people perceive figurative language at the boundaries of conflicting frames of reference, in which an expectation of one frame has been inappropriately violated in a way that is appropriate in the other. In this respect, experts can tease apart the fine distinctions between one specialized form of figurative language and another, in ways that casual speakers find it unnecessary to do.

The question here is why do we focus on representing figurative language patterns based on social media examples rather than on prototypical examples? Mainly, due to language is not a static phenomenon; rather, it is continuously changing. Such changes, that come from oral language, are easily registered and generalized in written language by taking advantage of the new technological platforms, especially, the web-based platforms. In this respect, social media are the best examples concerning the impact of such technologies on language and social habits: communication, for instance, is slightly changing and acquiring wider scope because of the existence of new ways of interacting. As our media

increasingly become more social, the problem of (figurative) language will become even more pressing. Therefore, in this investigation we opted for analyzing figurative language in terms of dynamic, living, and current examples, rather than in terms of static, ad hoc, and literary examples of humor and irony, respectively. In addition, our interests are addressed to apply our findings in real systems. Hence, it would be useless supporting a figurative language model based on, for instance, Quevedo's irony, rather than on people's irony. In this respect, part of the challenge of recognizing figurative language in user-generated contents is to avoid misclassification in online contents, as well as to be able to mine fine-grained knowledge from such contents. Thus, from a NLP perspective, the relevance of this investigation lies on the fact of dealing with non-factual information that is linguistically expressed, and therefore, it is extremely useful in the automatic mining of new knowledge; i.e. sentiments, attitudes, moods, feelings, etc., which are inherent to humor and irony, and on a broader level, to language and social communication.

1.3 Objective

Figurative language is in some way inherent to discourse, whatever the type of text (Vernier and Ferrari [176]). The problem of automatically detecting figurative language cuts through every aspect of language, from pronunciation to lexical choice, syntactic structure, semantics and conceptualization. As such, it is unrealistic to seek a computational silver bullet for figurative language, and a general solution will not be found in any single technique or algorithm. Rather, we must try to identify specific aspects and forms of figurative language that are susceptible to be computationally analyzed, and from these individual treatments attempt to synthesize a gradually broader solution. In this context, our objective is to deeply analyze two figurative devices: humor and irony, in order to detect textual patterns to be applied in their automatic processing, especially, in their automatic identification at textual level. Thus, we aim to propose two specific models concerning two independent tasks: humor recognition and irony detection.

Each model is intended to represent the most salient attributes of verbal humor and verbal irony, respectively: or at least, what speakers believe to be humor and irony in a social media text. In order to achieve this objective, three overall tasks must be performed:

I. Collect objective data concerning independent examples of humor and irony.

II. Represent each device by means of textual patterns that are suggestive of humor and irony, respectively.

III. Assess the set of patterns by analyzing their ability to automatically differentiate humorous from non-humorous texts, and ironic from non-ironic texts.

In addition, this objective deals with some conceptual issues that are addressed throughout the thesis.

i. Literal language and figurative language are windows to cognitive processes that are linguistically verbalized. The meanings encoded by linguistic symbols refer to projected realities (Jackendoff [70]). In the analysis and representation of (figurative) language, the meaning cannot be derived only from lexicon, but from its use as well. Therefore, an integral vision of language, in which its grammatical substance is as important as its social use, is basic to understand how figurative meaning is conveyed.

ii. Overlapping is quite common in figurative language (Triezenberg [165]). Indeed, it appears quite often in examples of humor and irony. For instance, irony is a common mechanism to produce a humorous effect, and *vice versa*, humor is usually an effect of ironic expressions. In the absence of formal linguistic boundaries to accurately separate such devices, the task of defining a model capable of representing both phenomena must take into account fine-grained patterns, in such a way they allow the identification of particularities supported by generalities.

iii. Figurative language is fuzzy enough to be computationally, and even linguistically, represented. Specialized literature, in this respect, defines humor and irony in fine-grained terms. However, such granularity cannot be

directly mapped from theory to praxis due largely to the idealized communicative scenarios that such granularity entails. Concerning our approach, such fine-grained scenarios do not match with the scenarios registered in our data. Hence, it is necessary to represent the core of both devices the less abstract as possible, in order to describe deeper and more general attributes of both phenomena; rather than only particular cases that 100% match with prototypical descriptions.

iv. Humor and irony are typical devices in which both literal and non-literal meaning might be simultaneously active (Cacciari [26]). Moreover, there are not linguistic marks to denote that any expression is funny or ironic. For instance, although there is a general agreement with respect to verbal irony's main property: opposition; such opposition usually lacks of an explicit negation marker. Therefore, any attempt to computationally model these phenomena must be robust enough to properly deal with such theoretical and practical issues.

1.4 Thesis Overview

This thesis is conceptually organized as follows:

In Chapter 2 we will describe the linguistic background as well as the theoretical issues regarding literal and figurative language. We will emphasize the importance of considering language as a dynamic system, rather than a static one. Thus, examples of both linguistic realities: literal and figurative, will be given. Finally, both humor and irony will be conceptually described and discussed in detail. In Chapter 3 we will introduce the related work concerning figurative language processing. First, the framework in which this thesis is developed will be described. Then, the challenges that any computational treatment of figurative language faces will be outlined. In addition, the state-of-the-art concerning the computational treatment of humor and irony will be detailed.

In Chapter 4 we will describe, both conceptually and pragmatically, our humor recognition model. Hypotheses, patterns, experiments, and results will be presented. Moreover, evaluation data sets will be introduced. Finally, we will

discuss model's implications. In Chapter 5, in turn, we will present our irony detection model. First, operational bases, as well as aims, will be outlined. Then, experiments and results will be explained. Like in the previous chapter, all the evaluation data sets will be introduced. Lastly, results and further implications will be discussed.

In Chapter 6 we will describe how both models are assessed in terms of their applicability in tasks related to information retrieval, sentiment analysis, and trend discovery. Such evaluations are intended to represent real scenarios concerning the treatment of figurative language beyond the data sets employed in Chapters 4 and 5. Finally, in Chapter 7 we will outline the main conclusions of this thesis, as well as its contributions and lines for future work.

2

Figurative Language

His research is about as ground-breaking as a foam jackhammer.

Veale et al. [175]

This chapter will be focused on describing some **theoretical issues regarding language**. In particular, we will concentrate on discussing similarities and differences concerning literal language and figurative language. Furthermore, we will talk about some of the most relevant figurative devices cited in the specialized literature. Two specific figurative devices will be analyzed in detail: humor and irony. Based on linguistic arguments, we will outline the difficulty of automatically dealing with these phenomena. To this end, examples regarding their usages in natural language will be given. Finally, overall definitions of both devices, taking into consideration further discussions, will be also established.

2.1 Background

Language, in all its forms, is the most natural and important mean of conveying information. However, given its social nature, it cannot be conceptualized only in terms of grammatical issues. In this respect, while it is true that grammar regulates language in order to have a non-chaotic system, it is also true that language is dynamic, and accordingly, is a live entity. This means that language is

not static; rather it is in constant interaction between the rules of its grammar and its pragmatic use. For instance, the idiom *"all of a sudden"* has a grammatical structure which is not made intelligible only by knowledge of the familiar rules of its grammar (Fillmore et al. [50]), but by inferring pragmatic information as well. This latter provides the knowledge that, in the end, gives sense to the idiom.

Emphasizing the social aspect of language, modern linguists deem language as a continuum of symbolic structures in which lexicon, morphology, and syntax form a continuum which differs along various parameters but can be divided into separate components only arbitrarily (Langacker [86, 87]). Language thus is viewed as an entity whose components and levels of analysis cannot be independent nor isolated. On the contrary, they are embedded in a global system which depends on cognitive, experiential, and social contexts, which go far beyond the linguistic system proper (Kemmer [79]).

This vision, according to the cognitive linguistics bases, entails a close relation between semantics and conceptualization (cf. Langacker [86, 87], Goldberg [60, 61], Fillmore [49]); i.e. apart from grammar, the linguistic system is dependent on *cognitive domains* in which both referential knowledge (e.g. lexical semantic information) and inferential knowledge (e.g. contextual and pragmatic information) are fundamental elements to understand what it is communicated. Let us consider the following example:

1) "I really need some antifreeze in me on cold days like this".

Example 1 is fully understandable only within a cognitive domain in which the sense is given by figuring out the analogy between $antifreeze$ (referential knowledge: antifreeze is a liquid) and liquor (inferential knowledge: antifreeze is a liquid, liquor is a liquid, antifreeze is a liquor)[1]. These cognitive domains are based on conventional images of the reality (semantic knowledge), as well as on the discursive use of such reality (pragmatic knowledge). Together, they constitute the core of what it is communicated: the meaning.

[1]Ferdinand de Saussure [43] argued that meaning is an abstract representation, stable and independent of its pragmatic use. If this was completely true (beyond a perfect linguistic system), then the example would be senseless due to the abstract representation of antifreeze cannot exceed its semantic boundaries.

According to this point of view, language is the mean by which the reality is verbalized and acquires, thus, its complete meaning (see Lakoff and Johnson [85], Langacker [86]). Based on this integral vision of language, in which its grammatical substance is as important as its social referents, we will subscribe the arguments to describe, analyze and support our approach. In particular, we will focus on highlighting the role of the cognitive processes[2] that impact on the linguistic system when expressing figurative language.

2.2 Literal Language

Traditionally, language has been described from dichotomous points of view: *langue vs. parole*, signifier *vs.* signified, synchrony *vs.* diachrony, paradigmatic *vs.* syntagmatic, oral *vs.* written, an so on. In this section, another dichotomy will be discussed: literal language *vs.* figurative language. The objective is to define and describe these linguistic realities (exemplifying their similarities and dissimilarities) in order to set the minimum bases (at linguistic level) for their automatic differentiation.

The simplest definition of literal language is related to the notion of *true*, *exact* or *real* meaning; i.e. a word (isolated or within a context) conveys one single meaning (the one conventionally accepted), which cannot be deviated. In Saussure's terms, literal meaning is corresponded with a perfect dichotomy of signifier and signified (cf. [43]). Some experts, in addition, have noticed certain properties of literal meaning: it is direct, grammatically specified, sentential, necessary, and context-free (see Katz [77], Searle [151], Dascal [40]). Hence, it is assumed that it must be invariant in all contexts. According to Ariel [2], literal meaning is generated by linguistic knowledge of lexical items, combined with linguistic rules. Therefore, it is determinate, explicit, and fully compositional. For instance, the word *flower* can only refer to the concept of plant, regardless of its use in different communicative acts or discourses (e.g. botany, evolution, poetry).

[2] As they are understood in Cognitive Grammar: metaphor, metonymy, mental imagery, etc. (see Fillmore [49], Langacker [86], Lakoff [84], Goldberg [60].)

Although these properties suppose a perfect symmetry between the word and its meaning, such symmetry rarely appears in the reality. Let us consider the Chomskyan dichotomy competence *vs*. performance. Putting into context these terms, competence refers to the linguistic system (components and rules) which allows both speaker and hearer to produce and understand an infinite number of sentences (regardless of whether they are grammatically correct or not). Performance, instead, is the capability of using such linguistic system. It does not depend exclusively on rules but on extra-linguistic factors such as memory, distractions, shifts of attention and interest, etc., which determine the meaning at communicative level (cf. Chomsky [30]). In accordance with these concepts, literal meaning has two faces: one which depends on the competence (which corresponds with the properties described so far), and another one which depends on the performance. This latter is deviated from the concept of literalness due to the meaning does not depend entirely on what it is conventionally accepted, but on communicative processes and information that, according to Chomsky, is out of the linguistic system. For instance, *there* entails its linguistic meaning (e.g. the one registered in a dictionary), which is complemented with information given by its use in specific situations (e.g. point of reference). Thus, its literal meaning will depend on interpreting the linguistic meaning within a communicative context.

This "value-add" shows up the role of the communication when setting up, in a comprehensive way, the properties of literalness. In this respect, Grice argued that literal meaning is what it is said ([63]). However, what it is said does not always correspond with what it is interpreted. Glucksberg and McGlone [59] emphasized such distinction by stating a difference between linguistic decoding and linguistic interpretation. According to these authors, before any type of interpretation can be generated, utterances must be decoded (phonologically for spoken language, orthographically for written text), lexically and syntactically, at least to some minimal extent. Once decoded, utterances must then be interpreted: literally, figuratively, or both ([59]). Let us consider example 2 to clarify their point:

2) "In 1995 my girlfriend was a student".

The process of decoding will indicate that this sentence is fully understandable in terms of its components (words) and syntax. Lexically, there is no doubt with respect to the meaning of each word. However, to be able to fully understand it, some communicative and contextual gaps must be filled. For instance, it is necessary to perform processes such as inferences, implications, or assumptions, in order to communicatively determine who is the pragmatic subject:

3) In 1995 my girlfriend [X] was a student [A]; today

   a) she is my girlfriend, but not a student anymore [X = X', A $\neq$ A'];

   b) she is a student, but not my girlfriend anymore [A = A', X $\neq$ X',];

   c) she is neither my girlfriend nor a student [X $\neq$ X', A $\neq$ A'].

Depending on our selection, there will be three possible interpretations. Each will provide valuable information to determine the pragmatic subject, and lastly, the meaning. Furthermore, each will keep the core of the sentence unchangeable; i.e. its literalness will not be deviated, rather, will be particularized.

Thus, by taking into consideration both processes (decoding and interpretation), the concept of literal meaning would seem to be more complex than the simple alignment one word, one meaning. On one hand, as we have previously discussed, language entails dynamism, interaction, change, live (see Section 2.1). Word meaning, on the other hand, is not "fixed" but is rather a function of perspective (Sikos et al. [153]). Based on these criteria, as well as of the arguments given along this section, it is necessary to redefine our initial concept of literalness by setting it off in terms of lexicon, context (both socially as linguistically), communication, and pragmatic motivations. Therefore, we conclude this section by stating our definition of literalness.

Literal language : refers to the notion of symmetry between what it is said and what it is decoded. Stable meaning. Conceptual.

Literal meaning : refers to the result of what it is interpreted. The core of the meaning is in the lexicon. However, this is complemented with contextual, communicative and pragmatic information. There is no meaning deviation. If so, this is NOT intentional, it is due by errors when interpreting what it is decoded. Processual.

2.3 Figurative Language

In the context of a dichotomous view of language, figurative language could be regarded as the simple oppositeness of literal language. Thus, whereas the latter is assumed to communicate a direct meaning, the former is more related to the notion of conveying indirect or veiled meanings. For instance, the word *flower*, which literally refers only to the concept of plant, speaking figuratively can refer to several concepts, which not necessarily are linked to plants. Therefore, it can be used instead of concepts such as beauty, peace, purity, life, and so on, in such a way its literal meaning is **intentionally** deviated in favor of secondary interpretations[3].

Although, at first glance, this distinction seems to be clear and sufficient on its own, figurative language involves basic cognitive processes rather than only deviant usage (Peters [118]). Therefore, it is necessary going deeper into the mechanisms and processes that differentiate both types of languages.

In accordance with classical perspectives, the notions of literalness and figurativity are viewed as pertaining directly to language; i.e. words have literal meanings, and can be used figuratively (Katz [77], Searle [151], Dascal [40]). Figurative language, thus, could be regarded as a type of language that is based on literal meaning, but is disconnected from what people learn about the world [or about the words] based on it [them] [15]. Thus, by breaking this link, literal meaning loses its primary referent and, accordingly, the interpretation process becomes senseless. Let us consider Chomsky's famous example to explain this issue:

4) "Colorless green ideas sleep furiously" [29].

Beyond grammatical aspects, in example 4 is possible to observe how the decoding process is achieved easily enough. Either phonologically or orthographically, Chomsky's example is fully understandable in terms of its linguistic constituents. However, when interpreting, its literal meaning is completely nonsensi-

[3]It is worth noting that such secondary interpretations are not guaranteed. Their success will depend on several factors, both linguistic as extra-linguistic. This issue will be discussed later in this section.

cal. For instance, the bigrams [colorless green] or [green ideas] are sufficiently disconnected from their conventional referents for being able to produce a coherent interpretation. Thus, in order to make the example understandable, secondary interpretations are needed. If such interpretations are successfully activated, then figurative meaning is triggered[4] and, accordingly, a more coherent interpretation can be achieved. Based on this explanation, literal meaning could be deemed as denotative, whereas figurative meaning, connotative; i.e. figurative meaning is not given a priori; rather, it must be implicated.

On the other hand, according to Katz et al. [76], much figurative meaning is based on learned convention, such as with idioms, and proverbs. Therefore, its use is not lexicalized[5] (Li and Sporleder [88]), although is pragmatically motivated[6]. In this respect, figurative language plays an important role on communication due to the need of performing mental processes such as reasoning and inferencing (Peters [118]), which require additional cognitive effort (Gibbs [53]). Moreover, Lönneker-Rodman and Narayanan [89] point out that figurative language can tap into conceptual and linguistic knowledge (as in the case of idioms, metaphor, and some metonymies), as well as evoke pragmatic factors in interpretation (as in indirect speech acts, humor, irony, or sarcasm). In accordance with the assumptions given by these authors, an expected conclusion is to conceive the processes of interpreting figurative language much more complex than the ones performed when interpreting literal language. Let us consider examples 3 and 4 to put these assumptions into context. Whereas example 3 (In 1995 my girlfriend was a student) is semantically understandable on its own, in terms of pragmatics it requires of infering facts, as well as implicating relations, to fully interpret its underlying meaning. In contrast, in example 4 (Colorless green ideas sleep furi-

[4]According to Sikos et al. [153], understanding figurative language often involves an interpretive adjustment to individual words; i.e. not all the constituents of the example trigger a figurative meaning on their own; rather, this is usually triggered by manipulating individual words.

[5]Not conventionally accepted. For instance, when the sense is not registered in a dictionary.

[6]In this respect, authors such as Searle [151] or Grice [63], note that the standard pragmatic model assumes that understanding what speakers figuratively implicate in context demands pragmatic information that is more difficult to access than the semantic knowledge used to determine literal meaning (Gibbs [53]).

ously) these same processes are not enough to find out its meaning, not even at semantic level[7]. Apart from inferences and implications, it would be necessary to find out how the information given can be associated with new conceptual frames (mostly extra-linguistic). These frames will then make sense in specific contexts by adjusting their prototypical referents to the ones of the new frame, in such a way a coherent sense can be achieved[8]. Thus, whereas example 3 would be communicating a simple semantic statement (particularized with pragmatic information), example 4 should first be set into a frame that will provide the necessary pragmatic information to support a coherent semantic sense. In this way, its figurative meaning would be unveiled. Of course, in case it had one (see footnote 7).

Although the arguments given in this section provide sufficient elements to determine what figurative language is, the main question still remains: How to differentiate between literal language and figurative language (theoretically and automatically)?

The examples given so far have shown some of their main characteristics; however, based on that information, there is no way of totally affirming that example 1 is more figurative than example 4, or example 3 is the most literal of all. Finally, all of them could be examples, either of literal or figurative language. To be able to provide arguments for differentiating both linguistic realities, a crucial extra-linguistic element (with linguistic repercussion) must be highlighted: **intentionality**. Beyond mechanisms to explain why figurative language requires much more cognitive efforts to correctly interpret its meaning, the most important issue is that these examples are, in the end, sequences of words with semantic meaning. Perhaps, such meaning is very clear (literalness), or perhaps could be senseless (figurativity); but lastly, this difference could be explained in terms of performance and competence or even as a matter of correctness. However, in a more comprehensive conception of language (see Section 2.1), this difference

[7]It is worth stressing that this sentence is an intentional example of semantic senseless, whose meaning (either literal or figurative) is supposed to not exist; i.e. it does not intend to communicate anything at all, except nonsense. However, here it is used to precisely exemplify the nonsensical effect produced by figurative contents. Most of them, finally, are senseless on their own, and need a pragmatic anchor to correctly interpret their meanings.

[8]Cf. Fillmore [49] about the main mechanisms of his Frame Semantics Theory.

would be motivated by the need of maximizing a communicative success (cf. Sperber and Wilson [156]). This need would be then the element that will determine what type of information has to be profiled. If a literal meaning is profiled, then certain intention will permeate the statement. This intention will find a linguistic repercussion by selecting some words or syntactic structures to successfully communicate what it is intended. In contrast, if the figurative meaning is profiled, then the intention will guide the choice of others elements to ensure the right transmission of its content. It is likely that such content cannot be accomplished, but in this case, the failure will not lay on the speaker's intention, rather, on the hearer's skills to interpret what it is communicated figuratively[9]. Let us observe the following sentence to clarify this point.

5) "The rainbow is an arc of colored light in the sky caused by refraction of the sun's rays by rain" (cf. WordNet [104] v. 3.0).

6) "The rainbow is a promise in the sky".

Whereas in example 5 the intention is to describe what a rainbow is, in example 6 the intention is to communicate a veiled meaning, motivated and understandable by a specific conceptual frame. In each statement the speaker has a communicative need, which is solved by maximizing certain elements. Thus, in the first example, the communicative success is based on making a precise description of a rainbow (note that all the words in this context are very clear in terms of their semantic meaning), whereas in the second, is based on deliberately selecting elements that entail secondary and nonliteral relations: [rainbow - promise], [promise - sky].

[9]In this respect, Moreno [108] points out that this need of communicative success, that Sperber and Wilson discuss in their work, relies on assumptions about how the human mind has evolved in the direction of increasing efficiency and is now set up in such a way that its attention and cognitive resources tend to be automatically directed to the processing of information which seems relevant at the time. This relevance-driven processing of stimuli [continues] is exploited in human communication and comprehension where the hearers investment of effort, attention and cognitive resources is oriented to deriving the interpretation that the speaker intended to convey.

2. FIGURATIVE LANGUAGE

Once argued that intentionality is one of the most important mechanisms to differentiate literal from figurative language[10], it is worth noting that language on its own provides specific linguistic devices to deliberately express different types of figurative contents: metaphor, allegory, irony, similes, analogy, and so on. These devices will be discussed and exemplified later in this chapter. This section is concluded by stating some precisions regarding the concepts described so far.

Figurative language refers to the use of linguistic elements (words, phrases, sentences) to intentionally deviate the literal language symmetry. Conceptual.

Figurative meaning refers to the result of figuring out the secondary meaning, and then interpreting it within a specific frame. Processual.

Frame refers to the notion of context: linguistic in terms of semantics, and social in terms of pragmatics.

Semantics refers to direct meanings, situationally and discursively independent.

Pragmatics refers to indirect meanings, situationally and discursively dependent.

Finally, hereafter when the terms literal language and figurative language appear in the document, they will denote both conceptual as processual aspects, unless the notion of meaning, either literal or figurative, appears to specify their use. With respect to the term frame, this will be arbitrarily used instead of the term context, and vice versa.

[10] Despite the characteristics of each type of language, the concept of intentionality will be one of the most important elements to later explain how we tackle the problem of figurative language processing; in particular, when the different evaluation corpora will be described.

2.4 Figurative Devices

A figurative device is part of a major class commonly known as figure of speech. These figures are linguistic statements in which one, or various of their constituents, deviate(s) its/their literal meaning in favor of a figurative interpretation. There are two main categories in which they can be classified: tropes and schemes. The former are devices with an unexpected twist in the meaning of words, as opposed to the latter, which only deal with patterns of words[11]. Some examples of schemes are parallelism (similarity by virtue of corresponding), antithesis (juxtaposition of contrasting words or ideas), ellipsis (omission of words), alliteration (sound that is repeated to cause the effect of rhyme). In the case of tropes, some devices are similes (when something is like something else), puns (play of words with funny effects), irony (opposition between what it is decoded and what actually must be interpreted), or oxymoron (use of contradictory words).

While it is true that all these devices communicate figurative meanings, it is also true that not all of them are ordinarily used by people. Some of them are more circumscribed to literary usages. For instance, puns are more common in ordinary natural language scenarios than erotemas (rhetorical questions), or hypophoras (answering rhetorical questions), which are more related to literary scenarios.

It is worth stressing that this distinction (which is not a general rule) is based on our goal of figurative language processing in social media examples (i.e. ordinary and colloquial statements); rather than literary examples of figurative language (see 1.1).

In the following sections are thus listed and exemplified some of these figurative devices. The list is not exhaustive, and does not intend to be. Rather, it intends to summarize the figures of speech based on the previous operational distinction. Devices such as onomatopoeia (words that sound like their meaning), oxymoron (terms that normally contradict each other), parable (extended

[11] Since the main property of schemes relies on word order (e.g. anaphora, cataphora, climax, hyperbaton, etc.), and our interest is more related to word meaning, they will not be deemed in the rest of the document.

metaphor told as an anecdote to illustrate or teach a moral lesson), paradox (use of contradictory ideas to point out some underlying truth), and many others more, are not listed here for being consistent with this distinction. Nonetheless, in Appendix A, a list with some of these devices , and their corresponding definition and exemplification, is given.

2.4.1 Metaphor

The simplest definition of this device is related to the concept of comparison. Such comparison, either explicit or implicit, is literally false (Katz et al. [76]). For instance, according to Feldman and Peng [48], metaphors such as "that flat tire cost me an hour", "you need to budget your time", or "don't spend too (much/little) time", are examples of comparisons in which a concept (time) is implicitly compared to other one (money). Most times, these devices are used by people for conceptualizing *abstract* concepts in terms of the apprehendable, or to express ideas that are inexpressible by literal language (Cacciari [26], Ortony [111]). To be able to link both conceptual realities, it should exist a systematic set of correspondences between the constituent elements of these two domains (Lönneker-Rodman and Narayanan [89]). However, this is not a straightforward process. For instance, in their famous work *Metaphors we live by*, Lakoff and Johnson [85] point out that although most people consider metaphors as devices of the poetic imagination and the rhetorical flourish, they belong to ordinary language, and accordingly, they are systematic. Metaphorical expressions [continue] are partially structured in a systematic way; therefore they can be extended in some ways but not others. Thus, the metaphors in which time is compared to money, it does not mean that money can be always used instead of the concept of time. Following authors' example, if you spend your time trying to do something and it does not work, you cannot get your time back. There are no time banks.

In addition, various authors consider that metaphors are based on cognitive procedures (see [85, 155, 87, 61, 115]) which make them possible to communicate concepts in a cost effective way by exploiting semantic relatedness (Peters [118]). Furthermore, according to Pierce et al. [121], some studies of figurative language processing have shown that metaphorical meanings are automatically

activated by a semantic link. Therefore, when these metaphors are produced in our native language, we are capable of understanding them effortlessly (Saygin [147]). However, according to Veale and Hao [171], metaphors can sometimes be so enigmatic and so challenging (due they convey elaborate meanings) that can only be interpreted by listeners that share the appropriate conceptual frame. For instance, *women are dangerous things* entails a metaphoric (and metonymic[12]) relation which is fully, and truly, understandable only by knowing how the speakers of Dyirbal[13] categorize all the objects of the world in terms of associations motivated by their reciprocal properties (cf. Lakoff [84]).

Finally, metaphors can also be classified into different fine-grained categories. Some of them are: *discourse* metaphors which are verbal expressions containing a construction that evokes an analogy negotiated in the discourse community (Zinken [185]). *Orientational* metaphors which are comparison that are not given in terms of another concept but instead organize a whole system of concepts with respect to one another (Lakoff and Johnson [85]). *Ontological* metaphors which are ways of looking at events, activities, emotions, ideas, etc., as entities and substances (Lakoff and Johnson [85]).

2.4.2 Metonym

The concept of metonym is related to the process of representing the whole for the part (Marschark [96]), as well as to the process of using one entity instead of another one ([89]). For instance, the word *university* can mean a building, the organization that is located in that building, and the people working for that organization ([118]).

Like metaphor, metonym is one of the basic characteristics of cognition. This is due to people usually take one well-understood or easy-to-perceive aspect of something and use it to stand either for the thing as a whole or for some other aspect or part of it (Lakoff [84]). Let us consider Lakoff and Johnson' example: "the ham sandwich just spilled beer all over himself ([85])". This sentence is

[12]This concept will be explained in Section 2.4.2

[13]This is an Australian aboriginal language spoken in northeast Queensland by the Dyirbal tribe. Information taken from http://en.wikipedia.org/wiki/Dyirbal_language.

understandable within a frame in which the entity *ham sandwich* stands for the one who asked that food; i.e. the *customer*.

It is worth noting that, unlike metaphor (where the concepts do not share a common domain), metonym requires that the related entities to share the same conceptual domain ([89]). This issue is addressed by Bergen [15] when remarks that in most metonymies, a word or set of words that identify a referent (the trigger) are used to identify a second referent (the target) that is pragmatically related in some way to that first referent. In addition, Papafragou [115] argues that metonymy is not so much a mapping between concepts as a novel way of referring to an external entity. In this respect, Pexman notes that metonym may also function as a type of cultural shorthand, allowing speakers to present themselves as witty and quick thinking, thereby acquiring a social function ([120]).

2.4.3 Simile

Unlike metaphors, similes are defined as direct comparisons (Glucksberg and McGlone [59]) in which one thing is like another different thing. Veale and Hao [171] explain that similes are not categorization statements, in terms of they do not share common properties to map the comparison through a well-defined link. Rather, such link must be inferred from our knowledge of the external world. Thus, their example "a wedding is like a funeral", is interpretable (funnily) by mapping some salient properties of *funeral* that can be applicable to *wedding*, so that the sentence makes sense after figuring out the fact that weddings occur in a (solemn) church, and are sometimes forced (unfortunately) for non-romantic (sad) reasons.

Moreover, according to Veale and Hao [172], simile is widely viewed as a less sophisticated conceptual device than metaphor, not least because similes are explicitly marked and are frequently more obvious about the meanings they carry[14]. Nonetheless, this *naivete* that underlies similes makes them suitable elements for acquiring the category-specific knowledge required to understand metaphor ([172]). For instance, Marschark [96] supports this assumption by demonstrating

[14]Situation that does not always happen. It is sufficient to recall the previous example to realize that similes are not so transparent and direct as one could a priori think.

how similes are fairly obvious regarding their use in sing language, where this obviousness is the basis to build more complex sings.

Usually, the similes are identified by the presence of discursive markers such as *like*, *as*, or *than*. For instance: "he fights *like* a lion", or "he was *as* brave *as* a lion in the fight".

2.4.4 Idioms

Traditionally, idioms have been defined as nonliteral statements whose meaning cannot be derived from the meanings of their individual compositional parts; however, they are conventionally accepted, and thus, pragmatically interpretable (see [48, 74, 15]). For instance, "spilled the beans" does not mean to slop beans, nor to reveal information about beans, but telling a secret; i.e. the interpretation is given by placing what it is decoded within a frame which makes it sense.

In addition, Deg and Bestgen [44] talk about three important properties of idioms: i) a sequence with literal meaning has many neighbors, whereas a figurative one has few; ii) idiomatic expressions should demonstrate low semantic proximity between the words composing them; and iii) idiomatic expressions should demonstrate low semantic proximity between the expression and the preceding and subsequent segments.

On the other hand, Bergen [15] notes that while idiomaticity contributes to figurativeness, it is not uniquely constitutive of it. In this respect, Glucksberg and McGlone [59] argue that idioms do not consist of a single type of expression but instead vary systematically from simple phrases such as by and large, to metaphors. Likewise, Moreno [107] suggests that most idioms lie along that continuum of looseness and as a result they vary in the extent to which the overall idiomatic meaning can be inferred from the meanings of the parts and their degree of transparency. The consequence is that people understand idioms but are not capable to find the path to explain where their referent, or conventional motivation, is.

Finally, Glucksberg and McGlone [59] stress that idioms are set apart from most other fixed expressions here described due to the absence of any discernable

relation between their linguistic meanings and their idiomatic meanings. Therefore, unfamiliar idioms are, in essence, no idiomatic; i.e. people will attempt to interpret them only compositionally (Katz et al. [76]).

2.5 Figurative Devices in this Thesis

While it is true that our general objective relies on figurative language processing, the scope of this thesis is circumscribed to two specific devices: humor and irony (see 1.2). In this section, therefore, both devices will be described in detail.

It is worth noting that devices such as pun or sarcasm, although are closely related to humor and irony, will not be described independently. Rather, they will be defined in terms of fine-grained ways of expressing both humor as irony, respectively. Furthermore, it is important to keep in mind that the devices here described, as well as the ones included in Appendix A, are not mutually exclusive, nor exclusive with respect to humor and irony; i.e. a metaphoric statement, for instance, does not exclude an ironic interpretation, nor an simile excludes a funny one.

2.5.1 Humor: A Multidimensional Phenomenon

One of the characteristics that defines us as human beings and social entities is a very complex, as well as very common concept: humor. This concept, which we could simply define by the presence of amusing effects, such as laughter or well-being sensations, plays a relevant role in our lives. Its function as a mechanism to release emotions, sentiments or feelings, impacts positively on human health. Furthermore, its cathartic properties, in a social context, make most people react to a humorous stimulus regardless of their beliefs, social status or cultural differences. Moreover, by means of analyzing its effects, humor provides valuable information related to linguistic, psychological, neurological and sociological phenomena. However, given its complex nature, humor is still an undefined phenomenon. Partly, because the stimuli that make people laugh cannot be generalized and formalized. Cognitive aspects as well as cultural knowledge, for instance, are some of the multi-factorial variables that should be analyzed in order to comprehend humor's underlying properties. Nonetheless, disciplines

such as philosophy (see Halliwell [65]), linguistics (see Attardo [5], Raskin [127]), psychology (see Ruch [143]), or sociology (see Hertzler [69]), have attempted to study humor in order to provide formal insights to explain humor's basic features. For instance, from a psychological point of view, Ruch [143] has analyzed the link between personality and humor appreciation, providing interesting observations about this property and the kind of necessary stimuli to produce a response. Some linguistic studies, on the other hand, have explained humor by means of semantic and pragmatic patterns. Attardo [5, 6] attempts to explain verbal humor[15] as a phenomenon which supposes the presence of some *knowledge resources*, such as language, narrative strategies, target, situation, logical mechanisms or opposition, to produce a funny effect. From a sociological point of view, cultural patterns are ones of the most studied features regarding humor appreciation. Hertzler [69] stresses the importance of analyzing the cultural background to be able to conceptualize humor as an entire phenomenon.

In addition of these disciplines, humor has been explained by means of several theories (Schmidt and Williams [148], Mihalcea [97]). According to Valitutti [169], they can be classified into three main classes[16]:

i. Superiority Theory. Based on the assumption that funniness is caused by the misfortunes of others. This fact reflects superiority. Some of the authors that support this theory are Plato, Aristotle, and Hobbes.

ii. Relief Theory. Base on psychological and physiological assumptions regarding the nature of humor, and how it impacts on our lives by releasing physic energy. Authors such as Freud, Mindess, and Fry, represent this approach.

iii. Incongruity Theory. The more linguistic theory. Based on the assumption that humor relies on incongruity, and of course, on its resolution. Schopenhauer, Attardo, and Raskin are some of its best exponents.

[15] Verbal humor, as opposed to visual, or physical humor, refers to the type of humor that is expressed linguistically.

[16] However, there are who suggest more classes: Surprise Theory (proposed by Descartes); Ambivalence Theory (proposed by Socrates); Configurational Theory (proposed by G. W. Hegel); Evolution Theory (proposed by Darwin, and supported by Institute for the Advancement of Human Behavior - IAHB); etc.

It is worth noting that not all these theories fit with the scope of this thesis. While it is true that they are intended to explain humor, it is also true that they not necessarily deal with verbal forms of communication. Let us observe the following image to clarify this point.

Figure 2.1: Example of visual humor[∓].

[∓]Image taken from http://friendsofirony.com.

In Figure 2.1 the funny effect is not given by interpreting linguistic information[17]. Rather, the effect is supported by interpreting nonverbal forms of communication within a specific frame.

Unlike these types of humor, verbal humor is defined in terms of linguistic ways of expression (see [9, 127, 5, 8, 161, 102]). This is the type of humor, therefore, that underlies our investigation (hereafter, when speaking about verbal humor, we will do it by referring only to humor).

In this respect, the Semantic Script Theory of Humour (Raskin [127]) first, and then, the General Theory of Verbal Humor (Attardo and Raskin [9], Raskin [127], Attardo [5, 8]), have described basic mechanisms of humor based on linguistic arguments (mostly semantic and pragmatic). According to their creators,

[17]Note that we do not talk about how funny the effect is; i.e. funniness is neither quantified nor qualified.

humor can be (broadly) interpretable based on the following features: *script opposition*, *incongruity* and *resolution*, *situation, target, genre*, and *language*. In addition, Nilsen [109] suggests that humor deals with features such as *ambiguity, exaggeration, understatement, hostility*, *incongruity* or *irony*, *situation-insight, sudden insight, superiority*, *surprise* or *shock*, *trick* or *twist*, and *word play*. Some of these features are illustrated in examples 7 to 11.

7) "I'm on a thirty day diet. So far, I have lost 15 days" (opposition, incongruity).

8) "Change is inevitable, except from a vending machine" (ambiguity).

9) "The sex was so good that even the neighbours had a cigarette" (language, exaggeration).

10) "Drugs may lead to nowhere, but at least it's a scenic route" (twist).

11) "I've got the body of a god ... unfortunately its Buddha" (incongruity, irony).

According to these approaches, humor is thus more likely to appear if some of these features are fulfilled, or are communicatively violated. Furthermore, based on such criteria, they have classified the ways of expressing humor into these features. The punchlines, for instance, are supposed to trigger script opposition (Attardo and Raskin [9]),

On the other hand, with respect to figurative language, humor is deemed as an ideal vehicle to lead figurative contents. Katz et al. [76], for instance, note that nonliteral sense is a key element to produce humorous effects. Veale and Hao [172] in turn, stress that some similes hinge on a new, humorous sense when interpreting them, as in "as fruitless as a butcher-shop", and "as pointless as a beach-ball". Allen [1] also reports this link between similes and humor when analyzes these devices and laughter in Don Quixote. However, irony and sarcasm are the devices that seem to be closer to humor. For instance, Gibbs and Izett [56], as well as Colston [34], point out that people use irony to achieve a complex set of social and communicative goals, being humorous one of the most important. Pexman [120], in turn, indicates that men are more likely than women to perceive humor when the stimulus is given by means of sarcasm or irony.

Finally, throughout the following chapters we will understand this device according to the following criterion.

Humor will be reduced to its most constricting conception (which is not so narrow at all); i.e. not as the general concept that impacts on different aspects of our lives (such as we initially defined it), rather, as a linguistic device that takes advantage of different resources (mostly related to figurative usages) to produce a specific effect: laughter.

