Improving multilingual sentiment analysis using linguistic knowledge

Marilena Di Bari

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy

University of Leeds

School of Languages, Cultures and Societies

September 2015

The candidate confirms that the work submitted is her own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The work in Chapter 5 of the thesis has appeared in publication as follows:

SentiML: functional annotation for multilingual sentiment analysis. Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: metadata, vocabularies and techniques in the Digital Humanities, held in conjunction with DocEng 2013, Florence (Italy), 2013, Di Bari M., Sharoff S., Thomas M.

I was responsible for the design and practical development of the annotation scheme, and statistics on the data. The contribution of the other authors was a collaborative discussion on annotation scheme, and the paper revision.

The work in Chapter 6 of the thesis has appeared in publications as follows:

Multiple views as aid to linguistic annotation error analysis, Proceedings of the 8th Linguistic Annotation Workshop (LAW VIII), ACL SIGANN Workshop held in conjunction with Coling 2014, Dublin (Ireland), 2014, Di Bari M., Sharoff S., Thomas M.

I was responsible for the experiments and the writing of the paper. The contribution of the other authors was the idea of the multiple views, suggestions on experiments and paper revision.

Comparing sentiment annotations in English, Italian and Russian, Abstract Book of Corpus Linguistics 2015, Lancaster (UK), 2015, Di Bari M.

Operationalizing appraisal multilingually, Book of Abstracts of ISFC 2015 Challenging Boundaries, 42nd International Systemic Functional Congress, Aachen (Germany), 2015, Di Bari M.

Acknowledgements

At the end of this long and rich path, my thanks go first to the University of Leeds for accepting me as a student, rewarding me with an SMLC fee-bursary and giving me access to a variety of high-level courses and initiatives that have been essential to build my academic skills.

Of course, none of this would have been possible without the trust that Dr. Serge Sharoff, my first supervisor, has put in me since the very beginning. I am very grateful to both him and my second supervisor, Dr. Martin Thomas, for giving me space and guidance throughout the years on a project that - I can proudly say - has fulfilled my desire of creating a computational tool for linguists. Most importantly, they have always been supportive and flexible with respect to my choices, and I feel I have learnt much from them. In particular, this work would not have this form if it was not for their invaluable insights.

The self-taught training on Computational Linguistics I had to go through and that started with Dr. Atwell's course in NLP has certainly not been for the faint-hearted, and thanking my fiancé (and very soon husband) Michele a billion times is not near enough for what he has meant for me. I cannot dare to imagine a more supportive companion than him, that would push me to start a PhD, trigger and keep alive my interest for programming and the Computer science field, and *make me arrive at the end* by being ALWAYS on my side. He is the one that has made everything possible, including having an automatic system as practical output of this work by assisting me with its implementation. I am so proud that guiding me in so many ways has not prevented him for standing out as PhD student at the University of Manchester. *Grazie con tutto il cuore*, I cannot wait to embark in even bigger adventures together.

Deep and many many thanks go to my family for understanding my choices, and not making me feel guilty for being far from them. I am very grateful to each of them for finding their own way to show me eternal and infinite love and support despite difficulties and changes. This helped me realise which kind of person I want to be, and that I want to be there for you.

My loving thanks go also to each of the members of Michele's family, who I feel lucky and proud to call also my own by now. It is just exceptional that since the very first day seven years ago we haven't collected anything but happy memories, love, respect to each other and, most importantly, the will to reinforce them.

My special thanks also go to Mavis Bracegirdle and Sean Connolly for being my mentors and true friends throughout. Without them I would not even dared to apply for a PhD in the UK. They not only made it possible with their constant moral and concrete encouragement, but most importantly they truly inspired as professionals and human beings, and helped me in so many ways when I had the privilege to spend my days with them.

I also feel blessed for getting invaluable professional and teaching experience. As for the first, I have to thank Dr. Serge Sharoff, Dr. James Wilson and Dr. Paul Stephenson for getting me involved in the Intellitext project, and the members of the Committee for Foreign Language Assistants in West Yorkshire, especially Nicky and Alessandra. As for the second, I have to sincerely thank staff in the Centre for Translation Studies, the Language Centre and the Italian department of the University of Leeds, and the Italian department of the University of Manchester, for trusting my skills and welcoming me in a familiar environment. Among all the managers and colleagues, I would like to thank in particular Alessandra, Patrizia and Simone, for being the best mentors and friends one could ask for. Thanks to your passion, enthusiasm and support I learnt so much that I now feel strong enough to use my preparation in whatever environment I might come across.

There are many other old and new friends that I would like to thank for sharing your knowledge, time and feelings with me. A big space is deserved by those in Leeds with whom I shared the PhD experience, especially Vivian for the many ways in which she has always been close to me, Marco for sharing our desks and working days, Mika for making up the time for very nice occasions despite her busy life, Hassan for being my best bad friend, Nicha for showing her affection even now that she is far away, but also also the co-organizers of the L@L PGR group, members of the editorial board of the 'Leeds Working Papers in Linguistics and Phonetics', with whom I have also organized the L3T conference with the help of Professor Jeremy Munday and Dr. Serge Sharoff. I also would like to thank my friends from the University of Manchester: the CDT first cohort, the whole GNTeam led by Dr. Goran Nenadic (particularly George for being always there, with the longest coffees and best film nights of our lives), Ignazio, Eleni, Chiara, Mercedes, Warren and George from the office 2.90 and Michele G. for our periodical Italian coffees. I would also thank people met during the conferences, especially Professor Maite Taboada for showing interest in my work and opening my mind on some of the potential future works.

Last, but not least, I am very grateful to my examiners for agreeing to spend time to read a work that I sincerely hope it is interesting and mature enough.

Abstract

The need for the automatic analysis of opinions in written texts, which has been growing in recent years in several domains, has made *Sentiment Analysis* a very popular field (Liu, 2012).

In this area, systems have been traditionally classifying sentences as positive or negative only in accordance to the sentiment that words most frequently assume (e.g. “angry” negative, “beautiful” positive).

Such strategies present two main limitations:

- Multiple opinions often appear in the same sentence, with each expressing an opposing sentiment on different subjects (e.g. a positive opinion is expressed on the plot of a film, but a negative one on the actors’ performance).
- The most frequent sentiment, collected in *sentiment dictionaries*, does not take into account the fact that context often alters the orientation. Sentiment dictionaries have also been demonstrated to have small coverage (Di Bari, 2015; Di Bari *et al.*, 2013).

As a consequence, I propose an automatic system based on deep linguistic knowledge given in particular by dependency parsing relations (Nivre, 2005) and by attributes taken from the *Appraisal framework* (Martin & White, 2005), a theory concerned with the language of evaluation, attitude and emotion within *Systemic Functional Linguistics* (Halliday, 1978).

As a basis for the creation of the automatic system, I tailored an annotation scheme called *SentiML* inspired by previous works (Bloom & Argamon, 2009; Bloom *et al.*, 2007a; Whitelaw *et al.*, 2005) and carried out the annotation task in three languages (English, Italian and Russian) by

using *MAE* (Stubbs, 2011). The resulting corpora consist of around 500 sentences and 9000 tokens for each language. The corpora contain both original texts and translations of different types: news, political speeches and TED talks (Cettolo *et al.*, 2012).

The foundation of *SentiML* lies in the fact that an opinion can be captured in a pair consisting of usually two words with different functions: a *target* as the expression the sentiment refers to, and a *modifier* as the expression conveying the sentiment. The pair consisting of the target and the modifier altogether is called *appraisal group*. Along with these main categories, the annotation includes their attributes, among which the most important are the appraisal type according to the *Appraisal framework* ('affect', 'appreciation', 'judgement') and the orientation ('positive' or 'negative', both out-of-context and contextual).

A detailed manual analysis of the translation strategies (Baker, 2002) and the appraisal types across the corpora, supported by insights from *Corpus Linguistics* has been carried out. The most interesting expressions found during such analysis have been automatically analysed afterwards with the aim of having a further evaluation of the system. Nonetheless, the main evaluation consists of a comparison with a rule-based system that makes use of already existing tools such as the part-of-speech (POS) tagger and the sentiment dictionary.

The main objective of this work is to demonstrate that the *Appraisal framework* and *Sentiment analysis* can successfully support each other. The additional consideration that this has been done not only for English, but in parallel for Italian and Russian (and as one of the first applications of the Appraisal Framework in these languages) and for different text types, makes the research unique. Moreover, because the methodology used to compare a variety of linguistic features (morphological, grammatical, lexical, syntactical) at work in sentiment analysis has been applied to three languages belonging to different families (Germanic, Romance and Slavonic), it is expected to be generalizable to other languages.

As far as the practical applications are concerned, the automatic system could be used in any field in which written opinions need to be analysed. In the meanwhile, the new individual resources such as the annotated corpora and the *Maltparser* models for Italian and Russian have been made publicly available[1].

[1] http://corpus.leeds.ac.uk/marilena/SentiML/

Abbreviations

SA	Sentiment Analysis
AF	Appraisal Framework
SFL	Systemic Functional Linguistics
CL	Corpus Linguistics
SL	Source Language
TL	Target Language
ST	Source Text
TT	Target Text
CDA	Critical Discourse Analysis

Note: Although the present thesis is written in British English, spelling in American English is preserved in the examples.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The problem

Recent years have seen the increase in the availability of data related to customers' opinions on a variety of subjects, and in the interest of exploiting them. As a consequence, a brand new field called *Sentiment analysis* (SA) (or *Opinion mining*) has been rapidly growing. Its aim is to analyse people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organisations, individuals, issues, events, topics, and their attributes (Liu, 2012). Sentiment analysis applications have spread to several domains, from consumer products, services, health care and financial services to social events and political elections (Liu, 2012).

Studies in this field have seen a change in their focus over the years. While at first they aimed at classifying entire documents into positive or negative without caring much about an accurate analysis, nowadays the challenges are represented both by "really" understanding whether a single opinion expresses (or implies) a positive or negative sentiment, and by going beyond such binary classification.

The problem is that most of the times, no matter which context is under analysis, opinions are not expressed as simple and direct assertions, but by using a number of stylistic devices (Shastri *et al.*, 2010) such as pronominal references, abbreviations, idioms and metaphors, not to mention slang, *SMSese*, mistakes and typos (Zagibalov *et al.*, 2010).

In addition to these features that we can define as "general" since they affect the majority of the studies in Natural Language Processing (NLP), Sentiment analysis has to face the possibility that multiple opinions on different entities occur in the same sentence, sometimes expressing opposing sentiments (Shastri *et al.*, 2010). For example, if an opinion is supposed to be on the plot of a movie, it is not unusual that the user starts talking about actors' performance or director's choices. Sentiment systems should then cope with such shift in topics. A further challenge is represented by the automatic identification of sarcasm, irony and humour, which may influence the correct detection in particular of positive opinions (Carvalho *et al.*, 2009).

In terms of how automatic classification works, until now many approaches have been relying on the *bag-of-words* model (Harris, 1954), which assumes that the word order has no significance (for example the term *home made* would be treated in the same way as *made home*) because it only looks at word forms and distributions (Harris, 1951).

The theory of which Harris, and later Chomsky, were exponents, called *formalism*, is in contrast to *functionalism* that inquires "how the arrangement of words, phrases, and utterances are determined by social and historical conditions, including authentic contexts of situation, and speakers' and hearers' status plans and goals" (de Beaugrande, 1997). In fact, Halliday's *Systemic Functional Linguistics* (SFL) is the study of how people exchange meanings through the use of language (Halliday, 1978). For Halliday a central theoretical principle is that any act of communication involves choices. *System networks* capture the sets of choices available to language users at each stratum of the linguistic system, and the internal organisation of language "embodies a positive reflection of the functions that language has evolved to serve in the life of social man" (Halliday, 1973).

The present work has its place among those in Sentiment analysis that have valued syntax as opposed to handling language as a *bag of words*. In addition, as I will explain in detail in the following sections, since the proposed methodology has been carefully designed to be applied to different languages, it has the further advantage of contributing to the almost complete lack of theoretical multilingual studies in Sentiment analysis and of available resources.

1.2 Aims and objectives

The main aim of this work is to apply a *functional approach* to the analysis of evaluative language in order to improve the accuracy in Sentiment analysis multilingually.

This has been done by designing a methodology to extract a variety of linguistic features important for the accurate classification of sentiment into positive and negative. Among these, particular relevance is given to syntactical relations under the form of *dependency relations* (Nivre, 2005). At the most basic level, dependency relations allow to elicit the target and the way in which it is evaluated, for example *princess* and *beautiful* in "The princess has been judged the most beautiful" (more details on the methodology will be provided in Section 1.3). At a more advanced level, they also allow to overcome some of the main issues in Sentiment analysis mentioned in Section 1.1, namely the change of sentiment due to the presence of negation and *polarity reversals*. In fact, the use of the dependency relations is a way to recognize negations and polarity reversals, and thus change the out-of-context sentiment of words, which is collected in *sentiment dictionaries*.

One of the other aims was to push further the boundary of the automatic analysis by going beyond the binary classification of opinions into positive or negative. For this reason, I grounded my research in the *Appraisal Framework* (AF) (Martin & White, 2005), the theory within Systemic Functional Linguistics (SFL) focusing on the language of evaluation, attitude and emotion. The use of SFL (and the AF) is not new in computational contexts in so far as "when confronted with real texts, the theory [is] sufficient to provide formal instructions for interpreting those texts; and when confronted with meanings, the theory [is] sufficient for providing formal instructions that motivate natural texts corresponding to those meanings" (O'Donnell & Bateman, 2005). In the case of the evaluative language, my hypothesis is that features derived from two of the main categories of the AF would support the scopes of Sentiment analysis. These are the *attitude sub-system*, since it focuses on *how* one expresses opinions, and the *graduation sub-system*, since it focuses on the intensity of the opinions. In Figure 1.1 an excerpt of the framework with the categories used in this research is shown, while a complete description of the AF and its contributions to this research will be given in Chapter 3 and Chapter 5. The way in which I have combined these features has been through the creation of an annotation scheme called *SentiML*, whose main characteristics will be anticipated in Section 1.3 and fully described later in Chapter 5.

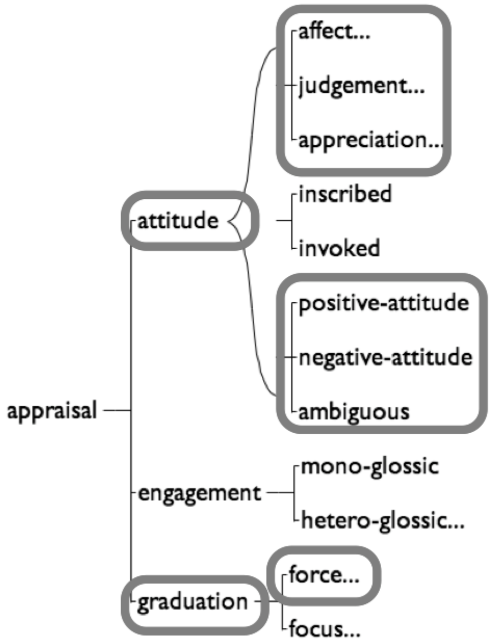

Figure 1.1: Excerpt of the Appraisal Framework. The highlighted concepts are those taken into account in this research.

SentiML has been designed to improve the accuracy in Sentiment analysis multilingually and so it is based on features general enough to be applied to languages belonging to different families. In this work I tested them on English, Italian and Russian belonging to the Germanic, Romance and Slavonic families respectively.

The manual annotation task has been carried out on the texts described in Section 1.4, chosen because belonging to different text types and because being originally-produced and translations. In addition, since manual annotation is a highly subjective task, I tried to avoid errors as much as possible by adopting two strategies: first of all, by relying on the speaker's (or the writer's) perspective and, second, by finding inconsistencies through the comparison between the manual annotations and those predicted by automatic classifiers.

The main practical outcome of my research has been an automatic system for English, Italian and Russian, based on rules for each of the linguistic features important for the accurate analysis and classification of appraisal groups. One of the main aspects of the classification is the orientation (i.e. positive, negative, ambiguous) of the appraisal groups, for which the system takes in input the sentiment of the individual words from the sentiment dictionaries, adjusts it according to the context and gives in output the overall correct one. The other main aspect of the classification is the attitude according to the AF (i.e. 'affect' judgement' and 'appreciation'). This last one, apart from representing a challenging task, is mostly new in the case of Italian and Russian where there are still very few works that have applied the Framework even just for the manual analysis of the data. The present work could then represent a unique contribution to both the linguistic analysis based on the Framework thanks to the annotated corpora, and the computational analysis thanks to the automatic system.

The performances of the system have been evaluated considering the manually-annotated corpora as gold standard as well as a baseline.

To summarise, the aims have been to:

1. Investigate and propose a method of automatic connection of each sentiment expression to its target.

2. Design a scheme to integrate the main categories of the Appraisal Framework and other advanced linguistic features to specify the attributes of the appraisal expressions.

3. Apply the Appraisal Framework for computational purposes to languages in which there are few or no works with the same scope, i.e. Italian and Russian.

4. Overcome subjectivity as much as possible by using the predictions of automatic classifiers to test the annotations.

5. Evaluate the effectiveness of the methodology in all the three languages by quantitatively and qualitatively measuring the results of the automatic system.

6. Investigate three different text types to generalize results.

7. Investigate independently-produced and translated texts to find similarities and differences among languages and their relative cultures.

These aims matched the following objectives:

1. Create an automatic system for Sentiment analysis based on the same methodology for three typologically different languages.

2. Create and distribute the specifically-designed annotation scheme *SentiML*, which can be applied to languages other than those under analysis in this work.

3. Create and distribute comparable corpora in English, Italian and Russian, collected and annotated to explore typological and cultural differences, which however can also serve other purposes due to their *reusability* and *multifunctionality* (McEnery *et al.*, 2006).

4. Provide an evaluation of the sentiment dictionaries in three languages from the perspectives of their coverage and accuracy.

5. Provide an evaluation of existing dependency-parsing resources in English, and of updated ones in Italian and Russian.

6. Provide a list of issues related to the application of dependency parsing for each of the languages.

7. Provide a comparison of the use of the dependency parsing on different text types across languages.

8. Evaluate the proposed methodology in two case studies: independently-produced texts and translated texts.

9. Substantiate the claims concerning the highly complex and heterogeneous linguistic systems of English, Italian and Russian in the final discussion by relying both on previous studies in *Translation Studies* and evidence from *Corpus Linguistics*.

1.3 Hypotheses and research questions

In this Section I will provide further details on the hypotheses mentioned before. The fundamental research question that motivates my research is "How far is it possible to analyse explicit opinions in order to bring together both a linguistic and a computational perspective?".

In order to address this question, I started from the definition of appraisal expressions adopted by previous works (Bloom & Argamon, 2009; Bloom *et al.*, 2007a; Nakagawa *et al.*, 2010; Zhao *et al.*, 2011), i.e. expressions comprising a *source*, an *attitude*, and a *target*. For example, in “I found the movie quite monotonous”, the speaker (the source) expresses a negative attitude (“quite monotonous”) towards “the movie” (the target). However, I decided to use pairs rather than triples in order to cover those cases in which either the subject or the object was stated, for example in “The stars sparkle so much tonight!”. The pairs consist of usually two words with different functions: a *target* as the expression the sentiment refers to, and a *modifier* as the expression conveying the sentiment. For example in the sentence “The chief is not just angry, he is scared” the target is *chief* and the modifiers are *angry* and *scared*. Such pairs are called *appraisal groups*. In the example above there are two of them: “chief angry” and “chief scared”.

The second research question is “What are the linguistic features of evaluative language that can lead to a successful automatic analysis of sentiment across multiple languages?”.

In order to provide an answer, I started from the set of features specified in (Bloom *et al.*, 2007a): attitude, orientation, force, polarity and target type. I included them in my annotation scheme (described in Chapter 5), by assigning them to either targets, modifiers or appraisal groups. The hypothesis has been that each of these features would have allowed to elicit and test the challenges in Sentiment analysis mentioned in Section 1.1, namely the change in the original sentiment of words due to the presence of negations, polarity reversals and the influence of the context (see Section 2.2 for a full description).

The third research question is “How far is the automatic classification of opinions into the main categories of the Appraisal Framework within Systemic Functional Linguistics possible and useful?”.

I have used the categories ‘affect’, ‘appreciation’ or ‘judgement’ from the subsystem *attitude* of the Appraisal Framework by (Martin & White, 2005). If successful automatic classification of the sentiment of opinions is possible, the advantages will in addition be the achievement of a new goal since it will lay the foundations for a more linguistically-informed analysis of the targets of opinions in the future.

The approach needed to answer all these research questions has to be *functional* because the *bag-of-words* approach mentioned in Section 1.1 would have not allowed to connect each sentiment expression to its target. In addition, a functional approach, by eliciting the syntactic relations among words, allows to extract the linguistic features useful for the correct classification of the sentiment of the appraisal expressions. The choice of using *dependency-based* relations rather than *constituency-based* relations has been motivated by their usefulness for the task. In a dependency tree, words and their POS-tags are linked through labelled arcs that express grammatical functions (see Figure 1.2).

Dependency relations have been used in a few studies in Sentiment analysis (see Section 2.5). In my case, they are particularly useful because they “are close to the semantic relationships needed for the next stage of interpretation” (Nivre, 2005). To exemplify this, I will use the sentence “Let both sides unite to heed in all corners of the earth the command of Isaiah to "undo the heavy burdens...and to let the oppressed go free."” taken from Kennedy’s Inaugural speech. In this sentence, the appraisal groups would be “sides unite”, “heed command”, “Isaiah command”, “undo burdens”, “heavy burdens”, “go free”[1]. The dependency representation in Figure 1.2 shows that the dependency parser would connect the appraisal groups in the way they are meant to, along with grammatical connections very similar to those that one has in mind when linking targets and modifiers in appraisal groups: *sides* is subject of *unite*, *command* is object of *heed*, *Isaiah* is modifier of *command*, *burdens* is object of *undo*, *heavy* is modifier of *burdens*, and *free* is complement of *go*.

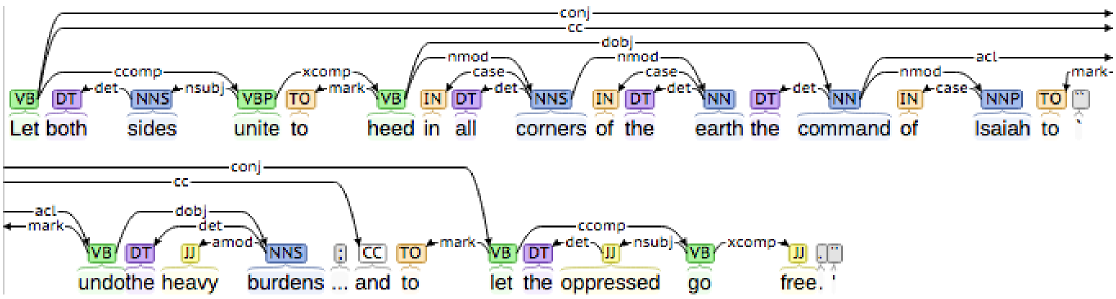

Figure 1.2: Dependency structure of the sentence “Let both sides unite to heed in all corners of the earth the command of Isaiah to "undo the heavy burdens...and to let the oppressed go free."”

[1]The group “oppressed free” has not been included because an appraisal group cannot consist of two adjectives, while “Isaiah command” has been included because the speaker’s perspective always dominates and in this case we can easily infer that the speaker considers it positive.

On the other hand, it is more difficult to find these links expressed in such a convenient form in the constituency representation in Figure 1.3[1]. In constituency-based relations, words are clustered in *phrases* that it is always possible to connect, but through a path of different length. In the example sentence, to connect *sides* (with `NNS` as POS-tag) to *unite* (with `VB` as POS-tag), the path `NP-S-VP` has to be covered, or to connect *heed* (with `VB` as POS-tag) to *command* (with `NN` as POS-tag), the path `VP-NP-NP` has to be covered.

Moreover, while it is possible to use rules to navigate the tree and find, for example, the noun connected to a particular adjective, the nature of their relations (e.g. modifier subject, complement) is not as explicit as in the dependency-based relations. In Section 2.5 I will explain in more detail which are the linguistic features part of the annotation scheme that I expect will benefit from the use of dependency-parsing relations.

[1]Stanford parser relations by `http://nlp.stanford.edu:8080/corenlp/process` and `http://nlp.stanford.edu:8080/parser/index.jsp`, with visualisation provided by the website `http://www.ark.cs.cmu.edu/parseviz/`

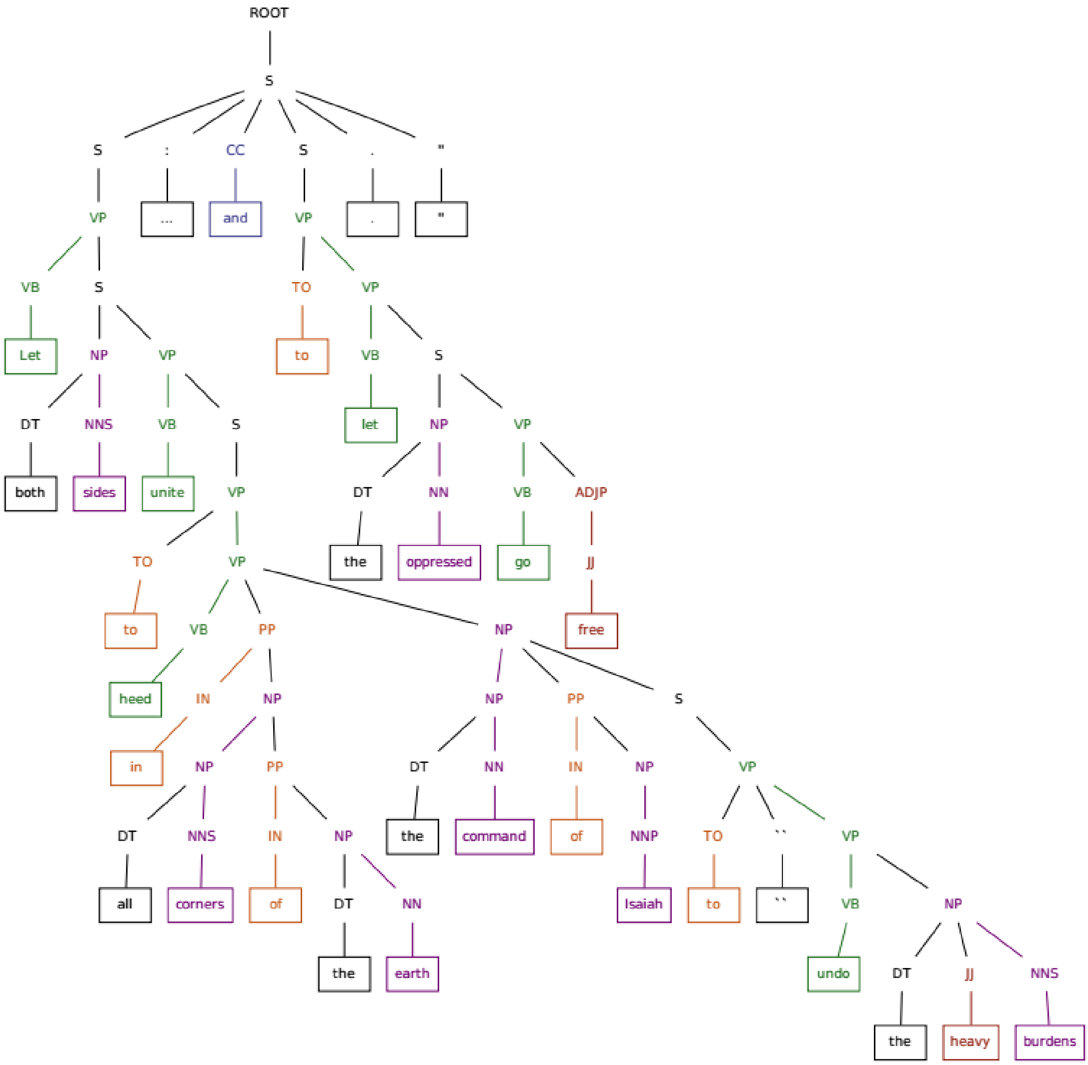

Figure 1.3: Constituency structure of the sentence “Let both sides unite to heed in all corners of the earth the command of Isaiah to "undo the heavy burdens...and to let the oppressed go free."”

Finally, in order to come up with generalizable results, I decided to explore different text types in the three languages. This would also allow me to test the possibility that the dependency parsers behave differently according to the text type and, in particular, the hypothesis that news are the most difficult to parse because of their formal register (see details in the following Section).

1.4 Data

I have chosen to build a multilingual *comparable* corpus (as defined by (Baker, 1995), consisting of texts belonging to three text types:

- **Political speeches**. American presidents' addresses in English[1], and their translations in Italian[2] and Russian[3].
- **TED (Technology, Entertainment, Design) Talks** in English, and their translations in Italian and Russian[4] (Cettolo *et al.*, 2012).
- **News**. Belonging to the *human rights* domain of the *MPQA opinion corpus* (Wilson, 2008) for English, to *Sole24ore* for Italian[5] and to Project Syndicate[6] and Global Voices[7] for Russian[8].

The corpus can also be defined as *unidirectional* since it includes political speeches and TED talks translated from English into the other two languages (Italian and Russian), but not the translations into English of the original Italian and Russian news.

The reasons behind the choice of these text types are especially related to the difference in *register*, summarised in the analysis of its components *field, tenor* and *mode* in Table 1.1 (Delin, 2000). By having such a spectrum of values (e.g. formal/informal as tenor, written/spoken as mode), I could ensure generalization.

	Political speeches	**Talks**	**News**
Field	Institutional, informative	Informative (for general knowledge)	Informative (for updating)
Tenor	Formal	Informal	Formal/informal
Mode	Written to be read	Spoken (not spontaneous)	Written

Table 1.1: Comparison of registers (through their components field, tenor and mode) across the text types political speeches, news and TED talks.

[1] `http://avalon.law.yale.edu/subject_menus/inaug.asp`
[2] `http://www.repubblica.it/2009/01/sezioni/esteri`
[3] `http://iipdigital.usembassy.gov/iipdigital-en/index.html`
[4] `http://www.ted.com/talks`
[5] `http://www.ilsole24ore.com/`
[6] `https://www.project-syndicate.org/`
[7] `http://globalvoicesonline.org/`
[8] Alternatives such as *Reuters* were excluded because, despite providing news on the same topics, these were not translations.

At the sub-level of *tenor*, which is the most relevant one in the case of the evaluative language, the difference can be seen in the speakers' status, plans and goals (Delin, 2000; Hunston & Thompson, 2000). This will be described in more detail in Chapter 3:

- Political speeches. Speakers in this case have a highly-regarded status. Their plan is first to thank the audience, then to create a common background (e.g. by reminding past important events) before introducing the present situation. The final part of their speech is dedicated to future plans and final greetings. Their goals can be to motivate, to alert, to reassure and to raise pride.

- TED Talks. Speakers in this case are experts and, like in the previous case, have to make an effort to keep the level of the attention high, especially considering that TED talks do not have a diffusion comparable to that of a speech to the nation.

- News. In this case journalists are generally not experts, with exceptions depending on the section they are writing in (for example in the "Broken hearts" section, the journalist might have a background in psychology). Since the audience reads news to stay updated, their style is necessarily different in comparison to the political speeches and talks.

These three text types also have the practical advantages of being publicly available, and featuring longer and better-formed sentences that were expected to work better with parsing, as opposed to phenomena such as slang, SMSese and typos commonly found in other text types like reviews.

The corpora consist of around 500 sentences and 9000 tokens per language, specifically 328 sentences in Italian, 459 sentences in Russian and 462 sentences in English. In Table 1.2 the partial number of words according to the text type as well as the total per language is shown. The annotation has been carried out using MAE (Stubbs, 2011), a multi-platform freely available annotation software. The SentiML scheme used for the annotation will be described in Chapter 5, while the annotation process and its results will be described in Chapter 6. In the meanwhile, in the following Chapter I will start clarifying the objects of Sentiment analysis and the previous works related to my research.

Text type	**# words**		
	English	Italian	Russian
Political	3782	3960	3408
News	2281	2316	3094
TED	2992	2804	2533
total	9055	9080	9035

Table 1.2: Partial number of words according to the text type as well as the total per language.

1.5 Thesis outline

The thesis will be organized as follows: Section 1.1 has given an overview of the topic and the motivations for the present research, while Section 1.2 has outlined Aims and objectives and Section 1.3 Hypotheses and research questions; Chapter 2 will provide a more detailed introduction to the field of Sentiment analysis and previous works related to the present research; Chapter 3 will provide an introduction to Systemic Functional Linguistics and the Appraisal Framework; in Chapter 4 the Appraisal Framework will be used as the basis for a manual analysis of the texts aimed at highlighting commonalities and diversities from different points of view.

Afterwards, Chapter 5 will describe the annotation scheme, Chapter 6 will show the results of the manual annotations in form of statistics and Chapter 7 will describe the automatic system. Finally, in Chapter 8 the results of the experiments run on the automatic system will be presented and discussed in comparison to the manual analysis, and Chapter 9 will serve as place for the final conclusions and the discussion of future works.

Chapter 2

Sentiment analysis and the use of dependency parsing

I will start this Chapter by clarifying the objects of Sentiment analysis, and then move to the description of how my work draws or is related to previous works in the field.

2.1 What are the "sentiments" under analysis?

In order to have in mind what are the "sentiments" under analysis, I will use the list of definitions provided by Pang & Lee (2008) since I find it particularly comprehensive:

- *Opinion* implies a conclusion thought out yet open to dispute ("Each expert seemed to have a different opinion").
- *View* suggests a subjective opinion ("Very assertive in stating his views").
- *Belief* implies often deliberate acceptance and intellectual assent ("A firm belief in her party's platform").
- *Conviction* applies to a party's firmly and seriously held belief ("The conviction that animal life is as sacred as human").
- *Persuasion* suggests a belief grounded in assurance (as by evidence) of its truth ("Was of the persuasion that everything changes").

- *Sentiment* suggests a settled opinion reflective of one’s feelings (“Her feminist sentiments are well-known”).

In Sentiment analysis, the above categories are simply called *opinions* and the way in which an opinion is analysed is by looking at its components. Consider the following examples taken by Liu (2012):

1. I bought a Canon g12 camera six months ago.
2. I simply love it.
3. The picture quality is amazing.
4. The battery life is also long.
5. However, my wife thinks it is too heavy for her.

In these, two key components can be identified:

> a target g and a sentiment s on the target, where g can be any aspect of the entity about which an opinion has been expressed, and s is a positive, negative, or neutral sentiment (or a numeric rating score expressing the strength/intensity of the sentiment such as 1-5 stars). Positive, negative, and neutral are called sentiment (or opinion) orientations (or polarities). For example, the target of the opinion in sentence 1 is *Canon g12*, and the target of the opinion in sentence 3 is *The picture quality* of Canon g12.

This is as far as I go with my research. I will explain the motivations of this choice in Chapter 5, when I describe the annotation scheme in detail. However, for the sake of completeness, I must mention that, apart from these two key components, Liu (2012) mentions three more:

- Opinion source (or holder), that in sentences (2), (3), and (4) is the author of the review, but in sentence (5) is the author’s wife.
- Time (in which the review has been written).

- Entity, i.e. the pair $e : (t, W)$ where t is a hierarchy of parts and W is a set of attributes of e. Each part or sub-part also has its own set of attributes. In practice, the target can often be decomposed and described in a structured manner with multiple levels, for example "picture quality of Canon G12" can be decomposed into an entity and an attribute of the entity and represented as a pair: Canon-G12, picture-quality.

The complete definition to which Liu (2012) arrives is thus that:

> An opinion is a quintuple, (e_i, a_{ij}, s_{ijkl}, h_k, t_l), where e_i is the name of an entity, a_{ij} is an aspect of e_i, s_{ijkl} is the sentiment of e_i, h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k. The sentiment s_{ijkl} is positive, negative, or neutral, or expressed with different strength/intensity levels [...]. When an opinion is on the entity itself as a whole [...], e_i and a_{ij} together represent the opinion target.

2.2 How are the objects of sentiment analysis usually identified?

Before looking at how the objects of Sentiment Analysis are usually identified, it must be specified that *Sentiment analysis* (or *Opinion mining*) nowadays covers fields with similar names such as *Opinion extraction, Sentiment mining, Subjectivity analysis, Affect analysis and Review mining*.

Due to the amount of work in these fields, convenient categorizations are based on **the unit of text** they focus on, and **the technique used** to conduct the analysis. I will explore them in their relation to the aim of this Section.

As for the first, this work falls under the category of the *fine-grained* ones, i.e. those aimed at classifying sentiment below the level of the sentence, as opposed to *coarse-grained*. One of the most important reasons for looking at appraisal-loaded expressions shorter than a sentence is that each sentence does not necessarily express coherent opinions on a single entity (Liu, 2012). For example, in "Although the service is not that great, I still love this restaurant", the opinion is positive about *the restaurant*, but negative about its *service*. Works in *fine-grained* sentiment analysis that specifically aim at identifying aspects of the entity (i.e. the service of the restaurant) belong to the sub-category of *aspect-based sentiment analysis*. Advanced works have also attempted to address the subtler task mentioned by Mejova (2009) of the identification of aspects even when they are non-explicit such as "volume" in "Camera is too large", as opposed to when they are explicit such as "battery life" in "Battery life is too short". However, the *aspect-based analysis* is not of concern for this work, since it aims at connecting explicit targets to their evaluations.

By having short appraisal expressions as goal, the present work also avoids one of the issues in coarse-grained SA represented by the wrong classification of entire sentences based on words that do not actually express any sentiment, like in the case of "Can you tell me which Sony camera is good?" and "If I can find a good camera in the shop, I will buy it" (Liu, 2012).

Conversely, it focuses on a more accurate classification of these groups, by paying attention to negation, which has been widely recognised as one of the issue to take into account (Benamara *et al.*, 2012; Choi & Cardie, 2008; Jia *et al.*, 2009; Liu, 2010; Wiegand *et al.*, 2010; Wilson *et al.*, 2005). In fact, apart from being local (e.g. "not good"), negation can also involve long-distance dependencies (e.g. "does not look very good") or the subject (e.g. "no one thinks that it is good") (Wilson, 2008), and of course does not include expressions such as *not* in "not only...but also".

Another issue that is taken into account is reversed polarities. For example, in the sentence "The medicine kills cancer cells", although the phrase "cancer cells" has negative sentiment, the word *kills* reverses it by making the group "kills cancer cells" positive (Nakagawa *et al.*, 2010).

In order to do so, my research gives importance to the sentence structure, which connects us to the second categorization of works in SA, i.e. the technique used.

As mentioned in Chapter 1, this work attempts at overcoming the limitations linked to the *bag-of-words* approach, in which words are considered as features (called *-grams*) and are represented in an unordered collection often along with their frequencies. When the TF-IDF (Term Frequency - Inverse Document Frequency) weighting scheme is also applied, the most frequent terms are considered more informative. In both cases, the sentence (or the document) is judged as positive or negative based on whether it has a preponderance of positive or negative words.

One of the ground-breaking works in this category has been Liu & Seneff (2009)'s, who proposed to extract "adverb-adjective noun phrases" (e.g. "very nice car") based on the parsing relations. They assigned sentiment scores based through a heuristic method that computes the contribution of adjectives, adverbs and negations to the sentiment degree, based on the ratings of the reviews where these words occurred.

Afterwards, Qu *et al.* (2010) introduced a "bag-of-opinions" representation of documents to capture the strength of n-grams with opinions, which is different from the traditional bag-of-words representation in so far as it addresses some of the issues mentioned earlier: each opinion is a triple consisting of a sentiment word, a modifier and a negator. For example, in "not very good," *good* is the sentiment word, *very* is the modifier, and *not* is the negator. Also in this case the sentiment score of each opinion is learnt from an opinion lexicon and the review ratings.

An *opinion lexicon* (or *sentiment dictionary*) is a collection of words with their out-of-context polarity, e.g. *good, wonderful* and *amazing* are positive sentiment words, whereas *bad, poor and terrible* are negative sentiment words. Most sentiment words are adjectives and adverbs, but nouns and verbs can also be used to express sentiments.

My research has in common to these works in SA the fact that it relies on sentiment dictionaries to identify and classify the sentiment words which, according to Liu (2012), are not surprisingly the most important indicators of sentiments. Among the multitude of works that have made use of sentiment lexicons, the first have been *WordNet-Affect* (Strapparava *et al.*, 2004) and *SentiWordnet* (Esuli & Sebastiani, 2006), both expansions of WordNet (Fellbaum, 1998) with polarity (positive-negative), and the latter also with objectivity (subjective-objective) labels. SentiWordnet has been used afterwards by (Denecke, 2008) in a multilingual context to classify documents, by Ohana & Tierney (2009) to classify film reviews, by Dang *et al.* (2010) for product reviews and by Taboada *et al.* (2011) for sentence-level analysis.

However, my research makes use of the *NRC Word-Emotion Association Lexicon* (Mohammad, 2011), used by others in the field (Kennedy *et al.*, 2012; Perrie *et al.*, 2013). Its annotations in English were manually undertaken through *Amazon's Mechanical Turk*, and the *Roget Thesaurus*[1]. The lexicon was chosen among others such as AFFIN because it has entries for approximately 24200 word–sense pairs, corresponding to

[1] `http://www.gutenberg.org/ebooks/10681`

14200 word types. In addition, it is not specific for domain such as the *Lexicoder Sentiment Dictionary* for the political domain (Young & Soroka, 2012), the *Opinion Lexicon* for customers reviews (Hu & Liu, 2004), AFINN for microblogs (Nielsen, 2011) and ANEW with emotion words (Bradley & Lang, 1999) or SentiWordnet (Esuli & Sebastiani, 2006).

As in the case of few of the works mentioned above, I am using syntactic dependencies for the identification of the appraisal expressions, and the sentiment dictionaries to classify them. Dependency-based features (explained in detail in Section 2.5) have been used quite extensively in the literature (Argamon *et al.*, 2007; Bloom & Argamon, 2009; Bloom *et al.*, 2007a; Nakagawa *et al.*, 2010; Nasukawa & Yi, 2003; Taboada & Grieve, 2004; Wilson *et al.*, 2004; Zhao *et al.*, 2011), and I will describe in detail in Sections 2.5 and 5.1 how my work differ from the others.

Finally, while my automatic system is based on rules to extract those features, an alternative adopted more and more is supervised machine learning, which consists of 3 phases: feature extraction, training and testing. For the training phase, an annotated corpus has to be given to the system in order to examine the features associated to each word/phrase/sentence (depending on the type of the annotation) and learn how to classify them. The system builds a model that is tested on the testing set, whose data are new but present the same characteristics of the training set. Among the first to use it in Sentiment Analysis on the general domain there are Pang *et al.* (2002), Mullen & Collier (2004) and Pang & Lee (2008). The choice of having my system based on rules rather than on machine learning was motivated by the relatively small size of the dataset that would have made the learning process extremely challenging. In addition, I believed that with a rule-based approach I could have had more insights over the influence of each feature on the overall results.

2.3 Previous work on fine-grained sentiment analysis in English

The most studied texts in Sentiment Analysis have been product reviews (Liu, 2010), (Denecke, 2008; Miao *et al.*, 2008), movie reviews (Hu & Liu, 2004; Mullen & Collier, 2004; Popescu & Etzioni, 2005) and book reviews (Zagibalov *et al.*, 2010), mainly

Rank	Song lyrics	Song titles	Blogs	SOTU
1	Love (7.37%)	Love (7.39%)	Good (4.89%)	People (5.49%)
2	Time (4.18%)	Time (4.19%)	Time (4.72%)	Time (4.09%)
3	Baby (2.75%)	Baby (2.75%)	People (3.94%)	Present (3.45%)
4	Life (2.59%)	Life (2.60%)	Love (3.31%)	World (3.10%)
5	Heart (2.14%)	Heart (2.15%)	Life (3.13%)	War (2.98%)

SOTU State of the Union addresses

Table 2.1: Top five most frequently occurring ANEW words in five corpora

because they already have a clearly specified topic and a star rating system. For the sake of completeness, it is also worth mentioning that other researched text types have been novels, fairy tales, e-mails, letters, suicide notes (Mohammad, 2012), tweets (starting from Go *et al.* (2009)), text messaging (Neviarouskaya *et al.*, 2007) and quotations by the European Commission[1].

In the following Sections, I will give details only of the works related to the text types under analysis in my research (i.e. political speeches, news and TED talks) for English, while for Italian and Russian I will provide a broader overview because resources and studies are much fewer.

2.3.1 Political speeches

Among those works that have focused on political speeches, Dodds & Danforth (2010) quantified happiness levels for a diverse set of texts including the *State of the Union addresses*. Such classification has been manually done on some words previously identified as bearing meaningful emotional content, basing on the *Affective Norms for English Words* (ANEW) study (Bradley & Lang, 1999). In Table 2.1 their main findings on all the four corpora are shown.

According to their analysis, while blogs evince a more social aspect with *people* and *life* in the top five, the nature of State of the Union addresses is reflected in the disproportionate appearance of *world* and *war*. They also analysed the percentage in the use of words during the years. In Figure 2.1 they show three graphs corresponding to events such as September 11, 2006; Valentine's Day, 2008; and US Presidential Election

[1] `https://ec.europa.eu/jrc/en/language-technologies`

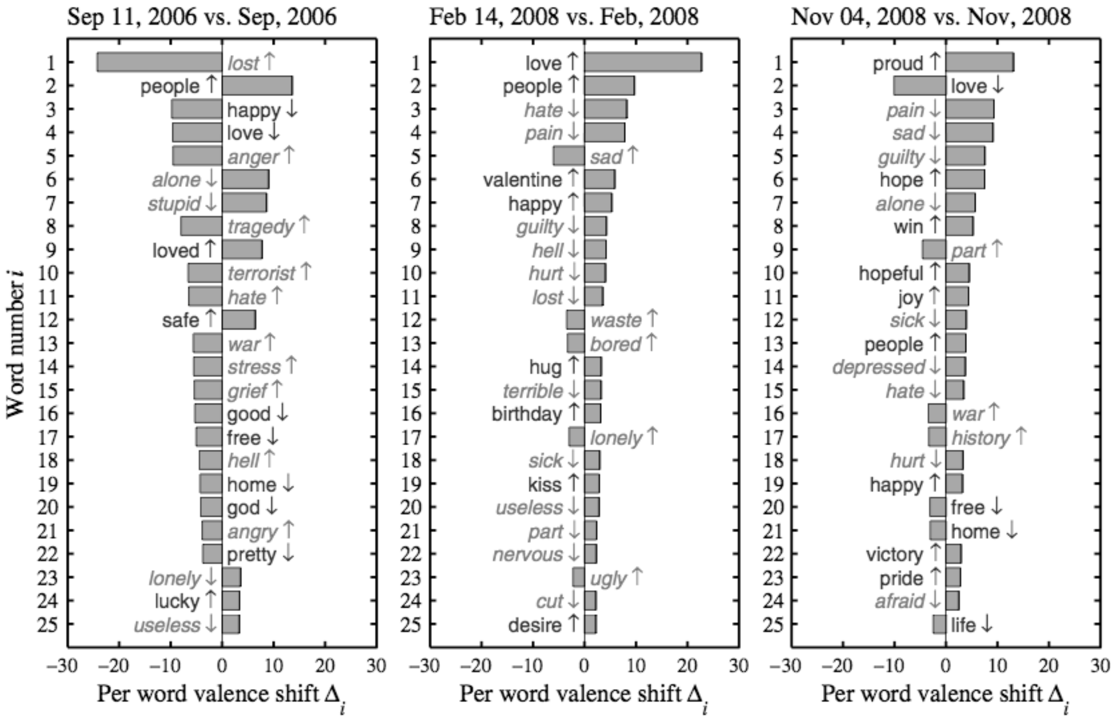

Figure 2.1: Graphs in Dodds & Danforth (2010) showing the use of words for 11 September 2006, Valentine's Day 2008 and US Presidential Election Day on 4 November 2008

Day, November 4, 2008. The first panel in Figure 2.1 shows that the negative words of the fifth anniversary of the 9/11 attacks are *lost, anger, hate* and *tragedy*. The impact of these words is augmented by a decrease in frequency of *love* and *happy*, while in the third panel we can see that the strongest word for the 2008 US Election is *proud*.

Finally, by using the State of the Union addresses for the United States as a starting point for assessing the emotional temperature of the United States over its 220 year history, they also calculated the US presidents with the highest scores of happiness value (see Figure 2.2): Kennedy (with 6.41), Eisenhower (with 6.38), and Reagan (with 6.38). For many presidents the authors also highlighted important words in their speeches.

This is somehow similar to the analysis of voters' opinions as reaction to political debates and campaigns in order to find out the importance of a specific politician during that period and predict the following election results (Adamic & Glance, 2005; Kato *et al.*, 2008; Pang & Lee, 2008).

Other recent research has compared the styles of political leaders from annotating

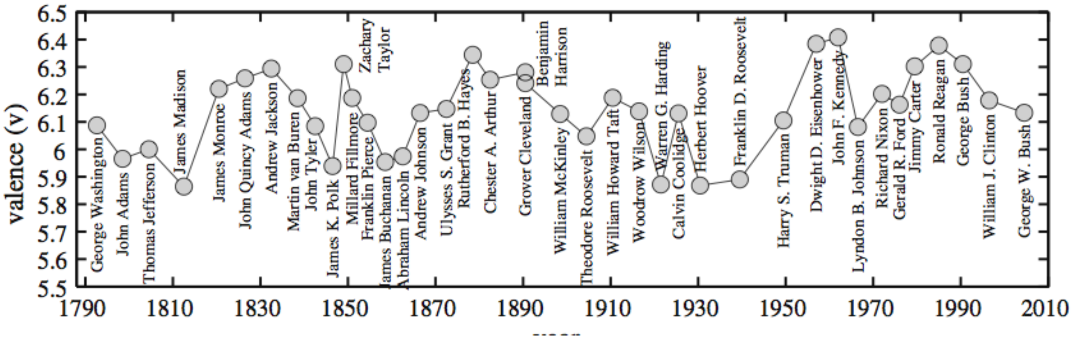

Figure 2.2: Happiness score of State of the Union addresses

political speeches. For example, Dyson (2008) used the analysis of high vs. low complexity words to study British Prime Ministers' responses to foreign policy questions in the House of Commons from 1945-2008. Yu *et al.* (2008) used supervised learning algorithms to classify U.S. Congressional speeches based on the political party affiliation of the speaker (i.e. Democrat or Republican). Muhlberger & Zhang (2008) provided a statistical analysis to speeches of John McCain and Barack Obama in a part of the 2008 presidential race talks in an effort to extract the candidates' ideological mental models.

2.3.2 News

As for news, Sentiment analysis in this field is quite extensive: Bloom *et al.* (2007b) proposed a system which uses lexical shallow parsing to find adjectival "appraisal groups" in sentences, which convey a positive or negative appraisal of an item, and used a simple heuristic to detect opinion holders.

Godbole *et al.* (2007) implemented a lexicon-based system for news and blogs analysis through a method to expand candidate seed lists opinion words through WordNet.

Bautin *et al.* (2008) performed sentiment analysis on automatic translations of news from nine languages into English, by expanding their seed list of opinion words through WordNet as in the previous case.

