Annotating Satire in Italian Political Commentaries with Appraisal Theory

Michele Stingo, Rodolfo Delmonte
Department of Language Studies & Department of Computer Science
Ca' Bembo 1075 – Ca' Foscari University – 30123 Venezia (Italy)
delmont@unive.it

1 Introduction

We present work carried out on journalistic political commentaries in two Italian newspapers, by two well-known Italian journalists, Maria Novella Oppo, a woman, and Michele Serra, a man[1]. Political commentaries published on a daily basis consists of short texts not exceeding 400 words each. Sixty-four texts come from Michele Serra's series titled "L'Amaca", published daily on the newspaper "La Repubblica" between 2013 and 2014; usually the targeted subjects are politicians, bad social habits and in general every trendy current event. Forty-nine texts come from Maria Novella Oppo's series titled "Fronte del video", published daily on the newspaper "L'Unità" in a previous span of time, say from 2011 to 2012; the targeted subjects are usually politicians and televised political talk shows.

The two journalists have been chosen for specific reasons: Oppo is a master in highly cutting and caustic writing, Serra is less so. Both are humorous, however, Oppo is more witty in building the overall logical structure of the underlying satiric network of connections. Oppo borders sarcasm, Serra never does so. Oppo's texts are slightly longer than Serra's.

Italy is not included in the upper part of the list of countries where the freedom of information is very strong. Because of her trenchant and stinging style, Oppo has been publicly attacked by some of her favourite targets, politicians including Berlusconi and Beppe Grillo, who reacted bitterly to her commentaries by including her in a black list of journalists criticizing his movement called Movimento5Stelle. In particular, she has been attacked – and heavily offended - by Grillo who claimed that since she has always been working for the same newspaper for all her career, she will be out of her job in case the Parliament approves the law that precludes newspapers from receiving public financial support. And she will be obliged to find a new job.

http://www.beppegrillo.it/2013/12/giornalista_del_giorno_maria_novella_oppo_lunita.html

Criticisms to the attack has come from many sources as for instance in

http://www.articolo21.org/2013/12/grillo-giu-le-mani-da-maria-novella-oppo/

Berlusconi wanted to sue the newspaper l'Unità where Oppo published her commentaries, but the action didn't result in a real lawsuit for defamation or libel action, and were regarded a case of wrongful prosecution.

In order to focus on the specific features connotating political satire, the human annotation has been carried out on these 112 texts using a reduced (and modified with new criteria, where needed) version of the Appraisal Framework (Martin & White, 2005). Below and in the next two sections we will delve into a precise description of both the framework and the specific criteria devised for text annotation.

2. Satire and the Appraisal Framework

The decision of adopting Appraisal Theory (hence APTH) is based on the fact that previous approaches to detect irony - a word we will use to refer to satire/sarcasm - in texts have failed to explain the phenomenon. Computational research on the topic has been based on the use of shallow features, so as in (Carvalho et al., 2009), (Burfoot & Baldwin, 2009), (Davidov et al., 2010), (Reyes & Rosso, 2011), (Owais et al., 2015), in order to train statistical model with the hope that when optimized for a particular task, they would come up with a reasonably acceptable performance. However, they would not explain the reason why a particular Twitter snippet or short Facebook text has been evaluated as containing satiric/sarcastic expressions. Except perhaps for features based on text exterior appearance, i.e. use of specific emoticons, use of exaggerations, use of unusually long orthographic forms, etc. which however is not applicable to the political satire texts. These texts are long texts, from 200 to 400 words long and do not compare with previous experiments.

The other common approach used to detect irony, in the majority of the cases, is based on polarity detection. So-called Sentiment Analysis is in fact an indiscriminate labeling of texts either on a lexicon basis or on a supervised feature basis (Gianti et al., 2012), (Wang, 2013), (Bosco et al., 2015), (Hernandez Farias et al., 2015), (Özdemir & Bergler, 2015), where in both cases, it is just a binary decision that has to be taken. This is again not explanatory of the phenomenon and will not help in understanding what is it that causes humorous reactions to the reading of an ironic piece of text. It certainly is of no help in deciding which phrases, clauses or just multiwords or simply words, contribute to create the ironic meaning.