2.5.2 Irony: A Veiled Phenomenon

Like most figurative devices, irony is difficult to pin down in formal terms, and no single definition ever seems entirely satisfactory. So to begin with, let us consider three obvious examples of irony in everyday situations:

12) Going to your car in the morning, you notice that one of your tires is completely flat. A friendly neighbor chimes in with: "Looks like you've got a flat". Marveling at his powers of observation, you reply "Ya think?".

13) A man goes through the entrance to a building but fails to hold the door for the woman right behind him, even though she is visibly struggling with a heavy box. She says "Thank You! anyway".

14) A professor explains and re-explains Hegel's theory of the State to his class of undergraduates. "Is it clear now", he asks. "Clear as mud", a student replies.

15) After seeing a stereotyped romantic movie, the guy says: "I never believed love at first site was possible until I saw this film".

These examples suggest that pretense plays a key role in irony: speakers craft utterances in spite of what has just happened, not because of it. The pretense in each case alludes to, or echoes, an expectation that has been violated (cf. Clark and Gerrig [32], Sperber and Wilson [155]), such as the expectation that others behave in a civil fashion (example 12), or speak meaningfully and with clarity (example 14). This pretense may seem roundabout and illogical, but it offers a sharply effective and concise mode of communication. Irony allows a speaker

to highlight the expectation that has been violated while simultaneously poking fun at, and often rebuking, the violator. Additionally, an underlying sensation of false message (or *negation* of what it is expressed) permeates what must be interpreted (e.g. "thank you!" instead of "fuck you" (example 13)).

Now, beyond these examples, experts point out that irony is essentially a communicative act that expresses an opposite meaning of what was literally said (Wilson and Sperber [181]). However, this is only one type of irony. In the specialized literature we found two primaries types of irony: *verbal irony* and *situational irony*.

Verbal irony is a playful use of language in which a speaker implies the opposite of what is literally said (Curcó [39], Colston and Gibbs [36]); i.e. a type of indirect negation (Giora [57]); or expresses a sentiment in direct opposition to what is actually believed, as when Raymond Chandler in *Farewell, My Lovely* describes Moose Malloy as "about as inconspicuous as a tarantula on a slice of angel food". According to some pragmatic frameworks[18], certain authors focus on fine-grained properties of this concept to correctly determine whether a statement is ironic or not. For instance, Grice [63] requires that an utterance intentionally violate a conversational maxim if it is to be judged ironic. Wilson and Sperber [181] assume that verbal irony must be understood as echoic, that is, they argue that irony deliberately blurs the distinction between use and mention. Utsumi [168] suggests that an ironic environment, which establishes a negative emotional attitude, is a prerequisite for considering an utterance as ironic.

Situational irony, in contrast, is an unexpected or incongruous quality in a situation or event (cf. Lucariello [91]), such as a no-smoking sign in an ashtray (see Figure 2.1), or a vegetarian having a heart-attack outside a McDonald's. Moreover, other authors distinguish fine-grained types of ironies: dramatic irony (Attardo [7]); discourse irony (Kumon-Nakamura et al. [83]); tragic irony (Colston [35]); etc. Here, like with humor, our scope is limited to verbal irony, but we do not reject the possibility to apply our linguistic model to situational irony, not least because much of the irony in our data sets exhibits precisely this type of irony.

[18]Unlike humor, most studies regarding either verbal irony or situational irony are focused on linguistic approaches.

Despite irony seems to be clearly defined in terms of its specialized linguistic properties, such properties are rarely observed when common people use this device. In this respect, people have their own concept of irony, which seldom matches with the properties suggested by the experts. Instead, it is mixed with other concepts. Let us consider the following examples:

16) "I feel so miserable without you, it's almost like having you here".

17) "Don't worry about what people think. They don't do it very often".

18) "Sometimes I need what only you can provide: your absence."

19) "I am giving this product [a t-shirt] 5 stars because not everyone out there is a ladies' man. In the hands of lesser beings, it can help you find love. In the hands of a playa like me, it can only break hearts. That's why I say use with caution. I am passing the torch onto you, be careful out there folks."

According to some examples of irony given by different people, examples 16 to 19 could be either ironic, or sarcastic, or even satiric. In these examples, irony (if they are really ironic) is perceived as a mixture of sarcasm and satire, whose effect is not only based on expressing an opposite meaning, but a humorous one as well. However, beyond the fact of what device better represents each example, we want to highlight the fact that, for many people, there is not a clear distinction with respect to the boundaries for differentiating between irony and other related devices (e.g. sarcasm). For instance, we could note that several of the above examples might be both ironic as sarcastic, (e.g. "clear as mud" (example 14), or "people do not do think very often" (example 17)). Nonetheless, theoretically speaking, we can argue that irony tends to be a more sophisticated mode of communication than sarcasm: whereas the former often emphasizes a playful pretense, the latter is more often concerned with biting delivery and savage put-downs. Thus, while irony courts ambiguity and often exhibits great subtlety, sarcasm is delivered with a cutting or withering tone that is rarely ambiguous. However, these differences rely indeed on matters of usage, tone, and obviousness, rather than only on theoretical assumptions.

In this respect, even the experts do not clearly define the boundaries among these devices. For instance, Colston [35] and Davidov et al. [41], consider that

sarcasm is a term commonly used to describe an expression of verbal irony; whereas Gibbs [54] argues that sarcasm along with jocularity, hyperbole, rhetorical questions, and understatement, are only types of irony. In contrast, Kumon-Nakamura et al. [83] talk about a type of sarcastic irony which is opposed to the non sarcastic one; while Attardo [7] stresses that sarcasm is just an overtly aggressive type of irony. Moreover, Burfoot and Baldwin [24] suggest that satirical texts, specifically news articles, tend to incorporating irony and non sequitur in an attempt to provide a humorous effect; whereas Gibbs and Colston [55] indicate that irony is usually compared to satire and parody.

In accordance with these statements, it is obvious how the boundaries among these figurative devices are not clearly differentiable. Therefore, in this thesis and according to our objective, we will understand irony in the following terms.

Irony is a verbal expression whose formal constituents, i.e. words, attempt to communicate an underlying meaning which is opposite to the one expressed. In addition, we differentiate between *aim* and *effect*. The aim of irony, according to our definition, is to communicate the opposite of what is literally said; whereas the effect may be a sarcastic, satiric, or even funny interpretation that undoubtedly profiles negative connotations.

In this context, it is convenient to treat irony and related devices as different facets of the same phenomenon. Therefore, devices such as sarcasm, satire, hyperbole, or litotes, will be deemed as specific extensions of a general and broad concept of irony.

2.6 Summary

In this chapter we have treated three important issues. First, the importance of considering language as a dynamic system, rather than a static one. By analyzing language in these terms, several phenomena can be understood and explained in a comprehensive way (Section 2.1). In particular, the types of phenomena that we treat in this thesis: humor and irony.

Then, in Sections 2.2 and 2.3, similarities and differences regarding two specific linguistic realities (literal and figurative) were described. Both literal language as figurative language were described in order to lay the linguistic foundations that

differentiate both types of languages. Furthermore, in Section 2.4, examples of figurative devices (metaphor, metonym, similes, and idioms) were given.

Finally, the devices that support this thesis were treated in Section 2.5. Humor and irony were described and exemplified along Sections 2.5.1 and 2.5.2, respectively. Their definitions, as well as the ones of other important concepts, were given along this chapter as well.

3

Figurative Language Processing

For example, if the user has asked the agent to contact "John", and there are several Johns to which the user might be referring, the agent might respond: "Do you want John 'not today' Bannerman or John 'beers on Friday?' Smith?"

BINSTED [16]

This chapter will be focused on introducing the **state-of-the-art** with respect to figurative language processing. We will outline challenges, as well as benefits of considering the inclusion of figurative devices concerning a computational framework. Furthermore, we will broadly exemplify some tasks in which the automatic processing of figurative devices is involved. Finally, we will describe the related work regarding the automatic processing of humor and irony.

3.1 Natural Language Processing

This thesis relies on Natural Language Processing (NLP). Broadly speaking, this field is intended to cover any type of computer manipulation of natural language (Bird et al. [20]), regardless of whether it is spoken or written. In accordance

with some points of view, different but overlapping fields converge in this concept: computational linguistics in linguistics, NLP in computer science, speech recognition in electrical engineering, and computational psycholinguistics in psychology (Jurafsky and Martin [72]). Regardless of the field in which NLP is conceptualized, its major goal is to reproduce the natural transmission of information by modeling the speaker's production and the hearer's interpretation on a suitable type of computer (Hausser [67]). This is supposed to be achieved by developing algorithms that can recover the intended meaning of a sentence or utterance based on its spoken or textual realization (Clark et al. [31]).

Moreover, since its interdisciplinary nature, NLP relies on different techniques and approaches concerned with artificial intelligence, machine learning, pattern recognition, linguistics, corpus linguistics, statistics, and so on. Thus, by using linguistic resources and applying diverse techniques, any NLP system should be capable to determine the structure of an utterance (Manning and Schütze [92]): from phonetics to speech recognition and speech understanding; from morphology and syntax to semantics and discourse.

Some of the tasks that have been investigated (with better or worse results) in NLP are: part-of-speech tagging (concerned with determining the grammatical category of a word or sequences of words); morphological segmentation (concerned with splitting words into their minimal morphological segments); parsing (concerned with representing and analyzing the syntactic structure of a sentence or phrase); speech recognition (concerned with determining the textual representation of the speech); machine translation (concerned with translating from one language to another); named entity recognition (concerned with mapping words to information beyond grammatical categories, for instance, proper names); word sense disambiguation (concerned with identifying the precise meaning of a polisemic word within a specific context); co-reference resolution (concerned with solving anaphoric and cataphoric relations within a text); question answering (concerned with giving the correct answer given a human-language question); sentiment analysis (concerned with identifying elements within a document that profile subjective or objective information, as well as positive or negative polarity).

Finally, apart from solving practical issues, many of the cited tasks, as well as many others that were not mentioned, are concerned with providing information

regarding the cognitive processes that underlie the human communication; i.e. according to Clark et al. [31], given its psycholinguistic branch, NLP should reflect how people process language.

3.2 Figurative Language Processing

Figurative Language Processing (FLP) may be deemed as a subfield of NLP in which the major goal is not only focused on modeling natural language but on finding formal elements to computationally process figurative usages of natural language. According to our definition given at the end of Section 2.3, figurative language refers to the intentional deviation of literal meaning in favor of second interpretations. This is mostly achieved by employing different devices, such as metaphor or irony, whose communicative function relies on profiling nonliteral meanings. In this respect, FLP supports its models on the analysis of specific linguistic statements which imply abstract layers of human communication to be fully, and correctly, interpreted (see Chapter 2). In this context, its target is closer to provide insights regarding how people process language, and then, how use it to communicate more elaborated linguistic realities. For instance, consider the cognitive linguistic point of view which points out towards the link between language faculty and our general cognitive processes (Langacker [86, 87]), and especially, Lakoff and Johnson's arguments regarding how figurative devices (in particular metaphor and metonym) are central mechanisms to cognition ([85]). Moreover, Sikos et al. [153] refer works in which figurative language has been studied with neurophysiological methods in order to obtain empirical information for understanding the neural basis of complex cognition.

Nonetheless, many challenges underlie FLP, both in terms of linguistic as computational aspects. Based on the arguments given in Chapter 2 is doubtless that figurative language entails more complicated scenarios than literal language: from theoretical issues to a practical consolidation. Let us consider the following examples related to sentiment analysis, to clarify this point:

20) "This movie is crap".

21) "It's not that there isn't anything positive to say about the film. There is. After 92 minutes, it ends".

One of the most difficult problems when assigning either positive or negative polarity in sentiment analysis tasks is to determine what is the truth value of a certain statement. In case of literal language (example 20) the existent techniques achieve good results; instead, in case of figurative language (example 21), the result may be a consequence of simply finding out what types of words prevail in the surface of such statement. In such cases, the same automatic techniques lose effectiveness because the profiled and real meaning is deviated from its literal interpretation, or in terms of cognitive linguistics, is in ground. This fact might be evident for humans; i.e. after processing the information of example 21, some people[1] could realize that a negative polarity permeates it due to the presence of irony. In addition, here we are taking for granted that both examples are profiling different linguistic realities; i.e. whereas example 20 is literal, example 21 is figurative. Nevertheless, this is not always a fact. This is the prototypical and most common situation that we will face in real tasks. Moreover, in absence of valuable information such as tone, gesticulations, or context, both examples might be reduced to a literal interpretation. Their labels would then depend on many variables, except on taking into account their figurative purpose[2].

Despite the challenges are huge, there are many approaches that deal with FLP. They are focused on solving particular issues: from automatically discriminating literal from figurative language, to create models for automatically detecting certain figurative devices. For instance, Bogdanova [22] bases her approach to figurative language detection on the fact that the sense of a word significantly differs from the sense of the surrounding text. To her, this is an insight about a word is used figuratively. In the same vein, Li and Sporleder [88] use gaussian mixture models in order to automatically detect figurative language. They assume that figurative language exhibits less semantic cohesive ties with the context than literal language. In turn, Rentoumi et al. [128] propose a methodology for sentiment analysis of figurative language which applies word sense disambiguation and Hidden Markov Models. By combining n-gram graphs based method,

[1] It is doubtless that not all of us are capable to identify figurative usages of language any time. As referred in the previous chapter, this faculty depends on many factors, many of them are not even related to our linguistic competence.

[2] It is obvious that the performance of any task that involves natural language cannot be always accurate. However, we think that the more knowledge can be provided, the better results can be achieved.

they assign polarity to word senses. On the other hand, Sikos et al. [153] use experimental techniques to understand the cognitive and neural mechanisms that subserve figurative interpretations. According to the authors, in order to process figurative language, the brain may be organized in such a way that the two cerebral hemispheres work in parallel, each with somewhat different priorities, competing to reach an appropriate interpretation.

Finally, it is worth noting that the fact of considering a computational approach of figurative language may be useful for several tasks. Mihalcea and Strapparava [101], for instance, note that entertainment, and especially, edutainment are perfect scenarios for automatic humor processing. Saygin [147] indicates that metaphors are efficient mechanisms to analyze how bilingual people process language. This figurative device may then be useful concerning with learning a second language, as well as concerning with machine translation tasks (Li and Sporleder [88]). Moreover, some figurative devices may impact on tasks beyond NLP. For instance, Gibbs and Izett [56] note that irony is widely employed in advertising. Sikos et al. [153], in contrast, reach more neurological boundaries by arguing that our right hemisphere plays a key role when processing figurative language.

3.3 Advances on FLP

The interest for automatically processing issues related to figurative language is not new in NLP. Some examples were given in the previous section. In this section, we will focus on presenting some of the most relevant research works related to FLP. In particular, like in Chapter 2, we will concentrate on a few devices that are closer to our objective of analyzing language in terms of social media examples, rather than of literary examples of figurative language (see 1.3) [3].

[3] Due to this thesis is focused on two specific devices (humor and irony), the devices outlined in this section are broadly described; i.e. the information given cannot be properly considered as a state-of-the-art; rather, only an overall outline regarding advances on FLP.

3.3.1 Metaphor Processing

According to the arguments given in Section 2.4.1, metaphor can be conceptualized as a simple comparison. Based on this fact, their automatic processing should be achieved effortless. However, this is not straightforward. There are many factors that must be taken into consideration when defining a model to automatically process metaphors. For instance, Veale and Hao [171] indicate that any attempt to computationally deal with metaphor should start by considering metaphor as part of our conceptual structure, as well as a way of knowledge representation. In this respect, various approaches have shown how the task of automatically processing metaphor can yield interesting results. Rentoumi et al. [128], for instance, addressed this task from a sentiment analysis point of view. They pointed out that expanded senses and metaphors can be used as expressive subjective elements since they display sentiment implicitly. Saygin [147], in turn, approached the task by analyzing the role that metaphor plays in translation. According to her results, when people translated sentences to and from their native language (Turkish) and their second language (English), upon encountering a metaphorical usage, both the underlying metaphor and the literal meaning are likely to be active in people's perception, even though it is clear from the context which meaning is intended. Whereas in Veale [170], author approached metaphor processing by means of an information retrieval task. In contrast, by analyzing metaphors in a corpus of newspaper texts, Zinken [185] noted that metaphors follow a regular pattern when the comparisons (analogies in his terms) are made. He found in his corpus that such analogies are form-specific; this means, they are bound to particular lexical items. On the other hand, Pierce et al. [121] showed that, at cognitive level, metaphor processing tends to be automatic. They argued that metaphor processing is triggered automatically by violations of semantic expectancies that cause people to consider a wider semantic neighborhood. In this context, their concept of *working memory* plays a crucial function due it speeds the process through which metaphoric meanings are automatically identified.

3.3.2 Metonym Processing

With respect to automatic metonym processing, the approaches have been scarcer. Given the common frames that metonym and metaphor share, most of researches

are focused on processing the latter. The former, however, has been tackled from points of view that involve the use of specific linguistic resources such as thesaurus, as well as corpus-based analysis.

In the case of the first point of view, the research described by Peters and Wilks [119] is focused on exploiting the taxonomic information that is explicitly present in WordNet to select instances of metonymy. By capturing semantic information, authors classify senses into groups of senses that represent a more coarse-grained, underspecified level of semantic description.. In the same vein, Peters [118] uses WordNet, especially, its hierarchical structure to infer underlying knowledge to detect metonymies. His approach was based on identifying words with systematically related senses and their glosses in order to capture a semantic relation between the senses. In contrast, Markert and Nissim [93] addressed the task of metonymy resolution by employing corpus information to be able to find empirical data, and on this basis, to discriminate between literal and metonymic usages of a word. The same authors evaluate five different methods regarding figurative language resolution in one of the shared tasks of SemEval-2007[4]. In this contest, one of the major issues was related to the importance of metonym processing beyond common examples (see Markert and Nissim [94]).

3.3.3 Similes Processing

Simile is one of the figurative devices that more commonly appears in our daily communication. However, despite this property would suppose various research works about similes[5], the reality is different. Some of the few approaches that deal with similes were performed by Veale and Hao [171], and Veale and Hao [172]. In the former, authors demonstrated that the markedness of similes allows for a large case-base of illustrative examples to be easily acquired from the web. On this basis, they presented a system that used these examples both to understand property-attribution metaphors as to generate apt metaphors for a given target on demand. In the latter, authors used Google search engine to retrieve explicit similes conforming to the pattern "as ADJ as a—an NOUN". By analyzing a large quantity of similes based on this pattern, they noted how web users

[4]For detailed information refer to http://nlp.cs.swarthmore.edu/semeval/.

[5]Since similes appear quite often, then the possibility of having or building corpora to assess models is bigger than for devices such as metonym or irony.

often use this device to express ironic content. Moreover, according to their view, this knowledge would allow a cognitive agent to gradually develop a more sophisticated understanding of irony..

3.3.4 Idioms Processing

Idiomatic expressions are considered as ad-hoc examples of the use of figurative content. According to the arguments given in Section 2.4.4, idioms have a property that makes them more suitable to be detected than other figurative devices: idioms are fixed expressions, thus, they are conventionally accepted. Unlike metaphoric, metonymic, or ironic expressions, whose ways of expression have no limits, people can effortlessly recognize when a statement has to be interpreted as an idiom; i.e. in terms of its possibilities to be verbalized, an idiomatic expression is finite.

In this context, Li and Sporleder [88] exploited this property of having a degree of syntactic and lexical fixedness to stress that such properties are useful when identifying potential idioms, for instance [they say], by employing measures of association strength between the elements of an expression. In turn, Feldman and Peng [48] addressed the task of automatic idioms identification by stating that idioms are elements which appear to be inconsistent within a representative set of data. Based on semantic criteria, authors pointed out that idiomatic expressions have low semantic proximity between the words composing them, as well as between the expression and their preceding and subsequent segments. Therefore, they are likely to be outlier within a general dataset, and accordingly, be easily identified.

In a similar vein, Deg and Bestgen [44] presented a procedure for the automatic retrieval of idiomatic expressions from large text corpora. Their procedure combined text segmentation techniques and latent semantic analysis. Although such procedure is not perfect, authors achieved a considerable reduction of data in terms of candidates to be idioms. From such reduction, they found that 20.9% of the remaining data is idiomatic, and nearly 60% is phraseological in nature; i.e. its meaning cannot be derived from the meanings of their individual components either.

3.4 Related Work on Humor Processing

In the following sections we will describe the state-of-the-art regarding the computational processing of humor and irony. With respect to the former, the specialized literature considers a subfield of NLP called computational humor. This subfield is intended to create models that can simulate and understand humor. Most of the models, on purpose, are based on a specific type of humor: verbal humor (see footnote 15 in Section 2.5.1).

Two major approaches are involved when referring computational humor: humor generation and humor recognition. Each is described below.

3.4.1 Humor Generation

The aim of humor generation is to study lexical features which could be formalized in order to simulate their patterns and generate a funny effect. One of the first approaches to automatically generate humor was described by Zrehen and Arbib [186]. Authors noted that comic effect is largely due to an alliteration effect that is discovered while the joke is read; therefore, it is possible to devise a neural network that allows the recognition of this information to generate humor. In a similar way, Binsted [17], and Binsted and Ritchie [18] showed the importance of linguistic patterns, especially phonetic and syntactic ones, for automatically generating funny instances. Example 22 illustrates some of the linguistic elements that can facilitate the humor generation:

22) "What do you use to talk to an elephant? An elly-phone".

In this example we can observe how structural features, codified through linguistic information, are used to automatically generate a text with funny connotations. Note, for instance, how $elly - phone$ has phonological similarity with telephone. Moreover, $elly - phone$ is related, phonologically and "semantically", to the word which gives its right sense: elephant. This type of funny question answering structure, called punning riddle, takes advantage of linguistic patterns in order to produce an amusing effect. Furthermore, in the research works described in [17, 18, 19], authors noted that features like these ones may be simulated by rules to automatically generate funny sentences like example 23:

23) "What do you call a depressed engine? A low-comotive".

More complex characteristics have been also studied to represent and generate funny patterns. The findings reported by Stock and Strapparava [159] demonstrated how incongruity and opposite concepts are important elements for producing funny senses. By means of combining words, which socially represent opposite referents, authors have automatically produced new funny senses for acronyms such as MIT (Massachusetts Institute of Technology):

24) "Mythical Institute of Theology".

Although, at first glance, the fact of generating humor seems to be effortless, the reality is quite different. Beyond phonological information, as well as very simple syntactic templates, the task entails the identification of supplementary information, which usually is not given along with the linguistic patterns. In this respect, according to Attardo [8], humor is not only a linguistic phenomenon, although it commonly relies on this type of knowledge. The punning riddle given by the author clearly exemplifies this fact.

25) "What do you get when you cross a mafioso with a postmodern theorist? Someone who will make you an offer you cannot understand".

Example 25 stresses that humor, even in these simple structures, is more than learning some linguistic patterns, rather, it implies the activation of different mechanisms, both linguistic as social, to be able to interpret the funniness of certain joke[6]; i.e. humor is not given only by interpreting literal meanings but by adding extra-linguistic information which gives sense to the joke (see Section 2.1). In this respect, some other researchers have provided empirical evidence to create more robust systems for generating humor. For instance, Tinholt and Nijholt [164] addressed their research by evaluating the role of cross-reference ambiguity in utterances for generating humor. According to the authors, the cross-reference ambiguity is a hint at humor, and it may be useful to automatically generate punchlines. Based on the General Theory of Verbal Humor (see Attardo [5]),

[6]In this respect, Ruch [144] highlights the importance of considering individual differences, as well as the targeted recipient when analyzing and creating humorous systems.

Hempelmann [68], in turn, proposed a method to evaluate and select phonologically possible and better imperfect puns for use in computational on-the-fly pun generation in human-computer interfaces. In contrast, Augello et al. [10] suggested a humorist conversational agent capable to generate humorist expressions, proposing to the user riddles, telling jokes, and ironically answering to the user. Although this approach suggests an important degree of inferences as well as knowledge (need to be able to recognize situations appropriate for humour; choose a suitable type of humour for the situation, including a target if necessary; and generate an appropriately humorous piece of text (Binsted [16])), some researchers are focused on providing extra-linguistic knowledge in order to generate better instances of humor. Valitutti [169], for instance, stressed the importance of considering affective information in humor generation. He developed a pun generator that took advantage of affect-based verbal humorous expressions to achieve its goal.

3.4.2 Humor Recognition

The aim of humor recognition is, from the analysis of linguistic and extra-linguistic information, to identify triggers of humor that can be learned in order to automatically discriminate a funny instance from a *serious* one. By applying machine learning and pattern recognition techniques, as well as by using linguistic resources, the scientific community has approached the challenge of automatically recognizing humor with encouraging results (see Mihalcea [97]). In this respect, most of investigation is focused on the analysis of particular funny structures: one-liners. These short structures are syntactically very simple, so the humorous effect relies on more complex features. Consider example 26:

26) "Infants don't enjoy infancy like adults do adultery".

This one-liner contains phonological information which helps to produce humor, but this is not everything. There is also a pun that plays an oppositional role between the meaning of the words. Together, they produce a funny effect. These types of surface elements have provided evidence for characterizing humor in terms of formal components (which may automatically be recognized). For instance, Mihalcea and Strapparava [101, 102] applied machine learning techniques to identify humorous patterns in one-liners. Some of the elements they reported

are alliteration, antonymy or adult slang. In addition, they suggested semantic spaces which are triggers of humor: human centric vocabulary (example 27), negative orientation (example 28), and professional communities (example 29):

27) "Of all the things **I** lost, **I** miss **my** mind the most".

28) "Money **can't** buy your friends, but you do get a better class of **enemy**".

29) "It was so cold last winter that I saw a **lawyer** with his hands in his own pockets".

Furthermore, taking advantage of phonological information, Purandare and Litman [125] approached this task by analyzing humorous spoken conversations from the TV show *Friends*. They labeled all the utterances followed by laughs as humorous, and then examined their prosodic information to establish a schema to recognize humor. The work of Sjöbergh and Araki [154], on the other hand, is focused on finding patterns in syntactic and semantic layers. According to their results, devices such as similarity, style or idiomatic expressions, are sources in which humor tends to appear. The research carried out by Buscaldi and Rosso [25] also pointed in this direction. By employing features such as n-grams, bag of words, or sentence length, they noted that it is possible to discriminate humorous from non humorous sentences with acceptable accuracy (∼80%). Stark et al. [158], in turn, developed a model that exploits incongruity to produce the funny effect. Their model is based on two main concepts: the connector (part of the set-up of the one-liner) and the disjunctor (the punch line). According to their results, the system is capable to select the best disjunctor from a list of alternatives, and mainly, such disjunctor agrees with human judgements In addition, Mihalcea et al. [103] also explored how incongruity resolution can improve the humor recognition models. By applying several measures of semantic relatedness, along with a various joke-specific features, authors achieved interesting results in the task.

When considering bigger structures such as news articles or blogs, the research described by Mihalcea and Pulman [98] evidenced how negative polarity plays a very important role when characterizing humor; whereas Taylor and Mazlack [163] indicated that it is possible to recognize jokes based on statistical language recognition techniques; especially when their syntactic structure is quite similar (e.g. punning riddles, or knock knock-based jokes). Friedland and Allan [51], in

contrast, based their model on information retrieval assumptions. They proposed the tasks of joke retrieval as a domain where standard language models may fail. Therefore, authors exploited the structure of jokes to develop two domain-specific alternatives to retrieve the jokes: 1) selecting the punch line; 2) interchanging their words, if and only if, they belong to a same category (e.g. countries, professions). On this basis, the set of elements to identify a joke clearly increases.

Last, but not least, in accordance with the conclusions stated in these cited research works, incongruity, idiomatic expressions, common sense knowledge, ambiguity and irony, are sources to investigate, beyond surface information, deeper characteristics of humor.

3.5 Related Work on Irony Processing

To begin with, it is apt to cite what Aristotle thought about irony: both in speaking or writing, irony is a sign of sophistication, at the very least in the use and understanding of language. In addition, Gibbs and Izett [56] considered that irony is inherently elitist in setting apart an elite (one who understands and employs irony), from the masses (those who neither use nor understand irony). Now then, in a NLP context, irony is one of the figurative devices that more interest is causing on the scientific community. This is due to irony, in addition to the foregoing, represents a source of valuable knowledge to be exploited in different tasks. Based on our definition of irony, and especially, with respect to our dichotomy *aim-effect* (see Section 2.5.2), consider, for instance, how the funny, critical, or persuasive *effect* that is produced by ironic contents could be addressed to tasks as diverse as advertising, forum management, online marketing or product tracking.

However, the challenges that irony supposes are huge. For instance, Katz et al. [76] advert that irony tends to be more difficult to comprehend than metaphor because irony requires the ability to recognize, at least, a second-order meta-representation. Thus, if irony entails more complex meta-representational reasoning to be correctly interpreted, then, its automatic processing is still far away to be achieved. In the following section, nevertheless, we will describe some of the approaches that have dealt with this amazing figurative device.

Before proceeding, it is worth noting that, usually talking, irony is closely

related to devices such as sarcasm, satire or litotes. However, according to our aims, they are here deemed as different facets of the same phenomenon. Therefore, the research works that have approached fine-grained issues regarding their automatic processing will be also described.

3.5.1 Irony Detection

The computational approaches which deal with more abstract uses of figurative language, such as irony, tend to be more restricted, and scarcer, than the ones regarding humor. Nonetheless, they are current hot topics in NLP due to the advances in fields such as sentiment analysis and opinion mining, as well as the prevalence of irony in online texts and social media. In this respect, one of the first computational attempts to formalize irony was described by Utsumi [168]. However, his model was too abstract to represent irony beyond the ambit of an idealized hearer-listener interaction. More recently, from the perspective of computational creativity, Veale and Hao [173] have attempted to throw light on the cognitive processes that underlie verbal irony. By analyzing a large quantity of humorous similes of the form "as X as Y" from the web, authors noted how web users often use figurative comparisons as a mechanism to express ironic opinions. In addition, Carvalho et al. [27] determined some clues for automatically identifying ironic sentences. Such clues were based on the fact of detecting emoticons, onomatopoeic expressions, as well as punctuation and quotation marks. Based on this simple approach, authors achieved interesting results in the task. Veale and Hao [174], in turn, recently presented a linguistic approach to separating irony from non-irony in figurative comparisons. In this research work, authors demonstrated how the presence of ironic markers like "about" can make rule-based categorization of ironic statements a practical reality, at least in the case of similes, and described a system of linguistically-coded heuristics for performing this categorization. Finally, in a framework of computational generation of resonant expressions, Hao and Veale [66] conducted various experiments over a corpus of ironic similes in which authors found that most of these ironic comparisons use a ground with positive sentiment to impart a negative view (∼70%). Their insights to detect irony in these figurative devices, actually, were implemented in a creative information retrieval system that is available on the web.

3.5.2 Sarcasm and Satire Detection

Although at first glance the terms irony, sarcasm and satire seem to be concepts perfectly distinguishable each other, when they are used in real communicative scenarios, such perfection is rarely accomplished. Examples 16 to 19 in the last chapter clearly illustrate this point. Despite these devices, according to the specialized literature, have their own characteristics, these are not discriminating enough to guarantee there is not overlapping when they are used by non-experts speakers. For instance, Katz [75] states that sarcasm, but not irony in general, involves the ridicule of a specific person or group of people. However, in the cited examples, this property seems to be equally important in the four examples. Furthermore, like in humor, these examples take advantage of unexpected situations to convey their meaning. This is clearer in examples 16 and 18, where the expected ending in both examples, given the initial chunks, would suggest a different and "sweeter" final.

These facts highlight people's perception with respect to the fuzzy boundaries to conceptually separate these devices, and thus, their daily uses: they seem not to exist. Where does irony end, and where does sarcasm (or satire) begin? It could be argued, for instance, that irony courts ambiguity and often exhibits great subtlety, whereas sarcasm is delivered with a cutting or withering tone that is rarely ambiguous. However, in the end, these fine-grained differences are not taken into account by people. Beyond subtle and fine-grained features, people have their own concept of these figurative devices, which likely do not match with the ones suggested by the experts (mostly when people have to use them in non-prototypical scenarios). Therefore, instead of facing *pure* examples of irony, for instance, what we will finally face it will be a mixture of expressions pretending to be ironic, but being sarcastic, satiric, or even humorous[7].

Despite the conceptual and pragmatic problems that these facts suppose, there are a few approaches that are directly focused on sarcasm and satire rather than on irony. With respect to satire, Burfoot and Baldwin [24] approached the task of automatically classifying satirical newswire articles. By means of lexical and semantics features, represented by headlines, profanity (offensive language), and

[7]Here is where our dichotomy *aim-effect* plays its role: by stating that irony has an aim, and this aim causes an effect; it is easier to categorize sarcastic or satiric expressions just like ironic expressions with a particular effect (or reading).

slang, they could separate satirical from "true" (sic) newswire articles achieving, with their best score, a F-measure of 0.798. Regarding the former, the approaches to automatically detect sarcasm are a little bit broader. On one hand, Tsur et al. [166], as well as Davidov et al. [41], addressed their research in order to find elements to automatically detect sarcasm in online products reviews and tweets, respectively. Based on a semi-supervised approach, they suggested surface features such as content words (words regarding information about the product, company, title), frequent words, or punctuation marks, to represent sarcastic texts. According to their results, the achieved F-measure scores are significantly positive (0.788 and 0.827, respectively). On the other hand, González-Ibáñez et al. [62] reported a method for identifying sarcasm in Twitter. Authors investigated the impact of lexical and pragmatic features on the task. However, according to their results, neither the human judges nor the machine learning techniques performed very well.

Finally, although these approaches have demonstrated that both humor and irony can be handled in terms of computational means, it is necessary to improve the mechanisms to represent their characteristics, and especially, to create a feature model capable to symbolize, the less theoretical as possible, both linguistic and social knowledge in order to describe deeper and more general properties of these phenomena. For instance, most of results here described are text-specific; i.e. they are centered either on one-liners, punning riddles, or on similes, newswire articles, or products reviews; thus, their scope regarding different instances in which figurative language appears, might be limited. Therefore, part of our objective is to identify salient components, for both humor as irony, by means of formal linguistic arguments (i.e. words and sequences of them), in order to gather a set of more general attributes to characterize these figurative devices.

3.6 Summary

In this chapter we have first established the framework in which this thesis is developed. In Section 3.1 we introduced some basic concepts regarding this framework. In addition, in Section 3.2 we outlined the challenges that underlie any computational treatment of figurative language, as well as some of the possible tasks in which figurative language could be applied.

In Section 3.3 we broadly described some of the research works that have dealt with figurative language processing. Basically, we focused on citing some approaches that have investigated the automatic processing of devices such as metaphor, metonym, similes, and idioms.

Finally, in Section 3.4 and 3.5 we exposed the state-of-the-art regarding the computational treatment of humor and irony, respectively.

4

Automatic Humor Recognition

Children in the back seats of cars cause accidents, but accidents in the back seats of cars cause children.

DATA SET H2
MIHALCEA AND STRAPPARAVA [99]

This chapter will be focused on describing our **Humor Recognition Model (HRM)**. First, our initial assumptions will be introduced. Then, experiments and results will be presented. In addition, we will introduce the different data sets in which we have tested HRM, as well as the linguistic resources that we have employed. Likewise, evaluation will be focused on analyzing the advantages of considering ambiguity in humor processing, as well as on showing how HRM improves taking into account surface features. Finally, results will be discussed.

4.1 Initial Assumptions

As noted in Section 2.5.1, humor is a multidimensional phenomenon in which several factors interact for producing laughter. However, given its subjective and multivariable origin, humor is also a challenging subject for any scientific or humanist field. In this respect, from NLP's point of view, in Section 3.4 we described the efforts concerning automatic humor processing. In this chapter, we

will concentrate on giving evidence about how to characterize humor in terms of linguistic constituents (mostly ambiguity-based) in order to automatically recognize humorous patterns at textual level.

Our underlying assumption, apart from providing more complex patterns to automatically recognize this phenomenon, is to show how humor can provide valuable knowledge to be used beyond systems to generate it or recognize it[1]. Let us consider the statistics given in Figure 4.1 in order to clarify this point.

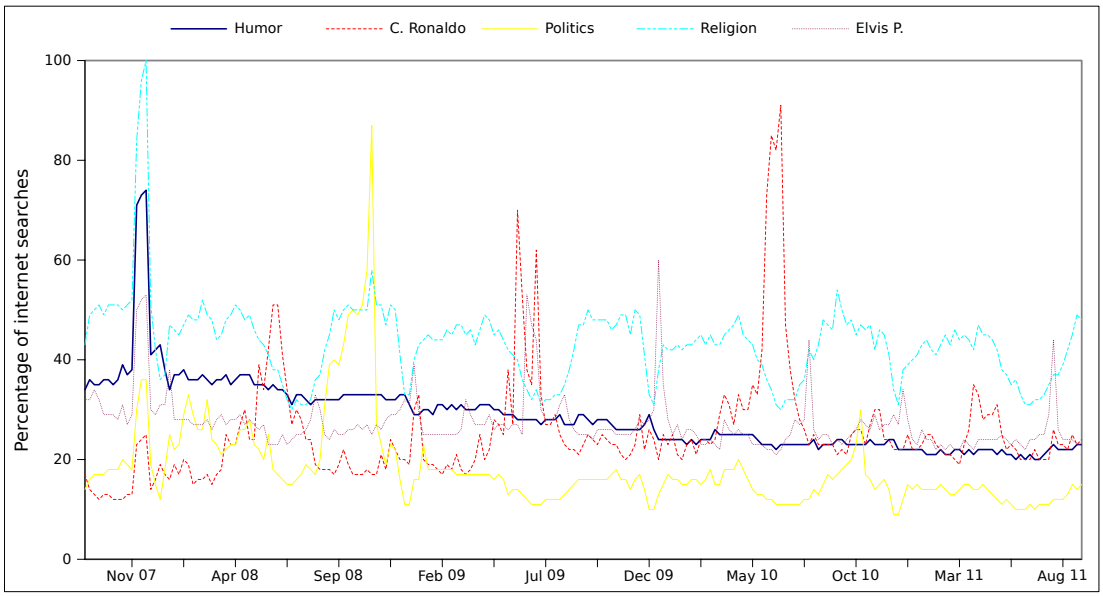

Figure 4.1: Frequency of Internet searches related to 5 different social subjects during four years (September 2007 - 2011) around the world. Statistics retrieved from Google Insights.