Balahur *et al.* (2010) performed sentiment analysis at document level by focusing on three subtasks that they found to be specific to the news text type: definition of the target; separation of the good and bad news content from the good and bad sentiment expressed on the target; and analysis of clearly marked opinion that is expressed explicitly. As lexicons they used *WordNet Affect* (Strapparava *et al.*, 2004), *SentiWordNet* (Esuli & Sebastiani, 2006), *MicroWNOp* (Cerini *et al.*, 2007) and the in-house built *JRC Tonality*.

Moreo *et al.* (2012) followed Wilson *et al.* (2004) in making use of a lexicon with strength of the opinions (i.e. very negative and very positive) and three classes: (i) a set of hierarchically-related Objects (O); (ii), a set of object Features (F); (iii) a set of Valuations (V). They used both specialised lexicons on the most recurrent discussion topics in news such as sports, politics, economics, current events and entertainment, and a generic lexicon. In addition, before performing sentiment analysis, they also applied two heuristics: disambiguation analysis and frequency analysis. The importance of disambiguation has also been put forward by Bloom & Argamon (2009) and Ebert & Schütze (2014).

Finally, Curran & Koprinska (2013) have focused on the sentiment analysis of quotes in news.

2.3.3 TED Talks

Research on TED talks has only been carried out in the past two years. To my knowledge, the only two works in sentiment analysis have been conducted by Pappas & Popescu-Belis (2013) and Pappas *et al.* (2014). In the first one, the authors focused on the corpus of TED comments crawled by them rather than the talks. Interestingly, their approach is similar to mine since they used a rule-based classifier cross-checked against human annotations. The lexicon they used is the MPQA polarity lexicon, which has been applied to entire sentences. In the second one, their model assigns weights to each of the sentences or paragraphs both of TED comments and talks to uncover their contribution to the aspect ratings and then it predicts the aspect ratings.

Not related to sentiment on written talks is the work of Sugimoto *et al.* (2013) on the relationship between TED presenters and videos to study the impact of the videos.

2.4 Previous work in Russian and Italian

As far as Russian is concerned, few works have been done and mostly in the context of the tasks “Dialog” and “Romip (Information Retrieval Seminar)” mainly on product reviews and news. Before these, only Ermakov (2009) (cited by Chetviorkin &

Loukachevitch (2013)) had proposed a system to extract opinions about different aspects of cars from a Russian blog. In 2011 Chetviorkin & Loukachevitch (2011) automatically extracted a lexicon of 3200 opinion words in Russian by using news and reviews and Pazelskaya & Solov'ev (2011) manually built a lexicon consisting of 15000 words They built a rule-based system for detecting sentiment in Russian texts in the field of mass media.

In 2012 Solov'ev *et al.* (2012) assigned negative and positive with intermediate values (weak, medium and strong) by using TF/DF measure. Then a filter function was used to get an emotive summary of a document, while Kan (2012) proposed a rule-based approach.

Not in the context of shared tasks, Zagibalov *et al.* (2010) analysed reviews related to the same books in English and in Russian, and Steinberger *et al.* (2012) created a multilingual parallel news corpus projecting sentiment annotation from one language to many others. Not having access to POS-taggers and parsers for all the languages, they used sentiment dictionaries: they added up positive and negative sentiment scores in six-word windows around the entities, distinguishing two positive and two negative levels of sentiment words. Enhancers and diminishers added or removed 1 point, negation inverted the value, except for negated high positive ("not very good" is not equivalent to "very bad"). The sentiment dictionaries were created by using a triangulation method, i.e. sentiment word lists in English and Spanish were translated into a third language. The introduction of errors through word sense ambiguity was limited by taking the intersection of both target language word lists. According to their evaluation, approximately 90% of these intersection words were correct. However, the results show that their classification system did achieve a good accuracy in none of the languages.

As far as Italian is concerned, sentiment analysis on social media has always been the most studied area: Bosca *et al.* (2012) have presented *Linguagrid*, a framework that deals with the identification of textual snippets containing opinions and sentiments, then produces a list of the sentiment relations detected in the text along with the target and the polarity of the opinion detected.

Bosco *et al.* (2013b) have created a human annotated corpus of Italian tweets including labels for sentiment polarity and irony. In 2014, in the context of "Evalita 2014" a SENTIment POLarity Classification Task on Italian tweets (SENTIPOLC) was organized (Basile *et al.*, 2014), and there are also industrial projects such as *Blogmeter* by

Bolioli *et al.* (2013) about Italian top brands and *Voices from the Blog* by Ceron *et al.* (2014). Marrazzo (2014) has analysed the Italian social media with particular focus on the opinions on political issues and politicians.

Maisto & Pelosi (2014) have worked on customer reviews, by building a lexicon to recognize different features and the opinions expressed on them. Sorgente *et al.* (2014) have worked on different aspects of movie reviews such as acting, story, soundtrack by manually annotating a corpus, while Casoto *et al.* (2008) on the same topic, but using machine learning.

Elia *et al.* (2015) have manually annotated adjectives in the Italian dictionary in *Nooj* (Silberztein, 2003) with their polarity and intensity, and expanded their polarity lexicon automatically to adverbs and nouns which, especially in the case of nouns, was a risky move that I preferred not to make in my research, but that in their case was backed-up by some reasoning on the exceptions. However, in terms of categories that would change the prior polarity of words according to the context, we have in common *Contextual Valence Shifters* (in my case reversals), negation (~polarity), intensification, modality and comparison (~force), and opinionated idioms (~multi-word expressions). Their final output is different from mine because it is a network of local grammars ready to be used for the sentiment annotation.

2.5 How can dependency relations help?

In Section 1.3 I have explained why relying on the information related to the sentence structure is unavoidable when dealing with the study of the language. In particular, I have also described why I believe that dependency-parsing relations more than constituency-based ones have a direct link to the appraisal-group approach, and thus could constitute an important factor in the increase of the accuracy.

We have seen that, for example in the sentence “Let both sides unite to heed in all corners of the earth the command of Isaiah to "undo the heavy burdens...and to let the oppressed go free."” taken from Kennedy’s Inaugural speech, the appraisal groups according to the annotation scheme *SentiML* would be the same that the dependency parser would connect, along with grammatical connections very similar to those that one has in mind when linking targets and modifiers in those appraisal groups.

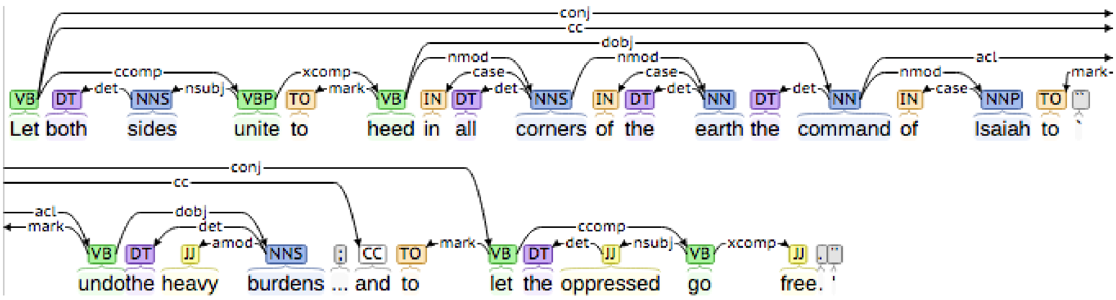

Figure 2.3: Dependency structure of the sentence “Let both sides unite to heed in all corners of the earth the command of Isaiah to "undo the heavy burdens...and to let the oppressed go free."”

In this Section I will focus more on the insights that dependency-parsing relations could bring in terms of linguistic features across the three languages.

First of all, the specified grammatical combinations for an appraisal group to be manually and automatically annotated are expected to be matched to a good extent by the dependency parsers. These are:

- A noun with an adjective.
- A pronoun with an adjective.
- A verb with an adverb.
- A noun with a verb.
- A pronoun with a verb.

In the case of English, both the lack of inflections in the language and the advance in the available resources might play a lead factor in the achievement of good performances.

Role	Attribute	Possible values
Target	Type	Person, Thing, Place, Other
	Orientation	Neutral, Positive, Negative, Ambiguous
Modifier	Orientation	Neutral, Positive, Negative, Ambiguous
	Attitude	Affect, Judgement, Appreciation
	Force	Normal, High, Low, Reverse
	Polarity	Unmarked, Marked
Appraisal group	Orientation	Neutral, Positive, Negative, Ambiguous

Table 2.2: Attributes for target, modifier and appraisal group along with all possible values and default ones underlined.

Second, in terms of linguistic features useful for SA that should be captured by the attributes in the annotation scheme (see Table 2.2 for an outline), parsing should allow to identify some among the most challenging issues, namely reversals (thanks to the attribute *force*) and negation (thanks to the attribute *polarity*). They both have an influence on the orientation taken from the sentiment dictionaries:

- Force is the intensity of the appraisal. Force is largely expressed via modifiers such as *very* (increased force), or *slightly* (decreased force), but may also be expressed lexically, for example *greatest* vs. *great* vs. *good*. When reverse, for example "avoid war", the orientation changes, i.e. it becomes the opposite of *war*.

- Polarity is marked if the appraisal is scoped in a polarity marker (such as *not*), or unmarked otherwise. When affected by negation, for example the expression "not good", the orientation changes, i.e. it becomes the opposite of *good*.

As I will explain in Chapter 7, the value 'reverse' for force and 'marked' for polarity are activated by specific rules used by the automatic system.

In addition, even the accuracy for other attributes such as the *target type* and the *attitude* according to the Appraisal Framework is expected to be strongly dependent on the accuracy of the parsing relations. I will explain their connections in Chapter 7, and test them in Chapter 8.

The following sections will be dedicated to explore the works in the literature that have made use of dependency parsing for Sentiment Analysis for all the three languages. Those on which the annotation scheme is based will also find space in Chapter 5.

2.5.1 Previous work with dependency parsing in English

Among those works that have used dependency relations, Wilson *et al.* (2005)'s is relevant to mine since they have used both modification features (dependency features) and structure features (dependency-tree-based patterns). Modification features involve relationships with the word immediately before or after, and four categories are considered: nouns, adjectives, adverbs, intensifiers. Conversely, structure features are determined by starting with the word instance and climbing up the dependency tree towards the root, looking for particular relationships, words or patterns.

At the same time, Bloom *et al.* (2007a) and Popescu & Etzioni (2005) first proposed the concept of "appraisal expression". In this way the real polarity can be identified, in opposition to those not carrying any sentiment (e.g. "I am well" vs. "Well, I am going home"). The authors used the dependency parser to construct a linkage lexicon by hand. This resulted in a better accuracy comparing to surface patterns, but in a low coverage because the manually-compiled patterns were only ten.

Ng *et al.* (2006) introduced new grammatical combinations from which I drew inspiration, namely subject-verb (SV) and verb-object (VO) relations, in addition to the adjective-noun (AN) relation used by Kushal & Lawrence (2003) and Popescu & Etzioni (2005). They also used a dependency parser, *MINIPAR* by Lin (1998), to extract these relations. Another important finding was that dependency relations were actually not improving accuracy because they did not provide additional useful knowledge when applied to bigrams and trigrams. Although this was somewhat surprising, their final hypothesis on these low performances was that, by stemming, MINIPAR returned dependency relations in which all the verb inflections were removed.

In a later work by Zhao *et al.* (2011), the advantage is that the linkage specifications manually annotated in the previous works by Bloom *et al.* (2007a) and Popescu & Etzioni (2005) were acquired automatically. These are paths connecting two words in a parse tree that also describe the relation between them. They used the polarity lexicon *HowNet* by Dong & Dong (2003) to identify the polarity words, and the constituency *Charniak parser* by Charniak (2000) to connect the polarity words to nouns or pronouns that were candidate targets in each sentence. Afterwards, they generalised the syntactic paths with identical constituents and similar POS-tags and chose the most frequent ones (see example in Figure 2.4).

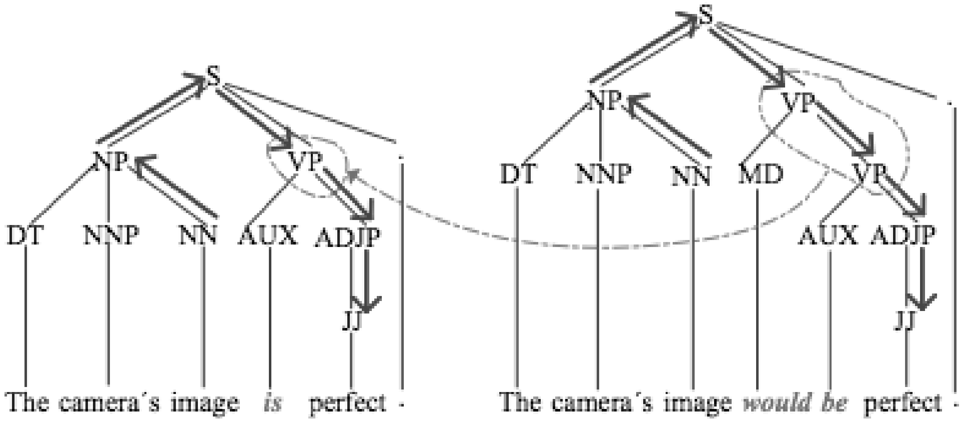

Figure 2.4: Syntactic paths containing a sequence of identical constituents between the polarity word *perfect* and its target *image*
.

Another work involving a dependency tree-based method for Japanese and English has been done by Nakagawa *et al.* (2010). Although not relevant to my work in terms of technique, either because a supervised machine-learning algorithm (*Conditional Random Fields*) has been used instead of a rule-based approach, and in terms of focus because it falls under the sentence-level sentiment analysis task (while mine in fine-grained sentiment analysis), this work is interesting because of the way in which the issue of the reversals and negation have been addressed: the sentiment polarity of each dependency sub-tree is represented by a hidden variable, and the polarity of the whole sentence is calculated in consideration of interactions between the hidden variables. For example, in the sentence in Figure 2.5 “It prevents cancer and heart disease”, *cancer* and *heart disease* individually carry a negative polarity. However, the polarities are reversed by modifying the word *prevents*, and the dependency subtree “prevents cancer and heart disease” has positive polarity.

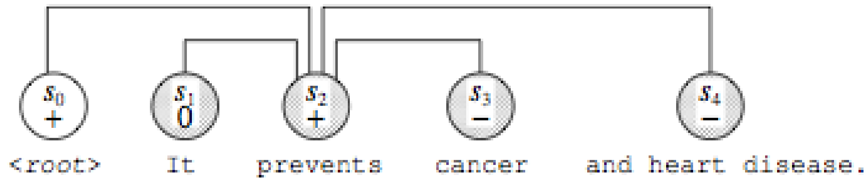

Figure 2.5: Probabilistic model based on dependency tree

The prior polarity of the root is the innate sentiment polarity of a word contained in the phrase, which can be obtained from sentiment polarity dictionaries. When a phrase contains multiple words in the dictionaries, the registered polarity of the last (nearest to the end of the sentence) word is used, which in my opinion was limiting. However, building polarity reversing word dictionaries (called *context valence shifters*) containing such words as *decrease* and *vanish* that reverse sentiment polarity was a clever idea that has been used also by others (e.g. Argamon *et al.*, 2007; Choi & Cardie, 2009; Ikeda *et al.*, 2008; Kennedy & Inkpen, 2006; Moilanen & Pulman, 2007; Ohana & Tierney, 2009; Polanyi & Zaenen, 2006; Shaikh *et al.*, 2007; Shanahan *et al.*, 2006; Wang & Manning, 2012).

The English polarity reversing word dictionary was constructed from the *General Inquirer dictionary* by Stone (1966). Then word dictionaries were categorised into two categories: "function-word negators" such as *not* and "content-word negators" such as *eliminate*. The polarity reversal of a phrase handles only the "content-word negators".

Similarly, Jia *et al.* (2009) used dependency parsing to identify the scope of each negated item in English, while Councill *et al.* (2010) used a conditional random field model informed by a dependency parser in English.

2.5.2 Previous work with dependency parsing in Russian and Italian

As for Russian, although dependency parsing for this language was theorised since 1988 in Melvcuk (1988), works in NLP have started after the creation of the first Russian treebank *SYNTAGRUS* by Boguslavsky *et al.* (2000).

However, the first works in sentiment analysis have been in the contest of "ROMIP11", a specific evaluation campaign for Russian, in which in particular Pak & Paroubek

(2012) presented a system with features based on n-grams, POS-tags and dependency parsing, followed by terms weighting for the optimisation.

As for Italian, the use of dependency parsers is wider, also thanks to the shared task "CoNLL" and "Evalita" since 2006. Among the parsers presented, there are *Desr* (Attardi *et al.*, 2007, 2009; Bosco & Mazzei, 2011; Bosco *et al.*, 2008) and *Maltparser* (Bosco & Mazzei, 2011; Bosco *et al.*, 2008; Lavelli *et al.*, 2009), while most recently Bosco *et al.* (2013a) have worked to build an Italian Stanford dependency treebank.

It is thus obvious that, despite the advantages brought by the dependency parsers, works in Italian and Russian are still very few. Chapter 3 will be now dedicated to additional insights that Systemic Functional Linguistics could bring to Sentiment Analysis.

Chapter 3

Sentiment Analysis and Systemic Functional Linguistics

While in Chapter 2 I have provided an overview of the previous works in Sentiment Analysis in English, Italian and Russian by giving special attention to the contributions of dependency-parsing relations to the field, I will now focus on the contributions of Systemic Functional Linguistics (and in particular the Appraisal Framework), in accordance to the hypotheses stated in Chapter 1.

3.1 Systemic Functional Linguistics and the Appraisal Framework

Systemic Functional Linguistics (SFL) was developed in the 1960s by the British linguist M.A.K. Halliday, who had been influenced by the work of the Prague School and British linguist J.R. Firth, and it is defined as the study of the relationship between language and its functions in social settings (Halliday, 1994). For Halliday a central theoretical principle is that any act of communication involves choices that are mapped using "system networks". Such choices:

- Depend on the aspects of the context in which the language is being used.
- Can be charted on different levels (or strata) of language:

1. Text, which includes in order graphology/phonology, lexico-grammar level and discourse-semantic level.
2. Context, which includes the context of situation represented by register (field, tenor and mode) and the context of culture, which includes genre and ideology.

Systemic linguistics is also "functional" in so far as, at the discourse-semantic level, it looks at the three *macro-functions* that "the language has evolved to serve in the life of social man" (Halliday, 1973):

1. *Ideational*. It is present in all the uses in which the adult typically engages, as it represents the potential for expressing a content in terms of the speaker's experience and that of the community.

2. *Interpersonal*. It embodies all uses of language to express social and personal relation, including all forms of the speaker's intrusion into the speech situation and the speech act.

3. *Textual*. It fills the requirement that language should be operationally relevant in real contexts.

The power of SFL is that we can conduct a bottom-up analysis of the text, by linking the lowest level (lexico-grammar) to the medium (discourse-semantic) and the highest (register) (Halliday & Webster, 2009):

1. Field, i.e. the nature of the social action, can be linked to the ideational meta-function.

2. Tenor, i.e. the nature of the social relationship amongst those involved in the action, can be linked to the interpersonal meta-function.

3. Mode, i.e. the mode of contact for the actors in the discourse event, can be linked to the textual meta-function.

Evaluation is related to tenor, and it is defined by Hunston & Thompson (2000) (p. 5) as "the broad cover term for the expression of the speaker or writer's attitude or stance towards, viewpoint on, feelings about the entities or propositions that he or she is talking about".

The function of "constructing and maintaining relations between the speaker or the writer and the hearer or reader" is one of the three performed by the evaluation (also simultaneously) according to Hunston & Thompson (2000) (p. 6), and one of the reasons why evaluation is a topic worthy of study. The other two functions are to reflect the systems of values of the speaker and their community, and to organize the discourse.

This had been anticipated in Section 1.4, when the analysis of tenor for the text types under analysis was done. In this context I will add a few more considerations related to how the relationship between the writer and reader is built in the case in which the assumption of shared attitudes, values and reactions is valid:

- Political speeches. The evaluation is here used to organize the discourse (Hunston & Thompson, 2000) (pp. 10-13) (e.g. through an introduction, a description of the current situation with reminders to the past, followed by the goals for the future), but especially to express opinions whose acceptance by the hearers is assumed, and more in general, to transmit an ideology as sets of values (Hunston & Thompson, 2000) (p. 8).
- TED Talks. The evaluations are expressed by the speakers in order to transmit an ideology, and in a way that is difficult for the hearers not to accept or to question their validity (Hunston & Thompson, 2000) (pp. 8-9).
- News. According to Hunston & Thompson (2000) (pp. 9-10), in newspapers a subtle level of *manipulation* is retrievable through a not obtrusively placement of the evaluation.

Given the important role that evaluation carries through these three functions, it is not surprising that many studies have focused on it throughout the years, naming it in different ways according to the chosen parameters such as *subjectivity, evidentiality, stance, affective meaning, connotative meaning, appraisal* (Hunston & Sinclair, 2000; Munday, 2012).

Appraisal is the one I am taking into account in this research, and it is associated to the *Appraisal Framework* (AF), a framework concerned with the language of evaluation, attitude and emotion and developed starting from the the model of tenor in SFL by

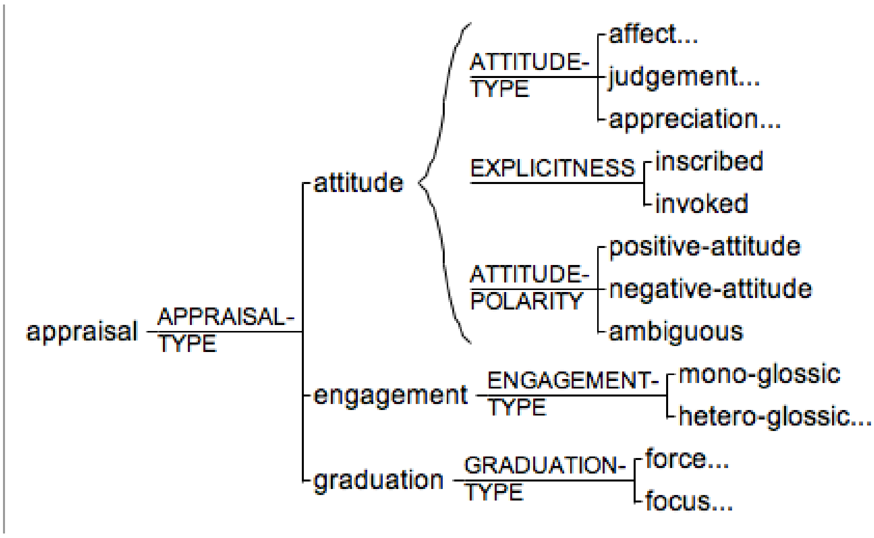

Figure 3.1: Excerpt of the Appraisal Framework showing the sub-systems *attitude*, *engagement* and *graduation*.

Martin & White (2005). The AF consists of three sub-systems that operate in parallel: *attitude*, *engagement* and *graduation* (see Figure 3.1). Since they are among the foundations of my research, I will now briefly describe them, and then explain what their contributions to my project are.

3.1.1 Attitude: emotion, ethics and aesthetics

The Attitude sub-system looks at how one expresses private state, i.e. one's emotion and opinions and describes three areas of private state:

1. *Affect*, which deals with personal emotions and opinions (e.g. *happy*, *sad*).
2. *Judgement*, which concerns the author's attitude towards people's behaviour (e.g. *heroic*, *craven*).
3. *Appreciation*, which considers the evaluation of things (e.g. *striking*, *inherent*).

An attitude is further qualified by its explicitness and polarity.

3.1.2 Engagement: appraisal of appraisals

Engagement considers the positioning of oneself with respect to the opinions of others. It deals with the linguistic constructions by which authors construe their point of view and the resources used to adopt stances towards the opinions of other people. Engagement can be either mono-glossic (one may anticipate the responses of an intended audience and include counter-responses in the original text) or hetero-glossic (an author will acknowledge and agree or disagree with the stances of others who have previously appraised a subject). Examples of Engagement are epistemic modal expressions (e.g. "He might have finished his studies by now"), evidential expressions (e.g. "Apparently, he has recovered from his illness") or denials (e.g. "This hotel is not near the sea as you said").

3.1.3 Graduation: strength of evaluations

Graduation investigates how the use of language functions to amplify or diminish the attitude and engagement conveyed by a text. Graduation is a general property of both attitude and engagement. In attitude it enables authors to convey greater or lesser degrees of positivity or negativity, while graduation of engagements scales authors' conviction in their utterance. Graduation is divided into two subsystems. *Force* alters appraisal propositions in terms of its intensity, quantity or temporality, or by means of spatial metaphor. It is sub-divided into:

- Intensification, which can apply to a quality (e.g. "slightly sad") or to a process (e.g. "greatly disturbed me").
- Quantification, which covers quantities (e.g. *few, many*), proximity (*recent, distant*) and distribution (*fast, broad*).

Focus considers the resolution of semantic categories, i.e. how binary relationships can be turned into scalar ones (e.g. *real, genuine(ly), effective(ly), sort of*).

3.2 Affect, Judgement and Appreciation

The definitions of Affect, Judgement and Appreciation will be now given in detail, supported either by my examples or provided in Martin & White (2005). When not taken from the annotated corpus, my examples are additional data belonging to the same text types although not annotated.

While the principle for the annotations throughout the project has been the writer or speaker's view, subjectivity could not be avoided either in what is considered "evaluative" and its labels, especially those related to the positive-negative/good-bad parameter (Hunston & Sinclair, 2000) (pp. 22-26) and the AF (see below instances and provoked judgement, affect and appreciation).

Another note has to be made on the categories of the AF that I included. As will be clear from the description of the annotation scheme in Chapter 5, apart from 'affect' judgement' and 'appreciation' as values of the attribute *attitude*, the subsystem *engagement* has been considered in the attribute *force*, whose values have been annotated as 'high', 'low' or 'reverse' according to the modal verbs and the adverbs part of the evaluative expression.

The decision of using only the broad categories 'affect', 'judgement' and 'appreciation', which is quite limiting especially considering the richness of the framework, is based on my initial hypothesis that the accurate automatic classification of these across the three languages would have represented already a challenging task. In terms of the classification, I will explain in Chapter 7 that, like in the case of Taboada & Grieve (2004), it is based on a match between the opinion and the category, i.e. opinions on a thing (appreciation), on a person (judgement) and on one's self (affect). However, in Section 9.3 dedicated to "Future works" I will also explain how I believe that this match could be improved by using phrase patterns rather than the categories thing, person and one's self, or that more sub-categories from the Framework could definitely become the new goal of the automatic classification.

3.2.1 Affect: definition and what to bear in mind during the sentiment annotation

Under the category of *Affect*, expressions indicating personal emotions and opinions fall:

- Verbs of emotion such as *to love/to hate, to frighten/to reassure, to interest/to bore, to enrage/to placate* (e.g. "Your offer pleases me", "I hate chocolate").

- Adverbs such as *happily/sadly* (e.g. "Sadly the government has decided to abandon its commitment to the comprehensive school system").
- Adjectives of emotion such as *happy/sad, worried/confident, angry/pleased, keen/ uninterested* (e.g. "I'm sad you've decided to do that", "I'm happy she's joining the group", "She's proud of her achievements", "He's frightened of spiders").

There is also an important differentiation between **authorial** and **non-authorial sentences**.

Authorial instances involve the writer/speaker, indicating how they have responded emotionally to the person, thing, happening or situation being evaluated, by strongly foregrounding his/her subjective presence in the communicative process. For the evaluation to carry any rhetorical weight, the reader must see this personalised response as in some way relevant, significant, valid, justified or at least understandable. Thus, by the use of such Affect, the writer bids to establish an interpersonal bond with the reader to the extent that the reader agrees with, understands or at least sympathises with that emotional reaction. This functionality can be illustrated by the following extract from a TED talk entitled "A life lesson from a volunteer firefighter" in which the author, Mark Bezos, describes his own experiences as head of development for a non-profit called *Robin Hood* and firefighter. Affect words are underlined.

> In both my vocation at Robin Hood and my avocation as a volunteer firefighter, I am witness to acts of <u>generosity</u> and <u>kindness</u> on a monumental scale, but I'm also witness to acts of <u>grace</u> and <u>courage</u> on an individual basis. And you know what I've learned? They all <u>matter</u>.

By appraising events in such affectual terms, the speaker invites his audience to share that emotional response, or at least to see that response as appropriate and well motivated. When that invitation is accepted, solidarity between the speaker and the listener is enhanced. Once such an empathetic connection has been established, there is the possibility that the listener will be more open to the broader ideological aspects of the speaker's position:

> So as I look around this room at people who either have <u>achieved</u>, or are on their way to achieving, <u>remarkable</u> levels of <u>success</u>, I would offer

> this reminder: don't wait. Don't wait until you make your first million to make a difference in somebody's life. If you have something to give, give it now. Serve food at a soup kitchen. Clean up a neighborhood park. Be a mentor. Not every day is going to offer us a chance to save somebody's life, but every day offers us an opportunity to affect one.

Non-authorial instances are those in which it is not the author's emotions which are described but those of other human individuals or groups. The following excerpt is from the TED talk “Why we love, why we cheat” by Helen Fisher.

> And this graduate student was madly in love with another graduate student, and she was not in love with him. And they were all at a conference in Beijing. And he knew from our work that if you go and do something very novel with somebody, you can drive up the dopamine in the brain, and perhaps trigger this brain system for romantic love. So he decided he'd put science to work, and he invited this girl to go off on a rickshaw ride with him. [..] Apparently they go all around the buses and the trucks and it's crazy and it's noisy and it's exciting. [...] So off they go and she's squealing and squeezing him and laughing and having a wonderful time. An hour later they get down off of the rickshaw, and she throws her hands up and she says, "Wasn't that wonderful?" And, "Wasn't that rickshaw driver handsome?"

In this instance the writer is not evaluating, or at least not with respect to the couple of students. The writer presents herself as merely reporting on the emotional reactions of both - she is not taking responsibility (at least not directly) for any positive (or negative) assessment, which might be suggested or invoked by such a reporting of emotions. However, the fact that the story ends in this particular way contributes to the speaker's general purpose of presenting love as mysterious. This strategy works when the source of the reported Affectual value is presented as reliable or reasonable in his/her emotional responses, and is consistent with the text's overall evaluative position.

Martin & White (2005) also point out that many types of discourse (especially Public discourses) do not function simply as cases for isolated individuals. As has been discussed in the Critical Discourse Analysis field (starting from van Dijk (1996, 2008), they often stand in for generalised social types or groupings, e.g. embattled teachers, the homeless, asylum seekers, victims of crime, drug addicts, business leaders, scientists. A reader who sympathises with the emotional response attributed to a given social type is thus predisposed to legitimate the social position represented by that social type.

For example, in the passage below the speaker, Jackson Katz, in "Violence against the women - it's a men's issue" talks about himself as part of the general category of *men* as opposed to *women*:

> The first is that it gives men an excuse not to pay attention. Right? A lot of men hear the term "women's issues" and we tend to tune it out, and we think, "Hey, I'm a guy. That's for the girls," or "That's for the women." And a lot of men literally don't get beyond the first sentence as a result. It's almost like a chip in our brain is activated, and the neural pathways take our attention in a different direction when we hear the term "women's issues".

The presentation of the two categories - men vs. women - works as an introduction to present the men's point of view from the inside, and then prove it wrong:

> But there's so many men who care deeply about these issues, but caring deeply is not enough. We need more men [...] with the courage, with the strength, with the moral integrity to break our complicit silence and challenge each other and stand with women and not against them. By the way, we owe it to women. There's no question about it. But we also owe it to our sons. We also owe it to young men who are growing up all over the world in situations where they didn't make the choice to be a man in a culture that tells them that manhood is a certain way. [...] We that have a choice, have an opportunity and a responsibility to them as well.

In my annotation the aim is to reflect the attitude of the person who actually is speaking or his/her surrogate in case their positions coincide.

A further feature, which is of the other two categories as well, is that attitude can be *implicit* (or invoked) or *explicit*. I will explain in Chapter 5 that my research mainly deals with explicit attitude.

An example of implicit affect has been found in Kennedy's inaugural discourse:

> Let the word go forth from this time and place, to friend and foe alike, that the torch has been passed to a new generation of Americans born in this century, tempered by war, disciplined by a hard and bitter peace, proud of our ancient heritage and unwilling to witness or permit the slow undoing of those human rights to which this Nation has always been committed, and to which we are committed today at home and around the world.

The sentiment of the expression “tempered by war” depends on the context, and in this case it is a positive evaluation of the new generation of Americans.

3.2.2 Judgement: definition and what to bear in mind during the sentiment annotation

Judgement is related to the language that criticises or praises, condemns or applauds the behaviour - the actions, deeds, sayings, beliefs, motivations - of human individuals and groups. Perhaps the most obvious examples of judgement involve assessments by reference to systems of:

- Legality/illegality
- Morality/immorality
- Politeness/impoliteness

Values of judgement involve evaluations by which the person judged will be lowered or raised in the esteem of their community, but which do not have the same legal, religious or moral implications as the first set. Here there are assessments of:

- *Normality* with terms such as *eccentric, maverick, conventional, traditional*.
- *Competence* with terms such as *skilled, genius, knowledgeable, stupid, brilliant, incompetent, powerful*.
- *Psychological disposition*, both as distinguishing trait of the personality or a temporary behaviour, with terms such as *brave, cowardly, determined, obstinate, zealous, committed, lazy, immoral, virtuous, sinful, lascivious, innocent, unjust, fair-minded, law-abiding, murderous, cruel, brutal, compassionate, caring, dishonest, honest, deceptive, fraudulent*.

Below I offer an example of judgement at work in a piece of news part of the annotated corpus, entitled "Computer selected and disseminated without FBIS editorial intervention":

> Recently, North Korea strongly denounced comments made by U.S. President George W. Bush during his Seoul visit last month accusing the North Korean leadership of starving its people while developing weapons of mass destruction. [...] The report comprehensively blamed the North Korean authorities for committing wrong-doings in terms of human rights.

It is also possible that judgement is implicit. In their website, Martin and White [1] offer the example of a commentator that may inscribe a value of negative capacity by explicitly accusing the government of "incompetence" as opposite to an example such as "the government did not lay the foundations for long term growth" in which it is implicit.

As in the case of affect, implicit judgement relies upon conventionalised connections between actions and evaluations and, as such, it is highly subject to the reader's position. In some instances, the ethical evaluation evoked by some 'factual' description has become so naturalised or taken-for-granted in a given cultural situation that it is likely to be regarded as explicit rather than as implicit. Quoting Fairclough (1989) (p. 64), "conventions routinely drawn upon in discourse embody ideological assumptions which come to be taken as mere 'common sense'".

Another example comes from a TED talk entitled "Photos that changed the world" by Jonathan Klein:

> In the 1960s and 1970s, the Vietnam War was basically shown in America's living rooms day in, day out. News photos brought people face to face with the victims of the war, a little girl burned by napalm, a student killed by the National Guard at Kent State University in Ohio during a protest. In fact, these images became the voices of protest themselves.

Nowadays the moral evaluation associated with such an action is so firmly established in our culture as to be virtually automatic (e.g. *victims, burned, killed*). This

[1] http://www.grammatics.com/appraisal/AppraisalGuide/Framed/Frame.htm

means that the author did not need to use more explicit attitude words such as *murder* to make his point. Similarly, the expression “solemn oath” can be seen as a convention.

For similar reasons, Martin and White mention that the way particular words are interpreted may also depend on the social and ideological position of the reader. The actual meaning of a word, its specific judgement value, will often be determined by where it occurs in the text and by what other judgements have been previously made in the text. They propose the example of *militant*: from a left-wing, union oriented perspective, the term has obvious positive connotations - “to be militant” is to have a praiseworthy determination to pursue the interests of the working class; from a right-wing, management perspective, *militancy* is commonly associated with a negative value, connoting a hard-line, obstinate determination to frustrate management initiatives wherever possible. This is also the case of “revolutionary beliefs”, found during my annotation. It might be considered as cliché or not, and such perspective might vary also across languages.

Finally, it is important to point out the difference between affect and *provoked* judgement. In this case, I will use another example from Martin and White’s website[1], taken from a newspaper commentary by Norman Tebbit on the *Daily Mail on Sunday* (Feb 4 2001) entitled “Crocodile tears for the men of steel”:

> The Prime Minister is not just angry. He is scared.

Angry and *scared* are instances of affect because they indicate the author’s personal opinion, but they are also instances of implicit negative provoked judgement since they aim at implying that he is either incapacitated or cowardly.

3.2.3 Appreciation: definition and what to bear in mind during the sentiment annotation

Appreciation involves positive and negative assessments of objects, artefacts, processes and states of affairs rather than human behaviour. The most obvious values of appreciation are concerned with what is traditionally known as *aesthetics*, with positive or negative assessments of the form, appearance, construction, presentation or impact of objects and entities.

[1] http://www.grammatics.com/appraisal/AppraisalGuide/Framed/Frame.htm

An example of Appreciation can be seen in the TED talk “Photos that changed the world” by Jonathan Klein:

> Unfortunately, some very important images are deemed too graphic or disturbing for us to see them. I’ll show you one photo here, and it’s a photo by Eugene Richards of an Iraq War veteran from an extraordinary piece of work, which has never been published, called “War is Personal".

On their website, Martin and White also specify that in some instances human participants may also be appreciated when they are described as *beautiful, handsome, ugly, lopsided, gangly, striking* and so on. Such evaluations do not represent instances of judgement because they do not involve assessments of behaviour, being *beautiful* or *ugly* in this physical sense not a question of morality. There is still the possibility, however, that in the right context a term such as *beautiful* can take on moral associations and hence operate as a value of judgement, for example in “She was always kind, considerate and forgiving - truly one of the most beautiful human spirits I ever encountered”.

Finally, they underline that appreciation shares with judgement this property of being oriented towards the ‘appraised’ rather than the subjective ‘appraiser’, typical of affect. To say that “the building bores me” (Affect) is to offer an individualised evaluation that depends entirely on my own, singular state of mind or emotional disposition. Crucial here is the fact that the emotional reaction (depress, bore, etc.) has been detached from any human experiencer of the emotion and been attached to the evaluated entity as if it were some property which the entity objectively and intrinsically possesses. To say that “the building is boring” (Appreciation) is to offer an evaluation of a different order. Conversely, Judgement involves consciousness, volition or intentionality. This means that values of judgement (at least in their adjectival form) can be slotted into the collocational frames of the type, as we can see from the following examples taken from the website:

- It was corrupt of the Minister to accept these payments.
- It was dishonest of you not to tell her.
- It was brave of Mary to stand her ground
- It was clever of you to hide your wallet in the vegetables

- It was eccentric of you to wear that hat.

This is not possible of values of Appreciation. Thus the following would be incongruous:

- It was beautiful of the sunset to light up the sky like that.

- It was ugly of the scar to gape like that.

In the annotated texts, I found this ambiguity in the case of the piece of news entitled “Computer selected and disseminated without FBIS editorial intervention” previously seen:

> The U.S. State Department on Tuesday (KST) rated the human rights situation in North Korea "poor" in its annual human rights report, casting dark clouds on the already tense relationship between Pyongyang and Washington.

In my opinion, this should be considered as judgement since it involves assessments of right and wrong and there is a sense of ‘blame’ to the agent who is thereby evaluated (North Korea).

3.3 Beyond participants and circumstances: an investigation of processes

In this Section I will look at verbs as processes. This concept is related to that of the metafunctions seen in Section 3.1. In this work, the ideational metafunction is the one we are most interested in, especially in its experiential realization. It is represented by a ‘process’ (realized by a verbal group), the ‘participants’ involved (realized by nominal groups) and their ‘circumstances’ (usually realized by adverbial groups).

The choice of exploring verbs is due to the fact that I wanted to counter-balance the importance given to both nominal and adverbial groups, since they represent the main objects of my annotation scheme.

In particular, among the vast variety of the type of processes (material, mental, relational, behavioural and verbal), I am interested in the sub-categories most related to evaluation, i.e. 'cognition' (*to know, to think, to believe, to realize*) and 'affection' (*to like, to love, to hate*) of the mental ones[1], as well as the part related to psychological behaviour in the 'behavioural' ones (*to blame*).

My hypothesis is that, while 'cognition' processes are the most prominent in political speeches and news, 'affection' are moderately frequent in political speeches and more frequent in TED talks, but not in news, and 'behavioural' frequent in all the text types.

Despite being aware of the limitation that processes are not always mappable to word forms, in my English corpus of political speeches, news and TED talks, and across other bigger reference corpora, I did a count of some verbs, whose frequency and collocates were likely to yield to interesting conclusions:

- WIT3, Web Inventory of Transcribed and Translated Talks (Cettolo *et al.*, 2012), a corpus of more than 900 TED talks.
- CORPS (Guerini *et al.*, 2008), a corpus of 3600 political speeches.
- Spinn3r Dataset (Gordon & Swanson, 2009), a corpus of web blogs.

Table 3.1 shows the corpus size of all corpora, which have been used to count the normalized frequency of the items. In the case of my corpus, SentiML, the normalized frequency has been counted out of the total number of words, and the text type in which the verb appears has been specified.

	Corpus size (# words)
SentiML	9055
WIT3	2.35M
CORPS	8M
Spinn3r	13M

Table 3.1: Size of the corpora used in the analysis of the processes.

[1] 'Perception' verbs such as *to see, to feel, to hear* have been left out because not relevant to evaluation.

	Frequency (# times)				Normalized frequency (# times/corpus size)			
	My corpus	**WIT3**	**CORPS**	**Spinn3r**	**My corpus**	**WIT3**	**CORPS**	**Spinn3r**
Mental processes: cognition								
To think	9 (TED)	11105	14552	17474	0.10%	0.47%	0.18%	0.13%
To realize	/	570	933	1377	/	0.02%	0.01%	0.01%
To understand	5 (1 news, 4 political)	2042	3069	2531	0.06%	0.09%	0.04%	0.02%
Mental processes: affection								
To like	/	16048	12899	37979	/	0.68%	0.16%	0.29%
To love	2 (1 TED, 1 political)	2124	2243	10681	0.02%	0.09%	0.03%	0.08%
To regret	/	18	103	334	/	0%	0%	0%
Behavioural processes								
To blame	2 (1 news, 1 political)	78	281	546	0.02%	0%	0%	0%

Table 3.2: Frequency and normalized frequencies of the verbs used as representative for the mental and behavioural processes across several English corpora.

Mental processes: cognition

As representative verbs for the category ‘cognition’ of mental processes, I have chosen *to think, to realize, to understand*. Table 3.2 shows both their frequency and normalized frequency, while their usage is described in more detail below.

To think was found 9 times in SentiML, only in the case of TED talks. We find a similarity with the corresponding corpus in terms of text types, WIT3, which has the biggest percentage (0.47%) among the three reference ones. In terms of collocations, no pattern has been found to be specific of any text type: in Spinn3r *think* is followed by the pronoun *I, you, we* (e.g. “think I got up”, “think I prefer”) or a person/object, by the prepositions *about/of*, by the conjunction *that* (e.g. “that they are gods?”) and the adverb *what* (e.g. “what we need”). The most common left collocate (i.e. preceding) collocate is the pronoun *I* (I think), followed by the negated form *don’t*.

To realize has no occurrences in SentiML, and it is also quite rare in the reference corpora with a percentage between 0.01 and 0.02%. It is mostly followed by *that* (e.g. “realize that I am comfortable with”), a direct object or an adverb such as *now, later*. The structures “happen/come to + realize” are also present. In CORPS, which is a corpus of political speeches, it is preceded by *must/have to* (e.g. "we must realize that the educational account"), *begin to, will, should* and negated forms. In Wit3 very common are the informal forms "you realize" and "don't realize".

To understand appears 5 times in SentiML, once in the news ("The human communities have [. . .] understood that any progress") and 4 times in speeches ("That we are in the midst of crisis is now well understood", "We understand that greatness is never a given", "What the cynics fail to understand is that the ground has shifted beneath them", "They understood that our power alone cannot protect us"). In WIT3, with the biggest percentage of all (0.09%), it mostly has *I, we, you* as subjects.

In Spinn, it is followed by a *wh*-question (*why, what, how, where*) but also a direct object/person/verb (*me, stories, things, stuff, having an illness*), adverbs (*more, right now*) and conjunctions (*that, if*). It is preceded by modal verbs (*can/could, may/might, would/will*) and other common verbs (*need, seem, start to/try to, help to*) and it appears also in the negated form. In CORPS it is used mostly in the same way, but with different objects (e.g. *peace, policies, decisions*).

Mental processes: affection

As representative verbs for the category 'affection' in mental processes, I have chosen *to like, to love, to regret*. Table 3.2 shows their both their frequency and normalized frequency, while their usage is described in more detail below.

To like has no occurrences in SentiML. In Spinn it is mostly followed by the infinitive/-ing form (e.g *to keep, ending*), direct object/person and in the negated form. In Wit3 it appears 0.68% of times with different objects (e.g. *hormones, bacteria, fireflies*).

To love is twice in SentiML, once in a TED talk ("misery truly does love company") and once in political speeches ("the land we love)". Overall it is preceded by a number of persons and followed by direct objects/persons (mostly *you, it*) or infinitive/-ing form (e.g. *to eat, doing*). In Wit3 it has different objects (e.g. *company, stories, fish, farmers*).

To regret has no occurrences in SentiML, and very few in the reference corpora. It is followed by infinitives/-ing form (e.g. *to say, being, doing, going, not asking*), by *that* ("regret that the exclusions"), direct object ("regret this signing"). In CORPS different objects are *staying close, the move, my nine years of service*, while in Wit3 mostly "regret the decision", "regret to say".

Behavioural processes

As representative verbs for the category behavioural processes, I have chosen *to blame*. Table 3.2 shows its frequency and normalized frequency, while its usage is described in more detail below.

To blame is used once in the SentiML news ("The report comprehensively blamed the North Korean authorities") and once in speeches ("blame their society's ills"). It appears very rarely in the reference corpora, mostly as transitive verb (e.g. "blame somebody/something" including objects and verb in -ing form) and it is frequently negated. In CORPS it is frequently followed by the preposition *in* (e.g. "in the Congress", "in the White House"), *for* (e.g. "for this", "for not explaining"), direct object /person (e.g. *the other party, them, you, people*).

From the analysis done on SentiML and other reference corpora the hypothesis that 'cognition' processes are the most prominent in political speeches and news is partly confirmed, since they are the most frequent in political speeches (both in SentiML and CORPS), but they also frequently appear in TED talks (both in SentiML and WIT3).

In addition, the hypothesis that the 'affection' processes would be moderately frequent in political speeches and more frequent in TED talks, but not in news, is confirmed in SentiML, but not in the reference corpus Spinn3r for news since it has a total 0.37%, lower than in the other two.

As for the 'behavioural' processes being frequent in all the text types, unfortunately no significant conclusions can be drawn especially on SentiML because only one verb has been taken into account. However, we can still conclude that *blame* is surprisingly not frequent in the reference corpora either.

In the following Chapter I will give more space to the other two languages, namely Italian and Russian, by exploring differences and similarities in the texts according to the categories of the Appraisal Framework. The analysis, like in this Chapter, will be supported by counts of interesting patterns in the corpora.

Chapter 4

Commonality and diversity: Appraisal Framework, Corpus Linguistics and Translation Studies at work

In this chapter I intend to explore differences and similarities in the corpus data in English, Russian and Italian focusing on translations of President Obama's 2009 Inauguration speech. I will analyse their evaluative language under the main categories of the Appraisal Framework used for the annotation (i.e. affect, judgement and appreciation) described in Chapter 3. I will make use of Corpus Linguistics (CL) in the case of interesting patterns from the point of view of the translation choices, like several previous studies (Baker, 1995; Laviosa, 2002; Oakes & Ji, 2012; Tognini-Bonelli, 2001), including those specifically in the political domain (Romagnuolo, 2009) and in SFL (Neale, 2006).

As a result, I will obtain a deeper and more comprehensive picture of the texts under analysis, which will support the quantitative analysis on both the manually-annotated data in Chapter 6 and the automatically-annotated data in Chapter 8.

Munday (2012)'s work will be taken as primary reference for the discussion since it provides an excellent overview of other background works, as well as some practical analyses. The motivation for not considering several layers inside each of the categories of attitude (e.g. security, happiness, inclination, satisfaction for affect, etc.) like Munday does is the different goal that this work has: while deepness has been always preferred in traditional analyses inside the AF framework, the simplification of my annotation consciously sacrifices that in order to be useful as test for a system that can automatically apply such labels. However, I share the same goal of analysing any shifts in the value systems according to the AF, as well on the translation strategies level. As Munday says (p. 44) “such value shifts are complex because (i) they might be due to a number of reasons (cross-cultural differences, deliberate textual manipulation, degree of competence or some other form of translator preference)” and (ii) “they may be expressed by subtle linguistic markers”.

On the other hand, my work is an attempt to push down the boundaries among various disciplines such as cross-linguistic variation, language typology, contrastive linguistics, translation studies and multilingual computational linguistics, as proposed by Teich (1999), with a model “using the representational categories SFL sets up as parameters along which cross-linguistic variation can be described” because it “uses categories that are cross-linguistically relevant” and lends itself “as an anchor for translational concepts” (Teich, 2001) (p. 193).

Two of the essential requirements proposed by Teich (2003) are satisfied by the present analysis, namely that:

- The analysis should have in the background a contrastive typology that provides information about the major contrasts and commonalities between the grammatical systems of the languages under investigation. This is particularly true in the case of English, Russian and Italian being them very different in a number of features (morphological, syntactical and cultural) because they belong to different families (Germanic in the case of English, Romance in the case of Italian and Slavonic in the case of Russian).

- The analysis should be corpus-based. The use of CL is motivated by the fact that “one of the central axioms of SFL is the recognition of ‘meaning potential’, and the most productive means for observing meaning ‘potential’ is through ’instantiation’ - or ‘evidence”’ provided by corpora (Neale, 2006) (p. 145).

  In fact, CL matches the fact that “the theoretical framework [SFL] has been developed on naturally-occurring examples” (Thompson & Hunston, 2006) (p. 2).

> It must also be noted that, apart from the support given by the CL in this chapter, more quantitative statistics will be presented in the next one.

My work represents one of the few examples in which the Framework has been systematically applied for both Italian (Manfredi, 2011; Pounds, 2010) and Russian (Bateman *et al.*, 2000), as opposed to other languages such as French and Spanish in which this has been already done (Caffarel-Cayron, 2006; Lavid *et al.*, 2010; Taboada & Carretero, 2012).

As far as the link between SFL and Translation Studies is concerned, Translation Studies have been recognized not to be "a new direction in SFL, but work is expanding rapidly in many places around the world, reflected in research projects, publications and also in translation courses informed by SFL" (Matthiessen, 2009), and particularly multilingual corpora by Neumann & Hansen-Schirra (2005); Pagano *et al.* (2004); Teich (2003). In my case, the analysis has been done on a subset of the annotations across the pairs English-Italian and English-Russian, and it will be based on the work of Taylor (1998) (pp. 47-64), who well summarises Malone (1988)'s theoretical framework, as well as on Baker (2002). Back translation will be provided in brackets.