By adopting Appraisal analysis, we intended not only to describe but also to compute with some specificity the linguistic regularities which constitute the evaluative styles or keys of political journalistic texts. The theory put forward by White and Martin(2005) (hence M&W) makes available an extended number of semantically and pragmatically motivated annotation schemes that can be applied to any text in order to draw precise conclusions.

[1] Permission to republish excerpts from their articles has been granted personally by the authors.

In particular, one preliminary hypothesis would be being able to ascertain whether the text under analysis is just a simple report, a report with criticism, a report with criticism and condemnation. This is something that can be established in a totally safe and stable manner by simply counting and comparing the type of categories and subcategories present in the annotation of the text. In the book by M&W there's a neat distinction between three types of voices: ‘reporter voice’, ‘correspondent voice’ and ‘commentator voice’. Only the commentator voice has the possibility to condemn, criticize and report at the same time, and since we assume that satire, and even more, sarcasm have a strong component made of social moral sanction, we are automatically selecting this as the target of our research hypothesis.

In APTH, the evaluative field called Attitude is organized into three subclasses, Affect, Appreciation and Judgement, and it is just the latter one that contains subcategories that fit our hypothesis. We are referring first of all to Judgement which alone can allow social moral sanction, and to its subdivision into two subfields, Social Esteem and Social Sanction. In particular, whereas Social Esteem extends from Admiration/Admire vs Criticism/Criticise, Social Sanction deals with Praise vs Condemn. As reported in M&W p.52 "… Judgements of esteem have to do with ‘normality’ (how unusual someone is), ‘capacity’ (how capable they are) and ‘tenacity’ (how resolute they are); judgements of sanction have to do with ‘veracity’ (how truthful someone is) and ‘propriety’ (how ethical someone is). Social esteem tends to be policed in the oral culture, through chat, gossip, jokes and stories of various kinds – with humour often having a critical role to play... Sharing values in this area is critical to the formation of social networks (family, friends, colleagues, etc.). Social sanction on the other hand is more often codified in writing, as edicts, decrees, rules, regulations and laws about how to behave as surveilled by church and state – with penalties and punishments as levers against those not complying with the code. Sharing values in this area underpins civic duty and religious observances."

The texts we have annotated show the use of any type of judgement, expressed directly by the writer. As M&W (p.170) define it, the "commentator voice" is an evaluative style typically only of commentary, opinion and editorials. "It is typical of this category in being primarily concerned with assessments of social sanction, but with also making some reference to assessments of social esteem."(ibid.p.170) And further on, we read, "It would seem that within broadsheet journalistic discourse, this function of ‘sanctioning’ – whether it be via attitudinal assessments or via directives (modals of obligation) – is confined to the one journalistic role, that of commentator. Even though the correspondent voice writer may argue and evaluate, they typically refrain from either mode of ‘sanctioning’."(ibid.p.181)

So eventually in our texts we are dealing with the "commentator voice", which may consist of authorial social sanction, plus authorial directives (proposals), in addition to criticism.

3 Annotating Italian Political Journalistic Texts

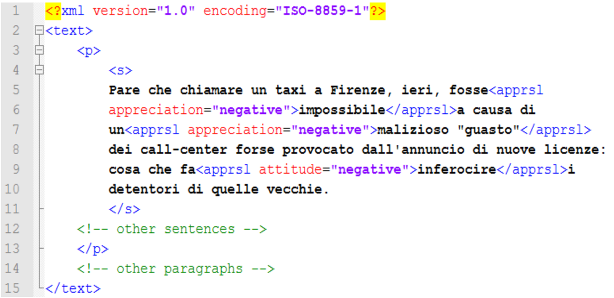
For out annotation work we limited ourselves to using one single subsystem. The Attitude subsystem describes the author’s feelings as they are conveyed within the text, and it is articulated into three main semantic regions with their relative positive/negative polarity, namely:

- Affect: describes proper feelings and any emotional reaction within the text aimed towards human behaviour/process and phenomena.
- Judgement: considers the ethical evaluation on people and their behaviours.
- Appreciation: represent any aesthetic evaluation of things, both man-made and natural phenomena.

