Detecting Satire in Italian Political Commentaries

Rodolfo Delmonte, Michele Stingo

Department of Language Studies & Department of Computer Science
Ca' Bembo 1075 – Ca' Foscari University – 30123 Venezia (Italy)
delmont@unive.it, stingomichele@gmail.com

Abstract: This paper presents computational work to detect satire/sarcasm in long commentaries on Italian politics. It uses the lexica extracted from the manual annotation based on Appraisal Theory, of some 30K word texts. The underlying hypothesis is that using this framework it is possible to precisely pinpoint ironic content through the deep semantic analysis of evaluative judgement and appreciation. The paper presents the manual annotation phase realized on 112 texts by two well-known Italian journalists. After a first experimentation phase based on the lexica extracted from the xml output files, we proceeded to retag lexical entries dividing them up into two subclasses: figurative and literal meaning. Finally more fine-grained Appraisal features have been derived and more experiments have been carried out and compared to results obtained by a lean sentiment analysis. The final output is produced from held out texts to verify the usefulness of the lexica and the Appraisal theory in detecting ironic content.

Keywords: Semantic Annotation, Pragmatic Annotation, Appraisal Theory, Automatic Irony Detection, Literal vs Nonliteral Language

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 man1. 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, 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

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

highly cutting and caustic writing, Serra is less so. Both are humorous, both use sophisticated rhetorical devices 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. In order to focus on the specific features connotating political satire, manual annotation has been carried out on the 112 texts using at first a reduced version of the Appraisal Framework [1]. Following the annotation activity, a typological classification has been produced for all the entries contained in the automatically collected lexica (one for each author) composed of the annotated items/phrases (see also [2]; [3]; [4]). The classification has been carried out using three linguistic traits: namely idiomatic, metaphorical – these two being figurative uses - and none for the rest. This has been done in order to set apart figurative uses the author chose for a specific item/phrase from nonfigurative ones(see [5]). All the annotations have been done by the second author and counterchecked by the first author.

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 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 [6]. These latter texts are long texts, from 200 to 400 words long and do not compare with previous experiments.

In the majority of the cases, the other common approach used to detect irony 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 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 it is 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 (see [7]; [8]).

By adopting the 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 & Martin [1] (hence M&W) makes available an extended number of semantically and pragmatically motivated annotation schemes that can be applied to any text. 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. In the book by M&W there's a neat distinction between these three types of voices: ‘reporter voice’, ‘correspondent voice’ and ‘commentator voice’. Since 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, this is our option and our first 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 etc.2 So 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. The second hypothesis is that both commentators are 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.

There are three possible strategies writers can use to produce humorous effects: the superiority presumption, [9], relief presumption ([10]; [11]) and incongruity presumption [12]. The first speculative contribution was proposed by [13], further revised by [14], [15] and [17] as a general theory of verbal humour. The hypothesis we will now formulated is based on the contribution that our new annotation traits can bring to the detection task. The superiority presumption assumes that the object of the ironic process be sanctioned, so here we refer to the Judgement Social Sanction/Esteem Negative classified items of our lexicon. The relief presumption could be based again on the use of the previous features in addition to Positively marked features. The relief is given by laughter, i.e. by humorous meaning which generates positive energy. This physic energy is built anytime we need to suppress negative feelings in our psyche and every time we release this energy, by virtue of jokes related to taboos and cultural values induced by society (namely when we suppress the mental censorship mechanism), we experience laughter and a psychological benefit is reached. This may be obtained by the use of figurative language, i.e. the use of a word/phrase/expression with the opposite meaning it usually has. Finally the incongruity presumption can be again achieved by combining Positive and Negative Judgement/Appreciation features with strong socially related nuances. As to the satiric discourse we rather deem the incongruity presumption [17] to be more adequate to explain the humurous mechanism. In particular, at the heart of this approach there is an opposition between two dimensions, and in order for a text to be processed as humorous – in addition to the opposition feature, the dimensions have to share a common part, so that it is possible a shift from one dimension to another. First of all we present general data about the annotations:

---

2 As reported in M&W p.52.

Table 1. General data about the corpus

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

When we collapse polarity in the two main categories we obtain the picture reported in Table 2. below. As can be noted, differences in total occurrencies of Negative Judgements are very high and Oppo has the highest. Also Positive Judgements shows a majority of cases annotated for Oppo's texts.