Figure 4.1 summarizes, according to Google, the amount of worldwide queries regarding five different subjects during four years. Despite humor is not the most frequent query, it is evident that it is a common subject when people make searches on Internet; i.e. of 100% of queries, around 32% are related to retrieve pages in which humor is the main subject. If considering the applicability of humor in practical tasks (beyond only retrieving funny videos, for instance), then the advances on automatic humor processing could find interesting targets in tasks such as opinion mining (cf. Esuli and Sebastiani [47]), sentiment analysis

[1]Recall that we explicitly refer to verbal humor (see Footnote 15 in Chapter 2).

(cf. Strapparava and Mihalcea [161]), blogs analysis (cf. Balog et al. [12]), information retrieval (cf. Feldman and Peng [48]), irony detection (cf. Reyes et al. [141]), and so on.

Furthermore, although humor is a window through which is possible to determine fine-grained information related to many aspects of human behavior, our interest is focused on detecting specific patterns with respect to how people use language to express humorous statements[2]. In this respect, various research questions have to be answered. For instance:

i. How to identify a phenomenon whose primary attributes rely on information that transcends the scope of linguistic arguments?[3]

ii. What are the formal elements to determine that a statement is funny?

iii. What are the linguistic constituents that should be taken into account when creating a model to automatically discriminate a funny statement from a serious one?

iv. If humor is not only a linguistic phenomenon, then how useful is to define a humor recognition model based on linguistic knowledge?

With regard to these questions, due to humor relies on many factors (physical, cognitive, social, cultural, and linguistic), it is doubtless that it cannot be defined only by means of linguistic arguments. However, as stated in Section 2.1, language is the most important vehicle by which non-linguistic information is conveyed. Therefore, we consider that the most suitable way of (computationally) handling humor's complexity is by means of linguistic patterns. Moreover, it is worth noting that our approach is not focused on qualifying how funny or unfunny a statement is; i.e. we will not judge funniness of humorous statements; rather we will concentrate on learning what elements make them funny. In this respect, we are focused on analyzing ambiguity. This linguistic mechanism can provide valuable knowledge in order to define a robust model of explicit and implicit patterns to automatically recognize humor. Such model, furthermore, can be

[2]Unfortunately, such patterns usually do not appear at surface level; i.e. they have to be inferred (and automatically learned) from the available data.

[3]Humor, as described in Chapter 2, is not only a linguistic phenomenon.

enhanced by considering recurrent humor topics, such as sex, religion, body parts, or ethnic minorities (see Nilsen [109]).

4.2 Humor Recognition Model

In this section we will describe the set of patterns to represent humor at textual level. Such patterns are selected based on the analysis of ambiguity throughout different linguistic layers.

4.2.1 Ambiguity

According to Binsted [17], two important linguists (Pepicello and Green [117]) claimed that humor is closely related to ambiguity. Actually, they affirmed that specific humor devices, such as punning riddles, depend on ambiguity to produce their comic effect. This ambiguity, they suggested, is in the language of the joke itself (such as the phonological ambiguity in a punning riddle), or in the situation the riddle describes. In addition, Mihalcea and Strapparava [102], pointed out how ambiguity, apart from leading surprise, is used as the mechanism to trigger the humorous effect. In this respect, ambiguity is especially used to suggest false semantic connections that generate humor.

Although such claims stress out the role of ambiguity as a major mechanism to produce humor, in this context of NLP, ambiguity is still work in progress. So far, results are important for tasks in which the target is literal language, for instance, POS tagging or word sense disambiguation. However, results are quite different regarding tasks in which FLP is involved (that is why ambiguity is regarded as future work in most of such tasks). Therefore, the question is how to capitalize the advances in the treatment of ambiguity beyond literal language. In particular, taking into consideration that ambiguity in figurative language, and especially in humor, usually entails knowledge beyond the word or the sentence. Let us consider example 30 below to clarify this point.

30) "Jesus saves, and at today's prices, that's a miracle!".

Unlike example 22 in Section 3.4.1, in which humor was given by phonological ambiguity (elly-phone vs. telephone), in this example humor is given by exploiting semantic and pragmatic ambiguity. The funny effect relies on turning the

figure of the sentence, i.e. "Jesus saves", in the *ground*, thereby modifying the literal meaning of the whole sentence. This process generates ambiguity and, consequently, the humorous effect. In other words, example 30 entails two interpretations: the first one is related to the literal meaning of preserving someone from harm or loss. The second interpretation shifts this meaning from its primary referent related to religion (i.e. the *figure*), to a secondary referent related to economy (i.e. the *ground*). This latter interpretation has to be promoted as primary referent. Finally, the funny effect is achieved by setting this interpretation within a frame related to the economic crisis: we all spend a lot of money. If someone (Jesus or whoever) can do the contrary, then that fact is a miracle.

This example stresses out that humor is a phenomenon which tends to exploit different types of ambiguities in order to achieve its effect. Therefore, we are focused on taking advantage of this property to integrate ambiguity in our HRM. To this end, below we describe four types of ambiguities used to trigger humor. Each depends on a specific linguistic layer[4]. Furthermore, each layer represents a pattern that is computationally estimated by means of statistical measures (below in Sections 4.3.2.1 to 4.3.2.4).

4.2.2 Lexical Ambiguity

Lexical ambiguity refers to the fact of having more than one meaning or sense registered in a dictionary. For instance, in "John lives near the bank", the noun *bank* can refer to either a building where money is stored, or the shore of a river (Binsted [17]).

We understand this type of ambiguity in terms of predictable sequences of words. Our hypothesis suggests that we use words based on conceptual frames (see Fillmore [49]). Such frames are responsible for imposing the use of certain sequences of words when a specific meaning is profiled. So, if *bank* has to be interpreted in terms of financial issues, then its conceptual frame will impose the use of predictable words such as money, cashier, checks, some named entities,

[4]Phonological layer is not here considered due to most research works on humor processing are focused on finding phonological patterns (see Binsted [17, 16], Purandare and Litman [125], Strapparava and Mihalcea [161]). In addition, our objective is focused on figurative language processing concerning only textual patterns (see 1.1).

etc. If this conceptual frame is violated (for instance, by using non-predictable words), then the meaning is deviated, and likely, humor is produced.

In order to estimate lexical ambiguity, we employ a measure called **perplexity**. This measure predicts, in terms of language models, the quality of linguistic representation given two probabilistic models (Jurafsky and Martin [72]). Therefore, our initial assumption is that humorous texts tend to maximize the degree of perplexity since they take advantage of lexical ambiguity to produce their effect.

4.2.3 Morphological Ambiguity

This type of ambiguity is given with regard to word's internal structure. For instance, the sentences "The book is read," and "The book is red" are morphologically ambiguous, since "read" is phonetically identical with "red" in its past participle form (Binsted [17]). Moreover, this ambiguity directly impacts on syntax. For instance, in absence of context, an isolated word such as *lay* can play different functions: noun [a narrative song with a recurrent refrain]; verb [put into a certain place or abstract location]; or adjective [characteristic of those who are not members of the clergy] (cf. WordNet [104] v. 3.0)[5]. Our hypothesis regarding morphological ambiguity relies on this latter fact. Humor is produced by using ambiguous words that alter the literal (and logic) meaning of any statement. Thus, funniness is triggered by a meaning shift. Consider the different functions that a word such as *lie* can play: verb or noun. A funny statement could exploit this ambiguity by profiling in $figure$ the meaning related to the verb (to be postrated), rather than the meaning related to the noun (prevarication).

Although this ambiguity seems to be difficult to represent due to it entails a deep linguistic analysis, we simplify the task by estimating the number of **POS tags** that a word in context can have. In this way, we have elements to analyze whether or nor a funny statement takes advantage of such ambiguity to generate its effect. Furthermore, by representing morphological ambiguity with POS tags we can obtain hints at the underlying mechanism of humor.

[5]Available at: http://wordnet.princeton.edu/.

4.2.4 Syntactic Ambiguity

Syntactic ambiguity is concerned with the fact of having different, and logical, interpretations at sentence level. For instance, in "John looked over the car", syntactic ambiguity is given by the possibility of having two distinct parse trees (Binsted [17]). In this respect, syntactic ambiguity is not given in terms of evaluating well-formed sentences, but in terms of finding out all the possible interpretations without violating syntactic rules. This fact is very important due to any statement, either literal or figurative, is supposed to be syntactically correct. Otherwise, it is more difficult to interpret the meaning of what it is conveyed. Based on this assumption, we aim to represent syntactic ambiguity in terms of syntactic complexity; i.e. instead of computing all the possible trees that any funny statement can have, we are focused on analyzing the complexity of its syntactic dependencies.

To this end, we employ a measure proposed by Basili and Zanzotto [13]: **sentence complexity**. This measure, beyond representing the most likely syntactic tree, captures aspects like average number of syntactic dependencies (what represents relevant information regarding our approach). By representing syntactic ambiguity with this measure, we will not evaluate the correctness of the syntactic tree, but its syntactic complexity. Our hypothesis is that humor tends to maximize the value of sentence complexity due to humorous statements, although syntactically correct, present at least two possible interpretations: the senseless interpretation and the funny interpretation.

4.2.5 Semantic Ambiguity

Semantic ambiguity occurs when a single word profiles multiple senses. For instance, in example 30, the word *save* profiles two different senses: one related to religion and one related to money[6]. In the case of this example, only one of these senses is responsible for enabling the funny interpretation. However, unlike lexical ambiguity, when this interpretation is activated, its conceptual frame does not only impose the use of certain words, but it also imposes certain extra-linguistic

[6]Sometimes such senses are not even registered in a dictionary due to they are (conventionally) deviated from their original referent. For instance, in colloquial communicative acts, the word fag is deviating its meaning in favor to others referents.

referents to correctly understand what it is communicated. In this respect, semantic ambiguity is more complicated to be detected. Therefore, we define a new measure to estimate this type of ambiguity: **semantic dispersion**. Such measure is intended to represent the possible frames that are activated when interpreting a funny statement. Each frame corresponds with a unique sense, if and only if, such sense is a synset registered in the WordNet ontology (Miller [104]). For instance, according to WordNet, *save* belongs to eleven different synsets. So, this is the number of possible frames that can be activated by this word. This information is tuned up by computing the hypernym distance of *save* with respect to all the synsets it belongs to. In this way, we obtain a numerical value related to the ambiguity produced by semantic information[7].

Our hypothesis, finally, relies on the fact discussed in Section 4.2.1: humor is often caused by shifting the ground sense. If this sense is profiled, then the primary and logical sense is broken, and accordingly, humor is produced.

4.3 Evaluation of Ambiguity-based Patterns

In this section we will assess the applicability of the four patterns above described to automatically recognize humorous texts. To this end, several experiments will be performed. Below, we first describe the data sets to be employed in this task, and then present the results obtained for each pattern.

4.3.1 Data Sets H1 - H3

Since humor is a very subjective phenomenon, the task of collecting humorous examples (positive data) is really challenging. Therefore, in order to avoid the subjectivity of collecting examples of humor based on personal judgments, we decided to use examples a-priori considered to be funny. To this end, we used three data sets employed in investigations related to humor. By considering this approach, apart from avoiding personal judgments regarding what we consider a funny statement, we obtained two benefits: i) it is unnecessary a manual annotation (nor agreement measures) because this information is given by user-generated

[7]Further details regarding the estimation of this measure are given in Section 4.3.2.4

tags, for instance, humor, funny, joke, etc.[8]; ii) according to our objective, we can extend the scopes of this research to others types of texts that contain figurative language.

The data sets are listed below:

a) Data set **H1** (Emoticorpus): Italian quotations. It contains 1,966 examples collected by automatically retrieving quotations, aphorisms and proverbs from the Italian Wikiquote. Used first in Buscaldi and Rosso [25]. Available at: http://users.dsic.upv.es/grupos/nle.

b) Data set **H2** (Humor Recognition): Humorous one-liners. It contains 16,000 examples collected by means of a bootstrapping process (see Mihalcea and Strapparava [99]). Used first in Mihalcea and Strapparava [100].

c) Data set **H3** (CesCa Project[9]): Humorous stories in Catalan produced by children between 6 and 16 years old. It contains 4,039 examples. Collected by means of direct interviews. Used first in Reyes et al. [137].

Each data set is summarized along with further details in Table 4.1.

In addition, some data sets are also used for two specific purposes: i) obtain negative examples; ii) create reference corpora. These data sets will be introduced when describing the experiments in which they are employed.

4.3.2 Evaluation

In this section we will describe the experiments carried out concerning the ambiguity-based patterns described in Sections 4.2.2 to 4.2.5.

[8]The use of user-generated tags is a common method to allow users the automatic labeling of their posts, web comments, tweets, etc. Unlike traditional methods, such as interviews or polls, in which the interviewee can rarely provide personal answers or judgments (most of these methods are multiple-choice questions), the use of user-generated tags allow people to **intentionally** focus their contributions on particular topics, as well as to provide their personal judgments by means of a descriptor (tag). Later in this chapter, as well as in Chapter 5, we will exemplify how people use this method.

[9]Project devoted to provide the educative community with a fundamental tool to know pupils' linguistic usage in Catalan. See http://clic.ub.edu/corpus/en/cesca-en.

Table 4.1: Detailed information regarding data sets H1 to H3.

	H1	**H2**	**H3**
Language	Italian	English	Catalan
Positive examples	471	16,000	1,867
Negative examples	1,495	—	2,172
Type of humor	Quotations	One-liners	Children's stories
Source	Wikipedia	Internet	Interviews
Labeling	Manual	Automatic	Manual
Availability	Public	Private	Private

4.3.2.1 Lexical Layer: Perplexity

In order to measure lexical ambiguity, a common statistical measure taken from language models was employed: perplexity. This measure indicates how well a given statistical model matches a test corpus. The perplexity (PPL) of a language model on a test set is a function of the probability that a language model assigns to a test set (Jurafsky and Martin [72]). For instance, the trigram "*the C5 anaphylatoxin*" will have higher perplexity than the trigram "*the red car*"; i.e. given a representative language model, it is more likely to find the sequence [the + red + car] than the sequence [the + C5 + anaphylatoxin].

According to Formula 4.1, for a test set $W = w_1, w_2 \dots w_N$, the perplexity is the probability of the test set, normalized by the number of words [see 72]:

$$PP(W) = P(w_1 w_2 \dots w_N) - \frac{1}{N} \tag{4.1}$$

In order to compute the perplexity concerning H1 - H3, we first created three reference language models. To this end, the SRILM Toolkit (Stolcke [160]) was employed. The reference language model is supposed to be representative enough to report common occurrences such as "*the red car*", or rare occurrences such as "*the C5 anaphylatoxin*". Three external data sets were employed to create the reference language models:

a‘) concerning H1, the Italian version of the Leipzig Corpus was used (Quasthoff et al. [126]). It contains 300,000 sentences collected from Italian newspapers.

b‘) concerning H2, Google N-grams were used (Brants and Franz [23]). This data set contains 95,119,665,584 sentences collected from public web pages stored in Google’s data centers.

c‘) concerning H3, the Catalan version of the Leipzig Corpus was used. It also contains 300,000 sentences collected from various Internet sites.

Each reference language model was trained with trigrams. Different smoothing methods[10] were employed: backoff, Good Turing, interpolation, etc. However, the results here reported correspond with the experiments in which interpolation and Kneser-Ney discount were applied. The following phase involved the comparison of the reference language model against two test distributions: one with positive data, and one with negative data[11]. Each was integrated with the same amount of positive and negative examples: H1: 471 examples. H2: 16,000 examples. H3: 1,867 examples. With respect to the negative examples, in Table 4.1 is evident how they are more than the positive ones, except for H2 which does not contain negative examples. In the case of H1 and H3, the negative examples were randomly undersampled to 471 and 1,867 examples, respectively. In the case of H2, the 16,000 negative examples were automatically retrieved from Internet[12].

After comparing the reference language model against the positive and negative test distribution, every data set was represented with its perplexity ratio. This ratio was obtained by dividing the perplexity of each data set by the size of the data set (471, 16,000, and 1,867, respectively). Results[13] are given in Table 4.2.

4.3.2.2 Morphological Layer: POS Tags

Although words in context are unlikely to be ambiguous, there are many situations in which this assumption is not satisfied. Observe example 31:

[10] This term refers to the fact of addressing the poor estimates that are due to variability in small data sets (see Jurafsky and Martin [72]).

[11] Prior to making this comparison, all words were stemmed, as well as all stopwords were eliminated.

[12] In this process we looked for retrieving examples with length similar to the one-liners.

[13] Although many experiments were performed (modifying parameters such as smoothing, order representation, or distribution), here are only reported the experiments with best results.

Table 4.2: Perplexity ratios.

	H1	**H2**	**H3**
	Positive | Negative	Positive | Negative	Positive | Negative
PPL	**3.08** | 2.33	**0.07** | 0.06	**0.63** | 0.34
OOVs‡	562 | 906	738 | 1446	205 | 742

(‡) Out Of Vocabulary words (regarding the reference language model)

31) "Why is coffee like soil? It is ground".

According to Pepicello and Green [117], the funny effect in this example is profiled by morphological ambiguity. This is clear when realizing that *ground* can be either noun (synonym of soil) or verb (past participle of grind). Such type of ambiguity is here computed by estimating, for all the words, their probability of playing different roles in context. Those roles are represented by POS tags. Thus, considering previous example, *ground* is likely to play the role of noun as well as the role of verb.

Table 4.3: Average of POS tags per example.

	H1	**H2**	**H3**
	Positive | Negative	Positive | Negative	Positive | Negative
Nouns	0.05 | 0.07	**0.08** | 0.07	**0.18** | 0.16
Verbs	**0.10** | 0.09	**0.11** | 0.08	**0.13** | 0.12
Adjectives	0.03 | 0.03	0.02 | 0.02	0.03 | 0.03

The process to obtain POS tags was performed by labeling both positive as negative test distributions[14] with the FreeLing toolkit (Atserias et al. [4])[15]. Apart from obtaining all POS tags, we obtained their different probabilities to play various roles in a sentence. This fact is highly important since words' meaning

[14] Recall that such distributions contain the same amount of examples for all the experiments: H1: 471 examples. H2: 16,000 examples. H3: 1,867 examples, respectively. Furthermore, in this experiment words were not previously stemmed, only stopwords were eliminated.

[15] Available at: www.lsi.upc.edu/~nlp/freeling.

is determined by context. Therefore, instead of basing morphological ambiguity only on the most likely POS tag, all the possible POS tags were considered. In this way, it is easier to statistically prove that *ground* may play the roles above mentioned with different thresholds of probability.

In Table 4.3 POS tags average per example is given. Such average was computed by summing the number of POS tags in categories noun, verb, and adjective, respectively, and then dividing by sentence length average. Finally, the result was normalized by the size of each data set. In addition, Figure 4.2 graphically shows the probability of assigning different POS tags according to the context.

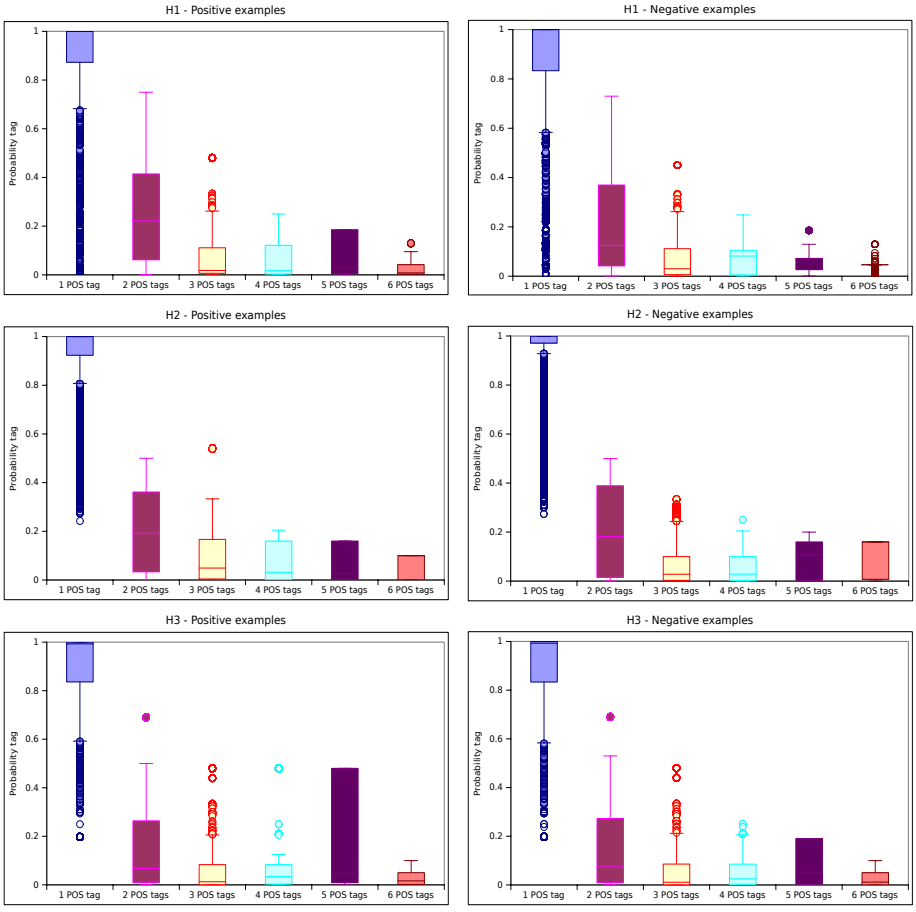

Figure 4.2: Probability of assigning POS tags concerning positive and negative examples in H1 - H3.

4.3.2.3 Syntactic Layer: Sentence Complexity

Syntactic ambiguity was computed in terms of finding out how complex any statement, either humorous or serious, is. In this respect, syntactic dependencies such as clauses or phrases, are formal structures to determine the syntactic complexity (see example 32).

32) "Children in the back seats of cars cause accidents, but accidents in the back seats of cars cause children".

The experiment consisted in running a syntactic parser in order to obtain the syntactic representation of all the examples. It is worth noting that H3 was not considered due to the resource employed (Chaos parser by Basili and Zanzotto [13]) does not contain any syntactic module for Catalan[16]. Therefore,

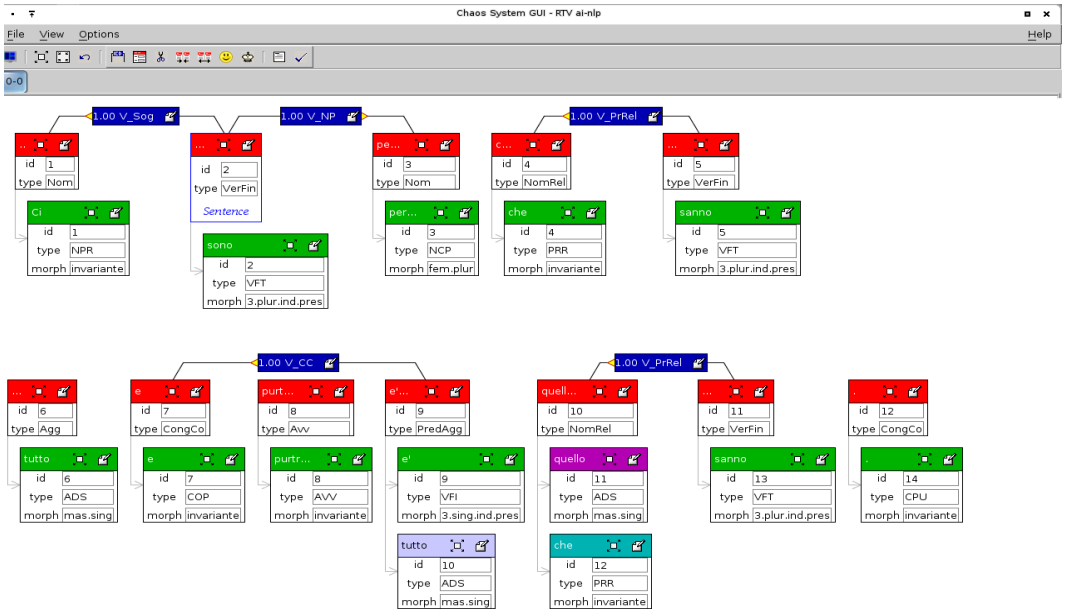

Figure 4.3: Chaos parser: Example of syntactic representation.

the results are focused only on H1 and on H2. Figure 4.3 illustrates the syntactic representation for the following sentence: "ci sono persone che sanno tutto e purtroppo è tutto quello che sanno"[17].

[16] There are few resources that include Catalan grammar, such as FreeLing; however, they were not considered due to the algorithm employed to estimate sentence complexity mandatorily requires a specific type of syntactic representation.

[17] There are people who know everything, and unfortunately, that is all they know.

Once obtained the syntactic representations, Formula 4.2 was employed in order to estimate the syntactic complexity:

$$SC = \forall t_n \frac{\sum v_l + \sum n_l}{\sum cl} \tag{4.2}$$

Such formula was introduced by Basili and Zanzotto [13]. It computes the number of verbal links v_l and nominal links n_l with respect to the number of clauses cl for any text t_n. Table 4.4 contains sentence complexity average concerning H1 and H2. Average was obtained by summing all the single scores and dividing by the size of each data set.

Table 4.4: Results concerning sentence complexity.

	H1	**H2**
	Positive | Negative	Positive | Negative
Sentence complexity	**1.84** | 1.72	0.99 | 2.02

4.3.2.4 Semantic Layer: Sense Dispersion

By estimating semantic ambiguity, we aim t assess whether or not this type of ambiguity is a key value for distinguishing figurative language, in particular, humor. Let us consider the following header to put this assumption in context: *The assembly passed and sent to the Senate a bill requiring dog owners in New York City to clean up after their dogs, in penalty of $100 fine. The bill also applies to Buffalo.* $Buffalo$'s ambiguity relies on the fact that $Buffalo$ can be either a city or a bison. If such ambiguity was only related to two different types of bison, or two cities, the ambiguous effect would disappear. Therefore, we think that the degree of ambiguity is key to humor; i.e. a word with senses that differ significantly between them is more likely to be used to create humor than a word with senses that differ slightly.

In order to estimate semantic ambiguity we employed a new measure called semantic dispersion. Given any noun, such measure computes the distance among all its senses and their first common hypernym. For instance, the noun *killer* has

four synsets (cf. WordNet v. 3.0). Taking into account only the synsets s_0 and s_1, their first common hypernym is *physical entity*. The distance (in terms of the number of nodes) to reach such hypernym is 6 and 2, respectively. Thus, *killer*'s semantic dispersion is the sum of those distances divided by 2. Now, considering all *killer*'s synsets, there are six different combinations whose distances to their first common hypernym produce the semantic dispersion of 6.83[18].

The experiment consisted in computing the semantic dispersion for all the nouns in H1 - H3. Since this measure is based on WordNet hierarchy, we used the Italian, English and Catalan versions of such resource, respectively. Semantic dispersion was computed by applying Formula 4.3:

$$\delta(w_s) = \frac{1}{P(|S|, 2)} \sum_{s_i, s_j \in S} d(s_i, s_j) \tag{4.3}$$

where S is the set of synsets $(s_0 \dots s_n)$ for noun w; P(n,k) is the number of permutations of n objects in k slots; and $d(s_i, s_j)$ is the distance of the hypernym path between synsets (s_i, s_j). In addition, Formula 4.4 was applied in order to obtain semantic dispersion average per sentence:

$$\delta_{Sn} = \sum_{w_s \in W} \delta(w_s) \tag{4.4}$$

where W is the set of nouns for sentence Sn. Finally, semantic dispersion average per word (δ_W) was calculated by dividing δ_{Sn} by the length of each data set. Results for both δ_{Sn} and δ_W are given in Table 4.5.

Table 4.5: Semantic dispersion at sentence and word level.

	H1	**H2**	**H3**
	Positive | Negative	Positive | Negative	Positive | Negative
$\overline{\delta}_{Sn}$	**4.62** | 3.73	**7.63** | 6.94	**4.83** | 4.16
$\overline{\delta}_W$	**0.74** | 0.42	1.85 | 2.25	**1.14** | 0.94

Preliminary findings are analyzed in the following section.

[18] In other words, semantic dispersion is a way of quantifying the differences among the senses of any polisemic noun.

4.3.3 Discussion of Preliminary Findings

The following comments are based on the results obtained in the experiments above described. Although results are only reported in terms of tables and figures, some interesting findings can be inferred from them. On one hand, concerning lexical ambiguity, results reported in Table 4.2 show that humor in Romance languages such as Italian (H1) and Catalan (H3) seems to be less predictable than English (H2). For instance, perplexity in H1 and H3 is clearly higher when testing positive examples than when testing negative ones (although such difference in perplexity is marginal concerning H2, the pattern does not disappear). Such result makes evident that, given two different distributional schemes (humorous and serious), the structures that better exploit lexical ambiguity are the humorous ones. According to our hypothesis, perplexity in H1 - H3 shows that figurative language, especially humor, is probabilistically more ambiguous in terms of predictable sequences of words than literal language. That is, in terms of language models, it is more likely to predict the word $w{+}_1$ given a humorous statement than given a serious one. Moreover, looking at the Out Of the Vocabulary words (OOVs), we will realize that the frequency of OOVs is higher regarding negative examples (almost double in H2 and H3). Based on this difference one would expect higher perplexity regarding negative examples. However, this does not occur. We think that this fact makes evident how humor is intrinsically more difficult to be classified.

Concerning morphological ambiguity, Table 4.3 shows that humorous statements, on average, often use verbs and nouns to produce such ambiguity. In contrast, adjectives are used concisely. This fact is consistent with example 31, in which humor is produced by profiling a noun in figure instead of a verb. With respect to Figure 4.2, it is evident that higher probabilities are closer to 1 when words may be labeled with up to two tags. In contrast, in cases when words may have more than 3 tags, probability is closer to 0. Moreover, positive examples tend to contain more words that may be labeled with different tags, except for H2 (5-6 POS tags). Based on this behavior, we can infer that, given an isolated word w_i, the probability to be assigned to various categories can break logical meaning, thereby producing ambiguity and, therefore, humorous effects.

On the other hand, our third experiment aimed at verifying whether or not the

syntactic complexity could provide useful information about the impact of syntax on humor. Results in Table 4.4 show two different scenarios. Concerning H1, sentence complexity is slightly higher regarding humorous examples. However, concerning H2, the result is completely contrary; i.e. according to the results, non-humorous statements are syntactically more complex than humorous ones. Although the latter result is opposite to our hypothesis, it is consistent whit Mihalcea and Strapparava's claims regarding one-liners' simple syntactic structure. Moreover, such result is even more consistent with pragmatic goals: there is always a communicative goal. To achieve such goal, it is necessary the communication by means of well-formed expressions (as possible). Therefore, from a syntactic point of view, humor, and accordingly one-liners, are well-formed structures that would tend to exploit others types of linguistic strategies to produce the funny effect.

Finally, the role of semantic ambiguity as a trigger of humorous situations seems to be more relevant. According to the results reported in Table 4.5, it is clear that semantic dispersion is higher concerning positive examples, for both sentence representation and word representation (except for this latter representation in H2). Such results strengthens our hypothesis regarding the role of semantic ambiguity on humor generation. Moreover, results are consistent with the ones regarding syntactic ambiguity: if sentence complexity is lower concerning humorous examples, then one might think that humor is usually produced by exploiting lexical, morphological or semantic ambiguity, rather than syntactic ambiguity. Based on these results, we may infer that semantic strategies are relevant for generating hollows of ambiguity. Those hollows are intentionally used for producing second interpretations, and accordingly, humor.

4.4 Adding Surface Patterns

Although the patterns described so far seem to be relevant facing the task of automatically recognizing humorous statements, it is also necessary to enhance HRM's scope by considering patterns beyond linguistic ambiguity[19]. Moreover, it is indispensable to assess the model beyond the conceptual representation. Thus,

[19]It is senseless to think that a few patterns are sufficient to satisfactorily achieve such task.

in this section we will describe new patterns (here called surface patterns) as well as new evaluations that are focused on automatic classification tasks.

4.4.1 Humor Domain

According to Nilsen [109], humor tends to rely on taboos and censorship such as sex, religion, swear words, women, gays, ethnic minorities, and so on. This trend was also commented by Mihalcea and Strapparava [101]. They reported that elements such as human centric vocabulary (e.g. pronouns), professional communities (e.g. lawyers), or human weakness (e.g. beer) are recurrent instances in their corpus of humorous one-liners. In this respect, *humor domain* is our pattern to represent features such as the above mentioned. Conceptually, by means of *humor domain* we look for integrating some of the most relevant features described in the specialized literature concerning humor processing. Thus, the following features were considered:

- *adult slang*: focused on obtaining all the words with the tag "sexuality" in WordNet Domains (Bentivogli et al. [14]), as described by [102];
- $wh-templates$: focused on representing syntactic information in terms of wh-phrases;
- *social relationships*: focused on retrieving all the nouns concerning the synsets *relation*, *relationship* and *relative* in [104], as described by Mihalcea and Pulman [98];
- *nationalities*: focused on identifying ethnic information.

Each feature is illustrated in examples 33 to 36, respectively.

33) "Artificial **Insemination**: procreation without recreation".

34) "**What** are the 3 words you never want to hear while making love? Honey, I'm home!"

35) "A **family** reunion is an effective form of birth control. "

36) "In **Canada** we have two seasons . . . six months of winter and six months of poor snowmobile weather "

4.4.2 Polarity

According to Superiority Theory (see Section 2.5.1), humor is a kind of malice towards those who are considered relatively powerless. Therefore, every humorous situation has a winner and a loser ([169]). In the same vein, Attardo [8] cited Freud's claims about innocent and tendentious jokes. Both perspectives undoubtedly profile the presence of negative elements concerning humor generation. In this respect, by means of *polarity* we aim to represent the degree of negativeness (or positiveness) in humor from a sentiment analysis point of view.

Two different resources were used in order to represent *polarity*: SentiWordNet (Esuli and Sebastiani [47])[20] and Macquarie Semantic Orientation Lexicon (MSOL) (Saif et al. [145])[21]. The former proposes a set of graduated tags to label the following categories: nouns, verbs, adjectives and adverbs[22]. The latter, MSOL, contains 76,400 entries (30,458 positive and 45,942 negative ones). It is based on Roget-like thesaurus for expanding their positive and negative seeds. According to their authors, MSOL reaches a higher-coverage regarding phrase polarity than SentiWordNet.

4.4.3 Templates

Like most figurative language, humor is a phenomenon closely related to creativity. However, such creativity is not given in all the humorous statements; rather, in most cases is "plagiarized"; i.e. when a joke is really funny, people tend to copy the pattern that makes them laugh. Such pattern is then generalized by changing certain elements but keeping the core intact. For instance, consider all the jokes that exploit the sequence: nationality 1 + nationality 2 + nationality 3. Such sequence is quite productive. It may be filled with as many nationalities as countries exist, just by modifying either order or nations involved. This type of generalization is intended to be identified by means of the pattern concerning *templates*. Our underlying assumption relies on language systematicity: success-

[20] Available at: http://sentiwordnet.isti.cnr.it/.

[21] Available at: www.umiacs.umd.edu/saif/WebPages/ResearchInterests.html.

[22] In order to avoid ambiguity regarding such graduated tags, only tags with score ≥ 375 were considered. For instance, *adventive* has the positive score of 0.0, and the negative of 875; whereas *rash* has the positive score of 625 and the negative of 0.25. Scores ≤ 375 are discarded due they do not provide sufficient information to define orientation.

fully communicative patterns tend to be easily adopted by people in order to guarantee social linguistic functions.

Templates are identified by measuring Mutual Information Oakes [110]. With such measure we attempted to identify recurrent templates based on the probability of evaluating two or more words as a unique entity.

4.4.4 Affectiveness

The final surface pattern is related to humor's psychological side; i.e. emotional aspects of humor. According to Valitutti [169], the pleasurable humorous feeling arises from the disparagement toward some target character or category of people, to the activation level of the nervous autonomic system when a humorous response is produced. In this respect, $affectiveness$ is a pattern to quantify the degree of emotional content profiled by humorous statements.

Such content is computed according to WordNet-Affect categories (Strapparava and Valitutti [162]). Those categories are intended to represent how speakers convey emotional content by means of selecting certain words and not others. They are: attitude, behaviour, cognitive state, edonic signal, emotion, mood, physical state, emotional response, sensation, emotion-eliciting situation, and trait[23].

4.5 HRM Evaluation

Experiments here described were performed in order to obtain empirical evidence regarding HRM's strengths to automatically distinguish humorous statements from non-humorous ones. To this end, a new data set was built. Such data set (H4) was integrated with 19,200 blogs organized in 8 subsets, each contains 2,400 texts. We decided to use blogs due to they are heterogeneous sites in which humor is not text-specific; i.e. humor is expressed in several ways: jokes, gags, punning riddles, one-liners, comments, discussions, and so on. In this respect, HRM's scope would not be limited to one specific type of humor expression.

[23]Detailed information regarding the concepts symbolized by these categories is described in [162].

4.5.1 Data Set H4

Data set H4 was automatically collected from LiveJournal.com. Two requirements were established: i) the mandatory presence of user-generated tags[24] such as humor, joke, funny, laughter, and so on; concerning the positive subset (*humor*); ii) the presence of user-generated tags regarding moods concerning negative subsets. In this respect, we considered the following user-generated tags: *angry*, *happy*, *sad*, *scared*, *miscellaneous*, and *general*), as well as one more subset retrieved from Wikipedia (*wikipedia*). The latter subsets (*general* and *wikipedia*) were harvested without considering any tag. They are intended to be control subsets. Below are listed H4's characteristics according to every subset. In addition, Appendix B illustrates what kind of information is contained in such data set.

d) Data set **H4** (Blog Analysis): One-liners, long and short jokes, discussions about humor, web comments, personal posts, etc. It contains 19,200 documents. Collected by automatically retrieving documents labeled with user-generated tags. English language. Used first in Reyes et al. [136]. It is publicly available at: `http://users.dsic.upv.es/grupos/nle`.

Prior carrying out any experiment, H4 was evaluated in terms of the measures described by Pinto et al. [122] regarding the assessment of text corpora. They are: *shortness*, *broadness*, and *stylometry*. Evaluation was performed in order to minimize the noise produced by the automatic blog retrieval[25]. Results are given in Table 4.6.[26]

According to the values registered in Table 4.6, *shortness* is low, both in terms of documents and vocabulary. The vocabulary and document ratio (VDR) measure indicates that all the subsets entail high complexity. Concerning *broadness*, the unsupervised vocabulary based (UVB) measure indicates that, broadly, all the subsets tend to restrict their topics to specific contents, especially, regarding the subsets *happy* and *humor*. With respect to *stylometry*, the stylometric evaluation measure (SEM) yields interesting information related to style. This seems

[24] See footnote 8.