4.1 Previous works

Among those studies mentioned by Munday (2012), that of Cavaliere & Abbamonte (2006) seems to be particularly interesting both because of the proximity of subjects to mine and the use of Italian as target language. They highlight a difference between the English Source text (ST) of the UNICEF *State of the World's Children* Report 2004 and the Italian Target text (TT) in fulfilling the communicative purposes of the text: the declared purpose was to inform and the ultimate one was to attract attention to the girls' urgent need of education. While emotion was an important feature in the English text, it was not always shared by the Italian version. They did not actually use the Appraisal Framework, but rather the Applied Descriptive Translation Studies framework, described in Bollettieri Bosinelli & Ulrich (1999) as a "descriptive, target-oriented , functional and systemic approach", which allowed them to discover that "in the documents under investigation many types of strategies, both covert and overt, are at work". They also dealt with affective positioning, translation strategies and commitment/distancing.

Italian and British news have been analysed by Pounds (2010), whose findings of particular interest for my research have been that subjective and explicit language often characterises Italian news more than their British counterparts.

Finally, the political domain has been among the most researched in the field of Critical Discourse Analysis (CDA), to the point of having an entire branch devoted to it, Political Discourse Analysis by Schäffner (2004). An interesting warning that Schäffner makes is that in this domain "what may look like a mistranslation or a translation loss at a first glance (or from a linguistic or text-specific point of view) will actually turn out to highlight the socio-political or ideological structures, processes, norms and constraints in which translations were produced (and received)" (ibid., p. 142).

4.2 Political speeches: The case of Obama's inaugural speech

Munday (2012) offers an analysis of Obama's first inaugural speech of 2009, whose "wealth of interpretings and translations in so many different languages provides an unusual opportunity to analyse the strategies adopted in the construction of the TTs" and "lends itself well to exemplify appraisal analysis specifically because of the inherently evaluative and ethical tone" (p. 42).

Munday (2012) also mentions the common practice of the 'invisibility' of the translator of political texts (Bielsa & Bassnett, 2008), (Schäffner, 2008), which has to be carefully analysed in the light of what previously said by Schäffner about the covered ideology.

An interesting piece of information he gives at the very beginning is that in Russia only edited excerpts were broadcast from a studio with a domestic commentator, and they were treated as a minor story by most Russian TV channels (BBC, 2009). Indeed, an interesting point is that "these forms of attempted censorship are fascinating in themselves, particularly as the Internet and social media now have the power to permeate previously hermetic societies and potentially stabilize the political landscape"

(Munday, 2012) (p. 44). This applies to Russia, where no censorship is carried out on websites.

My English ST has been taken from the Avalon Project online archive with major documents in American Law, History and Diplomacy[1]. I also compared it to that of the Digital Archive managed by the U.S. Department of State to make sure that no editorial control had been done on it.

The Russian TT has been also found among the official ones made available on Digital Archive website managed by the U.S. Department of State[2].

However, since for Italian no translation was provided on the website, the TT has been taken from the newspaper *La Repubblica*[3].

For both Russian and Italian, other translations have been additionally taken into consideration in some interesting cases[4]. This is motivated by the awareness that each translation represents only a point of view and, more generally, the product of a number of factors, such as ideology, experience, conventions, linguistic knowledge and, potentially, mistakes.

Since the "TT" notation will be used to indicate the primary translation for both Italian and Russian, where two alternatives are presented, these secondary translations will be indicated as "TT2".

4.2.1 Affect

According to Munday's analysis, there are surprisingly few instances of happiness in the speech, while it is rather security that dominates. Examples of negative words or expressions are *crisis, war, violence and hatred, the challenges we face, those who seek to advance their aims by inducing terror and slaughtering innocents* as opposed to positive intentions such as *we intend to move forward, we are ready to lead again, duties that we seize gladly* or expressions referred to the military like *willingness to*

[1] http://avalon.law.yale.edu/21st_century/obama.asp

[2] http://iipdigital.usembassy.gov/st/english/texttrans/2009/01/20090120130302abretnuh0.2991602.html

[3] http://www.repubblica.it/2009/01/sezioni/esteri/obama-insediamento/testo-discorso-italiano/testo-discorso-italiano.html

[4] http://presportal.ru/rechi-liderov/inauguracionnaya-rech-baraka-obamy/, http://www.corriere.it/Speciali/Esteri/2009/Discorso_Obama/

find meaning in something greater than themselves and previous enemies like *willing to unclench your fist*.

At the very beginning there are a couple of examples of what I annotated as affect: *I stand here today humbled by the task before us....grateful for the trust you have bestowed*. However, as Munday (2012) quotes, Martin & White (2005) qualify them as hybrids as they "construe an emotional reaction to behaviour we approve or disapprove of" and are "affectual inscriptions invoking judgement" (ibid., p.68). The Italian TT mantains the interpreting of affect by offering the literal translation *umile per il compito che ci aspetta, grato per la fiducia che mi avete accordato*, which includes the use of diffusion as necessary translation strategy for the grammatical structure *che ci aspetta* (is waiting for us) instead of the more economic *before us*. In the Russian TT *humbled by* corresponds to ощущая огромную важность (feeling the huge importance of) in TT1 and полностью сознавая грандиозность (fully aware of the enormity of) in TT2. In the same way, *grateful* has been rendered as испытывая признательность (feeling gratitude) in both TTs, with the addition of глубокую (deep) in TT2. These can be seen as examples of impersonalisation, since it was the translators' choice not to use the corresponding adjectives in the target languages.

At line 19, both the Russian TTs offer an occurrence of affect where there was none in English (*But know this, America*), and consequently in Italian (*Ma America, sappilo*), by adding Но я хочу, чтобы Америка знала (I want America to know). The reason for the use of the amplification strategy here might be the translator's awareness - either conscious or unconscious - that the use of imperative for mental processes such as *to know* is unlikely. This is confirmed by the fact that only 4 occurrences of знай followed by это (this), have been found in the Russian Internet Corpus[1] (consisting of 160 million words as snapshot of the Russian language on the Internet) and 5 occurrences in the Russian National corpus[2]. However, probably with the aim of being faithful to the ST, this rule is broken later: *know that your people will judge you on what you can build, not what you destroy* that becomes мы говорим: знайте, ваши народы будут судить вас по тому, что вы построите, а не по тому, что вы разрушите (we say, know that your people will judge you on what you build, not what you destroy).

Opposite to this case is line 26, where *[Our journey has never been one of] settling for less* in Italian and Russian loses some emphasis by being translated as *non ci siamo*

[1]http://corpus.leeds.ac.uk/ruscorpora.html
[2]http://corpus.leeds.ac.uk/internet.html

mai accontentati (we were never easy to please) and довольствовались малым (we were never satisfied with little).

At line 65 we see an interesting example of divergence for the English phrase *to do as we please*. The Italian and Russian TTs use the formal alternative *come più ci aggrada* and что мы пожелаем (as we wish).

In the light of the cross-linguistic and cross-cultural analysis of appraisal, it is also noticeable that at line 77 the figurative image in *we will extend a hand* is kept: *vi tenderemo la mano* and мы протянем вам руку.

Finally at line 106, there is a case of omission in Russian when *we delivered it safely* has been translated as передали его (we gave it), whereas in Italian the same number of words is maintained although with a slight different construction: *abbiamo consegnato intatto* (we gave it intact) of what would have been the literal equivalent *in modo sicuro*.

4.2.2 Judgement

The evidence from my annotation confirms Munday's about the amount of judgement evaluation being higher than affect.

At the very beginning we find that for *the task before us*, the Russian translation reports поставленных перед нами задач (the tasks assigned to us). Here the translator opted for a cliché, a "stop-phrase" commonly used, as evident from a corpus-based search in the Russian Internet corpus, in which 74 occurrences were found in which поставленных, перед, задач are in the same sentence (although sometimes in different order or with an object other than нами (us)), and 55 in the Russian National corpus. The strategy adopted in the following part goes in a sort of opposite direction: *for the trust you have bestowed* becomes признательность за оказанное мне доверие (for the trust accorded to me), with the effect of hiding the action taken by the American people to elect Obama.

In the same line, the English *our ancestors* is translated in Italian TT1 with the more familiar term *i nostri padri* (our fathers) while the Italian TT2 and the Russian one keep the literal translation *i nostri antenati*/нашими предками.

This is followed by some criticisms. The first against a *far-reaching network [of violence and hatred]* (line 10), kept in Italian (*una rete [di violenza e di odio] di portata globale*) through the use of diffusion and reordering of the equivalent of *far-reaching* (*di portata globale*); conversely in Russian we see the use of the substitution strategy, since the object is changed into the people who created such network: тех, кто всюду сеет [насилие и ненависть] (those who everywhere sow [violence and hatred]).

The second criticism is against American schools (line 13): *our schools fail too many*. In the Italian TTs *fail* is translated as *lasciano indietro, trascurano* (leave behind, overlook) and in the Russian TT2 часто не справляются со своей задачей (often do not do their job), while the effect is diminished in the Russian TT1:не дают достаточных знаний многим учащимся (do not provide sufficient skills to many students).

This seems to be a case supporting Cavaliere & Abbamonte (2006)'s hypothesis about the difference in the way the ST and the TT fulfil the communicative purposes of the text. Very much as in their example of the UNICEF report, it might be that, because the ultimate goal in English is to attract attention to urgent needs and emotion represents an important feature, this is not shared by the TTs, and in particular by the Russian on in which the translator is simply reporting a piece of information.

The third criticism is against *worn-out dogmas*, an expression that is perfectly rendered in Italian (*dogmi stanchi*) and Russian (избитым догмам).

Line 32 represents an interesting example in so far as the phrase *till their hands were raw so that we might live a better life* referring to ancestors is kept in the Italian TT, which however seems to mark more their strong will with *fino a massacrarsi le mani per permettere a noi di vivere una vita migliore* (until mangling their hands to allow us to have a better life); this translation seems somehow better than the literal one provided in TT2 *fino ad avere le mani in sangue, perché noi potessimo avere un futuro migliore*. The Russian TT prefers to use two figurative images до мозолей на руках (to calluses on their hands) and в стремлении к лучшей жизни (in the quest for a better life) .

At line 36, *Our workers are no less productive than*, we found a minor change in the Russian Сегодня производительность нашего труда не ниже (Today the productivity of our work is no less), where the noun *ability* is preferred.

The same term *capacity* has been translated differently a few lines afterwards in *our capacity remains undiminished*. Munday notes the same in Spanish, although the strategies adopted in Italian is to use plural (*le nostre capacità*) followed by the verb *to be* and the adverb *still* (*sono ancora*) instead of *to remain*, and finally the adjective *intatte* (intact). The Russian TT also makes some changes by using the singular as in English, but not of the equivalent Наш потенциал (Our potential) followed by the verb in the past не приуменьшился (did not diminish).

Afterwards Munday (p.50) notes a series of omissions of the adjectives in Spanish, not done in Italian and Russian: *swift action, enduring convictions, hard work*. In these two languages there are instead two examples of non equation. The first example is *forge a hard-earned peace*, again another example of necessary diffusion in Italian *forgiare una pace duramente guadagnata*, and a quite remarkable divergence in Russian: доведем до конца трудную работу по достижению мира (bring to the end the hard work for the achievement of peace).

The second example of non equation is *quiet force [of progress]*, which the Italian TT reports as *forza pacifica [del progresso]* (peaceful force), not a wrong choice in this context; the Russian TT makes use of addition, so marking even more the force: скрытой движущей силой [прогресса] (hidden driving force [of progress]).

An important shift noticed by Munday in the Spanish TTs happens in Russian as well: *We will begin to responsibly leave Iraq to its people* is translated as Мы начнем ответственный процесс ухода из Ирака, оставляя страну иракцам (We will start the responsible process of withdrawal from Iraq, leaving the country to Iraqis). The reasons for this amplification might be both a linguistic preference of the Russian language, and the will to mark the long-term effect of such action, i.e. that Iraqis will have the power on their country again.

Fascism and *communism* are then analysed as examples of non-core lexis that invoke judgement. In English these are preceded by the verb *faced down*, well translated in Italian by *hanno sgominato* and in Russian победили.

Other examples of invoked judgements highlighted by Munday (p.53) are linked to the American history (industry, search for land, suffering of slaves and tough conditions of farmers if the Midwest):

> **English ST**: For us, they toiled in sweatshops and settled the West; endured the lash of the whip and plowed the hard earth.
>
> **Italian TT**: Per noi, hanno subito lo sfruttamento sul lavoro e si sono stabiliti nell'Ovest. Hanno sopportato la frusta e arato la terra dura.
>
> (For us, they suffered the work exploitation and settled the West. They endured the whip and plowed the hard earth).
>
> **Russian TT**: Ради нас они гнули спины в условиях потогонной системы и осваивали Запад, терпели удары кнута и распахивали целинные земли.

(For us they curved backs in the conditions of the sweating system and settled the West, endured the lash of the whip and plowed the virgin lands).

Settled the West and *Plowed the hard earth* have word-for-word translations in both languages, while *endured the lash of the whip lash* is omitted in Italian, and *they toiled in sweatshops* is better rendered in the Russian TTs and the Italian TT that keep the figurative images of the sweatshops; the Italian TT2 prefers a more generic sentence about work exploitation that does not achieve the same effect (e.g. *hanno faticato nelle fabbriche* (they worked hard in the factories)). We also notice the difference in punctuation across the three languages since the Italian TT2 keeps the semi-colon, while the Italian TT prefers a full stop and the Russian TT prefers a colon.

Munday also reports two examples of metaphors. The first is *[We will] wield technology's wonders*, whose translations lose this aspect: Italian uses *ricorreremo alle meraviglie della tecnologia* (we will draw upon technology's wonders). A good alternative might have been *impiegheremo*. Russian uses применим достижения технического прогресса (we will apply technology's wonders). Finally, in both TTs the effect given by the alliteration of *will...wield...wonders* pointed out by Munday is definitely lost.

The second metaphor is *[We will] harness the sun and the winds and the soil to fuel our cars*. This time the Italian TT successfully uses the best equivalent *Imbriglieremo il sole e i venti e il suolo per alimentare le nostre auto*, whereas the Russian TT made use of a different structure Мы добьемся того, что энергия солнца, ветра и земли будет приводить в движение наши автомобили (We will make sure that the energy of the sun, wind and earth will drive our cars) .

At line 76, *blame* in *blame their society's ills on the West* is translated in Italian as *scaricare* in *scaricano sull'Occidente i mali delle loro società* (to shift their society's ills) instead of using the direct equivalent *incolpare* (to blame), in the attempt of denigrating this behaviour. Russian maintains more the similarity to the ST: обвиняют Запад в проблемах, существующих в их обществах (to accuse the West of the problems existing in their societies).

Finally, at line 98 we find a preference in the Russian and Italian TTs for an image instead of *How far we have travelled* that becomes "what a long way we have come" in both cases (какой долгий путь мы проделали and *quanta strada abbiamo fatto*).

4.2.3 Appreciation

My analysis in Italian and Russian confirms Munday's hypothesis (p. 47) that some basic words (e.g. *crisis, war, violence*) form an evaluative core that is most likely to be realized uniformly.

In addition, Munday (pp. 47, 55) mentions the fact that the omission of adjectives or adverbs in the following expressions in Spanish (*raging storm, quiet force of progress, precious gift, great gift of freedom, full measure of happiness*) has a consequence on the intensity of the evaluation (called *graduation* in the Appraisal Framework, and 'force' in my annotation schema).

In Italian and Russian *raging storm, precious gift, great gift of freedom* maintain their force. *To pursue their full measure of happiness* is subjected to condensation in both languages: in Italian *perseguire la felicità* (to pursue happiness), in Russian на свою долю счастья (in their own portion of happiness).

Like in Spanish, for *sapping of confidence* the Italian TT has chosen *perdita di confidenza* (loss of confidence), while the Russian TT has used a verb слабнет уверенность (to lose confidence).

Afterwards Munday reports other metaphors :

- *gathering clouds* translated faithfully into *nubi tempestose* and мрачных туч, with no difference of force as in the case of *raging storm*.
- *stale political arguments* translated as *stessi argomenti politici ammuffiti* (the same mouldy political arguments) and застарелым политическим спорам (inveterate political arguments).
- *we have tasted the bitter swill of civil war and segregation* translated faithfully as *abbiamo assaggiato l'amaro sapore della Guerra civile e della segregazione razziale*, but not in Пройдя через ужасы гражданской войны и сегрегации (Going through the horrors of the civil war and segregation).

An interesting reflection is made on the terms *patchwork* whose negative or positive appraisal depend on context. In this case, because the instance is *our patchwork heritage is a strength, not a weakness*, it is positive. The translations of the sentence are:

> **Italian TT**: il nostro retaggio disomogeneo e discontinuo è una forza e non una debolezza

(our inhomogeneous and discontinuous heritage is a strength, not a weakness)

Russian TT: многообразие общества – это наша сила, а не слабость

(the diversity of the society is our strength, not a weakness)

We notice that both languages struggle to convey the ambiguous meaning of the term: Italian has opted for an amplification ("inhomogeneous and discontinuous"), whereas Russian for a replacement of the whole phrase ("diversity of the society"), by eliminating *heritage* as well. However, we can say that the final positive effect is kept in the Italian and Russian by the literal translation of the second part ("is a strength, not a weakness").

Difficult cases are also represented by the triplet *lost, shed and shuttered* because all the verbs indicate closure, but *shed and shuttered* are non-core items that provide intensification through their semantic strength and alliteration. In Italian I found a similarity to one of the TT in Spanish, which uses the same verb in the same tense (*hanno perso*) for the first two verbs and a good non-core TL phrase as equivalent (*chiudere i battenti*, which literally means "to close shutters"). The Russian TT alters both the structure and the tense, by using "People lose their home and job, businesses close" in the present form (Люди теряют жилье и работу, закрываются предприятия).

At line 23, there is an example of what Munday calls 'simple' epithet: *to choose our better history*, that is translated faithfully in Italian (*di scegliere la nostra storia migliore*), less in Russian: сделать выбор в пользу лучшего будущего (to make a decision in favour of a better future). The choice of Russian of replacing the term *history* with *future* seems to make somehow vanish the reference to a glorious past as well as to future, and we also notice the addition of the preposition *in favour*.

Another example of 'simple' epithet might be *hard choices* translated as *scelte difficili* (difficult choices) and решительный выбор (decisive choice).

Very much like an example seen in the Affect section, there are a couple of examples in which the Italian TT adds an item to load the appraisal:

- In *This is the journey we continue today* (line 34), the verb *to want*, *Questo è il cammino che noi oggi vogliamo continuare* (This is the journey that today we want to continue).

- In *Unpleasant decisions* (line 39), the adverb *more*, *decisioni più spiacevoli* (more unpleasant decisions).

The same happens in Russian:

- *The choice between our safety and our ideals* (line 60) in which *faithfulness* is added, выбор между безопасностью и верностью идеалам (the choice between safety and the faithfulness of ideals).
- *Those ideals still light the world* (line 62) in which the adjective *key* is added, Заложенные в ней идеалы продолжают оставаться маяком (The key ideals in it continue to represent a light for people in the world).

The opposite cases are translations in Russian that convey less appraisal:

- *For expedience's sake* (line 62): ради сиюминутных выгод (for momentary advantages).
- *A moment that will define a generation* (line 84): момент, который станет определяющим для нашего поколения (a moment that will be determinant for our generation).
- *it is precisely this spirit that must inhabit us all* (line 84): именно этой идеей следует проникнуться всем нам (exactly this idea should inspire all of us).
- *[Farms] flourish* (line 780: приносили урожаи (bring abundance).

In the following cases the syntax and the vocabulary is quite changed:

> **English ST**: But those values upon which our success depends - hard work and honesty, courage and fair play, tolerance and curiosity, loyalty and patriotism - these things are old. These things are true. (line 91)
>
> **Italian TT**: Ma i valori da cui dipende il nostro successo - lavoro duro e onestà, coraggio e fair play, tolleranza e curiosità, lealtà e patriottismo - sono valori antichi. Sono verità.
>
> (But those values upon which our success depends - hard work and honesty, courage and fair play, tolerance and curiosity, loyalty and patriotism - are ancient values. They are truth).

Russian TT: Но неизменными остаются те ценности, от верности которым зависит наш успех: это честность и трудолюбие, отвага и стремление к справедливости, терпимость и любознательность, преданность и патриотизм. Эти ценности подлинны.

(But those values upon which our success depends remain unchanged: it is honesty and hard work, courage and commitment to justice, tolerance and curiosity, loyalty and patriotism. These values are true)

In this case, while Italian prefers *antichi* (ancient) as equivalent of *old*, the Russian version prefers *remain unchanged* that represents a common positive collocation for values, when *old* might be negative. Interestingly but not surprisingly in the light of the tendency of Italian of incorporating English words and expressions, both the Italian TTs used the loan *fair play* (Baker, 2002) (p. 33). In addition, it seems that the Russian TT does not respect the pause typical of spoken language, in this case before the verb to allow the listing of the values. The preference for the final *values* instead of the more colloquial *things* seems to confirm this hypothesis.

English ST: which sees us through our darkest hours (line 86)
Italian TT: che ci hanno guidato nei nostri momenti più bui
(that guided us in our darkest moments)
Russian TT: – именно эти качества помогают нам справляться с самыми серьезными испытаниями
(it is these qualities help us to cope with the most severe tests)

In this case, we see two different choices in the TTs quite far from the ST, both in verbs *to guide, to cope* instead of *to see* and images *darkest moments, the most severe tests* instead of *darkest hours*.

To conclude this section, I also would like to point out a couple of examples (*in the year of America's birth, founder of our nation*) of what I would consider as implicit/implied sentiment.

4.2.4 Translation differences

I will now mention some other among the most evident and interesting examples from the point of view of translation strategies. When possible, these will be mapped to the *translation universals* proposed by Baker *et al.* (1993) as the features "which typically occur in translated texts rather than original utterances and which are not the result of interference from specific linguistic systems".

In my corpora, I found that simplification as "the tendency to simplify the language used in translation" (Baker, 1996) (p. 176) is retrievable when the following strategies have been used:

Omission

To nourish starved bodies and feed hungry minds, with the Italian TT simplified at the stylistic level by omitting repetitions and redundant information (Laviosa, 1998) in *per nutrire i corpi e le menti affamate* (to feed hungry bodies and minds). On the other hand, the Russian TT is simplified at the lexical level using using more informal lexis and showing a preference for high-frequency words (Laviosa, 1998, 2002) with чтобы накормить голодных и дать свободу изголодавшимся по ней (to feed the hungry and give freedom starved for it).

Addition

1. *The generosity and cooperation* he has shown *throughout this transition*, with the Russian TT having longer but more explicit phrases such as его великодушие и сотрудничество на протяжении всего периода передачи власти (his generosity and cooperation during the extent of all the period of transaction of power).

2. *[Power] grows through its prudent use*, simplified in both the Italian and Russian TT at the syntactical level with *cresce quanto più lo si usa con prudenza* (grows more when one uses it with prudence) and растет, если пользоваться ею осмотрительно (if it is used prudently).

3. *The tempering qualities of humility and restraint* in Italian becomes *umiltà e dal ritegno che ci caratterizzano* (humility and restraint that characterize us) and in Russian скромность и способность к самоконтролю (humility and ability of self-control).

4. *To strengthen its shield of the new and the weak and to enlarge the area in which its writ may run* simplified at the lexical level in Italian and Russian through the use of less formal language: *a rafforzarla come scudo dei paesi nuovi e dei paesi deboli e ad ampliare l'area in cui la sua parola può avere valore di legge* and in

Russian пусть она усиливает свою мощь, ограждающую молодые и слабые государства, и расширяет сферу своего влияния (to increase its power, enclosing the young and weak states, and expanding its sphere of influence).

5. *Molti riconoscono una forte spinta alla nascita del movimento ambientalista* (Many recognize a strong drive for the rise of the environmental movement)/ and Многие люди приписывают главную причину зарождения экологического движения (Many people attribute the main cause of birth of the environmental movement).

Examples of normalisation, as "the tendency to conform to patterns and practices that are typical of the target language" (Baker, 1996) (p. 183), are:

1. *So it has been* that in Russian becomes Так было всегда (It has always been like this). This is a necessary choice since the English present perfect is always rendered in Russian with the past tense, accompanied by an adverb to give the sense of imperfect.

2. *Many people credit a lot of the birth of the environmental movement* necessarily requiring an expansion: *riconoscono una forte spinta* (recognize a strong drive) and приписывают главную причину (attribute the main cause). Also, strangely enough, while Italian keeps the positive connotation by the use of the equivalent of *drive*, Russian uses *cause*.

Metaphors, consisting of the group of those already in the ST:

1. *They have something to tell us today, just as the fallen heroes who lie in Arlington whisper through the ages* has its metaphor kept in the Italian *Questi uomini hanno qualcosa da dirci oggi, proprio come gli eroi caduti sepolti ad Arlington mormorano attraverso il tempo* (These men have something to tell us today, just as the fallen heroes buried in Arlington whisper through the ages). Они нам многое могли бы поведать, как и павшие герои, покоящиеся на Арлингтонском кладбище (They tell us a great deal could tell as the fallen heroes at rest at Arlington Cemetery).

2. *The ground has shifted beneath them* is literally translated in both: *è venuto a mancare il terreno sotto i loro piedi* (the ground beneath their feet failed) and земля под ними сдвинулась (the ground has moved under them).

3. *In the light of day* where the Italian TT opts for keeping the metaphor translated as *alla luce del sole* (in the light of the sun), while in Russian the metaphor is lost and simplification takes place: полной прозрачности (in full transparency).

4. *Without a watchful eye*, whose image of the eye is kept in Italian *senza un occhio rigoroso*, but not in the Russian без надлежащего надзора (without proper oversight).

5. *Roll back the specter of a warming planet* is translated faithfully in the Italian TT as *respingere lo spettro di un pianeta che si surriscalda* (roll back the specter of a planet that is warming up), but not in the Russian TT that prefers a more concrete concept: принимать меры по борьбе с глобальным потеплением (take measures for the fight against the global warming).

And of the group of metaphors created in the TT:

1. *To lead* is translated in Italian as *aprire la strada* (to open the path), while a more literal translation is given in the Russian TT стать лидером (become leader).

2. *For us, they packed up their few worldly possessions and travelled across oceans in search of a new life* is kept in Italian and another one is created before that *Per noi, hanno messo in valigia quel poco che avevano e hanno attraversato gli oceani in cerca di una nuova vita* (For us, they put in their luggages the little they had and travelled across oceans in search of a new life) the Russian TT does not move away from the ST apart from the addition of *in the hope*: Ради нас они когда-то собрали свои небогатые пожитки и пустились в путь через океан в надежде начать новую жизнь (For us, they once gathered their modest belongings and set off across the ocean in the hope of starting a new life).

3. *Care they can afford* is actually translated almost literally in Italian if not for the use of the adjective (*cure accessibili*, accessible care), whereas in Russian an image is used: медицинское обслуживание, которое им по карману (medical service that fits their pockets).

Both simplification and convergence, i.e. "the relatively higher level of homogeneity of translated texts with regard to their own scores on given measures of universal features" (Laviosa, 2002) or "less variance in textual features in translated texts" (Olohan, 2004), take place in **punctuation**:

1. *My fellow citizens:* In the English version used this phrase is followed by a colon, whereas the Italian uses the semi-colon (*Concittadini,*) and the Russian TT an exclamation sign (Дорогие соотечественники!).

2. *We renew our pledge of support to prevent it from becoming merely a forum* is split in Russian through a semi-colon:мы снова обещаем поддержку; эта организация не должна превратиться в форум (again we promise to support; this organization should not become a forum). This is not always a preference in Russian as demonstrated by the following example: *For they have forgotten what this country has already done; what free men and women can achieve when imagination is joined to common purpose, and necessity to courage*. Они забыли, чего нашей стране уже удалось достигнуть, чего могут добиться свободные мужчины и женщины, объединенные общей целью, способные мечтать и совершать мужественные поступки..

   Splitting is also a common practice in the Italian version, but through the use of a full stop (e.g. *The state of the economy calls for action, bold and swift, and we will act not only to create new jobs, but to lay a new foundation for growth* becomes *Lo stato dell'economia richiede un'azione, forte e rapida, e noi agiremo. Non solo per creare nuovi posti di lavoro, ma per gettare le nuova fondamenta della crescita*.

Linguistic preferences can also be seen as examples of convergence:

- **Word order**. English has a fairly fixed word order and meaning is expressed through the addition of words or movement of words within limited boundaries. An example of this can be *That we are in the midst of crisis is now well understood*. The new information, the *rheme That we are in the midst of crisis* has been given priority to the detriment of the theme *is now well understood* in what would be the expected order: *It is now well understood that we are in the midst of crisis*. Such *marked* choice definitely is an example of Halliday's concept of *Token/Value* in Halliday & Matthiessen (2004) (pp. 230-234). The token *in the midst of crisis* has a value *(it is) now well understood*, so the clause is defined as *encoding*. Such consideration about the directionality of the clause is important to determine its voice, i.e. to relate it to the author's stance towards the topic. According to Halliday's definition this would be *operative* since it leads to a marked focus of the information. In practical terms, the fact that Obama states it at the beginning of his speech straight after the conventional formulae aimed at arousing patriotism indicates that it will be one of the key themes. The choice of changing the order cannot be attributed to a distraction of spoken language either, since presidents' speeches are always carefully edited in order to fulfil several functions throughout (Reiss, 1971/2000). In this case, because the text has to fulfill mainly the function of being "form-focused", the TTs would ideally keep the perspective of the ST author. Russian does it: Тот факт, что мы переживаем кризис, сегодня осознают все (The fact that we are experiencing a crisis, today everybody realizes it), whereas the Italian TT prefers the standard order *È ormai ben chiaro che ci troviamo nel mezzo di una crisi* (It is now very clear that we are in the midst of a crisis). This is somehow a surprising choice since Italian does not have strict order rules, which has been in fact not taken in the Italian TT2: *Che siamo nel mezzo della crisi ora è ben compreso*.

- **Possessive adjectives** There is a preference for possessive adjectives in English, which is not always shared by the other two languages. For example, at line 60 *the choice between our safety and our ideals*, is translated as выбор между безопасностью и верностью идеалам (the choice between safety and the faithfulness of ideals) and *la scelta tra sicurezza e ideali* (the choice between safety and ideals).

- **Nouns** The Italian tendency of using nouns is respected in a few cases such as *the risk-takers, the doers, the makers of things*, which is translated as *i coraggiosi, gli uomini d'azione, i fautori di grandi cose*. In this case Russian prefers verbs, but it is not always the case: кто рискует и действует, кто занимается созидательным трудом (who risks and acts, who is engaged in creative work).

As mentioned earlier, this qualitative analysis has its complementary quantitative part in the following chapter, where the amount of the phenomena outlined above is specified and explained. However, before moving to the quantitative analysis, I will finish this Chapter by analysing the other two text types apart the political discourses, i.e. TED talks and news.

4.3 Considerations in the other text types: the case of TED talks and news

Since a number of the translation features pointed out in the case of Obama's inaugural speech have been found in the other political speech for which annotation was carried out (Kennedy's inaugural), I will move to the other two text types under analysis: TED talks and news.

For TED talks a consideration is that both Italian and Russian TTs are translations of the video transcripts in English. As a consequence, the features characteristic of the spoken language are often avoided. I do not only mean pauses (that are frequently not reported in the English transcripts either), but to register-related choices (Baker, 2002) (pp. 54-56). These are occasionally also linked to the source language (SL) preferences. For example, *we made the world more cooperative* both the Russian and the Italian translations correspond to "we extended cooperations in the world" (Мы расширили взаимосвязи в мире/abbiamo esteso rapporti di cooperazione); or their translations for *You know in 1950* correspond to "While in 1950" (Если в 1950-м, Mentre nel 1950). Register is an aspect taken into consideration also by Teich (2001) (p. 216), who suggests that sometimes *free* translation is preferred "even if literal translations are often systemically possible".

This sometimes can result also in a non-functional equivalence of concepts as in the case of *than any other single act* чем что-либо еще (than anything else), in which we notice the use of a hyperonym for *act*.

Another case is that of repetitions that do not need to fulfil their function of fixing concepts in the audience's minds when used in texts that most likely serve as support to the video in English for the audience at home. An example is *split the world, tore the world apart, divided the world* rendered in both Russian and Italian as *split, torn and divided the world* (раскололи, разорвали и разделили мир/separò, smembrò e divise il mondo) or *but we're also aware that* simply translated in Russian as но (but). Additions and omissions are also an example: *tragically reminded* becomes напомнило (reminded), *they put a lot of focus and attention* becomes они привлекли внимание (they attracted the attention).

Sometimes such additions are justified by the fact that the SL word is semantically complex (Baker, 2002) (p. 19). For example, *without darkening me* is translated as не оставляет меня в темноте (doesn't leave me in the dark) or *haunts me* is translated as не дает мне покоя (gives me no peace).

Like in the political domain, there are clichés that have been translated as cultural substitution (ibid., p. 29), such as *the best is yet to come*, for which the Russian TT could have chosen, like Italian, a functional equivalent (лучшее еще впереди, *il meglio deve ancora venire*) instead of лучшее уже на подходе (the best is yet on the way), whose occurrence is low on Google and only one in the Russian National Corpus.

Another interesting case is that of the fixed expression *bright-eyed and bushy-tailed* that means “very energetic”. The SL concept is not lexicalized in the target language (TL) (Baker, 2002) (p. 18), so in Russian a literal translation is preferred: с блеском в глазах (with brightness in the eyes) and у нас хвост трубой (we have tails up). In the Italian TT we find a translation by paraphrasis using unrelated words (ibid., pp.38-41, 80-84): *siamo impazienti* (we are impatient).

However, linguistic preferences are often respected, for example preference for nominal phrases in Russian: *to produce new ideas* translated as к созданию новых идей (creation of new ideas), *reasons why I'm optimistic* translated as причины моего оптимизма (reasons for my optimism), *those reactions have caused change to happen* translated as эти реакции были причиной перемен (cause of change).

In the case of news, I chose to have the original texts in either Russian or Italian with English translations. It was interesting to notice that the English translations of Russian and Italian STs are usually not problematic, apart from minor changes in the word order that can still be considered under the *literal* translation strategy (Teich, 2001). I will mention below the most interesting features for the appraisal analysis:

1. вы ей полостью отдаетесь (you are devoted to it) in English and Italian is equivalent to *you are good at* (in cui si è bravi).

2. “дорогой обратно” (a kind of "road back”) is kept as *a sort of way back* and *passi all'indietro* (steps backward).

3. продолжались постоянные горячие споры (constant passionate arguments were ongoing) is lowered in register and expanded in the English *very hot discussions among those who supported and opposed it were ongoing*, and compressed in the Italian *si assisteva a molte discussioni* (many discussions were ongoing).

4. как вы находите выход (how will you look for a way out) is replaced by the verb *to solve* in the other TTs (how do you solve/come risolverebbe).

5. In the following cases, the TTs increased the force of the sentiment: firstly by adding an adjective to the original во времена Сталинских чисток (during the time of Stalin's purges) *in the darkest days of Stalin's purges*, secondly by choosing more-loaded words for создалось "определенное" мнение (have a certain reputation), which becomes *Russian hackers are very infamous* in English and *sono tristemente noti* (are sadly known) in Italian.

   Similarly, the strategy of using a different word and the quotes is adopted here: называемых специалистов (so-called specialist) *cosiddetto "esperto"*, *"pundits"*.

6. в решение существующей проблемы in English is *to help solve this problem* and in Italian *per contribuire alla soluzione di questo problema* (to contribute to the solution of this problem).

4.4 Concluding remarks

To summarize, I can say that:

- Appreciation followed by judgement are the most assigned *attitudes* across the texts.
- Among the translation strategies used in Italian and Russian with respect to English, *addition* has been extensively used mostly to match the formality of the TL (see the examples provided throughout the sections and in particular in Section 4.2.4, while omission to a less extent (see Section 4.2.4).
- Figurative images and metaphors were mostly kept in both TLs, with preference for different collocations or additions shown in the TTs if available (see Sections 4.2.2, 4.2.3, 4.2.4, 4.3). We might say that, although reasonably feeble because occurring in translations, these findings in contrastive grammar could be expressions of a **cultural preference**. For example, this preference could be expressed through a more subjective and explicit language in Italian (Pounds, 2010) as expression of a different *image of the world* based on emotionality and passionateness (Stubbs, 1996a; Wierzbicka, 1992). While some scholars like Šmelev (2002) see it reflected only on the lexicon and phraseology, others like Rylov

(2003) consider syntax as well. In particular, Rylov (2003) has focused on the word order in Italian and Russian by bringing the hypothesis that "the possibility to present the facts of reality in opposite order (from the predicate to the subject) and vary the position of the remainder items is one of the most important characteristics of the Russian and Italian linguistic image of the world, in so far as it reflects our vision of the situation". However, this last hypothesis on the word order is not substantiated in my corpora.

- The force of appraisal is mostly kept apart from a few cases in which it was diminished or amplified (see Sections 4.2.3 and 5).

- There are also some expressions that, although carrying sentiment in context, at fist glance are not really loaded with it, e.g. "implemented in practise". For this reason, I looked at their collocations to check their semantic prosody in bigger corpora, i.e *UkWac* [1] and a British newspapers corpus [2] for British English, and the *Corpus of Contemporary American English* (COCA) [3] for American English.

  The first one is *implement*, whose most frequent collocates in UKWac are *legislation, reform, programme/program, recommendation*, and in British newspapers and COCA *policy, government, decision*. There are also *efficiency, proposal, measure*, but with very low frequency. Due to the lack of negative collocates, I would judge *implement* as positive.

  The second is *common*. Across the corpora, many collocates are similar, although some with different frequency (e.g. *language, factor, problem, sense, denominator, ground, cause, theme, knowledge, law, currency, experience, interest*). However, the British newspaper corpus reports some specific ones such as *practise, consent, thread, misconception, complaint*, of which some shared with UkWac. In general, the prosody seems neutral with a tendency towards the negative.

  The last one is *important*. Many collocates are shared by the different corpora (e.g. *thing, issue, role, factor, aspect, element, question, step, contribution, decision*), although some such as *discovery, game, event* are only present in newspapers and COCA. I would say that, since *important* accompanies neutral and positive collocates, it has a positive prosody.

[1] http://www.sketchengine.co.uk/
[2] http://corpus.leeds.ac.uk/itweb/htdocs/Query.html#
[3] http://corpus.byu.edu/coca/

These corpus-based findings show once more that in sentiment analysis one cannot avoid to look at the collocations rather than at individual words. On the other hand, we have also started to qualitatively explore the differences and similarities across the languages and the different text types especially in the light of the categories of the Appraisal Framework, which is something that will be supported by a quantitative analysis in Chapter 6. Only at that point, the research question related to whether some features are generalizable will find an answer.

Chapter 5

Manual annotation: principles

After looking in Chapter 4 at the differences and similarities in the texts under analysis from different perspectives, especially according to the categories of the attitude subsystem of the Appraisal Framework, in this Chapter I will clarify the value of each of these categories in the annotation. I will start by looking at the works that have led to the creation of the *SentiML* annotation scheme, and then move to its detailed description with the support of real cases.

5.1 The SentiML annotation scheme: the foundations

As anticipated in Chapter 1, the SentiML annotation scheme had to respond to the primary requirement of producing machine-readable annotated texts multilingually, from which linguistic features aimed at an accurate analysis of sentiment could be extracted.

Among the works that dealt with this task by including features belonging to the Appraisal Framework, there are those by Neviarouskaya & Aono (2012); Neviarouskaya *et al.* (2010); Read *et al.* (2007a) in English, Zagibalov *et al.* (2010) in English and Russian, Carretero & Taboada (2014); Taboada & Carretero (2012) in English and Spanish on reviews of different types (e.g. movies, products and services, books). The primary ones taken into account for the creation of SentiML have been Bloom & Argamon (2009); Bloom *et al.* (2007a) (preceded by Whitelaw *et al.* (2005) and Argamon *et al.* (2007)) because they specified clear boundaries and components of sentiment expressions, as well as a full set of relevant features for a software system.

As for the boundaries and components, my work is based on the definition of ap-

$$\underset{\text{``happy''}}{\begin{pmatrix} \text{Attitude:} & \text{affect} \\ \text{Orientation:} & \text{positive} \\ \text{Force:} & \text{neutral} \\ \text{Focus:} & \text{neutral} \\ \text{Polarity:} & \text{unmarked} \end{pmatrix}} \Rightarrow \underset{\text{``very happy''}}{\begin{pmatrix} \text{Attitude:} & \text{affect} \\ \text{Orientation:} & \text{positive} \\ \textbf{Force:} & \textbf{high} \\ \text{Focus:} & \text{neutral} \\ \text{Polarity:} & \text{unmarked} \end{pmatrix}} \Rightarrow \underset{\text{``not very happy''}}{\begin{pmatrix} \text{Attitude:} & \text{affect} \\ \textbf{Orientation:} & \textbf{negative} \\ \textbf{Force:} & \textbf{low} \\ \text{Focus:} & \text{neutral} \\ \textbf{Polarity:} & \textbf{marked} \end{pmatrix}}$$

Figure 5.1: Analysis of the appraisal group "not very happy" taken from Whitelaw *et al.* (2005).

praisal expression given by Bloom *et al.* (2007a) and Bloom & Argamon (2009), i.e. an expression comprising a *source*, an *attitude*, and a *target*. For example, in "I found the movie quite monotonous", the speaker (the source) expresses a negative attitude ("quite monotonous") towards "the movie" (the target).

This definition differs from their initial one in Whitelaw *et al.* (2005), where groups consisted only of modified adjectives such as "not very happy" (see Figure 5.1).

The following set of features of the appraisal groups was also specified:

Attitude type, the type of appraisal being expressed ('affect', 'appreciation', or 'judgement').

Orientation, whether the attitude is 'positive' (e.g. *good*) or 'negative' (e.g. *bad*).

Force, the intensity of the appraisal. Force is largely expressed via modifiers such as *very* (increased force), or *slightly* (decreased force), but may also be expressed lexically, for example *greatest* vs. *great* vs. *good*. This category corresponds to some aspects covered by Graduation and Engagement in the AF.

Polarity, if the appraisal is scoped in a polarity marker (such as *not*), or unmarked otherwise. The appraisal is affected by negation, for example "not good" has the opposite orientation of *good*. This category is covered by engagement.

Target type, a domain-dependent semantic type for the target. This attribute takes on values from a domain-dependent taxonomy, representing important (and easily extractable) distinctions between targets in the domain.

These find a match with the aspects named by (Hunston & Sinclair, 2000) (pp. 14-25) as those important to identify evaluation: *lexis* and *grammar*. As for the first, the very clearly evaluative lexical items specified are adjectives, adverbs, nouns and verbs, while for the latter the list includes intensifiers, comparators (negatives, futures, modals, questions, superlatives and comparatives) and hedges as vague language (e.g. *sort of, about*).

Bloom & Argamon (2009) also establishes an explicit point of contact to the evaluative groups containing adjectives related to the *local grammar* identified in Hunston (2000):

1. It + link verb+ adjective group + clause (as in “It was certain that he was much to blame”).

2. THERE + link verb + SOMETHING/ANYTHING/NOTHING + adjective group + ABOUT/IN + noun groups/-ing clause (as in “There is nothing sacrosanct about this unit of analysis”).

3. Link verb + adjective group + *to*-infinitive clause (as in “Horses are pretty to look at”).

4. Link verb + adjective group + *that*-clause (as in “I’m fairly certain that he is an American”).

5. Pseudo-clefts (as in “What’s very good about this plays is...”).

6. Patterns with general nouns (as in “The surprising thing about chess is that computers can play it so well”).

Bloom *et al.* (2007a) built a system for extracting attitude and target only in adjectival appraisal expressions such as “The Matrix is a good movie”. This was done on the basis of a hand-built lexicon, dependency parsing and a word-sense disambiguation module. While on this work they focussed on adjectival expressions only, in Bloom & Argamon (2009) they included nominal appraisal groups.

Argamon *et al.* (2007) focused on attitude type and force by proposing a method for their automatic system by applying supervised learning to WordNet glosses (I will explain in Chapter 7 that WordNet has also been used in my work, although by relying on rules rather than on supervised algorithms).

The SentiML scheme has been designed to unify the above mentioned works by being applicable to appraisal groups in different languages. We have seen in Chapter 1 that appraisal groups consist of a *target* as the expression the sentiment refers to, and a *modifier* as the expression conveying the sentiment. For example in the sentence “The chief is not just angry, he is scared” the target is *chief* and the modifiers are *angry* and *scared*, and the appraisal groups are “chief angry” and “chief scared”.

Appraisal groups consist of grammatical categories other than adjectives (i.e. nouns, verbs, adverbs and pronouns) following the combinations that I will describe in Section 5.5. I will explain the advantages of the scheme in the following Section, and then move to the detailed description of its categories and the rules to annotate them.

5.2 What are the advantages of the SentiML corpus annotation?

SentiML has been designed as an XML-based scheme and MAE (Stubbs, 2011), a freely available multi-platform annotation environment, has been used to implement the annotation. The scheme is designed to allow fast multi-layer annotation of appraisal groups once targets and modifiers have been annotated (see Figure 5.2).

Table 5.1 shows the annotation of the expression “They definitely sparked outrage”. The expression contains: (I) one appraisal group (“sparked outrage”) with contextual orientation ‘negative’, (II) one modifier (*sparked*) with attitude ‘judgement’, out-of-context orientation ‘ambiguous’, force ‘high’ (because of the presence of *definitely*) and polarity ‘unmarked’ (because the verb *sparked* is in its affirmative non negated form), (III) one target (*outrage*) with type ‘thing’ and out-of-contet orientation ‘negative’.

Category	**Expression**	**Orientation**	**Type**	**Attitude**	**Force**	**Polarity**
Modifier	sparked	Ambiguous	n/a	Judgement	High	Unmarked
Target	outrage	Negative	Thing	n/a	n/a	n/a
Appraisal group	sparked outrage	Negative	n/a	n/a	n/a	n/a

Table 5.1: Appraisal group derived from the sentence “They definitely sparked outrage”

The XML output of the expression “They definitely sparked outrage” is then shown in Figure 5.3.

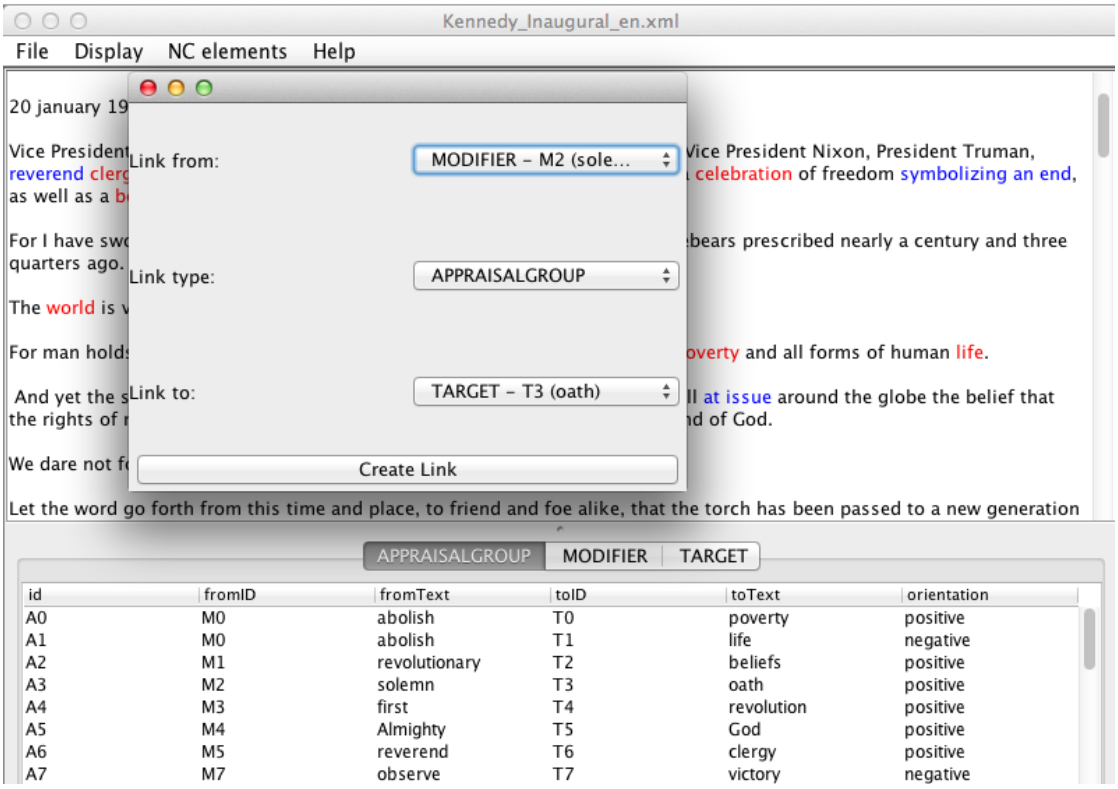

Figure 5.2: Creation of Appraisal Groups in MAE

```
<?xml version="1.0" encoding="UTF-8" ?>
<AppraisalAnnotation>
  <TEXT><![CDATA[
    They definitely sparked outrage.
  ]]></TEXT>
  <TAGS>
    <APPRAISALGROUP id="A0" fromID="M0" fromText="sparked" toID="T0"
     toText="outrage" orientation="negative" />
    <MODIFIER id="M0" start="21" end="28" text="sparked"
     attitude="judgement" orientation="ambiguous"
     force="high" polarity="unmarked" />
    <TARGET id="T0" start="29" end="36" text="outrage" type="thing"
             orientation="negative" />
  </TAGS>
</AppraisalAnnotation>
```

Figure 5.3: XML output of the sentence “They definitely sparked outrage” annotated in SentiML format.

The choice of using pairs (i.e. usually one word for target and one for modifier) rather than more complex expressions is motivated by the nature of my annotation, which is "problem-oriented" (according to McEnery *et al.* (2006) (p.43)) because it is aimed at gaining insights into the research question "What are the linguistic features of evaluative language that can lead to a successful automatic analysis of sentiment across multiple languages?". In fact, while the annotation of appraisal groups consisting of only *one* target and *one* modifier seems rather limited from a linguistic point of view, it turns out to be quite reliable and inclusive from a computational one. I will show some concrete examples of how far these limitations go in Section 5.6, but in the meanwhile some examples will start clarifying the usefulness of using pairs. Examples have been taken from Read *et al.* (2007a) to show the possibility of an overlap with a different scheme that uses more detailed AF categories, but does not specify an annotation span. The first example is:

> It is tempting to point to the bombs in London and elsewhere, to the hideous mess - QUALITY in Iraq, to recent victories of Islamists, to the violent and polarised rhetoric - PROPRIETY and answer yes.

In this example, the SentiML appraisal groups matching the last annotated expression would be "violent rhetoric" and "polarised rhetoric". These would then cover the original annotation "violent and polarised rhetoric" making the span consistent with "hideous mess" (with obvious advantages for the manual and automatic annotation), as well as allow both groups to be assigned the value PROPRIETY.

The second example is:

> The design was deceptively simple – COMPLEXITY.