Table 2. Annotations split by polarity

Writers	JudgNegat	JudgPost	ApprNegat	ApprPost
Serra	**577**	**216**	**678**	**385**
Oppo	**824**	**260**	**442**	**188**

On the contrary, in the Appreciation class differences are all in favour of Serra, both for Negative and Positive polarity values. Finally, we can see that Oppo's commentaries are based mainly on Judgement categories and their polarity is for the majority of cases Negatively marked. Also Appreciation has a strong Negative bias as can be gathered from Table 2. On the contrary, Serra's commentaries are more based on Appreciation and polarity is almost identically biased.

3 Experiments to validate Nonliteral language

The first approach to better understand the semantic/pragmatic features of our texts has been that of automatically deriving a lexicon from the annotated texts and then proceed to some further investigation. We extracted some 3500 annotations overall, one third has been identified as belonging to figurative language, that is idiomatic expressions and similes, metaphors and metonymies. The remaining 2/3, i.e. 2300 has been assigned to the neutral category NONE. However this classification was not satisfactory and so we started detecting literal from nonliteral expressions at first using automatic procedures. We produced a lexicon of Appraisal Categories related to lexical entries as they are listed in the book by M&W. We came up with some 500 entries which we then used to retag the 3500 lexical entries. We wrote a simple script that took each lexical entry, produced the lemmata for every semantic word, and then tried to match it with the Appraisal lexicon. The results have been very poor and we only managed to cover 10% of all entries. So we decided to manually retag the 2300 neutral entries dividing them up into three subcategories: a. Literal meaning – whenever the appraisal category coincided with the literal meaning of the entry; b. Nonliteral meaning – whenever the appraisal category was not related to the literal meaning associated to the entry; c. Semantically hard to compute literal meaning – whenever the meaning of the entry required some compositional analysis to recover the literal meaning and there was not a one-to-one correspondence between the entry and the appraisal category. We ended up by reclassifying 16% of the 2300 None as belonging to category b, i.e. 244 new entries as nonliteral; and another 22.96% as

semantically hard, i.e. 528 entries. The new organization of the two lexica is now as follows:

Table 3. Semantic subdivision of lexical entries

	None	Idiomatic	Figurative	Nonliteral	Sem_hard	Totals
Oppo	711	201	422	153	187	**1674**
Serra	816	143	449	91	341	**1840**
Totals	1527	344	871	244	528	**3514**

Now proportions are reversed and literal language covers only 43% of all lexical entries. As to appraisal classification, lexicon values repeat the opposition we found in counting annotations in texts: Oppo's lexicon has a majority of Negative Judgements, Serra's lexicon has a majority of Negative Appreciations. Serra's Positive Appreciations are almost the double of Oppo's, whereas Positive Judgements are comparable:

Table 4. Subdivision of lexica with fine grained Judgement subclasses

	Judgmt. Negat.	**Judgmt. Posit.**	**Apprec. Negat.**	**Apprec. Posit.**	**Negative Esteem**	**Negative Sanction**
Oppo	***742***	275	396	181	375	363
Serra	554	214	***618***	343	275	274
Totals	1296	489	1014	524	650	637

In order to comply with our interpretation of commentators' role, we expected then to have an internal subdivision of Negative Judgement showing a high percentage of Negative Sanction. So we proceeded in the reclassification of all Judgement lexical entries into the new two subcategories, Sanction and Esteem. All these values are referred to the types listed in the new lexica and they only represent potential new automatic annotations which however need to be tested on the corpus. The subdivision of Negative Judgements between Sanction and Esteeem is strongly in favour of Oppo with slight differences in distribution between the two classes.

3.1 Computing Nonliteral language

We will now delve into the experimental part of the work which is strictly related to the fine-grained classification and the subdivision of lexical entries into Literal and Figurative language, which should allow better performances as far as irony detection is concerned (see [18]; [19]; [20]). We then set up our algorithm for irony detection with the following instructions:

```
-  SEARCH inside a sentence all annotations
       - of type Judgement_Sanction_Negative
       - or Judgement_Esteem_Negative
       - or Appreciation_Negative
```

```
        - or Emotion_Negative
- ELSE none found
   END
        AND
        - together with annotations with the
                opposite polarity, Positive
                EITHER
                - belonging to LITERAL type
                - belonging to NONLITERAL type
                output=TRUE
        ELSE
        output=FALSE
        END
```

where True indicates a possible condition for irony detection and False the opposite. The combination of all the different parameters has given as a result six different outputs which confirm all the hypothesis we put forward in the previous section.