[25] Before estimating those measures, all stopwords were eliminated. The list of stopwords was enhanced with words such login, username, copyright, LiveJournal, next, top, as well as words related to html content, in order to not bias H4's characteristics .

[26] The Watermarking Corpora On-line System (WaCOS) is available online at: `http://nlp.dsic.upv.es/watermarker/`.

Table 4.6: H4 characteristics. Measures: corpus vocabulary size (CVS); document and vocabulary length (DL and VL, respectively); vocabulary and document length ratio (VDR); unsupervised vocabulary based measure (UVB); stylometric evaluation measure (SEM).

	Humor	Angry	Happy	Sad	Scared	Miscellaneous	General	Wikipedia
Words	1,577.16	1,314.55	1,114.41	1,193.92	1,342.98	1,027.32	843.44	1,934.07
CVS	219.25	132.83	161.33	119.90	145.42	122.56	107.00	162.30
DL	720.50	604.39	542.56	567.73	625.81	483.44	410.44	937.96
VL	503.27	411.10	382.99	384.47	418.42	341.92	301.68	516.18
VDR	0.95	0.94	0.94	0.94	0.94	0.94	0.95	0.91
UVB	9.29	6.91	9.27	6.78	7.48	7.75	7.80	6.90
SEM	0.37	0.39	0.40	0.38	0.37	0.40	0.46	0.40

to be fairly common, in terms of stylistic expression, in each subset. Based on the values above described, we can conclude that all the subsets in H4, apart from containing information related to humor, are distinguishable each other enough to be used for our purposes[27].

4.5.2 Evaluation

HRM's evaluation consists in analyzing the capabilities of automatically classifying texts into the data set they belong to. On one hand, each of the 19,200 documents was represented as a frequency-weighted term vector[28] (considering both ambiguity-based patterns as surface patterns) according to a humor average score obtained by applying the following bag-of-patterns framework:

1 Let $(p_1 \dots p_n)$ be HRM' patterns, concerning both ambiguity-based and surface patterns.

2 Let $(b_1 \dots b_k)$ be the set of documents in H4, regardless of the subset they belong to.

[27] Unlike data sets H1 - H3, which were assessed in investigations prior ours, H4 is a new data set which must be assessed in terms of its applicability for researches in humor. Therefore, it was necessary to apply a quality control (the measures above cited) in the information retrieved in order to guarantee its relevance for humor processing.

[28] All documents were first preprocessed by removing stopwords. Stemming was applied as well. Porter algorithm was employed (Porter [123]).

3 If $b_k \left(\frac{\sum_{p_1 \ldots p_n}}{|B|} \right) \geq 0.5$, then humor average for b_k was $= 1$.

4 Otherwise, humor average was $= 0$.

Humor average is intended to evaluate whether the set of patterns are linguistically correlated to the ways in which people employ words when producing humorous contents, regardless of the use of user-generated tags. Humor average per subset is graphically represented in Figure 4.4.

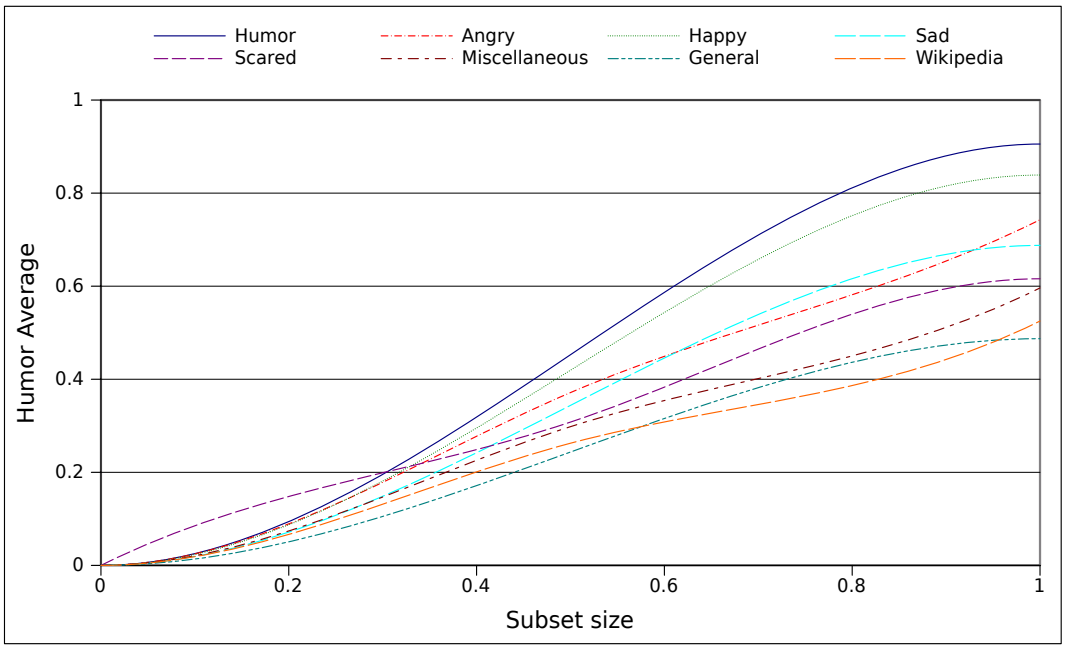

Figure 4.4: Humor average per subset.

On the other hand, two popular algorithms were used to classify all blogs: Naïve Bayes (NB) and Support Vector Machines (SVM)[29]. All the classifiers were trained considering 70% for training, and 30% for test. HRM was evaluated as a single entity; i.e. no pattern was evaluated isolately. Results in terms of accuracy, precision, recall, and F-measure, are detailed in Table 4.7 and Table 4.8.

[29]Such classifiers were chosen because most research works on humor processing use them for evaluation.

4.5.3 Results and Overall Discussion

The latter classification experiments were carried out in order to understand how difficult is to automatically detect humorous patterns beyond text-specific examples. In this respect, by means of this task we aimed at evaluating the entire system of patterns to accurately classify non text-specific instances into the subset they belong to.

On one hand, humor average results (illustrated in Figure 4.4) corroborate the relevance of the set of patterns to identify common sequences of elements used by people when producing humorous contents. Although expected, humor average is closer to 1 concerning subset humor. However, subsets concerning moods happy, angry, and sad, seem to have important humorous content. A priori, this could be correlated to humor's psychological branch, as well as to humor's relief effects: why do we laugh?; do we laugh for not suffering?; how is humor cognitively processed?; what makes us laugh?

On the other hand, according to the results registered in Table 4.7 and Table 4.8, important implications regarding the usefulness of the patterns above described to represent humor can be noted. For instance, concerning subset humor, accuracy is clearly higher for both NB and SVM classifiers. In contrast, accuracy is always lesser concerning all the negative subsets, especially, concerning subset general. This is supported by precision, recall, and F-measure scores. Although such results do not achieve 100% of accuracy (our best result achieves 89.63%), they are comparable with the results obtained by Strapparava and Mihalcea [161], Sjöbergh and Araki [154], Buscaldi and Rosso [25] concerning short texts; and with the ones described by Mihalcea and Pulman [98] concerning longer texts. Moreover, it is worth noting how the model seems to correctly discriminate blogs regarding subsets happy and wikipedia; i.e. positive mood and specialized contents, respectively. However, regarding subsets such as sad, angry or scared, accuracy considerably decreases. This is likely due to the close relationship among the negative moods (and underlying contents) profiled by such subsets. The rest of negative subsets (miscellaneous and general) were the worst classified, being the latter the one which achieved lowest accuracy (51% with SVM). This fact makes evident how difficult this task is. Due such subsets are not focused on particular topics, contents are spread and the possibility of finding specific infor-

mation is low. It is just like making a query based only on a single word: result will be irrelevant.

Table 4.7: Results obtained with NB classifier.

	Accuracy	Precision	Recall	F-measure
Humor	**87.17%**	**0.87**	**0.87**	**0.87**
Angry	69.37%	0.70	0.69	0.69
Happy	84.85%	0.85	0.85	0.85
Sad	66.13%	0.67	0.66	0.66
Scared	69.67%	0.70	0.70	0.69
Miscellaneous	62.18%	0.64	0.62	0.61
General	53.05%	0.55	0.53	0.49
Wikipedia	75.19%	0.77	0.75	0.75

Table 4.8: Results obtained with SVM classifier.

	Accuracy	Precision	Recall	F-measure
Humor	**89.63%**	**0.90**	**0.90**	**0.90**
Angry	71.40%	0.71	0.71	0.71
Happy	83.87%	0.84	0.84	0.84
Sad	66.13%	0.67	0.66	0.66
Scared	69.67%	0.70	0.70	0.69
Miscellaneous	62.63%	0.71	0.63	0.58
General	51.86%	0.55	0.52	0.44
Wikipedia	76.75%	0.78	0.77	0.77

Although the previous results note HRM's proficiency for classifying different types of humor, at least regarding the data sets here employed, not all the patterns are equally relevant. That is why all the patterns were reassessed in terms of their relevance for representing humor. To this end, an information gain filter was applied. Results are given in Table 4.9.

Table 4.9: Information gain results on HRM's patterns.

Ranking	Pattern	Feature
1	Lexical ambiguity	PPL
2	Domain	Adult slang, wh-templates, relationships, nationalities
3	Semantic ambiguity	Semantic dispersion
4	Affectiveness	Emotional content
5	Morphological ambiguity	POS tags
6	Templates	Mutual information
7	Polarity	Positive/Negative
8	Syntactic ambiguity	Sentence complexity

According to the results registered in Table 4.9, it is evident that some patterns are more relevant for this task. Such relevance can be related to the type of content profiled by texts. Perhaps considering different sources the relevance would be different. Nonetheless, the important point is that HRM works as a whole; i.e. not single pattern is humorous per se, but all together provide a valuable linguistic inventory for detecting humor at textual level.

Finally, we would like to stress some remarks regarding every pattern.

Results obtained by estimating **lexical ambiguity** proved, according to our initial assumption, that ambiguity underlies humor. Therefore, humorous statements are less predictable and, probabilistically, more ambiguous than non humorous ones. This means that, given two different distributional schemes, humorous discourses profile a broader range of combinations to trigger ambiguous situations when generating the funny effect.

Morphological ambiguity seems to be another important feature to represent humor. Although its information gain is not so high, when conceptually evaluating data sets H1 - H3 such pattern showed its relevance for characterizing humorous statements. Instead, concerning **syntactic ambiguity**, it is quite clear that this pattern is useless for characterizing humor (at least regarding data sets H1, H2, and H4). Sentence complexity showed that humorous statements are less complex than serious texts; except in H1. However, in such data set, sen-

tence complexity score could be a matter of syntactic rules: syntax in Romance languages such as Italian (H1) is not as rigid as in English (H2, H4). Therefore, such pattern could be language-dependent.

Concerning **semantic ambiguity**, according to the results obtained, it can be deemed as an important source of humorous situations. By profiling, at least two possible interpretations, it is more likely to generate hollows of ambiguity that contribute to produce more complex meanings to produce humorous effects.

Results concerning **humor domain** corroborated the relevance of features previously tested in various research works to automatically recognize humor. Its relevance is even clearer when confronting information gain results: it is ranked in second place. Therefore, it is a hit the inclusion of such features in any humor model.

With respect to **polarity**, it is worth noting how positive words are more representative concerning funny texts. This fact is contrary to results described in Mihalcea and Pulman [98]. They indicate the relevance of negative information to generate humor. However, in our case, such relevance is not fullfiled. Perhaps, this is due to the type of texts.

Regarding **templates**, results provide some hints concerning the presence of recurrent sequences such as the ones described in Section 4.4.3. Such sequences indicate that humorous texts often exploit the same productive template to successfully produce the funny effect.

Finally, **affectiveness** proved that humor takes advantage of emotional content to convey or generate its effect. That is why its relevance concerning information gain results. Such result can be interpreted as a successful way of communicating ad hoc stimuli, through which, people easily create favorable contexts to express humor.

4.6 Summary

In this chapter we have described our HRM and the linguistic patterns used to evaluate it. First, in Section 4.1 we provided our initial assumptions concerning the task of automatically recognizing humor. Then, in Section 4.2 we detailed and exemplified HRM's patterns. In particular, in this section we were focused on describing the role of linguistic ambiguity in humor.

Moreover, in Section 4.3 we described all the experiments regarding linguistic ambiguity, as well as their evaluation. Data sets H1 - H3 were introduced. In addition, in Section 4.4 we outlined the need of including surface patterns.

In Section 4.5 those patterns, as well as ambiguity-based patterns, were assessed by means of a classification task. The new data set (H4) was introduced and described. Finally, in Subsection 4.5.2 we detailed the evaluation, and then, in Subsection 4.5.3, results were presented and discussed.

5

Irony Detection

The effect that this t-shirt has on women is pretty impressive. Unfortunately its natural healing powers reversed my vasectomy and I impregnated nine women in two weeks before I realized. They all had twin boys. Now I have 18 sons and spend most of my money on child support and condoms.

DATA SET I1
REYES AND ROSSO [132]

This chapter will be focused on describing our **Irony Detection Model (IDM)**. First, operational bases, as well as aims, will be introduced. Then, experiments and results will be explained. In addition, we will introduce the data sets in which IDM is going to be assessed. In this respect, we will highlight the challenges of compiling a data set with ironic examples, as well as the lack of resources concerning automatic irony processing. Finally, results and further implications will be discussed.

5.1 Irony: Beyond a Funny Effect

As described in Section 2.5.1, irony and humor tend to overlap their effects: an ironic statement can easily cause a funny reaction, as well as a joke can exploit irony to produce laughter (cf. Gibbs and Izett [56], Colston [34], Pexman [120]). However, although such devices can share some similarities (that make suppose a kind of logic entailment between them), they cannot be treated as the same device, neither theoretically nor computationally. For instance, some HRM's patterns could be useful when dealing with ironic examples (the ones that exploit humor), but they would not be enough to cover the instances in which irony is not humor-dependent[1]. Let us think of the major characteristic of such device: negation.

Irony is one of the most subtle figurative devices used to, in a refined way, deny what it is literally said. Most people concur that irony relies on negation or opposition (see Chapter 2). However, unlike literal language, such negation is not formally marked; i.e. there is not any explicit negation marker underlying ironic statements. Such fact, apart from making evident the major difference between irony and humor, represents a significant challenge due to the lack of formal elements to identify what it is being negated. In this respect, although many authors have provided arguments to deal with negation, most of them are focused on literal language[2]. Hence, such arguments can hardly be mapped to the irony detection task. In particular, due to literal language always profiles a formal linguistic constituent to indicate that something is being negated.

In this chapter, thus, we will be focused on providing arguments to: (a) prove that HRM's patterns are not sufficient to correctly address the irony detection task (although some of them can be useful); (b) show how to deal with non-

[1]Later in this chapter we will describe some experiments concerning the poor applicability of HRM over the ironic examples.

[2]For instance, the investigation described by Giora et al. [58] in which the authors assessed whether or not the information introduced via negation markers is retained or suppressed when mentally representing the negated concepts. Along the same line, Kaup et al. [78] investigated the amount of milliseconds to process affirmative and negative sentences in which a target entity and a contradictory predicate were being mentioned. Likewise, concerning NLP, Morante and Daelemans [106] described a metalearning approach to process the scope of negation in biomedical texts.

factual information that is inherent to the role of negation in irony; (c) represent the core of the concept of irony by means of linguistic patterns.

Last, but not least, it is worth noting that it is unrealistic to seek a computational silver bullet for irony, and a general solution will not be found in any single technique or algorithm. Rather, we must try to identify specific aspects and forms of irony that are susceptible to computational analysis, and from these individual treatments, attempt to synthesize a gradually broader solution. In this context, we will be focused on describing a model capable of representing the most obvious attributes of irony in a text, or at least what speakers believe to be irony, in order to automatically detect possible ironic statements in user-generated contents such as opinions, comments, or reviews.

5.2 Target

Since irony is common in texts that express subjective and deeply-felt opinions, its presence represents a significant obstacle to the accurate analysis of sentiment in such texts (cf. Wiegand et al. [180] and Councill et al. [38]). Therefore, a successful model of irony could play both a direct and an indirect role to this task. In this respect, one of the most difficult problems when assigning either positive or negative polarity in sentiment analysis tasks[3] is regarded to determining what is the truth value of any statement. In particular, due to irony allows to change the truth value of such statement. In the case of literal language (e.g. "this movie is crap") the existent techniques achieve good results; instead, when the meaning in ground is totally different to the meaning in figure, the result may be a consequence of simply finding out what types of words prevail in the surface of the statement (e.g. "It's not that there isn't anything positive to say about the film. There is. After 92 minutes, it ends"). In those cases, the same automatic techniques lose effectiveness because the profiled and real meaning is in ground, or *veiled* due to the presence of irony. The shift meaning due to this figurative

[3]According to the terminology discussed in Pang and Lee [112], we differentiate opinion mining from sentiment analysis based on the fact that the former suggests an emphasis on extracting and analyzing judgments on various aspects of given items; whereas sentiment analysis is focused on the specific application of classifying reviews as to their polarity (either positive or negative).

device may be evident for humans; i.e. given the proper frame, we can easily interpret that a negative polarity permeates the last example. However, how do we do to define a computational model capable of recognizing the negated meaning in which irony relies? The question seems to be nearly impossible to be computationally solved. Nonetheless, we attempt to investigate a first approach which could provide insights into the figurative uses of textual elements to communicate irony.

In this respect, the expected result could represent fine-grained knowledge to be applied in tasks as diverse as sentiment analysis (cf. Reyes and Rosso [129] about the importance of determining the presence of irony in order to set a fine-grained polarity), opinion mining (cf. Sarmento et al. [146], where the authors note the role of irony for minimizing the error when discriminating negative from positive opinions), or even advertising (cf. Kreuz [81] as well as Gibbs and Izett [56], concerning the function of irony to increase message effectiveness).

5.3 Basic Irony Detection Model

We are proposing a new model that is organized according to six operational patterns. Such patterns are intended to capture low-level properties of irony based on conceptual descriptions found in the specialized literature; i.e. we intend to extract the core of the most defining characteristics of verbal irony, according to several formal studies such as the ones cited in Section 2.5.2, and then, transfer this core to our model by mapping it through *textual patterns*. The set of patterns are listed below.

i. **N-grams**; concerning with finding frequent sequences of words based on n-grams of different orders.

ii. **Descriptors**; concerning with providing tuned up sequences of words based on discriminating irrelevant information.

iii. **POS n-grams**; concerning with establishing morphosyntactic templates given a POS representation.

iv. **Polarity**; concerning with evaluating the underlying polarity of ironic statements as reported in Section 4.4.2.

v. **Affectiveness**; concerning with representing irony in terms of emotional content as reported in Section 4.4.4.

vi. **Pleasantness**; concerning with measuring the degree of pleasure produced by irony.

5.3.1 Data Set I1

Like humor, irony (and most figurative language) is very subjective and often depends on personal appreciation. Therefore, the task of collecting ironic examples (positive data) is quite challenging. In addition, as noted in Chapter 2, the boundaries to differentiate verbal irony from situational irony, or even from sarcasm or satire, are very fuzzy indeed: non-expert people usually use an intuitive and unspoken definition of irony rather than one sanctioned by a dictionary or a text-book. Hence, such task becomes any harder. Given this scenario, we have opted for collecting a data set with statements that are a priori labeled as ironic by social media users. In this respect, we decided to rely on the wisdom of the crowd and retrieve a set of customer reviews from Amazon web site. Such reviews are considered to be ironic by customers, as well as by many journalists, both in mass and social media. According to their personal appreciation, all these reviews deal with irony, sarcasm and satire (hence, they are consistent with our definition of irony). It is worth noting that all the reviews were posted by means of an online viral effect, which in most cases, increased the popularity and sales of the reviewed products, thereby achieving a cult status. The *Three Wolf Moon T-shirt* is the clearest example. This item became one of the most popular products, both in Amazon as in social networks, due to the ironic reviews posted by many users[4]. For instance, visit the following web sites in order to evaluate the scope of the reviews posted for ironically commenting this t-shirt: Youtube[5], Wikipedia[6], BBC [7], or ABC[8].

[4] According to Google search engine, there are more than one million of results when searching this product. In addition, there are more than 10,000 blogs concerning the viral effect caused by the reviews above mentioned.

[5] http://www.youtube.com/watch?v=QPB45AUmchM.

[6] http://en.wikipedia.org/wiki/Three_Wolf_Moon.

[7] http://news.bbc.co.uk/2/hi/8061031.stm.

[8] http://abcnews.go.com/WN/story?id=7690387&page=1.

Likewise, we decided to use such reviews due to the importance of Amazon in electronic commerce: such importance is not only supported by its business schema, but by trusting in the opinions posted by its customers. Those opinions impact, either positively or negatively, on other customers interested in the products offered by Amazon. The fact of considering such opinions in order to mine deeper information to detect irony could be capitalized for labeling opinions beyond the positive or negative polarity; rather, for obtaining fine-grained information to be employed, for instance, for a better decision making (see Kim et al. [80] and Jøsang et al. [71] about the role of trust on decision making).

Data set **I1** contains 3,163 ironic reviews concerning five products published by Amazon. The list of products is given below:

- ≻ *Three Wolf Moon T-shirt*. Amazon product id: B002HJ377A
- ≻ *Tuscan Whole Milk*. Amazon product id: B00032G1S0
- ≻ *Zubaz Pants*. Amazon product id: B000WVXM0W
- ≻ *Uranium Ore*. Amazon product id: B000796XXM
- ≻ *Platinum Radiant Cut 3-Stone*. Amazon product id: B001G603AE

In addition, all the reviews whose customer rating, according to the Amazon rating criteria, was lesser than four stars were removed in order to filter out reviews with non ironic content. Two pragmatic facts support this decision: i) viral effect, and ii) ironic intent. The former is related to the fact of posting reviews whose main purpose, and perhaps the only one, is regarded to describe superficial or non-existent properties about certain product or topic. Such fact produces an effect on users, who automatically copy the strategy. Based on this assumption, it is unlikely to find *real* reviews in a scenario like this because every user is contending to show who posts the most "original" (in our case, ironic) review. In turn, ironic intent is related to the concept of negation: irony is employed in order to negate what is communicating. Therefore, if users are ironically commenting any product, they will not do it by rating it with the lowest score (one or two stars in Amazon); rather, they will rate it with the highest score (four and five stars).

After applying this filter, **I1** was reduced to 2,861 ironic reviews. In addition, three negative data sets complement **I1**. They were automatically collected from the following sites in order to assess IDM's capabilities: Amazon.com, Slashdot.com, and TripAdvisor.com. Each contains 3,000 documents. The products selected from Amazon (**AMA**) were: Bananagrams (toy), The Help by Kathryn Stockett (book), Flip UltraHD Camcorder (camera), I Dreamed A Dream (CD), Wii Fit Plus with Balance Board (Videogame console). The set collected from Slashdot (**SLA**) contains web comments categorized as funny in a community-driven process. Finally, the last negative set was taken from the TripAdvisor (**TRI**) data set (Baccianella et al. [11]). It contains opinions about hotels. Table 5.1 summarizes both positive as negative data sets.

Table 5.1: Detailed information regarding data set I1.

	I1	**AMA**	**SLA**	**TRI**
Language	English	English	English	English
Size	2,861	3,000	3,000	3,000
Type	Reviews	Reviews	Comments	Opinions
Source	Amazon	Amazon	Slashdot	TripAdvisor
Availability	Public	Public	Private	Public

Data set **I1** is available at: `http://users.dsic.upv.es/grupos/nle/`.

5.3.2 HRM and Irony

Prior to assessing the basic IDM over **I1**, we performed a classification task applying HRM. The goal was to evaluate HRM's capabilities to accurately classify instances of irony, given the close link (in terms of effects) between humor and irony. To this end, the documents in **I1** were converted in terms vectors according to the criteria exposed in Section 4.5.2. HRM was evaluated as a single entity; i.e. no pattern was evaluated isolately. SVM was employed due to the best results were achieved with their algorithm. The classifier was trained considering 70% for training, and 30% for test. Results in terms of accuracy are detailed in Table 5.2.

As noted in this table, the accuracy is acceptable only concerning AMA (which is composed by funny web comments). However, for the ironic set (AMA), the

Table 5.2: Results applying HRM over I1.

	Accuracy
AMA	57,62%
SLA	73.28%
TRI	48.33%

result is quite poor: it hardly achieves 57% of accuracy. Considering the task here involved, it is evident the need of different, and perhaps, more complex patterns to automatically detect ironic instances.

Finally, while the results are as expected, we cannot obviate that some HRM's patterns seem to be relevant for the task. For instance, according to the results obtained by applying an information gain filter, we realized that *humor domain* and *lexical ambiguity* are the patterns that better perform in the classification. Therefore, their inclusion in IDM as a new pattern (funniness) is necessary to represent the relationship between humor and irony. The final list is thus integrated by seven patterns: **N-grams**, **descriptors**, **POS n-grams**, **funniness**, **polarity**, **affectiveness**, and **pleasantness**.

5.4 Basic IDM representation

In the following sections we will describe the experiments employing the basic IDM.

5.4.1 N-grams

This pattern is focused on representing irony in the simplest way: with sequences of n-grams (from order 2 up to 7) in order to find a set of recurrent words that could be commonly used to express ironic contents. Note that all the documents were preprocessed. First, stopwords were removed, then, all the reviews were stemmed. In addition, all the irrelevant terms were eliminated by applying a

$tf - idf$ measure ([92]). The measure is calculated according to Formula 5.1:

$$tfidf_{i,j} = tf_{i,j} \cdot idf_i = tf_{i,j} \cdot log = \frac{|D|}{|\{d_j|t_j \in d_j\}|} \tag{5.1}$$

where $|D|$ is the number of documents in D, and $|\{d_j|t_j \in d_j\}|$ is the number of documents in D containing t_i. This measure assesses how relevant a word is, given its frequency both in a document as in the entire corpus. Irrelevant words such as *t-shirt, wolf, tuscan, milk,* etc., were then automatically eliminated. The complete list of filtered words, stopwords included, contains 824 items. Examples of the most frequent sequences are given in Table 5.3.

Table 5.3: Statistics of the most frequent word n-grams.

Order	Sequences	Examples
2-grams	160	opposit sex; american flag; alpha male
3-grams	82	sex sex sex; fun educ game
4-grams	78	fun hit reload page; remov danger reef pirat
5-grams	76	later minut custom contribut product
6-grams	72	fals function player sex sex sex
7-grams	69	remov danger reef pirat fewer shipwreck surviv

5.4.2 Descriptors

Two metrics were implemented in order to provide tuned up sequences of words: *keyness*, concerning with the extraction of the most representative words; *clustering*, concerning with grouping similar words. Keyness is estimated by means of comparing the frequency of each word in a corpus against its frequency in a reference corpus (Google N-grams were used as reference corpus). This value is computed from the Log Likelihood test (Dunning [46]).

SenseClusters was used in order to group similar words into clusters. This toolkit integrates several clustering algorithms that operate either directly in the objects feature space or in the objects similarity space (Karypis [73])[9]. In addition, it implements various metrics for identifying similar contexts when building

[9]Available at: http://senseclusters.sourceforge.net/.

Table 5.4: SenseCluster parameters per experiment.

Space	Cl. Method	Cr. Function	LSA	Order	Cluststop
Vector	RB/Direct	UPGMA	Yes	Bi/Co	All
Similarity	Agglo/RBR/Graph	H2	Yes	Uni/Bi	Gap
Vector	Direct	H2	Not	Co	None
Similarity	RB	I2	Not	Bi	Pk

the clusters (Kulkarni and Pedersen [82]). Four experiments were performed to group similar words. Each requested five clusters based on different criteria. Table 5.4 summarizes the process[10]; whereas Table 5.5 shows the six more descriptive words obtained by applying the metrics above described.

Table 5.5: Descriptors obtained with keyness and clustering metrics.

Word	Keyness	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	design	dog	girl	contain	walk
2	human	wash	pick	dairi	pull
3	garment	dream	kid	cow	street
4	word	attract	mom	pour	sit
5	hope	teeth	watch	internet	hit
6	spirit	husband	woman	enjoi	arm

5.4.3 POS n-grams

Irony does not depend on specialized discourses. Therefore, it cannot be defined only in terms of words because the ways of expressing irony are as many as the words of any language. This pattern, thus, is intended to symbolize an

[10] Abbreviations in this table indicate: Cl. Method represents the clustering method employed; RB represents repeated bisections; Agglo represents agglomerative clustering; RBR represents repeated bisections globally optimized; Graph represents graph partitioning-based clustering; Cr. Function represents criterion function employed; LSA represents Latent Semantic Analysis representation; Order represents contexts; Bi represents bigrams; Uni represents unigrams; Co represents co-occurrences; Cluststop represents cluster stopping measure; Gap represents adapted gap statistics; Pk represents pk measures. Detailed explanation about these parameters is given in [73].

abstract structure through sequences of POS tags (hereafter, POS-grams) instead of only words. First, a statistical substring reduction algorithm (Lü et al. [90]) was employed in order to eliminate redundant sequences. For instance, if the sequences "he is going to look so hot in this shirt" and "he is going to look hot in this shirt" occur with similar frequencies in the corpus, then, the algorithm removes the last one because is a substring of the first one. After carrying out such algorithm, all reviews were labeled with FreeLing ([4]). The N-best sequences of POS-grams, according to orders 2 up to 7, are given in Table 5.6.

Table 5.6: Statistics of the most frequent POS-grams.

Order	Frequency	Sequences
2-grams	300	dt nn; nn in; jj nn; nn nn
3-grams	298	dt nn in; dt jj nn; jj nn nn
4-grams	282	nn in dt nn; vb dt jj nn
5-grams	159	vbd dt vbg nn jj
6-grams	39	nnp vbd dt vbg nn jj
7-grams	65	nns vbd dt vbg nn jj fd

5.4.4 Funniness

Although the results described in Section 5.3.2 are quite poor, we cannot obviate the relationship between humor and irony. Actually, various research works have provided evidence about it (for instance, see [152, 120, 34]). Therefore, it is important to consider, at least, some patterns to represent the role that humor plays as a effect of ironic statements. In this respect, funniness is intended to characterize irony in terms of the best humor patterns reported in the section previously cited: *lexical ambiguity and humor domain*. The process to represent funniness in **I1** consisted in representing irony as conducted in Section 4.3.2.1 and Section 4.4.1, respectively.

5.4.5 Polarity

As noted throughout this thesis, one of the most important properties of irony relies on conveying negative information through positive words. This pattern, thus, is intended to be an indicator about the correlation between positive and negative words in irony. Polarity representation was carried out according to the parameters reported in Section 4.4.2. However, instead of using both SentiWordNet and MSOL resources, we opted for selecting only the latter. This is due to MSOL's higher-coverage regarding phrase polarity.

5.4.6 Affectiveness

In order to enhance the quality of the information related to the expression of irony, we considered to represent information linked to emotional content. Therefore, like in previous patterns, we used the mechanism described in Section 4.4.4 to represent irony in terms of subjective contents such as emotions, feelings, moods, attitudes, and so on.

5.4.7 Pleasantness

The last pattern is an attempt to represent ideal cognitive scenarios to express irony. This means that, like with words, the contexts in which irony appears are quite numerous. Since it is impossible to make out all the contexts in which irony can appear, we defined a measure to represent favorable and unfavorable ironic contexts. This is done by estimating the degree of pleasure profiled by ironic contents. In order to represent pleasantness, we used the Dictionary of Affect in Language (Whissell [177]). This dictionary assigns a score of pleasantness to ∼9,000 English words. Such scores were obtained from human ratings. The range of scores goes from 1 (unpleasant) to 3 (pleasant). For instance, Whissell's Dictionary notes that the word $flower$ generally produces a pleasant affect (pleasantness = 2.75); in contrast, $crazy$ is quite unpleasant (pleasantness = 1.6).

5.5 Evaluation

In this section we will describe IDM's evaluation. This was done by means of a classification task. Two underlying goals were analyzed: i) pattern relevance; and ii) possibility of automatically finding ironic content.

Classifiers were evaluated by comparing the positive set against each of the three negative ones (AMA, SLA and TRI, respectively). All the texts were represented as frequency-weighted term vectors according to a representativeness ratio. Such ratio was estimated with Formula 5.2:

$$\delta(d_k) = \frac{\sum_{i,j} fdf_{i,j}}{|d|} \tag{5.2}$$

where i is the i-th pattern ($i = 1 \ldots 7$); j is the j-th feature of i; $fdf_{i,j}$ (*feature dimension frequency*) is the feature frequency j of pattern i; and $|d|$ is the length of the k-th document d_k. For patterns such as funniness, polarity, affectiveness, and pleasantness, we used an empirical representativeness threshold ≥ 0.5. Value $= 1$ was assigned if $\delta(d_k)$ exceeded the threshold, otherwise a value $= 0$ was assigned. For instance, let us consider example 37 to clarify this process:

37) "I was searching for clothes that speak to me... These pants not only spoke to me, they entered my soul and transformed me."

In such example *pant* and *soul* are features in pattern funniness; *cloth*, *speak* (twice), *enter*, and $transform$, are features in pattern pleasantness; and *search*, *speak* (twice), *pant*, *enter*, *soul*, and $transform$, are features that belong to pattern affectiveness. After summing all features j for patterns i, we obtain a frequency of 14, which is then normalized relative to the length of d_k. Its δ is 0.60. Thus, its representativeness ratio is $= 1$.

Figure 5.1 graphically shows how features j are distributed, according to their representativeness ratio, throughout each pattern i.

A different criterion was determined for patterns such as n-grams, descriptors, and POS-grams because we were not only interested in knowing whether or not the sequences appeared in the texts, but also in obtaining a measure to represent the degree of similarity among them. In order to define a similarity score, we used Jaccard similarity coefficient ([92]). According to Formula 5.3, similarity is obtained by comparing the overlapping between two sets given the union of both:

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{5.3}$$

Finally, classification accuracy was assessed employing three classifiers: NB, SVM, and decision trees (DT). They were trained with 5,861 instances (2,861 positive and 3,000 negative). 10-fold cross validation method was used as test. Global accuracy, precision, recall, and F-measure, were used to evaluate the performance of the classifiers. Results in Table 5.7.

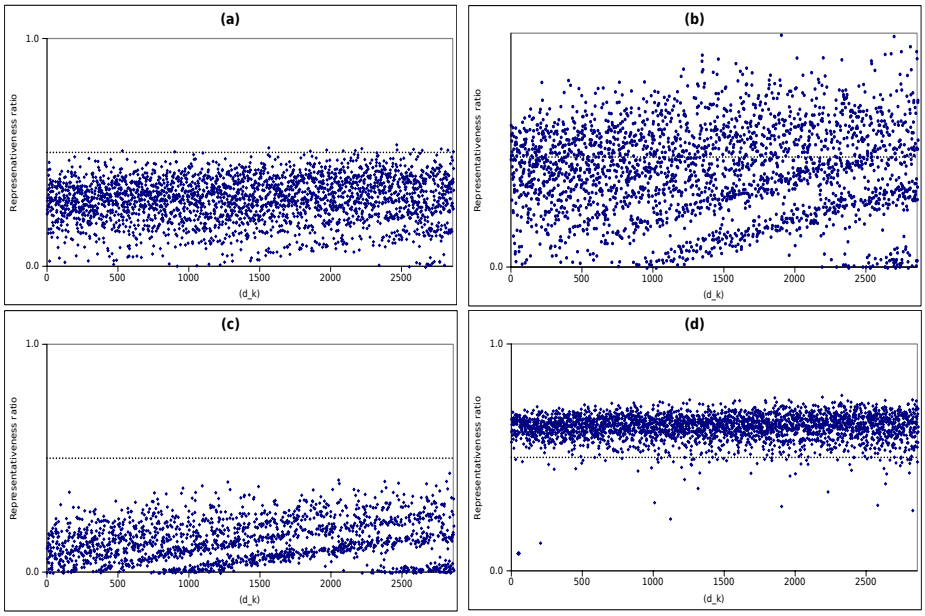

Figure 5.1: Representativeness ratios of patterns *funny* (a), *polarity* (b) (positive in red and negative in blue, respectively), *affectiveness* (c), and *pleasantness* (d). Axis x represents the ironic reviews whereas axis y depicts its representativeness ratio. Dotted line symbolizes representativeness threshold.

5.5.1 Discussion

Regarding the first goal (pattern relevance), our a-priori aim of representing irony in terms of seven general patterns seems to be fairly acceptable. On one hand, representativeness ratios do not provide sufficient information to undoubtedly state that the set of patterns are representative of ironic contents. Figure 5.1

graphically shows how the patterns are distributed according to their representativeness ratio. In addition, it makes evident that only pleasantness clearly exceeds the representativeness threshold. In contrast, although polarity often exceeds it, its performance is not constant for all reviews. The rest of patterns hardly reaches the threshold, being funniness the pattern which is closer to such threshold. On the other hand, according to the results shown in Table 5.7, although seems to be acceptable, they are not as expected: accuracy goes from 72% up to 89%; whereas a classifier that labels all texts as non-ironic would achieve an accuracy around 54%. Precision, recall, and F-measure rates support what mentioned: most classifiers obtained scores > 0.7. This means that the capabilities for differentiating an ironic review from a non-ironic one are not completely satisfactory. In addition, it is important to note how the model is not constant with the three negative sets. For instance, set TRI achieves the best results with all classifiers. In contrast, sets AMA and SLA obtain worse results. This behavior impacts on the learning process. For instance, note in Figure 5.2 how the learning is achieved with less instances regarding set TRI, whereas sets AMA and SLA require many more examples.

Table 5.7: Classification results.