The choice to rule-out the others two subsystem (Engagement and Graduation) and the features of the three sub-categories of the Attitude subsystem, was made mainly to maintain the notational work on a manageable level, and also because we were more interested in a coarse quantitative substantiation of the authors‘ opinions within the analyzed texts, rather than conducting a fine-grained analysis about their construction or graduation. In other words, we wanted to assess how descriptive a plain recognition of evaluative sequences is without further detailed information.

3.2 Using the XML format

The annotation work on the texts has been accomplished using the Extensible Markup Language due to its flexibility and because of the possibility to use specifically devised tags. Following there is a snippet of the XML annotation.



```xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<text>
    <p>
        <s>
        Pare che chiamare un taxi a Firenze, ieri, fosse<apprsl
        appreciation="negative">impossibile</apprsl>a causa di
        un<apprsl appreciation="negative">malizioso "guasto"</apprsl>
        dei call-center forse provocato dall'annuncio di nuove licenze:
        cosa che fa<apprsl attitude="negative">inferocire</apprsl>i
        detentori di quelle vecchie.
        </s>
    <!-- other sentences -->
    </p>
    <!-- other paragraphs -->
</text>
```

The tags we used for the annotation include a tag for <text> contains the whole text of the article; <p> serve to mark paragraphs, and <s> to mark sentences. However, every time the article was published as a unique block of text, we structured the article content, first identifying the sentences and then grouping them in relation to their meaning: when a sentence was strictly related to one or more of the following/previous propositions, they were clustered together within the same paragraph. Focusing on the annotation of the evaluative sequences instead, every time we found an evaluative word (or sequence of words) within a political satire article, we delimited the item/phrase within the tags <apprsl></apprsl>. Subsequently, following the general indications mentioned above provided by (M&W 2005), we assigned one of the three subcategories – affect, judgement and appreciation – as attribute of the tag <apprsl>, also providing the positive/negative sentiment orientation as value of the attribute.

3.3 Linguistic Criteria for Annotation

Since the text typology we annotated showed a lot of complex linguistic features, and because no previous work was found on labelling long satiric texts using the Appraisal Framework – nor in the computational linguistics framework, we had to address the annotation task using brand-new criteria specifically designed for the purpose of isolating as many evaluative items/sequences as possible. The criteria, in relation to their most relevant linguistic aspect, are grouped in one of the following set of notational principles, namely lexical, semantic and syntactic set.

Lexical criteria: these notational principles mostly correspond to the indications contained in (M&W, 2005):
≥ Whenever an item implicitly or explicitly indicates or presumes an emotive reaction, a mood or a feeling related to the author or to others subjects mentioned by the author, use the tag <apprsl> with the attitude attribute and its relative polarity.
≥ Whenever an item indicates or presumes a judgement on people, groups or actions related implicitly/explicitly to people or groups, use the tag <apprsl> with the judgement attribute and its relative polarity.
≥ Whenever an item indicates or presumes an evaluation on abstract entities, natural phenomena, artificial processes or man-made things, use the tag <apprsl> with the appreciation attribute and its relative polarity.
≥ The polarity orientation assignment is based on the literal meanings of the evaluative item.
≥ In case of doubtful polarity orientation, it is allowed to assign the polarity looking at the previous or current phrasal context where the evaluative item appears.