As the authors explain in their article, this sentence is aimed at appreciating the simplicity of the design, so annotating single tokens such as *deceptively* and *simple* would be wrong. They use this example to show how necessary it is to annotate larger units of appraisal-bearing language. SentiML would go even further by allowing the annotation of the target of the opinion as "design simple" but with reverse force due to the presence of *deceptively*. SentiML also covers the case of multiple subjects such as in the sentence "Mario and Lucia are nice" by splitting them in the two groups "Mario nice" and "Lucia nice". This is possible even in inflected languages such as Italian and Russian, although the plural adjective then is associated to a singular noun, e.g "Mario e Lucia sono bravi (pl.)" -> "Mario bravi (pl.)". We will see in Chapter 7 that, while this might create some confusion in the manual annotation, it does not represent an issue in the automatic matching of targets and modifiers thanks to the dependency parsing relations. Dependency parsing relations are also useful in the case in which the agreement is in the verb, which in English happens in the third person singular of the present tense such as in "Mario and Lucia sparkle" -> "Mario sparkle", while in Italian and Russian for all persons and tenses.

As for adverbs, I will explain in Section 5.4.1 that SentiML is flexible because in selected cases it allows adverbs to be included in the appraisal groups.

The third example is:

> (version 1) Like him, Vermeer – or so he chose to believe– was an artist neglected – SATISFACTION and wronged – SATISFACTION by critics and who had died an almost unknown.
>
> (version 2) Like him, Vermeer – or so he chose to believe – was an artist neglected and wronged – PROPRIETY by critics and who had died an almost unknown.

In this example, even if the annotation of more detailed AF categories was considered, the strict annotation span allowed by SentiML would have avoided the possibility of choosing. SentiML would have allowed, in fact, only an annotation similar to version 1 (i.e. "artist neglected" and "artist wronged"), which is considered by the annotator as the most correct one because it manages to reflect the artist's dissatisfaction with the way he is treated by critics. Even if one might object that the annotation provided in version 2 is more accurate because the critics are being reproached for their treatment of the artist (a point raised by the authors), at least SentiML would have avoided subjectivity in choosing the span, which according to Read *et al.* (2007a) themselves is a problem with their coding scheme. Subjectivity could have been relegated to the next level, the attitude selection (e.g. appreciation vs. judgement, SATISFACTION vs. PROPRIETY). It must be mentioned that another corpus annotated with a larger span is *MPQA* by Wilson (2008). However, the MPQA does not make any distinction between the AF attitude types (i.e. affect, judgement and appreciation) because it considers the general category *sentiment*.

As mentioned in Section 5.1, with respect to the previous works, the scheme is flexible enough not to limit appraisal groups to adjectives, but to include several combinations of nouns, pronouns, verbs and adverbs. For example: “children love” (noun and verb), “perfectly manage” (adverb and verb), “experience crisis” (verb and noun) (see Section 5.2.3 for the complete list of combinations).

In addition, both the scheme and the annotation tool cover the case in which the same word needs to be annotated as modifier in a group and as target in another, for example *peace* in the sentence “the instruments of war have far outpaced the instruments of peace”. *Peace* is modifier in the appraisal group “peace instruments” (positive) and target in the appraisal group “outpaced peace” (negative) (see Figure 5.4).

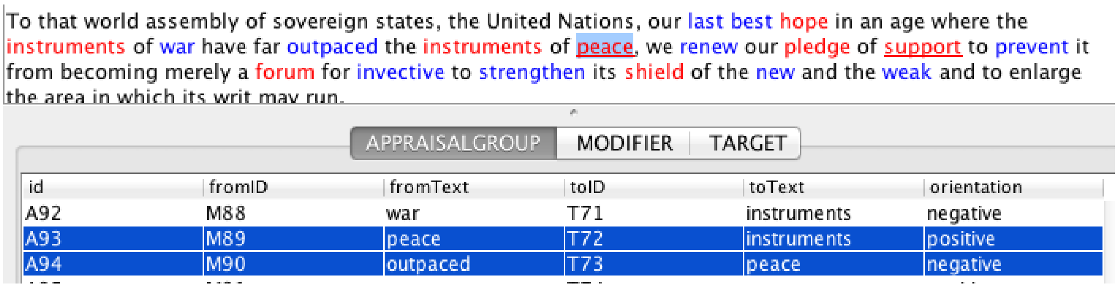

Figure 5.4: Word *peace* annotated both as modifier and target in MAE.

Apart from its inclusiveness, SentiML has been designed to be used in a variety of contexts. First and foremost, to be applied to different languages in a way that “successive studies can be compared and contrasted on a common basis” (McEnery *et al.*, 2006) (p.30). This is true both for future studies in Sentiment analysis, but also in the light of the fact that corpus annotation “considerably extends the range of research questions that a corpus can readily address” (McEnery *et al.*, 2006) (p. 29). In the case of the SentiML annotated corpus, all the text types included (political speeches, news and TED talks) are prone to this *reusability*. In particular TED talks, whose continuous growth in number and topics, variety of formats (videos and transcripts) and availability of translations in other languages, represent a new and exciting resource for studies with different focus or studies with similar focus but different methodology. Moreover, considering that both Italian and Russian are less resourced than English, *reusability* and *multifunctionality* (McEnery *et al.*, 2006) (p.30) are advantages that definitely apply to these languages. By *reusability* I mean that selected annotations could be used, for example, for a study in Sentiment Analysis based on a different methodology. By *mul-*

tifunctionality I mean that the annotated corpora could be used for different purposes, such as language teaching (either of an individual language or as basis for contrastive studies).

Another important aspect is that my corpora are easily expandable with other documents. Interesting additions would be represented by translations in a different direction, i.e. Italian and Russian into English.

As for the challenges of corpus annotation, subjectivity is certainly one of them, especially in the case of sentiment annotation. In fact, even if I have explained that this has been avoided at the level of the boundaries (span) of the appraisal groups, this is not easily solvable in two cases: first of all in the identification of the groups to be annotated and, afterwards, in the values assigned to the layers of the annotations. Particularly challenging are the target and modifiers' out-of-context orientation, which has been included to allow a comparison with the contextual one (see Section 6.2.2), and also the attitude type according to the AF. In all cases, I must stress that the speaker or the writer's perspective at the moment in which the opinions were stated has been taken into account as much as possible. In addition, in the case of the orientation layer (commonly called *polarity*), I deliberately decided not to treat the system as inherently probabilistic (e.g. positive 0.9/negative 0.1) as suggested by Halliday (1992) in order to avoid a further level of difficulty. As for the other features, i.e. target type, force, polarity and contextual orientation, I would dare to say that annotations are pretty standard and no big differences were to be expected with other potential annotators.

As I will explain in Chapter 8, the solutions to the subjectivity issue that I have adopted (apart from considering the speaker's or the writer's perspective), have been revising the annotations by starting from the plain texts as suggested in McEnery (2005). These have been available throughout the project because the annotation tool MAE keeps the annotations separate from the content of the input documents, by using stand-off annotations in accordance with the *ISO Linguistic Annotation Framework* (Ide & Romary, 2004) as shown in Figure 5.5. Another solution has been to compare my manual annotations to the predictions of automatic classifiers to find inconsistencies. Finally, the analysis done by Munday (2012) of Obama's inaugural speech has also been very useful to clarify a few ambiguous cases.

In Section 5.4 I will also show that the practical one of using an annotation tool that does not allow taking multi-word expressions consisting of words not immediately next to each other, which however does not affect badly the annotation procedure due to the

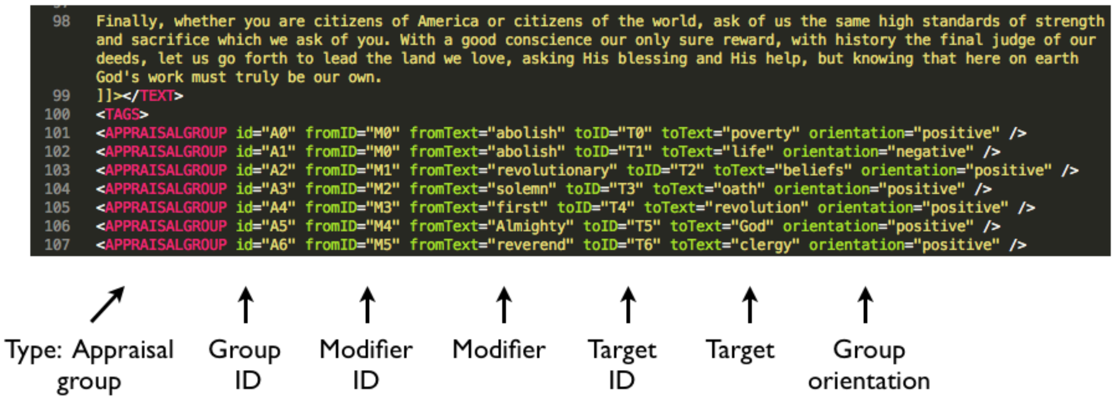

Figure 5.5: Example of SentiML XML annotation output

quite low frequency of the phenomenon.

Since corpus annotation "represents a record of analysis open to scrutiny and criticism" as opposed to leaving the analysis hidden (McEnery *et al.*, 2006) (p.30), the DTD file to be used with MAE, the guidelines and the original and annotated texts have been made publicly available since the beginning[1].

In the following sections the three categories **targets, modifiers** and **appraisal groups**, and their attributes, will be presented. Table 5.2 shows them at glance, with underlined those assigned as default in MAE, but changeable through the drop-down menu.

Role	Attribute	Possible values
Target	Type	Person, <u>Thing</u>, Place, Other
	Orientation	<u>Neutral</u>, Positive, Negative, Ambiguous
Modifier	Orientation	<u>Neutral</u>, Positive, Negative, Ambiguous
	Attitude	<u>Affect</u>, Judgement, Appreciation
	Force	<u>Normal</u>, High, Low, Reverse
	Polarity	<u>Unmarked</u>, Marked
Appraisal group	Orientation	<u>Neutral</u>, Positive, Negative, Ambiguous

Table 5.2: Attributes for target, modifier and appraisal group along with all possible values and default ones underlined.

[1] `http://corpus.leeds.ac.uk/marilena/SentiML`

5.2.1 Targets

A target is any entity (object, person or concept) that is implicitly or explicitly regarded as positive or negative by the author of the text. The following scenarios can be found:

1. One target with one feature. For example, in "This article is useless", there is one target (*article*) with a feature (*useless*).

2. Different targets with their features. For example, in "I bought a good camera from that website, but a very bad phone", there are two targets (*camera* and *phone*) and their related features (*good* and *bad*).

3. Different targets with their features referring to one entity. For example, in "The camera has a good quality, but a considerable weight", the two targets (*quality* and *weight*) have got their own features (*good* and *considerable* respectively), but they refer to the same object (*camera*).

In terms of its logical function, a target can be either a subject or an object, depending on which carries the sentiment (e.g. in "we share beliefs", only *beliefs*). From a grammatical point of view, it can be a:

- **Common noun**. By far, the most common category (e.g. children, cat).
- **Proper noun**. Sub-categories include person (e.g. Olivia), place (e.g. China), animal (e.g. Sparky). In multi-word expressions such as "the United States of America", annotating *America* is enough, whereas *U.S.* and *United States* are annotated as they are.
- **Verb**. Usually when the modifier is an adverb (e.g. talk quickly).

5.2.2 Modifiers

A modifier is what *modifies* the target. It can be:

- **Adjective**. When the target is a noun, for example "[beautiful]$_{\mathrm{M}}$ [car]$_{\mathrm{T}}$[1]". Apart from qualifying adjective, we can also have possessive adjectives (e.g. "[our]$_{\mathrm{M}}$ [star]$_{\mathrm{T}}$").

[1] From now onwards, the notation [word]$_{\mathrm{M}}$ will be used for modifiers, and [word]$_{\mathrm{T}}$ for targets.

- **Verb**. When the target is a noun, for example "[obtain]$_{\mathrm{M}}$ [victory]$_{\mathrm{T}}$ " or "[victory]$_{\mathrm{T}}$ [obtained]$_{\mathrm{M}}$ ".
- **Adverb**. When the target is a verb and the adverb is too important to be implicit, for example "[foolishly]$_{\mathrm{M}}$ [sought]$_{\mathrm{T}}$" (more details will be given in Section 5.4.1).
- **Noun**. When the target is another noun and they are linked by a preposition, for example "[alliance]$_{\mathrm{T}}$ for [progress]$_{\mathrm{M}}$" which, for the purpose of the annotation, is equivalent to say "progress alliance".

5.2.3 Appraisal groups

An appraisal group represents an opinion on a specific target. For this reason, it is defined as the link between the target and the modifier. It has to match one of the following combinations:

- **A noun with an adjective**. For example, in English [good]$_{\mathrm{M}}$ [plan]$_{\mathrm{T}}$, in Russian огромную важность (huge importance), and in Italian *decisioni spiacevoli* (unpleasant decisions).
- **A pronoun with an adjective**. For example, in English [they]$_{\mathrm{T}}$ (are) [beautiful]$_{\mathrm{M}}$, in Russian Мы здоровы (we (feel) great), and in Italian *lui (e') folle* (he is a fool).
- **A pronoun and a verb**. For example, in English [I]$_{\mathrm{T}}$ [like]$_{\mathrm{M}}$, in Russian мы пожелаем (we wish), and in Italian *noi accogliamo* (we embrace).
- **A noun with a noun when linked by prepositions**. For example, in English [stigmatization]$_{\mathrm{T}}$ of [people]$_{\mathrm{M}}$, in Russian выходные с пользой (benefits from holiday) and in Italian *libertà di parola* (freedom of speech). The usual prepositions are *of*, *for*, *in*, *against*, *with*, *towards*, *between*.
- **A verb with an adverb**. For example, in English [strongly]$_{\mathrm{M}}$ [support]$_{\mathrm{T}}$, in Russian прекрасно справляется (perfectly manage), and in Italian *giustamente indicato* (rightly marked).
- **A noun with a verb**. For example, in English [children]$_{\mathrm{T}}$ [love]$_{\mathrm{M}}$, in Russian переживаем кризис (we experience a crisis), and in Italian *illuminare il mondo* (to light the world).

5.3 Annotating targets

I will now give more suggestions on how to carry out the annotation task of the different categories. As for targets, an important thing to bear in mind is that the target tag should be assigned to one word per time matching the grammatical categories seen in Section 5.2.1. Figure 5.6 shows an example of annotation in MAE.

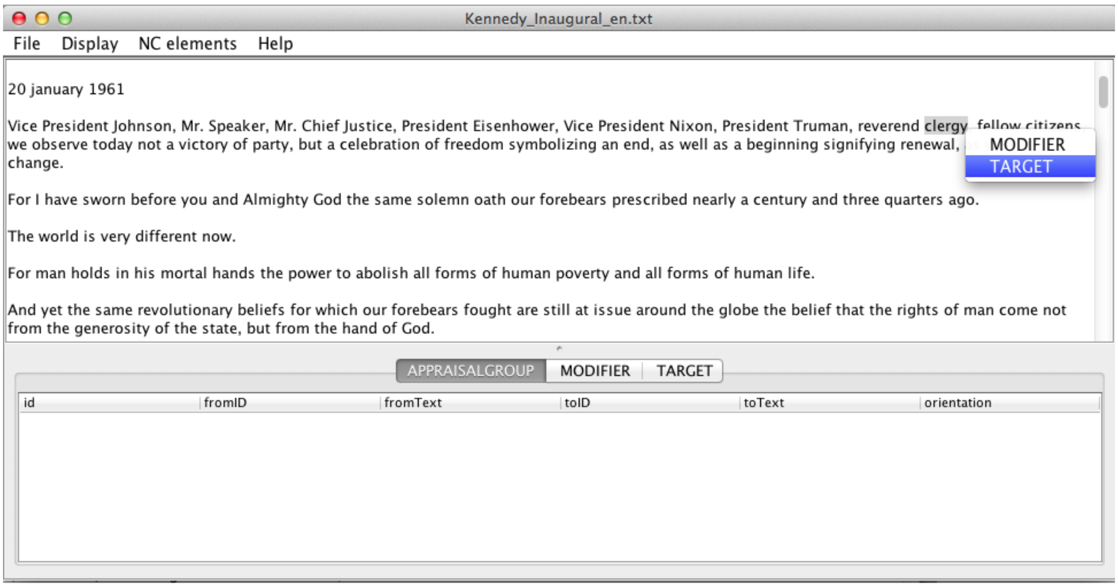

Figure 5.6: Target annotation in MAE according to SentiML

When two targets have the same modifier(s), for example in the expression "Luke and Isabelle are nice", the groups are "[Luke]$_{T}$ [nice]$_{M}$" and "[Isabelle]$_{T}$ [nice]$_{M}$".

The cases in which the tag must not be used are:

- **Non-sentiment words.** We do not annotate words that carry no sentiment in a given context, even where they feature in a sentiment dictionary. For example:
  - "Worldly possessions" in *"They packed up their few worldly possessions and traveled across oceans in search of a new life"*.

  In any case, **the author's perspective (as far as it can be interpreted from the outside) is what the annotations are based on**. For example in:

  Samuel Pisar, an *Auschwitz survivor* said

"Auschwitz survivor" has the only scope of presenting the person in the context of the talk, so we cannot really assign a positive or negative connotation to *survivor* or *Aushwitz*.

When the author's opinion is retrievable, despite being not expressed explicitly like in following examples, the text should be annotated:

- "*People* with disease are stigmatized". In this case, the negative opinion is on the fact that this group of people is badly treated, so we need to annotate "people" as part of the group "$[\text{people}]_{\text{T}}$ $[\text{stigmatized}]_{\text{M}}$", but not the group "disease people".
- "The *Holocaust* teaches us that nature, even in its cruelest moments, is benign in comparison with man, when he loses his moral compass and his reason". In this case, *Holocaust* is implicitly negative.
- "In the 1960s and 1970s, the *Vietnam War* was basically shown in America's living rooms day in, day out". In this case the fact that a negative event was constantly shown is considered negative itself. So we annotate the target *war* and include it into two groups "$[\text{Vietnam}]_{\text{M}}$ $[\text{war}]_{\text{T}}$" and $[\text{war}]_{\text{T}}$ $[\text{shown}]_{\text{M}}$".
- "This is practically the *model* of TED". Because this statement was made during a TED presentation, we understand that "$[\text{TED}]_{\text{M}}$ $[\text{model}]_{\text{T}}$" has positive connotation.
- "The *demand for pharmaceuticals* is going to increase tremendously". Intuitively (and confirmed by the broader context of the sentence) we need to link the target *demand* to the two negative groups "$[\text{pharmaceuticals}]_{\text{M}}$ $[\text{demand}]_{\text{T}}$" and "$[\text{demand}]_{\text{T}}$ $[\text{increase}]_{\text{M}}$".
- "*Artists and innovators*, many of the people you've seen on this stage". Because of the implicit admiration towards *artists* and *innovators*, we annotate these two targets.
- "We will see an *Einstein* in Africa in this century". Because of the implicit admiration towards *Einstein*, we annotate it.
- "Is China drinking *our milkshake*?" In this case, the target "milkshake" has to be annotated as part of the negative group "$[\text{drinking}]_{\text{M}}$ $[\text{milkshake}]_{\text{T}}$".

- **Actions that are simply statements of facts** (although including them might be already a marked choice by the speaker). For example:
  - "The leader rejected the report".
  - "Where the answer is yes, we intend to move forward. Where the answer is no, programs will end".
  - "Those of us who manage the public's dollars will be held to account".

5.3.1 Attributes of targets

The SentiML annotation scheme allows the annotation of attributes for each category. Targets have two attributes:

Type. This attribute captures the type of target and has five possible values: 'person', 'thing', 'place', 'action' and 'other'. Animals are included in the category 'thing'. Countries, cities, provinces and natural geographical points (e.g. rivers, lakes, mountains) are usually annotated as 'place', whereas *world* can be either 'thing' or 'place', depending on whether it carries the action or not. The 'other' value is only used when an adjective is marked as target (e.g. "[easily]$_{\mathrm{M}}$ [imaginable]$_{\mathrm{T}}$").

Orientation. This attribute captures the prior orientation of a target and it has four possible values: 'positive', 'negative', 'neutral' and 'ambiguous'. For example, *peace* is positive regardless of the context, whereas *pessimism* is negative regardless of the context. The 'ambiguous' value is given if the orientation depends on the context (e.g. the word *challenge* in "promising challenge" and "unfair challenge"). In this case, the appropriate orientation is annotated in the appraisal group (e.g. "promising challenge" is marked as positive and "unfair challenge" is marked as negative). Quite common is also the case in which words do not seem ambiguous at first (for example *growth* can give the idea of being always positive, but it is actually negative when modified by *slow*, or in medicine in the case of "tumor growth"). The 'neutral' value is assigned to targets that have no connotation, very often personal pronouns (e.g. we) and places (e.g. America).

5.4 Annotating modifiers

The modifier tag usually corresponds to one word matching the grammatical categories in Section 5.2.2. By single word I also mean hyphenated ones such as *low-budget*.

However, there are special cases in which one word does not carry the entire sentiment:

- **Phrasal/multi-word verbs**. They are annotated as single tokens and embedded in a group afterwards (e.g. "[cast off]$_{\mathrm{M}}$ [worries]$_{\mathrm{T}}$"). Other examples are "get rich", "put forward", "worried about".

  Multi-word verbs such as "symbolizing end", "signifying renewal", "proclaim an end" are included.

  Because MAE allows annotating only words close to each other, in case the phrasal verb is split (e.g. "*carried* us *up*", "*tore* the world *apart*") but it still conveys the meaning, either the verb (e.g. carried) or the preposition (e.g. apart) can be annotated.

  Another example is "pull ourselves out of this abyss" in which the group "[out]$_{\mathrm{M}}$ [abyss]$_{\mathrm{T}}$" has been annotated.

- **Multi-word expressions**. These are annotated as single tokens (e.g. "at issue", "at odds", "in practice", "out of control", "under pressure", "in the light", "go off a cliff").

- **Modified modifiers**. Modifiers modified by adverbs of *degree* (e.g. significantly, slightly) or *affirmation* (e.g. absolutely, seemingly). For example:

  – "Not very happy", in which we keep the information related to *not* and *happy* as marked polarity and high force respectively (see Section 5.4.1).

  – "Far more iron", in which "far more" is considered high force of *iron*.

  – "Far worse destructive power", in which we annotate two groups "[far worse]$_{\mathrm{M}}$ [power]$_{\mathrm{T}}$" and "[destructive]$_{\mathrm{M}}$ [power]$_{\mathrm{T}}$" because *worse* is a high-force modifier by itself.

  – "Strongly supporting", in which *supporting* has high force because of *strongly*.

  – "Too firmly focused", in which the modifier *focused* is double modified.

The cases in which the modifier tag must not be used are the following:

- **Auxiliary verbs**
  - *To be* when followed by a non-finite verb (e.g. "it is working", "it was chosen").
  - *To have* when followed by a past participle (e.g. "crisis has reminded", "work has to be done").
  - *To do* when followed by an infinitive (e.g. "I don't like", "you did know").
  - *Used to, to dare, need* (e.g. "we dare not [meet]$_{\mathrm{M}}$ a powerful [challenge]$_{\mathrm{T}}$").

  This category also includes **modal** verbs:

  - *Can/could* (e.g. "he can really sing").
  - *May/might* (e.g. "that may be a problem").
  - *Must* (e.g. "we must change").
  - *Will/would* (e.g, "we will do").
  - *Shall/should* (e.g. "you should stop that").
  - *Ought* (e.g, "they ought to respect the law").
- **Catenative verbs**
  - *To get* (e.g. "she got chosen for the job").
  - *To keep* (e.g. "she kept disturbing, to keep our [education]$_{\mathrm{T}}$ system [globalized]$_{\mathrm{M}}$").
  - *To start/begin* (e.g. "she started to blame").
  - *To help* (e.g. "she helped improving").

    In the case in which two types of verbs follow each other, a preference is given to the one carrying the sentiment. For example:

    * "Trade [walls]$_{\mathrm{T}}$ began to come [tumbling down]$_{\mathrm{M}}$"
    * "Have [caused]$_{\mathrm{M}}$ [change]$_{\mathrm{T}}$ to happen"
- **Copulas/Operators**
  - *To be* followed by a noun (e.g. "heritage is a strength") or an adjective (e.g. "she is ready")

  - *To have* in the sense of possess (e.g. "she has money")

- **Words not placed next to each other**. For example *never* and *before* in the expression "never seen our planet from this perspective before".

  No annotation should be done also in the case of phrasal verbs whose meaning is ruined when the second part is split (e.g. "pick ourselves up", "dust ourselves off").

- **Modifier not linked to anything in the sentence**. For example in "The common disease or the rare disease? Common.", the modifier *common* makes a sentence by itself.

5.4.1 Attributes of modifiers

Modifiers have four attributes:

Orientation. This attribute refers to the *prior orientation* of a modifier and it has four possible values: 'positive', 'negative', 'neutral' and 'ambiguous'. For example *beautiful* is positive regardless of the context, whereas *horrid* is negative regardless of the context. The 'ambiguous' value is used when the orientation depends on the context (e.g. for the verb *wishes* in the cases "wishes well" and "wishes ill"). Sometimes the same word takes a different orientation according to their grammatical category (e.g. *light* used as modifier in "light the world" is positive, whereas *light* used as target in "give light" is ambiguous).

Attitude. According to the AF, attitude has three possible values: 'affect', 'judgement' and 'appreciation'. The 'affect' value is used for personal states (e.g. "I'm optimistic") and opinions (e.g. "if we are divided, we won't achieve much"). The 'judgement' value is used for others' behaviour (e.g. "children are unwilling to obey"), whereas the 'appreciation' value is used for the evaluation of things (e.g. "solemn oath").

Because any grammatical category (noun, verb, preposition, adverb) can fall under the *modifier* label, we choose appreciation, judgement and affect according to what such word refers to (e.g. "out of control" would be classified as appreciation if linked to an action or to judgement if linked to a person). For the same reasoning, those words or expressions that involve a human participation or response such as *crisis, freedom, justice, courage, determination, violence, corruption, long rugged path* (including metaphors such as "coldest of months") should be annotated as judgement, rather than appreciation. At the same time, those connected to personal circumstances such as "in the midst of crisis", "at war", "our enduring spirit", "our capacity" should be annotated as affect, rather than appreciation.

Polarity. This attribute captures the information linked to the presence of a negation. Polarity refers to the positive and negative poles in the mood system - the realization of tenor at the clause level (see Section 3.1).

It has two possible values: 'marked' and 'unmarked'. It is 'marked' when there is a negation (e.g. "we do not observe a victory"), 'unmarked' otherwise (e.g. "we like this place") . Apart from being local (e.g. "not good"), negation can also involve long-distance dependencies (e.g. "does not look very good") or the subject (e.g. "no one thinks that it is good", "nothing so satisfying"). As previously seen, auxiliary verbs give the polarity (e.g. "system cannot tolerate"), but they are not annotated.

Force. This attribute refers to the intensity of the modifier. It has four possible values: 'high', 'low', 'normal' and 'reverse'. In SFL, it corresponds to the circumstance 'degree' (see Section 3.3). The value 'normal' represents the standard input and is used when the modifier is not modified by anything (e.g. *good)*.

The 'high' value is used if the modifier is preceded by adverbs of high intensity such as *very* and *extremely* (e.g. "very good", "extremely good"), if it is included in expressions such as "not only ..., but ..." (e.g. "not only good, but amazing") or if it expresses high intensity itself (e.g. *best* as opposed to *good*).

The value 'low' is used if the modifier is preceded by adverbs of low intensity such as *less*, *little* and *poorly* (e.g. "less good") or it expresses low intensity itself (e.g. *worse* as opposed to *bad*).

The 'reverse' value is used in presence of words called *reversals* because they reverse the prior orientation of their targets (e.g. the verb *abolish* in "abolish taxes"). Apart from verbs (e.g. *to decrease*, *to limit*, *to diminish*, *to remove*), reversals can be nouns (e.g. *termination*), prepositions (e.g. *without*, *despite*) and condition operators (e.g. *if*, *even though*).

For example "If the world were as wealthy as the United States now", the group "world wealthy" has reverse force, whereas "United States wealthy" has normal force. Same for "If you'd been a little more optimistic", in which *optimistic* takes reverse force in the group "you optimistic".

An important thing to bear in mind is that force signals are not annotated. For example in "She, better than most, knew the power of an image", *knew* has got high force, because of *better*.

However, in those cases in which the adverb radically modifies the force of the group like in the expression "those who foolishly sought power", a new group is created (foolishly sought), apart from the main one (sought power), provided that it falls into one of the combinations mentioned in Section 5.5. In "foolishly sought", *foolishly* has high force, whereas *sought* has normal force.

Building a new group for them was a decision that I made after trying another annotation style, in which the verb in the main group ($[\text{sought}]_M$ $[\text{power}]_T$) would have the force of the adverb *foolishly*, and realised that no time was saved.

Adverbs that do not need to be included in an appraisal group because they contribute to the *force* usually indicate *intensity*:

- Those expressing low force: *scarcely, merely, somewhat, badly, unsurprisingly, so*.
- Those expressing high force: *extremely, a lot, well, again, more than ever before/ even more than ever before*.

Adverbs that should be put in groups usually indicate mode: *safely, tirelessly, grudgingly, gladly, easily, at least, comprehensively, peacefully, rightly, responsibly*.

5.5 Annotating appraisal groups

The appraisal group tag has to be used with the grammatical categories in Section 5.2.3, which are expanded here with examples in English:

- **A noun with an adjective**. For example "$[\text{wonderful}]_M$ $[\text{example}]_T$". If a target has more than one sentiment-loaded modifier, like in the expression "cultural and spiritual origins", one group for each modifier is created ("$[\text{cultural}]_M$ $[\text{origins}]_T$", "$[\text{spiritual}]_M$ origins_T"). This rule also applies to modifiers not close to the target, for example in the sentence in Figure 5.7 in which several adjectives refer to "Americans": "$[\text{Americans}]_T$ $[\text{tempered}]_M$", "$[\text{Americans}]_T$ $[\text{disciplined}]_M$", "$[\text{Americans}]_T$ $[\text{proud}]_M$" and "$[\text{Americans}]_T$ $[\text{unwilling}]_M$"(their annotation is shown in Figure 5.8).

Let the word go forth from this time and place, to friend and foe alike, that the torch has been passed to a new generation of Americans born in this century, tempered by war, disciplined by a hard and bitter peace, proud of our ancient heritage and unwilling to witness or permit the slow undoing of those human rights to which this Nation has always been committed, and to which we are committed today at home and around the world.

Figure 5.7: Sentence 1, with several modifiers (i.e. *tempered, disciplined, proud, unwilling*) referring to the target *Americans*.

APPRAISALGROUP | MODIFIER | TARGET

id	fromID	fromText	toID	toText	orientation
A14	T13	Americans	M15	tempered	positive
A15	T13	Americans	M14	disciplined	positive
A16	T13	Americans	M16	proud	positive
A17	M17	hard	T14	peace	negative
A19	M18	bitter	T14	peace	negative
A20	M19	ancient	T15	heritage	positive
A18	M20	human	T16	rights	positive
A21	M21	slow	T17	undoing	negative
A22	M22	permit	T17	undoing	negative
A23	T13	Americans	M23	unwilling	positive
A24	T18	Nation	M24	committed	positive
A25	T19	we	M6	committed	positive

Figure 5.8: Annotation of the appraisal groups in Sentence 1.

- **A noun with a verb**. For example “[children]$_{\mathrm{T}}$ [love]$_{\mathrm{M}}$”.
- **A verb with an adverb**. For example “[rightly]$_{\mathrm{M}}$ [marked]$_{\mathrm{T}}$” (see Subsection 5.4.1).
- **Nouns linked by a preposition**: For example the expression “victims of war” is annotated as “[war]$_{\mathrm{M}}$ [victims]$_{\mathrm{T}}$. The usual prepositions are *of, for, in against, with, towards, between*. This category also includes the possessive case such as “man’s power”, “mankind’s war”, “expedience’s sake”, “women’s rights”.
- **Superfluous modifier**. For example:
  - “To set aside childish things”, in which “set aside things” does not have a completely different orientation with respect to “set aside childish things”.
  - “To choose our better history”, in which “choose history” has the same orientation as “choose better history”.
- **Numbers, percentages and related words**. The are annotated only if they carry sentiment for the speaker. For example:
  - “There is twice as much light available for everyone”, with the group“[twice]$_{\mathrm{M}}$ [light]$_{\mathrm{T}}$”.

- "Several hundred documents", with the groups"[several]$_{\mathrm{M}}$ [documents]$_{\mathrm{T}}$" and "[hundred]$_{\mathrm{M}}$ [documents]$_{\mathrm{M}}$".
- "Income less than 1,000 dollars per year", with the group"[incomes]$_{\mathrm{T}}$ [less]$_{\mathrm{M}}$".
- "Less than 1/10th of one percent of the world's population are scientists and engineers", with the group "[less]$_{\mathrm{M}}$ [percent]$_{\mathrm{T}}$". The other ways of annotating this would have been "less population" (wrong subject) or "less scientists/less engineers" (grammatically incorrect).
- "1996, less than one million new university students in China, per year. 2006, over five million", with the groups"[one]$_{\mathrm{M}}$ [million]$_{\mathrm{T}}$" (negative) and "[five]$_{\mathrm{M}}$ [million]$_{\mathrm{T}}$" (positive).
- "Drug treats 1,000 people, 100,000 people, or a million people", with the groups"[drug]$_{\mathrm{T}}$ [treats]$_{\mathrm{M}}$" and "[treat]$_{\mathrm{M}}$ [people]$_{\mathrm{T}}$". In this case, because it does not make much sense to annotate the numbers (1,000, 100,000 and 100,000), we do not annotate the other occurrences of *people* other than the first.
- "Billions of our processors are off-line", with the group"[processors]$_{\mathrm{T}}$ [off-line]$_{\mathrm{M}}$"

The cases in which the appraisal group tag must not be used are:

- **Non-sentiment modifier**. For example, "growth of six percent". Only *growth* is annotated as target.
- **Combinations outside the fixed ones**. For example:
  - "Merely as a struggle" or "beyond doubt" because of the combination "adverb + noun".
  - "Expect to see'' because of the combination "verb + verb".
- **Group assuming a wrong attitude**. For example in the sentence

  "to proclaim an end to false promises"

  the group "[end]$_{\mathrm{M}}$ [promises]$_{\mathrm{T}}$" would be negative, so not equivalent to "end false promises".

- **Group with no sense**. Unlike in "greatest economic disaster" , which can be easily split into "[greatest]$_{\mathrm{M}}$ [disaster]$_{\mathrm{T}}$" and "[economic]$_{\mathrm{M}}$ [disaster]$_{\mathrm{T}}$", in the following cases the groups are not meaningful on their own:
  - In "Full measure of happiness", the groups "measure of happiness" and "full measure" cannot be annotated because they do not make sense without each other.
  - In "First World War", only the group "World War" can be annotated, not "First War".
- **Same group in the same sentence**. For example in the sentence

  "They threw up walls, political walls, trade walls, transportation walls, communication walls, iron curtains"

  only one group "[walls]$_{\mathrm{T}}$ [threw up]$_{\mathrm{M}}$" is created because all the other occurrences (i.e. political walls, trade walls, transportation walls, communication walls) do not convey different meanings when put in a group without their modifiers.
- **Numbers and related words with no sentiment**. As always, we take into account the author's perspective, like in the following case:

  "Even at that rate by 2100, average GDP per capita in the world will be 200,000 dollars".

5.5.1 Attributes of appraisal groups

Appraisal groups have just one attribute:

Orientation. This attribute refers to the *contextual* orientation of the appraisal group, i.e. it considers the context in which the word appears, and it has four possible values: 'positive', 'negative', 'neutral' and 'ambiguous'. For example, "[hungry]$_{\mathrm{M}}$ [minds]$_{\mathrm{T}}$" has a positive contextual orientation, whereas "[hungry]$_{\mathrm{M}}$ [children]$_{\mathrm{T}}$" a negative one.

There are cases in which the **orientation completely depends on the context**. For example:

- "[Hungry]$_{\mathrm{M}}$ [minds]$_{\mathrm{T}}$", a positive appraisal group.

- "Provoke us to step up" with the appraisal group "[provoke]$_{\mathrm{M}}$ [us]$_{\mathrm{M}}$" as ambiguous. Conversely, the group "[us]$_{\mathrm{T}}$ [step up]$_{\mathrm{M}}$ would be positive.

Other cases in which the **prior orientation of a word is changed** when put in a group. For example:

- "All of this has been tremendous for the world" with the appraisal group "[all]$_{\mathrm{T}}$ [tremendous]$_{\mathrm{M}}$" as positive (see Figure 5.9 for the context).

We globalized the world.
And what does that mean?
It means that we extended cooperation across national boundaries.
We made the world more cooperative.
Transportation walls came tumbling down.
You know in 1950 the typical ship carried 5,000 to 10,000 tons worth of goods. Today a container ship can carry 150,000 tons.
It can be manned with a smaller crew, and unloaded faster than ever before.
Communication walls, I don't have to tell you, the internet, have come tumbling down.
And of course the iron curtains, political walls have come tumbling down.
Now all of this has been tremendous for the world.
Trade has increased.

Figure 5.9: Context for the sentence "All of this has been tremendous for the world".

- "Lack of freedom", which is negative although *freedom* is positive *a priori*.

- "Prohibition of terrorism", which is positive although "terrorism" is negative *a priori*.

The 'neutral' value is actually never used during the annotation, as we are interested only in targets which carry sentiment (details have been given in Section 5.2.1).

The 'ambiguous' value can be assigned for two reasons:

Modifier cannot be included. For example:

- "[Human]$_{\mathrm{M}}$ [power]$_{\mathrm{T}}$" in the expression "human destructive power". Conversely, the group "[destructive]$_{\mathrm{M}}$ [power]$_{\mathrm{T}}$" would have negative orientation.

- "[All]$_{\mathrm{T}}$ [deserve]$_{\mathrm{M}}$ in the expression "all deserve a chance". Conversely, the group "[deserve]$_{\mathrm{M}}$ [chance]$_{\mathrm{T}}$" would be positive.

- "[Seek]$_{\mathrm{M}}$ [pleasures]$_{\mathrm{T}}$" in the expression "seek only the pleasures of riches". Conversely, the groups "[seek]$_{\mathrm{M}}$ [riches]$_{\mathrm{T}}$" would have negative orientation.

- "[Increasing]$_{\mathrm{M}}$ [demand]$_{\mathrm{T}}$" in the expression "increasing demand for ideas".

- "$[\text{They}]_T$ $[\text{met}]_M$" in the expression "they will not be met easily". Conversely, the group "[met]T $[\text{easily}]_M$" would have positive orientation.

Unsure connotation given by the speaker. For example in:

- "America bigger than the sum of individual ambitions, greater than all the difference of birth and wealth", it is not clear how *ambitions* and *differences* are perceived by the speaker

- "How many Ramanujans are there in India today toiling in the fields, barely able to feed themselves?", it is not clear whether "feed themselves" is positive or negative.

In case the orientation given by the author is clear, but it does not sound right outside the context, it must not be changed. For example in:

- "Misery does love company" the group "$[\text{love}]_M$ $[\text{company}]_T$" is marked as negative.

- "US losing leadership", the group "$[\text{US}]_T$ $[\text{losing}]_M$" is positive and "$[\text{losing}]_M$ $[\text{leadership}]_T$" is negative.

Groups that are ambiguous will assume the orientation of the context. For example:

- "$[\text{Showered upon}]_M$ $[\text{favours}]_T$", negative according to the context.

- "$[\text{Idea}]_T$ $[\text{passed on}]_M$", positive.

- "$[\text{War}]_T$ against $[\text{network}]_M$" in the expression "Our nation is at war, against a far-reaching network of violence and hatred". The groups "war against violence" and "war against hatred" have not been annotated because only logical groups, not grammatical.

- "$[\text{Band}]_T$ of $[\text{patriots}]_M$ in the expression "a small band of patriots huddled".

5.6 Special cases

During the annotation task, a number of cases have been found not to be very easy to annotate for different reasons:

External references. Especially in TED talks, it might happen to find external references to the screen (e.g. this photo, first few bars here). In this case, the annotation is as usual if the target carries sentiment.

Difficulty in extracting sentiment through couples. For example:

- "I think it's very unlikely they were far from the minds of Americans", which can be annotated as "[they]$_{\mathrm{T}}$ [far]$_{\mathrm{M}}$" with marked force and negative orientation, as it was equivalent to "they were not far".
- "I wish you were here", which can be annotated as "[wish]$_{\mathrm{M}}$ [you]$_{\mathrm{T}}$ ".
- "The horrifying images from Abu Ghraib as well as the images from Guantanamo have a profound impact", with the groups "[horrifying]$_{\mathrm{M}}$ [images]$_{\mathrm{T}}$, [Abu Grahib]$_{\mathrm{M}}$ [images]$_{\mathrm{T}}$, [Guantanamo]$_{\mathrm{M}}$ [images]$_{\mathrm{T}}$, [profound]$_{\mathrm{M}}$ [impact]$_{\mathrm{T}}$.
- "Thank you to all photographers", annotated as "[thank]$_{\mathrm{M}}$ [photographers]$_{\mathrm{T}}$".

Too long modifiers. For example, "in a short span of time" and "from generation to generation" are too long and cannot be annotated. In addition, in the first case, it is impossible to select "in time" because of software limitations.

Adjective followed by a noun with words in the middle. For example "resistant to growth", annotated as "[growth]$_{\mathrm{T}}$ [resistant]$_{\mathrm{M}}$".

Useful negation markers. For example in the sentence

"heritage is a strength, not a weakness"

one group contains *not* ("[not]$_{\mathrm{M}}$ [weakness]$_{\mathrm{T}}$"), because otherwise the only verb *is* would have carried both positive and negative polarity.

Useful force markers, especially because of the lack of subject. For example in:

- "And over the next 18 years have almost tripled", the group "[almost]$_{\mathrm{M}}$ [tripled]$_{\mathrm{T}}$" should be annotated because the subject is missing.
- "Absolutely incredible", the group has necessarily to be that one because the sentence consists of these two words.
- "Sometimes even poorer than their grandparents had been", the group"[sometimes]$_{\mathrm{M}}$ [poorer]$_{\mathrm{T}}$" should be annotated because the subject is missing.

Punctuation. Annotating words separated by colon and semi-colon is allowed, but not when separated by full stops, exclamation marks or interrogation marks. For example in:

- "Some data on tariffs: coming down to", the group "[tariffs]$_{\mathrm{T}}$ [coming down]$_{\mathrm{M}}$" is annotated because it is not just a fact for the speaker.

- "America bigger than the sum of our individual ambitions; greater than all the differences", the group "[America]$_{\mathrm{T}}$ [greater]$_{\mathrm{M}}$" is created (see Figure 5.10).

APPRAISALGROUP | MODIFIER | TARGET

id	fromID	fromText	toID	toText	orientation
A419	T363	America	M389	bigger	positive
A420	M390	individual	T364	ambitions	ambiguous
A421	T363	America	M391	greater	positive
A422	M392	birth	T365	differences	ambiguous
A423	M393	wealth	T365	differences	ambiguous
A424	M394	faction	T365	differences	ambiguous

Figure 5.10: Appraisal groups annotated in sentences separated by semi-colon.

- "HIV/AIDS" in the sentence below are separated by double dash.

  "Images have [[power]$_{\mathrm{T}}$ to [shade light]$_{\mathrm{M}}$]$_{\mathrm{AG}}$ on [understanding]$_{\mathrm{M}}$ on [suspicion]$_{\mathrm{T}}$, [ignorance]$_{\mathrm{T}}$, and in particular – I've given [[a lot of]$_{\mathrm{M}}$ [talks]$_{\mathrm{T}}$]$_{\mathrm{AG}}$ on this but I'll just show one image – the issue of [HIV]$_{\mathrm{T}}$/[AIDS]$_{\mathrm{T}}$".

Appraisal groups with marked polarity or reverse force appearing to be of opposite orientation. For example:

- "Not a [bad]$_{\mathrm{M}}$ [thing]$_{\mathrm{T}}$". Although positive, marked as negative.

- "No more [kept]$_{\mathrm{T}}$ [in dark]$_{\mathrm{M}}$". Although positive, marked as negative.

- "Not [compromised]$_{\mathrm{M}}$ [principles]$_{\mathrm{T}}$". Although positive, but marked as negative.

- "Not [shrink from]$_{\mathrm{M}}$ [responsibilities]$_{\mathrm{T}}$". Although negative, marked as positive.

- "Not [one]$_{\mathrm{T}}$ of [shortcuts]$_{\mathrm{M}}$". Although negative, marked as positive.

- "Not [one]$_{\mathrm{T}}$ of [settling for less]$_{\mathrm{M}}$". Although negative, marked as positive.

- "Not a [path]$_{\mathrm{T}}$ for [faint-hearted]$_{\mathrm{M}}$". Although negative, marked as positive.

Co-reference: When an element is referring to something else mentioned in the sentence before or after, as in the short paragraph

"Let's begin with some images. They're iconic, perhaps cliches"

the pronoun ("they"), rather than the actual subject ("images"), is annotated in two groups ("[they]$_T$ [iconic]$_M$" and "[they]$_T$ [cliches]$_M$"). This strategy is used only when two or more sentences are involved. If it happens in the same sentence, e.g. "a nation, whether it is good or bad", annotating the pronoun (it) instead of the actual subject (nation) is preferred.

Logical link. For example in the following sentence:

"People were actually getting poorer than their parents. And sometimes even poorer than their grandparents had been."

the groups "parents poorer" and "grandparents poorer" should not be annotated because the link is not direct. However, here are the cases in which a direct link works:

- "It is fortunate that we are becoming less of an idea leader", in which the groups "[we]$_T$ [fortunate]$_M$" and "[less]$_M$ [leader]$_T$" are annotated, in the first case in order not to lose the information carried by *fortunate*.
- "Best efforts to help, because it is right", in which the annotated group is "[help]$_T$ is [right]$_M$".
- "The many who are poor, the many who are rich", in which the annotated groups are "[many]$_T$ [poor]$_M$", "[many]$_T$ [rich]$_M$".

Idioms and fixed expressions. It is still advisable to find appraisal groups like in the following cases:

- "The best is yet to come" as "[best]$_T$ to [come]$_M$".
- "On the cutting edge" as "[cutting]$_M$ [edge]$_T$".
- "We are not there" as "[we]$_T$ [there]$_M$".
- "Settling for less" as "[settling]$_T$ [for less]$_M$".

In alternative, if it is important to link the target, they can be annotated as single unit (e.g. "[parents]$_T$ [had the last say]$_M$").

5.7 Non-binary structures that can still be annotated with SentiML

This Section pertains to those expressions that can be defined as "non-binary" since their annotation is not straightforward, for example "The rights of man come not from the generosity of the state, but from the hand of God". In this case, while the appraisal groups "man rights", "state generosity" and "God hand" are easy to identify because they all consist of two nouns linked by the preposition *of*, the group "come generosity" is not a real group because the verb is not connected to its target.

The advantages are that these non-binary expressions are not frequent, and that most of the times they can still be included in the annotation (unless they should not be annotated as in the case of "come generosity"). Some of the most interesting examples are:

- "One form of colonial control shall not have passed away merely to be replaced by a far more iron tyranny", whose binary groups would be "colonial control" (carrying negative orientation), "control passed away" (carrying positive orientation), and "iron tyranny" (having high force); "control replaced" would be ambiguous, but it could still be annotated, while "replaced (by) tyranny" should be avoided because the verb has not *tyranny* as subject.
- "The instruments of war have outpaced the instruments of peace", in which the verb *outpaced* could not be linked to *instruments* of either *war* and *peace*, but we could still annotate "war instruments" and "peace instruments".
- I think it's very unlikely they were far from the minds of Americans, in which neither "they far" and "far minds" make sense on their own, but rather "were far" could be annotated with marked polarity as if it was "they were not far".
- "Imagination is joined to common purpose, and necessity to change", in which only "common purpose" and "necessity change" fall under the specified combinations. Unfortunately the expressions "imagination joined to purpose" and "imagination joined to necessity" cannot be annotated.
- "Globalization is increasing the demand for ideas, the incentive to create new ideas", in which "demand for ideas", "create ideas" and "new ideas" should be

annotated and marked as positive according to the author's perspective (even if new ideas might not be necessarily be good).

- "Images that provoke us to step up and do something, in other words to act", in which "images provoke", "us step up", "us do", "do something", "us act" all are sensible groups, although there is the need to skip *provoke* to annotate them in this way.

Despite the limitations illustrated in this Chapter, I have demonstrated how in theory the annotation scheme represents a reliable tool to answer all the research questions formulated in Section 1.3:

1. "How far is it possible to analyse explicit opinion in order to bring together both a linguistic and a computational perspective"
2. "What are the linguistic features of evaluative language that can lead to a successful automatic analysis of sentiment across multiple languages?"
3. "How far is the automatic classification of opinions into the main categories of the Appraisal Framework within Systemic Functional Linguistics possible and useful?"

In the following Chapters I will find an answer to all of them, by presenting both quantitative and qualitative results.

Chapter 6

Manual annotation: results

After describing the advantages and the principles of the annotation in the previous Chapter, I will start this Chapter by presenting the error analysis conducted in order to automatically detect borderline cases and inconsistencies in the manual annotation. Afterwards I will present the statistics related to the manually-annotated corpora in English, Italian and Russian.

6.1 Manual annotation: the error analysis phase

Inspired by the co-training strategy, I designed and implemented a way to speed up the manual error analysis based on a number of machine learning models trained on different *views* of the same data. The predictions obtained by these models were then automatically compared in order to bring to light highly uncertain annotations and systematic mistakes in the classification of *targets* and *modifiers*.

This approach was similar in terms of variation from the standard procedures to Snow *et al.* (2008). For each expert annotator (six in total) they trained a system using only the judgements provided by those annotators, and then created a test set using the average of the responses of the remaining five labellers on that set. The result were six independent expert-trained systems. The difference with my methodology is that I trained six independent classifiers based on the judgements of only one human annotator (myself), and used the average of the responses of the six classifiers as the most predicted class to be compared to the gold standard.

Jin *et al.* (2009) also used the strategy of selecting the labelled sentences agreed upon by their classifiers and achieved good performances in the task of identifying opinion sentences.

Finally, this methodology is also similar to one of those mentioned in Yu (2014). The author used the traditional co-training strategy, i.e. providing a small pool of unlabelled data to two classifiers with confidence rates in order to obtain automatically labelled examples that would be added to an initial set of labelled ones. Subsequently, this final large set is used to train the two classifiers, and a combination of them (constructed by multiplying their predictions) is eventually the one used to label new documents. Five strategies were applied to obtain the views: (a) using unigrams and bigrams as features, (b) randomly splitting the feature set in two, (c) using two different supervised learning algorithms because they would provide useful examples to each other since based on different learning assumptions; (d) randomly splitting the training set, and (e) applying a character-based language model (CLM) and a bag-of-words model (BOW). We extended the third strategy by using three classifiers and two different views for each of them, and by applying this to the task of annotation validation rather than semi-supervised learning.

6.1.1 Description of the methodology adopted

The use of multiple supervised machine learning classifiers and the analysis of their predictions in parallel to automatically identify disagreements ultimately leads to the discovery of borderline cases in the annotation, an expensive task in terms of time when carried out manually.