In Fig.1 below main differences can be found when Figurative language is used and Negative features are involved. In particular When SocialSanction_Negative and SocialEsteem_Negative Figurative annotations are used together with any Negative annotations in the same sentence, Oppo's texts show a great jump up when compared to Serra's - that's the orange cylinders. On the contrary, when Only Positive Figurative annotations are used together with any Positive annotation in the same sentence, we see that Serra's values are higher, light green. Using all Negatives Figurative annotations with Negatives again favours Oppo's texts - light blue, whereas Negatives Figurative with Positive annotations favours Serra's - light red.

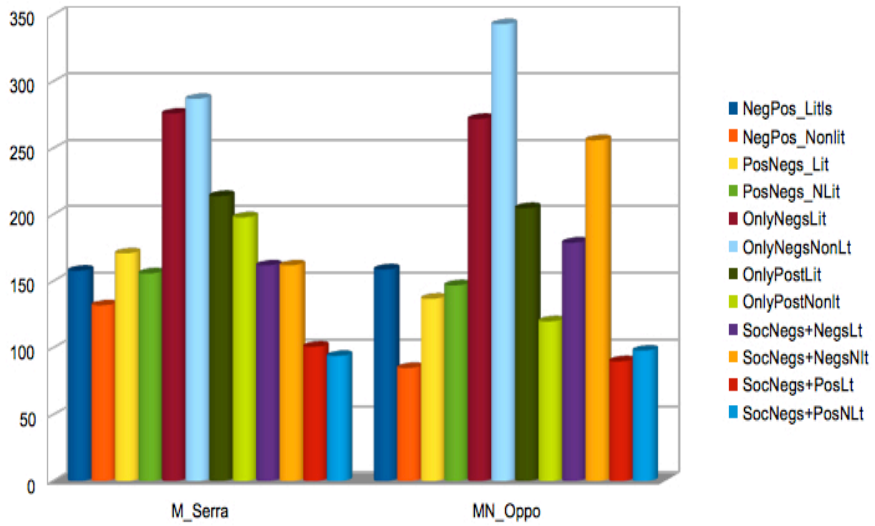

Fig 1. 12 experiments with new lexica

This distribution of the data confirms our previous hypothesis: Oppo's text are more

close to sarcasm, while Serra's text are less so and just satiric. Oppo's appraisal configuration for best irony detection requires the presence of Negatively marked Judgements, socially biased, and also with a preference for literal meaning. On the contrary, Serra's texts are characterized by the preference of purely Positively marked words/phrases with a strong bias for nonliterality.

3.2 Experimenting with new texts

We now report results obtained with held out texts for the two journalists. We ran our automatic annotation algorithm based on the lexica created from the manual annotation and further modified, on 20 texts, ten for each author, to verify whether the setup we derived from our previous analysis is directly applicable to any new text or not. Oppo's texts contain 118 sentences, Serra's texts contain 96 sentences. Oppo's texts have been automatically assigned 100 annotations; Serra's texts, only 66. From Fig.2 below we have some confirmations but also some new data.

The experiments have been organized using different setups both for the lexica and for the irony detection. At first, we used separate lexica from each corpus, then joined them into one single lexicon made of 3514 entries. We also selected different strategies for irony detection – which we mark with TRUE - on the basis of our previous computation. We used all negatives – this strategy favouring Oppo - choosing those with literal meaning in combination with all negatives. Then we selected all positives - this strategy favouring Serra - this time choosing those with nonliteral meaning in combination with positives. As can be clearly seen, best irony detection results have been obtained when lexica have been joined together. However, there are remarkable differences. When we use specific lexica we see important improvements in the number of annotations, in particular in the case of Serra's texts. With Oppo's texts, we get more TRUE detection cases when Serra's lexicon is used compared with Oppo's lexicon. Remember that when we use Serra's lexicon, we also modify our strategy for irony detection to Positive+Nonliteral. Generally speaking, however, it is always Oppo's texts and lexica that produce the highest number of Judgements Negative. Strangely enough, Oppo's texts are also characterized by a great number of Judgement Positive, in fact the highest number. Then, contrary to expectations, TRUE decisions in Serra's texts are determined by the Positive strategy which obtains higher results than the Negative one. In the case of Oppo, we see a slightly higher number of TRUE when the Positive strategy is applied.