Furthermore the phrasal contexts often served not only as clue for the polarity assignation, but they themselves contained evaluative sequences and thus we had to annotate chains of lexical items as single evaluative units. This aspect reflects the discursive nature of long satiric texts, so a number of semantic and syntactic criteria were needed so as to enhance the notational analysis.

Semantic criteria:
- Anytime one or more verb/noun modifiers are found, when they do not represent meaningful evaluation by themselves, they are annotated together with the part of speech that they contribute to modify.
- Any instance of evaluation conveyed by means of a multiword expression, is annotated as a single appraisal unit.
- Any instance of evaluation conveyed by means of rhetorical or figurative language, is annotated as a single appraisal unit. When possible the evaluations are embedded so as to include appraisal units into bigger evaluative unit, in order to fully capture figures of speech such as oxymora, apagoges, rhetorical questions, interjections and the like.

Syntactic Criteria:
- Without exceeding the length of the proposition, it is allowed to annotate phrases as single appraisal unit up until a clause-level, whenever they express opinions or evaluations. Additionally, for those cases where complex phrasal structures were found, we limited ourselves to the annotation of the most evaluative part within the overall sequence, so as to avoid overproduction of long annotation.
- Again, when possible, the clauses have been de-structured so that through embedding we were able to capture the evaluation on a clause-level in greater detail.
- It is allowed to annotate evaluative sequences on a clause level even beyond the punctuation marks limits. However, these annotations were very rare.
- In case of dyad/triad of items, whenever they share the same attribute and the same polarity orientation, they are annotated as single evaluative units.
- In case of more than three items in a row that share the same attribute and the same polarity orientation, they were annotated separately.

3.4 Embedded classifications

We created embedded classifications in order to account for the dependency existing between two adjacent phrases in the definition of the literal/nonliteral meaning of the sentence. We counted 220 such embeddings for Serra's texts and 146 for Oppo's texts. Consider a few examples taken from Serra's texts:

<apprsl appreciation="positive">Di scienza si vive</apprsl>, ma<apprsl appreciation="negative">di "allarmi" si muore</apprsl>,<apprsl appreciation="negative">ne ammazza più l'<apprsl attitude="negative">ansia</apprsl>del colesterolo</apprsl>. / One can live of science, one dies from alarms, more get killed by anxiety than by cholesterol.

In this case, the polarity of the embedded annotations is identical as it is for the majority of the cases in Serra's texts. But look at one of the non-identical cases:

Chiunque ci abbia provato, almeno negli ultimi due secoli,<apprsl judgement="positive">ha vinto qualche battaglia<apprsl judgement="negative">ma alla fine ha perduto la guerra</apprsl></apprsl>. / All those who have tried, at least in the last two centuries, have won some battle but at the end have lost the war.

And here below two examples taken from Oppo's texts where we see that the same technique is used:

Intanto, Berlusconi<apprsl judgement="negative"> dilaga in prima persona</apprsl>, <apprsl judgement="negative">sotto forma di una<apprsl appreciation="negative">generale regressione nazionale</apprsl></apprsl>, / In the meantime, Berlusconi floods everywhere in first person under the guise of a general national regression.

where we find in both case the same polarity – negative – but a different category, the first Judgement and the second Appreciation. Now a second example where polarity is also reversed but also category is modified:

E tutto per le<apprsl appreciation="negative">famose cene<apprsl appreciation="positive"> eleganti</apprsl> </apprsl>, ragazza alla quale,<apprsl judgement= "positive">al massimo, venivano pagati il viaggio e un<apprsl appreciation="negative"> abituccio</apprsl>di circostanza</apprsl>. / And all for the famous elegant dinners, a girl to whom at most the trip was reimbursed and a valueless courtesy dress

A further example that contains a well constructed definition of irony:

Ovvio che lo stile è<apprsl appreciation="positive"> molto diverso</apprsl>: da parte del professore<apprsl judgement="positive">nessuna volgarità</apprsl> e<apprsl judgement="positive">tanto meno barzellette<apprsl appreciation="negative"> sconce</apprsl></apprsl>;<apprsl judgement="positive" >soltanto una ironia<apprsl appreciation="positive">così sottile che <apprsl appreciation="negative">sembra la lama di un coltello<apprsl appreciation="positive">ben affilato</apprsl></apprsl></apprsl></apprsl>.
/Obviously, the style is very different: from the side of the professor, no vulgarity and not even dirty puns and jokes, just a subtle irony, so sharp that it seems the edge of a knife well sharpened

4 Results

The starting hypothesis was that both commentators were characterized by a high number of Judgements and possibily, negative ones. Then we also hypothesized that there should be an important difference between the two corpora, Oppo's being the one with the highest number. This hypothesis has been borne out by the results of the annotation as can be seen in the distribution of categories in the tables presented below. First of all general data about the annotations:

	NoSents	No.Toks	No.Annots
Oppo	**514**	**14350**	**1651**
Serra	**561**	**14641**	**1849**

Table1: Serra's annotations split by polarity

When computing general data for main categories the picture was the one in Fig. 1 below. There is a clear difference in the use of appraisal evaluative classes in our two authors: Serra seems to prefer Appreciations, Oppo on the contrary favours the use of Judgements.

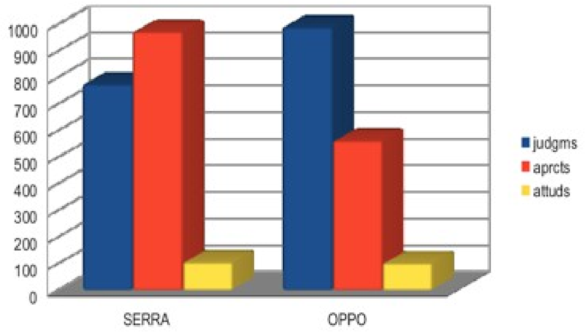


Fig1. : Total Annotations divided by main appraisal classes

However, when we collapse polarity with the categories we obtain the picture reported in the Tables below:

Serra	**JudgNega**	**JudgPos**	**ApprNega**	**ApprPos**
totals	577	216	678	385
mean	9.0	3.4	10.6	6.0
dev.stand	5.0694	2.6934	4.566	3.6274

Table2: Serra's annotations split by polarity

Oppo	**JudgNega**	**JudgPos**	**ApprNega**	**ApprPos**
totals	824	260	442	188
mean	17.2	5.4	9.2	3.9
dev.stand	5.289	3.637	3.978	2.727

Table3: Oppo's annotations split by polarity

In Table2. and 3. we report data related to the two main categories collapsed separately by polarity. As can be noted, differences in total occurrencies of Negative Judgements are very high now and Oppo has the highest. Also Positive Judgements shows a majority of cases annotated for Oppo's texts.

On the contrary, with the Appreciation class the difference is in favour of Serra, both for Negative and Positive polarity values. Standard Deviations are higher for Serra's data but this may be due to the disparity of total occurrencies, which in the case of Positive polarity is over the double and in the case of Negative polarity it is about one third higher. Eventually, we can see that Oppo's commentaries are based mainly on Judgement categories and their polarity is for the majority of the cases Negatively marked. Also Appreciation has a strong Negative bias as can be gathered from Table 3. On the contrary, Serra's commentaries are more based on Appreciation and polarity is almost identically biased.

5 Conclusions

As previous scientific literature on the topic suggests – (Taboada & Grieve, 2004), (Fletcher & Patrick 2005), (Khoo et al., 2012), (Read & Carrol, 2012), (Hall & Sheyholislami, 2013) – using (a reduced version of) the Appraisal framework proved to be a useful tool for the completion of manual annotation and for further automatic operations. Yet we were not able to represent properly some of the evaluative sequences because of the high level of complexity of the textual structure.

One of the main issue was represented by cases of linguistic cohesion realized through nominal anaphora. Additionally, if we consider that the anaphora is not always realized within a sentence, a further level of complexity is added to the representation of evaluative information: the need to take into account discourse or text level anaphora.