In particular, the goal is to highlight:

- Predictions with a number of different labels that might be signals of inconsistencies in the annotation and highly difficult cases to be annotated. These must be manually analysed afterwards. The analysis of those disagreements in conjunction with the gold annotations also provides insights about the efficacy of the features provided to the classifiers for the learning phase.

- Cases with high agreement that are most likely signal of a reliable annotation scheme. Conversely, if all the classifiers agree on a wrong annotation, this is a strong signal of ambiguity in the annotation schema and/or guidelines.

I will now present the methodology as a series of steps, and for each of them provide the specific decisions made for the testing of *SentiML*.

The first step consists in splitting the annotated corpus into a training set and a test set. I decided to use 90% of the English corpus for training purposes and 10% for test purposes.

Afterwards features for the machine learning phase have to be prepared. The optimal set of features to model the annotation task varies from problem to problem. I used the following:

- Word features, representing the numeral identifier, word form, lemma and POS-tag of each word.
- Contextual features, representing the lemma and POS-tags of the preceding and succeeding words.
- Dependency-based features, representing the reference to the word on which the current token depends in the dependency tree (*head*) along with its lemma, POS-tag and relation type.
- Number of linked modifiers, representing the number of adjectives and adverbs linked to the current word in the dependency tree.
- Role, representing the predicted role (modifier or target) of the current token in conveying sentiment. The predictions are computed using fixed syntactic rules.
- Gazetteer-based sentiment. I used the *NRC Word-Emotion Association Lexicon* (Mohammad, 2011) to represent the *a-priori* (out-of-context) sentiment of each word.

Once the features are ready, two or more feature partitions (called *views* in the co-training strategy) have to be defined in order to be as orthogonal as possible (Abney, 2007). I opted for a linguistically-grounded dichotomy: lexical features (word features, role and gazetteer-based sentiment) versus syntactic features (contextual and dependency-based features, number of linked modifiers). The training set and the test set are split accordingly.

At this point, machine learning classifiers are chosen. These need to be confidence-rated, i.e. able to provide a confidence rate for each prediction. In my experiments

I selected Naïve Bayes (NB), Radial Basis Function Network (RBF) and Logistic Regression (LR)[1]. These models rely on very different strategies, which makes the analysis more reliable. I discarded Support Vector Machines since in my preliminary experiments they had very low F-measure (a range between 0.09 and 0.11 across modifiers and targets), not comparable to the other models reaching between 0.32 and 0.55.

A model for each combination of view and classifier has to be now produced and tested on the test set. I performed a 10-fold cross-validation. In the test phase, I opted for a numerical threshold of 0.67 to consider the predictions reliable. A prediction with a confidence lower than the threshold is considered uncertain.

For each instance several potentially different predictions are obtained, in my case six. In order to decide one candidate prediction, the agreement score per class has to be calculated and the most predicted class selected.

Only the predictions different from the gold annotations are considered: the higher the agreement score, the more the instance is interesting in the context of my analysis. The final step consists in manually investigating such cases to shed light on the errors. In this experiment I opted for the use of a simple protocol based on the following classification schema:

- W (wrong), when I stand by the gold, despite the classifiers disagree with it (e.g. *flourish* ("make your farm flourish") classified as 'target').
- A (ambiguous), when I can consider the suggested category possible along with the gold (e.g. *consume* ("consume resource") classified as 'target'). Those are most likely the cases in which the guidelines need to be better or the annotation method could have been simpler.
- M (to modify), when the gold needs to be amended, because it is incorrect (e.g. only *much* ("much more") should have been annotated).

I stress the advantage of having as result a drastically reduced number of instances to be examined with respect to the entire set.

The methodology is summarised in Figure 6.1 (part 1) and Figure 6.2 (part 2). The models, the datasets and the error analysis are publicly available in order to ensure reproducibility[2].

[1] The implementation provided by *WEKA* tool (`http://www.cs.waikato.ac.nz/~ml/weka/`) has been used for all of them.

[2] `http://corpus.leeds.ac.uk/marilena/public/SentiML/`

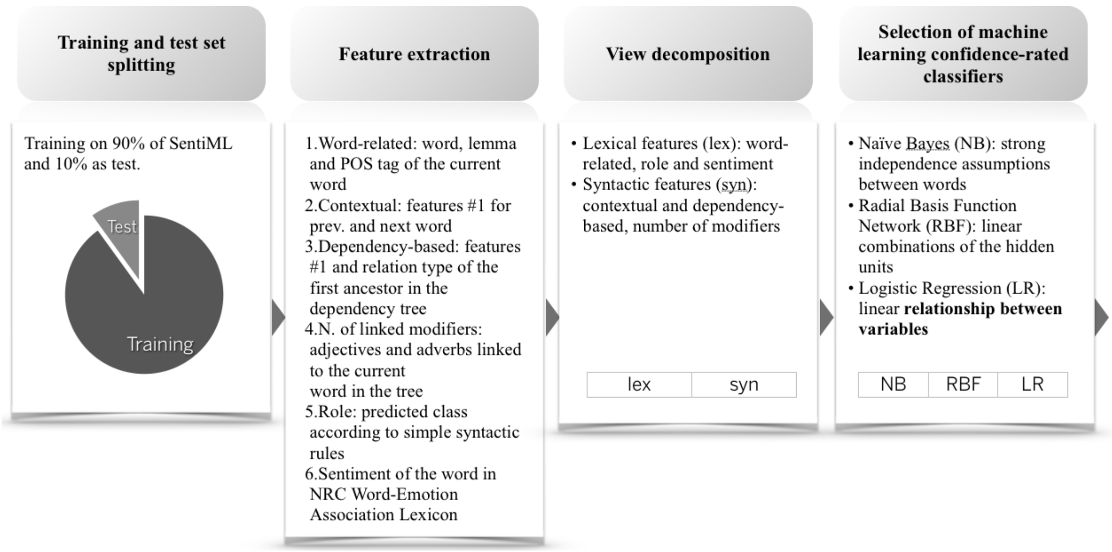

Figure 6.1: For the automatic detection of borderline cases first 2 views and 3 confidence-rated classifiers were chosen.

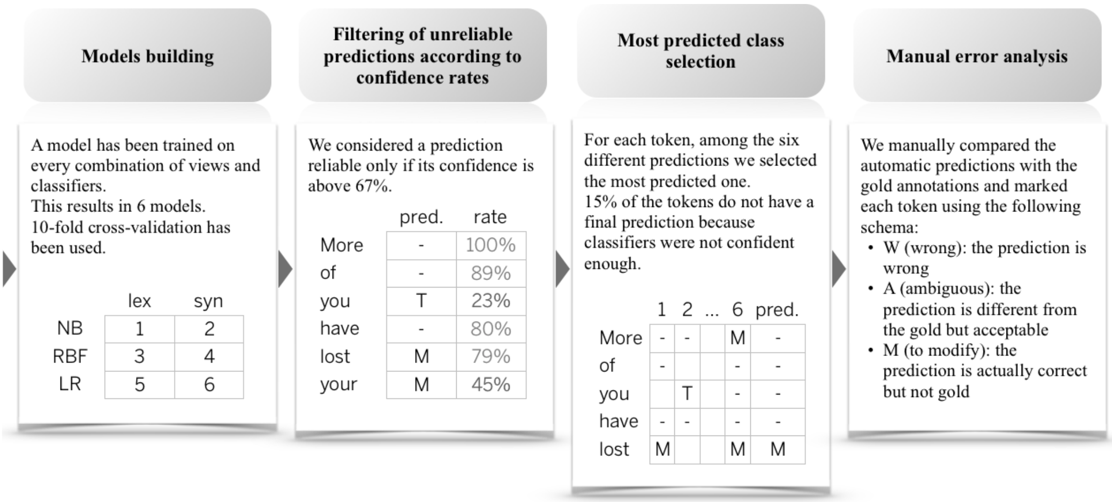

Figure 6.2: For the automatic detection of borderline cases the confident predictions given by the views of the classifiers were manually analysed.

6.1.2 Results

Figure 6.3 shows the percentage of predictions matching 67% (or above) of the confidence rate for each classifier. In general they are higher in the case of the syntactic feature set (87% for NB, 88% for RBF and 78% for LR) than for the lexical feature set (97% for NB, 69% for RBF and 76% for LR).

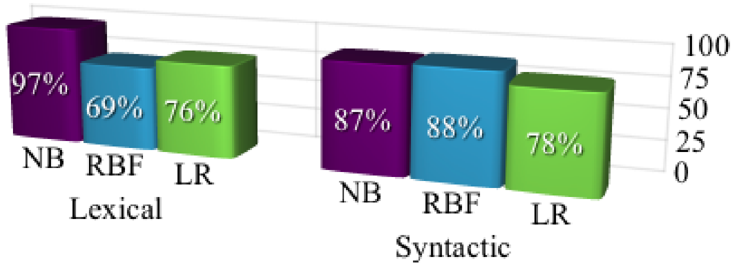

Figure 6.3: The figure shows that, with 0.67 as confidence rate, many predictions would go beyond it and considered *confident*.

Table 6.1 shows the performances of the six models obtained from the training of each combination of view and classifier, mentioned in the Section 6.1.1. F-measures for modifiers range across 0.32 and 0.54 for modifiers, and 0.43 and 0.55 for targets. There is no huge difference in performances between the lexical and the syntactic feature sets, which is good in the light of data sparseness.

The performance on the the empty class (no category assigned) was good, as 76% was predicted out of 77%, whereas the performance on the modifiers was 4% out of the gold 12% and the performance on the targets was 5% out of the gold 11%. Although the annotation allows each token to be annotated both as modifier and target in two different appraisal groups, I have not reported the performances for the MT class as the cases are not significant. Finally, there were 15% of cases in which the classifiers were not confident.

Feature set	**Classifier**	**Modifier**			**Target**		
		Precision	**Recall**	$\mathbf{F}_{\beta=1}$	**Precision**	**Recall**	$\mathbf{F}_{\beta=1}$
Lexical	NB	**0.71**	0.10	0.48	**0.82**	0.12	0.43
	RBF	0.52	**0.56**	**0.54**	0.51	**0.59**	**0.55**
	LR	0.59	0.42	0.49	0.61	0.48	0.54
Syntactic	NB	0.46	0.48	0.47	**0.82**	0.12	0.43
	RBF	0.49	0.35	0.40	0.55	0.50	0.53
	LR	0.58	0.22	0.32	0.60	0.41	0.49

Table 6.1: Performance of the classifiers Naïve Bayes (NB), RBF Network (RBF) and Logistic regression (LR) trained on two views, lexical and syntactic.

As far as the manual classification of mistakes is concerned, an important result is that in 79% of cases (1630 instances) out of the total test instances (2066) the most predicted class matched the gold standard. As mentioned in the premises of the methodology, these predictions can be regarded as expression of either the efficacy of the annotation schema, the guidelines, the features used for the machine learning step or their combination.

The remaining 21% (436 instances) in which the most predicted class differed from the gold standard represented my final set of instances on which to conduct the error analysis: the label *Wrong* was assigned 49% of times (214 instances), the label *Ambiguous* in which the predictions were different but acceptable was assigned 35% of times (153 instances), and the label *Modify* for the cases in which the gold standard was wrong was assigned 16% (69 instances). Figure 6.4 graphically summarises these percentages.

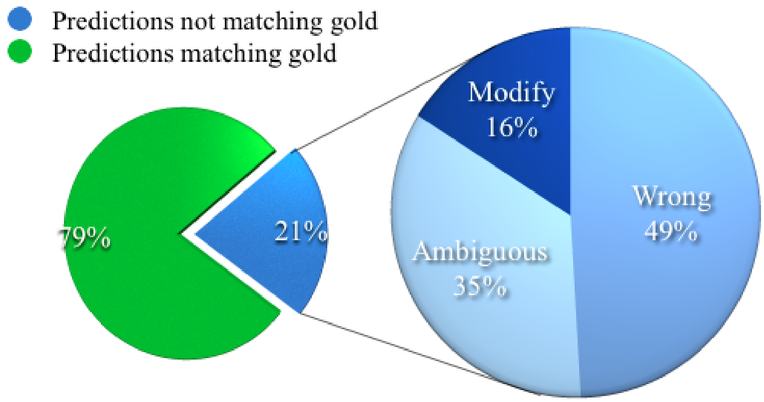

Figure 6.4: The set of instances in which the prediction did not match the gold (21% of the test set) was categorised in *Wrong, Modify* and *Ambiguous*.

Wrong was mostly assigned when the modifier or the target was correctly identified, but not the counterpart in the pair (e.g. *way forward, blame society, wrong side*). It was also assigned when a word should have been identified because of a strong sentiment word (e.g. *destroy, flourish*), and only the first of two or more targets was identified (e.g. *women and children, the city and the country*).

Ambiguous was assigned when an adverb was annotated as modifier (e.g. *through corruption, seize gladly, tragically reminded*) since these are ambiguous cases for human annotation too in so far as it is the annotator's decision to include adverbs if they think this is important for the sentiment. Other cases in which the label has been used is with compound modifiers (e.g. *face to face, in the face of*), phrasal verbs (e.g. *turn back, carried forth, came forth*) and difficult couples to link (e.g. *instruments with which we meet them [challenges]*). Finally this label was also used in cases in which the prediction was sensible, but not as good as the gold (e.g. in *enjoy relative plenty*, the gold standard was *enjoy plenty* and the classifiers predicted *relative plenty*).

Modify was assigned when another modifier had been wrongly annotated by the annotator, instead of modifying the value of the force of the current one (e.g. in *much more*, only *more* should have been annotated with 'high' force), in the case of couples with no sentiment (e.g. *future generations, different form*), of not previously identified couple (e.g. *stairway filled with smoke, icy river*) or couples that could have been annotated in an easier way (e.g. *provoke us to step up and do something, image resonates with us*).

6.1.3 Two annotation iterations

The method just described allowed me to have further feedback on the annotation scheme and the annotation process that I have carried out (see Chapter 5). Since I was the sole annotator, I could not make use of the standard solution represented by the inter-annotator agreement (IAA) (Artstein & Poesio, 2008). Cases in which this has been used involve previous works in the field of appraisal such as Wilson (2008) and Read *et al.* (2007b), in which the guidelines did not specify the span of the annotations.

However, following the best practices outlined in Pustejovsky & Stubbs (2012), I revised my annotations after a reasonably long period of time, and kept the last version separate from the first.

The revision was done also considering the outcomes of the error analysis described in the previous Section 6.1.2. In particular I corrected the attitude type and added any appraisal groups that had not been previously annotated, by trying to be consistent across the languages on these two aspects[1].

Nonetheless, some inconsistencies were innate characteristics of the texts under analysis. In particular I am referring to the *translation universals* (Baker *et al.*, 1993), on which some reflections have been done in Chapter 4:

- Simplification as "the idea that translators subconsciously simplify the language or message or both" (Baker, 1996) (p. 176), of which an example from my data

[1]Only the last iteration has been used as "gold standard" for testing of the automatic system.

is *we're naive, we're bright-eyed and bushy-tailed* translated as *siamo ingenui ed impazienti* (we are naive and impatient).

- Normalisation as "the tendency to conform to patterns and practices that are typical of the target language, even to the point of exaggeration" (Baker, 1996) (p. 183), of which an example is *Seeing the planet like this for the first time, its smallness, its fragility*, with the gerund kept in Italian although in the past (*all'averlo visto*), but not in Russian because less common (увидели нашу планету (saw our planet)) and with the two final nouns translated with adjectives: *piccolo e fragile* and такой маленькой, такой хрупкой (so small, so fragile).

The other dimension I am looking at apart from the annotation process is the categorization of the data according to the annotation scheme (i.e. modifiers, targets and appraisal groups), which I will discuss extensively in the following section. Both the findings categorised under these dimensions will be extremely valuable for giving linguistically-motivated explanations for the variety of the results related to the automatic system across the languages (see Section 8.2.4).

6.2 Statistics on the annotated data

6.2.1 Data related to the identification phase

In this section I will give different types of statistics on the annotated data, as complementary to the qualitative analysis in Chapter 4. While this investigation had been already conducted on a non-complete English corpus (Di Bari *et al.*, 2013), this time it has the advantage of being on complete corpora in all languages.

I will start by looking at the most general and yet essential aspect: how much of the data is sentiment-related. This is a piece of information that it is possible to obtain in two ways: through (i) the number of appraisal groups, and (ii) the percentage of the words making the appraisal groups (out of the total number of words).

Number of appraisal groups. As clear from Table 6.2, English has more annotated appraisal groups (1209) with respect to Italian (1081) and Russian (1108). It is also possible to see that, among the types of text, political speeches have more appraisal groups in all the languages, whereas news the lowest.

Language	Type	Appraisal groups	Targets	Modifiers
EN	Political	624	519	551
	News	236	194	197
	TED	349	326	297
	tot	1209	1039	1045
IT	Political	486	411	437
	News	254	203	244
	TED	341	292	323
	tot	1081	906	1004
RU	Political	599	510	542
	News	221	191	214
	TED	288	246	264
	tot	1108	947	1020

Table 6.2: Number of annotated categories according to language and text type.

Percentage of the words making the appraisal groups out of the total number of words. Figure 6.5 shows that English has **27%** (with 2418 words out of 9055) in comparison to Italian and Russian that have **24%** (with 2162 words out of 9080, and 2216 words out of 9035 respectively).

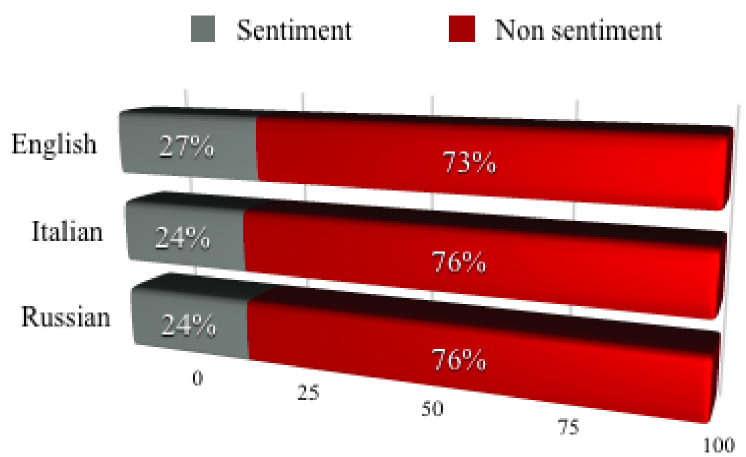

Figure 6.5: Percentage of sentiment words according to language

From Figure 6.6 (with Table 6.3 showing the exact numbers leading to these percentages) we can see that Russian is the most volatile with both the lowest percentage for news and the highest for political speeches, while TED is the most consistent among the text types.

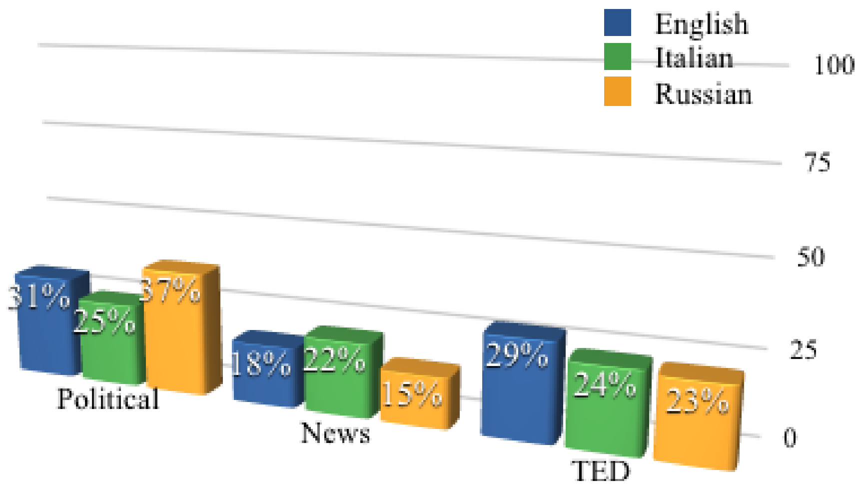

Figure 6.6: Percentage of sentiment words according to text type

Language	Type	# of appraisal groups	# of words included in appraisal groups/ # of words	percentage of words included in appraisal groups
EN	Political	624	1187 / 3782	31%
	News	236	424 / 2281	18%
	TED	349	665 / 2992	29%
IT	Political	486	982 / 3960	25%
	News	254	526 / 2316	22%
	TED	341	685 / 2804	24%
RU	Political	599	1248 / 3408	37%
	News	221	475 / 3094	15%
	TED	288	581 / 2533	23%

Table 6.3: Number of words included in appraisal groups for each text type across languages.

6.2.2 Data related to the attributes

In this section I will provide the statistics for each of the attributes.

Orientation. This attribute refers to the values 'positive', 'negative', 'neutral' and 'ambiguous' for appraisal groups, targets and modifiers.

Orientation values are shown both in Table 6.4 and Figure 6.7. What emerges is that positive appraisal groups are more common than the negative ones in all three languages, by covering a range of 59-67% vs. 29-36%; the almost complete lack of neutral appraisal groups is aligned to the expectation that appraisal groups should not be neutral since they carry sentiment (either positive, negative or ambiguous), whereas the same explanation cannot be applied to targets and modifiers in so far as they can be encapsulated in sentiment groups despite being neutral individually. This happens especially in the case of targets because they represent the object of the appraisal expression, but it is mostly modifiers that carry the appraisal. In particular, we see that Italian and Russian have similar percentages for targets, which might be due to the similar nature of the lexicon contained in the originally-produced news.

Language	Category	Positive	Negative	Neutral	Ambiguous
EN	Appraisal groups	744	440	2	23
	Targets	165	200	477	215
	Modifiers	294	189	178	481
IT	Appraisal groups	723	345	0	13
	Targets	247	146	334	186
	Modifiers	299	141	143	467
RU	Appraisal groups	736	362	0	10
	Targets	284	124	400	151
	Modifiers	382	178	120	367

Table 6.4: Orientation values for appraisal groups, targets and modifiers across languages.

Another pattern across languages is represented by modifiers having very similar percentages this time in English and Italian. This is probably due to the role of their language typology in the greater number of phrasal (or multi-word) expressions. In fact, the equivalent expressions for these in languages with different morpho-syntactic structures like Russian often involve prefixation (Kruijff *et al.*, 2000; Mudraya *et al.*, 2005; Sharoff, 2004):

- die down - замирать
- find out - выяснить
- take apart/portare via – разбирать
- scrape together - наскрести

but not always: bring together - сводить вместе, leave behind - оставлять позади, to go after gold - отправиться за золотом, andare dietro all'oro; to go after a prize - вступить в борьбу за призовое место, andare dietro a un sogno.

On the other hand, generally speaking, Italian does not lack prefixed forms, although they might not be as common as the constructions "adverb + verb" (Iacobini & Masini, 2007)):

- move ahead - продвигаться or двигаться вперёд, procedere or portare avanti.
- go around - обойти, andare attorno or girare.

In the manually annotated English corpus, and most frequently in TED talks, I counted 60 (not unique) occurrences, both hyphenated expressions (e.g. *bushy-tailed, bright-eyed, well-known, anti-poverty*) and not (e.g. *in dark, in the light of day, face to face, off a cliff, becoming wealthy, gets rich*).

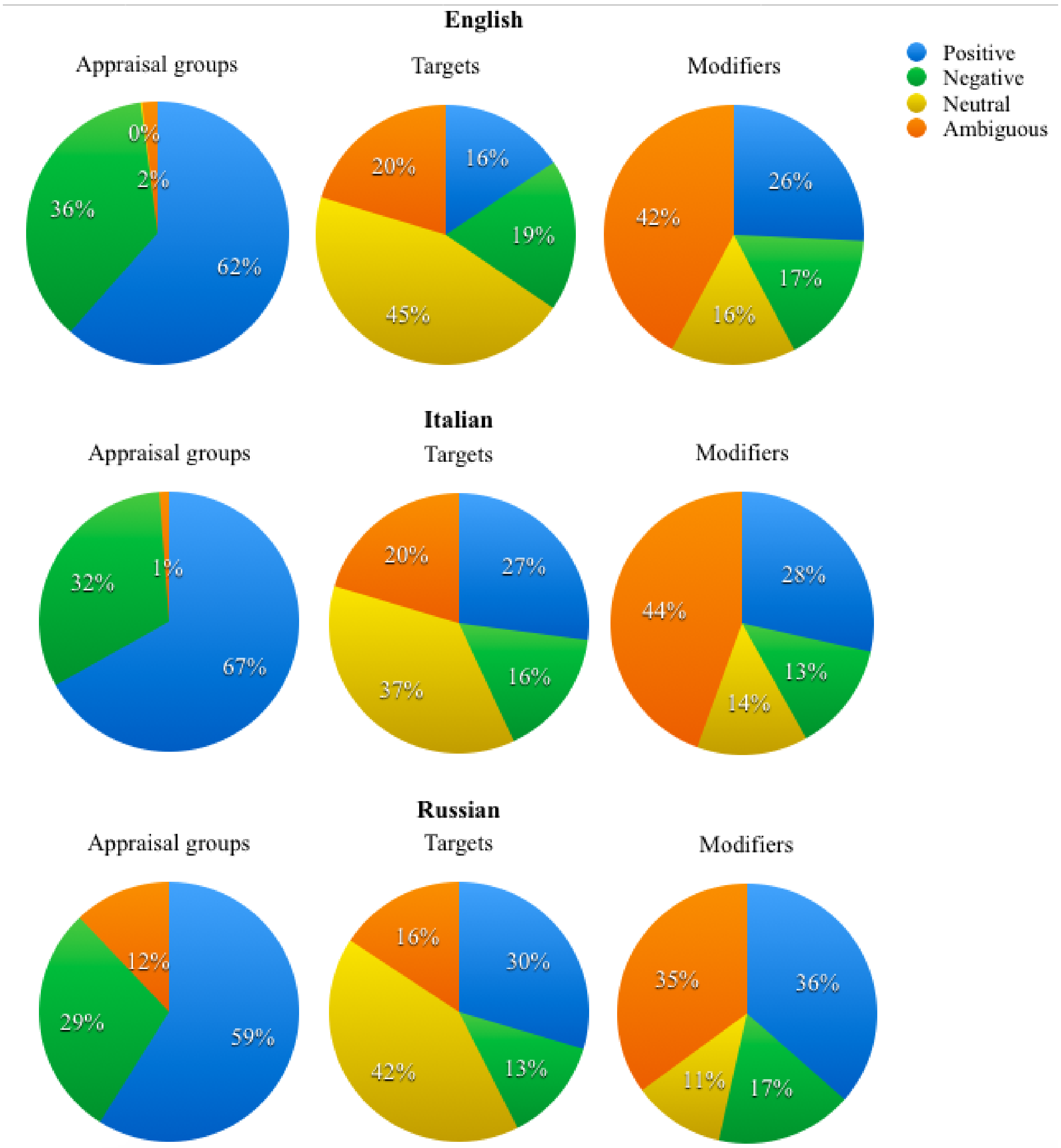

Figure 6.7: Comparison of the orientation values in their percentage form across languages.

An evaluation of dictionary orientation vs. manually-annotated orientation

Because I am using a sentiment dictionary, I conducted research in order to find out how 'realistic' the orientation of the individual words recorded in it is. The problem is that sentiment dictionaries report the orientation that word is frequently assigned, which means that there are very few cases of words with both ‘positive’ and ‘negative’ value. This represents a first considerable difference with my manual annotations in which ‘ambiguous’ targets and modifiers appear around 16-20% and 35-42% respectively.

Another important comparison that it is worth making is the orientation in the sentiment dictionary vs. the orientation of the appraisal group the word belongs to (orientation that is by definition *contextual*). For my analysis on the overall datasets, I have used the “NRC Word-Emotion Association Lexicon” (Mohammad, 2011), whose annotations were manually done through “Amazon’s Mechanical Turk”, while the “Roget Thesaurus”[1] was used as source for the target terms. The lexicon has entries for about 24200 word–sense pairs, corresponding to 14200 word types. Apart from the sentiment that has ‘positive’ and ‘negative’ values, each word could also have a value coming from emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust).

I used the English dictionary as source to create the correspondent Italian and Russian dictionaries[2] and then made the comparison on the partially-annotated English dataset (Di Bari *et al.*, 2013). As soon as the annotations were completed, the comparison was conducted on the final datasets in English, Italian and Russian.

As evidence for my hypothesis, I have discovered that the words included in my appraisal groups were present in the sentiment dictionaries only **35.33%** of times in English, **29.39%** of times in Italian and **10.29%** of times in Russian (see Figures 6.8, 6.9 and 6.10).

Such percentages suggest that we can rely on the dictionary only to a certain extent if we aim at more accuracy. In particular, the low accuracy in Russian might be due to a number of mistranslations from the English dictionary.

[1] `http://www.gutenberg.org/ebooks/10681`

[2] The translation was carried out by using *Google translate*. All the words were first lemmatised and trimmed to avoid repetitions of singular and plural, as well as repetition of the same word with wrongly-annotated extra spaces.

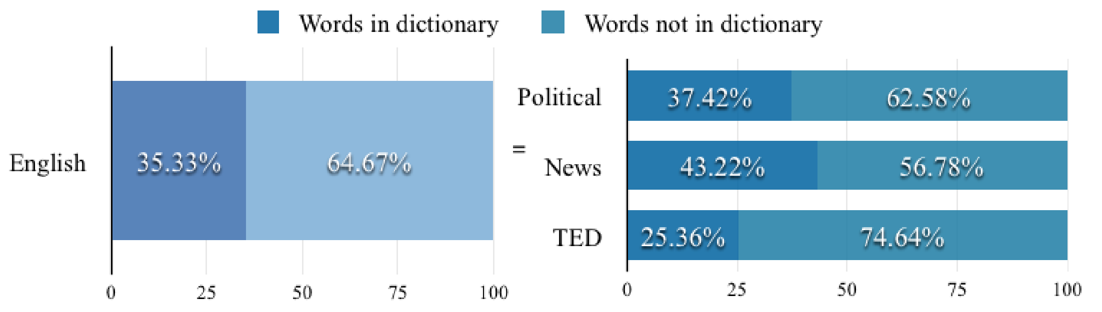

Figure 6.8: Percentage of words part of appraisal groups present in the dictionary for the entire English dataset and for specific text types.

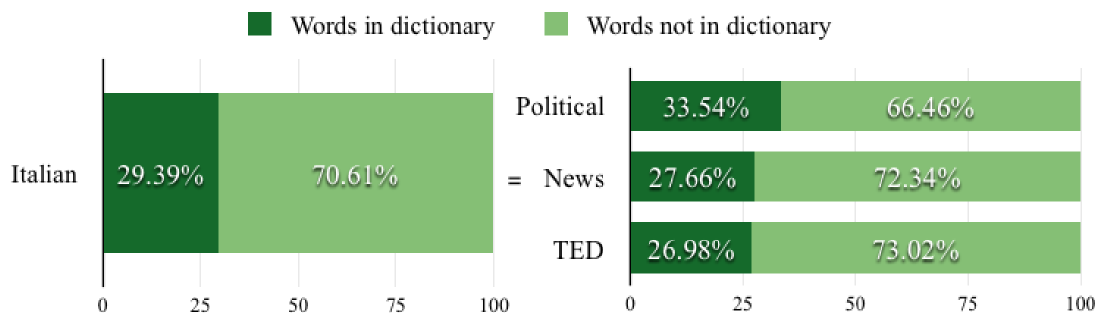

Figure 6.9: Percentage of words part of appraisal groups and present in the dictionary for the entire Italian dataset and for specific text types.

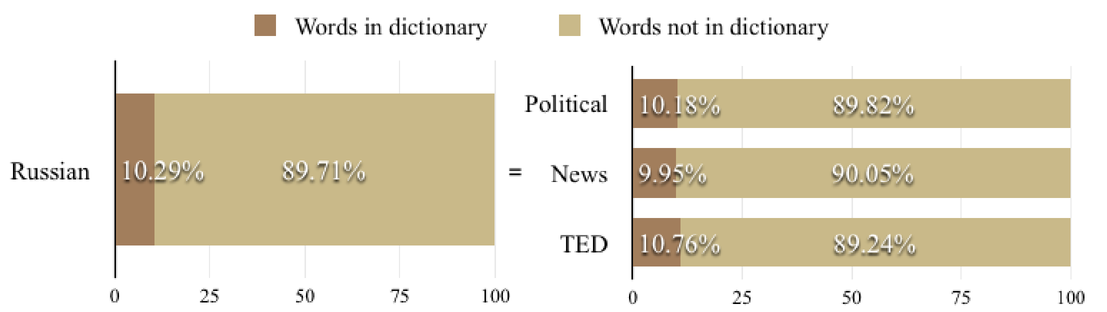

Figure 6.10: Percentage of words part of appraisal groups and present in the dictionary for the entire Russian dataset and for specific text types.

Figures 6.8, 6.9 and 6.10 show these percentages for each text type and for the overall dataset in each language (where the percentages for the overall datasets are the average of those related to the text types). Data in English seem to show the greatest variations. As far as the linguistic analysis of the data is concerned, a considerable amount of adjectives such as *anti-terror*, *better*, *bitter*, *brave*, *wisely* and *weak* were missing in the dictionary, although for some of them nominal forms were provided (e.g. *bitterness*, *braveness*). Multi-word expressions were also not included at all in the dictionary.

I then focused on the words included in the dictionaries. In order to carry out a systematic analysis, I decided to divide them in 3 categories:

- **Agreeing words**: words whose dictionary orientation agrees with that of the appraisal group they are taken from.

- **Disagreeing words**: words whose dictionary orientation does not agree with that of the appraisal group they are taken from.

- **Ambiguous words**: words that already have both positive and negative values in the dictionary.

Table 6.5 shows the percentages in which the prior orientation and contextual orientation are agreeing or disagreeing, as well as the percentage related to the words with both ‘positive’ and ‘negative’ values (thus defined as ‘ambiguous’) in the dictionary. While the Table gives in detail also the percentages for the text types, Figure 6.11 only serves as quick comparison among languages.

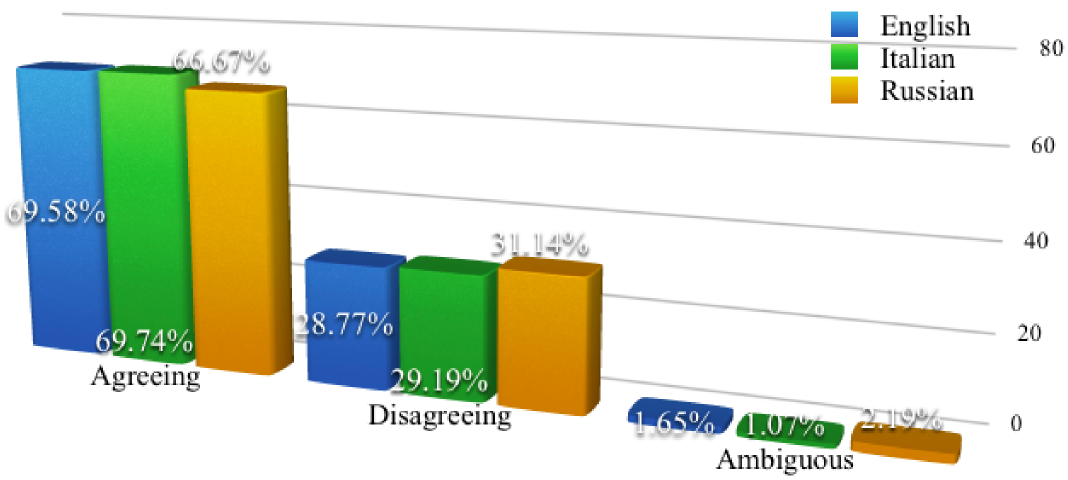

Figure 6.11: Percentage of words part of appraisal groups present in the sentiment dictionary.

Language	Orientation category	Text type			
		Political	News	TED	Complete dataset
EN	Agreeing	73.23%	62.26%	68.36%	69.58%
	Disagreeing	25.91%	36.76%	27.12%	28.77%
	Ambiguous	0.86%	0.98%	4.52%	1.65%
IT	Agreeing	71.47%	70.21%	66.30%	69.74%
	Disagreeing	27.30%	27.66%	33.70%	29.19%
	Ambiguous	1.23%	2.13%	0%	1.07%
RU	Agreeing	67.21%	65.91%	66.13%	66.67%
	Disagreeing	31.15%	29.55%	32.26%	31.14%
	Ambiguous	1.64%	4.54%	1.61%	2.19%

Table 6.5: In the context of the comparison between their prior orientation and the contextual orientation assigned to their appraisal group, words can have either agreeing or disagreeing orientation, as well as be ambiguous a priori. For each of these categories the percentage is shown.

From Figure 6.11 it is, in fact, possible to see that agreeing words cover between 66% and 69% of the total times words were found in the dictionary. The list generally includes reasonable out-of-context positive words (e.g. *love, liberty, leisure, bless*), as well as out-of-context negative words (e.g. *criticism, hypocrisy, hostile, blame*).

Disagreeing words, i.e. the percentage in which the prior orientation given in the dictionary is different from the correct one given by the context, is **28.77%** of times in English, **29.18%** in Italian and **31.14%** in Russian. This is the most important piece of information since it is the one that allows a meaningful comparison. Words included in this group are *maximum, important, demand, balance*, and some that seem to be positive or negative *a priori* (e.g.,*freedom*, *discrimination*, *liberty*, *peace*, *enemy*, *deserve*, *effort*) but in fact with orientation depending on the context: *useless effort, deserve to suffer*. Another category is that of the so-called “reversals” such as *abolish*, *attack*, *oppose*, *question* (see complete list in Chapter 5) usually found in the case of nouns linked by a preposition such as *infringement of liberty, lack of freedom, execution of citizen, war/campaign against terrorism, trade of sex, prohibition of terrorism*.

Finally, ambiguous words only account for 1-2%.

Attitude. This is the second most important attribute, and the difference across the languages is of much interest: in English the most common value is ‘judgement’ (assigned 540 times), followed by ‘appreciation’ (assigned 482 times) and ‘affect’ (assigned 130 times), whereas in Italian and Russian it is 'appreciation' (assigned 646 and 567 times respectively), followed by 'judgement' (assigned 294 and 325 times respectively) and 'affect' (assigned 110 and 155 times respectively). In terms of percentages, Figure 6.12 shows that across languages 'judgement' covers from 28 to 47%, 'appreciation' from 42 to 62% and 'affect' from 10 to 15%.

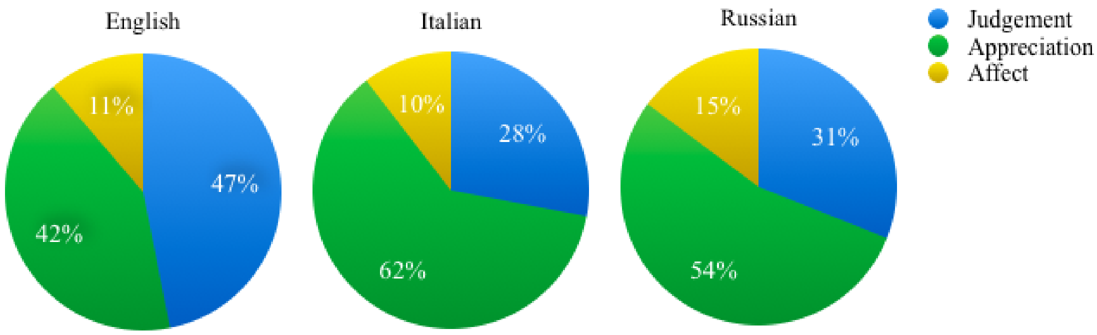

Figure 6.12: Comparison of the attitude values in their percentage form across languages.

The difference in the proportions is due to different factors. First of all, the topics and the language of the news, which represent an original section in each language: in English they are on human rights, which explains the abundance of attitudinal evaluation in which human behaviour is criticised or praised by reference to some set of social norms (184 occurrences of 'judgement') as opposed to assessments of objects, processes and states of affairs (only 37 occurrences of 'appreciation'); Italian and Russian news are more related to economics, which explains their similar percentages (64 of 'judgement' and 185 of 'appreciation' in Italian and 95 of 'judgement' and 111 of 'appreciation' in Russian).

However, most interestingly, the abundance of 'judgement' in English is also retrievable in the non-originally produced documents represented by political speeches and TED talks. In particular, political speeches have 249 occurrences of 'judgement' in English vs. 156 in Italian and 174 in Russian, while TED talks have 107 occurrences in English vs. 74 in Italian and 56 in Russian. It is an interesting phenomenon in so far as it would work as further evidence in support of the linguistic analysis described in Chapter 4, in which I have demonstrated that there are substantial differences in the case of the terms used as equivalents, additions/omissions, invoked attitude as well as in the rendering of metaphors and idiomatic expressions.

On the other hand, the possibility that such differences in the proportions are also due to any inconsistencies in the manual annotation across the languages despite the two iterations described in Section 6.1.3 must be brought forward.

Polarity. Unlike the previous case, the amount of negations is similar across the languages: polarity is, in fact, 'marked' 52 times (vs. 1101) in English, 45 times (vs. 1005) in Italian, 68 times (vs. 979) in Russian. Figure 6.13 shows that this corresponds to 4-6%. As previously explained, any considerations on the linguistic preferences of each language derived from this result must take into account that there are a number of translations that populate the corpora, along with the originally-produced texts.

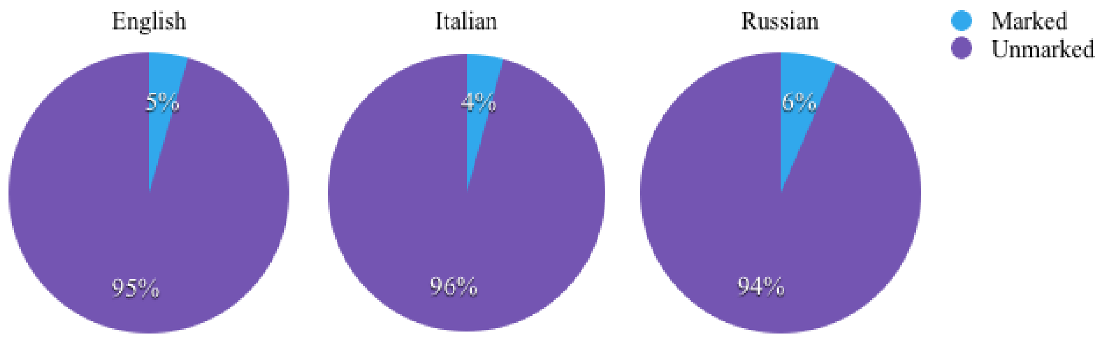

Figure 6.13: Comparison of the polarity values in their percentage form across languages.

Linguistically speaking, it was interesting to see that *to do* was negated almost always in its auxiliary function and accompanying appraisal verbs (e.g. *I do not shrink from this responsibility, people do not believe, images don't change the world*), and only once in its main function (*there is little we cannot do*)[1]. Also, by looking at the occurrences of *to be*, they were mainly in the passive form (*this will not be finished, this will not be met easily, we are not darkened, human rights must not be compromised, our spirit cannot be broken*), and only few times it was negated as main verb (e.g. *that's not that far, that civility is not a sign of weakness, it has not been the path for faint-hearted*). Among the other verbs frequently negated and related to appraisal *to fear* appears three times in political speeches (*we should not fear other countries becoming wealthy, the report is feared, let us never fear to negotiate*). Semantic groups commonly negated include the verb *to use* (*measures to crack down on terrorism should not be used to justify*

[1]Collocations have been provided by *SketchEngine*

rights abuses, the United States did not use the term "axis of evil", not to use the war against terror as an excuse) and *not knowing* (*not knowing that the economy was about to go off a cliff, not knowing that we were about to enter the twentieth century, what would you have predicted not knowing this?*). Interestingly, the group *not only* did not lead the system to wrong classifications: in fact in *we honor them not only because they are guardians* the verb *are* has 'unmarked' polarity, while *we will act not only to create new jobs* was not spotted.

Force. The order in the values of force (i.e. 'normal', 'high', 'reverse', 'low') is the same across the languages with 'normal' as the most common, in accordance to the expectations. According to Figure 6.14, 'normal' has been assigned 82-86% of times (corresponding to 991 times in English, 906 times in Italian and 856 times in Russian), 'high' 9-10% (corresponding to 105 times in English, 97 times in Italian and 106 times in Russian), 'reverse' 4-7% (corresponding to 43 times in English, 43 times in Italian and 78 times in Russian) and 'low' 0-1% (corresponding to 14 times in English, 4 times in Italian and 7 times in Russian).

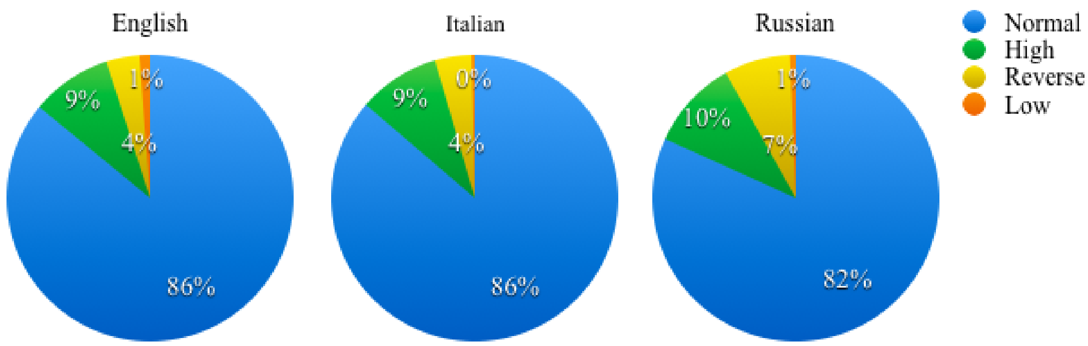

Figure 6.14: Comparison of the force values in their percentage form across languages.

In the case of *force*, two main categories should be looked at: adverbs and comparative/superlative forms.

In the case of 'high', among the adverbs there are some "expected" ones such as *very, extremely*, but also a few "modal adjuncts" (Stubbs, 1996b) cited in White (2003), Salager-Meyer (1994) cited in Kussmaul (1997) and "emotionally charged intensifiers" like *probably, ultimately, extremely, particularly, unexpectedly, surprisingly*, especially in TED talks (with 59 occurrences in English, 44 in Italian and 45 in Russian) but in general also in the other text types (on average less than 20 occurrences):

"Such an unfavourable report for North Korea is likely to produce a negative effect on Pyongyang-Washington relations" (news)

"That time has surely passed" (political speeches)

"It is ultimately the faith and determination" (political speeches)

"Absolutely incredible (TED talks)"

As for the comparative of majority and superlative, some examples from TED talks are *smaller crew, greater function, poorer people, greater revenues, larger markets, larger budgets, surest route*, while from political speeches *maximum danger, better history, greater cooperation*. Another much less frequent category apart from adverbs and comparative/superlative forms is represented by adjectives that already convey a 'high' force, such as *tremendous growth, reverend clergy, Almighty God*.

In the case of 'low', there are both "cues" such as *perhaps, maybe* (Salager-Meyer, 1994) and "downtoners" such as *barely, nearly, slightly* (Hyland, 1995) cited in (Kussmaul, 1997). Alike 'high', this value is more frequent in TED talks with examples such as *slowly began, less leader, little bit of growth, less needed*, but present in political speeches as well with a few interesting examples: *scarcely imagine, slow undoing, less inventive*.

As far as the 'reverse' value is concerned, it is more frequent in political speeches. In general the most frequent occurrence is "without + noun" (e.g. *without legitimacy*) or "without + verb'' (e.g. *without lessening*). However, also a considerable number of verbs such as *combat, fight, abolish, undermine, stop, halt, oppose* have the reverse function (see also the related description of the attribute *force* in Chapter 7). The much higher percentage in Russian in comparison to the other languages is due to both the abundance in news and political speeches (19 and 40 occurrences respectively).

Target type. The most common type is 'thing' in all languages and also assigned in similar percentages: 840 times in English corresponding to 82%, 746 times in Italian corresponding to 84% and 728 times in Russian corresponding to 79% (see Figure 6.15). Some of the most common targets are, in fact, *world, rights, growth, power, war, ideas, cooperation, freedom*. The second most common is 'person', assigned 157 times in English corresponding to 15%, 116 times in Italian corresponding to 13% and 156 times in Russian corresponding to 17%, and has at the top of the list *we, they, people, you*. The second to last is 'place', assigned 23 times in English, 13 times in Italian and 34 times in Russian, and the last one is 'other', assigned mostly in the case of abstract concepts such as *urgency, time, transparency* 4 times in English, 12 times in Italian and 2 times in Russian. They both account for 1-4% and 0-1% respectively.

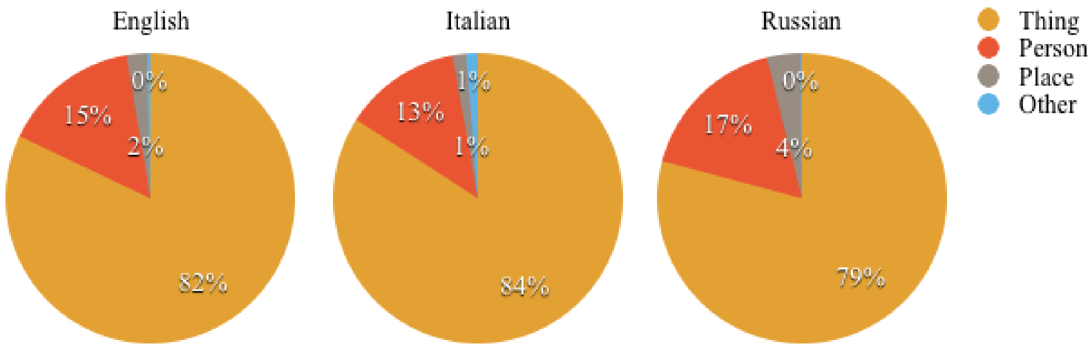

Figure 6.15: Comparison of the values for the target types in their percentage form across languages.

The statistics shown in this Chapter will be useful to conduct the final evaluation of the automatic system in Chapter 8. However, first the system will be described in the following Chapter.

Chapter 7

Automatic annotation: the pipeline

This Chapter will be entirely dedicated to the description of the modules that constitute the automatic system used in this research.

The system takes in input individual sentences (from the command line), files or directories, and extracts modifiers (along with the attributes *orientation, attitude, force* and *polarity*), targets (along with the attributes *orientation, type*) and appraisal groups (along with the attribute *orientation*). In the case of the texts belonging to the *SentiML* corpus, they have been manually aligned across the three languages on the sentence level, by keeping concepts together even in case of different punctuation across the three languages, in order to facilitate the comparison.

The pipeline consists of six modules (Figure 7.1):

1. **POS tagger**: This module applies part-of-speech (POS) tags, by using *TreeTagger* (Schmid, 1994)[1]. First it processes the input in order to have one word per row, and then applies the tagging. The result is the word in the first column, the POS in the second column and the lemma in the third column.

2. **CoNLL formatter**: This module adds some fields to the POS-tagging format required by the CoNLL format [2] explained in Table 7.1.

3. **Dependency parser**: This module applies *Maltparser* [3] to the sentence and populate the CoNLL format with actual values. In order to apply the parsing, models trained in advance on the language are necessary.