		Accuracy	Precision	Recall	F-Measure
NB	AMA	72.18%	0.745	0.666	0.703
	SLA	75.19%	0.700	0.886	0.782
	TRI	**87.17%**	0.853	0.898	0.875
SVM	AMA	75.75%	0.771	0.725	0.747
	SLA	73.34%	0.706	0.804	0.752
	TRI	**89.03%**	0.883	0.899	0.891
DT	AMA	74.13%	0.737	0.741	0.739
	SLA	75.12%	0.728	0.806	0.765
	TRI	**89.05%**	0.891	0.888	0.890

With respect to the second goal (possibility of automatically finding ironic documents), we applied an information gain filter in order to verify patterns'

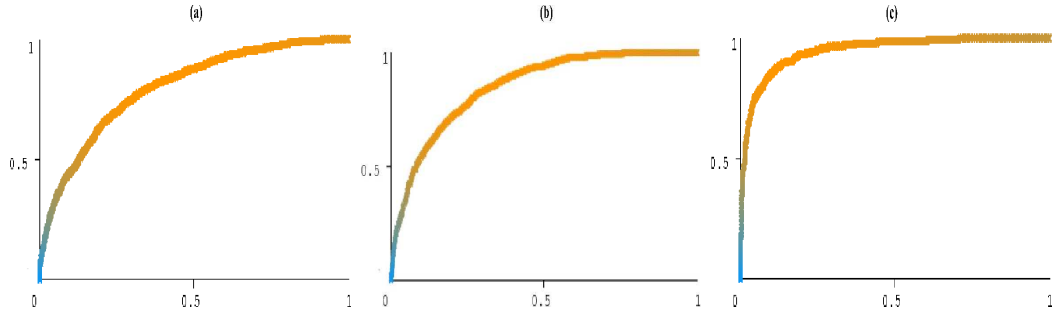

Figure 5.2: Learning curve according to sets AMA (a), SLA (b), and TRI (c).

relevance for determining ironic documents in terms of the different narrative *discourses* profiled by each negative data set. In Table 5.8 are detailed the most discriminating patterns per set. Based on the results depicted in this table, it is evident that pattern's relevance varies according to the negative set. For instance, when classifying the set AMA, it is clear that POS-grams (order 3), pleasantness and funniness, are the most informative ones; in contrast, pleasantness, n-grams (order 5) and funniness, are the most relevant regarding set SLA; whereas n-grams (order 2, 3 and 4) are the most discriminating when set TRI is classified.

Table 5.8: Most discriminating patterns per set.

AMA	SLA	TRI
3POS-grams	Pleasantness	2grams
Pleasantness	5grams	3grams
Funny	Funny	4grams
2POS-grams	Affectiveness	Pleasantness
4POS-grams	Positive words	5grams
Positive words	2POS-grams	Funny
Negative words	3POS-grams	Negative words
Affectiveness	Negative words	Positive words
5POS-grams	Descriptors	6grams
7POS-grams	7grams	Descriptors

5.5.2 Pattern Analysis

In this section we would like to highlight some observations with respect to each pattern. Regarding **n-grams**, it is important to note the presence of some interesting sequences that are common to all three sets. For instance: *pleasantly surprised*. However, we cannot define irony only in terms of sequences like these ones because they could represent domain-specific information, such as the bigram *customer service*. Similar situation concerning **descriptors**. Most sequences depend on the information included in the training set. Therefore, their usefulness is quite low.

With respect to **POS-grams**, the fact of focusing on abstract templates (instead of only on words) seems to be more effective. For instance, the sequence *noun + verb + noun + adjective* would represent more information than the sum of simple words: *[grandpa/hotel/bed] + [looks/appears/seems] + [years/days/months] + [younger/bigger/dirtier]*. However, the relevance of such sequences could be language-dependent; i.e. POS-grams are intended to represent prototypical templates given POS information, but POS information is obtained by means of applying either a deep or shallow syntactic parser; hence, their relevance could be co-related to syntactic restrictions.

Funniness seems to be a relevant pattern to express irony. However, its relevance could be supported by the type of information profiled in the positive set. Considering the comic trend in the reviews posted by Amazon's customers, it is likely that many of the features belonging to this pattern appeared in such reviews. For instance, in the following example the words in italics represent funny features: "I cannot write this review and be any happier with my purchase. It replaced at least one or two of my *family guy* t-shirts and is perfectly designed to hide my pit *stains* after *playing* twelve hours of xbox. I am an attractive *guy*. Slender, weak, and I have never shaved in my 19 years, but *sexy* as hell, and I cannot tell you how many *women* have flocked to me since my purchase". However, it is important to stress out that this pattern is equally discriminating for all sets, funny web comments included.

Concerning **polarity**, although in MSOL the number of negative words is higher than the number of positive words (more than 15,000 words of difference; cf. Section 4.4.2), the latter are more frequent in the ironic documents. This fact

supports the assumption about the use of positive information in order to express an underlying negative meaning: "The cool$_{\text{POS}}$, refreshing$_{\text{POS}}$ taste$_{\text{POS}}$ of the milk$_{\text{POS}}$ washed away my pain$_{\text{NEG}}$ and its kosher$_{\text{POS}}$ source$_{\text{POS}}$ of calcium$_{\text{POS}}$ wash away my fear$_{\text{NEG}}$".

Regarding **affectiveness**, its relevance is not as important as we have a-priori considered, despite it is one of the patterns used to discriminate set SLA: "Man, that was *weird* ... I think is *funny*, because there's a *good* overlap". However, if we take into account the whole accuracy for this set, then we can conclude that its relevance is minor. Nonetheless, we still consider that affective information is a valuable factor which must be taken into account in order to provide rich knowledge related to subjective layers of linguistic representation.

Instead, the role played by **pleasantness** is significant. Despite this pattern is not the most discriminating, its effectiveness for increasing classification accuracy is remarkable. For instance, consider the following ironic sentence: "I *became* the *man* I *always dreamed* I *could be all* those *nights staying up* late watching *wrestling*", where most of its constituents are words whose pleasantness score is ≥ 2.5; i.e. these words (in italics) should tend to communicate information related to favorable contexts to express irony.

5.6 Complex Irony Detection Model

In Section 5.3 we described IDM's basic version. It is understood as basic due to most patterns are represented in terms of low-level properties of irony. Moreover, due to the close relationship between humor and irony, such version incorporates many patterns that showed to be useful in humor processing. Therefore, its scope could be limited to detect ironic contents that, according to our definition of irony, are closer to irony's effect (humorous interpretation) than irony's aim (communicate opposition).

In this section we will describe an IDM that is focused on representing irony in terms of more complex patterns. This new version recaptures the essence of the basic IDM by making a fine-grained representation of some its patterns[11].

[11] Some of the patterns described in Section 5.3, such as polarity, affectiveness, or pleasantness, as well as n-grams, are now improved and incorporated to this new version.

The complex IDM is organized according to four types of conceptual patterns: **signatures, unexpectedness, style,** and **emotional scenarios**.

They are intended to capture both low-level and high-level properties of textual irony. In this respect, the patterns are given in terms of textual elements in order to represent the core of the phenomenon; in particular, those aspects that lead people to explicitly tag any text as ironic. Every pattern, save for unexpectedness, is represented with three dimensions; unexpectedness is represented with just two dimensions. The dimensions are listed and discussed below.

i. **Signatures**: concerning pointedness, counter-factuality, and temporal compression.

ii. **Unexpectedness**: concerning temporal imbalance and contextual imbalance.

iii. **Style**: as captured by character-grams (c-grams), skip-grams (s-grams), and polarity skip-grams (ps-grams).

iv. **Emotional contexts**: concerning activation, imagery, and pleasantness.

All these patterns are described in the following sections.

5.6.1 Signatures

This pattern is focused on exploring irony in terms of specific textual markers or *signatures*. It is largely characterized by typographical elements such as punctuation marks and emoticons, as well as by discursive elements that suggest opposition. Formally, we consider signatures to be textual elements that throw focus onto certain aspects of a text. For instance, from a shallow perspective, quotes or capitals are often used to highlight a concept or an attribute (e.g.“ I HATE to admit it but, I LOVE admitting things”); in contrast, from a deeper perspective, adverbs often communicate contradiction. (e.g. “Saying we will destroy terrorism is *about* as meaningful as saying we shall annihilate mocking”).

Signatures is represented in three dimensions: *pointedness*, *counter-factuality*, and *temporal compression*. **Pointedness** is focused on explicit marks which, according to the most relevant properties of irony, should reflect a sharp distinction

concerning the conveyed information. The set of features here considered are punctuation marks such as ., . . . , ;, ?, !, :, ,, emoticons[12], quotes, and capitalized words. In contrast, **counter-factuality** is focused on implicit marks; i.e. discursive terms that hint at opposition or contradiction such as *about*, *nevertheless*, *nonetheless*, or *yet*. We use some adverbs that hint at negation, as well as their synonyms in WordNet, to represent this dimension[13]. The last dimension, **temporal compression**, is focused on identifying elements related to opposition in time; i.e. terms that indicate an abrupt change in a narrative. These elements are represented with a set of temporal adverbs such as *suddenly*, *now*, *abruptly*, and so on. The complete list of elements to capture both counter-factuality and temporal compression is given in Appendix C.

5.6.2 Unexpectedness

Irony often exploits incongruity, unexpectedness and the ridiculous to ensure that an insincere text is not taken literally by a listener. Lucariello [91] proposes the term *unexpectedness* to represent the "imbalances in which opposition is a critical feature". She notes that surprise is a key component of irony, and even goes as far as to claim that unexpectedness underlies all ironic situations. In this respect, we conceive **unexpectedness** as a pattern to capture both temporal and contextual imbalances in an ironic text. Lucariello defines such imbalances in terms of oppositions or inconsistencies within contexts or situations, or between roles, or across time-frames (e.g. "The wimp who grows up to be a lion tamer", or "A kiss that signifies betrayal"; cf. [91]). This pattern is represented in two dimensions. First, **temporal imbalance** is used to reflect the degree of opposition in a text with respect to the information profiled in present tense and past tense. Unlike temporal compression, here we are focused on analyzing divergences related only to verbs (e.g. "I *hate* that when you *get* a girlfriend most of the girls that *didn't* want you all of a sudden *want* you!). **Contextual imbalance**, in contrast, is intended to capture inconsistencies within a context. In order to represent contextual imbalances, we estimate the semantic similarity of a text's concepts to each other. Resnik measure, implemented in WordNet::Similarity module (Pedersen

[12]The complete list with emoticons is given in Appendix C.
[13]Version 3.0 was used.

et al. [116])[14], is used to calculate the pair-wise semantic similarity of all terms in a text. Normalized semantic relatedness score is determined by summing the highest scores (across different senses of the words in the text) and dividing the result by the length of the text. The driving intuition here is: the smaller the semantic inter-relatedness of a text, the greater its contextual imbalance (suggesting an ironic text); the greater the semantic inter-relatedness of a text, the lesser its contextual imbalance (suggesting a non-ironic text). Lastly, we calculate the contextual imbalance of a text as the reciprocal of its semantic relatedness (that is, 1 divided by its semantic relatedness score).

5.6.3 Style

According to one dictionary definition, style is a "distinctive manner of expression". It is this type of fingerprint, imparted by the stylistic characteristics of text, that allows people (and machines) to discriminate, for instance, Shakespeare's work from that of Oscar Wilde. Within the current framework, the concept of style refers to recurring sequences of textual elements that express relatively stable features of how a text is appreciated, and which might thus allow us to recognize stylistic factors that are suggestive of irony. **Style** is captured in the current model using three types of textual sequences: **character n-grams (c-grams)**, **skip-grams (s-grams)**, and **polarity s-grams (ps-grams)**.

First, **c-grams** captures frequent sequences of morphological information, such as affixes and suffixes (e.g. -ly. Cf. Stamatatos [157]). In order to obtain the best c-grams sequences, we consider sequences of 3 to 5 characters. In the second type of sequence, **s-grams**, we widen the scope to consider whole words. But instead of looking for sequences of adjacent words (simple n-grams), we look for word sequences that contain (or skip over) arbitrary gaps; hence the name skip-grams (cf. Guthrie et al. [64] and Chin-Yew and Och [28]). For instance, in the sentence "There are far too many crazy people in my psychology class", a typical 2-gram is represented by the sequences *there are*, whereas a 2-sgram, with a 1 token gap, would be *there far*. Gaps are limited to 2 or 3 word skips, because longer sequences are not very common.

The last sequence type, **polarity s-grams**, provides sequences of abstract

[14] Available at: http://wn-similarity.sourceforge.net/.

categories based on s-grams; i.e. we can produce an abstract structure for a text from sequences of positive and negative terms instead of specific content words or characters. The intuition here is that one generally employs positive terms to communicate a negative meaning when using irony; for example, there is usually a positive ground in an ironic comparison that conveys a critical meaning (cf. Veale and Hao [173]). As in the case of s-grams, the gaps in ps-grams are limited to 2-word and 3-word skips only. To provide tags for s-grams, as well as to observe the distribution of positive and negative terms in each text, we use MSOL. As an example of this representation, consider the text "I need more than luck. I need Jesus and I'm an atheist…". If considering only 2-word skips, the abstract representation provided by the terms labeled with positive or negative polarity is the following sequence of tags (after removing stop words): pos_{need} pos_{jesus} $neg_{atheist}$.

5.6.4 Emotional Contexts

Language is one of our most natural mechanisms of conveying information about emotional states. Textual language provides specific tools on its own, such as the use of emoticons in web-based content to communicate information about moods, feelings, and our sentiments toward others. Online ironic expressions often use such markers to safely realize their communicative effects (e.g. "I feel so miserable without you, it is almost like having you here :P"). **Emotional contexts** capture information that goes beyond grammar, and beyond the positive or negative polarity of individual words. Rather, this pattern attempts to characterize irony in terms of elements that symbolize abstractions such as overall sentiments, attitudes, feelings and moods, in order to enhance the schema of favorable and unfavorable contexts for the expression of irony.

Adopting a psychological perspective, we represent emotional contexts in terms of the categories described by Whissell [178], namely *activation*, *imagery*, and *pleasantness*. These categories (or *dimensions* in our terminology) attempt to quantify the emotional content of words in terms of scores obtained from human raters. **Activation** refers to the degree of response, either passive or active, that humans exhibit in an emotional state (e.g. *burning* is more active than *basic*). **Imagery** quantifies how easy or difficult is to form a mental pic-

ture for a given word (e.g. it is more difficult to mentally depict *never* than *alcoholic*). **Pleasantness** (as described in Section 5.4.7) quantifies the degree of pleasure suggested by a word (e.g. *love* is more pleasant than *money*). In order to represent these dimensions, we use Whissell's Dictionary of Affect in Language. The range of scores go from 1.0 (most passive, or most difficult to form a mental picture, or most unpleasant) to 3 (most active, or easiest to form a mental picture, or most pleasant). For instance, *flower* is passive (activation = 1.0), easily representable (imagery = 3.0), and generally produces a pleasant affect (pleasantness = 2.75); in contrast, *crazy* is more active (1.33), moderately representable (2.16), and quite unpleasant (1.6); whereas *lottery* is very active (3.0), moderately representable (2.14), and mostly pleasant (2.4).

Finally, to aid understanding every pattern along with its dimensions, Appendix D provides examples from each one.

5.6.5 Data Set I2

A new data set with ironic texts was integrated in order to assess IDM's capabilities[15]. Like in **I1**, we have opted for collecting an evaluation data set with examples that are a priori labeled as ironic by their users. In this respect, although the manual annotation is supposed to be the best way of obtaining reliable information in corpus-based approaches, in tasks like this one, such approach is hard to be achieved. First, there are not formal elements to accurately determine the necessary components to label any text as ironic. Then, in the case that we had a prototype of ironic expressions, its discovery is a time-consuming manual task (according to Peters and Wilks [119], this is a reason for the restricted number of attested instances of figurative language in texts). Finally, (as claimed throughout these chapters) linguistic competence, personal appreciation, moods, and so on, make irony quite subjective; therefore, any annotation agreement faces the complexity of standardizing annotation criteria. That is why we decided to use examples labeled with user-generated tags, which are **intentionally** focused on particular topics[16]. By opting for this approach, we eliminate the inconveniences

[15]Like in the last chapter, we built a new data set in order to widen IDM's scope, as well as to assess it with another kind of instances.

[16]Recall the role of intentionality in the process of communicating the figurative intent. Cf. Section 2.3.

above mentioned: such examples are self-annotated (thus, it is not necessary the presence of "human annotators" to manually (and subjectively) collect and label positive examples). Moreover, positive examples can be retrieved effortless taking advantage of their tags (thus, it is likely having thousands of examples in a short time). Lastly, by applying the model to examples with user-generated tags, according to our objective, we broaden our analysis beyond literary uses of irony[17].

In this context, we focused on one of the current trendsetters in social media: the Twitter micro-blogging service. The membership criterion for including a tweet in **I2** is that each should contain a specific *hashtag* (i.e. the user-generated tag according to Twitter's terminology). The hashtags selected are #irony, in which a tweet explicitly declares its ironic nature, and #education, #humor, and #politics, to provide a large sample of potentially non-ironic tweets. These hashtags were selected because when using the #irony hashtag, users employ (or suggest) a family-resemblance model of what it means (cognitively and socially) for a text to be ironic. In this respect, a text so-tagged may not actually be ironic by any dictionary definition of irony, but the tag reflects a tacit belief about what constitutes irony. Based on these criteria, we collate an evaluation data set **I2** of 40,000 tweets, which is divided into four parts, comprising one self-described positive set and three other sets that are not so tagged, and thus assumed to be negative. Each set contains 10,000 different tweets (though all tweets may not be textually unique). We assume therefore that **I2** contains 10,000 ironic tweets and 30,000 largely non-ironic tweets. Some statistics[18] are given in Table 5.9. It is worth noting that all the hashtags were removed. No further preprocessing was applied at this point.

In addition, Monge Elkan distance was employed in order to estimate the overlap between the ironic set and each of the three non-ironic ones. This measure, according to Monge and Elkan [105], allows for gaps of unmatched characters, [and thus], it should perform well for many abbreviations, and when fields have

[17] Although we have opted for this approach, in Chapter 6 we detail a task in which IDM was assessed on a manual labeling.

[18] Type-level statistics are not provided because these tweets contain many typos, abbreviations, user mentions, etc. At this point, there was no standardization processing to remove such misspellings. Therefore, any statistics regarding types would be biased.

Table 5.9: Statistics in terms of tokens per set concerning data set I2.

	#irony	#education	#humor	#politics
Vocabulary	147,671	138,056	151,050	141,680
Nouns	54,738	52,024	53,308	57,550
Adjectives	9,964	7,750	10,206	6,773
Verbs	29,034	18,097	21,964	16,439
Adverbs	9,064	3,719	6,543	4,669

missing information or minor syntactical differences. Therefore, it should help us minimizing the likelihood of noise arising from the presence of typos, common misspellings, and the abbreviations that are endemic to short texts. Moreover, since we are dealing with tokens instead of types, the metric was computed using the approach outlined by Cohen et al. [33]. In such implementation, authors considered a scheme in which the substrings are tokens. Formula 5.4 describes the algorithm[19]; whereas results are shown in Table 5.10.

$$sim(s,t) = \frac{1}{k} = \sum_{i=1}^{K} \max_{j=1}^{L} sim'(A_i, B_j) \tag{5.4}$$

Monge Elkan distance approaches 1.0 as the data sets share more of their vocabulary. Results in Table 5.10 thus suggest a significant difference between the vocabularies of the four tweet sets. As one might expect, this difference is least pronounced between sets #irony and #humor. After all, irony is most often used to communicate a humorous attitude or insight, as in examples 38 and 39 in which both tweets were tagged as #irony:

38) Just think: every time I breathe a man dies. —A friend: Have you tried to do something about bad breath?

39) I find it humorously hypocritical that Jeep advertises on TV about how we shouldn't watch tv in favor of driving their vehicles.

[19]Prior to computing the distance between texts, all words were stemmed using Porter algorithm, and then all the stopwords were eliminated. Accordingly, the distance measure better reflects the similarity in core vocabularies rather than similarity in shallow forms.

Table 5.10: Monge Elkan distance among sets.

	sim(s,t)
(#irony, #education)	0.596
(#irony, #humor)	0.731
(#irony, #politics)	0.627
(#education, #humor)	0.593
(#education, #politics)	0.605
(#humor #politics)	0.648

5.7 Evaluation

IDM was evaluated in two ways: i) by considering the appropriateness or representativeness of different patterns to irony detection; and ii) by considering the empirical performance of the model on a tweet classification task. Both considerations are evaluated in separate and independent experiments. When evaluating representativeness we look to whether individual patterns are linguistically correlated to the ways in which users employ words and visual elements when speaking in a mode they consider to be ironic. The classification task, in contrast, evaluates the capabilities of the model as a whole, focusing on the ability of the entire system of patterns to accurately discriminate ironic from non-ironic tweets.

5.7.1 Representativeness of Patterns

In the first experiment, each of the 40,000 tweets is converted into a vector of term frequencies[20] according to the representativeness ratio described in Section 5.5. This ratio is intended to provide a global insight into the effectiveness of the model for actually identifying patterns in the ways that users employ the four conceptual patterns when genuinely speaking ironically. Furthermore, such ratio is employed due to we need to know that the model is not simply detecting artifacts of the ways that users employ the #irony hashtag, or worse, artifacts of the way they use the #education, #humor, or #politics hashtags. By characterizing tweets

[20] All tweets underwent preprocessing, in which terms were stemmed and both hashtags as stopwords were removed.

with this ratio, we obtain global insights about the distribution of patterns in all sets; this will allow us to determine those which are more likely to express ironic meanings[21].

The *representativeness* of a document d_k is estimated using Formula 5.2. Like in in Section 5.5, if $\delta(d_k)$ is ≥ 0.5, then document d_k is assigned a representativeness value of 1 (i.e. pattern i is representative of d_k); otherwise, a representativeness value of 0 (not representative at all) is assigned.

40) "I love ugly people LIKE you :)".

For instance, in example 40 appear $LIKE$ and :) which belong to the dimension *pointedness* in the *signatures* pattern; *love* and *people* belong to the dimension *pleasantness* in the *emotional contexts* pattern; and the sequence of tags *neg pos* for words *ugly people* belong to the dimension *ps-grams* in the *style* pattern. After summing the frequencies of these elements we obtain a score of 5, which is then normalized relative to the length of the tweet (i.e. 7) that gives a global δ of 0.71. This suggests that the previous patterns are representative of this tweet. The overall representativeness per set (shown in Table 5.11) is obtained by summing every single δ and smoothing this score by dividing it by the size of the set; i.e. 10,000 documents.

As shown by results in Table 5.11, all dimensions, except pointedness and temporal imbalance, seem to be sufficiently indicative to represent ironic tweets from educational, humorous and political tweets. On a set level then, there appear to be patterns used in a text that correlate with the ways in which people use irony. Consider, for instance, counter-factuality dimension, whose textual elements are terms that suggest contradiction. It is evident that terms that suggest this dimension appear most often in the set #irony. In addition, ironic tweets do not score well overall on semantic relatedness, which means they score well on the contextual imbalance dimension. This, in turn, supports our hypothesis about the reduced inter-word semantic relatedness of ironic tweets. This is clearer in Table 5.12, in which semantic relatedness per set is given.

[21] In Appendix E we present the probability density function associated with the representativeness ratio in order to make clear that our model really captures some aspects of irony (as opposed to the alternative where the classification of irony could be a by-product if IDM captured idiosyncratic features of the negative sets).

Table 5.11: Overall pattern representativeness per set.

	Irony	Education	Humor	Politics
Signatures				
Pointedness	0.314	0.268	**0.506**	0.354
Counter-Factuality	**0.553**	0.262	0.259	0.283
Temporal Compression	**0.086**	0.054	0.045	0.046
Unexpectedness				
Temporal Imbalance	0.769	0.661	**0.777**	0.668
Contextual Imbalance	**1.121**	0.994	0.788	0.904
Style				
c-grams	**0.506**	0.290	0.262	0.395
s-grams	**0.554**	0.195	0.144	0.161
ps-grams	**0.754**	0.481	0.494	0.534
Emotional Scenarios				
Activation	**1.786**	1.424	1.482	1.324
Imagery	**1.615**	1.315	1.378	1.218
Pleasantness	**1.979**	1.564	1.660	1.394

Moreover, with respect to dimensions of *style* and *emotional scenarios* patterns, the scores achieved for each indicate a greater presence of textual elements related to these dimensions in the ironic set, especially as regards the scores for *s-grams* and *pleasantness* dimensions.

Graphs depicted in Figure 5.3, on the other hand, show the distribution of positive and negative words in terms of their position in the tweet (X axis) and their overall representativeness ratio (Y axis). It is interesting to note how the preponderance of negative terms in set #irony is concentrated in the first 7 words of the texts, whereas the frequency of positive terms is lower but relatively constant across texts. In sets #education and #politics, in contrast, the distribution seems to be just the contrary: more positive terms are found in the first 6 words of a text, while negative terms appear with relative constancy and a lower frequency throughout the text. In set #humor, the negative terms tend to appear

Table 5.12: Semantic relatedness per set.

#irony	**0.892**
#education	1.006
#humor	1.270
#politics	1.106

with higher frequency between word positions 3 and 8, while positive terms tend to occur between word positions 1 and 4. This behavior is supposed to hint at that part of the utterance in which irony produces its effect, and on which the greatest energy should be placed.

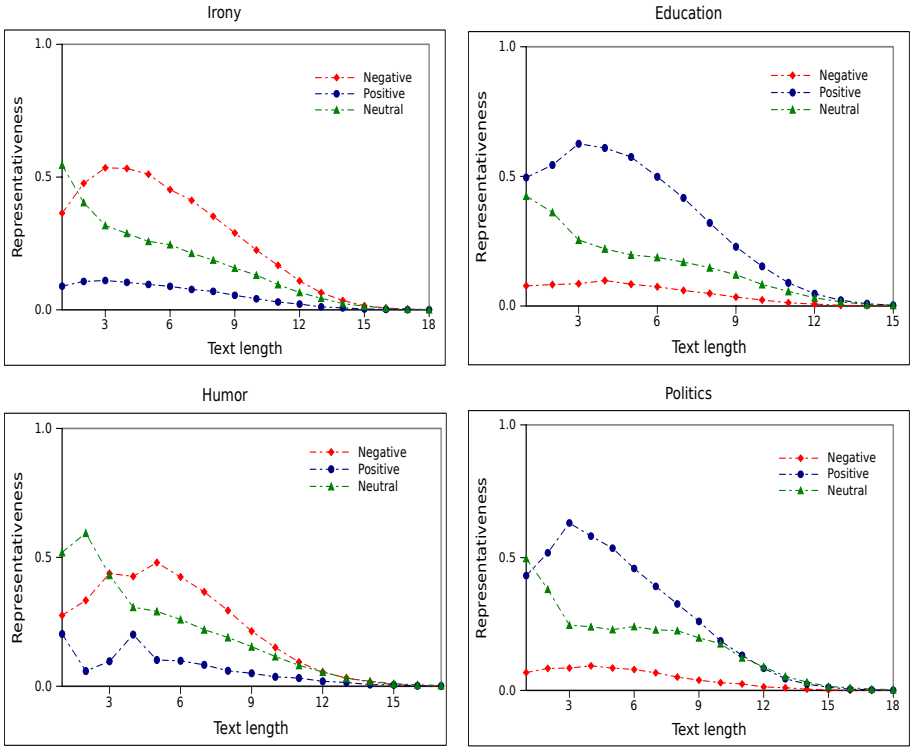

Figure 5.3: Distribution of positive, negative and out of vocabulary (neutral) terms in I2.

5.7.2 Classification Tasks: Results and Discussion

We use two different classifiers to evaluate the ability of the model to automatically discriminate each text set. We perform a series of binary classifications, between #irony *vs.* #education; between #irony *vs.* #humor; and between #irony *vs.* #politics. In each case, the patterns are added incrementally to the classification processing in order to determine their relative value to the classifier. Thus, the classifiers first use *signatures* pattern; *unexpectedness* pattern is then added; and so on. Two distributional scenarios are evaluated: i) balanced distribution, comprising 50% positive texts and 50% negative texts; ii) imbalanced distribution, with a more realistic mix of 30% ironic texts and 70% non-ironic texts. NB and DT algorithms are used to perform the classification. We choose these particular algorithms for two reasons: first, we use NB since our experiments are focused on the presence or absence of patterns. These are represented by boolean attributes that are treated as independent variables assigned to the class with maximum probability (Witten and Frank [183]); and second, DT are used in order to analyze the sequences of decisions regarding the relevance of such patterns, and to be able to make further inferences about them.

Classifiers, for both balanced and imbalanced distributions, were tested using 10-fold cross validation. Results shown in Figure 5.4 indicate an acceptable performance on the automatic classification. The model evidently improves its performance in almost all cases (with the exception of *emotional scenarios* pattern) each time a new pattern is added (e.g. the accuracy increases after considering at least two or three patterns). Based on the accuracy, a trivial classifier that labels all texts as non-ironic would achieve the accuracy of 50%; our entire model, instead, achieves an accuracy higher than the baseline (over 75%). This suggests that the four conceptual patterns cohere as a single framework that is able to clearly discriminate positive (ironic) tweets from negative (non-ironic) tweets. Similar results are reported by Carvalho et al. [27]. By exploring oral and gestural features to detect irony in user comments, they achieve accuracies ranging from 44.88% to 85.40%. Concerning Figure 5.5, results are not as good as in the balanced distribution. A classifier that labels all texts as non-ironic would achieve an accuracy of 70%, whereas in this case we see that our model hardly exceeds this baseline when considering just a couple of patterns (from 68% to

74%). When the entire model is considered, the obtained accuracy is 6% higher than the baseline. This evidences the difficulty of identifying irony in data sets where the positive examples are very scarce; i.e. it is easier to be right with the set when ironic instances statistically appear quite often than with the one where they barely appear. This situation, nonetheless, is the expected in real scenarios in which the absence of positive data, or the lack of labeled examples, is the main practical difficulty. However, this first approach has shown some advances when dealing with distributional issues. Our efforts, thus, must be addressed to find more discriminating patterns that allow us to increase current accuracy on both balanced and imbalanced scenarios.

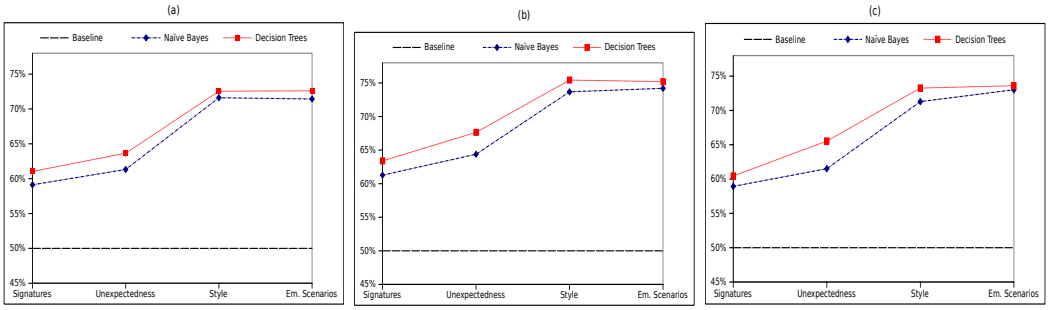

Figure 5.4: Classification accuracy regarding irony *vs*. education (a), humor (b), and politics (c), considering a balanced distribution.

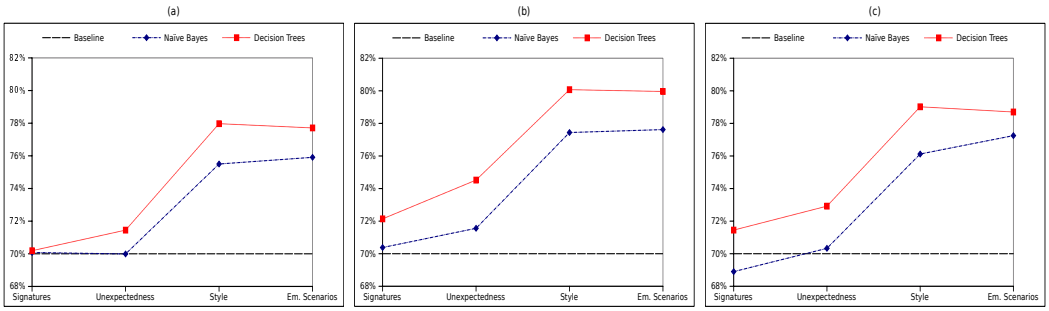

Figure 5.5: Classification accuracy regarding irony *vs*. education (a), humor (b), and politics (c), considering an imbalanced distribution.

Table 5.13, on the other hand, presents the results obtained in terms of precision, recall and F-Measure on both balanced and imbalanced distributions. These results support our intuitions about irony. While the results reported by Davidov

et al. [41] and Burfoot and Baldwin [24] relate to analyses of different figurative devices, such as sarcasm and satire respectively, and are thus not entirely comparable to the current results, such a comparison is nonetheless warranted. Our model obtains F-Measures that are comparable to, or better than, either of these previous approaches. For instance, the former study reports a highest F-Measure of 0.545 on a corpus collected from Twitter; while the latter reports a highest F-Measure of 0.798 for a corpus of newswire articles. In the current study, the highest F-Measure obtained is a score of 0.768 in the balanced distribution.

Table 5.13: Precision, Recall and F-Measure regarding i) balanced distribution, and ii) imbalanced distribution.

		Precision	Recall	F-Measure
		i | ii	i | ii	i | ii
NB	#education	0.73 | 0.60	0.66 | 0.62	0.69 | 0.61
	#humor	**0.79** | 0.64	0.68 | 0.59	**0.73** | 0.62
	#politics	0.75 | 0.60	**0.69** | 0.60	0.72 | 0.60
DT	#education	0.76 | 0.70	0.66| 0.52	0.70 | 0.60
	#humor	**0.78** | 0.75	**0.74** | 0.47	**0.76** | 0.58
	#politics	0.75 | 0.69	0.71 | 0.52	0.73 | 0.59

To further assess IDM's capabilities, an additional variation of the classification task was undertaken. It is based on considering positive set (irony) against all three negative ones (education, humor, politics). Classification was performed using DT, and evaluated using 10-fold cross validation. We considered both a balanced distribution (10,000 positive instances and 3,333 of each negative set) and an imbalanced distribution (10,000 positive instances and all 30,000 negative instances). Results show a similar behavior to those previously observed. When using a balanced distribution, the accuracy is lower but precision, recall and F-measure are all significantly higher (72.30%, 0.736, 0.695, 0.715, respectively). Conversely, when using an imbalanced distribution, the accuracy is higher but precision, recall and F-measure are not (80.44%, 0.661, 0.447, 0.533, respectively). These results support our belief that a system of linguistic patterns can capture fine-grained elements used by people when communicating what they believe to

be ironic statements[22].

While IDM operates with a system of patterns, yet each pattern can be analyzed in terms of information gain to determine its individual contribution to the discrimination power of the model. Figure 5.6 presents the results of an information gain filter (Y axis) on each of the dimensions of our four patterns (X axis)[23].

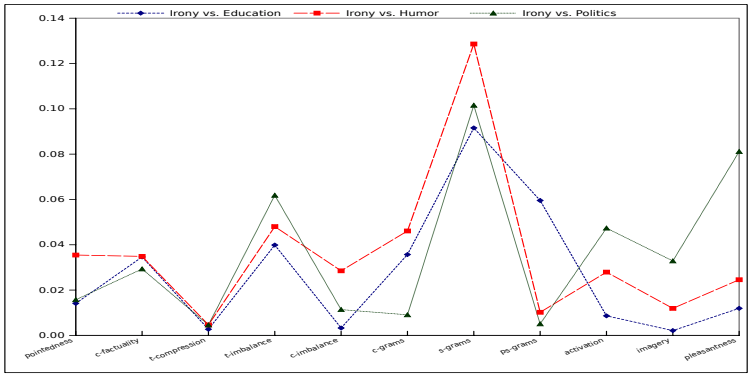

Figure 5.6: Relevance of every single dimension according to its information gain value.

Information gain results show that there are dimensions that appear to be not so useful in the discrimination of ironic tweets (e.g. *temporal compression* dimension of signatures, *contextual imbalance* dimension of unexpectedness, *ps-grams* dimension of style, and *imagery* dimension of emotional scenarios). However, the apparent minor usefulness is a function of the types of texts that are to be discriminated. Consider, for instance, *ps-grams* dimension: while it exhibits a very low information gain when discriminating #irony *vs.* #humor and #irony *vs.* #politics, this score increases significantly when discriminating #irony *vs.* #education. A similar situation holds with respect to *contextual imbalance* dimension: when considering the discrimination of #irony *vs.* #humor, the score is

[22]Finally, the complex IDM was evaluated by applying a SVM classifier to **I1** in order to compare the results obtained with the basic IDM. The results, in terms of accuracy, showed a slight improvement concerning the ironic set (AMA): whereas the basic IDM reached 75.75% of accuracy, the complex IDM achieved 83.04%. This corroborates IDM's capabilities to detect some common patterns concerning the verbalization of irony.

[23]Only the information gain values for the balanced distribution are displayed. The imbalanced case is not considered here since the values follow a similar distribution.

acceptable, whereas on the remaining two negative sets, the score is unacceptably low. In addition, there are dimensions that exhibit a strong relevance to the irony task (e.g. *temporal imbalance* dimension of unexpectedness, *s-grams* dimension of style, and *pleasantness* dimension of emotional scenarios). Once again, this relevance is also a function of the types of texts that are to be discriminated. This behavior suggests that these patterns cohere well together, so that while no single pattern captures the essence of irony, all four together provide a useful linguistic framework for detecting irony at a textual level.

5.8 Summary

In this chapter we have described two different models concerning the irony detection task: the basic IDM and the complex IDM.

In Section 5.1 we began by setting up IDM's scope in terms of tasks as diverse as sentiment analysis, opinion mining, or advertising. Then, in Section 5.3 we detailed the basic IDM. Moreover, in Subsection 5.3.1 we introduced a data set concerning ironic examples: data set **I1**. Lastly, in Section 5.5 we reported experiments and evaluation regarding the basic IDM.

The complex IDM was described in Section 5.6. The novel data set **I2** concerning different types of ironic examples was introduced in Subsection 5.6.5. Finally, in Section 5.7 we detailed all the experiments and mechanisms to comprehensively assess IDM's capabilities.

6

Applicability of Figurative Language Models

#Toyota's new slogan: moving forward (even if u don't want to)

DATA SET I2
REYES ET AL. [142]

This chapter will be focused on **assessing both HRM as IDM** on tasks such as information retrieval, opinion mining, sentiment analysis, or trend discovery. Such tasks are intended to represent real scenarios concerning FLP. First, we will concentrate on evaluating the applicability of HRM, and then, of IDM. After each task, some final remarks will be given.

6.1 Aim

Throughout this chapter we will describe three different tasks in which the proposed models will be evaluated. Our aim is to corroborate models' capabilities facing scenarios beyond the evaluation data sets previously described (H1, H2, H3, H4 and I1, I2, respectively). The set of tasks involved in this evaluation deals with information retrieval, opinion mining, sentiment analysis, and trend discovery. In addition, we will outline the applicability of our models concerning tasks such as machine translation, vandalism detection, or analysis of political

speech.