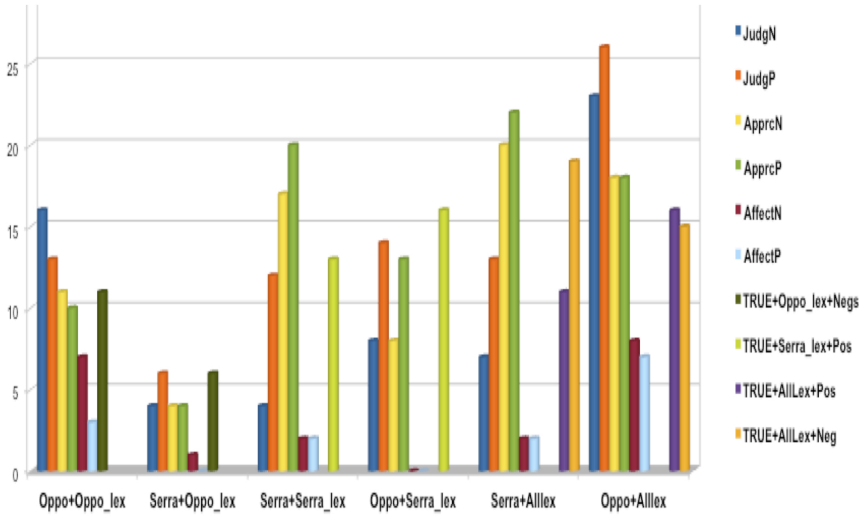

Fig.2 Evaluation of satire detection algorithm with held out texts

So it would seem that this experiment does only partially confirm our hypotheses. However we need to consider that the lexica produced from previous manual analysis do not cover completely the new texts in that the number of automatic annotations obtained is only a small percentage: 100/118, 66/96, i.e. not even one per sentence. On the contrary, in the previous manual work, we had an average of 3.2 annotation per sentence.

To improve recall, we then collected all lexical items contained in the book by M&W and we used them with the lexicons with shallow analysis as before, and we labeled them all as having literal meaning in association with each appraisal category. Results are reported in Fig.3 below. First of all, number of automatic annotations now increased to 103 for Serra, and 146 for Oppo: still not comparable to manual annotations but certainly much better than before - we are now halfway from the target of 3.2 annotations per sentence. Coming now to the new automatic classification of test texts, Appraisal categories are divided up as follows, where N=negative and P=positive, Af=affect, Ap=appreciation, Jg=judgement, and Sct=Sanction, Est=Esteem:

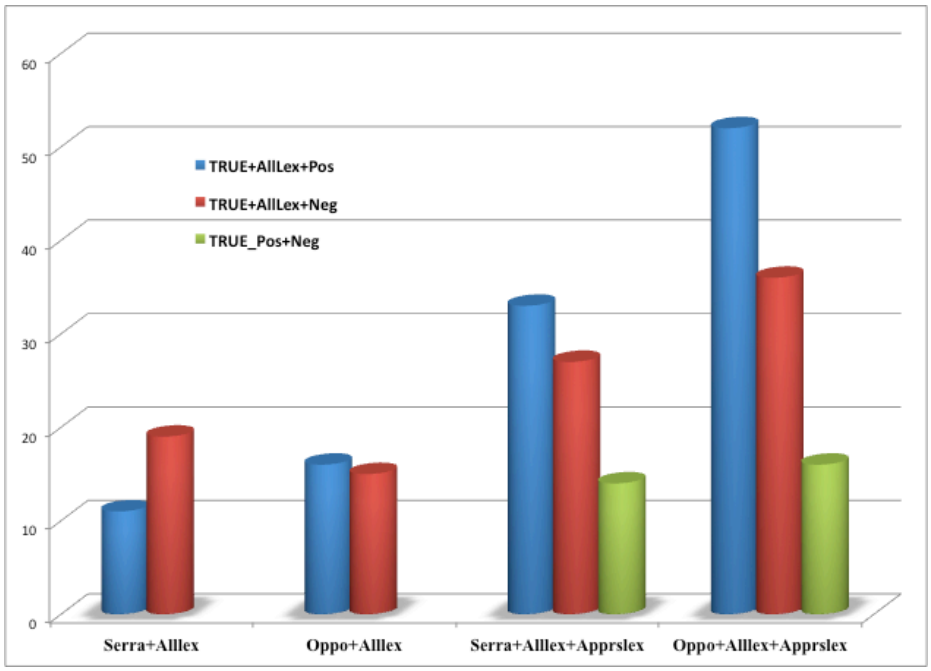

Fig. 3. Irony detection using augmented lexica

In this figure we present final results for irony detection using appraisal theory by simply checking three possible combinations of polarity values: only Positives, marked PP, only Negatives, marked NN and then Positives and Negatives marked NP.