[1] `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`
[2] `http://ilk.uvt.nl/conll/`
[3] `www.maltparser.org`

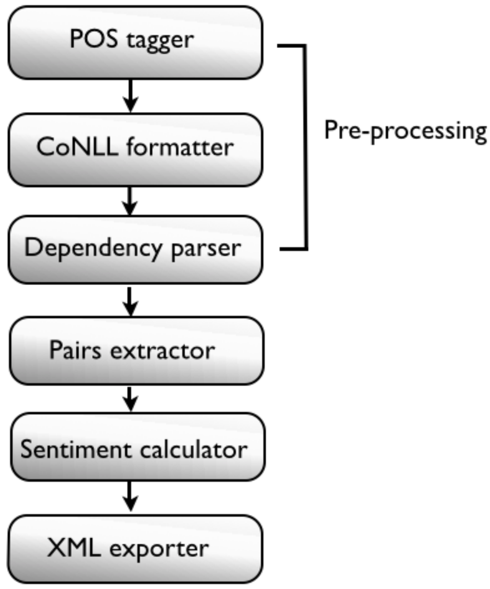

Figure 7.1: The pipeline to extract appraisal groups and their attributes takes in input either TXT documents or individual sentences, and returns XML documents.

For English, one of the available models trained on the Wall Street Journal section of the Penn Treebank and QuestionBank has been used[1], whereas Italian and Russian models had to be created *ex novo*[2] and have been made available for download[3].

The Italian model has been trained on the corpus PAISÀ (Lyding *et al.*, 2014), a collection of about 380 thousand Italian web texts coming from 1000 websites. PAISÀ consists of about 250 million tokens fully annotated in CoNLL format and is both searchable online or available for download[4].

Afterwards, because the tag-set used in PAISÀ is *ISST-Tanl* [5], but the one of the POS-tagger and MaltParser is TreeTagger[6], a conversion has been necessary in order to parse each sentence and populate the CoNLL (described in Table 7.1). Ta-

[1] `http://www.maltparser.org/mco/english_parser/engmalt.html`.The model is *English engmalt.linear.1.7.mco* and uses linear SVMs

[2] The 1.7.2 version of MaltParser available at `http://www.maltparser.org` has been used

[3] `http://corpus.leeds.ac.uk/marilena/index.html`

[4] `http://www.corpusitaliano.it/it/contents/description.html`

[5] `http://medialab.di.unipi.it/wiki/Tanl_POS_Tagset`

[6] `ftp://ftp.ims.uni-stuttgart.de/pub/corpora/italian-tag-set.txt`

N.	Name	Description
1	ID	Token counter, starting at 1 for each sentence
2	FORM	Word form
3	LEMMA	Lemma
4	CPOSTAG	Coarse-grained part-of-speech tag
5	POSTAG	Fine or coarse-grained part-of-speech tag
6	FEATS	Any features, here replaced by underscore
7	HEAD	Head of the current token
8	DEPREL	Dependency relation to HEAD
9	PHEAD	Projective head, here replaced by underscore
10	PDEPREL	Dependency relation to PHEAD, here replaced by underscore

Table 7.1: CoNLL format. For each of the ten fields, the position, the name and the description are provided.

ble 7.2 reports such mapping, for which I took into consideration the fine-grained rather than the coarse-grained tags, although these were all mostly consisting of one letter.

Despite this different granularity, the TreeTagger tag-set did not have some tags: first, those for main, modal and auxiliary verbs (while in Tanl there is a tag for each of them); second, the tag for the past participle, which would have contributed to a more accurate automatic match of the pairs with respect to the manual annotation, especially because no modal and auxiliary verbs have been annotated (a detailed discussion will be provided in the following module, n.4, "Pairs extractor"); third, the tag for the negative adverb *non*. Conversely, different tags are available in the Tanl tag-set whether a punctuation mark is located in the sentence (FF) or at the end of the sentence (FS). In this last case, I decided to replace them with F in all the cases.

POS-tag		Explanation	POS-tag		Explanation
TreeTagger	**Tanl**		**TreeTagger**	**Tanl**	
ABR	SA	abbreviation	PRO:refl	PC	reflexive pronoun
ADJ	A	adjective	PRO:rela	PR	relative pronoun
ADV	B	adverb	VER:cimp	V	verb subjunctive imperfect
CON	C	conjunction	VER:cond	V	verb conditional
DET:def	RD	definite article	VER:cpre	V	verb subjunctive present
DET:indef	RI	indefinite article	VER:futu	V	verb future tense
FW	S	foreign word	VER:geru	V	verb gerund
NOM	S	noun	VER:impe	V	verb imperative
NPR	SP	name	VER:impf	V	verb imperfect
NUM	N	numeral	VER:infi	V	verb infinitive
PON	F	punctuation	VER:pper	V	verb participle perfect
PRE	E	preposition	VER:ppre	V	verb participle present
PRE:det	EA	preposition+article	VER:pres	V	verb present
PRO	P	pronoun	VER:refl:infi	V	verb reflexive infinitive
PRO:demo	PD	demonstrative pronoun	VER:remo	V	verb simple past
PRO:indef	PI	indefinite pronoun	SENT	FS	sentence marker
PRO:inter	PQ	interrogative pronoun	SYM	X	symbol
PRO:pers	PE	personal pronoun	INT	I	interjection
PRO:poss	PP	possessive pronoun	LS	X	list symbol

Table 7.2: Mapping between TreeTagger and Tanl tag-sets

The Italian parsing model has been built on more than 2% of PAISÀ (i.e. 250 thousand sentences with an average of 22 words per sentences out of 13.1 million) because of limitations in the RAM memory of the machine used to build the model, and the presence of sentences with wrong parsing trees mainly because of missing ID numbers.

I also solved the problems with Unicode in a way that the system would accept apostrophes, accents, interrogation marks without need of escaping them (exclamation marks and quotes are an exception).

The Russian MaltParser model also uses the TreeTagger POS tagger, and it is based on the *MTE* tag-set[1] and the *SynTagRus* corpus (Sharoff & Nivre, 2011).

4. **Pairs extractor**: This module finds in the parsed sentence the pairs of modifiers and targets corresponding to the following combinations:

   - A noun with an adjective. For example, "$[\text{good}]_{\text{M}}$ $[\text{plan}]_{\text{T}}$".
   - A pronoun with an adjective. For example, in a relational clause like "$[\text{they}]_{\text{T}}$ $[\text{brave}]_{\text{M}}$".
   - A verb with an adverb. For example, "$[\text{strongly}]_{\text{M}}$ $[\text{support}]_{\text{T}}$".
   - A noun with a verb. For example, "$[\text{children}]_{\text{T}}$ $[\text{love}]_{\text{M}}$".
   - A pronoun with a verb. For example, "$[\text{she}]_{\text{T}}$ $[\text{smiles}]_{\text{M}}$".

   With respect to the annotation scheme, the automatic system is not able to retrieve nouns linked by a preposition and multi-word units (see Section 5.5).

   To do so, it relies on the POS-tag information previously extracted. It returns a result **only if the modifier or the target are found in the dictionary**. In addition, lemmas rather than word forms are taken.

   The above rules have been applied to all languages, but the tagsets available were of different granularity (in Italian both *TreeTagger* and *Tanl* have been used). Tables 7.3, 7.4 and 7.5 show examples in English of each combination, along with the exact POS-tags for each language.

[1] http://corpus.leeds.ac.uk/mocky/ru-table.tab

Pair (and example)	English		Russian		Italian	
	POS-tags	**Explanation of POS-tags**	**POS-tags**	**Explanation of POS-tags**	**POS-tags**	**Explanation of POS-tags**
Adjective+Noun						
Beautiful/ more beautiful/ the most beautiful + overview/ overviews/ Anna/ opera	JJ, JJR, JJS + NN, NNS, NP, NPS, FW	standard, comparative, superlative + common singular, common plural, proper singular, proper plural, foreign	A+N	all types+ all types	ADJ+ NOM, NPR, FW ****** S, SP+ V	all types+ common, proper, foreign ****** common, proper+ all types
Noun + Verb						
Boy/boys/ Laura/data + outperform/ outperforms/ outperformed/ are outperforming	NN, NNS, NP, NPS, FW + VB, VBD, VBG, VBN, VBP, VBZ	common singular, common plural, proper singular, proper plural, foreign + present, past, gerund, past participle	N+V	all types+ all types	NOM, NPR, FW+ VER, VER:cpre, VER:cimp, VER:cond, VER:pres, VER:remo, VER:impf, VER:futu, VER:impe, VER:geru, VER:infi, VER:refl:infi, VER:pper, VER:ppre ****** S, SP+ V	common, proper, foreign+ base, subjunctive present, subjunctive imperfect, conditional, present, simple past, imperfect, future, imperative, gerund, infinitive, reflexive, past participle, present participle ****** common, proper+ all types

Table 7.3: Grammatical categories of the appraisal groups with examples and POS-tags for each language (part 1). In Italian first *TreeTagger* then *Tanl* POS-tags are shown.

Pair (and example)	**English**		**Russian**		**Italian**	
	POS-tags	**Explanation of POS-tags**	**POS-tags**	**Explanation of POS-tags**	**POS-tags**	**Explanation of POS-tags**
Pronoun +Adjective						
He/they + happy/ happier/ the happiest	PP + JJ, JJR, JJS	personal + standard, comparative, superlative	P+A	all types+ all types	PRO, PRO:demo, PRO:indef, PRO:pers, PRO:inter, PRO:poss, PRO:refl, PRO:rela+ ADJ ****** P, PD, PI, PE, PQ, PP, PR, PC + A	general, demonstrative, indefinite, personal, interrogative, possessive, reflexive, relative+ all types ****** general, demonstrative, indefinite, personal, interrogative, possessive, relative, locative + all types
Verb + Adverb						
Live/lived/ is living + actively/ more actively/ most actively	VB, VBP, VBZ, VBD, VBG, VBN + RB, RBR, RBS	base, present, 3rd person singular present, past, gerund, past participle + simple, comparative, superlative	V+R	all types+ all types	VER, VER:cpre, VER:cimp, VER:cond, VER:pres, VER:remo, VER:impf, VER:futu, VER:impe, VER:geru, VER:infi, VER:refl:infi, VER:pper, VER:ppre+ ADV ****** V+ R	base, subjunctive present, subjunctive imperfect, conditional, present, simple past, imperfect, future, imperative, gerund, infinitive, reflexive, past participle, present participle+ all types ****** all types + all types

Table 7.4: Grammatical categories of the appraisal groups with examples and POS-tags for each language (part 2). In Italian first *TreeTagger* then *Tanl* POS-tags are shown.

Pair (and example)	English		Russian		Italian	
	POS-tags	**Explanation of POS-tags**	**POS-tags**	**Explanation of POS-tags**	**POS-tags**	**Explanation of POS-tags**
Pronoun+ Verb						
She/we + smile/ smiled/ are smiling	PP + VB, VBP, VBZ, VBD, VBG, VBN	personal + base, present, 3rd person singular present, past, gerund, past participle	P+V	all types+ all types	PRO, PRO:demo, PRO:indef, PRO:pers, PRO:inter, PRO:poss, PRO:refl, PRO:rela + VER, VER:cpre, VER:cimp, VER:cond, VER:pres, VER:remo, VER:impf, VER:futu, VER:impe, VER:geru, VER:infi, VER:refl:infi, VER:pper, VER:ppre ****** P, PD, PI, PE, PQ, PP, PR, PC + V	base, demonstrative, indefinite, personal, interrogative, possessive, reflexive, relative+ base, subjunctive present, subjunctive imperfect, conditional, present, simple past, imperfect, future, imperative, gerund, infinitive, reflexive, past participle, present participle ****** general, demonstrative, indefinite, personal, interrogative, possessive, relative, locative + all types

Table 7.5: Grammatical categories of the appraisal groups with examples and POS-tags for each language (part 3). In Italian first *TreeTagger* then *Tanl* POS-tags are shown.

In the case of targets, the Pairs Extractor module fills their *type* attribute by searching the most likely synset in *Wordnet*[1]. The English Wordnet has been integrated in the pipeline by using the *Wordnet* library from the NLTK library for Python [2]. Since there are no correspondent libraries for Italian and Russian, I first thought

[1] `http://wordnet.princeton.edu/`

[2] `http://www.nltk.org/api/nltk.corpus.html`

of using the Italian and Russian *Swadesh wordlists* [1] (Abney & Bird, 2010) in the "Nltk corpus package" (Bird, 2006). However, because of their small coverage (only about 200 words), I translated the words in the Italian and Russian pairs into English [2], and then matched them to their synsets in the English Wordnet.

According to the type of synsets, the target *type* and the modifier *attitude* are computed in sequence: the synsets 'person', 'entity', 'location' are associated to one of the values specified for the target type in the annotation scheme, i.e. 'person', 'thing', 'place' respectively, while 'action' is assigned to verbs; afterwards, the following rules assign to each of these target types an attitude ('affect', 'judgement', 'appreciation') (see Table 7.6 for the summary fo the rules):

- If the lemma is *I, we* or its translations *io, noi,* я, мы, the type is 'person' and the attitude is 'affect'.
- In any other case, if the type is 'person', the attitude is 'judgement'.
- If the type is 'action', the attitude is 'judgement'.
- If the type is 'thing' or 'place', the attitude is 'appreciation'.

Wordnet synset/POS tag	**Type**	**Attitude**
Person	Person	Judgement
Verb	Action	Judgement
I/We	Person	Affect
Entity	Thing	Appreciation
Location	Place	Appreciation

Table 7.6: Chains of *Wordnet* synsets (or POS tags), types and attitudes as computed by the system.

5. **Sentiments identifier**: This module filters the couples that actually bring sentiment and, for these, calculates the overall sentiment. Here the sentiment value is taken from the *NRC Word-Emotion Association Lexicon* (Mohammad, 2011), whose annotations were manually undertaken through "Amazon's Mechanical Turk", and the "Roget Thesaurus"[3].

[1] `http://www.nltk.org/book/ch02.html`

[2] The translation was carried on by using Google translate available at `https://translate.google.co.uk/`

[3] `http://www.gutenberg.org/ebooks/10681`

As mentioned in Chapter 2, the lexicon was chosen among others because it has entries for approximately 24200 word–sense pairs, corresponding to 14200 word types. In addition, it is not specific for domain such as the *Lexicoder Sentiment Dictionary* for the political domain (Young & Soroka, 2012), the *Opinion Lexicon* for customers reviews (Hu & Liu, 2004), *AFINN* for microblogs (Nielsen, 2011) and *ANEW* with emotion words (Bradley & Lang, 1999) or *SentiWordnet* (Esuli & Sebastiani, 2006).

I built a *json* lexicon in which each English key is associated to its feature set (taken from the *NRC dictionary*). The Russian and Italian versions of the lexicon have been populated with the translations of the English words. The translation has been carried out by using Google translate[1].

Despite the good quality of the equivalents proposed considering the lack of context, I had to manually correct some mistakes:

- Infinitive verbs translated as conjugated (e.g. *meritano* (they deserve) instead of *meritare* (to deserve)) or nouns (e.g. *riproduzione* (reproduction) instead of *riprodurre* (to reproduce)), due to the fact that infinitives do not contain *to* in the English version of the *NRC dictionary*.
- English versions left in the translations (e.g. *decanter* instead of *caraffa*) or not adapted (e.g. *demos* changed into *demo*).
- Duplicates, for which I decided to always take the first occurrence.

I also counted the words included in the dictionary for which both the values 'negative' and 'positive' would be false (although their other attributes for emotions would not be), so that would be considered 'neutral' and they were about 8,000 out of 14,000.

Finally I checked that the polarity reversals would not be included in the dictionary because they are not supposed to be encapsulated in the appraisal groups according to the guidelines in Chapter 5.

The first task of this module, which is to filter the couples that actually bring sentiment, consists of two stages:

(a) Checking if the modifier and/or the target are in dictionary.

[1] `https://translate.google.co.uk/`

(b) Taking the sentiment of those retrieved in the dictionary.

If none of the words are in dictionary, the couple is not chosen.

The second task, which is to get the sentiment of the appraisal group, is carried out by performing the calculations in Table 7.7. The resulting sentiment of the appraisal group is the product of those of the modifier and the target, provided by the *NRC dictionary*. For example, if both the modifier and the target are positive/negative/neutral/ambiguous, the appraisal group will be assigned a positive/negative/neutral/ambiguous value; if one of them is negative, the appraisal group will be assigned a negative value; if one of them is neutral/ambiguous, the appraisal group will be assigned the value of the other, i.e. positive or negative; and finally, if the pair consists of neutral and ambiguous, the the appraisal group will be assigned an ambiguous value.

Modifier	**Target**	**Appraisal group**
Positive	Positive	Positive
Positive	Negative	Negative
Negative	Positive	Negative
Negative	Negative	Negative
Neutral	Neutral	Neutral
Positive	Neutral	Positive
Neutral	Positive	Positive
Negative	Neutral	Negative
Neutral	Negative	Negative
Positive	Ambiguous	Positive
Ambiguous	Positive	Positive
Negative	Ambiguous	Negative
Ambiguous	Negative	Negative
Neutral	Ambiguous	Ambiguous
Ambiguous	Neutral	Ambiguous
Ambiguous	Ambiguous	Ambiguous
Not in dictionary	In dictionary	Modifier’s sentiment
In dictionary	Not in dictionary	Target’s sentiment

Table 7.7: Calculations of the overall sentiment of couple of modifiers and targets, performed by the module *Sentiments*.

6. **XML exporter**: This is not a separate module, but it rather represents the final part of the script that runs the entire pipeline.

   In order for the final XML document to resemble the output required by the annotation tool MAE, it was necessary to add the fields for the ID and the start/end of the segments to the DTD (document type definition) that I created and used for my own annotation (e.g. *fromText toText, start and end*) [1].

   Each XML document (either standard output of MAE or exported) contains the raw text in the first part, and the annotations in the second part. For targets, the attributes in output are *type* and *orientation*: the *type* is taken from *Wordnet* (see Module 4, *Pairs*), whereas the *orientation* is taken from the dictionary (see Module 5, *Sentiments*).

   For modifiers the attributes in output are:

   - *Attitude*, (i.e. 'affect', 'judgement' and 'appreciation'), associated to the target *type* (see Module 4, *Pairs*).
   - *Force*, presence of an adverb of intensity, with values 'normal', 'low, 'high', 'reverse'. In the three languages the rules for the activation of the values 'low' and 'high' are based on 13 adverbs (e.g. *very, lot, definitely, extremely* vs *less, little, poorly, slightly*), along with the case in which the current lemma was *good* and *bad* in their comparative/superlative form (i.e. *better, best, worse, worst*). Instead, for the value 'reverse' a list of 14 verbs/nouns (e.g. *change, abolish, stop, oppose*) has been considered, first only according to the description of the annotation scheme in previous works and then populated with examples found during the annotation; these items appear 24 times in the corpus.
   - *Polarity*, presence of negation, with values 'marked', 'unmarked'. The value assigned by the system is 'marked' only when the previous word is the lemma *not*. No more complex rules have been designed, in order to be general for all the languages.
   - *Orientation*, taken from the dictionary (see Module 5, *Sentiments*).

[1] The DTD is available for download at `http://corpus.leeds.ac.uk/marilena/SentiML/SentiML.dtd`

For appraisal groups, the attribute in output is *orientation* that is calculated on the basis of the orientation of the modifier and the target (see Module 5, *Sentiments*).

The pipeline has been tested on the corpora described in Section 1.4 and whose statistics have been provided in Section 6.2. In Chapter 8 I will present the results of the evaluation.

Chapter 8

Automatic annotation: results

In this chapter I will conduct a full evaluation of the system described in Chapter 7 both on the corpora described in Section 1.4 and on non-annotated data from the same text types. The evaluation, which will be related to the performances in the identification of the appraisal groups and their classifications, will serve the purpose of answer the research questions formulated in Section 1.3.

8.1 Performance measures

For each category, precision, recall and F1 measure have been measured in both a strict and lenient way. In order to be included in the lenient measurement, the modifiers and the targets recognized by the system had to overlap with the annotated ones at least for one character. This measurement has been particularly useful in the case of multi-word expressions (most times annotated as modifiers) because, when in the manual annotation an expression consisting of two or more words (next to each other) was annotated (e.g. *disagree with*), but only one word was annotated by the system (e.g. *disagree*), the lenient measure would count that as a match. As for the appraisal groups, the lenient measurement would include those consisting of a modifier and/or a target that, in turn, have been included in the lenient measurement (e.g. *disagree with him*). A similar reasoning is applied to the appraisal groups in the case of the strict measurement.

In the case of the measurement of precision, recall and F1 for a system with several sets of data (in my case political speeches, news and TED talks), in the literature two methods are used (Manning *et al.*, 2008):

- **Micro-average**, in which the individual true positives (TP), false positives (FP) and false negatives (FN) of the system for different sets are summed up:
  - TP1, FP1 and FN1
  - TP2, FP2 and FN2
  - TP3, FP3 and FN3

  Then the calculations are performed as follows:
  - Micro-average of precision = $\frac{TP}{TP+FP}$
    where $TP = \sum TP_i$ and $FP = \sum FP_i$
  - Micro-average of recall = $\frac{TP}{TP+FN}$
    where $FN = \sum FN_i$
  - Micro-average F1-Score = $2 * \frac{precision*recall}{precision+recall}$

  This method can be a useful measure when the sets are composed of heterogeneous documents (like in my case in which news are much shorter than political speeches and TED talks), because it is as if the entire corpus was in one big document.

- **Macro-average**, in which the average of the precision (P) and recall (R) of the system on different sets is considered:
  - Macro-average precision = $\frac{P}{3}$
    where $P = \sum P_i$
  - Macro-average recall = $\frac{R}{3}$
    where $R = \sum R_i$
  - Macro-average F1-Score = $2 * \frac{precision*recall}{precision+recall}$

  Since macro-averaging gives equal weight to each class (by doing the micro-averaging of each set and then the mean among them), this method is strongly influenced by the size of each class.

For the reason explained above, I decided to present in this Chapter all the results related to the micro-average, while those related to the macro-average can be found in Appendix A.

8.2 Results on annotated data

I will now present and discuss the results individually for each language (in the first place in English, followed by Russian and Italian), and then conclude with a comparison of the three. The results will be related to the performances in the identification of appraisal groups, and the classification of all the attributes (i.e. *orientation, attitude, force, polarity, type*), both on the overall datasets as sum of the text types (i.e. political speeches, news and TED talks) and on them individually.

However, a note concerning the grammatical combinations used by the system has to be made first. In order to fulfil the greater scope of linguistically supporting the automatic analysis of evaluative language, one of the practical tasks has been that of automatically matching the rules used by the system to the manual annotations as much as possible. The rules are: "adjective + noun", "noun + verb", "pronoun + adjective", "pronoun + verb", "verb + adverb". Tables 7.3 and 7.4 show examples for each of the rule. However, as mentioned in the "Pairs extractor" module in Section 7, two differences between the system and the manual annotations are represented by the fact that there is no rule for the system to annotate (i) nouns linked by a preposition (e.g. *acts of courage*) and that (ii) multi-word units are hardly spotted by the system.

The rules have been chosen because all strongly linguistically motivated, including "verb + adverb", which I found particularly challenging to apply even during the manual annotation. The reason was that groups matching this rule had to be annotated only when the adverb could not have been replaced by 'high' or 'low' force, for example in the case of *comprehensively, safely, tirelessly, grudgingly, gladly, peacefully*.

After carrying out the manual annotations, I provided to the system a list of these adverbs (available in Section 5.4.1), although I expected the system to have difficulty because it would have been easy for example to identify *comprehensively conjugate* even though not carrying appraisal (as opposed to *comprehensively blame, comprehensively illustrate, comprehensively beaten*).

When, after trying different clusters of the grammatical combinations mentioned above, I found that the best set in terms of F1 measure was generally the one with all the rules except for "verb + adverb", I decided that it was fair to allow the system not to use it. The testing corpus still had appraisal groups matching it, so both the linguistic choices behind it and the evaluation process were not compromised.

Other two considerations that have to be done are that I will give more focus to the appraisal groups category since the results on targets and modifiers are of importance only if related to them, and that I will always refer to the lenient values.

8.2.1 English

I will start by discussing the results related to the identification phase in English. Table 8.1 shows the performances obtained by the automatic system in the identification of modifiers, targets and appraisal groups on the overall English dataset. Performances are quite high in terms of F1 (0.32). In the case of the appraisal groups, it is interesting to notice that precision (0.41) is higher than recall (0.26), which suggests that the grammatical combinations have an influence.

	Overall	
	Strict	Lenient
	MODIFIERS	
Precision	0.49	**0.51**
Recall	0.31	**0.32**
F1	0.38	**0.39**
	TARGETS	
Precision	0.57	**0.57**
Recall	0.33	**0.33**
F1	0.42	**0.42**
	APPRAISAL GROUPS	
Precision	0.40	**0.41**
Recall	0.25	**0.26**
F1	0.31	**0.32**

Table 8.1: Identification of modifiers, targets and appraisal groups in the overall English dataset.

Table 8.2 shows the performances for each text type. Political speeches obtain the best results, with a precision for appraisal groups (0.47) higher than the average (0.41), due to both high precision in the identification of modifiers and targets (0.59 and 0.65 respectively). One of the reasons for such a high precision might be the redundancy of some vocabulary in this text type compared to news and TED talks. In particular, as mentioned in Section 1.4, the news belong to the human rights section of the MPQA corpus but have different focus, and the TED talks are on two completely different topics (the power of images and economics). In fact, the F1 for news is slightly lower (0.27) than for political speeches (0.33) and TED talks (0.30).

	Political		News		TED	
	Strict	Lenient	Strict	Lenient	Strict	Lenient
	MODIFIERS					
Precision	0.56	**0.59**	0.24	**0.28**	0.37	**0.38**
Recall	0.31	**0.32**	0.27	**0.32**	0.31	**0.32**
F1	0.40	**0.41**	0.25	**0.30**	0.34	**0.35**
	TARGETS					
Precision	0.65	**0.65**	0.42	**0.44**	0.44	**0.44**
Recall	0.33	**0.33**	0.46	**0.50**	0.29	**0.29**
F1	0.44	**0.44**	0.44	**0.46**	0.35	**0.35**
	APPRAISAL GROUPS					
Precision	0.45	**0.47**	0.22	**0.25**	0.32	**0.33**
Recall	0.25	**0.25**	0.26	**0.30**	0.26	**0.27**
F1	0.32	**0.33**	0.23	**0.27**	0.29	**0.30**

Table 8.2: Identification of modifiers, targets and appraisal groups across text types in English.

Moving to the discussion of the attributes for each category, Table 8.3 reports the results of the automatic identification of the orientation of modifiers, targets and appraisal groups on the overall dataset.

Interestingly, while the classification of the *orientation* ('neutral', 'positive', 'negative', 'ambiguous') of individual words such as modifiers and targets gives low results (0.19 and 0.23 respectively), the value increases quite considerably for the appraisal groups by reaching 0.46. The same pattern can also be seen in the specific text types in Table 8.4: the accuracy in the *orientation* of the appraisal groups is 0.48 in political speeches, 0.46 in news and 0.40 in TED talks.

The attribute *attitude* has also a particular high value with 0.81 as average and its maximum in news with 0.86 as opposed to 0.77 in political speeches and 0.83 in TED talks. Such accuracy is due to the connection with *type* (i.e. 'person', 'action, 'thing', 'other') shown in Table 7.6: first the target type ('person', 'thing', 'place', 'other') is classified by connecting the word to its synset in Wordnet; afterwards to each target *type*, an *attitude* is associated ('affect', 'judgement' or 'appreciation') by using the following rules:

	Overall	
	Strict	Lenient
	MODIFIERS	
Orientation	0.19	**0.19**
Attitude	**0.81**	0.81
Force	**0.88**	0.88
Polarity	**0.98**	0.98
	TARGETS	
Orientation	**0.23**	0.23
Type	**0.91**	0.91
	APPRAISAL GROUPS	
Orientation	0.45	**0.46**

Table 8.3: Attributes of modifiers, targets and appraisal groups in the English overall dataset.

- If the lemma is *I, we* (or their translations *io, noi,* я, мы), the type is 'person' and the attitude is 'affect'.
- In any other case, if the type is 'person', the attitude is 'judgement'.
- If the type is 'action', the attitude is 'judgement'.
- If the type is 'thing' or 'place', the attitude is 'appreciation'.

Type performs very well with 0.91 for the overall dataset, as we see from Table 8.3.

As for the other modifiers' attributes, they also perform very well: *force* (i.e. presence of an adverb of intensity, so 'normal', 'low, 'high', 'reverse') and *polarity* (i.e. presence of negation, so 'marked', 'unmarked) have 0.88 and 0.98 respectively for the overall dataset (Table 8.3), with no substantial differences across the different text types (Table 8.4). In terms of *force*, the adverbs most contributing to the classification are the expected ones, namely *more* (e.g. "more often men and women obscure, more the icy currents") and *most* (e.g. "the most prosperous nation, the most sacred oath") for the value 'high', and *less* (e.g. "services no less needed") for the value 'low'.

As (De Haan, 2002) points out, we have to consider the cases in which the negation не is placed before the modal (not before the main verb), which is covered by the annotation scheme, for example не должна стать (it should not be), не должна превратиться (it should not become). There is also no occurrence of the modal нельзя (cannot).

Apart from the individual accuracy of the attributes, I was interested in seeing how

	Political		News		TED	
	Strict	Lenient	Strict	Lenient	Strict	Lenient
	MODIFIERS					
Orientation	0.19	**0.19**	0.19	**0.21**	0.19	**0.19**
Attitude	**0.78**	0.77	**0.87**	0.86	**0.83**	0.83
Force	**0.89**	0.88	0.93	**0.93**	**0.84**	0.83
Polarity	**0.98**	0.97	0.96	**0.97**	0.98	**0.98**
	TARGETS					
Orientation	**0.23**	0.23	**0.38**	0.35	**0.11**	0.11
Type	0.89	**0.89**	**0.91**	0.91	**0.92**	0.90
	APPRAISAL GROUPS					
Orientation	0.47	**0.48**	0.46	**0.46**	0.40	**0.40**

Table 8.4: Attributes of modifiers, targets and appraisal groups across text types in English.

the attributes would connect among each other and whether there was any positive contribution to the classification of the *orientation* of the appraisal groups as specified in the "Aims and Objectives" in Chapter 1. For this reason, I designed the system in a way that the *orientation* of the appraisal group would have been swapped (i.e. 'positive' into 'negative', and vice versa) when:

- The value of *force* was 'reverse' because of the presence of "reversals" such as *abolish, stop, halt, without, against*.
- The value of *polarity* was 'marked' because of a negation

For my experiments, I set the value of *force* on 'normal' and of *polarity* on 'unmarked' (which would also mean that the *orientation* would not have been swapped), and compared the performances of the two settings. Unfortunately, it turned out that no difference was visible due to the fact that none of the conditions for the rules were present in the identified appraisal groups and so the rules did not fire.

8.2.2 Russian

I will start by discussing the results related to the identification phase in Russian. Table 8.5 shows the identification of modifiers, targets and appraisal groups on the overall Russian dataset.

	Overall	
	Strict	Lenient
	MODIFIERS	
Precision	**0.50**	0.47
Recall	0.34	**0.36**
F1	0.41	**0.41**
	TARGETS	
Precision	**0.42**	0.40
Recall	0.31	**0.33**
F1	0.36	**0.36**
	APPRAISAL GROUPS	
Precision	**0.45**	0.44
Recall	0.29	**0.31**
F1	0.35	**0.37**

Table 8.5: Identification of modifiers, targets and appraisal groups in the overall Russian dataset.

Performances are comparable to those in English (shown in Table 8.1) in terms of appraisal groups, since both languages have precision higher than recall. The peaks in Russian are even higher: 0.44 for precision (vs. 0.41), 0.31 for recall (vs. 0.26) and 0.37 for F1-measure (vs. 0.32). This is likely to be due to the accuracy in identifying the noun phrases at the parsing stage. In fact, the dependency parser uses a fine-grained tag-set at this stage that works well with the morphology in Russian.

While a frequent problem found with the English parsing is with copulas since the ROOT is usually assigned to the adjective instead of the verb (for example in *our health-care is too costly*, *costly* is ROOT and every word is directly connected to it), frequently there is not such problem in Russian, in which *is* is omitted, and the ROOT function is assigned to either a noun or an adjective, in this case in Система (слишком) дорого, Система (system) is the ROOT, and the appraisal group Система дорого (system costly) is successfully identified.

Other times, even if the English dependency tree is correct, the group is not retrieved, e.g. in *they will not be met easily, petty grievances*, while in Russian the opposite would happen: мелким обидам was chosen because they were connected despite обидам (grievances) being ROOT.

This demonstrates that the fact that the rules used by the system have been written by using coarse-grained POS-tags makes no difference.

Table 8.6 shows the identification of modifiers, targets and appraisal groups across text types. The highest result in F1 measure is achieved by TED talks with 0.40 vs. 0.37 in political speeches and 0.23 in news. In particular for news this depends on the much lower precision being 0.18 vs. 0.30 in political speeches and 0.37 in TED talks.

	Political		News		TED	
	Strict	Lenient	Strict	Lenient	Strict	Lenient
	MODIFIERS					
Precision	0.55	**0.56**	0.20	**0.21**	0.32	**0.44**
Recall	0.35	**0.35**	0.32	**0.36**	0.26	**0.38**
F1	0.43	**0.43**	0.24	**0.27**	0.29	**0.41**
	TARGETS					
Precision	0.47	**0.47**	0.23	**0.23**	0.35	**0.39**
Recall	0.31	**0.31**	0.38	**0.41**	0.31	**0.36**
F1	0.37	**0.37**	0.29	**0.30**	0.33	**0.38**
	APPRAISAL GROUPS					
Precision	0.50	**0.51**	**0.19**	0.18	0.34	**0.43**
Recall	0.29	**0.30**	0.30	**0.31**	0.26	**0.37**
F1	0.37	**0.37**	**0.23**	0.23	0.29	**0.40**

Table 8.6: Identification of modifiers, targets and appraisal groups across text types in Russian.

As for the attributes of modifiers, targets and appraisal groups on the overall dataset, they are shown in Table 8.7.

In terms of *orientation*, results are comparable to English since the appraisal groups have 0.38 vs. 0.46 in English, despite the low results for modifiers and targets (0.14 and 0.21 respectively). Also in this case the performances on the individual text types shown in Tables 8.8 do not differentiate much.

Moving to the *attitude* (i.e. 'affect', 'judgement' and 'appreciation'), the good connection between this attribute and *type* is confirmed in this language as well, since the accuracy for *attitude* is 0.68 and for type 0.80. In this sense, we must also take into account two limitations: the first is the lack of a Wordnet for Russian, either because not available like in the case of *RussNet* (Azarowa, 2008) or *Yet Another RussNet* (Braslavski *et al.*, 2014), or because not having a complete status like in the case of *Russian Wordnet* (Gelfenbeyn *et al.*, 2003).

The second is the fact that the Python library *Swadesh* (Abney & Bird, 2010) in the

	Overall	
	Strict	Lenient
	MODIFIERS	
Orientation	**0.15**	0.14
Attitude	0.65	**0.68**
Force	0.83	**0.84**
Polarity	0.96	**0.97**
	TARGETS	
Orientation	0.20	**0.21**
Type	0.78	**0.80**
	APPRAISAL GROUPS	
Orientation	0.38	**0.38**

Table 8.7: Attributes of modifiers, targets and appraisal groups in the Russian overall dataset.

Nltk corpus package (Bird, 2006) was not complete enough to cover the words included in the appraisal groups of my corpora. For this reason, I had to automatically translate the appraisal groups into English[1] in order to use the English Wordnet.

The other modifiers' attributes also perform very well. In Table 8.7, the highest values are for *force* and *polarity* being 0.84 and 0.97 respectively.

	Political		News		TED	
	Strict	Lenient	Strict	Lenient	Strict	Lenient
	MODIFIERS					
Orientation	**0.16**	0.15	0.11	**0.11**	**0.20**	0.12
Attitude	**0.61**	0.61	**0.78**	0.75	0.68	**0.79**
Force	0.86	**0.86**	**0.84**	0.84	0.60	**0.79**
Polarity	0.95	**0.95**	**0.98**	0.97	0.96	**0.98**
	TARGETS					
Orientation	0.21	**0.21**	**0.18**	0.15	0.15	**0.25**
Type	0.79	**0.79**	**0.80**	0.79	0.69	**0.82**
	APPRAISAL GROUPS					
Orientation	**0.38**	0.38	**0.42**	0.37	0.31	**0.39**

Table 8.8: Attributes of modifiers, targets and appraisal groups across text types in Russian.

[1]The translation was carried on by using Google translate available at `https://translate.google.co.uk/`

As for the contribution of *force* and *polarity* to the *orientation* of the appraisal groups, because of the lack of any spotted modifier to the lemma не/нет or the reversal adverbs, the rules did not fire.

8.2.3 Italian

In the case of Italian, since some important information could have been lost during the conversion from the TreeTagger tag-set used to Tanl tag-set during the parsing phase (see Section 3), the same rules for the identification of the appraisal groups were written in both tag-sets. As previously described in Section 7 and in Tables 7.3 and 7.4, the TreeTagger tag-set is fine-grained, while Tanl is coarse-grained.

	Overall			
	TreeTagger		**Tanl**	
	Strict	Lenient	Strict	Lenient
	MODIFIERS			
Precision	0.32	0.32	0.34	**0.35**
Recall	0.16	0.18	0.21	**0.23**
F1	0.21	0.23	0.26	**0.27**
	TARGETS			
Precision	0.43	0.42	**0.45**	0.44
Recall	0.21	0.22	0.28	**0.29**
F1	0.28	0.29	0.34	**0.35**
	APPRAISAL GROUPS			
Precision	0.13	**0.20**	0.14	0.19
Recall	0.07	0.14	0.10	**0.14**
F1	0.09	0.16	0.11	**0.16**

Table 8.9: Identification of modifiers, targets and appraisal groups in the overall Italian dataset.

Table 8.9 shows that the Tanl tag-set is better in terms of F1 measure of all the categories with respect to TreeTagger, although by considering the lenient measure for the appraisal groups there is practically no difference. As for the comparison of the results of Table 8.9 to those in English (in Table 8.1) and Russian (in Table 8.5), the F1 for modifiers in Italian is much lower than English and Russian (0.27 vs. 0.39 and 0.41), but similar for targets (0.35 vs. 0.42 and 0.36). However, what really stands out is that F1 for appraisal groups is lower (0.16 vs. 0.32 and 0.37). In order to check whether this was due to the noise possibly created by some rules, I did two tests: first, each of the rules were individually applied in order to understand their weight in the identification phase; second, from the overall set, each of them was eliminated in turn, following the same reasoning of *feature ablation* in studies such as those by (Prakash *et al.*, 2007), (Burkett & Klein, 2008) and (Recasens *et al.*, 2013). This unfortunately did not lead to any improvement, but only confirmed the hypothesis that the most important rules are "noun + adjective", followed by "noun + verb".

Table 8.10 shows the identification of modifiers, targets and appraisal groups across text types. F1 for appraisal groups is considerably higher for news in which it reaches 0.27 as opposed to 0.08 in political speeches and 0.16 in TED talks. It is worth noticing that in the case of news, all lenient values are much higher than the corresponding strict ones (F1 strict is in fact 0.11, so more in line with the other text types). This somehow goes against my expectation that news would have been more difficult to parse because of their more formal register.

	Political				News				TED			
	TreeTagger		Tanl		TreeTagger		Tanl		TreeTagger		Tanl	
	Strict	Lenient	Strict	Lenient	Strict	Lenient	Strict	Lenient	Strict	Lenient	Strict	Lenient
	MODIFIERS											
Precision	0.31	0.32	0.32	**0.33**	0.22	**0.27**	0.22	**0.27**	0.40	0.41	0.41	**0.41**
Recall	0.14	0.15	0.18	**0.19**	0.21	**0.28**	0.21	**0.28**	0.17	0.17	0.26	**0.26**
F1	0.20	0.20	0.23	**0.24**	0.21	**0.27**	0.21	**0.27**	0.23	0.24	0.32	**0.32**
	TARGETS											
Precision	0.46	0.47	0.48	**0.49**	0.27	**0.29**	0.27	**0.29**	0.46	0.46	0.44	0.45
Recall	0.22	0.22	0.28	**0.29**	0.28	**0.33**	0.28	**0.33**	0.18	0.18	0.27	**0.27**
F1	0.29	0.30	0.35	**0.36**	0.28	**0.31**	0.28	**0.31**	0.26	0.26	0.33	**0.34**
	APPRAISAL GROUPS											
Precision	0.11	**0.12**	0.10	0.11	0.11	**0.26**	0.11	0.25	0.18	0.18	0.19	**0.20**
Recall	0.06	0.06	0.06	**0.07**	0.11	**0.28**	0.11	**0.28**	0.07	0.07	0.12	**0.13**
F1	0.07	0.08	0.08	**0.08**	0.11	**0.27**	0.11	0.27	0.07	0.10	0.15	**0.16**

Table 8.10: Identification of modifiers, targets and appraisal groups across text types in Italian.

The low F1 in political speeches is also inexplicable since the F1 for modifiers and targets (0.24 and 0.36) in this type of texts is not much different from those of news (0.27 and 0.31) and TED talks (0.32 and 0.34). In terms of comparison with English and Russian, from Tables 8.2 and 8.6 we can see that the system performs better in terms of appraisal groups when applied to the news dataset in Italian (with 0.27) and English (with 0.33), but not in Russian (with 0.23) as opposed to political speeches (0.32 in English, 0.37 in Russian, 0.08 in Italian) and TED talks (0.27 in English, 0.40 in Russian, 0.16 in Italian).

	Overall			
	TreeTagger		**Tanl**	
	Strict	Lenient	Strict	Lenient
	MODIFIERS			
Orientation	**0.18**	0.16	0.18	0.17
Attitude	**0.80**	0.78	0.72	0.71
Force	0.82	0.82	0.82	**0.83**
Polarity	0.99	**0.99**	0.99	0.99
	TARGETS			
Orientation	**0.19**	0.17	0.18	0.17
Type	**0.83**	0.80	0.77	0.76
	APPRAISAL GROUPS			
Orientation	**0.54**	0.52	0.51	0.50

Table 8.11: Attributes of modifiers, targets and appraisal groups in the overall Italian dataset.

As far as the attributes are concerned, from Table 8.11 we can see that *orientation* results are higher for the appraisal groups with 0.52 as opposed to 0.46 in English and 0.43 in Russian (see Tables 8.3 and 8.7).

The good connection between *type* (i.e. ‘person’, ‘action, ‘thing’, ‘other’) and *attitude* (e.g. ‘affect’, ‘judgement’ and ‘appreciation’) is confirmed in this language as well, since the accuracy is 0.78 for *attitude*. Like in Russian, the translation of the appraisal groups from Italian into English had to be done in order to match them to their English Wordnet synsets, since no Wordnet is publicly available for Italian (*MultiWordNet* by (Pianta *et al.*, 2002), *EuroWordNet* by (Vossen *et al.*, 1997) and its more updated version *ItalWordNet* by (Roventini *et al.*, 2000).

The target *type*, in fact, performs well with 0.80 on the overall dataset, and 1.00 on news (see Table 8.12).

All the other modifiers' attributes perform very well: *force* (i.e. 'normal', 'low, 'high', 'reverse') and *polarity* (i.e. 'marked', 'unmarked) being 0.83 and 0.99 respectively.

As for the contribution of *force* and *polarity* to the *orientation* of the appraisal groups, the rule to spot negation (i.e. preceding lemma being *non*) was activated twice, but did not influence the *orientation* anyway.

	Political				News				TED			
	TreeTagger		Tanl		TreeTagger		Tanl		TreeTagger		Tanl	
	Strict	Lenient	Strict	Lenient	Strict	Lenient	Strict	Lenient	Strict	Lenient	Strict	Lenient
	MODIFIERS											
Orientation	0.22	0.22	**0.23**	0.22	**0.20**	0.15	**0.20**	0.15	0.12	0.12	**0.13**	0.13
Attitude	0.72	**0.72**	0.67	0.68	**0.90**	0.81	**0.90**	0.81	**0.82**	0.80	0.68	0.67
Force	0.85	**0.86**	0.84	0.85	**0.90**	0.88	**0.90**	0.88	0.70	0.71	0.77	**0.77**
Polarity	0.99	0.99	0.99	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.99	0.99
	TARGETS											
Orientation	0.23	0.22	**0.24**	0.23	**0.18**	0.12	**0.18**	0.12	**0.12**	**0.12**	0.11	0.11
Type	**0.76**	0.73	0.73	0.71	**1.00**	0.91	**1.00**	0.91	**0.80**	**0.80**	0.71	0.72
	APPRAISAL GROUPS											
Orientation	**0.50**	0.48	0.44	0.42	**0.53**	0.51	**0.53**	0.51	**0.59**	**0.59**	0.56	0.53

Table 8.12: Attributes of modifiers, targets and appraisal groups across text types in Italian

.

8.2.4 An evaluation of the automatic system performances

At this point, a comparison of the performances of the automatic system to another more *basic* system not based on strong linguistic reasoning is useful. Because of the substantial difference in terms of methodology and/or data of this study with respect to other studies in the automatic recognition of *attitude* and providing a baseline (e.g. Dotti, 2013), I had to build a suitable one. The requirement was that this more basic system had to work as a **meaningful basis** for testing the performances of my system on all categories (targets, modifiers, appraisal groups) and attributes (force, polarity, type, attitude, orientation). In order for the comparison to be meaningful, the building assumptions had to represent a balance between a very basic baseline and an advanced system.

In order to demonstrate the difference of the performances of similar tools on the same data, the basic system has been built by using previously existing resources and tools (also used, along with others, in the parsing-based system): the TreeTagger POS-tagger and the NRC sentiment dictionary (the original in English and its automatic translations in Italian and Russian). As such:

1. All adjectives have been assigned the function of modifiers and all nouns the function of targets. This results in the number of modifiers and targets in a sentence to be not necessarily the same.

2. Afterwards, for those words contained in the original *NRC sentiment dictionary* the *orientation* ('positive' or 'negative') has been taken. For those not in the dictionary, the *orientation* has been labelled 'neutral'.

3. Neighbour targets and modifiers have been linked in order to form the appraisal groups, i.e. each target has been linked to the closest modifier. In order to account for standard differences across languages, a suitable modifier could be found either before or after the target: for example, in English adjectives always precede nouns (e.g. *beautiful eyes*) unless in predicate position (e.g. *the eyes are beautiful*); in Italian they are quite flexible, with a general preference for "noun - attributive", usually depending on the adjective (e.g. *occhi belli* or *begli occhi*); in Russian, you can have прекрасные глаза (beautiful eyes), but also the order "noun - attributive" usually for statements in present tense, where additional verbs are not needed (e.g. глаза (-) прекрасны(е) - these eyes are beautiful).

For this reason, the absolute value has been taken.

4. For *orientation*, the 'ambiguous' value has not been considered. This assumption applies both to modifiers and targets in the few cases in which a word is both positive and negative in the dictionary (in that case 'negative' has been used), and also to appraisal groups consisting of "positive + negative" (in that case 'negative' has been chosen as final *orientation* of the group).

5. For the other attributes (attitude, force, polarity, target type), values have been randomly assigned.

Although the assumptions below are clearly not basic, especially when it comes to the use of the sentiment dictionary, I will refer to this new system as "baseline" or "POS-based system" (or a combination of both, according to the necessity) from now on. I will base the discussion on the strict values, micro-averaged.

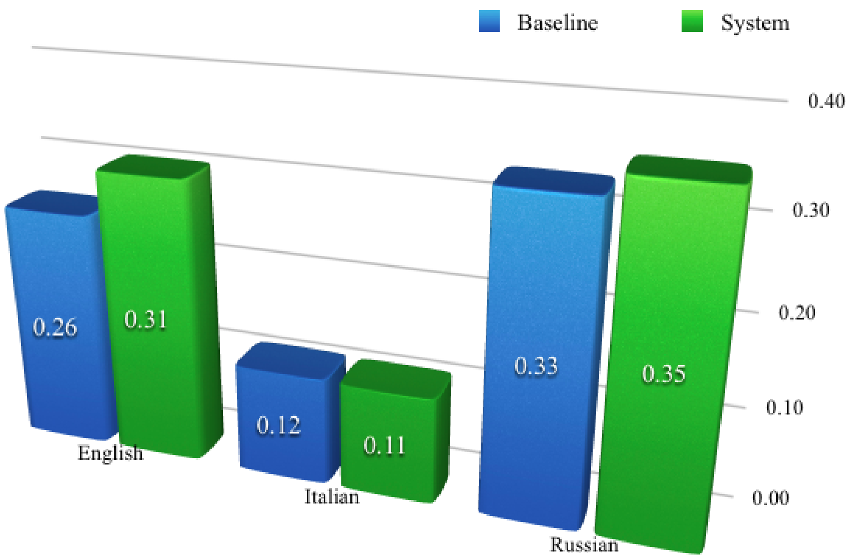

Figure 8.1: Results of the comparison between the POS-based baseline and the parsing-based system in the task of the identification of the appraisal groups.

I will start the comparison by comparing the POS-based system and the parsing-based system from the point of view of the identification. From the bar chart 8.1 showing the F1 measure of the appraisal groups, we can see that **the parsing-based system**

outperforms the baseline in English and Russian, with 0.31 vs. 0.26 in English and 0.35 vs. 0.33 in Russian, and **has practically the same F1 in Italian** with 0.11 vs. 0.12.

	English		**Italian**		**Russian**	
	BASE	SYSTEM	BASE	SYSTEM	BASE	SYSTEM
	MODIFIERS					
Precision	0.60	0.49	0.44	0.34	0.53	0.50
Recall	0.27	0.31	0.28	0.21	0.29	0.34
F1	0.37	0.38	0.34	0.26	0.38	0.41
	TARGETS					
Precision	0.47	0.57	0.38	0.45	0.39	0.42
Recall	0.65	0.33	0.69	0.28	0.71	0.31
F1	0.55	0.42	0.49	0.34	0.50	0.36
	APPRAISAL GROUPS					
Precision	0.40	0.40	0.15	0.14	0.49	0.45
Recall	0.19	0.25	0.10	0.10	0.25	0.29
F1	0.26	**0.31**	**0.12**	0.11	0.33	**0.35**

Table 8.13: Comparison of the F1 obtained by the POS-based baseline and the parsing-based system in the task of the identification of the appraisal groups.

The values related to precision and recall contributing to the calculation of this F1 are shown in Table 8.13, from which we can see that that the parsing-based system in general has higher recall than the baseline. However, the fact that the baseline has better F1 than the parsing-based system in the identification of modifiers and targets might be due to the following reasons:

- The baseline relies only on part-of-speech taggers, which represent by now a very stable tool in all languages regardless of the method adopted to build them (see Christodoulopoulos *et al.* (2010) for English, Attardi & Simi (2009); Magnini *et al.* (2008) for Italian and (Sharoff *et al.*, 2008) for Russian, to cite just a few).
- The strongest rules of my system have been used: (i) proximity of targets and modifiers (ii) grammatical combination “noun + adjective”.