In any case, labelling entire sentences allowed us to capture these kinds of evaluative items, but in future research it would be convenient to find a different analytic tool beside the Appraisal framework to deal with anaphora resolution, since labelling long sentences surely led us to produce noisy data within the notational work.

References

Bosco, C., Patti, V., & Bolioli, A. (2015). Developing

Corpora for Sentiment Analysis: The Case of Irony and Senti-TUT (Extended Abstract). In Q. Yang, & M. Wooldridge (Ed.), Proc. of 24th International Joint Conference on Artificial Intelligence, IJCAI 2015 (pp. 4158 - 4162). Buenos Aires, Argentina: AAAI Press.

Burfoot, C., & Baldwin, T. (2009). Automatic satire detection: are you having a laugh? Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (pp. 161--164). Suntec, Singapore: Association for Computational Linguistics.

Carvalho, P., Sarmento, L., Silva, M., & de Oliveira, E. (2009). Clues for detecting irony in user- generated contents: oh...!! it's so easy;-). Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion (pp. 53-56). Hong Kong: ACM.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. CoNLL '10 Proceedings of the Fourteenth Conference on Computational Natural Language Learning (pp. 107-116). Stroudsburg, PA, USA: Association for Computational Linguistics.

Fletcher, J., & Patrick, J. (2005). Evaluating the utility of appraisal hierarchies as a method for sentiment classification. Proceedings of the Australasian Language Technology Workshop, (pp. 134-142). Sydney.

Gianti, A., Bosco, C., Patti, V., Bolioli, A., & Di Caro, L. (2012). Annotating Irony in a Novel Italian Corpus for Sentiment Analysis. In L. Devillers, B. Schuller, A. Batliner, P. Rosso, E. Douglas- Cowie, R. Cowie, & C. Pelachaud (Ed.), Proc. of the 4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (ES3@LREC'12) (pp. 1-7). Istanbul, Turkey: ELRA.

Hall, C., & Sheyholislami, J. (2013). Using Appraisal Theory to Understand Rater Values: An Examination of Rater Comments on ESL Test Essays. The Journal of Writing Assesement, Volume 6 (1) .

Hernandez Farias, D., Sulis, E., Patti, V., Ruffo, G., & Bosco, C. (2015). ValenTo: Sentiment Analysis of Figurative Language Tweets with Irony and Sarcasm. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015) (pp. 694-698). Denver, Colorado, USA: Association for Computational Linguistics.

Khoo, C., Nourbakhsh, A., & Na, J. (2012). Sentiment analysis of online news text: A case study of appraisal theory. Online Information Review 36(6).

Martin, J., & White, P. R. (2005). Language of Evaluation, Appraisal in English. London & New York: Palgrave Macmillan.

Owais, S., Nafis, T., & Khanna, S. (2015). An Improved Method for Detection of Satire from User- Generated Content. International Journal of Computer Science and Information Technologies, Vol. 6 (3), pp. 2084-2088.

Ozdemir, C., & Bergler, S. (2015). CLaC-SentiPipe: SemEval2015 Subtasks 10 B,E, and Task 11. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), (pp. 479-485). Denver, Colorado: Association for Computational Linguistics.

Read, J., & Carrol, J. (2012). Annotating expressions of Appraisal in English. Language Resources and Evaluation, Volume 46, 421-447.

Reyes, A., & Rosso, P. (2011). Mining subjective knowledge from customer reviews: a specific case of irony detection. WASSA '11 Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (pp. 118-124). Stroudsburg, PA, USA: Association for Computational Linguistics.

Taboada, M., & Grieve, J. (2004). Analyzing appraisal automatically. In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (pp. 158-161). AAAI Press.

Wang, P.-Y. A. (2013). #Irony or #Sarcasm A quantitative and qualitative study based on Twitter . Proc. 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27). Taipei, Taiwan.