As can be noticed, best results are NN combinations and as before, they are higher in Oppo's texts. Then come the PP combinations and finally the NP which are however much lower. In this case, Oppo's True cases are over 50, which when compared to number of sentences makes almost 50% of them. In the case of Serra's True they only reach 36 sentences, which is a much lower percentage when compared to number of sentences, only 37.5%.

Table 5. Classification of test texts into Appraisal categories

Authors	Af N	Af P	Ap N	Ap P	Jg N	Jg P	Sct.	Est.	Sct N	Sct. P	Est. N	Est. P
Oppo	9	11	21	27	27	40	15	22	7	8	20	2
Serra	3	3	21	27	12	20	9	12	3	6	9	3
Total	12	14	42	54	39	60	24	34	10	14	29	5

Negative Esteem seems to be used a lot more than Sanction which is however used in the opposite manner, more Positive evaluations than negative ones. Here we must remind that we have decided to treat all new lexical entries derived from M&W as

semantically literal, but we have seen from previous analysis that this may only be true for 40% of all data.

5 Conclusion

We have shown that by using the framework of the Appraisal theory it is possible to highlight features of ironic texts and to use these features to detect satire/sarcasm automatically. The results obtained are still work in progress and we are continuing the manual annotation work to include more fine-grained distinctions. We have been able to show that Oppo and Serra stylistic devices are different in a significant manner, and that this difference is clearly borne out by the categories derived from Appraisal theory. In particular, we have succeeded in showing how Oppo's texts constitute more cutting political comments than Serra's text, speaking in general terms. This stylistic characteristic is strictly derivable from and related to the use in their comments of more Judgement rather than Appraisal lexical material for Oppo, while the opposite applies to Serra.

Future work will be devoted to increase the number of experiments. In particular, we want to try to show correlations existing between automatic and manual annotations, using test texts where however manual verification is needed to check how many nonliteral uses have been done with the specific Attitude related categories. Annotating texts using M&W theoretical framework is hard and it requires specific linguistic training. In addition, classifying political commentaries requires a lot of world knowledge due to the habit of commentators to refer to real life events and use them as a comparison to comment on the current political issue. This aspect could be covered by accessing LOD data and by using ground truth description to match satiric distorted ones. Another important element that has not yet been part of the automatic evaluation is constituted by the need to corefer events and people, again a difficult task to accomplish.

References

1. Martin, J., & White, P. R. (2005). Language of Evaluation, Appraisal in English. London & New York: Palgrave Macmillan.
2. 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.
3. 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.
4. Khoo, C., Nourbakhsh, A., & Na, J. (2012). Sentiment analysis of online news text: A case study of appraisal theory. Online Information Review 36(6).
5. Sarmento, L., Carvalho, P., Silva, M., & de Oliveira, E. (2009). Automatic creation of a reference corpus for political opinion mining in user-generated content. Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion (pp. 29-

36). Hong Kong: ACM.

6. 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.
7. 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.
8. Özdemir, 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.
9. Gruner, C. (1997). The Game of Humor: A Comprehensive Theory of Why We Laugh. New Brunswick, NJ: Transaction Publishers.
10. Freud, S. (1905). Der Witz und seine Beziehung zum Unbewußten. Leipzig; Vienna: Franz Deuticke.
11. Minsky, M. (1981). Jokes and their Relation to the Cognitive Unconscious. In L. Vaina, & J. Hintikka, Cognitive Constraints on Communication: Representations and Processes. Reidel.
12. Koestler, A. (1964). The Act of Creation. London: Hutchinson & Co.
13. Raskin, V. (1985). Semantic Mechanisms of Humor. Dordrecht - Boston - Lancaster: D. Reidel.
14. Attardo, S., & Raskin, V. (1991). Script theory revis(it)ed: joke similarity. HUMOR: International Journal of Humor Research, 4 ((3/4)), 293–347.
15. Attardo, S. (1994). *Linguistic Theories of Humor.* Berlin – New York: Mouton de Gruyter.
16. Attardo, S. (1997). The semantic foundations of cognitive theories of humor. *10*(4), 395–420.
17. 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.
18. Birke, J., & Sarkar, A. (2007). Active learning for the identification of nonliteral language. FigLanguages '07 Proceedings of the Workshop on Computational Approaches to Figurative Language (pp. 21-28). Rochester, New York: Omnipress Inc.
19. Turney, D., Neuman, Y., Assaf, D., & Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 680-690). Edinburgh, UK: Association for Computational Linguistics.
20. 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.