6.2 Humor Retrieval

This task is focused on the retrieval of humorous texts; more precisely, on the retrieval of funny comments on web items. If funny comments are retrieved accurately, they would be of a great entertainment value for the visitors of a given web page (see Section 4.1). To this end, we use a new large-scale data set for humor retrieval: the Slashdot news web site which contains human-annotated funny comments on a large scale.

6.2.1 Web Comments Data Set

Evaluation data set consists of about 3.8 million comments retrieved from the Slashdot news web site. It includes all comments on articles published between January 2006 and June 2008. Comments on Slashdot are categorized in a community-driven process. The comment categories include the following user-generated tags: *funny, informative, insightful, interesting, off-topic, flamebait,* and *troll*. This data set has first been used by Potthast [124] for measuring the descriptiveness of web comments.

The following comments are concrete examples about how the Slashdot community, depending on the meaning they want to communicate, categorize their personal comments by means of the previous tags.

41) *Re:Number of movies (Score:5, Insightful).*

"I believe that prior to this particular month, HD-DVD was consistently ahead of Blu-Ray. Declaring a winner based on a single months' worth of statistics (especially at this early point when both formats are in their infancy) is utterly idiotic."

42) *Re:Number of movies (Score:1, Interesting).*

"True. However, it can be used as a tool to gage the trend to try to predict WHERE the winning format will fall."

43) *Re:Number of movies (Score:2, Funny).*

"So let me get this straight: A single data point can be used as a "tool" to gage the trend? No shit?"

44) *Re:Number of movies (Score:2, Funny).*

"6 months of data is a single data point? No shit? It's not a single data point. It's the volume of title sales over 6 months. RTFA and maybe... just MAYBE click the links."

The amount of comments on Slashdot does not allow for every comment to be categorized, so that we restrict ourselves to the 1.068,953 categorized comments. They are divided into four classes: funny, informative, insightful, and negative. The latter contains comments from categories off-topic, flamebait, interesting and troll. The funny class is the smallest of the four; it contains 159,153 comments. In order to avoid problems related to class imbalance, samples of 150,000 comments from each of the other three classes are employed in the experiments; i.e. 600,000 comments in total.

6.2.2 Experiments and Discussion

The experiments are carried out with two algorithms: NB and DT. All comments were represented by means of HRM's patterns, according to the humor average score. Training sets contain 100,000 comments per class; whereas test sets contain 50,000 comments per class. All classifiers consider the classes funny *versus* informative (c_1), insightful (c_2), and negative (c_3), respectively. Table 6.1 comprises the results.

Table 6.1: Classification accuracy of funny vs. informative (c_1), insightful (c_2), and negative (c_3), respectively.

	NB	DT
c_1	73.54%	74.13%
c_2	79.21%	80.02%
c_3	78.92%	79.57%

Results in Table 6.1 show that HRM's patterns present a similar discriminative power in distinguishing funny comments from non-funny comments (80% *vs.* 85% in H4). In this respect, although non-funny training examples are of the same text type as the funny ones (web comments in which topic, vocabulary, or target audience share a common source: Slashdot.com), HRM shows interesting capabilities in classifying test sets. Consider, for instance, that a classifier that labels all web comments as non-funny would achieve 50% of accuracy. In addition, note that humor in web comments is produced by exploiting different linguistic mechanisms. For instance, humor in one-liners is often caused by phonological information; whereas humor in comments is introduced with a response to a comment of someone else; i.e. humor relies on making clear a discrepancy between two particular points of view. In this respect, as noted in Section 4.5.3, HRM seems to correctly represent humor beyond text-specific examples.

6.2.3 Final Remarks

This task evaluates the performance of HRM in the field of web comments. We distinguish 600,000 web comments using all HRM's patterns, Results show that HRM has a similar performance in distinguishing funny comments from informative, insightful, and negative comments. Despite funny and non-funny training examples share more common aspects than differences, as well as the issue that funny comments are often an answer either to the commented item or to another comment (i.e. humor is self-contained), HRM's capabilities seem to be discriminating enough to achieve up to 80% of accuracy.

6.3 Sentiment Analysis

This task is focused on evaluating IDM on sentiment analysis issues. Let us consider the following scenario: enterprises can have direct access to negative information and, based on that information, to plan actions in order to revert the negative image. However, it is more difficult to mine relevant knowledge from positive information which implies a negative meaning. In this scenario, our model should be capable to mine this knowledge by detecting the fragments that potentially involve ironic content. In this respect, the expected result is to

provide information to experts, either at sentence level or at document level, who will decide whether or not such information is really ironic[1].

6.3.1 Sentiment Analysis Data Sets

Three different data sets that have been already employed in tasks related to sentiment analysis, as well as one which contains satiric examples, were used in this task[2]. They are:

1) The polarity dataset v2.0 described by Pang and Lee [114]. Hereafter *movies*2. This data set contains 1,000 positive and 1,000 negative processed reviews[3].

2) The polarity dataset v1.1 described by Pang et al. [113]. Hereafter *movies*1. This is the cleaned version which is integrated with 700 positive and 700 negative processed reviews[4].

3) The English book review corpus. Hereafter *books*. This corpus is described by Zagibalov et al. [184]. It contains 750 positive and 750 negative reviews[5].

4) The newswire and satire news articles described by Burfoot and Baldwin [24]. Hereafter *articles*. This data set is integrated with 4,000 real and 233 satire news articles[6].

[1] Like most tasks that involve information beyond grammar; i.e. subjective, social or cultural knowledge (e.g. machine translation), we believe that the irony detection task implies a human assessment to validate results as well as to learn from errors. Further improvements of the model would suppose a less human involvement.

[2] Davidov et al. [41] employed two data sets specifically designed for sarcastic sentences recognition. We contacted authors in order to obtain such data sets and to evaluate our model. Unfortunately, Dmitry Davidov has tragically passed away last year. Due to this sad event, authors are incapable of sharing the data sets.

[3] Available at http://www.cs.cornell.edu/People/pabo/movie-review-data.

[4] Ibid.

[5] Available at http://www.informatics.sussex.ac.uk/users/tz21.

[6] Available at http://www.csse.unimelb.edu.au/research/lt/resources/satire.

6.3.2 Automatic Evaluation

The first phase consisted of representing the documents by means of the complex IDM. The following phase was focused on obtaining the **documents** with higher probability to have ironic content. These documents were obtained by applying Formula 6.1:

$$\gamma(d_k) = \frac{\delta(d_k)}{tdf} \tag{6.1}$$

where *tdf* (*total dimension of features*) is the number of textual patterns of the model; i.e. $fdf = 8$. The underlying hypothesis is: the higher γ of document d_k, the higher is the probability of having ironic contents along the whole document[7]. According to this formula, the documents with highest γ value per set were: document *cv270_6079.txt* (set movies2); document *cv173_tok-11316.txt* (set movies1); document 233 (set books); and document *training-1581.satire* (set articles).

The final phase consisted of obtaining the **sentences** that could likely be ironic. In this case, we reduced our scope to the 50 documents per set with highest γ value; i.e. 200 documents in total considering the four data sets.

To this end, we first split the 200 documents in isolated sentences. Then, after modifying one parameter, Formula 5.2 was applied. The modification lay on eliminating the highest and lowest values of i in order to avoid biased δ values. Finally, in order to identify the sentences with higher probability to be ironic[8], Formula 6.1 was applied. The 100 sentences with highest γ value were then considered to be ironic; i.e. 400 sentences in total.

Figure 6.1 shows the results after applying Formula 6.1. X axis represents every single document within its respective set. Y axis represents its γ value. The dotted line represents the minimum γ value after which a document is considered as potentially ironic. This minimum γ value was determined by obtaining the mean between the highest and the lowest value of each set. In Appendix F are given several examples regarding the sentences with higher γ value.

According to this figure, there are some facts to be highlighted: the highest γ values are centered on the sets *movies*2 and *movies*1, then the set *articles*, and finally, the set *books*. This fact is related to word length in each data set. The

[7]Note that we do not consider the whole document to be completely ironic. Instead, we highlight the possibility to have fragments or sentences that can be considered to be ironic.

[8]No matter if two or more sentences belong to a same document.

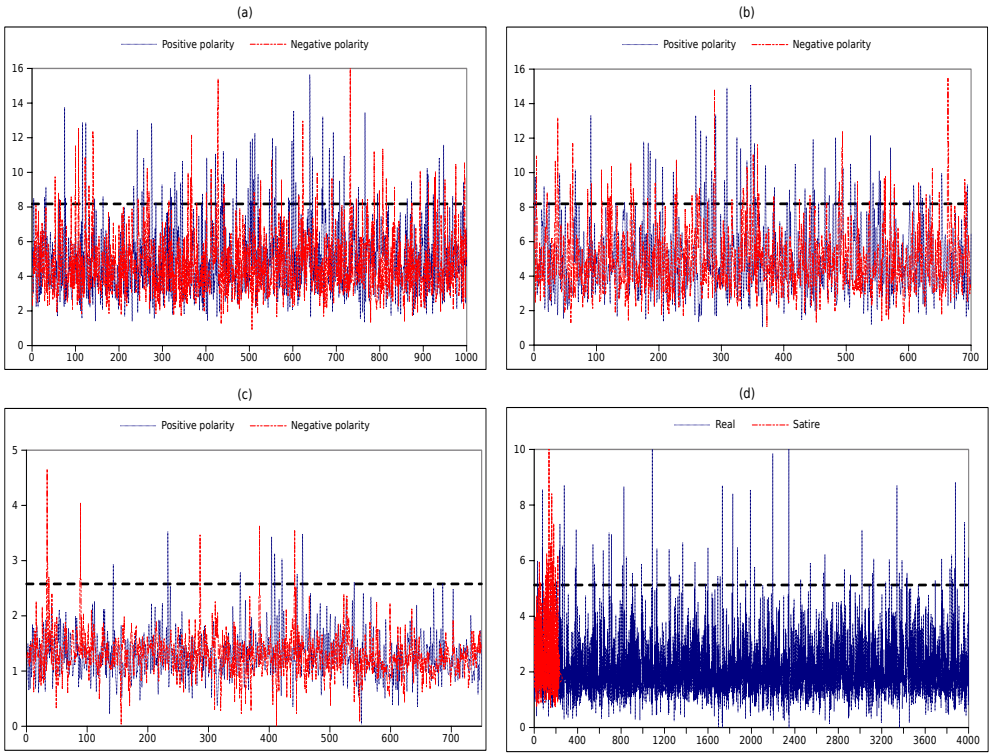

Figure 6.1: γ values per set: *movies*2 (a); *movies*1 (b); *books* (c); *articles* (d).

amount of words per document drastically varies across the sets: from an average length of 787 words in the set *movies*2 to an average length of 57 words in the set *books*. Nonetheless, this variation does not affect the quality of the documents candidates to have ironic content due to the documents are normalized according to their length (see Formula 5.2).

The most complex documents, discursively speaking, are the documents in sets movies. This is given by their length, which entails more elaborate narrative sequences. Now then, focusing only on these sets, it is observable how the documents labeled with positive polarity are the only ones in which the ironic content tends to often exceeds the minimum γ value. This behavior could provide some evidence about the presence of figurative content (either by using irony, sarcasm, satire, or even humor) when people try to consciously negate their literal words. According to this argument, the positive documents could be projecting a negative meaning in ground, completely different to the positive meaning profiled in surface.

Furthermore, most documents, regardless of the set they belong to, do not exceed the minimum γ value. About 90% of documents, or even more (see graph (c)), are far away from this minimum. This fact indicates that only very few documents might have ironic content. This is the expected situation because this kind of content does not appear constantly; i.e. there is not a balanced distribution between literal and figurative content. For instance, for every 10 literal web comments, one might expect that 1 or 2 had figurative content. This is clearer when analyzing graph (d). Despite there are only 233 satiric articles, most of them are close to the minimum γ values. Contrary situation with the real articles: they are 4,000 documents, but most are far away from the minimum.

According to the last argument, and considering the set *articles* (graph (d)) a kind of gold standard because its data is labeled as satiric or real[9], it is evident that IDM is representing underlying information linked to the way in which figurative content is expressed by people. Despite not all 233 documents exceed the minimum γ value, most of them steadily appear close to this minimum, unlike the real documents. This is clearer when considering the 50 most ironic documents belonging to this set: according to the model prediction, 34 documents belonged to documents labeled with the tag satire; whereas the remaining 16 documents belonged to documents labeled with the tag real. Thus, we might think that IDM is identifying elements that are commonly used to verbalize this type of content.

6.3.3 Manual Evaluation

Two manual evaluations were also performed in order to assess the results previously described. The first evaluation consisted of assessing the 400 sentences by two human annotators[10]. They were asked to evaluate whether or not those sentences might have any ironic meaning. Apart from their own concept of irony, no theoretical background was requested or offered. All the sentences were evaluated in isolation; i.e. their contexts were not provided. Each annotator evaluated 200 sentences (50 sentences per set). Furthermore, in order to estimate the degree of agreement between annotators, the Krippendorff α coefficient was calculated. According to Artstein and Poesio [3], this coefficient calculates the

[9] Unlike the others sets which are only labeled with positive or negative polarity.

[10] Both annotators are bilingual and they work as English-Spanish translators.

expected agreement by looking at the overall distribution of judgments without regard to which annotator produces the judgments. Table 6.2 presents their evaluation, for which Krippendorff α coefficient of 0.490 was noted. The percentages of *approved* sentences were obtained by dividing the amount of sentences marked as ironic by the annotators, by the total of sentences evaluated (50 per set).

Table 6.2: Manual evaluation in terms of isolated sentences.

	Annotator 1	Percentage	Annotator 2	Percentage
*Movies*2	13	26%	18	36%
*Movies*1	12	24%	18	36%
Books	8	16%	6	12%
Articles	17	34%	24	48%

The second evaluation consisted of assessing those sentences alongside the whole document they belong to. Thus, each annotator had to evaluate 25 documents per set. After reading the whole documents, they had to decide whether or not: i) the document was completely ironic; ii) the document contained any fragment (sentence or phrase) which may be considered to be ironic. In this case, apart from their own concept of irony, we provided our definition of irony stated in Section 2.5.2. Table 6.3 presents their evaluation, for which Krippendorff α coefficient of 0.717 was noted. The percentages of *approved* documents were now obtained by dividing the amount of documents marked as ironic by the total of documents evaluated (25 per set).

Table 6.3: Manual evaluation in terms of whole documents.

	Annotator 1			Annotator 2		
	i) document	ii) fragment	Percentage	i) document	ii) fragment	Percentage
*Movies*2	Not	16	64%	Not	14	56%
*Movies*1	Not	22	88%	Not	20	80%
Books	Yes	2	8%	Not	1	4%
Articles	Not	24	96%	Not	22	88%

According to the information depicted in both tables, we can infer the follow-

ing facts: on one hand, results given in Table 6.2 are quite poor. Each annotator evaluated 50 sentences per set, and the highest value achieved is 48%; i.e. less than half of them would be ironic. Results show that the problem of automatically classifying sentences as ironic is very challenging. For instance, it is completely senseless that only 6 of 50 sentences (the worst result) may come to be regarded as ironic when the purpose is just the contrary. Considering the sentences that are supposed to be more likely regarded as ironic (due to they come from the documents labeled as satiric of the *articles* set), evaluation evidences that IDM has some difficulties in identifying sentences which leave no doubt with respect to their ironic ground to any human. Moreover, based on annotators' comments, it is also evident that, except in very clear cases, an isolated sentence is often not sufficient to correctly decide whether or not such sentence is ironic. After manually analyzing some of these sentences, we could realize how hard is to figure out what is their ground meaning, especially, because of the lack of context. In absence of elements to map the information provided by the sentence, the fact of considering a sentence as ironic is almost a random process. For instance, sentence *"I never believed love at first site was possible until I saw this film"* could project both an ironic as a positive meaning. Similarly, sentence *"The plot, with its annoying twists, is completely inane"* could be profiling both a negative as an ironic meaning. If the context is not accessible to the annotators, they will hardly have elements to appreciate the existence of ironic content based only on an isolated sentence. Therefore, their evaluation will mostly depend on grammatical issues that leave no room to figurative interpretations.

On the other hand, second evaluation was performed based on these issues: if isolated sentences are not sufficient to determine the existence of ironic content, then we should try with entire documents. In this case, the results given in Table 6.3 show a clear improvement. Despite the results are not excellent (consider that only 1 document of 200 was regarded to be completely ironic[11], as well as the very low percentage of ironic content with respect to the documents belonging to the set *books*), it is evident how, when considering the whole document instead of isolated sentences, the spectrum to really appreciate irony clearly increased: 96% and 88% in the documents belonging to the set *articles*, as well

[11]However, it is unlikely to expect more ironic documents in these data sets because they were not compiled with the purpose of detecting irony.

as 88% and 80% in the documents of the set *movies*1. This fact shows the need of considering context and information beyond grammar for tasks such as this one. By examining the entire documents, annotators are able to access to very valuable information which makes sense as a whole, thereby achieving a complete overview of the meanings profiled. Result, accordingly, is that annotators now have elements to adequately judge whether or not an ironic content exists in the documents provided by IDM. Perhaps the participation of experts for evaluating results will increase: it is quite different to evaluate just a few sentences than entire documents, but it is also different to evaluate only some documents guided by the presence of such sentences than evaluating a complete data set.

Finally, by providing our definition of irony to the annotators, the scope of documents with ironic content substantially increased. This directly impacts on the scenario of applicability: a sentiment analysis task. According to the arguments given in Chapter 2, except in prototypical examples, the boundaries to correctly separate figurative phenomena are quite fuzzy. This is clearer when dealing with user-generated contents where people mix ironic remarks with observations about ironic, sarcastic or even funny situations; i.e. polarity depends on factors beyond the semantic of the words. If we intend to find out the underlying polarity of any document, we must spread the spectrum of phenomena related to the topic we are interested in. By considering phenomena related to irony (e.g. sarcasm and satire, which in many cases are considered part of it, or subclasses), we allowed annotators to have more elements to correctly make their decision. In addition, results depicted in Table 6.3 show some very interesting facts concerning the amount of documents with ironic content: 60.5% (121 of 200); 69 of them belonged to documents labeled with the positive polarity tag (documents labeled with the satiric tag are also considered here); whereas the remaining 52 belonged to the ones labeled with the negative polarity tag (documents labeled with the real tag are considered here). This means that ironic content does not always occur in the documents in which it is supposed to; i.e. irony should occur quite often in the documents labeled with the positive polarity tag due to its main characteristic is to produce an effect that denies the surface information. Now, when considering other kinds of effects (funny, disrespectful, sarcastic, etc.), the spectrum of sources to find ironic content increased. In this case, the definition provided to the annotators allowed to access to other sources in which figurative

content profiles **negative connotations**, regardless of it appeared in a document labeled as positive or negative.

6.3.4 Final Remarks

This task was focused on assessing IDM on a sentiment analysis scenario. Two kinds of results were obtained: isolated sentences and entire documents. These results were assessed by two human annotators on two key strata: i) determining whether or not the sentences could be regarded as ironic based only on the information provided by the sentence itself; ii) determining whether or not, by considering also the context of each sentence, the entire documents could be regarded as being completely ironic or having ironic content. Despite the two evaluations showed some model weakness, in particular with respect to the first stratum (it is quite hard to perceive irony based only on a sentence which belongs to a whole narrative). It is necessary to stress, however, that according to the evaluations obtained in the second stratum (when taking into consideration the context), the capabilities to correctly determine the presence of irony in the documents substantially increased.

6.4 Trend Discovery and Online Reputation

This task is focused on assessing IDM facing a trend discovery scenario. In this respect, consider that large companies have the most to gain from the appreciation of irony in social media, since these media are increasingly being used to comment on products and services and thereby encourage or discourage new customers. If a company can look beyond the distortional effect of irony, it can more accurately gather valuable marketing knowledge from the opinions of its users.

6.4.1 Toyota Data Set

We built a new data set concerning the case of a specific enterprise and its marketing problem: Toyota has of late encountered a variety of hardware problems to do with braking and acceleration, real or merely perceived, that have seriously

affected its reputation for quality and safety[12]. With respect to this topic, we have collected 500 tweets via the following attributes:

i. the **#toyota** tag;

ii. the positive emoticon **:)** and the negative emoticon **:(**.

All 500 tweets must contain the #toyota hashtag. To provide further focus, and to help us verify some assumptions regarding the contexts in which irony appears, such tweets should also contain either a positive or negative emoticon. Our test set thus contains 250 tweets with a positive emoticon and 250 labeled with a negative emoticon.

6.4.2 Human Evaluation

This experiment will allow us to test IDM's applicability to tweets that are not explicitly tagged as ironic by their senders. To this end, we compare the number of tweets identified as ironic by humans to the number predicted by IDM. We first obtain human judgments with respect to the presence or absence of ironic contents in the set #toyota. This step is performed by 80 annotators[13], who manually tagged the 500 tweets. They were asked to assign a value of 1 if they considered a tweet to be ironic, and a value of 0 if they considered it to be non-ironic. Like in the previous task, no theoretical background was requested or offered, and no dictionary definition of irony was provided. Instead, annotators were asked to rely on their own intuitions about what constitutes irony in a short text (we expect these intuitions to largely agree with the intuitions that lead a sender to mark a tweet with the hashtag #irony). Every annotator tagged 25 different tweets, and every tweet was tagged by 4 different annotators. In order to estimate the degree of agreement between the four annotators of each tweet, the Krippendorff α coefficient was calculated in each case. Table 6.4 presents overall statistics for the manual tagging of tweets, for which a Krippendorff α coefficient of 0.264 was noted. This value, according to the criteria exposed in Artstein and Poesio [3], indicates a fair reliability with respect to the generalization of

[12]This problem affected Toyota during the last months of 2009 and the beginning of 2010.

[13]Only 55 annotators were native speakers of English, while the remaining 25 were post-graduate students with sufficient English skills.

these annotations. Nonetheless, those authors also indicated that the purpose of reliability studies is not to find out whether annotations can be generalized, but whether they capture some kind of observable reality. According to this point of view, one of the main problems of the task is that irony remains a somewhat subjective concept, so that human annotators tend to disagree substantially. This, of course, is precisely the reason some tweeters feel the need to annotate their messages with an explicit indication of the presence of #irony.

Table 6.4: Statistics regarding annotators judgments.

	Tweets
Total tweets	500
Ironic tweets	147
Non ironic tweets	353
Ironic tweets$^{\mp}$	
4 annotators agree	28
3 annotators agree	39
2 annotators agree	80

(∓) **Considering only the 147 tweets annotated as ironic.**

We assume a tweet is ironic when at least two of its four human annotators classify it as such. Following this criterion, 147 of the 500 #toyota tweets are ironic. Of these 147 tweets, 84 belonged to tweets labeled with the positive emoticon #:); whereas 63 belonged to tweets labeled with the negative emoticon #:(. This difference supports the general assumption that irony more often relies on a positive ground to produce its critical effect[14]. Moreover, only in 28 tweets was there complete agreement among four annotators with respect to their assigned tags, while in only 39 tweets was agreement observed between three of the four annotators. In 80 tweets there was agreement between just two

[14]Recall that set #toyota set is artificially balanced, and contains 250 tweets with a positive emoticon and 250 tweets with a negative emoticon. Each emoticon serves a different purpose in an ironic tweet. Irony is mostly used to criticize, and we expect the negative emoticon will serve to highlight the criticism, while the positive emoticon will serve to highlight the humor of the tweet.

annotators[15]. Now, taking into consideration both Krippendorff α coefficient and the amount of ironic tweets, we will realize that the difficulty of recognizing irony, which somewhat perversely, is often greater than the difficulty of understanding irony. Quite simply, one does not always need to understand the concept of irony to understand the use of irony. Moreover, because irony requires a knowledge of cultural and social stereotypes and other pragmatic factors, the perception of irony tends to be subjective and personal.

Once the ironic tweets (relevant documents) are obtained, our model is applied to all 500 tweets in order to evaluate its performance to retrieve the documents with ironic content (147 tweets according to the human annotation). First, we determine three separate levels of representativeness (A, B, C) in order to cluster the texts into different groups for subsequent analysis. Each level is established by modifying the cutoff threshold in Formula 5.2 according to the following schema:

- Level A. Representativeness = 1 if $\delta_{i,j}(d_k) \geq \mathbf{0.8}$; otherwise = 0.
- Level B. Representativeness = 1 if $\delta_{i,j}(d_k) \geq \mathbf{0.6}$; otherwise = 0.
- Level C. Representativeness = 1 if $\delta_{i,j}(d_k) \geq \mathbf{0.5}$; otherwise = 0.

Then, for each level, we count how many retrieved documents matched with the relevant documents. Table 6.5 presents the results in terms of precision, recall and F-Measure. Taking the 147 tweets previously described as the total number of relevant documents to be retrieved, the results concerning precision are really low (they hardly exceed the 50% for each level); however, the results concerning recall are more satisfactory (from 40% to 84%). In this respect, such results seem to be very dependent on the level of representativeness. For instance, at the most discriminating level (A), the recall achieved is 40%, and the number of tweets retrieved is 59, of which 9 are tweets on which all four human annotators are in agreement, 16 are tweets on which three of the annotators agree, and 34 are tweets on which just two of the annotators agree. At the middle discriminating level (B), the number of tweets retrieved increased to 93 (recall = 63%), of which 14, 26, and

[15]It is important to mention that 141 tweets were tagged as ironic by just single annotators. However, these tweets were not considered in order to not bias the test. It is senseless to take a tweet as ironic when only one annotator tagged it as ironic, if 3 annotators said it was non-ironic.

53 agree with the judgments of four, three, and two annotators, respectively. At the lowest discriminating level (C), the number of relevant documents retrieved with ironic content increased to 123 (recall = 84%), of which 22, 32, and 69 agree with the judgments of respective annotators.

Table 6.5: Irony retrieval results.

Level	Tweets retrieved	Precision	Recall	F-Measure
A	59	56%	40%	0.47
B	93	**57%**	63%	0.60
C	123	54%	**84%**	**0.66**

In terms of precision, it is evident the need of improving the model. However, if considering the results concerning recall, the model shows some applicability to real-world problems. Though the performance of the model is not ideal when the representativeness level is close to 1, it seems clear that some of its features can capture recurrent linguistic patterns that characterize the use of irony in social media.

6.4.3 Final Remarks

This task was focused on applying IDM on short online texts. Though often repetitive and inane, these types of social texts are receiving increased attention as a carrier of influential customer opinions and feedback. In this respect, the comparison of human judgments with automatic classifications yields intriguing insights into how humans think about irony. Certainly, anyone who examines how the #irony hashtag is used in Twitter will know that humans do not have a single, precise notion of irony; rather, we seem to possess a diffuse, fuzzy, family-resemblance model of what it means for a text to be ironic. If we are capable of representing part of this fine-grained knowledge, then the implications of processing irony in real applications will be significant. For instance, the creation of indexes for obtaining the most ironic topics can be viewed as a trend discovery task, while characterization of information posted by bloggers can be seen as an application of online reputation. Each perspective in turn requires the

ability to extract fine-grained knowledge for decision making.

6.5 Further Tasks

Apart from the tasks above described, the models here proposed could represent further benefits concerning FLP in different tasks. In this respect, below we summarize two tasks in which both HRM and IDM were applied with interesting implications.

6.5.1 Towards a Humor Taxonomy

HRM could also represent further benefits for tasks in which information representation is quite relevant. For instance, in the process of building ontologies such as WordNet. In this respect, by analyzing several humorous examples we have realized that humor is often produced by two main referents that can be automatically identified: internal and external. On one hand, internal referents are intended to represent humor based on lexical patterns (for instance, phonological information). On the other hand, external referents are intended to represent humor based on extra-linguistic patterns (for instance, cultural information, beliefs, or social behaviors). Although HRM does not work based on explicitly differentiating both referents, it can incorporate a module to label each referent in order to provide fine-grained information.

Once implemented this module, we applied HRM over the corpus of one-liners used by Mihalcea and Strapparava in order to assess the viability of automatically building a humor taxonomy based on such information. We used this corpus due to the lack of a gold standard concerning humor processing.

Results show that one-liners can be represented in two classes: *low level* and *high level*. Low level class comprises texts in which humor is mainly produced by patterns such as humor domain, polarity, or affectiveness; i.e. prototypical information concerning humor topics and humor targets. For instance, as pointed out by Mihalcea and Strapparava [102], jokes about sexuality or self-referential. High level class, in contrast, comprises texts in which humor is predominantly caused by linguistic mechanisms such as ambiguity or incongruity. Now, from the two classes above mentioned, we built a general structure that roughly represents

humor's topics. In Figure 6.2 we illustrate how such structure can represent a preliminary humor taxonomy. As noted in this figure, we can identify nodes such as:

- ≻ stereotypes, humor about ethnic groups;
- ≻ pronominal, self-referential humor;
- ≻ white humor, positive polarity orientation;
- ≻ black humor, negative polarity orientation.

And deeper nodes such as:

- ≻ contextual, humor based on exaggeration, incongruity or absurd;
- ≻ intra-textual, humor based on linguistic ambiguity;
- ≻ extra-textual, humor based on pragmatic and cultural information.

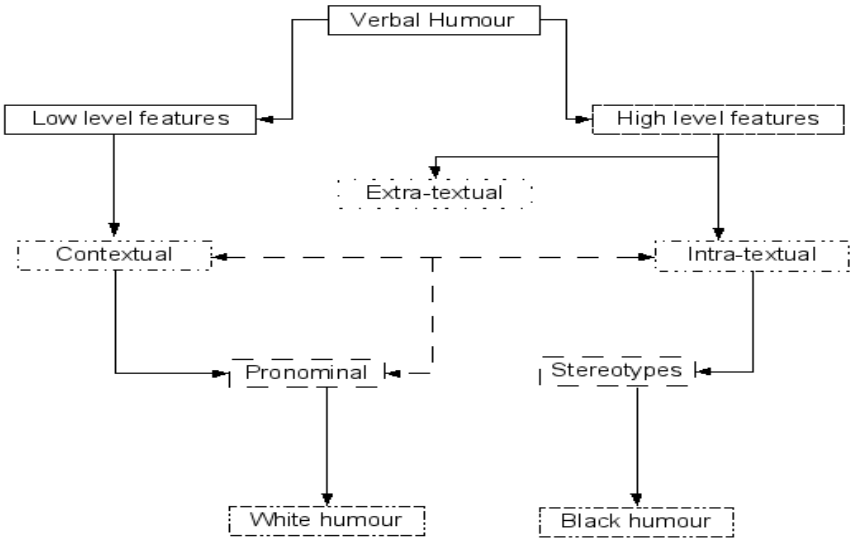

Figure 6.2: Preliminary humor taxonomy

6.5.2 Satire Detection

We have highlighted throughout the previous sections the difficulty to capture, by means of linguistic elements, the essence of irony. Phenomena such as linguistic and social factors impact on the perception of irony, making the task of automatically identifying ironic texts quite complex. Moreover, the close relationship between irony and sarcasm makes the task even more difficult. Here, we outline a preliminary approach to detect satire from specific IDM's patterns. In this respect, Burfoot and Baldwin [24] approached this task by means of lexical and semantics features (see Section 3.5.2). We, in contrast, represent their corpus of satiric articles by means of patterns such as polarity, affectiveness, incongruity, and emotional contexts.

The experiment consisted of representing the 233 satiric articles, as well as 700 randomly selected non satiric (or *real*, following the terminology employed by the authors)[16] with the patterns above mentioned. The aim was focused on assessing the relevance of such patterns to accurately retrieve satiric instances based only on such representation. All 933 instances were transformed in vectors by applying Formula 5.2. The vectorization was performed by assigning a value = 1 every time a feature appears in the document, regardless of the pattern it belongs to. These values were summed and divided by the number of features of the model in order to obtain the documents whose probability to be considered as satiric was higher. The final target was focused on retrieving as many satiric articles as possible.

Results are very promising. Considering 233 as the maximum of documents to retrieve, IDM predicted 193 satiric articles, failing in 40 articles; i.e. accuracy reaches 82.83%. In this respect, accuracy presents similar scores than the scores registered in Chapter 5. This means that some underlying patterns to express what people consider the core of figurative contents (either with respect to irony or satire), are adequately represented with our model.

Finally, although the latter two tasks are outlined in terms of preliminary approaches, their results show promising scenarios of applicability. For instance, concerning scenarios such as computer assisted translation (concerning with the

[16] In this case we are focused on keeping a relation 1 to 3 because figurative contents (either ironic, satiric, or sarcastic) do not appear in real contexts in a relation 1 to 1.

identification of common patterns that must not be translated literally); vandalism detection (concerning with the identification of racist or sexist texts (Blumentritt and Heredia [21])); analysis of political speech (concerning with the identification of common patterns influencing the semantics of political discourse (Vernier and Ferrari [176])); etc.

6.6 Summary

In this chapter we have described three complete evaluations concerning the applicability of both HRM as IDM in tasks related to information retrieval, sentiment analysis, trend discovery, and online reputation.

First, in Section 6.2 we assessed HRM in terms of an information retrieval task. The task was carried out in a data set of 600,000 web comments collected from Slashdot.com. In Section 6.3 and Section 6.4, IDM was evaluated. Fist, by means of a sentiment analysis task, and then, by means of a trend discovery task. In each task IDM was assessed with new data sets in order to verify its capabilities regarding non-labeled examples.

Finally, in Section 6.5, we outlined two preliminary approaches in which both models could be applied for FLP.

7

Conclusions

A conclusion is simply the place where you got tired of thinking.

Mihalcea and Strapparava [102]

In this thesis we have approached two tasks in which the automatic processing of figurative language has been involved: humor recognition and irony detection. Each task was undertaken independently by means of a linguistic pattern representation. In this respect, two models of figurative language were here proposed:

i. HRM (Humor Recognition Model);

ii. IDM (Irony Detection Model).

Both models go beyond surface elements to extract different types of patterns from a text: from lexicon to pragmatics. Since our target was focused on representing figurative language concerning social media texts, each model was evaluated by considering non-prototypical texts that are laden with social meaning. Such texts were automatically collected by chiefly taking advantage of user-generated tags. The data sets are freely available for research purposes. Two goals were highlighted while evaluating the models: representativeness and relevance. The former was intended to consider the appropriateness or representativeness of different patterns to humor recognition and irony detection, respectively; whereas the latter was focused on considering the empirical performance of each model on a text classification task. When evaluating representativeness, we looked to

whether individual features were linguistically correlated to the ways in which users employ words and visual elements (i.e. emoticons and punctuation marks) when speaking in a mode they consider to be figurative. The classification task, in contrast, evaluated the capabilities of the models as a whole, focusing on the ability of the entire system of patterns to accurately discriminate figurative from non-figurative texts.

According to the results described in Chapters 4- 6, our initial assumptions concerning the usefulness of this type of information in characterizing figurative language were confirmed. In addition, though HRM and IDM clearly leave room for improvement, they achieved encouraging results in terms of representativeness, classification accuracy, precision, recall and F-Measure.

Finally, it is worth noting that the patterns here proposed work better when they were used as part of a coherent framework rather than used individually; i.e. no single pattern was distinctly humorous or ironic, but all of them together provided a valuable linguistic inventory for detecting these types of figurative devices at textual level.

7.1 Contributions

Language reflects patterns of thought. Accordingly, to study language is to study patterns of conceptualisation (Kemmer [79]). In this respect, by analyzing two specific domains of figurative language, we aimed at providing arguments concerning how people conceive humor and irony in terms of their use in social media platforms. Such arguments were intended to represent formal features of each figurative device that could be implemented in computational models to foster the automatic processing of both humor and irony. In this section, we summarize our major findings whereas the details of each device, as well as their applicability, can be found in Chapters 4, 5, and 6.

I. As described in Chapter 2, figurative language is assumed to intentionally communicate indirect meanings. This type of language entails important challenges, not only for a computational processing, but for a linguistic representation as well. The linguistic and social factors that impact on the perception of figurativity make the task of automatically identifying

figurative texts quite complex, especially, due to textual instances lack of valuable information such as intonation and gestural information (present in oral communication), what substantially increases task complexity. In this respect, by representing humor and irony in terms of their conceptual use rather than only of their theoretical description, the models here discussed seem to efficiently capture the core of the most salient attributes of each figurative device.

II. According to the previous point, our figurative language representation is given by analyzing the linguistic system as an integral structure which depends on grammatical rules as well as on cognitive, experiential, and social contexts, which altogether, represent the meaning of what is communicated.

III. The current trends in NLP are increasingly focusing on the analysis of knowledge beyond formal language. By implementing fine-grained patterns, NLP systems are even more capable of mining valuable knowledge related to non-factual information that is linguistically expressed. Such knowledge is valuable for tasks in which the target is beyond the literal interpretation. For instance, in opinion mining (Ghose et al. [52]), sentiment analysis (Pang et al. [113], Wilson et al. [182]), or information extraction (Wiebe and Riloff [179]), where the implicit knowledge in texts is extremely useful for achieving good results. Despite such implicit knowledge is often expressed by means of figurative devices such as irony, sarcasm, humor, metaphor, and so on (see Chapter 3), the figurative meaning rarely appear registered in dictionaries. Thus, it must be inferred from context. In this respect, with this approach we provided a methodology to automatically identify figurative uses of language in order to foster FLP beyond the tasks here involved (see Chapter 6).

IV. Social media are replacing mass media. The result is that language is slightly changing to adapt to new ways of communication. Therefore, it is useless to propose a model based on prototypical instances of figurative language, or based on text-specific instances. Hence, the fact of concentrating on social media texts, whose intrinsic characteristics are quite different to the characteristics described in the specialized literature, increases the

scope of this investigation. For instance, humor in web comments is often an answer either to a commented item or to another comment; i.e. humor does not rely on prototypical jokes, but on exploiting mechanisms that likely are not considered by experts (e.g. semantic dispersion). In this respect, we are not dealing only with prototypical or literary examples of humor and irony. Rather, we are dealing with more general characteristics used by people to effectively communicate figurative intents (although such characteristics do not 100% correspond to the prototypical ones suggested by experts).

V. The lack of specific resources for figurative language is a fact to be highlighted. There are very few available corpora to assess any model. Manual annotation is a time-consuming manual task. That is why corpus-oriented research is restricted (Peters and Wilks [119]). When available, such corpora are mostly text-specific. Therefore, the possibility to evaluate new models is limited. In this respect, with our approach (that is focused on taking advantage of user-generated tags), we have reduced the constraints facing corpus-based research (see Chapters 4- 6). For instance, the subjectivity of determining figurativity at textual level is reduced by collecting examples that are intentionally labeled with a descriptor (user-generated tag) whose goal is to focus people's posts on particular topics.