A closer look at the data clarifies that this is mainly due to the low frequency of “long-distance dependency links”, where by *long-distance* a distance longer than 1 (as in the case of linearly neighbour target and modifier) is meant. In the few cases in which such links are found, examples of wrong dependency trees have been found in all languages. For example, in the English "the recriminations and the worn out dogmas, that for far too long have strangled our politics", the parser managed to connect *recriminations strangled* and *strangled politics*, but not *dogmas strangled*, whose head was instead *out*; or in the case of phrasal verbs such as in "to set aside childish things", *things* is connected to *childish* and *set*, but *aside* has *to* as head. In the Italian equivalent (*di mettere da parte gli infantilismi*), *infantilismi* (childish things) has also *mettere* (to put) as head but *da parte* (aside) consists of *da* rightly referring to *mettere* and *parte* referring to *da*. In the Russian equivalent отказаться от ребячества (depart from childishness), both the verb отказаться and the noun ребячества have the preposition от as head.

In Italian another example of wrong tree is *la fiducia che mi avete accordato* (the trust that you have conceded to me), the auxiliary *avete* has *fiducia* as head instead of the verb. However, there are exceptions such as in the English phrase *they will not be met easily* where *met* is ROOT and head of the remainder of the words in the sentence.

An example of well-parsed tree in all languages is the expression *threaten our planet*/*minacciare il nostro pianeta*/угрожают нашей планете with *planet* as object of *threaten* and, in case of Italian, also the article *il* along with the possessive adjective *nostro* like in English and Russian, having *pianeta* as head.

As highlighted in Chapter 4, we still have to take into account cases in which translations present different structures, e.g. *humbled by the task before us* corresponds to *umile per il compito che ci aspetta* (literally "humbled for the task that waits for us") in Italian and to ощущая огромную важность поставленных перед нами задач (feeling the huge importance of the tasks put before us) in Russian. Figure 8.2 shows the dependency trees for English, Italian and Russian. In the Russian one some parsing relations are not correct, a problem not found in the Italian and English ones.

Despite the errors in the tree and the challenging word order in the final part of the sentence (assigned to us tasks), the appraisal group громную важность (huge importance) has been found in Russian, *compito aspetta* (task awaits) in Italian, but none has been found in English.

Another interesting difference is that the word inflections in Italian helped the parser to assign to the adjective *umile* (humbled) (singular) the verb *trovo* (find) (1st person singular) as head, in the sentence *oggi mi trovo qui, umile* (today I stand here, humbled). By contrast, in English *humbled* (singular and plural) has a wrong head, i.e. the noun *citizens* (plural) found earlier in the sentence (My fellow citizens: I stand here today

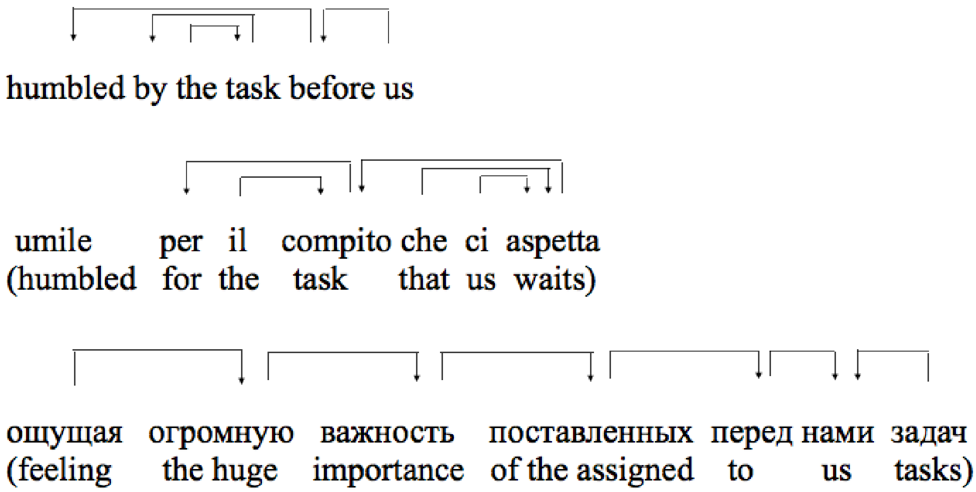

Figure 8.2: Dependency trees for English, Italian and Russian on the same sentence.

humbled). In Russian the correspondent ощущая (feeling) has also a wrong head, i.e. огромную (huge) (see Figure 8.2).

Copulas in Italian seem to work as well, e.g. in *questo è il prezzo e la promessa* (this is the price and the promise), the noun *prezzo* has *è* as head, while the noun *promessa* has *prezzo* as head instead of the verb. The hypothesis that the parser would have been misled by the singular verb *è* in cases in which it would have been grammatically more correct to use the plural to refer to both the nouns, was refuted by other two examples featuring the conjunction *e*:

- *Questo è il significato della nostra libertà e del nostro credo* (this is the meaning of our freedom and of our belief) in which *libertà* has *significato* as head, and *credo* has the verb *è* as head.

- *Di ogni razza e di ogni fede* (of every race and every faith), in which *razza* and *fede* have the first and the second preposition *di* as their heads respectively.

These few examples also made me test another hypothesis by Schwartz *et al.* (2012) cited in Nivre (2014) according to which parsers tend to prefer function words as heads. I found this to be true in all languages especially with prepositions: e.g. *in a new era*/*per una nuova era*/к новой эпохе (for a new era). In addition, in these cases the appraisal groups are usually identified (providing that at least one of the words is present in the sentiment dictionary).

An issue that I have found to be specific to Italian is the wrong classification of reflexive pronouns into personal (e.g. *mi* in “mi perdo” (I lose myself)).

However, when the combination “subject + verb + object + adjective” (e.g. “John makes me angry”), the English parser gave a perfect tree, and allowed the correct identification and classification of the appraisal group *me angry* as ‘negative’. In Italian the equivalent *Giovanni mi fa arrabbiare* would have a correct tree with the auxiliary *fa* as ROOT, but no identified appraisal groups. The same for its correspondent phrasal verb *rende arrabbiato*. In Russian the equivalent Иван рассердит меня has the wrong ROOT *Ivan*, but the appraisal group Иван рассердит (Ivan makes angry).

Finally, negation seems to be recognized and indicated by *neg*, apart with the verb *cannot* (also not manually annotated). However a good sign is that in the case of *not only, not just, not out* (e.g. *will act not only depended, not out of charity*), which might be easily misunderstood as negation, the parsing behaves well by recognizing them as one expression.

As for the last point, it is essential to add that, although the linear condition of proximity used by the baseline might not resemble the phenomena occurring in the natural language in few cases, it is still a valid one. On the contrary, precisely in order to cover all the cases, I decided to rely on using only modifiers and targets linked by dependency relations to build the appraisal groups. Unfortunately, despite the assumption of the dependency parsing was better than the linear proximity, my extrinsic evaluation has demonstrated that the dependency parsing models have to be extremely good in order for the entire tool to be reliable (see the case of Italian in Section 8.2.3) and, if this condition is not satisfied, the F1 in the identification of appraisal groups is only slightly better.

Another important aspect to bear in mind is that the parsing-based system is limited with respect to the manual annotation because rules do not cover two nouns linked by a preposition and multi-word units (see “Pairs extractor” module in Chapter 7). Out of 60 non-unique occurrences in English, the system has often annotated only one word of them, and 11 times including the wrong targets (e.g. *freedom symbolizing an end*, annotated as “symbolizing freedom”) apart from, for example, *far-reaching network, long-held human rights*.

As for the *orientation*, the evaluation method used so far for the results presented in this Chapter has been to first measure the correctness in the identification of the groups that carry appraisal, and afterwards on the correctly identified groups measure the correctness in the classification of the attributes (including *orientation*).

For this reason, in order to have a fair comparison, I will use a *score* that gives importance to the identification before measuring the *orientation*. Such score has been previously used in the context of the shared task *i2b2 challenge* on clinical texts (Sun *et al.*, 2013) and consists in:

$$score = F1_{Appraisalgroups} * Accuracy_{orientation} \tag{8.1}$$

The F1 measure achieved in the identification of the appraisal groups is multiplied by the accuracy achieved in the classification of their *orientation*. This prevents misleading conclusions in case a system is able to identify only few groups, but it still performs well in the classification of their *orientation*.

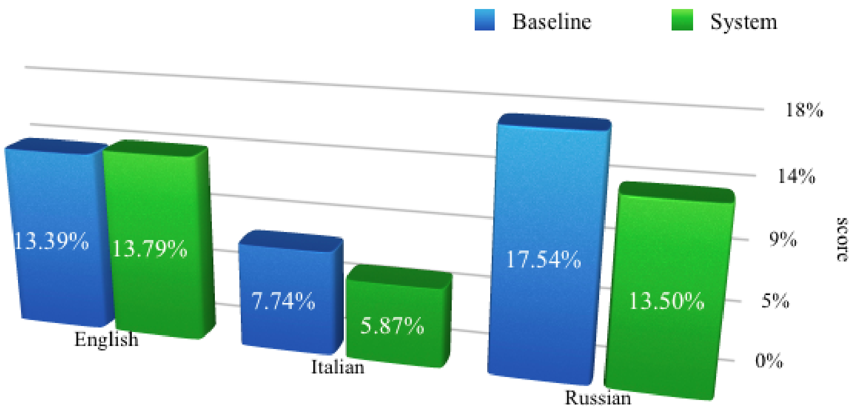

Figure 8.3: Results of the comparison between the POS-based baseline and the dependency-based system in the task of the classification of the orientation of the appraisal groups.

From the bar chart 8.3 it is visible that the performances of the baseline for the *orientation* are similar to those of the system for all languages. In particular, what emerges is that **the baseline slightly outperforms in the case of Italian** by 1.87%, **but also in Russian** by 4.04%. This last figure has a value, in so far as the Russian system already performed well with a score of 13.50% (vs. 13.79% in English) and higher than Italian (5.87%).

If the baseline outperforming in the case of *orientation* represents a rather unexpected result, from Table 8.14 we can see instead that **for all the other attributes the**

	English		Italian		Russian	
	BASE	SYSTEM	BASE	SYSTEM	BASE	SYSTEM
	MODIFIERS					
Orientation	0.34	0.19	0.34	0.18	0.39	0.15
Attitude	0.53	**0.81**	0.54	**0.72**	0.52	**0.65**
Force	0.22	**0.88**	0.25	**0.82**	0.25	**0.83**
Polarity	0.45	**0.98**	0.54	**0.99**	0.47	**0.96**
	TARGETS					
Orientation	0.62	0.23	0.54	0.18	0.52	0.20
Type	0.26	**0.91**	0.25	**0.77**	0.24	**0.78**
	APPRAISAL GROUPS					
F1	0.26	0.31	0.12	0.11	0.33	0.35
Orientation	0.52	0.45	0.66	0.51	0.53	0.38
SCORE	13.39%	**13.79%**	**7.74%**	5.87%	**17.54%**	13.50%

Table 8.14: Comparison of the attributes for the baseline and the system with final score for orientation of the appraisal groups.

parsing-based system is by far better in all languages with a difference that ranges from 0.13 to 0.66.

This difference being so consistent, we can safely draw some conclusions, despite of the fact that the values for all attributes (apart *orientation*) are randomly assigned by the baseline and so there is the possibility that different percentages are outputted according to other attempts. These are that (i) the rules used by the system for the attributes are very good, but that (ii) they have no influence on the correct classification of the *orientation* (see Section 8.2.3).

8.2.5 Manual annotation vs. automatic annotation

After looking at the figures related to the performances of the system, it is essential to investigate which are the real cases that the parsing-based system is able to correctly identify and classify from the point of view of *attitude* and *orientation* in English, Italian and Russian. A good dataset is represented by the appraisal groups highlighted in Obama's Inaugural speech in Chapter 4. I believe they are ideal because they cover different topics and are varied in terms of lexicon as well as register, and in some cases they encapsulate metaphors and idioms. In this respect, because appraisal groups coming from the other two text types (news and TED talks) highlighted in Chapter 4 share most of these features (or even fewer), the considerations for political speeches are likely to be generalizable. Nonetheless, I will focus on one piece of news, a political speech and a TED talk in English.

8.2.5.1 A comparison across languages

I will start with the comparison of the performances of the system on Obama's inaugural speech in the three languages, by dividing the outcome of the automatic annotation in relation to that of the manual analysis.

1. **Category 1: Both the *attitude* and *orientation* assigned by the system match those of the manual analysis**. This happens especially when the correct *attitude* is 'appreciation' and most times in English, like in the case of *great gift/abbiamo dono, full measure, prudent use, watchful eye* (positive), and *bitter swill/amaro sapore* (negative). A correctly identified and classified group in all languages because of its strong 'positive' prior orientation is *sacred oath/giuramento sacro/* священную присягу.

2. **Category 2: Only the *attitude* assigned by the system does not match that of the manual analysis**. As in the previous case, the majority of times the system assigns 'appreciation', which in these cases seems like a reasonable choice: for example for *greater effort/greater cooperation*, 'affect' had been assigned during the manual analysis because they were related to the speaker president Obama talking on behalf of all Americans, but 'appreciation' would be correct if analysed out of context. 'Appreciation' is also reasonable in the following cases, manually marked as 'judgement': *enduring spirit, vital trust/avete fiducia/*испытывая признательность, *better life/vita migliore/*лучшей жизни, *better history/storia migliore/*лучшего будущего, *hard-earned peace/pace guadagnata*, *rugged path, collective failure*.

3. **Category 3: Only the *orientation* assigned by the system does not match that of the manual analysis**. This happens because the classification is not straightforward, such as in:

    - *Hard choices/*решительный выбор, which has been classified as 'positive', or *hungry minds/menti affamate/*накормить голодных, which has been classified as 'negative' in all languages because *hungry* has 'negative' prior orientation.

- *Enduring convictions* that has been classified as 'negative' because of the lack of word sense disambiguation (WSD) in the dictionary. In fact, the prior orientation 'negative' in the sentiment dictionary related to *convictions* as "decision that someone is guilty" is not compatible to the word sense in which *convictions* means "a firmly held belief or opinion" and is 'positive':

  > Recall that earlier generations faced down fascism and communism not just with missiles and tanks, but with sturdy alliances and enduring convictions.

To this category a number of examples whose *orientation* has been classified as 'not in dictionary' belong: *patchwork heritage/retaggio disomogeneo, hard work, starved bodies*. And a final group consists of the wrong classification in English but not in other languages, for example *unpleasant decisions* wrongly classified as 'positive', but its Italian counterpart *decisioni spiacevoli* correctly classified as 'negative', or also *darkest hours* classified as 'neutral', while its Italian and Russian counterparts *momenti bui*/серьезными испытаниями classified as 'negative'.

4. **Category 4: They have not been identified at all in English, but in some cases in other languages**. Table 8.15 show examples belonging to this category. At first glance, the most likely cause of these missed identifications could be the fact that both the modifier and the target forming the group are not present in the dictionary. However, on second thought, this appears to be a rather simplistic and not reliable explanation because Italian and Russian dictionaries are translations of the English one. A more convincing cause might be a greater simplicity in the Italian and Russian expressions. In fact, in the case of *sapping of confidence*, the system is not supposed to retrieve two nouns linked by a preposition, but in Russian this problem does not exist because the structure becomes the verb and noun слабнет уверенность (weakens confidence).

   Table 8.15 also shows appraisal groups that, despite consisting of sentiment words and matching the expected POS-tags, have not been identified in any language. The last ones are examples of metaphors and fixed expressions: *give (a) hand, coldest (of) months, unclench (your) fist, raging storm.*

Special cases are represented by мы проделали (we travelled) wrongly identified in Russian despite not carrying any sentiment (and in fact classified as 'not in dictionary'), while in Italian *quanta strada* (how much path) carries sentiment and has been identified, but classified as 'not in dictionary', as well as *future generations/future world* wrongly identified and with *orientation* 'not in dictionary'.

5. **Category 5: They have not been highlighted during the manual analysis, but are interesting from the point of view of the *orientation*.** Table 8.16 shows two scenarios: those words that have been correctly classified with *orientation* not dependent on the context, and those words that have been correctly classified with *orientation* dependent on the context.

   For all groups, there is at least one appraisal word in each of the groups whose prior orientation has led to a correct contextual orientation: *wrong* 'negative' -> *wrong side* 'negative', *desert* 'negative' -> *far-off deserts* 'negative' vs. *timeless words, remember words, common humanity* correctly classified as 'positive'.

   I have created a separate category for the second group because they should be marked as 'ambiguous' in the sentiment dictionary. In particular *new foundation/new jobs*, marked as 'positive', could be 'negative' if referred to foundation of crime or drug-related jobs.

   Or in "What is demanded then is a return to these truths", *demanded return* has been marked as 'negative' despite being 'positive'.

   Those classified as 'not in dictionary' despite containing a strong sentiment word do not fall in this or the previous categories, e.g. *restore trust* (although *trust* was in the correctly classified group *vital trust*), *brave Americans/brave us*, *new era*. The same applies to those which have been wrongly identified because not carrying any appraisal, such as *gross product* classified as 'negative' because of *gross* as opposed to *Domestic product* classified as 'not in dictionary', *small village/small band*, *electric grids* and *subject indicators*.

As for the identification phase, we can conclude that all the categories report a very good number of identified appraisal groups, especially matching the grammatical combination "noun + adjective".

As for the classification, Category 1 reports cases in which not only appraisal groups have been identified, but both attributes have been classified correctly. The results of Category 2 focusing on *attitude* are also satisfying, and the same can be stated about Category 1, Category 3 and Category 5 focusing on *orientation*, and it is pretty clear that the inclusion in the sentiment dictionary is the reason behind the system's choice.

In terms of the other two languages, from a quick look at all categories but especially Category 4, we have a proof that overall the identification works, whereas the classification of the attributes is intermittently reliable (see Category 2 for *attitude* and Category 3 for *orientation*).

Appraisal groups identified		**Classification**
Italian	**Russian**	
	слабнет уверенность (weakens confidence)	Appreciation, Not in dictionary
pianeta surriscalda (planet overheats)	глобальным потеплением (global warming)	Appreciation, Not in dictionary
	ценности подлинны (values true)	Appreciation
	ответственный процесс (responsible process)	Appreciation, Positive
nubi tempestose (stormy clouds)	мрачных туч (gathering clouds)	Appreciation, Negative
	оставаться маяком (represent a light)	Appreciation, Positive
	зависит успех (success depends)	Appreciation, Positive
cure accessibili (affordable cares)		Appreciation, Positive
	приносили урожаи (bring abundance)	Appreciation, Positive
	павшие герои (fallen heroes)	Judgement, Negative
Appraisal groups not identified		
we seize gladly dying campfires as we please less productive undiminished capacity blame ills stale arguments civil war give (a) hand coldest (of) months unclench (your) fist raging storm		

Table 8.15: Examples of appraisal groups belonging to Category 4 (those highlighted in the manual analysis) are shown.

Extra appraisal groups correctly classified in context	
With orientation not dependent on context	
Positive	Negative
shown generosity strengthen ways enduring spirit noble idea passed gift prosperous nation rightful place sturdy alliances greater effort/greater cooperation/great gift/ delivered gift magnificent mall God-given promise/equal promise timeless words/remember words common humanity old friends former foes common defense decent wage	collective failure/make failure nagging fear/inevitable fear worn recriminations childish things struggled men bad habits old hatreds wrong side far-off deserts common danger alarmed city new threats/nuclear threats consumed arguments false choice uncertain destiny
With orientation dependent on context	
protecting narrow interests and putting off unpleasant decisions new foundation for growth we will act not only to create new jobs begin again the work of remaking America willingness to find meaning in something greater We remain a young nation We remain the most powerful nation bigger than the sum of our individual ambitions our goods no less needed than they were last week	the stale political arguments they packed up their few worldly possessions

Table 8.16: Examples of *extra* appraisal groups not highlighted in manual analysis belonging to Category 5 are shown.

8.2.5.2 A comparison across text types

I will now move to the discussion on the performances of the system across the three text types. This will hopefully be comprehensive enough to cover multiple objectives:

1. Analysis of the appraisal groups and their *orientation*, similarly to the analysis previously done in the case of Obama's inaugural discourse.

2. Analysis of the appraisal groups not identified, although expected.

3. Conclusions on any potential difference in the retrieval, depending on the text type.

Given that I have already looked at the other languages in the previous Section 8.2.5.1, in this Section I will focus on English. I will discuss only *orientation* since I have already discovered that 'appreciation' is likely to be assigned as *attitude* (see Category 2 of this Section and Section 6.2.2). As for the data, as mentioned in the introduction to this Section, I will analyse some selected paragraphs coming from my corpus and also previously used as support for the explanation of the Appraisal Framework in Chapter 3. In all of them I will highlight the appraisal groups that I intuitively expect the system to retrieve, so I will exclude those matching non-covered grammatical combinations. Out of the highlighted ones, I will underline those that have actually been correctly identified by the system.

The first text that I will look at belongs to political speeches, and is an extract from Kennedy's inaugural discourse:

> Let the word go forth from this time and place, to friend and foe alike, that the torch has been passed to a new generation of Americans born in this century, tempered by war, disciplined by a hard and bitter peace, proud of our ancient heritage and unwilling to witness or permit the slow undoing of those human rights to which this Nation has always been committed, and to which we are committed today at home and around the world.

In this case, out of 10 expected appraisal groups, the system retrieves 4: *hard peace, ancient heritage, slow undoing* and *human rights*, although only for *hard peace* and *ancient heritage* gives an *orientation* different from 'not in dictionary': 'positive' and correct in the first case, 'negative' and incorrect in the second case.

The second is a piece of news entitled "Computer selected and disseminated without FBIS editorial intervention":

> Recently, North Korea strongly denounced comments made by U.S. President George W. Bush during his Seoul visit last month accusing the

> North Korean leadership of starving its people while developing weapons of mass destruction. [...] The report comprehensively blamed the North Korean authorities for committing wrong-doings in terms of human rights. [...] The U.S. State Department on Tuesday (KST) rated the human rights situation in North Korea "poor" in its annual human rights report, casting dark clouds on the already tense relationship between Pyongyang and Washington.

In this case, out of the 13 expected appraisal groups, the system identifies 4 of them: *mass destruction, human rights, dark clouds, tense relationships* but surprisingly classifies all of them as 'not in dictionary' despite their strong sentiment words. The only exception is *mass destruction* classified as 'negative'. The system also retrieves some non-appraisal groups such as *last month, North authorities, human report* because containing words included in the sentiment dictionary.

The third text is taken from the TED talk entitled "Photos that changed the world" by Jonathan Klein:

> In the 1960s and 1970s, the Vietnam War was basically shown in America's living rooms day in, day out. News photos brought people face to face with the victims of the war, a little girl burned by napalm, a student killed by the National Guard at Kent State University in Ohio during a protest. In fact, these images became the voices of protest themselves. [...] Unfortunately, some very important images are deemed too graphic or disturbing for us to see them. I'll show you one photo here, and it's a photo by Eugene Richards of an Iraq War veteran from an extraordinary piece of work, which has never been published, called "War is Personal".

Here, out of 9 expected appraisal groups, 5 are identified, although only *student killed* is classified as 'negative', and *extraordinary piece* and *important images* as 'positive'. Moreover, the non-appraisal groups *little girl* and *published work* are justifiable in so far they could carry sentiment in other contexts.

The results just shown are satisfying for a number of reasons. In particular, by relating conclusions to the objectives of this section, I can state that:

1. The appraisal groups identified and their orientation classification are good, especially in the case of identification. The *orientation* strongly depends by whether

either the target or the modifier are included in the sentiment dictionary, as already noticed in the case of Obama's inaugural discourse.

2. Appraisal groups not identified, but expected, are a few, for example those including the words *disciplined, committed, denounced, starving, weapons, wrong-doings, poor, victims*. However, on a closer look at the dictionary, *disciplined, weapon, wrong-doing, poor* are not included, while for the others a wrong parsing could have been the cause.

3. Any difference in the retrieval according to the text type is likely, with more groups identified in the TED talk, both expected (5) and justifiable (2) out of 9; news is the worst in terms of performances with 4 out of 13, and political speech with 4 out of 10. However, these can only be regarded as suppositions because of the shortness of the texts.

8.3 Results on additional data

In order to conclude my overview on which are the real cases that the system is able to correctly identify and classify from the point of view of *orientation* and *attitude*, I wanted to extend the evaluation to additional data not manually annotated, in case the size of the corpus had the disadvantage of excluding interesting patterns. I could have chosen any data containing appraisal, but I decided to rely on the remaining sentences belonging to TED talks that supported the explanation of the Appraisal Framework in Chapter 3. Since such sentences are not included in my corpus, the objective is only to give a flavour of how the system behaves in the wild (real usage scenario) in English.

As in the previous section, I will highlight the expected appraisal groups (excluding those not matching the specified grammatical combinations) and underline those actually retrieved. Because *all attitudes* and *orientations* are considered, the highlighted expressions overcome in number those underlined during the manual analysis in Chapter 3, where different Sections were dedicated to the categories of the sub-system attitude of the AF.

The first is an extract from a TED talk entitled "A life lesson from a volunteer firefighter" in which the author, Mark Bezos, describes his own experiences as head of development for a non-profit called "Robin Hood" and firefighter.

> In both my vocation at Robin Hood and my vocation as a volunteer firefighter, I am witness to acts of generosity and kindness on a monumental scale, but I 'm also witness to acts of grace and courage on an individual basis. And you know what I've learned? They all matter.
>
> [...]
>
> So as I look around this room at people who either have achieved, or are on their way to achieving, remarkable levels of success, I would offer this reminder: don't wait. Don't wait until you make your first million to make a difference in somebody's life. If you have something to give, give it now. Serve food at a soup kitchen. Clean up a neighbourhood park. Be a mentor. Not every day is going to offer us a chance to save somebody's life, but every day offers us an opportunity to affect one.

It is worth mentioning that the groups *people achieved* and *be mentor* have been also correctly classified as 'positive', whereas *matter* and *opportunity* were marked as 'positive' but associated to a wrong modifier/target. One non-appraisal expression, *individual basis*, was also identified.

The second is an excerpt from the TED talk "Why we love, why we cheat" by Helen Fisher.

> And this graduate student was madly in love with another graduate student, and she was not in love with him. And they were all at a conference in Beijing. And he knew from our work that if you go and do something very novel with somebody, you can drive up the dopamine in the brain, and perhaps trigger this brain system for romantic love. So he decided he'd put science to work, and he invited this girl to go off on a rickshaw ride with him. [..] Apparently they go all around the buses and the trucks and it's crazy and it's noisy and it's exciting. [...] So off they go and she's squealing and squeezing him and laughing and having a wonderful time. An hour later they get down off of the rickshaw, and she throws her hands up and she says, "Wasn't that wonderful?" And, "Wasn't that rickshaw driver handsome?"

The system not only identified *something novel, laughing him, wonderful time, handsome driver*, but also marked them as 'positive'. The non-appraisal group *graduate student* was also wrongly identified.

The third is a TED talk titled "Violence against the women - it's a men's issue" by Jackson Katz.

> The first is that it gives men an excuse not to pay attention. Right? A lot of men hear the term "women's issues" and we tend to tune it out, and we think, "Hey, I'm a guy. That's for the girls," or "That's for the women." And a lot of men literally don't get beyond the first sentence as a result. It's almost like a chip in our brain is activated, and the neural pathways take our attention in a different direction when we hear the term "women's issues". [...] But there's so many men who care deeply about these issues, but caring deeply is not enough. We need more men [...] with the courage, with the strength, with the moral integrity to break our complicit silence and challenge each other and stand with women and not against them. By the way, we owe it to women. There's no question about it. But we also owe it to our sons. We also owe it to young men who are growing up all over the world in situations where they didn't make the choice to be a man in a culture that tells them that manhood is a certain way [...] We that have a choice have an opportunity and a responsibility to them as well.

In this extract, the system identified a very high number of appraisal groups, but only *have choice* was marked as 'positive'. At the same time, non-appraisal groups such as *first sentence* (negative), *neural pathways, young men, men growing, culture tells* have been identified, because each contain words included in the sentiment dictionary such as *first, pathways, young, growing, culture*.

Unlike the analysis of the corpus data in Section 8.2.5.2, I will not count the identified appraisal groups, but we can still see that there are not so many cases of bad missing identification apart from perhaps *romantic love, that wonderful, is exciting*. An explanation related to the grammatical combination being "noun+adjective" is not applicable since it is also the one matching most of the identified groups (like in the case of the corpus data in Section 8.2.5.2). Finally, in terms of precision, it is good since very few non-appraisal groups have been spotted.

From the point of view of *orientation*, it is not easy to state with precision whether the system comes up to the expectations. However, a good result is that, despite a few cases of 'not in dictionary', it is very unlikely that the 'positive' and 'negative' values are switched.

From the point of view of *attitude*, ‘appreciation’ is the most assigned; ‘judgement’ is often assigned to people, either expressed with the use of pronouns (*I am, I know*) or nouns (e.g. *Be a mentor, Americans born, little girl, student killed*), with the exception of *handsome driver*; and the only occurrence of ‘affect’ (and correct) is *we have a choice*.

8.4 A comprehensive interpretation of the results

I will conclude the Chapter by going through the main evaluation methods used, and summarise their results.

- **IDENTIFICATION OF THE APPRAISAL GROUPS**

  Comparison to baseline. The *baseline* system created *ad hoc* and presented in Section 8.2.4 has resulted in a valuable tool for testing the dependency-based system because it gave me more aspects to analyse in depth. As I have highlighted in Section 8.2.4, there are cases in which the parser is not able to link targets and modifiers (mainly copulas in English and Italian). For example, in “la ragazza e’ stupida” (the girl is stupid) the pair is not found. This sometimes contributes to a lack of identification of easy cases. For example, in “report on North Korea was not as severe as previous years” the parsing-based system identified *severe years* instead of *severe report* most likely because it was misled by the fact that *severe* has *years* as head, although *report* has *severe* as head. Conversely, the POS-based system would identify the group *severe Korea*, since it is restricted to find a suitable modifier for the target in the span of a sentence.

  Nonetheless, when able to properly identify the syntactic structure, the dependency-based system has the advantage of spotting long-distance dependency links. One could argue that other text types could work better with the features used, but the similar performances in the three used in this work would discourage such hypothesis.

  Results on corpus data. These are provided both by the figures presented and discussed in the Sections 8.2.1 for English, 8.2.2 for Russian and 8.2.3 for Italian, and by the scrutiny of the linguistic patterns done in Section 8.2.5.

Table 8.17 shows in parallel the performances in the identification of appraisal groups of English (originally in Table 8.1), Russian (originally in Table 8.5) and Italian (originally in Table 8.9).

	English		**Italian**		**Russian**	
	Strict	Lenient	Strict	Lenient	Strict	Lenient
	APPRAISAL GROUPS					
Precision	0.40	0.41	0.14	0.19	0.45	0.44
Recall	0.25	0.26	0.10	0.14	0.29	0.31
F1	0.31	0.32	0.11	0.16	0.35	0.37

Table 8.17: Identification of appraisal groups in the English, Italian and Russian overall datasets.

F1 for appraisal groups in Italian is low with 0.16 vs. 0.32 in English and 0.37 in Russian. **Russian is the best system in the identification phase, followed by English, and Italian**.

This proportion in the performances related to the identification step across the languages (i.e. English and Russian better than Italian) is also confirmed by the analysis done in Section 8.2.5.1, summarised specifically in Category 1 for English, and Category 4 for Italian and Russian. In fact from Table 8.15 it is immediately visible that the Italian system is not able to spot as many appraisal groups (in this case not identified in English either) as the Russian system.

In terms of differences across the text types, in Section 8.2.5.2 we had seen that in English TED talks are the best in terms of retrieval of appraisal groups (7 out of 9, equal to 78%), followed by political speeches (4 out of 10, equal to 40%) and news (4 out of 13, equal to 31%). These data are somehow aligned with the order emerging from Table 8.2 in which the highest F1 is for political speeches (0.33), followed by TED talks (0.30) and then for political speeches (0.27). As for the other languages, we can only rely on the figures of Tables 8.6 for Russian, and 8.10 for Italian from which the order is TED talks, political speeches and news for Russian vs. news, TED talks and political speeches in Italian. The exact figures are shown in Table 8.18.

A reason for the top performance of news in Italian and English might be that, unlike the Russian news, they have been taken from the same source and can be

	English		Italian		Russian	
	Strict	Lenient	Strict	Lenient	Strict	Lenient
	APPRAISAL GROUPS					
Political	0.32	0.33	0.08	0.08	0.37	0.37
News	0.23	0.27	0.11	0.27	0.23	0.23
TED	0.29	0.30	0.15	0.16	0.29	0.40

Table 8.18: Identification of appraisal groups in English, Italian and Russian across text types given on the basis of their F1 values.

assumed to belong to the same category: "Europe has lost its sympathetic soul", "The real bazooka is still not there", "A rotten system" from *Sole24ore* in the case of Italian, and human rights from *MPQA corpus* for English. It is also worth underlining that TED talks perform very well in all languages.

Results on additional data. In Section 8.3 we have seen that there are not so many cases of bad missing identification apart from perhaps *romantic love, that wonderful, is exciting*, although it cannot be stated that groups are always reliably identified either. Additionally, I concluded that the combination "noun+adjective" is the one matching most of the identified groups similarly to the corpus data in Section 8.2.5. Finally, in terms of precision, it is good to see that very few non-appraisal groups have been spotted.

- **CLASSIFICATION OF THE ORIENTATION OF THE APPRAISAL GROUPS**

  Comparison to baseline. As discussed in Section 8.2.4, the results of the comparison with the baseline in this task bring the performances of the parsing-based system into question.

  However, plausible explanations are that:

  - The features on which the annotation scheme has been based (in particular *force* and *polarity*) were simply not the right ones to influence the classification of the *orientation* (see Tables 8.3, 8.7 and 8.11).
  - The contextual orientation is too strongly influenced by prior orientation of the words.

  Results on corpus data. I will start by showing in Table 8.19 the accuracy in the classification of the *orientation* in parallel for all languages, both on the overall dataset and on the specific datasets (political speeches, news and TED talks).

	English		**Italian**		**Russian**	
	Strict	Lenient	Strict	Lenient	Strict	Lenient
	APPRAISAL GROUPS					
Overall	0.45	0.46	0.51	0.50	0.38	0.38
Political	0.47	0.48	0.44	0.42	0.38	0.38
News	0.46	0.46	0.53	0.51	0.42	0.37
TED	0.40	0.40	0.56	0.53	0.31	0.39

Table 8.19: Accuracy in the task of the classification of the orientation of the appraisal groups across languages and datasets.

From Table 8.19 we can see that results are comparable across the languages, as the lenient values for the overall dataset is 0.38 in Russian, 0.46 in English and 0.50 in Italian, ranges kept across the individual text types.

As for the detailed analysis of what is correctly classified, from Section 8.2.5.1 we can see in Category 1 that the cases in which not only appraisal groups have been identified, but both attributes have been correctly classified usually contain a word with strong prior orientation (e.g. *bitter swill, great gift, most sacred oath*). Category 3 and Category 5 confirm this hypothesis that most times there is a good reason behind the system's choices (see in particular the last group in Table 8.16). As for the other two languages, from a quick look at Category 3 we can conclude that the classification is intermittently reliable.

The sentences in Section 8.2.5.2, although in English only, are useful to guess any difference in the classification of the orientation across the text types. Actually, since only one group is correctly classified in the case of Kennedy's inaugural discourse and one in the piece of news "Computer selected and disseminated without FBIS editorial intervention", but threehe$f$cck$f$ in the TED talk "Photos that changed the world", the proportions in Table 8.19 are altered although the texts presented in the Section 8.2.5.2 are only excerpts and cannot be considered as definitive proof.

Results on additional data. In this case, despite a few cases of 'not in dictionary', the values 'positive' and 'negative' are never switched and they are usually not wrongly classified ('not in dictionary' is given instead).

- **CLASSIFICATION OF THE ATTITUDE OF THE APPRAISAL GROUPS**

  Although this is not formally an attribute of the appraisal groups (but of modifiers), I included it in the conclusions since it is based on a strong interrelation between modifiers and targets.

  Comparison to baseline. As seen in Section 8.2.4 and in particular from Table 8.14, the system outperforms the baseline in all languages, by achieving 0.88 vs. 0.53 in English, 0.72 vs. 0.54 in Italian and 0.65 vs. 0.52 in Russian.

  Results on corpus data. Table 8.20 show the accuracy in the task of the classification of attitude across languages and datasets.

	English		**Italian**		**Russian**	
	Strict	Lenient	Strict	Lenient	Strict	Lenient
	APPRAISAL GROUPS					
Overall	0.81	0.81	0.72	0.71	0.65	0.68
Political	0.78	0.77	0.67	0.68	0.61	0.61
News	0.87	0.86	0.90	0.81	0.78	0.75
TED	0.83	0.83	0.68	0.67	0.68	0.79

Table 8.20: Accuracy in the task of the classification of the attitude across languages and datasets.

  As discussed in the individual sections, these excellent values in the range of 0.71-0.81 (lenient on the overall dataset) confirm the clear working connection between *type* and *attitude* is present in all the languages. However, despite the acknowledgement that the simplification of this mapping (see Table 7.6) greatly helped, it is important to point out that the manual annotations themselves are open to discussion since no inter-annotator agreement could be measured (see Section 6.1).

  As for the specific values, by looking at Category 2 in Section 8.2.5.1, we can conclude that ‘appreciation’ is definitely the most assigned value, that the system is still able to assign ‘judgement’ to people, and ‘affect’ is very rare.

  Results on additional data. Like in the case of the corpus data, ‘appreciation’ is the most assigned, ‘judgement’ is still often assigned to people, either expressed with the use of pronouns (e.g. *I am, I know*) or nouns (e.g. *Be a mentor, Americans born, little girl, student killed*), and ‘affect’ is only assigned once.

In terms of research questions, in this Chapter I have managed to give an answer to all of them. In fact, I have shown that it is possible to annotate explicit opinions bringing together both a linguistic and a computational perspective by using appraisal groups, by evaluating the individual importance of the linguistic features chosen for Sentiment Analysis and applied multilingually, and demonstrated that it is possible to classify different types of opinions. I will complete this overview in the following Chapter by focusing more on the multilingual perspective, and showing how the obtained results fit with the other past, current and future works.

Chapter 9

Conclusions

Previous research that has specifically focused on the use of linguistic features for sentiment analysis, despite not being sparse in terms of techniques and domains in the case of the English language, turns out to be very limited for other languages. This work has attempted to fill in the gap by **conducting a systematic study to find out whether there are features that work across multiple languages, and test them on English, Italian and Russian through the creation of a computer software able to automatically extract them**.

Inspired by previous works (Bloom *et al.*, 2007a; Whitelaw *et al.*, 2005), the initial hypotheses of my project have been that an opinion can be captured in a pair (*appraisal group*) consisting of a *target* and a *modifier*, and that the extraction of linguistic features (designed by taking into account the morphological, grammatical, lexical and syntactical characteristics of these three languages) would lead to a more accurate classification of their sentiment in context. The methodology (and related automatic system) have also been designed to identify the attitude of the appraisal groups according to the Appraisal Framework within the Systemic Functional Linguistics.

In order to do so, I designed the annotation scheme *SentiML* (described in Chapter 5) to produce machine-readable annotated texts, and annotated a corpus consisting of originally-produced texts and translations belonging to three different text types (political speeches, news and TED talks) in all the three languages.

As I was the only annotator, I have tried and coped with the subjectivity issue in the annotations by inferring the authors' perspective as much as possible from their opinions, and by comparing the manual annotations to the predictions of automatic classifiers to find inconsistencies (see Section 5.2).

I then created an automatic system that could extract the linguistic features specified in the annotation scheme by relying on syntactical relations under the form of *dependency relations* (Nivre, 2005).

Dependency relations allow to elicit a variety of essential aspects for the task of sentiment analysis: first and foremost, the link between the target and its modifier, but also the presence of negation and polarity reversals.

Afterwards, I evaluated the automatic system in several ways. First of all, both the corpus data (see Section 8.2.5.2) and additional data (see Section 8.3) have been used in order to prevent any wrong considerations due to the corpus size. The additional data were chosen to be of the same text types of the corpus ones. In the related discussions, claims concerning the highly complex and heterogeneous linguistic systems of English, Italian and Russian were made on the basis of previous studies in Translation Studies and of evidence from Corpus Linguistics. Second, a comparison has also been made to a baseline using the sentiment dictionaries and rules based on POS-tags rather than on dependency-parsing relations.

In Chapter 4, the originally-produced texts and translations had been previously analysed to find similarities and differences among languages and their relative cultures, which has served to linguistically support the conclusive remarks based on the quantitative and qualitative analysis of the data. In particular, similarities and differences were highlighted in the context of the *attitude* categories of the Appraisal Framework and of translation-related considerations, including translation strategies (particularly omission, addition, metaphors, punctuation), presence of *universals* (Baker *et al.*, 1993) and linguistic and cultural preferences. Among the main differences, it was found that Russian and Italian sometimes express appraisal in a diminished or marked way with respect to English (a feature annotated as 'force') especially in political speeches, while among the similarities it was found that figurative images and metaphors are mostly kept in both the target languages.

Apart from the novelty in the approach applied to the sentiment analysis task, the present study can also be regarded as a follow-up of previous works in the appraisal field (Manfredi, 2011; Munday, 2012; Pounds, 2010; Thompson & Hunston, 2006), because it focuses on a wider range of text types than those previously analysed, and because it represents one of the first applications of the Appraisal framework (both for computational purposes and non) in Italian and Russian. Thanks to the nature of the methodology and the automatic system tested on languages belonging to different families (Germanic, Romance and Slavonic for English, Italian and Russian respectively), I believe that a further advantage of this work is its potential to be generalizable to other languages.

9.1 Main findings

The initial research questions and hypotheses described in Section 1.3 were tested:

The first question was "**How far is it possible to analyse explicit opinions in order to bring together both a linguistic and a computational perspective?**", to which the hypothesis of the usefulness of appraisal groups to encapsulate explicit opinions was linked. In Chapter 5 I have shown a number of advantages of the annotation scheme *SentiML* including the fact that it covers even complex linguistic patterns (see Section 5.7), by leaving out a very limited amount of cases (see Section 5.6).

As for the automatic identification, the dependency-parsing-based system has outperformed the POS-based baseline in English by achieving 0.31 of F1 (vs. 0.26) and in Russian with 0.35 (vs. 0.33), while it has similar performances in Italian with 0.11 (vs. 0.12). The difference in the outcome (which is visible also in the accuracy of some attributes) strongly depends on the quality of the tools. Particularly important are the POS-taggers, the dependency parsing models and the sentiment dictionaries. While I will describe the advantages and disadvantages found in the use of dependency parsers in the context of the following question, it must be mentioned here that the Russian and Italian sentiment dictionaries, despite consisting of automatic translations from the English words out of context, were useful both in the tasks of identification and classification of the appraisal groups (the lower performance in Italian is most likely to be attributed mainly to other causes that will be discussed later). The attention that has been given to the assessment of the quality of the dictionaries was due to the fact that the presence of the modifier and/or the target in them represents the most important feature in the identification and classification of the appraisal groups. As explained in Section 6.2.2, I have particularly questioned their coverage after the comparison with the manual annotations demonstrated that the words included in my appraisal groups were present in the sentiment dictionaries only 35.33% of times in English, 29.39% of times in Italian and 10.29% of times in Russian.

An additional important consideration related to the measurement of the accuracy in the identification is that translations sometimes present structures more responding to the specified grammatical rules and, as such, more appraisal groups are retrieved. On the other hand, there are also cases in which the subject (or the verb) lacks in the target languages and the appraisal groups cannot be annotated.

The second research question "**What are the linguistic features of evaluative language that can lead to a successful automatic analysis of sentiment across multiple languages?**" was strictly linked to the hypothesis that the chosen set of features would be the best for a good accuracy in the classification of the *orientation*. In order to account for differences across the languages, a number of considerations have been made. For example, a suitable modifier could be found either before or after the target because in English adjectives always precede nouns unless in predicate position (e.g. *the eyes are beautiful*), while in Italian and Russian they are quite flexible.

As for the influence of the granularity of the POS-tagsets, it was found that rules based on fine-grained POS-tags (in Russian and in Italian in the case of TreeTagger) do not have any advantage with comparison to those based on coarse-grained tags (in English and in Italian in the case of TANL).

Performances in the classification of the orientation were calculated according to a score that measures the correctness on the correctly identified groups in the previous phase (see details description in Section 8.2.4). The classification of the orientation turned out to be good, with 0.46 of F1 in English, 0.52 of F1 in Italian and 0.43 in Russian (as shown in Table 8.19).

However, this result was not necessarily better than the baseline, which outperforms by 1.87% for Italian and by 4.04% for Russian. This might be due to a number of reasons.

First of all, the fact that the baseline relies only on the accurate information related to the POS-tags (a manual check confirmed that even challenging cases such as reflexive verbs in Italian, and contracted verb forms in English are correctly tagged), and it is not influenced by the more unpredictable outcome of the dependency parsers (a subject that will be discussed later on).

Second, since the baseline is based on the "noun+adjective" rule, the highest percentage of groups are identified (even if not necessarily carrying appraisal), which results in the increased probability for the baseline to retrieve these words from the sentiment dictionary. The importance of this rule has been demonstrated by the feature ablation in Italian in Section 8.2.3, and in the comparison with the baseline in Section

8.2.4.

Other limitations have been that, because the system architecture had been designed to identify appraisal groups consisting of pairs, it was not possible to consider two nouns linked by a preposition or multi-word units. However, the manual annotations had been done independently from this constraint and the result has been a quite substantial difference between the automatic tagging and the manual annotations.

Additionally, the rules for the 'marked' value of *polarity* and the 'reverse' value of *force* did not fire because none of the conditions were found in the corpus. Their activation would have been a particular important factor to evaluate the quality of the dependency parsing when dealing with negations and polarity reversals. However, a manual check has demonstrated that negation is always recognized and indicated as *neg* by the parsers (apart with the verb *cannot*, which does not represent a problem because it is not to be included in appraisal groups), and adverbs such as *not only, not just, not out* (e.g. "will act not only depended, not out of charity"), which might be easily misunderstood as negation or reversals, are also recognized by the parsers as phrases. The individual attributes of *force* and *polarity* managed to outperform the random values of the baseline anyway, with a difference ranging from 0.13 to 0.66 in all languages.

However, the fact that the conditions for them to be activated were not met in the SentiML corpus resulted in having the orientation of the appraisal groups mainly influenced by the prior orientation provided by the dictionary for targets and modifiers. Nonetheless, this is adjusted according to the the rules specified in Table 7.7 (e.g. a negative modifier and a positive target would result in a negative appraisal group). A detailed comparison of the cases (correctly and wrongly identified and classified by the system) has been provided in the categorizations in Section 8.2.5.

The comparison to the POS-based baseline allowed to test and confirm the initial hypothesis that using the dependency parsing relations would have been beneficial. In fact, when accurate, dependency relations allow to retrieve more appraisal groups (see Section 8.2.4). A good accuracy is achieved in Russian also thanks to the fine-grained tagsets on which the parser is based. Italian should supposedly benefit from a fine-grained tagset as well, but unfortunately it cannot be demonstrated because the F1 achieved in the identification of the appraisal groups is low (0.16).

As for any potential differences in the accuracy of the dependency parser on different text types, it was found that the initial hypothesis of news being more difficult to parse was confirmed for English and Russian. In fact, news achieved the lowest accuracy (0.27 in English and 0.23 in Russian) vs. political speeches (0.33 in English and 0.37 in Russian) and TED talks (0.30 in English and 0.40 in Russian). Conversely, the hypothesis was not confirmed for Italian because the highest F1 was achieved exactly for the news text type (0.27) vs. TED talks (0.16) and political speeches (0.08) (see Table 8.18 in Section 8.4).

As for practical problems with the outcome of the parsers, we have seen that negation and polarity reversals are not among them (see reference above and detailed discussion in Chapter 8). Nevertheless, there are some others. First of all with copulas, mostly in English, because in Italian the parser is not misled even in the case in which the singular verb *è* (it is) is used to refer to two nouns instead of the more grammatically correct plural verb (e.g. *questo è il significato della nostra libertà e del nostro credo* (this is the meaning of our freedom and of our belief)). In Russian copulas do not represent an issue in the present tense because the verb is omitted in this case (глаза (-) прекрасные - these eyes are beautiful). Second, in Italian the wrong classification of reflexive pronouns into personal (e.g. *mi* in "mi perdo" (I lose myself)). Third, the wrong dependency trees of the "long-distance dependency links" (where by *long-distance* a distance longer than 1 is meant), which however are not frequent.

The third research question "**How far is the automatic classification of opinions into the main categories of the Appraisal Framework within Systemic Functional Linguistics possible and useful?**" was related to the hypothesis of automatically extracting the categories of the Appraisal Framework as good labels for different types of opinions. This hypothesis was confirmed since for the attribute *attitude*, the automatic system achieves 0.81 of accuracy in English, 0.72 in Italian and 0.65 in Russian (see Table 8.20), outperforming the baseline that achieves 0.53 in English, 0.54 in Italian and 0.52 in Russian. In terms of values, 'appreciation' is definitely the most assigned one, followed by 'judgement' that the system is mostly able to assign to people, whereas 'affect' is very rare.

Nonetheless, being this result connected to a direct match of *attitude* to *type*, and considering the potential mistakes in the application of the Appraisal Framework in the manual analysis, these results must not be strictly evaluated in terms of percentages, but rather as a first successful attempt to automatically recognize the AF attitude categories multilingually.

9.2 Resources produced

As for the resources produced, of significant importance is the development and distribution of the **multilingual comparable corpus** consisting of well-known text types such as political speeches and news, but also quite new such as TED talks. The corpus consists of approximately 500 sentences and 9000 tokens for each language (detailed figures are provided in Section 1.4).

As mentioned in Section 5.2, the corpus also has many other advantages linked to its *reusability* and *multifunctionality* (McEnery *et al.*, 2006). Apart from being used for other works in sentiment analysis where there is a shortage of manually annotated data, the corpus is also expected to be useful for SFL-related studies. These obviously include those focusing on the Appraisal Framework, where the availability of machine-readable annotated texts is very limited, especially in the form of multilingual comparable resource.

However, potential SFL applications are not limited to the AF. For example, Manfredi (2011) mentions that a number of works in Translation Studies share an SFL perspective, and herself sustains that SFL is a "tool for translators' education and training" since "the lexico-grammatical features found in different text types make students aware of how meanings are realized across languages" and how to "reproduce them in another language" (see the following Section for more suggestions on how to expand this research in this direction).

Machine translation is among other possible uses of both the manual analysis and the corpus. In Halliday (2005), some of his very early works related to machine translation are presented to show how their contributions to this field are still relevant despite the technological advances. This is due to the definition of machine translation in such works as "a problem in applied linguistics, specifically a problem requiring the application of those parts of General Linguistic Theory which deal with the systemic description and comparison of languages". I believe that my work could suggest some of the rules that the computer needs for "systematic relating these two descriptions [of the languages] one to another", rues that take into account the frequency of the terms, their surrounding text and the internal structure of the target language(s).

Another useful resource that I have publicly shared is the specifically-designed annotation scheme SentiML, which can be applied to languages other than those under analysis in this work.

However, the most complete outcome has been the creation of a computer software that, given a sentence or a full text, automatically identifies opinions and classifies the variety of linguistic features useful for Sentiment analysis described above. The system consists of several modules that start working in series with just one initial command. The modules are described in detail in Section 7 and include POS-tagging, CONLL format creation, dependency parsing, pairs extractor and XML output. As mentioned in Chapter 7, the newly created *Maltparser* models for Italian and Russian have also been made publicly available[1].

Given the nature of the topic - the analysis of sentiment - and the range of practical applications - in surveys, blogs and any means in which opinions are expressed in a written form, affecting several domains (e.g. engineering, politics, medicine and business), this would be a good resource that in the near future I intend to make usable through an interface for people without a programming background.

9.3 Future works

A number of aspects have emerged as interesting future works. They are not at the methodological level, since I have demonstrated that the approach based on using pairs achieve reasonably accurate results, and using longer spans such as triples or more would still come with the disadvantage of increasing the complexity in a research in which subjectivity is an ordinary issue.

The future works I am referring to are rather a series of practical improvements of the automatic system. First of all, the improvement of the sentiment dictionaries that I have demonstrated to be limited in accuracy and coverage.

As for the way in which they could be improved, I believe that using *seed* words to populate them with synonyms does not represent a good solution when the goal is a correct contextual orientation, since this is influenced by all the variables discussed in this work. Even the manual analysis could potentially not guarantee a perfect outcome, due to the subjectivity.

Other areas of improvement are definitely represented by the quality of the dependency parsing models, in particular the Italian one (see Section 8.2.3) and the inclusion of more advanced rules to identify polarity and force. If the current system were to be

[1] `http://corpus.leeds.ac.uk/marilena/SentiML/`

extended, a word-sense disambiguation component and a co-reference one would make it more complete, although at the moment there is no need for the second because only annotations at the sentence-level are allowed. In addition, since I have stressed that the same methodology and a similar automatic system could be applied to any other languages for which a POS-tag and dependency parsing systems are available, an alternative could be represented by a machine-learning system rather than rule-based as soon as the corpus size increases.

One aspect that I have demonstrated to have a big potential to be fruitful is the connection between Sentiment analysis and the Appraisal Framework. The advantages mainly relate to the SentiML annotation scheme since, as pointed out in Chapter 5, it allows an annotation span ideal both for the general attitude categories used in this project, but also for more detailed ones like those used by Read *et al.* (2007a). This expectation is also based on the fact that SentiML is flexible enough not to limit appraisal groups to adjectives, but to include several combinations of nouns, pronouns, verbs and adverbs. A possible exploitation of the co-text of sentiment words would be ideal, by matching the appraisal groups to a lexicon that includes peculiar examples of attitudes such as "beautiful person" (judgement) vs. "beautiful girl" (appreciation), as proposed by Bednarek (2009) who, however, warns about the difficulty in finding words that exclusively belong to either the *judging* lexis or the *appreciating* (for example in the case of *important, genuine, expected, possible* and *necessary*).

In the process of producing evidence for these claims, perhaps ways to better deal with implicit sentiment and sarcasm informed by the AF could be proposed.

As for the expansion to the general theory of Systemic Functional Linguistics, in the previous Section I have mentioned that initial work has been done in the thesis. Since the aim of SFL is to study how meanings are conveyed through the ideational, interpersonal and textual metafunctions, at the lexico-grammar level the *transitivity, mood/modality* and *theme/rheme* should be all respectively analysed (Halliday, 1994).

As for transitivity, in Section 3.3 I looked at the frequency and usage of some of the verbs that can be mapped to the processes most related to evaluation, i.e. 'cognition' (*to know, to think, to believe, to realize*), 'affection' (*to like, to love, to hate*) and 'behavioural' (*to blame*). However, a broader study of the processes, along with a link to participants and circumstances would be beneficial.

As for mood and modality, in my work they are encapsulated in the attributes *polarity* and *force* respectively. Statistics related to these attributes on the manually-annotated corpora have been already collected in Section 6.2.2, but it would be interesting to double check the claims on their frequency.

As for theme and rheme, in Section 4.2.4, under the linguistic preferences of each language for the word order, I have only just started to reflect on some of the marked choices found in the SentiML corpus, especially in the political speeches.

At the contextual level, a very brief comparison of *field, tenor* and *mode* as 'register' variables (Halliday, 1994) has been done in Section 1.4 which, however, could definitely be supported with more examples coming from the multilingual corpus and other evidence from disciplines such as Discourse analysis.

Finally, another venue that comes to mind when talking about *Sentiment analysis* is *Emotion analysis*, in which one of the most used set of emotions is Ekman (1992)'s: joy, sadness, anger, fear, disgust and surprise. However, the connection with my work is not immediate, mainly because of the difficulty to encapsulate expressions related to emotions in pairs (or even at the word level as stated by Mohammad (2012). The choice of a new span, new categories and new features would make the expansion of the *SentiML* annotation scheme not feasible.

In addition, all the above would result in an even increased level of subjectivity that would be impossible to face without the presence of multiple annotators. Yet, I believe that it would be very interesting if new research questions could be formulated, especially considering the lack of resources in Emotion analysis for Italian and Russian.

Appendix A

Automatic annotation: macro-average results

A.1 English

	Overall	
	Strict	Lenient
	MODIFIERS	
Precision	0.31	**0.36**
Recall	0.26	**0.33**
F1	0.28	**0.33**
	TARGETS	
Precision	0.42	**0.47**
Recall	0.37	**0.45**
F1	0.38	**0.44**
	APPRAISAL GROUPS	
Precision	0.27	**0.31**
Recall	0.24	**0.30**
F1	0.24	**0.29**

Table A.1: Identification of modifiers, targets and appraisal groups in the overall English dataset. For the macro- average definition see Section 8.1.

	Political		News		TED	
	Strict	Lenient	Strict	Lenient	Strict	Lenient
	MODIFIERS					
Precision	0.55	**0.59**	0.19	**0.26**	0.41	**0.43**
Recall	0.31	**0.33**	0.22	**0.31**	0.35	**0.36**
F1	0.40	**0.43**	0.20	**0.28**	0.38	**0.39**
	TARGETS					
Precision	0.62	**0.63**	0.34	**0.43**	0.44	**0.45**
Recall	0.35	**0.36**	0.40	**0.53**	0.30	**0.30**
F1	0.45	**0.45**	0.37	**0.47**	0.36	**0.36**
	APPRAISAL GROUPS					
Precision	0.46	**0.47**	0.17	**0.24**	0.37	**0.38**
Recall	0.27	**0.28**	0.20	**0.30**	0.30	**0.31**
F1	0.34	**0.35**	0.18	**0.26**	0.33	**0.34**

Table A.2: Identification of modifiers, targets and appraisal groups across text types in English (macro average).

	Overall	
	Strict	Lenient
	MODIFIERS	
Orientation	0.19	**0.23**
Attitude	0.74	**0.81**
Force	0.81	**0.90**
Polarity	0.87	**0.97**
	TARGETS	
Orientation	0.26	**0.28**
Type	0.82	**0.90**
	APPRAISAL GROUPS	
Orientation	0.41	**0.46**

Table A.3: Attributes of modifiers, targets and appraisal groups in the English overall dataset (macro average).

	Political		News		TED	
	Strict	Lenient	Strict	Lenient	Strict	Lenient
	MODIFIERS					
Orientation	0.19	**0.19**	0.18	**0.24**	0.22	**0.22**
Attitude	**0.77**	0.76	0.69	**0.83**	**0.84**	0.83
Force	**0.88**	0.88	0.78	**0.93**	**0.85**	0.84
Polarity	**0.97**	0.97	0.80	**0.97**	0.98	**0.98**
	TARGETS					
Orientation	**0.25**	0.25	0.32	**0.35**	**0.10**	0.10
Type	0.89	**0.89**	0.75	**0.90**	**0.93**	0.92
	APPRAISAL GROUPS					
Orientation	0.46	**0.47**	0.39	**0.47**	0.42	**0.42**

Table A.4: Attributes of modifiers, targets and appraisal groups across text types in English (macro average).

A.2 Russian

	Overall	
	Strict	Lenient
	MODIFIERS	
Precision	0.29	**0.38**
Recall	0.27	**0.36**
F1	0.27	**0.35**
	TARGETS	
Precision	0.28	**0.35**
Recall	0.28	**0.37**
F1	0.27	**0.34**
	APPRAISAL GROUPS	
Precision	0.27	**0.32**
Recall	0.23	**0.32**
F1	0.24	**0.32**

Table A.5: Identification of modifiers, targets and appraisal groups in the overall Russian dataset (macro average).

	Political		News		TED	
	Strict	Lenient	Strict	Lenient	Strict	Lenient
	MODIFIERS					
Precision	0.56	**0.58**	0.20	**0.24**	0.16	**0.40**
Recall	0.35	**0.35**	0.30	**0.38**	0.13	**0.34**
F1	0.43	**0.44**	0.24	**0.29**	0.15	**0.37**
	TARGETS					
Precision	0.45	**0.45**	0.23	**0.27**	0.17	**0.37**
Recall	0.29	**0.29**	0.23	**0.44**	0.15	**0.34**
F1	0.35	**0.35**	0.28	**0.33**	0.16	**0.36**
	APPRAISAL GROUPS					
Precision	0.50	**0.52**	0.19	**0.21**	0.17	**0.40**
Recall	0.29	**0.29**	0.27	**0.32**	0.13	**0.33**
F1	0.37	**0.37**	0.22	**0.25**	0.15	**0.36**

Table A.6: Identification of modifiers, targets and appraisal groups across text types in Russian (macro average).

	Overall	
	Strict	Lenient
	MODIFIERS	
Orientation	**0.13**	0.13
Attitude	0.61	**0.71**
Force	0.70	**0.81**
Polarity	0.83	**0.97**
	TARGETS	
Orientation	0.16	**0.19**
Type	0.67	**0.72**
	APPRAISAL GROUPS	
Orientation	0.34	**0.37**

Table A.7: Attributes of modifiers, targets and appraisal groups in the Russian overall dataset (macro average).

	Political		News		TED	
	Strict	Lenient	Strict	Lenient	Strict	Lenient
	MODIFIERS					
Orientation	**0.15**	0.15	0.14	**0.10**	0.10	**0.14**
Attitude	**0.60**	0.60	**0.81**	0.75	0.34	**0.76**
Force	0.87	**0.87**	**0.85**	0.83	0.30	**0.73**
Polarity	0.95	**0.95**	**0.98**	0.97	0.48	**0.97**
	TARGETS					
Orientation	0.19	**0.20**	**0.19**	0.16	0.08	**0.22**
Type	0.79	**0.79**	**0.81**	0.80	0.35	**0.78**
	APPRAISAL GROUPS					
Orientation	**0.37**	0.37	**0.45**	0.37	0.15	**0.37**

Table A.8: Attributes of modifiers, targets and appraisal groups across text types in Russian (macro average).

A.3 Italian

	Overall			
	TreeTagger		**Tanl**	
	Strict	Lenient	Strict	Lenient
	MODIFIERS			
Precision	0.39	**0.39**	0.29	0.33
Recall	0.14	0.14	0.21	**0.25**
F1	0.20	0.21	0.24	**0.28**
	TARGETS			
Precision	**0.47**	**0.47**	0.37	0.40
Recall	0.16	0.16	0.26	**0.30**
F1	0.24	0.24	0.30	**0.34**
	APPRAISAL GROUPS			
Precision	0.16	0.16	0.12	**0.16**
Recall	0.06	0.06	0.09	**0.14**
F1	0.08	0.08	0.10	**0.15**

Table A.9: Identification of modifiers, targets and appraisal groups in the overall Italian dataset (macro average).

	Overall			
	TreeTagger		**Tanl**	
	Strict	Lenient	Strict	Lenient
	MODIFIERS			
Orientation	**0.23**	0.22	0.18	0.19
Attitude	**0.81**	0.80	0.77	0.75
Force	0.78	0.78	**0.87**	0.85
Polarity	**1.00**	**1.00**	**1.00**	**1.00**
	TARGETS			
Orientation	0.11	0.11	**0.19**	0.16
Type	**0.85**	0.85	0.84	0.80
	APPRAISAL GROUPS			
Orientation	**0.56**	**0.56**	0.43	0.49

Table A.10: Attributes of modifiers, targets and appraisal groups in the overall Italian dataset (macro average).

	Political				News				TED			
	TreeTagger		Tanl		TreeTagger		Tanl		TreeTagger		Tanl	
	Strict	Lenient	Strict	Lenient	Strict	Lenient	Strict	Lenient	Strict	Lenient	Strict	Lenient
	MODIFIERS											
Precision	0.29	**0.30**	0.29	0.30	0.20	**0.28**	0.20	**0.28**	0.39	0.39	0.42	**0.43**
Recall	0.15	0.15	0.17	**0.18**	0.19	**0.27**	0.19	**0.27**	0.14	0.14	0.27	**0.27**
F1	0.20	0.20	0.22	**0.22**	0.19	**0.27**	0.19	**0.27**	0.20	0.21	0.33	**0.33**
	TARGETS											
Precision	0.43	0.44	0.44	**0.45**	0.26	**0.32**	0.26	**0.32**	**0.47**	**0.47**	0.46	0.46
Recall	0.23	0.23	0.27	**0.28**	0.25	**0.33**	0.25	**0.33**	0.16	0.16	0.28	**0.28**
F1	0.29	0.30	0.33	**0.34**	0.25	**0.33**	0.25	**0.33**	0.24	0.24	0.35	**0.35**
	APPRAISAL GROUPS											
Precision	0.11	**0.12**	0.11	0.12	0.09	**0.19**	0.08	0.18	0.16	0.16	0.17	**0.18**
Recall	0.06	0.06	0.07	**0.08**	0.27	**0.19**	0.08	**0.19**	0.06	0.06	0.12	**0.12**
F1	0.08	0.08	0.09	**0.09**	0.27	**0.19**	0.08	0.18	0.08	0.10	0.14	**0.14**

Table A.11: Identification of modifiers, targets and appraisal groups across text types in Italian (macro average).

	Political				News				TED			
	TreeTagger		Tanl		TreeTagger		Tanl		TreeTagger		Tanl	
	Strict	Lenient	Strict	Lenient	Strict	Lenient	Strict	Lenient	Strict	Lenient	Strict	Lenient
MODIFIERS												
Orientation	**0.22**	0.22	0.20	0.19	0.17	0.17	0.17	**0.19**	0.12	0.12	**0.18**	0.18
Attitude	0.72	**0.72**	0.68	0.69	**0.88**	0.83	**0.88**	0.83	**0.82**	0.80	0.68	0.68
Force	0.85	0.86	0.87	**0.87**	**0.92**	0.88	**0.92**	0.88	0.70	0.71	**0.79**	0.77
Polarity	0.99	0.99	0.99	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.99	0.99
TARGETS												
Orientation	0.23	0.22	**0.27**	0.26	**0.20**	0.13	**0.20**	0.13	**0.12**	**0.12**	0.10	0.10
Type	**0.76**	0.73	0.73	0.71	**1.00**	0.92	**1.00**	0.92	**0.80**	**0.80**	0.72	0.73
APPRAISAL GROUPS												
Orientation	**0.51**	0.49	0.43	0.42	0.37	**0.51**	0.37	**0.51**	**0.59**	**0.59**	0.54	0.52

Table A.12: Attributes of modifiers, targets and appraisal groups across text types in Italian (macro average).

References

ABNEY, S. (2007). *Semisupervised Learning for Computational Linguistics*. Chapman & Hall/CRC, 1st edn. 106

ABNEY, S. & BIRD, S. (2010). The human language project: building a universal corpus of the world's languages. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, 88–97, Association for Computational Linguistics. 135, 148

ADAMIC, L.A. & GLANCE, N. (2005). The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, 36–43, ACM. 21

ARGAMON, S., BLOOM, K., ESULI, A. & SEBASTIANI, F. (2007). Automatically Determining Attitude Type and Force for Sentiment Analysis. In *Proceedings of the 3rd Language and Technology Conference (LTC'07)*, 369–373, Poznan, Poland. 19, 30, 74, 76

ARTSTEIN, R. & POESIO, M. (2008). Inter-coder agreement for computational linguistics. *Comput. Linguist.*, **34**, 555–596. 111

ATTARDI, G. & SIMI, M. (2009). Overview of the evalita 2009 part-of-speech tagging task. In *Workshop Evalita*, Citeseer. 158

ATTARDI, G., DELL'ORLETTA, F., SIMI, M., CHANEV, A. & CIARAMITA, M. (2007). Multilingual dependency parsing and domain adaptation using desr. In *EMNLP-CoNLL*, 1112–1118. 31

ATTARDI, G., DELL'ORLETTA, F., SIMI, M. & TURIAN, J. (2009). Accurate dependency parsing with a stacked multilayer perceptron. *Proceedings of EVALITA*, **9**. 31

AZAROWA, I. (2008). Russnet as a computer lexicon for russian. *Proceedings of the Intelligent Information systems IIS-2008*, 341–350. 148

BAKER, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target*, **7.2**, 223–243. 11, 50

BAKER, M. (1996). Corpus-based translation studies: The challenges that lie ahead. *Benjamins Translation Library*, **18**, 175–186. 64, 65, 111, 112

BAKER, M. (2002). *In Other Words: A Coursebook on Translation*. Taylor and Francis. viii, 52, 63, 69, 70

BAKER, M. *et al.* (1993). Corpus linguistics and translation studies: Implications and applications. *Text and technology: In honour of John Sinclair*, **233**, 250. 64, 111, 182

BALAHUR, A., STEINBERGER, R., KABADJOV, M., ZAVARELLA, V., VAN DER GOOT, E., HALKIA, M., POULIQUEN, B. & BELYAEVA, J. (2010). Sentiment analysis in the news. *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'2010)*, 2216–2220. 22

BASILE, V., BOLIOLI, A., NISSIM, M., PATTI, V. & ROSSO, P. (2014). Overview of the evalita 2014 sentiment polarity classification task. *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14). Pisa, Italy*. 24

BATEMAN, J., SHAROFF, S., TEICH, E., KRUIJFF, G.J., KRUIJFF-KORBAYOVÁ, I. & SKOUMALOVÁ, H. (2000). Resources for multilingual text generation in three slavic languages. *Proceedings of LREC 2000, 2nd International Conference on Language Resources & Evaluation*. 52

BAUTIN, M., VIJAYARENU, L. & SKIENA, S. (2008). International sentiment analysis for news and blogs. 19–26. 22

BBC (2009). Bbc monitoring 2009. 53

BEDNAREK, M. (2009). Language patterns and ATTITUDE. *Functions of Language*, **16**, 165–192. 189

BENAMARA, F., CHARDON, B., MATHIEU, Y., POPESCU, V. & ASHER, N. (2012). How do negation and modality impact on opinions? In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, ExProM '12, 10–18, Association for Computational Linguistics, Stroudsburg, PA, USA. 17

BIELSA, E. & BASSNETT, S. (2008). *Translation in global news*. Routledge. 53

BIRD, S. (2006). Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, 69–72, Association for Computational Linguistics. 135, 149

BLOOM, K. & ARGAMON, S. (2009). Automated learning of appraisal extraction patterns. *Language and Computers*, **71**, 249–260. vii, 7, 19, 23, 74, 75, 76

BLOOM, K., GARG, N. & ARGAMON, S. (2007a). Extracting appraisal expressions. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, 308–315, Rochester, New York. vii, 7, 19, 28, 74, 75, 76, 181

BLOOM, K., STEIN, S. & ARGAMON, S. (2007b). Appraisal extraction for news opinion analysis at ntcir-6. In *Proceedings NTCIR-6, National Institute of Informatics*, Citeseer. 22

BOGUSLAVSKY, I., GRIGORIEVA, S., GRIGORIEV, N., KREIDLIN, L. & FRID, N. (2000). Dependency treebank for russian: Concept, tools, types of information. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, 987–991, Association for Computational Linguistics. 30

BOLIOLI, A., SALAMINO, F. & PORZIONATO, V. (2013). Social media monitoring in real life with blogmeter platform. In *ESSEM@ AI* IA*, 156–163, Citeseer. 25

BOLLETTIERI BOSINELLI, R.M. & ULRICH, M. (1999). The state of the art in translation studies: An overview. *Textus*, **XII**, pp. 219–241. 52

BOSCA, A., DINI, L., KOUYLEKOV, M. & TREVISAN, M. (2012). Linguagrid: a network of linguistic and semantic services for the italian language. In N.C.C. Chair), K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis, eds., *Proceedings of the Eight International Conference on Language*

Resources and Evaluation (LREC'12), European Language Resources Association (ELRA), Istanbul, Turkey. 24

BOSCO, C. & MAZZEI, A. (2011). The evalita 2011 parsing task: the dependency track. *Working Notes of EVALITA*, **2011**, 24–25. 31

BOSCO, C., MAZZEI, A., LOMBARDO, V., ATTARDI, G., CORAZZA, A., LAVELLI, A., LESMO, L., SATTA, G. & SIMI, M. (2008). Comparing italian parsers on a common treebank: the evalita experience. In *LREC*. 31

BOSCO, C., MONTEMAGNI, S. & SIMI, M. (2013a). Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 61–69, Citeseer. 31

BOSCO, C., PATTI, V. & BOLIOLI, A. (2013b). Developing corpora for sentiment analysis: The case of irony and senti-tut. *IEEE Intelligent Systems*, 55–63. 24

BRADLEY, M.M. & LANG, P.J. (1999). Affective norms for english words (anew): Instruction manual and affective ratings. 19, 20, 136

BRASLAVSKI, P., USTALOV, D. & MUKHIN, M. (2014). A spinning wheel for yarn: user interface for a crowdsourced thesaurus. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden*, 101–104. 148

BURKETT, D. & KLEIN, D. (2008). Two languages are better than one (for syntactic parsing). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 877–886, Association for Computational Linguistics. 151

CAFFAREL-CAYRON, A. (2006). *A systemic functional grammar of French: From grammar to discourse*. A&C Black. 52

CARRETERO, M. & TABOADA, M. (2014). *Graduation within the scope of Attitude in English and Spanish consumer reviews of books and movies*, 221–240. John Benjamins Publishing Company. 74

CARVALHO, P., SARMENTO, L., SILVA, M.J. & DE OLIVEIRA, E. (2009). Clues for detecting irony in user-generated contents: oh...!! it's "so easy" ;-). In *Proceedings of* *the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, TSA '09, 53–56, ACM, New York, NY, USA. 2

CASOTO, P., DATTOLO, A. & TASSO, C. (2008). Sentiment classification for the italian language: A case study on movie reviews. *Journal of Internet Technology*, **9**, 365–373. 25

CAVALIERE, F. & ABBAMONTE, L. (2006). *Lost in Translation: The Italian Rendering of UNICEF 'The state of the World's Children 2004' Report*, vol. Insights into Specialized Translation, 235–258. Bern: Peter Lang. 52, 57

CERINI, S., COMPAGNONI, V., DEMONTIS, A., FORMENTELLI, M. & GANDINI, G. (2007). Language resources and linguistic theory: Typology, second language acquisition, english linguistics. *Milano, IT*. 23

CERON, A., CURINI, L. & IACUS, S.M. (2014). Opinion mining e integrated sentiment analysis (isa). In *Social Media e Sentiment Analysis*, 27–52, Springer. 25

CETTOLO, M., GIRARDI, C. & FEDERICO, M. (2012). Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, 261–268, Trento, Italy. viii, 11, 46

CHARNIAK, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, NAACL 2000, 132–139, Association for Computational Linguistics, Stroudsburg, PA, USA. 28

CHETVIORKIN, I. & LOUKACHEVITCH, N. (2011). Extraction of domain-specific opinion words for similar domains. In *Proceedings of the RANLP 2011 Workshop on Information Extraction and Knowledge Acquisition*, 7–12, Association for Computational Linguistics, Hissar, Bulgaria. 24

CHETVIORKIN, I. & LOUKACHEVITCH, N. (2013). Evaluating sentiment analysis systems in russian. *ACL 2013*, 12. 23

CHOI, Y. & CARDIE, C. (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, 793–801, Association for Computational Linguistics, Stroudsburg, PA, USA. 17

CHOI, Y. & CARDIE, C. (2009). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, 590–598, Association for Computational Linguistics. 30

CHRISTODOULOPOULOS, C., GOLDWATER, S. & STEEDMAN, M. (2010). Two decades of unsupervised pos induction: How far have we come? In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 575–584, Association for Computational Linguistics. 158

COUNCILL, I.G., MCDONALD, R. & VELIKOVICH, L. (2010). 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*, 51–59, Association for Computational Linguistics. 30

CURRAN, T.O.J.R. & KOPRINSKA, P.A.I. (2013). An annotated corpus of quoted opinions in news articles. In *Proceedings of the Conference*, 516. 23

DANG, Y., ZHANG, Y. & CHEN, H. (2010). A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *Intelligent Systems, IEEE*, **25**, 46–53. 18

DE BEAUGRANDE, R. (1997). On history and historicity in modern linguistics: Formalism versus functionalism revisited. *Functions of Language*, **4/2**, 169–213. 2

DE HAAN, F. (2002). Strong modality and negation in russian. *R. Reppen, S. Fitzmaurice & D. Biber, Using Corpora to Explore Linguistic Variation. Amsterdam/Philadelphie: John Benjamins*, 91–110. 145

DELIN, J. (2000). *The Language of Everyday Life: An Introduction*. SAGE Publications. 11, 12

DENECKE, K. (2008). Using sentiwordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*, 507–512, IEEE. 18, 19

DI BARI, M. (2015). Comparing sentiment annotations in English, Italian and Russian. In *Abstract Book of Corpus Linguistics Lancaster 2015*. vii

DI BARI, M., SHAROFF, S. & THOMAS, M. (2013). SentiML: Functional annotation for multilingual sentiment analysis. In *DH-case 2013: Collaborative Annotations in Shared Environments: metadata, vocabularies and techniques in the Digital Humanities*, ACM International Conference Proceedings. vii, 112, 118

DODDS, P. & DANFORTH, C. (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, **11**, 441–456. xiv, 20, 21

DONG, Z. & DONG, Q. (2003). Hownet-a hybrid language and knowledge resource. In *Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on*, 820–824, IEEE. 28

DOTTI, F. (2013). Overcoming problems in automated appraisal recognition: The attitude system in inscribed appraisal. 5th International Conference on Corpus Linguistics (CILC2013). 156

DYSON, S.B. (2008). Text annotation and the cognitive architecture of political leaders: British prime ministers from 1945-2008. *Journal of Information Technology and Politics*, **5**, 7–18. 22

EBERT, S. & SCHÜTZE, H. (2014). Fine-grained contextual predictions for hard sentiment words. In *EMNLP*, 1210–1215. 23

EKMAN, P. (1992). An argument for basic emotions. *Cognition & Emotion*, **6**, 169–200. 190

ELIA, A., PELOSI, S., MAISTO, A., PAOLO II, V.G. & FISCIANO, S. (2015). Sentiment analysis through finite state automata. 25

ERMAKOV, A. (2009). Knowledge extraction from text and its processing: Current state and prospects. *Information technologies*, **7**, 50–55. 23

ESULI, A. & SEBASTIANI, F. (2006). Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06*, 417–422. 18, 19, 23, 136

FAIRCLOUGH, N. (1989). *Language and Power*. Language in social life series, Longman. 42

FELLBAUM, C. (1998). *WordNet: An electronic lexical database*. MIT Press, Cambridge, MA. 18

GELFENBEYN, I., GONCHARUK, A., LEHELT, V., LIPATOV, A. & SHILO, V. (2003). Automatic translation of wordnet semantic network to russian language. In *International Dialog 2003 Workshop*. 148

GO, A., BHAYANI, R. & HUANG, L. (2009). Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, **1**, 12. 20

GODBOLE, N., SRINIVASAIAH, M. & SKIENA, S. (2007). Large-scale sentiment analysis for news and blogs. 22

GORDON, A.S. & SWANSON, R. (2009). Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop*, San Jose, CA. 46

GUERINI, M., STRAPPARAVA, C. & STOCK, O. (2008). CORPS: A Corpus of Tagged Political Speeches for Persuasive Communication Processing. *Journal of Information Technology and Politics*, **5**, 19–32. 46

HALLIDAY, M. (1973). *Exploration in the Functions of Language*. Explorations in language study, Edward Arnold. 2, 33

HALLIDAY, M. (1992). Language as system and language as instance: the corpus as a theoretical construct. In J. Svartvik, ed., *Directions in corpus linguistics : proceedings of Nobel Symposium 82, Stockholm, 4 - 8 August 1991*, 61–77, Mouton de Gruyter, Berlin [u.a.]. 82

HALLIDAY, M. (1994). *An Introduction to Systemic Functional Linguistics*. London:Arnold, 2nd edn. 32, 189, 190

HALLIDAY, M. & WEBSTER, J. (2009). *Continuum Companion to Systemic Functional Linguistics*. Continuum companions, Continum. 33

HALLIDAY, M.A.K. (1978). *Language as social semiotic: the social interpretation of language and meaning*. University Park Press. vii, 2

HALLIDAY, M.A.K. (2005). *Computational and quantitative studies*, vol. 6. A&C Black. 187

HALLIDAY, M.A.K. & MATTHIESSEN, C.M.I.M. (2004). *An introduction to functional grammar*. Hodder Arnold London, 3rd edn. 67

HARRIS, Z. (1951). *Methods in Structural Linguistics*. University of Chicago Press, Chicago. 2

HARRIS, Z. (1954). Distributional structure. *Word*, **10**, 146–162. 2

HU, M. & LIU, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177, ACM. 19, 136

HUNSTON, S. (2000). Evaluation and the planes of discourse: status and value in persuasive texts. In S. Hunston & G. Thompson, eds., *Evaluation in Text: authorial stance and the construction of discourse*, 176–207, Oxford University Press, Oxford, England. 76

HUNSTON, S. & SINCLAIR, J. (2000). A local grammar of evaluation. In S. Hunston & G. Thompson, eds., *Evaluation in Text: authorial stance and the construction of discourse*, 74–101, Oxford University Press, Oxford, England. 34, 37, 75

HUNSTON, S. & THOMPSON, G. (2000). *Evaluation in Text : Authorial Stance and the Construction of Discourse: Authorial Stance and the Construction of Discourse*. Oxford University Press, UK. 12, 33, 34

HYLAND, K. (1995). The author in the text: Hedging scientific writing. vol. 18 of *Hong Kong papers in Linguistics and Language Teaching*. 125

IACOBINI, C. & MASINI, F. (2007). Verb-particle constructions and prefixed verbs in italian: typology, diachrony and semantics. In *On-line Proceedings of the Fifth Mediterranean Morphology Meeting (MMM5)*. 116

IDE, N. & ROMARY, L. (2004). International standard for a linguistic annotation framework. *Nat. Lang. Eng.*, **10**, 211–225. 82

IKEDA, D., TAKAMURA, H., RATINOV, L.A. & OKUMURA, M. (2008). Learning to shift the polarity of words for sentiment classification. In *IJCNLP*, 296–303. 30

JIA, L., YU, C. & MENG, W. (2009). The effect of negation on sentiment analysis and retrieval effectiveness. *Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09*, 1827. 17, 30

JIN, W., HO, H.H. & SRIHARI, R.K. (2009). Opinionminer: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, 1195–1204, ACM, New York, NY, USA. 104

KAN, D. (2012). Rule-based approach to sentiment analysis at romip 2011. *Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog*. 24

KATO, Y., KUROHASHI, S., INUI, K., MALOUF, R. & MULLEN, T. (2008). Taking sides: User classification for informal online political discourse. *Internet Research*, **18**, 177–190. 21

KENNEDY, A. & INKPEN, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, **22**, 110–125. 30

KENNEDY, A., KAZANTSEVA, A., INKPEN, D. & SZPAKOWICZ, S. (2012). Getting emotional about news summarization. In *Canadian Conference on AI*, 121–132, Springer. 18

KRUIJFF, G.J., TEICH, E., BATEMAN, J., KRUIJFF-KORBAYOVÁ, I., SKOUMALOVÁ, H., SHAROFF, S., SOKOLOVA, L., HARTLEY, T., STAYKOVA, K. & HANA, J. (2000). Multilinguality in a text generation system for three Slavic languages. In *Proceedings of the 18th conference on Computational linguistics - Volume 1*, COLING '00, 474–480, Saarbrücken, Germany. 115

KUSHAL, D. & LAWRENCE, S. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. 28

KUSSMAUL, P. (1997). *Text-Type Conventions and Translating: Some Methodological Issues*, vol. Text Typology and Translation of *Benjamins translation library*. John Benjamins Publishing. 124, 125

LAVELLI, A., HALL, J., NILSSON, J. & NIVRE, J. (2009). Maltparser at the evalita 2009 dependency parsing task. *Proceedings of EVALITA Evaluation Campaign*. 31

LAVID, J., ARÚS, J. & ZAMORANO-MANSILLA, J.R. (2010). *Systemic functional grammar of Spanish: A contrastive study with English*, vol. 1. A&C Black. 52

LAVIOSA, S. (1998). Core patterns of lexical use in a comparable corpus of english narrative prose. *Meta: Journal des traducteursMeta:/Translators' Journal*, **43**, 557–570. 64

LAVIOSA, S. (2002). *Corpus-based Translation Studies: Theory, Findings, Applications*. Approaches to translation studies, Rodopi. 50, 64, 66

LIN, D. (1998). Dependency-based evaluation of minipar. In *Proc. Workshop on the Evaluation of Parsing Systems*, Granada. 28

LIU, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 1–38. 17, 19

LIU, B. (2012). *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers. vii, 1, 15, 16, 17, 18

LIU, J. & SENEFF, S. (2009). Review sentiment scoring via a parse-and-paraphrase paradigm. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, 161–169, Association for Computational Linguistics, Stroudsburg, PA, USA. 18

LYDING, V., STEMLE, E., BORGHETTI, C., BRUNELLO, M., CASTAGNOLI, S., DELL'ORLETTA, F., DITTMANN, H., LENCI, A. & PIRRELLI, V. (2014). The paisa corpus of italian web texts. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, 36–43. 128

MAGNINI, B., CAPPELLI, A., TAMBURINI, F., BOSCO, C., MAZZEI, A., LOMBARDO, V., BERTAGNA, F., CALZOLARI, N., TORAL, A., LENZI, V.B. *et al.* (2008). Evaluation of natural language tools for italian: Evalita 2007. In *LREC*. 158

MAISTO, A. & PELOSI, S. (2014). Feature-based customer review summarization. In *On the Move to Meaningful Internet Systems: OTM 2014 Workshops*, 299–308, Springer. 25

MALONE, J. (1988). *The Science of Linguistics in the Art of Translation: Some Tools from Linguistics for the Analysis and Practice of Translation*. SUNY series in linguistics, State University of New York Press. 52

MANFREDI, M. (2011). Systemic functional linguistics as a tool for translation teaching: towards a meaningful practice. *Rivista Internazionale di Tecnica della Traduzione*, **13**, 49 – 62. 52, 182, 187

MANNING, C.D., RAGHAVAN, P. & SCHÜTZE, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. 140

MARRAZZO, F. (2014). Dai sondaggi alla sentiment analysis: le previsioni politico-elettorali nel racconto dei mass media. *Mediascapes journal*, 97–123. 25

MARTIN, J.R. & WHITE, P.R. (2005). *The language of evaluation*. Palgrave Macmillan, Basingstoke and New York. vii, 3, 7, 35, 36, 39, 55

MATTHIESSEN, C. (2009). *Ideas and New Directions*, vol. Bloomsbury Companion to Systemic Functional Linguistics, chap. 3. Bloomsbury Publishing. 52

MCENERY, T. (2005). *Corpus Linguistics*, vol. The Oxford handbook of computational linguistics, chap. 24. Oxford University Press. 82

MCENERY, T., XIAO, R. & TONO, Y. (2006). *Corpus-based Language Studies: An Advanced Resource Book*. Routledge applied linguistics, Routledge. 6, 79, 81, 83, 187

MEJOVA, Y. (2009). Sentiment analysis: An overview. Tech. rep., University of Iowa. 17

MELVCUK, I.A. (1988). *Dependency syntax: theory and practice*. SUNY Press. 30

MIAO, Q., LI, Q. & DAI, R. (2008). An integration strategy for mining product features and opinions. In *Proceedings of the 17th ACM conference on Information and knowledge management*, 1369–1370, ACM. 19

MOHAMMAD, S. (2011). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 105–114, Portland, OR, USA. 18, 106, 118, 135

MOHAMMAD, S. (2012). From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, –. 20, 190

MOILANEN, K. & PULMAN, S. (2007). Sentiment composition. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2007)*, 378–382. 30

MOREO, A., ROMERO, M., CASTRO, J. & ZURITA, J.M. (2012). Lexicon-based comments-oriented news sentiment analyzer system. *Expert Systems with Applications*, **39**, 9166–9180. 23

MUDRAYA, O., PIAO, S., LOFBERG, L., RAYSON, P. & ARCHER, D. (2005). English-russian-finnish cross-language comparison of phrasal verb translation equivalents. 115

MUHLBERGER, P. & ZHANG, W. (2008). Automated extraction of mental models from obama's and McCain's campaign speeches: A natural language processing and statistical approach. Chicago, IL. 22

MULLEN, T. & COLLIER, N. (2004). Sentiment analysis using support vector machines with diverse information sources. In *EMNLP*, vol. 4, 412–418. 19

MUNDAY, J. (2012). *Evaluation in Translation: Critical Points of Translator Decision-making*. Routledge. 34, 50, 52, 53, 54, 55, 82, 182

NAKAGAWA, T., INUI, K. & KUROHASHI, S. (2010). Dependency tree-based sentiment classification using crfs with hidden variables. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, 786–794, Association for Computational Linguistics, Stroudsburg, PA, USA. 7, 17, 19, 29

NASUKAWA, T. & YI, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd international conference on Knowledge capture*, 70–77, ACM. 19

NEALE, A. (2006). *Matching corpus data and system networks: using corpora to modify and extend the system networks for TRANSITIVITY in English*, vol. System and Corpus - Exploring connections, chap. 7, 143–163. Equinox London. 50, 51

NEUMANN, S. & HANSEN-SCHIRRA, S. (2005). The croco project: Cross-linguistic corpora for the investigation of explicitation in translations. *Proceedings from the Corpus Linguistics Conference Series*, **1**. 52

NEVIAROUSKAYA, A. & AONO, M. (2012). Analyzing sentiment word relations with affect, judgment, and appreciation. In *Proceedings of the 2nd Workshop on Sentiment Analysis where AI meets Psychology*, 37–52, The COLING 2012 Organizing Committee, Mumbai, India. 74

NEVIAROUSKAYA, A., PRENDINGER, H. & ISHIZUKA, M. (2007). Textual affect sensing for sociable and expressive online communication. 220–231. 20

NEVIAROUSKAYA, A., PRENDINGER, H. & ISHIZUKA, M. (2010). Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, 806–814, Stroudsburg, PA, USA. 74

NG, V., DASGUPTA, S. & ARIFIN, N. (2006). Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. *Proceedings of the COLING/ACL on Main*, 611–618. 28

NIELSEN, F.Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*. 19, 136

NIVRE, J. (2005). Dependency grammar and dependency parsing. Tech. rep., Växjö University: School of Mathematics and Systems Engineering. vii, 3, 8, 182

NIVRE, J. (2014). Universal dependency parsing. 160

OAKES, M. & JI, M. (2012). *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*. Studies in Corpus Linguistics, John Benjamins Publishing Company. 50

O'DONNELL, M. & BATEMAN, J.A. (2005). Sfl in computational contexts: a contemporary history. *Continuing Discourse on Language: A functional perspective*, 343–382. 3

OHANA, B. & TIERNEY, B. (2009). Sentiment classification of reviews using sentiwordnet. In *9th. IT & T Conference*, 13. 18, 30

OLOHAN, M. (2004). *Introducing corpora in translation studies*. Routledge. 66

PAGANO, A., MAGALHÃES, C. & ALVES, F. (2004). Towards the construction of a multilingual, multifunctional corpus: factores in the design and application of cordiall. *Tradterm*, **10**, 143–161. 52

PAK, A. & PAROUBEK, P. (2012). Language independent approach to sentiment analysis. *Proceeding of Dialog 2012*. 30

PANG, B. & LEE, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, **2**, 1–135. 14, 19, 21

PANG, B., LEE, L. & VAITHYANATHAN, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, 79–86, Association for Computational Linguistics. 19

PAPPAS, N. & POPESCU-BELIS, A. (2013). Sentiment analysis of user comments for one-class collaborative filtering over ted talks. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, 773–776, ACM. 23

PAPPAS, N., MARCONI, R. & POPESCU-BELIS, A. (2014). Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing*, EPFL-CONF-200899, 455–466. 23

PAZELSKAYA, A. & SOLOV'EV, A. (2011). A method of sentiment analysis in russian texts in russian texts. *Proceeding of Dialog 2011*. 24

PERRIE, J., ISLAM, A., MILIOS, E. & KESELJ, V. (2013). Using google n-grams to expand word-emotion association lexicon. In *Computational Linguistics and Intelligent Text Processing*, 137–148, Springer. 18

PIANTA, E., BENTIVOGLI, L. & GIRARDI, C. (2002). Multiwordnet: developing an aligned multilingual database. In *Proceedings of the first international conference on global WordNet*, vol. 152, 55–63. 153

POLANYI, L. & ZAENEN, A. (2006). Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications*, 1–10, Springer. 30

POPESCU, A.M. & ETZIONI, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, 339–346, Association for Computational Linguistics, Stroudsburg, PA, USA. 19, 28

POUNDS, G. (2010). Attitude and subjectivity in italian and british hard-news reporting: The construction of a culture-specific 'reporter'voice. *Discourse Studies*, **12**, 106–137. 52, 53, 71, 182

PRAKASH, R.S.S., JURAFSKY, D. & NG, A.Y. (2007). Learning to merge word senses. *EMNLP-CoNLL 2007*, 1005. 151

PUSTEJOVSKY, J. & STUBBS, A. (2012). *Natural Language Annotation for Machine Learning*. Oreilly and Associate Series, O'Reilly Media, Incorporated. 111

QU, L., IFRIM, G. & WEIKUM, G. (2010). The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, 913–921, Association for Computational Linguistics, Stroudsburg, PA, USA. 18

READ, J., HOPE, D. & CARROLL, J. (2007a). Annotating expressions of appraisal in English. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, 93–100, Stroudsburg, PA, USA. 74, 79, 80, 189

READ, J., HOPE, D. & CARROLL, J. (2007b). Annotating expressions of appraisal in English. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, 93–100, Stroudsburg, PA, USA. 111

RECASENS, M., DANESCU-NICULESCU-MIZIL, C. & JURAFSKY, D. (2013). Linguistic models for analyzing and detecting biased language. In *ACL (1)*, 1650–1659. 151

REISS, K. (1971/2000). Translation criticism, the potentials and limitations: Categories and criteria for translation quality assessment (trans. errol f. rhodes). In *The Translator*, vol. 6 of *Special Issue. Evaluation and Translation*, 347–356, St. Jerome Pub. 68

ROMAGNUOLO, A. (2009). Political discourse in translation political discourse in translation. a corpus-based perspective on presidential inaugurals. *Discourse Analysis and Translation Studies. Special issue of Translation and Interpreting Studies*, **4**, 1–30. 50

ROVENTINI, A., ALONGE, A., CALZOLARI, N., MAGNINI, B. & BERTAGNA, F. (2000). Italwordnet: a large semantic database for italian. In *LREC*. 153

RYLOV, J. (2003). Aspekty jazykovoj kartiny mira: ital'janskij i russkij jazyki. *Kafedra Romanskoj filologij*. 71, 72

SALAGER-MEYER, F. (1994). Hedges and textual communicative function in medical english written discourse. *English for Specific Purposes*, **13**, 149–170. 124, 125

SCHÄFFNER, C. (2004). Political discourse analysis from the point of view of translation studies. *Journal of language and politics*, **3**, 117–150. 53

SCHÄFFNER, C. (2008). 'the prime minister said...': voices in translated political texts. *Synaps*, **22**, 3–25. 53

SCHMID, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, vol. 12, 44–49, Citeseer. 127

SCHWARTZ, R., ABEND, O. & RAPPOPORT, A. (2012). Learnability-based syntactic annotation design. In *COLING*, vol. 24, 2405–2422. 160

SHAIKH, M.A.M., PRENDINGER, H. & MITSURU, I. (2007). Assessing sentiment of text by semantic dependency and contextual valence analysis. In *Affective Computing and Intelligent Interaction*, 191–202, Springer. 30

SHANAHAN, J., QU, Y. & WIEBE, J. (2006). *Computing Attitude and Affect in Text: Theory and Applications*. Springer international series on information retrieval, Springer. 30

SHAROFF, S. (2004). What is at stake: a case study of russian expressions starting with a preposition. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, 17–23, Association for Computational Linguistics. 115

SHAROFF, S. & NIVRE, J. (2011). The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In *Proc. Dialogue, Russian International Conference on Computational Linguistics*, Bekasovo. 131

SHAROFF, S., KOPOTEV, M., ERJAVEC, T., FELDMAN, A. & DIVJAK, D. (2008). Designing and evaluating a russian tagset. In *LREC*. 158

SHASTRI, L., PARVATHY, A.G., KUMAR, A., WESLEY, J. & BALAKRISHNAN, R. (2010). Sentiment extraction: Integrating statistical parsing, semantic analysis, and common sense reasoning. In *IAAI*. 1, 2

SILBERZTEIN, M. (2003). Nooj manual. *Available for download at: www. nooj 4nlp. net*. 25

ŠMELEV, A. (2002). Russkaja jazykovaja model' mira. *Jazyki Slavjanskoj Kul'tury*. 71

SNOW, R., O'CONNOR, B., JURAFSKY, D. & NG, A.Y. (2008). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, 254–263, Association for Computational Linguistics, Stroudsburg, PA, USA. 104

SOLOV'EV, A., ANTONOVA, A. & PAZELSKAYA, A. (2012). Using sentiment-analysis for text information extraction. *Proceeding of Dialog 2012*. 24

SORGENTE, A., FLEGREI, V.C., VETTIGLI, G. & MELE, F. (2014). An italian corpus for aspect based sentiment analysis of movie reviews. In *CLICIT2014*. 25

STEINBERGER, J., EBRAHIM, M., EHRMANN, M., HURRIYETOGLU, A., KABADJOV, M., LENKOVA, P., STEINBERGER, R., TANEV, H., VÁZQUEZ, S. & ZAVARELLA, V. (2012). Creating sentiment dictionaries via triangulation. *Decision Support Systems*, **53**, 689–694. 24

STONE, P.J. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. The MIT Press. 30

STRAPPARAVA, C., VALITUTTI, A. *et al.* (2004). Wordnet affect: an affective extension of wordnet. In *LREC*, vol. 4, 1083–1086. 18, 23

STUBBS, A. (2011). Mae and mai: Lightweight annotation and adjudication tools. In *Linguistic Annotation Workshop*, 129–133. viii, 12, 77

STUBBS, M. (1996a). *Text and corpus analysis: Computer-assisted studies of language and culture*. Blackwell Oxford. 71

STUBBS, M. (1996b). *Towards a modal grammar of English: a matter of prolonged fieldwork*, vol. Text and Corpus Analysis. Blackwell, Oxford. 124

SUGIMOTO, C.R., THELWALL, M., LARIVIÈRE, V., TSOU, A., MONGEON, P. & MACALUSO, B. (2013). Scientists popularizing science: characteristics and impact of ted talk presenters. 23

SUN, W., RUMSHISKY, A. & UZUNER, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, **20**, 806–813. 162

TABOADA, M. & CARRETERO, M. (2012). Contrastive analyses of evaluation in text: Key issues in the design of an annotation system for attitude applicable to consumer reviews in english and spanish. *Linguistics and the Human Sciences*, **6**, 275–295. 52, 74

TABOADA, M. & GRIEVE, J. (2004). Analyzing appraisal automatically. In *In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 158–161. 19, 37

TABOADA, M., BROOKE, J., TOFILOSKI, M., VOLL, K. & STEDE, M. (2011). Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, **37**, 267–307. 18

TAYLOR, C. (1998). *Language to Language: A Practical and Theoretical Guide for Italian/English Translators*. Cambridge University Press. 52

TEICH, E. (1999). System-oriented and text-oriented comparative linguistic research: cross-linguistic variation in translation. *Languages in Contrast*, **2**, 187–210. 51

TEICH, E. (2001). *Towards a model for the description of cross-linguistic divergence and commonality in translation*, vol. Exploring Translation and Multilingual Text Production: Beyond Content of *Volume 3 of Text, translation, computational processing*. Walter de Gruyter. 51, 69, 70

TEICH, E. (2003). *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*, vol. 5. Walter de Gruyter. 51, 52

THOMPSON, G. & HUNSTON, S. (2006). *System And Corpus: Exploring Connections*. Functional Linguistics, Equinox Publishing Limited. 51, 182

TOGNINI-BONELLI, E. (2001). *Corpus Linguistics at Work*. Studies in corpus linguistics, J. Benjamins. 50

VAN DIJK, T. (1996). *Discourse, power and access*, vol. Texts and Practices: Readings in Critical Discourse Analysis of *Linguistics: Media studies*, chap. 5. Taylor and Francis. 39

VAN DIJK, T. (2008). *Critical Discourse Analysis*, vol. The Handbook of Discourse Analysis, chap. 18, 349–371. Blackwell Publishers Ltd. 39

VOSSEN, P. *et al.* (1997). Eurowordnet: a multilingual database for information retrieval. In *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, 5–7. 153

WANG, S. & MANNING, C.D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 90–94, Association for Computational Linguistics. 30

WHITE, P.R. (2003). *Beyond modality and hedging: A dialogic view of the language of intersubjective stance*, vol. Text 23(2), 259–284. Walter de Gruyter. 124

WHITELAW, C., GARG, N. & ARGAMON, S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, 625–631, New York, NY, USA. vii, xiv, 74, 75, 181

WIEGAND, M., BALAHUR, A., ROTH, B., KLAKOW, D. & MONTOYO, A. (2010). A survey on the role of negation in sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, 60–68, Association for Computational Linguistics. 17

WIERZBICKA, A. (1992). *Semantics, culture, and cognition: Universal human concepts in culture-specific configurations*. oxford university Press. 71

WILSON, T., WIEBE, J. & HWA, R. (2004). Just how mad are you? finding strong and weak opinion clauses. In *In Proceedings of AAAI*, 761–769. 19, 23

WILSON, T., WIEBE, J. & HOFFMANN, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 347–354. 17, 27

WILSON, T.A. (2008). *Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. Ph.D. thesis, University of Pittsburgh. 11, 17, 80, 111

YOUNG, L. & SOROKA, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, **29**, 205–231. 19, 136

YU, B., KAUFMANN, S. & DIERMEIER, D. (2008). Classifying party affiliation from political speech. *Journal of Information Technology and Politics*, **5**, 33–48. 22

YU, N. (2014). Exploring co-training strategies for opinion detection. *Journal of the Asssociation for Information Science and Technology*. 105

ZAGIBALOV, T., BELYATSKAYA, K. & CARROLL, J. (2010). Comparable english-russian book review corpora for sentiment analysis. In *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, 67–72. 1, 19, 24, 74

ZHAO, Y., QIN, B., CHE, W. & LIU, T. (2011). Appraisal expression recognition with syntactic path for sentence sentiment classification. *Int. J. Comput. Proc. Oriental Lang.*, **23**, 21–37. 7, 19, 28