VI. By making freely available our data sets we are collaborating to the spread of researches related to figurative language, as well as palliating the lack of resources for FLP. In addition, we are showing that our corpus-based approach is useful for tasks in which the scarcity of data, the task subjectivity, the manual annotation, or the impossibility of making personal interviews, are impediments to be tackled.

VII. By deeply investigating ambiguity from different linguistic layers, as well as by considering surface patterns, we have shown that humor recognition accuracy substantially increases. In particular, due to humor here considered is not text-specific; i.e. HRM is useful for prototypical instances of humor, such as one-liners, as well as for more complex instances in which humor is self-contained (e.g. web comments, blogs). In this respect, the findings here reported are interesting due to they represent an original approach in

which underlying information, that people profile in their non-specialized texts when explicitly and implicitly expressing humor, is taken into account.

VIII. Irony is a sophisticated, subtle and ambiguous way of communication that has received little serious computational attention in the past. Though this is changing, perhaps because of the prevalence of irony in online texts and social media. However, its automatic processing is even more complex than humor processing. Despite most people (experts and non-experts) concur that irony conveys an opposite meaning; i.e. people say something that seems to be the opposite of what they mean (Colston and Gibbs [36]), such property is rarely observed in social media texts. For instance, people often mix ironic remarks with observations about ironic situations (see Chapters 2 - 6). We thus face the challenge that people possess a diffuse, fuzzy, family-resemblance concept of what it means for a text to be ironic. With IDM we have proposed, beyond a theoretical framework, a model that attempts to describe salient characteristics of irony. Thus, we align ourselves more with the intuitive view of irony (that an expectation has been violated in a way that is both appropriate and inappropriate) than with the strictly scholarly (and perhaps even scholastic) view. The result: an integral model that incorporates low and high level properties of irony based on formal linguistic elements.

IX. Completing the previous point, it is worth noting that another important issue concerning irony is related to negation. This grammatical category allows changing the truth value of a proposition. That is why its automatic processing is very important for several NLP tasks such as sentiment analysis, opinion mining, question answering or textual entailment[1]. However, if automatic negation processing is already quite complex when dealing with literal language, it becomes even more difficult and challenging when dealing with figurative language (e.g. consider the use of narrative strategies, such as tone, obviousness, or funniness, as well as the absence of a negation marker to realize the complexity that this task entails). Despite such inconveniences, the model here developed seems to be robust enough

[1]Consider, for instance, the international contests about negation and its automatic processing: http://www.cnts.ua.ac.be/BiographTA/qa4mre.html.

to identify complex features that throw focus onto the figurative uses of textual elements to accurately communicate ironic intent.

X. Figurative language is a widespread phenomenon in web content. As noted throughout the previous chapters, its automatic processing has important implications for many NLP tasks. For instance, in sentiment analysis (cf. Reyes et al. [136] about the importance of determining the presence of irony in order to assign fine-grained polarity levels), opinion mining (cf. Sarmento et al. [146], where the authors note the role of irony in discriminating negative from positive opinions), or advertising (cf. Kreuz [81], about the function of irony to increase message effectiveness in advertising). In this respect, one could ask whether HRM and IDM yield actual benefits in real-world applications. In Chapter 6 we showed how both models provide fine-grained knowledge concerning their applicability in tasks such as information retrieval, sentiment analysis, trend discovery and online reputation. The empirical insights here described demonstrated that our models should improve and facilitate hand-based retrieval, as well as accurate classification, of figurative content.

7.2 Future Work

Since figurative language is common in texts that express subjective and deeply-felt opinions, its presence represents a significant obstacle to the accurate analysis of sentiment in these texts. A successful model either of humor recognition or of irony detection can thus play both a direct and an indirect role for tasks as diverse as the ones described in Chapters 3-6.

In this respect, the main directions of future work are addressed to the mining of fine-grained knowledge that could be applied in tasks in which natural language, either literal or figurative, is involved. Some of them are listed below.

A. The main direction consists of improving the quality of textual patterns, as well as investigating new ones, in order to obtain a set of fine-grained patterns that may be used not only for having a more robust HRM or IDM, but for describing (figurative) language in such a way that our findings can impact other NLP tasks.

B. Another direction is related to the experiments reported in Chapter 6. In two of them we considered the practical applicability of IDM taking into account the comparison of human judgments with automatic classifications. In the near future we plan to manually annotate large-scale examples in order to compare the results here described.

C. On the other hand, anyone who examines the positive examples of irony will know that humans do not have a single, precise notion of irony; rather, we seem to possess a diffuse, fuzzy, family-resemblance model of what it means for a text to be ironic. This suggests that as another direction of future work, we should not just be focused on the quality and value of different linguistic patterns, though this of course will be an important topic. We shall also have to tackle the problem of how people think about irony, and recognize irony in their own texts and in those of others. This will require that we tease apart the categories of verbal irony and situational irony. Logically these are distinct categories; in real texts however, where people mix ironic remarks with observations about ironic situations, the two are very much intertwined.

D. In the same vein, it is evident that our data sets contain several types of irony. In this respect, once a broad sense of irony has been detected in a text, one can then apply other formal machinery to determine precisely which type of irony is at work. We relegate this fine-grained classification of an irony-laden text into distinct categories of irony to the realm of future work. Thus, if we are capable of classifying ironic instances according to their correct category, then the quality of results will considerably improve, and accordingly, their applicability. For instance, thinking of the appropriate use of figurative expressions in bilingual people (Schmitz [149], Deneire [45], Schmitz [150]).

E. The binary classification of figurative language here described might have multiple applications to be further addressed. For instance, according to Feldman and Peng [48], it is useful for indexing purposes and for increasing the precision of information retrieval systems, as well as for providing knowledge of which clauses should be interpreted literally and which figuratively regarding text summarization and machine translation systems.

F. Irony, satire, parody and sarcasm are overlapping figurative phenomena whose differences are a matter of usage, tone, and obviousness. For instance, sarcasm has an obviously mocking tone that is used against another, while irony is often more sophisticated, more subtle and ambiguous, and even self-deprecating. Although our aim was not focused on distinguishing among these figurative devices, but on recognizing statements that have non-literal meanings, we plan to address the fine-grained task of classifying instances of irony, sarcasm, and satire by applying more complex patterns.

G. It is necessary to come up with improved models capable to detect better figurative patterns in different types of texts. To this end, it will be indispensable the compilation of specific data sets for FLP. Task that is a challenge itself because of the subjectivity of determining figurativity at textual level.

H. Last but not least, it will be important to approach FLP from each of its angles considering also valuable information such as gestural information, tone, paralinguistic cues, etc (Cornejol et al. [37]), as well as trying to model figurative language taking into consideration the visual stimulus of brains responses when people have to process this particular type of language (such as in Mars et al. [95]).

7.3 Publications

In this final section we are listing all the publications related to this investigation.

[1] Reyes A., P. Rosso, D. Buscaldi 2012. From Humor Recognition to Irony Detection: The Figurative Language of Social Media. In Data & Knowledge Engineering 12 (2012): 1–12. DOI: 10.1016/j.datak.2012.02.005. Impact Factor 1.717. http://dx.doi.org/10.1016/j.datak.2012.02.005.

[2] Reyes A., P. Rosso 2012. Making Objective Decisions from Subjective Data: Detecting Irony in Customers Reviews. In: Journal on Decision Support Systems. In press. DOI: 10.1016/j.dss.2012.05.027 Impact Factor 2.135. http://dx.doi.org/10.1016/j.dss.2012.05.027.

[3] Reyes A., P. Rosso, T. Veale 2012. A Multidimensional Approach For Detecting Irony in Twitter. In Language Resources and Evaluation. Forthcoming. Impact Factor 0.615.

[4] Reyes A., P. Rosso 2012. Building Corpora for Figurative Language Processing: The Case of Irony Detection In: Proceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (in conjunction with LREC 2012). May 2012, Istanbul, Turkey. pp. 94-98.

[5] Reyes A., P. Rosso. 2011. Mining Subjective Knowledge from Customer Reviews: A Specific Case of Irony Detection. In Proceedings of the ACL 2011 Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), June 2011, Portland, OR. pp. 118-124.

[6] Reyes A., M. Potthast, P. Rosso, B. Stein. 2010. Evaluating humour patterns on web comments. In Proceedings of the 7th International Conference on Language Resources and Evaluation, 2010, pp. 1138-1141.

[7] Reyes A., D. Buscaldi, P. Rosso. 2010. The impact of semantic and morphosyntactic ambiguity on automatic humour recognition. In Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems (NLDB) 2009, Vol. 5723 of LNCS, 2010, pp. 130141.

[8] Reyes A., Rosso P., Buscaldi D. 2010. Finding Humour in the Blogosphere: The Role of WordNet Resources. In Proceedings of the 5th Global WordNet International Conference (GWN-2010), Bombay, India, pp. 56-61.

[9] Reyes A., P. Rosso, D. Buscaldi, 2009. Humor in the blogosphere: First clues for a verbal humor taxonomy. Journal of Intelligent Systems 18 (4) (2009) 311 331. Registered in the ERA Journal List and the CORE Extract of Journals (B).

[10] Reyes A., P. Rosso, D. Buscaldi. 2009. Affect-based patterns for Humour Recognition. In Proceedings of the 7th International Conference on Natural Language Processing, ICON 2009. Hyderabad, India. pp. 364-369.

[11] Reyes A., P. Rosso, D. Buscaldi. 2009. Linking Humour to Blogs Analysis: Affective Traits in Posts. In Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis, WOMSA 2009. Sevilla, Spain. pp. 100-109.

[12] Reyes A., P. Rosso, D. Buscaldi. 2009. Evaluating Humorous patterns: Towards a Humour Taxonomy. In Proceedings of the 4th Indian International Conference on Artificial Intelligence, IICAI-09. Tumkur, India, pp. 1373-1390.

[13] Reyes A., D. Buscaldi, P. Rosso. 2009. An Analysis of the Impact of Ambiguity on Automatic Humour Recognition. In Proceedings of the 12th International Conference Text, Speech and Dialogue, TSD 2009. LNAI (5729), Pilsen, Czech Republic, pp. 162-169.

[14] Reyes A., P. Rosso., A. Martí, M. Taulé. 2009. Características y rasgos afectivos del humor: Un estudio de reconocimiento automático del humor en textos escolares en catalán. In Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), (43):235-243.

Bibliography

[1] J. Allen. Smiles and laughter in don quixote. *Comparative Literature Studies*, 43(4):515–531, 2006. 29

[2] M. Ariel. The demise of a unique concept of literal meaning. *Journal of Pragmatics*, 34(4):361402, 2002. 13

[3] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, December 2008. 124, 129

[4] J. Atserias, B. Casas, E. Comelles, M. González, L. Padró, and M Padró. Freeling 1.3: Syntactic and semantic services in an open-source nlp library. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 48–55, 2006. 64, 93

[5] S. Attardo. *Linguistic Theories of Humor*. Mouton de Gruyter, 1994. 3, 27, 28, 44

[6] S. Attardo. *Humorous Texts: A semantic and pragmatic analysis*. Mouton de Gruyter, 2001. 27

[7] S. Attardo. Irony as relevant inappropriateness. In R. Gibbs and H. Colston, editors, *Irony in Language and Thought*, pages 135–174. Taylor and Francis Group, 2007. 1, 4, 31, 33

[8] S. Attardo. A primer for the linguistics of humor. In Victor Raskin, editor, *The Primer Of Humor Research*, pages 101–156. Mounton de Gruyter, 2008. 28, 44, 72

[9] S. Attardo and V. Raskin. Script theory revis(it)ed: Joke similarity and joke representation model. *Humor*, 4(3-4):293–347, 1991. 28, 29

[10] A. Augello, G. Saccone, S. Gaglio, and G. Pilato. Humorist bot: Bringing computational humour in a chatbot system. *Complex, Intelligent and Software Intensive Systems, International Conference*, 0:703–708, 2008. 45

[11] S. Baccianella, A. Esuli, and F. Sebastiani. Multi-facet rating of product reviews. In *Proceedings of the 31st European Conference on Information Retrieval*, volume 5478 of *Lecture Notes in Computer Science*, pages 461–472. Springer, 2009. 89

[12] K. Balog, G. Mishne, and M. Rijke. Why are they excited? Identifying and explaining spikes in blog mood levels. In *European Chapter of the Association of Computational Linguistics (EACL 2006)*, 2006. 5, 55

[13] R. Basili and F. Zanzotto. Parsing engineering and empirical robustness. *Journal of Natural Language Engineering*, 8(3):97–120, 2002. 59, 66, 67

[14] L. Bentivogli, P. Forner, B. Magnini, and E. Pianta. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In Gilles Sérasset, editor, *Multilingual Linguistic Resources (COLING 2004)*, pages 94–101, 2004. 71

[15] B. Bergen. Mental Simulation in Literal and Figurative Language Understanding. In Seana Coulson, editor, *The Literal And Nonliteral in Language and Thought*, pages 255–280. Peter Lang Publishing, September 2005. 16, 24, 25, 158

[16] K. Binsted. Using humour to make natural language interfaces more friendly. In *Proceedings of the AI, ALife and Entertainment Workshop*, 1995. 35, 45, 57

[17] K. Binsted. *Machine humour: An implemented model of puns*. PhD thesis, University of Edinburgh, Edinburgh, Scotland, 1996. 43, 56, 57, 58, 59

[18] K. Binsted and G. Ritchie. Computational rules for punning riddles. *Humour*, 10:25–75, 1997. 2, 43

[19] K. Binsted and G. Ritchie. Towards a model of story puns. *Humour*, 14:275–292, 2001. 43

[20] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly, Beijing, 2009. 35

[21] T. Blumentritt and R. Heredia. Stereotype processing and nonliteral language. In H. Colston and A. Katz, editors, *Figurative Language Comprehension: Social and Cultural Influences*, pages 261–281. Lawrence Erlbaum Associates, 2005. 136

[22] D. Bogdanova. A framework for figurative language detection based on sense differentiation. In *Proceedings of the ACL 2010 Student Research Workshop*, ACL '10, pages 67–72, Morristown, NJ, USA, 2010. Association for Computational Linguistics. 38

[23] T. Brants and A. Franz. Web 1t 5-gram corpus version 1. Technical report, Google Research, 2006. 63

[24] C. Burfoot and T. Baldwin. Automatic satire detection: Are you having a laugh? In *ACL-IJCNLP '09: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 161–164, 2009. 2, 33, 49, 114, 121, 135

[25] D. Buscaldi and P. Rosso. Some experiments in humour recognition using the italian wikiquote collection. In *3rd Workshop on Cross Language Information Processing, CLIP-2007, Int. Conf. WILF-2007*, volume 4578, pages 464–468, 2007. 46, 61, 77

[26] C. Cacciari. Why do we speak metaphorically? Reflections on the functions of metaphor in discourse and reasoning. In M. Marschark, editor, *Figurative Language And Thought*, pages 119–157. Oxfford University Press, 1998. 8, 22

BIBLIOGRAPHY

[27] P. Carvalho, L. Sarmento, M. Silva, and E. de Oliveira. Clues for detecting irony in user-generated contents: oh...!! It's "so easy" ;-). In *TSA '09: Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56, Hong Kong, China, November 2009. ACM. 2, 48, 112

[28] L. Chin-Yew and F. Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 605–612, Morristown, NJ, USA, 2004. Association for Computational Linguistics. 103

[29] N. Chomsky. *Syntactic Structures*. Mouton and Co, The Hague, 1957. 16

[30] N. Chomsky. *Aspects of the theory of syntax*. MIT Press, Cambridge, MA, 1965. 14

[31] Alexander Clark, Chris Fox, and Shalom Lappin, editors. *The Handbook of Computational Linguistics and Natural Language Processing*. Blackwell Handbooks in Linguistics. John Wiley & Sons, 2010. 36, 37

[32] H. Clark and R. Gerrig. On the pretense theory of irony. *Journal of experimental psychology: General*, 113 (1):121–126, 1984. 30

[33] W. Cohen, P. Ravikumar, and S. Fienberg. A Comparison of String Distance Metrics for Name-Matching Tasks. In *Proceedings of IJCAI-03 Workshop on Information Integration*, pages 73–78, August 2003. 107

[34] H. Colston. Social and cultural influences on figurative and indirect language. In H. Colston and A. Katz, editors, *Figurative Language Comprehension: Social and Cultural Influences*, pages 99–130. Lawrence Erlbaum Associates, 2005. 29, 84, 93

[35] H. Colston. On necessary conditions for verbal irony comprehension. In R. Gibbs and H. Colston, editors, *Irony in Language and Thought*, pages 97–134. Taylor and Francis Group, 2007. 3, 31, 32

[36] H. Colston and R. Gibbs. A brief history of irony. In R. Gibbs and H. Colston, editors, *Irony in Language and Thought*, pages 3–24. Taylor and Francis Group, 2007. 4, 31, 141

[37] C. Cornejol, F. Simonetti, N. Aldunate, A. Ibáñez, V. López, and L. Melloni. Electrophysiological evidence of different interpretative strategies in irony comprehension. *Journal of Psycholinguist Research*, 36:411–430, 2007. 144

[38] I. Councill, R. McDonald, and L. Velikovich. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 51–59, Uppsala, Sweden, July 2010. University of Antwerp. 85

[39] C. Curcó. Irony: Negation, echo, and metarepresentation. In R. Gibbs and H. Colston, editors, *Irony in Language and Thought*, pages 269–296. Taylor and Francis Group, 2007. 31

[40] M. Dascal. Defending literal meaning. *Cognitive Science*, 11(3):259–281, 1987. 13, 16

[41] D. Davidov, O. Tsur, and A. Rappoport. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL '10, pages 107–116, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. 32, 50, 114, 121

[42] F. de Quevedo. *La vida del Buscón llamado Don Pablos. Sueños y discursos*. Catedra, 1980. 1

[43] F. de Saussure. *Course in general linguistics*. Fontana, London, 1974. 12, 13

[44] L. Deg and Y. Bestgen. Towards automatic retrieval of idioms in French newspaper corpora. *Literary and Linguistic Computing*, 18:249–259, 2003. 25, 42

[45] M. Deneire. Humor and foreign language teaching. *Humor*, 8(3):285–298, 1995. 143

[46] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1): 61–74, 1993. 91

[47] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006)*, pages 417–422, 2006. 54, 72

[48] A. Feldman and J. Peng. An approach to automatic figurative language detection: A pilot study. In *Proceedings of the Corpus-Based Approaches for Figurative Language Colloquium*, Liverpool, UK, 2009. 22, 25, 42, 55, 143, 157

[49] C. Fillmore. Frame semantics. In *Linguistics in the Morning Calm. Selected Papers from SICOL-1981*, pages 113–137. Seoul, Hanshing Publishing Company, 1982. 12, 13, 18, 57

[50] C. Fillmore, P. Kay, and M. O'Connor. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538, 1988. 12

[51] L. Friedland and J. Allan. Joke retrieval: recognizing the same joke told differently. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008*, pages 883–892, 2008. 46

[52] A. Ghose, P. Ipeirotis, and A. Sundararajan. Opinion mining using econometrics: A case study on reputation systems. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 416–423. Association for Computational Linguistics, 2007. 139

[53] R. Gibbs. Evaluating contemporary models of figurative language understanding. *Metaphor and Symbol*, 16 (3):317–333, 2001. 17

[54] R. Gibbs. Irony in talk among friends. In R. Gibbs and H. Colston, editors, *Irony in Language and Thought*, pages 339–360. Taylor and Francis Group, 2007. 33

BIBLIOGRAPHY

[55] R. Gibbs and H. Colston. The future of irony studies. In R. Gibbs and H. Colston, editors, *Irony in Language and Thought*, pages 339–360. Taylor and Francis Group, 2007. 33

[56] R. Gibbs and C. Izett. Irony as persuasive communication. In H. Colston and A. Katz, editors, *Figurative Language Comprehension: Social and Cultural Influences*, pages 131–152. Lawrence Erlbaum Associates, 2005. 29, 39, 47, 84, 86

[57] R. Giora. On irony and negation. *Discourse Processes*, 19(2):239–264, 1995. 31

[58] R. Giora, N. Balaban, O. Fein, and I. Alkabets. Negation as positivity in disguise. In H. Colston and A. Katz, editors, *Figurative language comprehension: Social and cultural influences*, pages 233–258. Hillsdale, NJ: Erlbaum, 2005. 84

[59] S. Glucksberg and M. McGlone. *Understanding Figurative Language: From Metaphors to Idioms*. Oxford Psychology Series. Oxford University Press, New York, 2001. 14, 24, 25

[60] A. Goldberg. Making one's way through the data. In M. Shibatani and S. Thompson, editors, *Grammatical Constructions: Their Form and Meaning*, pages 29–53. Oxford, Clarendon Press, 1996. 12, 13

[61] A. Goldberg. Construction grammar. In E.K. Brown and J.E. Miller, editors, *Concise Encyclopedia of Syntactic Theories*. Elsevier Science Limited, 1997. 12, 22

[62] R. González-Ibáñez, S. Muresan, and N. Wacholder. Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Vol. 2*, pages 581–586. The Association for Computer Linguistics, 2011. 50

[63] H. Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and semantics*, volume 3, pages 41–58. New York: Academic Press, 1975. 14, 17, 31

[64] D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks. A closer look at skip-gram modelling. In *Proceedings of the Fifth international Conference on Language Resources and Evaluation (LREC-2006)*, pages 1222–1225, 2006. 103

[65] S. Halliwell. *Greek Laughter. A Study of Cultural Psychology from Homer to Early Christianity*. Cambridge University Press, New York, 2008. 3, 27

[66] Y. Hao and T. Veale. An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Minds Machines*, 20(4):635–650, November 2010. 48

[67] R. Hausser. *Foundations of Computational Linguistics*. Springer-Verlag, Berlin Heidelberg, 1998. 36

[68] C. Hempelmann. An ynperfect pun selector for computational humor. In *Proceedings of the First Annual Midwest Colloquium in Computational Linguistics*, 2004. 45

[69] J. Hertzler. *Laughter: A social scientific analysis*. Exposition Press, 1970. 3, 27

[70] R. Jackendoff. *Semantics and cognition*. MIT Press, Cambridge, Mass., 1983. 7

[71] Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007. 88

[72] D. Jurafsky and J. Martin. *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall, 2007. 36, 58, 62, 63

[73] G. Karypis. Cluto. A clustering toolkit. technical report 02-017. Technical report, University of Minnesota, Department of Computer Science., 2003. 91, 92

[74] A. Katz. Figurative language and figurative thought: A review. In M. Marschark, editor, *Figurative Language And Thought*, pages 3–43. Oxfford University Press, 1998. 25, 158

[75] A. Katz. Discourse and sociocultural factors in understanding nonliteral language. In H. Colston and A. Katz, editors, *Figurative Language Comprehension: Social and Cultural Influences*, pages 183–208. Lawrence Erlbaum Associates, 2005. 49

[76] A. Katz, Mark Turner, R. Gibbs Jr., and C. Cacciari. Counterpoint commentary. In M. Marschark, editor, *Figurative Language And Thought*, pages 158–192. Oxfford University Press, 1998. 17, 22, 26, 29, 47, 158

[77] J. Katz. *Propositional structure and illocutionary force: A study of the contribution of sentence meaning to speech acts*. Harvard University Press, 1980. 13, 16

[78] B. Kaup, J. Lüdtke, and R. Zwaan. Processing negated sentences with contradictory predicates: Is a door that is not open mentally closed? *Journal of Pragmatics*, 38: 1033–1050, 2006. 84

[79] S. Kemmer. About cognitive linguistics: Historical background. http://www.cognitivelinguistics.org/cl.shtml, 2010. Online on August 25, 2011. 12, 138

[80] D. Kim, D. Ferrin, and H. Raghav. A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision Support Systems*, 44(2):544–564, 2008. 88

[81] R. Kreuz. Using figurative language to increase advertising effectiveness. In *Office of Naval Research Military Personnel Research Science Workshop*, Memphis, TN, 2001. University of Memphis. 86, 142

[82] A. Kulkarni and T. Pedersen. Senseclusters: Unsupervised clustering and labeling of similar contexts. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 105–108. Association for Computational Linguistics, 2005. 92

BIBLIOGRAPHY

[83] S. Kumon-Nakamura, S. Glucksberg, and M. Brown. How about another piece of pie: The allusional pretense theory of discourse irony. In R. Gibbs and H. Colston, editors, *Irony in Language and Thought*, pages 57–96. Taylor and Francis Group, 2007. 1, 31, 33

[84] G. Lakoff. *Women, Fire and Dangerous Things: What Categories Reveal About the Mind*. University of Chicago Press, Chicago, 1987. 13, 23

[85] G. Lakoff and M. Johnson. *Metaphors we live by*. University of Chicago Press, Chicago, 1980. 13, 22, 23, 37, 158

[86] R. Langacker. *Foundations of Cognitive Grammar*. Stanford University Press, 1987. 12, 13, 37

[87] R. Langacker. *Concept, Image and Symbol. The Cognitive Basis of Grammar*. Mounton de Gruyter, 1991. 12, 22, 37

[88] L. Li and C. Sporleder. Using gaussian mixture models to detect figurative language in context. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 297–300, Los Angeles, California, June 2010. Association for Computational Linguistics. 17, 38, 39, 42

[89] B. Lönneker-Rodman and S. Narayanan. Computational approaches to figurative language, 2008. 17, 22, 23, 24

[90] X. Lü, L. Zhang, and J. Hu. Statistical substring reduction in linear time. In *Proceedings of IJCNLP-04, HaiNan island*, 2004. 93

[91] J. Lucariello. Situational irony: A concept of events gone away. In R. Gibbs and H. Colston, editors, *Irony in Language and Thought*, pages 467–498. Taylor and Francis Group, 2007. 3, 31, 102

[92] C. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999. 36, 91, 95

[93] K. Markert and M. Nissim. Corpus-based metonymy analysis. *Metaphor and Symbol*, 18(3):175–188, 2003. 41

[94] K. Markert and M. Nissim. Data and models for metonymy resolution. *Language Resources and Evaluation*, 43(2):123–138, 2009. 41

[95] R. Mars, B. Rogier, S. Debener, T. Gladwin, L. Harrison, P. Haggard, J. Rothwell, and S. Bestmann. Trial-by-Trial Fluctuations in the Event-Related Electroencephalogram Reflect Dynamic Changes in the Degree of Surprise. *J. Neurosci.*, 28(47):12539–12545, 2008. 144

[96] M. Marschark. Metaphors in sign language and sign language users: A window into relations of language and thought. In H. Colston and A. Katz, editors, *Figurative Language Comprehension: Social and Cultural Influences*, pages 309–334. Lawrence Erlbaum Associates, 2005. 23, 24

[97] R. Mihalcea. The multidisciplinary facets of research on humour. In *3rd Workshop on Cross Language Information Processing, CLIP-2007, Int. Conf. WILF-2007*, volume 4578, pages 412–421, 2007. 4, 27, 45

[98] R. Mihalcea and S. Pulman. Characterizing humour: An exploration of features in humorous texts. In *8th International Conference on Computational Linguistics and Intelligent Text Processing, CICLing 2007*, volume 4394 of *LNCS*, pages 337–347, 2007. 2, 46, 71, 77, 80

[99] R. Mihalcea and C. Strapparava. Bootstrapping for fun: Web-based construction of large data sets for humor recognition. In *Proceedings of the Workshop on Negotiation, Behaviour and Language (FINEXIN 2005)*, volume 3814, pages 84–93, 2005. 53, 61

[100] R. Mihalcea and C. Strapparava. Computational laughing: Automatic recognition of humorous one-liners. In *Proceedings of the Cognitive Science Conference (CogSci)*, Stresa, Italy, July 2005. 61

[101] R. Mihalcea and C. Strapparava. Technologies that make you smile: Adding humour to text-based applications. *IEEE Intelligent Systems*, 21(5):33–39, 2006. 39, 45, 70, 71

[102] R. Mihalcea and C. Strapparava. Learning to Laugh (Automatically): Computational Models for Humor Recognition. *Journal of Computational Intelligence*, 22 (2):126–142, 2006. 1, 2, 28, 45, 56, 71, 133, 137, 157

[103] R. Mihalcea, C. Strapparava, and S. Pulman. Computational models for incongruity detection in humour. In *Proceedings of the 11th International Conference, CICLing 2010*, pages 364–374, 2010. 46

[104] G. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 19, 58, 60, 71, 157, 158

[105] A. Monge and C. Elkan. The field matching problem: Algorithms and applications. In *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 267–270, 1996. 106

[106] R. Morante and W. Daelemans. A metalearning approach to processing the scope of negation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 21–29, Boulder, Colorado, June 2009. Association for Computational Linguistics. 84

[107] R. Vega Moreno. Idioms, transparency and pragmatic inference. *UCL Working Papers in Linguistics*, 17:389–425, 2005. 25

[108] R. Vega Moreno. *Creativity and Convention: The pragmatics of everyday figurative speech*. John Benjamins, 2007. 19

[109] A. Nilsen. *Living Language: Reading, Thinking, and Writing*. Longman, 1998. 29, 56, 71

[110] M. Oakes. *Statistics for Corpus Linguistics*. Edinburgh University Press, 1998. 73

[111] A. Ortony. Metaphor, language, and thought. In Andrew Ortony, editor, *Metaphor and thought*. Cambridge University Press, 2nd edition, 1993. 22

[112] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008. 85

[113] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, Morristown, NJ, USA, 2002. Association for Computational Linguistics. 4, 121, 139

[114] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, pages 271–278, 2004. 121

[115] A. Papafragou. Figurative language and the semantics-pragmatics distinction. *Language and Literature*, 5:179–193, 1996. 22, 24

[116] T. Pedersen, S. Patwardhan, and J. Michelizzi. Wordnet::similarity - measuring the relatedness of concepts. In *Proceeding of the 9th National Conference on Artificial Intelligence (AAAI-04)*, pages 1024–1025, Morristown, NJ, USA, 2004. Association for Computational Linguistics. 103

[117] W. Pepicello and T. Green. *The language of riddles*. Ohio State University Press, Columbus :, 1984. 56, 64

[118] W. Peters. *Detection and Characterization of Figurative Language Use in WordNet*. PhD thesis, University of Sheffield, Sheffield, England, 2004. 16, 17, 22, 23, 41

[119] W. Peters and Y. Wilks. Data-driven detection of figurative language use in electronic language resources. *Metaphor and Symbol*, 18(3):161–173, 2003. 41, 105, 140

[120] P. Pexman. Social factors in the interpretation of verbal irony: The roles of speaker and listener characteristics. In H. Colston and A. Katz, editors, *Figurative Language Comprehension: Social and Cultural Influences*, pages 209–232. Lawrence Erlbaum Associates, 2005. 24, 29, 84, 93, 158

[121] R. Pierce, R. MacLaren, and D. Chiappe. The role of working memory in metaphor production and comprehension. *Journal of Memory and Language*, 56(2): 172–188, February 2007. 22, 40

[122] D. Pinto, P. Rosso, and H. Jiménez. On the assessment of text corpora. In *Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems (NLDB) 2009*, 2009. 74

[123] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980. 75

[124] M. Potthast. Measuring the descriptiveness of web comments. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 724–725, 2009. 118

[125] Amruta Purandare and Diane Litman. Humor: Prosody analysis and automatic recognition for friends. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 208–215. Association for Computational Linguistics, 2006. 46, 57

[126] U. Quasthoff, M. Richter, and C. Biemann. Corpus portal for search in monolingual corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1799–1802, 2006. 62

[127] V. Raskin. *Semantic Mechanisms of Humor*. Dordrecht, The Netherlands: D, 1985. 27, 28

[128] V. Rentoumi, G. Giannakopoulos, V. Karkaletsis, and G. Vouros. Sentiment analysis of figurative language using a word sense disambiguation approach. In *Proceedings of the International Conference RANLP-2009*, pages 370–375, Borovets, Bulgaria, September 2009. Association for Computational Linguistics. 38, 40

[129] A. Reyes and P. Rosso. Linking humour to blogs analysis: Affective traits in posts. In *Proceedings of the 1st Workshop on Opinion Mining and Sentiment Analysis (WOMSA), CAEPIA-TTIA Conference*, pages 100–109, Sevilla, Spain, 13 November 2009. 86

[130] A. Reyes and P. Rosso. Mining subjective knowledge from customer reviews: A specific case of irony detection. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 118–124. Association for Computational Linguistics, 2011.

[131] A. Reyes and P. Rosso. Building Corpora for Figurative Language Processing: The Case of Irony Detection. In *Proceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (in conjunction with LREC 2012)*, pages 94–98, 2012.

[132] A. Reyes and P. Rosso. Making objective decisions from subjective data: Detecting irony in customers reviews. *Decision Support Systems*, In press. DOI: 10.1016/j.dss.2012.05.027 http://dx.doi.org/10.1016/j.dss.2012.05.027. 83

[133] A. Reyes, D. Buscaldi, and P. Rosso. An analysis of the impact of ambiguity on automatic humour recognition. In *Proceedings of the 12th International Conference Text, Speech and Dialogue (TSD) 2009*, volume 5729 of *LNAI*, pages 162–169, 2009.

[134] A. Reyes, P. Rosso, and D. Buscaldi. Evaluating humorous features: Towards a humour taxonomy. In *Proc. Workshop on Web 2.0 and Natural Language Engineering Tasks, 4th Indian Int. Conf. on Artificial Intelligence, IICAI-2009*, pages 1373–1390, 2009.

[135] A. Reyes, P. Rosso, and D. Buscaldi. Affect-based features for humour recognition. In *Proceedings of the 7th International Conference on Natural Language Processing ICON-2009*, pages 364–369, 2009.

[136] A. Reyes, P. Rosso, and D. Buscaldi. Humor in the blogosphere: First clues for a verbal humor taxonomy. *Journal of Intelligent Systems*, 18(4):311–331, 2009. 74, 142

[137] A. Reyes, P. Rosso, A. Martí, and M. Taulé. Características y rasgos afectivos del humor: Un estudio de reconocimiento automático del humor en textos escolares en catalán. *Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*, (43):235–243, 2009. 61

BIBLIOGRAPHY

[138] A. Reyes, D. Buscaldi, and P. Rosso. The impact of semantic and morphosyntactic ambiguity on automatic humour recognition. In *Proceedings of the 14th International Conference on Applications of Natural Language to Information Systems NLDB 2009*, volume 5723 of *LNCS*, pages 130–141, 2010.

[139] A. Reyes, M. Potthast, P. Rosso, and B. Stein. Evaluating humour features on web comments. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 1138–1141, 2010.

[140] A. Reyes, P. Rosso, and D. Buscaldi. Finding humour in the blogosphere: the role of wordnet resources. In *Proceedings of the 5th Global WordNet Conference*, pages 56–61, 2010.

[141] A. Reyes, P. Rosso, and D. Buscaldi. From humor recognition to irony detection: The figurative language of social media. *Data and Knowledge Engineering*, 74:1–12, 2012. DOI: 10.1016/j.datak.2012.02.005 http://dx.doi.org/10.1016/j.datak.2012.02.005. 55

[142] A. Reyes, P. Rosso, and T. Veale. A multidimensional approach for detecting irony in twitter. *Language Resources and Evaluation*, Forthcoming. 117

[143] W. Ruch. The perception of humor. In World Scientific, editor, *Emotions, Qualia, and Consciousness. Proceedings of the International School of Biocybernetics*, pages 410–425, 2001. 3, 27

[144] W. Ruch. Computers with a personality? Lessons to be learned from studies of the psychology of humor. In *Proceedings of the International Workshop on Computational Humor (TWLT14)*, pages 57–70, 2002. 44

[145] M. Saif, D. Cody, and D. Bonnie. Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus. In *Proceedings of the 2009 Conference on EMNLP*, pages 599–608, Morristown, NJ, USA, 2009. Association for Computational Linguistics. 72

[146] L. Sarmento, P. Carvalho, M. Silva, and E. de Oliveira. Automatic creation of a reference corpus for political opinion mining in user-generated content. In *TSA '09: Proceeding of the 1st international CIKM workshop on Topic-Sentiment Analysis for mass opinion*, pages 29–36, Hong Kong, China, November 2009. ACM. 86, 142

[147] A. Pinar Saygin. Processing figurative language in a multi-lingual task: Translation, transfer and metaphor. In *Proceedings of Corpus-Based and Processing Approaches to Figurative Language Workshop*, Lancaster University, UK, March 29 2001. 23, 39, 40

[148] N. Schmidt and D. Williams. The evolution of theories of humour. *Journal of Behavioral Science*, 1:95–106, 1971. 27

[149] J. Schmitz. New approaches to conceptual representations in bilingual memory: The case for studying humor interpretation. *Bilingualism: Language & Cognition*, 3 (1):28–30, 2000. 143

[150] J. Schmitz. Humor as a pedagogical tool in foreign language and translation courses. *Humor*, 15(1):89–113, 2002. 143

[151] J. R. Searle. Literal meaning. *Erkenntnis*, 13(1):207 – 224, 1978. 13, 16, 17

[152] C. Shelley. The bicoherence theory of situational irony. In R. Gibbs and H. Colston, editors, *Irony in Language and Thought*, pages 531–578. Taylor and Francis Group, 2007. 93, 157

[153] L. Sikos, S. Windisch Brown, A. Kim, L. Michaelis, and M. Palmer. Figurative language: "meaning" is often more than just a sum of the parts. In *Proceedings of the AAAI 2008 Fall Symposium on Biologically Inspired Cognitive Architectures.*, pages 180–185, Washington, DC., 2008. Association for the Advancement of Artificial Intelligence. 15, 17, 37, 39

[154] J. Sjöbergh and K. Araki. Recognizing humor without recognizing meaning. In *3rd Workshop on Cross Language Information Processing, CLIP-2007, Int. Conf. WILF-2007*, volume 4578 of *LNAI*, pages 469–476, 2007. 1, 46, 77

[155] D. Sperber and D. Wilson. On verbal irony. *Lingua*, 87:53–76, 1992. 22, 30

[156] D. Sperber and D. Wilson. Relevance theory. *Handbook of Pragmatics*, 42(5):607–632, 2002. 19

[157] E. Stamatatos. Intrinsic Plagiarism Detection Using Character n-gram Profiles. In *Proc. of the 3rd Int. Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN09)*, pages 38–46, 2009. 103

[158] J. Stark, K. Binsted, and B. Bergen. Disjunctor selection for one-line jokes. In *In: Proceedings of INTETAIN 2005*, pages 174–182, 2005. 46

[159] O. Stock and C. Strapparava. Hahacronym: A computational humor system. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 113–116, 2005. 2, 44

[160] A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of 7th International Conference on Spoken Language Processing, INTERSPEECH 2002*, pages 901–904, 2002. 62

[161] C. Strapparava and R. Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied Computing*, pages 1556–1560, 2008. 28, 55, 57, 77

[162] C. Strapparava and A. Valitutti. WordNet-affect: an affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, volume 4, pages 1083–1086, 2004. 73

[163] J. Taylor and L. Mazlack. Humorous wordplay recognition. *IEEE International Conference on Systems man and cybernetics*, 4:3306–33116, 2004. 46

[164] H.W. Tinholt and A. Nijholt. Computational humour: Utilizing cross-reference ambiguity for conversational jokes. In *7th International Workshop on Fuzzy Logic and Applications (WILF 2007)*, Lecture Notes in Artificial Intelligence, pages 477–483. Springer Verlag, 2007. 44

[165] K. Triezenberg. Humor in literature. In Victor Raskin, editor, *The Primer of Humor Research*, pages 523–542. Mounton de Gruyter, 2008. 7

[166] O. Tsur, D. Davidov, and A. Rappoport. ICWSM – A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In William W. Cohen and Samuel Gosling, editors, *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 162–169, Washington, D.C., 23-26 May 2010. The AAAI Press. 2, 50

[167] M. Turner. Figure. In M. Marschark, editor, *Figurative Language And Thought*, pages 44–87. Oxford University Press, 1998. 157, 158

[168] A. Utsumi. A unified theory of irony and its computational formalization. In *Proceedings of the 16th conference on Computational Linguistics*, pages 962–967, Morristown, NJ, USA, 1996. Association for Computational Linguistics. 2, 31, 48

[169] A. Valitutti. *Computational Production of Affect-Based Verbal Humorous Expressions*. PhD thesis, University of Trento, 2009. 4, 27, 45, 72, 73

[170] T. Veale. Creative language retrieval: a robust hybrid of information retrieval and linguistic creativity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 278–287, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. 40

[171] T. Veale and Y. Hao. Comprehending and generating apt metaphors: a web-driven, case-based approach to figurative language. In *Proceedings of the 22nd national Conference on Artificial intelligence - Volume 2*, pages 1471–1476, Vancouver, British Columbia, Canada, 2007. AAAI Press. 23, 24, 40, 41

[172] T. Veale and Y. Hao. Learning to understand figurative language: From similes to metaphors to irony. In *Irony CogSci 2007: the 29th Annual Meeting of the Cognitive Science Society Nashville, Tennessee*, Nashville, Tennessee, 2007. 24, 29, 41

[173] T. Veale and Y. Hao. Support structures for linguistic creativity: A computational analysis of creative irony in similes. In *Proceedings of CogSci 2009, the 31st Annual Meeting of the Cognitive Science Society*, pages 1376–1381, 2009. 1, 2, 48, 104

[174] T. Veale and Y. Hao. Detecting ironic intent in creative comparisons. In *Proceedings of 19th European Conference on Artificial Intelligence - ECAI 2010*, pages 765–770, Amsterdam, The Netherlands, 2010. IOS Press. 48

[175] T. Veale, K. Feyaerts, and G. Brône. The cognitive mechanisms of adversarial humor. *Humor, The International Journal of Humor Research*, 19(3):305–338, 2006. 11

[176] M. Vernier and S. Ferrari. Tracking evaluation in discourse. In *Workshop at EUROLAN 2007 on Applications of Semantics, Opinions and Sentiments (ASOS'07)*, Iasi, Romania, 2007. 6, 136

[177] C. Whissell. The dictionary of affect in language. *Emotion: Theory, Research, and Experience*, 4:113–131, 1989. 94

[178] C. Whissell. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological Reports*, 105(2): 509–521, 2009. 104

[179] J. Wiebe and E. Riloff. Finding mutual benefit between subjectivity analysis and information extraction. *IEEE Transactions on Affective Computing*, 2:175–191, 2011. 139

[180] M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pages 60–68, Uppsala, Sweden, July 2010. University of Antwerp. 85

[181] D. Wilson and D. Sperber. On verbal irony. In R. Gibbs and H. Colston, editors, *Irony in Language and Thought*, pages 35–56. Taylor and Francis Group, 2007. 3, 31

[182] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 0(0):1–35, 2009. 139

[183] I. Witten and E. Frank. *Data Mining. Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers. Elsevier, 2005. 112

[184] T. Zagibalov, K. Belyatskaya, and J. Carroll. Comparable English-Russian book review corpora for sentiment analysis. In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 67–72, Lisbon, Portugal, 2010. 121

[185] J. Zinken. Discourse metaphors: The link between figurative language and habitual analogies. *Cognitive Linguistics*, 18(3):445–466, 2007. 23, 40

[186] S. Zrehen and M. Arbib. Understanding jokes: A neural approach to content-based information retrieval. In *Agents*, pages 343–349, 1998. 43

BIBLIOGRAPHY

Appendices

Appendix A

Literary Devices

Below are listed and exemplified some of the most common figurative devices regarding literary usages. The list is not exhaustive. However, the devices here considered are often related to the two major devices treated in this thesis: humor and irony.

Allegory : This device is often understood as an extended metaphor in which a story is told to illustrate an important attribute of the subject (cf. Wikipedia). According to Turner [167], along with analogy and parable, allegory is often defined as having to do with abstract conceptual patterns but not so clearly with linguistic patterns, since their products can be expressed in many forms.

Alliteration : Word repetition and rhyme, which produce a comic effect (Mihalcea and Strapparava [102]). e.g. Infants dont enjoy infancy like adults do adultery.

Analogy : Analogy holds between two concepts when they participate in the same abstract relational structures (Shelley [152], Feldman and Peng [48]). e.g. My job is a jail.

Antiphrasis The use of a word in a sense opposite to its normal sense (especially in irony). e.g. This movies is really *good* when actually sucks (Miller [104]).

Euphemism : An inoffensive or indirect expression that is substituted for one that is considered offensive or too harsh. e.g. in Mexican Spanish, *asno*

(donkey) is used to refer a stupid person (Miller [104]).

Hyperbole : Simulation that is false and potentially impossible in terms of its scale (Bergen [15]). It is typically illustrated with superlative modifiers (Turner [167]). e.g. He is sure the best friend in the universe.

Imagery : Imagery is a cognitive function to organize (and, when necessary, to reorganize) simple mental units into higher-order units which makes sense as a meaningful whole (Katz et al. [76]).

Litotes : It is a figure of speech in which understatement is employed for rhetorical effect when an idea is expressed by a denial of its opposite, principally via double negatives (Miller [104]). e.g You are not wrong. You are correct.

Oxymoron : Terms that normally contradict each other conjoining contradictory terms (Miller [104]). e.g. Deafening silence.

Parable : Extended metaphor told as an anecdote to illustrate or teach a moral lesson (Turner [167]).

Paradox : A statement that contradicts itself (Miller [104]). e.g I always lie.

Proverb : Succinct or pithy expression of what is commonly observed and believed to be true (cf. Wikipedia). It is highly unlikely that, out of context, items would make contact with prestored conceptual metaphors, though it is possible that syntactic factors might suggest an item is a proverb (Katz et al. [76]). e.g. Strike while the iron is hot (Pexman [120]).

Synecdoche : Special case of metonymy what traditional rhetoricians have called synecdoche, where the part stands for the whole (Lakoff and Johnson [85]). e.g. There are a lot of good heads in the university. (= intelligent people)

Synesthesia : In synesthesia, perceptual stimuli presented in one modality (e.g. as sound) consistently map to another modality (e.g. as a visual analog). Among the aspects of this phenomenon is that cross-modal mapping is evident in some cases soon after birth, that is, well before the development of language (Katz [74]).

Appendix B

Examples of Blogs in H4

1) Now it's true that I have a tendency to dislike most of the things related to the 60s. I can't help it. But I still tried to understand what was so great about this book, there was no point, I hated this book from the very beginning to the end. Pynchon's story is probably full of irony but I couldn't possibly feel it when I was too busy trying to concentrate on this awful writing.

2) The ongoing ridiculous situation brewing between bloggers and the Associated Press has now taken a turn towards the enjoyably hilarious. We had already mentioned the fact that, despite the AP's complaints that bloggers quoting less than 100 words were violating fair use, the AP had a long history of quoting more than 100 words from bloggers – and not even linking back to the original blog. Now, the AP's own article about this brouhaha quoted (without linking) twenty-two words from TechCrunch. That's 18 words more than the supposed four word "limit" the AP has suggested. With an ironic chance that wide, TechCrunch's Michael Arrington couldn't resist, and asked his lawyer to send a DMCA take down notice to the Associated Press, along with a bill for $12.50 (directly off the AP's own pricing schedule). He admits that it's ridiculous, but that's what his actions are designed to present. By law, the AP should be required to take down the content before filing a response – though, since it's filing the response to itself, then perhaps it won't need to take down the content. Either way, this helps illustrate the insanity of the entire situation.

3) Okay, sports fans, it's official – Cleveland weather sucks ass! Opening day at Jacobs Field (where all of AG.com was headed tomorrow as a company reward/party) has been cancelled, since we are now expecting 3-6 inches of snow! They are rescheduling the game for Tuesday afternoon, so we will probably at least get to go... Though in a sharp twist of irony, the conference call that I asked people to reschedule is going to have to be rescheduled again (originally on Monday due to forgetfulness, it was moved to Tuesday afternoon –) d'oh!). Ha ha. Okay, maybe it's only funny to me, and a handful of people at work who (probably) don't read this.

4) And it's a darn good thing my Monday ended on such a good note, because Tuesday sure as hell started as a train wreck. I really need to get better at detecting when people are bullshitting me. I really hate to just assume that everyone is bullshitting me, because that level of cynicism doesn't really jive with my core Pollyanna-like personality. On the other hand, I think that tending to believe people might not be the best thing either.

5) My landlady's still surreal and psychotic but that's no real news. I think she's pissed that ALL four of the cats now simply hang out in my room all the time. All of them sleep on my bed (or, in Chloe's case, under it) and hang out in the bathroom while I'm chilling in the tub. She still thinks its weird that I read most nights and listen to music instead of being glued to the TV in a dark room like she is. She will get up at 8am on Saturday and spend ALL DAY IN FRONT OF THE TV. I think I just fear BECOMING LIKE HER. Giving up. Hiding. Coccooning. I understand it must be extremely tough to have a spouse pass away but... that was over a decade. I don't think she'd be dishonoring the spirit of her husband if she got out occasionally. If I can do it, she can. I'd like to win the lottery just to send her on a cruise or something.

6) I'd like to abolish the insidious terms Darwinism, Darwinist and Darwinian. They suggest a false narrowness to the field of modern evolutionary biology, as though it was the brainchild of a single person 150 years ago, rather than a vast, complex and evolving subject to which many other great figures have contributed. (The science would be in a sorry state if one man

150 years ago had, in fact, discovered everything there was to say.) Obsessively focusing on Darwin, perpetually asking whether he was right about this or that, implies that the discovery of something he didn't think of or know about somehow undermines or threatens the whole enterprise of evolutionary biology today.

7) You don't have to read all the way through. If you just skim read it then you get the general joist of it and it is mediocally funny. The point of the joke (i think) is that it is long and slightly boring (THATS THE POINT!!!!!) and this is one joke on this website that i actually felt was slightly funny. If they made the joke shorter then there wouldn't be a joke at all!!!!

8) An Englishman, an American and an Italian are having a conversation, praising their respective countries. The Englishman says: During the last war we had a ship so large, but so large that for docking maneuvers we needed 24 hours. The American reply: We had a ship so big that to move on it, there was a bus service. And the Italian: This is nothing. We had a ship so large that when at bow the war was over, stern even knew that was started.

9) A man and his wife were spending the day at the zoo. She was wearing a loose fitting, pink dress, sleeveless with straps. He was wearing his usual jeans and T-shirt. As they walked through the ape exhibit, they passed in front of a large, silverback gorilla. Noticing the wife, the gorilla went crazy. He jumped on the bars, and holding on with one hand and 2 feet he grunted and pounded his chest with his free hand. He was obviously excited at the pretty lady in the pink dress. The husband, noticing the excitement, thought this was funny. He suggested that his wife tease the poor fellow some more by puckering her lips and wiggling her bottom. She played along and the gorilla got even more excited, making noises that would wake the dead. Then the husband suggested that she let one of her straps fall to show a little more skin. She did and the gorilla was about to tear the bars down. Now show your thighs and sort of fan your dress at him, he said. This drove the gorilla absolutely crazy, and he started doing flips. Then the husband grabbed his wife, ripped open the door to the cage, flung her in

with the gorilla and slammed the cage door shut. Now. Tell him you have a headache.

10) The final piece of advice is writing humor takes time. To excel in humor is a lifetime job, and is not something that you can learn in a day or two. Don't think you can read a joke book and start writing funny stuff an hour later. You will have to teach yourself how to be funny. The process is mostly by trial and error, observing other people's comical situations, mistakes, laughing and applying it on yourself, etc. No one can teach you exactly how to write something funny, but the possibilities of creating humor on anything and everything are limitless.

11) Many companies hold information meetings in the office is not practicing humor, because they do not want to have one of the workers who will be offended. However, at the time the company can cross boundaries on what is acceptable and not acceptable.

12) Part of the problem with people telling funny jokes or humor is not acceptable is that if someone can not enjoy the job itself in the workplace will be a drab and unhappy workers.

Appendix C

Set of Features in Signatures

In this appendix is given the list of features concerning pattern *signatures*.

C. SET OF FEATURES IN SIGNATURES

Table C.1: Features in pattern signatures.

Emoticons	Counter-factuality	Temporal compression
:) :]	about therefore	abruptly
(: [:	almost though	impromptu
;) 8-]	although thus	later
(; [-8	approximately virtually	now
:o :>)	around well-nigh	out of the blue
o: (<:	but withal	recently
:-o <:o)	close yet	shortly
o-: (0:>	even	since
:-O 8-)	hence	soon
O-: (-8	herefore	sudden
:(xD	however	suddenly
): Dx	just	tomorrow
:(B-)	less	whenever
)-: (-B	merely	
^^ =))	more	
:-) ((=	most	
(-: :L	near	
;-) L-)	nearly	
(-; (-L	nevertheless	
:=) :-D	nigh	
(=: D-:	no	
;=) ;-D	non	
(=; D-;	nonetheless	
:P -_-	not	
P: ;S	notwithstanding	
:p S;	now	
p: ;s	only	
:D s;	roughly	
D: :s	simply	
:d s:	so	
d: ;s	some	
(H) s;	still	
:$ haha	then	
$: lol	thence	

Appendix D

Examples of Patterns Regarding the Complex Irony Detection Model

In this appendix are given some examples regarding how the model is applied over the tweets.

1) *Pointedness*

   - ≻ The govt should investigate him thoroughly; do I smell **IRONY**
   - ≻ Irony is such a funny thing **:)**
   - ≻ Wow the only network working for me today is 3G on my iPhone. **WHAT DID I EVER DO TO YOU INTERNET???????**

2) *Counter-factuality*

   - ≻ My latest blog post is **about** how twitter is for listening. And I love the irony of telling you about it via Twitter.
   - ≻ Certainly I always feel compelled, obsessively, to write. **Nonetheless** I often manage to put a heap of crap between me and starting ...
   - ≻ BHO talking in Copenhagen **about** global warming and DC is **about** to get 2ft. of snow dumped on it. You just gotta love it.

D. EXAMPLES OF PATTERNS REGARDING THE COMPLEX IRONY DETECTION MODEL

3) *Temporal compression*

- ≻ @$ryan_{c}onnolly$ oh the irony that will occur when they finally end movie piracy and **suddenly** movie and dvd sales begin to decline sharply.
- ≻ I'm seriously really funny when nobody is around. You should see me. But **then** you'd be there, and I wouldn't be funny...
- ≻ RT @$Butler_{G}eorge$: **Suddenly**, thousands of people across Ireland recall that they were abused as children by priests.

4) *Temporal imbalance*

- ≻ **Stop** trying to find love, it will find you;...and no, he **didn't** say that to me..
- ≻ Woman on bus **asked** a guy to turn it down please; but his music **is** so loud, he **didn't hear** her. Now she **has** her finger in her ear. The irony

5) *Contextual imbalance*

- ≻ DC's snows coinciding with a conference on global warming proves that God has a sense of humor.
  Relatedness score of **0.3233**
- ≻ I know sooooo many Haitian-Canadians but they all live in Miami.
  Relatedness score of **0**
- ≻ I nearly fall asleep when anyone starts talking about Aderall. Bullshit.
  Relatedness score of **0.2792**

6) *Character n-grams (c-grams)*

- ≻ **WIF**
  More about Tiger - Now I hear his **wif**e saved his life w/ a golf club?
- ≻ **TRAI**
  SeaWorld (Orlando) **trai**ner killed by killer whale. or reality? oh, I'm sorry politically correct Orca whale

- ≻ **NDERS**
  Because common sense isn't so common it's important to engage with your market to really u**nders**tand it.

7) *Skip-grams (s-grams)*

- ≻ 1-skip: *richest ... mexican*
  Our president is black nd the **richest** man is a **Mexican** hahahaha lol
- ≻ 1-skip: *unemployment ... state*
  When **unemployment** is high in your **state**, Open a casino tcot tlot lol
- ≻ 2-skips: *love ... love*
  Why is it the Stockholm syndrome if a hostage falls in **love** with her kidnapper? I'd simply call this **love**. ;)

8) *Polarity s-grams (ps-grams)*

- ≻ 1-skip: *pos-neg*
  Reading $\textbf{glasses}_{pos}$ have $\textbf{RUINED}_{neg}$ my eyes. B4, I could see some shit but I'd get a headache. Now, I can't see shit but my head feels fine
- ≻ 1-skip: *neg-neg-pos*
  $\textbf{Breaking}_{neg}$ $\textbf{News}_{neg}$: New $\textbf{charity}_{pos}$ offers people to adopt a banker and get photos of his new bigger house and his wife and beaming mistress.
- ≻ 2kips: *pos-pos-neg*
  Just heard the $\textbf{brave}_{pos}$ $\textbf{hearted}_{pos}$ English Defence $\textbf{League}_{neg}$ thugs will protest for our freedoms in Edinburgh next month. Mad, Mad, Mad

9) *Activation*

- ≻ I enjoy(2.22) the fact(2.00) that I just addressed(1.63) the dogs(1.71) about their illiteracy(0) via(1.80) Twitter(0). Another victory(2.60) for me.

- ≻ My favorite(1.83) part(1.44) of the optometrist(0) is the irony(1.63) of the fact(2.00) that I can't see(2.00) afterwards(1.36). That and the cool(1.72) sunglasses(1.37).
- ≻ My male(1.55) ego(2.00) so eager(2.25) to let(1.70) it be stated(2.00) that I'am THE MAN(1.8750) but won't allow(1.00) my pride(1.90) to admit(1.66) that being egotistical(0) is a weakness(1.75) ...

10) *Imagery*

- ≻ Yesterday(1.6) was the official(1.4) first(1.6) day(2.6) of spring(2.8) ... and there was over a foot(2.8) of snow(3.0) on the ground(2.4).
- ≻ I think(1.4) I have(1.2) to do(1.2) the very(1.0) thing(1.8) that I work(1.8) most on changing(1.2) in order(2.0) to make(1.2) a real(1.4) difference(1.2) paradigms(0) hifts(0) zeitgeist(0)
- ≻ Random(1.4) drug(2.6) test(3.0) today(2.0) in elkhart(0) before 4(0). Would be better(2.4) if I could drive(2.1). I will have(1.2) to drink(2.6) away(2.2) the bullshit(0) this weekend(1.2). Irony(1.2).

11) *Pleasantness*

- ≻ Goodmorning(0), beauties(2.83)! 6(0) hours(1.6667) of sleep(2.7143)? Total(1.7500) score(2.0000)! I love(3.0000) you school(1.77), so so much(2.00).
- ≻ The guy(1.9000) who(1.8889) called(2.0000) me Ricky(0) Martin(0) has(1.7778) a blind(1.0000) lunch(2.1667) date(2.33).
- ≻ I hope(3.0000) whoever(0) organized(1.8750) this monstrosity(0) realizes(2.50) that they're playing(2.55) the opening(1.88) music(2.57) for WWE's(0) Monday(2.00) Night(2.28) Raw(1.00) at the Olympics(0).

Appendix E

Probability Density Function for Patterns in Complex Irony Detection Model

In this appendix are shown 11 graphs in which we depict the probability density function associated with all IDM's patterns. All these graphs are intended to provide descriptive information concerning the fact that the model is not capturing idiosyncratic features of the negative sets; rather, it is really capturing some aspects of irony.

For all the graphs we keep the following representation: #irony (blue line), #education (black line), #humor (green line), #politics (brown line).

E. PROBABILITY DENSITY FUNCTION FOR PATTERNS IN COMPLEX IRONY DETECTION MODEL

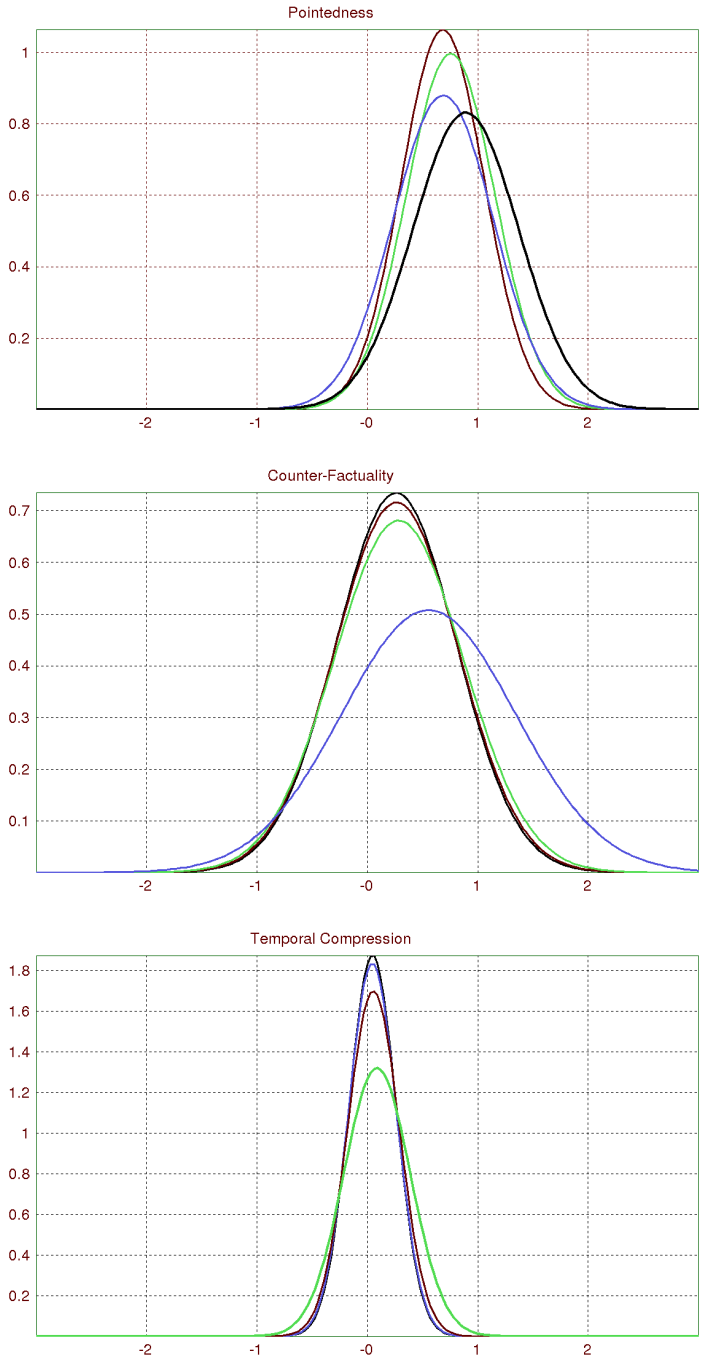

Figure E.1: Probability density function for dimensions in signatures.

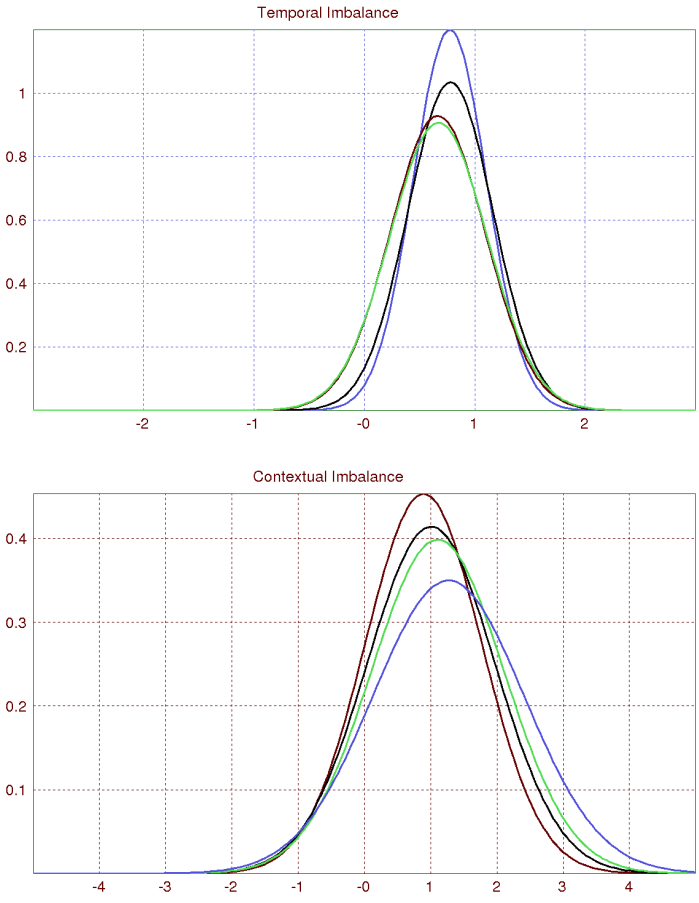

Figure E.2: Probability density function for dimensions in unexpectedness.

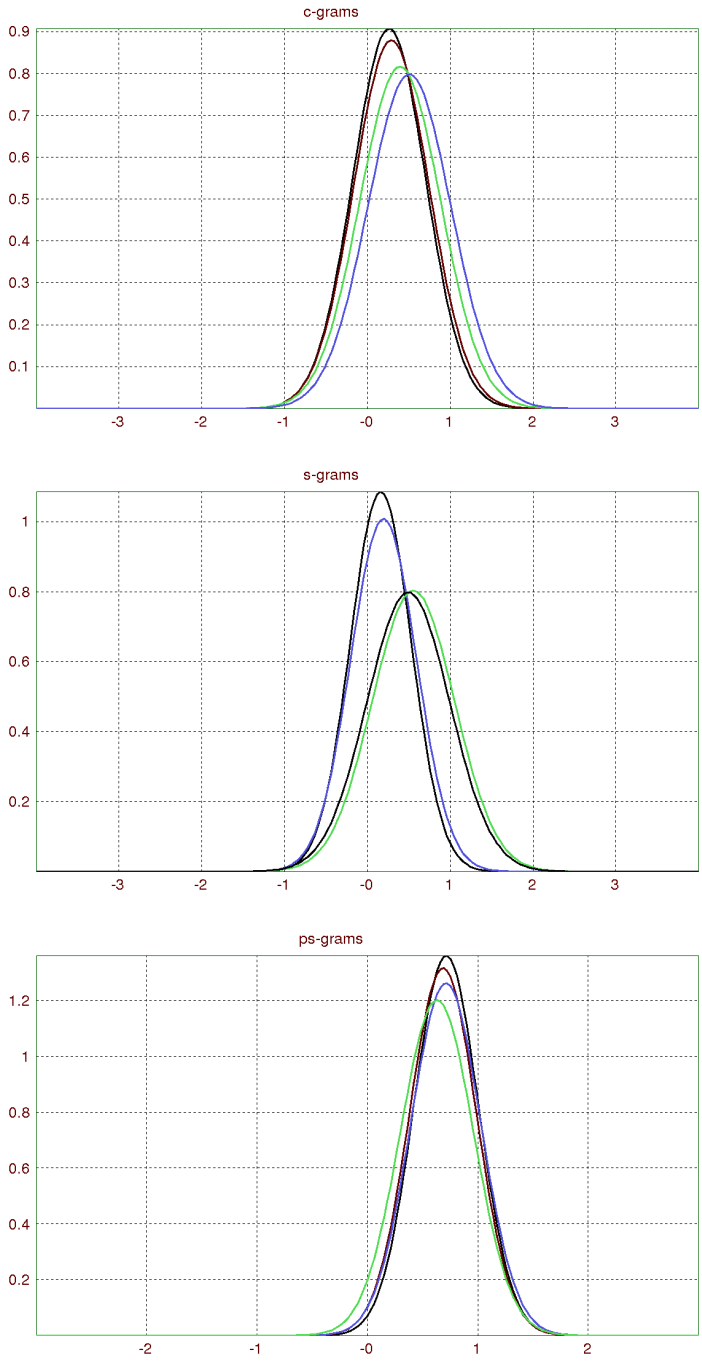

Figure E.3: Probability density function for dimensions in style.

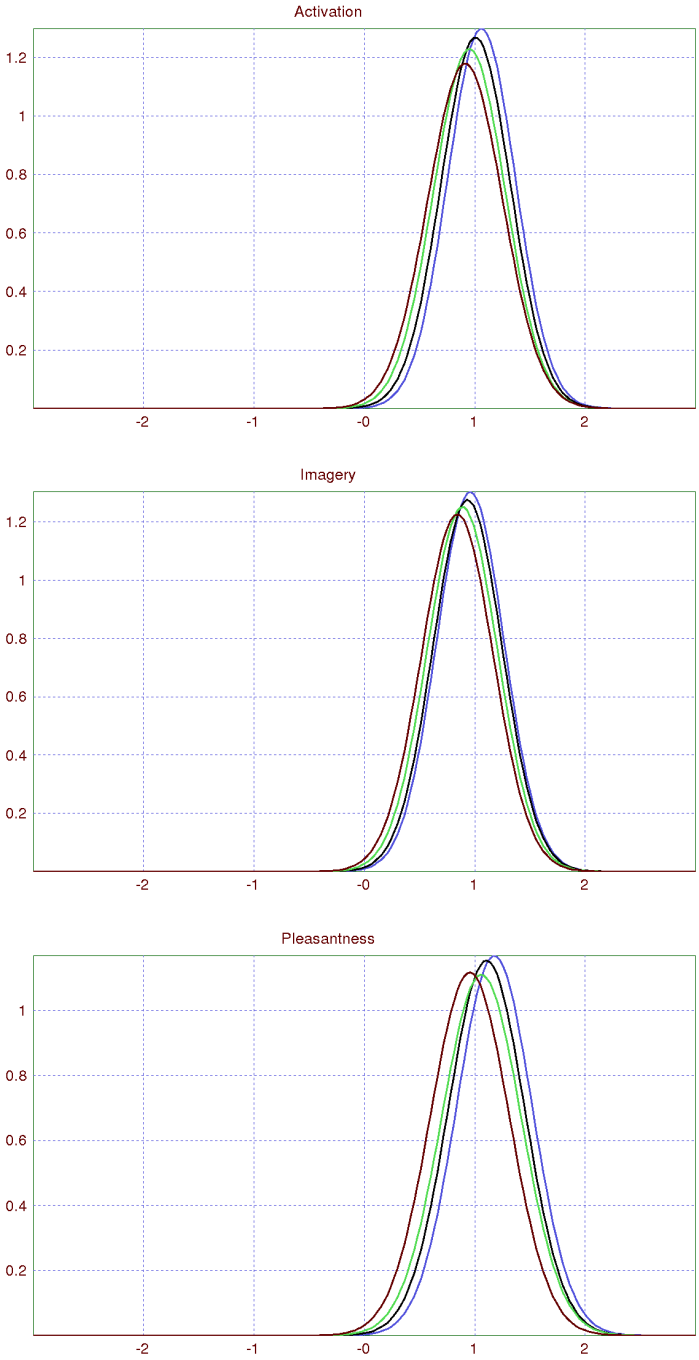

Figure E.4: Probability density function for dimensions in emotional contexts.

E. PROBABILITY DENSITY FUNCTION FOR PATTERNS IN COMPLEX IRONY DETECTION MODEL

Appendix F

Examples of the Most Ironic Sentences

According to IDM's predictions, here are presented some of the most ironic sentences. Each sentence has a document identifier. Such identifiers were kept as in the original data sets in order to facilitate their location.

1) *Movies*2

   - ≻ "Expecting them to give the viewer insights into the human condition is like expecting your car to vacuum your house " (doc. id. *cv116_28942.txt*).
   - ≻ "That degree of complexity combined with those very realistic looking dinosaur effects is just about as much as I require" (doc. id. *cv116_28942.txt*).
   - ≻ "Moulin Rogue is an original, and an original, even a flawed one, is a thing to be cherished" (doc. id. *cv275_28887.txt*).
   - ≻ "In some respects, Rush Hour is the ultimate exercise in cliched filmmaking. The hero is the renegade cop that prefers to work alone. The cop in question cannot solve the case until he gets in trouble. All chinese people are somehow involved in the criminal element. The duo must always be completely mismatched. The hero has to say some smart-assed comment before (and after) shooting someone. However, that doesn't necessarily make for a bad film" (doc. id. *cv402_14425.txt*).

- ≻ "Making her dramatic debut, after appearing in over 300 triple X adult films, porn star Nina Hartley takes command of her role with considerable assurance and a screen presence which puts many other contemporary 'straight' actresses to shame" (doc. id. *cv422_9381.txt*).
- ≻ "If I'd laughed any more, I might have needed an iron lung" (doc. id. *cv507_9220.txt*).
- ≻ "I never believed love at first site was possible until I saw this film" (doc. id. *cv513_6923.txt*).
- ≻ "Usually a movie is about something more than a soiled rug" (doc. id. *cv718_11434.txt*).
- ≻ "I remember really enjoying this movie when I saw it years ago. I guess my memory really sucks" (doc. id. *cv982_22209.txt*).
- ≻ "It's not that there isn't anything positive to say about the film. There is. After 92 minutes, it ends". (doc. id. *cv123_12165.txt*).
- ≻ "There's an enormous woman (played by transvestite porn star)" (doc. id. *cv142_23657.txt*).
- ≻ "However, isn't bad at all. The actors do the best they can with the bad material" (doc. id. *cv733_9891.txt*).

2) *Movies*1

- ≻ "The only actor in the movie with any demonstrable talent is a cute little prairie dog named Petey" (doc. id. *cv039_tok-11790.txt*).
- ≻ "This film needed that whole theatre-shaking: they needed to wake everybody up because they were so bored" (doc. id. *cv229_tok-9484.txt*).
- ≻ "Appreciate this movie for the few weeks it will be in theaters folks" (doc. id. *cv342_tok-24681.txt*).
- ≻ "I hated this movie for every second that I sat watching it, and I actively hate it now, days later, with the simpering, superficial, nauseatingly sentimental images forever plaguing my memories" (doc. id. *cv352_tok-15921.txt*).
- ≻ "It's too trashy to be good drama, but too dramatic to be good trash" (doc. id. *cv494_tok-11693.txt*).

- ≻ "I only wish that I could make that one hour and forty-five minutes of my life re-appear" (doc. id. *cv495_tok-18551.txt*).
- ≻ "In order to make the film a success, all they had to do was cast two extremely popular and attractive stars, have them share the screen for about two hours and then collect the profits" (doc. id. *cv176_tok-15918.txt*).
- ≻ "(Why, oh why, couldn't Lucas use computers to substitute better performers in the lead roles?)" (doc. id. *cv228_tok-8817.txt*).
- ≻ "Nostalgia appears to have a great appeal, but don't you think we could have more than 14 years before we yearn for the past?" (doc. id. *cv173_tok-11316.txt*).
- ≻ "The weak scenes could have been cut, but then there wouldn't have been much left" (doc. id. *cv198_tok-11090.txt*).
- ≻ "It's not a silent movie; there is lots of atmospheric music, occasional screams and weird sound effects, but nobody ever utters an audible word; unfortunately, is so bad that it's really bad" (doc. id. *cv524_tok-20616.txt*).
- ≻ "It seems that comedy is the main motive, and the violence is only intended to punctuate the laughs. Unfortunately, there are no laughs" (doc. id. *cv680_tok-12227.txt*).

3) *Books*

- ≻ "Essentially the entire plot can be summarised in a sentence of two, girl falls in love with boy, girl becomes damsel in distress, boy saves girl, end of....." (doc. id. *document 017 Negative*).
- ≻ "Yes that literally is the entire plot, but far worse than this is the complete lack of intelligent character design" (doc. id. *document 017 Negative*).
- ≻ "Harry goes to Hogwarts, bad guys try to kill Harry, battle with the bad guys, Harry triumphs - hurrah!" (doc. id. *document 043 Negative*).

- ≻ "In fact I could see myself possibly enjoying this book ten years ago. Than again, maybe not" (doc. id. *document 108 Negative*).

4) *Articles*

- ≻ "As we examine the passengers' cell-phone calls and flight recordings, we get a sense of the incredible courage displayed by these ordinary men and women" (doc. id. *014-test-0153.satire*).
- ≻ "Despite years of diplomatic stalemate in the Mideast crisis, Syrian officials appeared eager to mend troubled Arab-Israeli relations this week by participating in a second round of U.S.-led peace talks, which feature representatives from every country in the region, as well as a complimentary continental breakfast in the hotel lobby" (doc. id. *016-test-0165.satire*).
- ≻ "Unfortunately, most of the men and women who passed by seemed to speak only a bizarre Asian dialect unknown to me, and those who could communicate were more interested in selling me exotic cologne out of a duffel bag" (doc. id. *022-test-0294.satire*).
- ≻ "This is merely about improving liquidity, said King" (doc. id. *095-test-1483.satire*).
- ≻ "Virtually free, except for digging, pumping, processing, storage, by-product-disposal and shipping costs" (doc. id. *179-training-1407.satire*).
- ≻ "Maybe the one person who allowed Bush to ignore the opinions of 45 percent of America has a busy schedule" (doc. id. *144-training-0769.satire*).

Declaration

I herewith declare that I have produced this thesis without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This thesis has not previously been presented in identical or similar form to any other examination board.

The thesis work was conducted under the supervision of Dr. Paolo Rosso at Universitat Politècnica de València, Spain.

Valencia, Spain. July 2012.