A Corpus of Scientific Biomedical Texts Spanning over 168 years annotated for Uncertainty

***Bongelli, R., *Canestrari, C., *Riccioni, I., *Zuczkowski, A., *Buldorini, C., **Pietrobon, R., ***Lavelli, A., ***Magnini, B.**

* Università di Macerata, Dipartimento di Scienze dell'educazione e della formazione, P. le L. Bertelli 1, I-62100 Macerata (Italy)
{ramona.bongelli, c.canestrari, i.riccioni, zuko, cinzia.buldorini}@unimc.it

** Duke University, Department of Surgery, 2500 North Pavilion, Durham, NC 27710 (USA)
rpietro@duke.edu

***Fondazione Bruno Kessler, Via Sommarive 18, I-38123 Trento (Italy)
{lavelli, magnini}@fbk.eu

Abstract

Uncertainty language permeates biomedical research and is fundamental for the computer interpretation of unstructured text. And yet, a coherent, cognitive-based theory to interpret Uncertainty language and guide Natural Language Processing is, to our knowledge, non-existing. The aim of our project was therefore to detect and annotate Uncertainty markers – which play a significant role in building knowledge or beliefs in readers' minds – in a biomedical research corpus. Our corpus includes 80 manually annotated articles from the British Medical Journal randomly sampled from a 168-year period. Uncertainty markers have been classified according to a theoretical framework based on a combined linguistic and cognitive theory. The corpus was manually annotated according to such principles. We performed preliminary experiments to assess the manually annotated corpus and establish a baseline for the automatic detection of Uncertainty markers. The results of the experiments show that most of the Uncertainty markers can be recognized with good accuracy.

Keywords: Uncertainty markers; Biomedical texts; Evaluation

1. Introduction

The Certainty or Uncertainty of information communicated by scientific writers plays a significant role in establishing readers' knowledge, beliefs, and subsequent actions.

The Uncertainty, that is the focus of our study, is identified through a series of linguistic markers and can also assist in the computational interpretation of unstructured, free text data.

The detection of Certainty/Uncertainty markers and their linguistic scope in Natural Language Processing (NLP) have been receiving increasing attention in the NLP community. Distinguishing certain (= factual) and uncertain (= speculative) information in texts is of crucial importance in information extraction (IE).

Although there are a few analysis of corpora annotated for Uncertainty language (Crompton 1997, Hyland 1994, 1998, Salager-Meyer 1994), to our knowledge these annotations are not based on a comprehensive cognitive and linguistic theory of Certainty and Uncertainty communication. In addition, they tend to be small in their number of full-text scientific articles, making the validation of computational algorithms challenging. As an example of both issues, the BioScope corpus (Vincze et al. 2008)[1], is composed of only nine full text articles and has not been analyzed according to an explicit cognitive and linguistic theory of Certainty and Uncertainty communication.

In contrast with the above mentioned works and others, for example, the BioNLP'09 Shared Task on Event Extraction (Kim et al., 2009) and the CoNLL 2010 shared task (Farkas et al., 2010), our study, on the one hand, is set in a theoretical linguistic frame of reference (Petöfi's text theory 1973, 1980, 1981, 2004) and takes into account the international linguistic literature on evidentiality and epistemicity (topics strictly related to Certainty and Uncertainty); on the other hand, it is grounded on a cognitive and linguistic theory of Certainty and Uncertainty communication (Bongelli, Zuczkowski 2008) that provides abundant basis for a new classification. Such a theory could potentially allow for more detailed annotation guidelines, including lexical and morphosyntactic markers[2].

In order to fulfill the above mentioned gaps in the literature, the objective of the work reported in this article

[1] For a comprehensive literature review, see Agarwal, Yu 2010.

[2] Lexical markers are made of one or more words, e.g. in *perhaps*, in *my opinion* etc., while morphosyntactic (or grammatical) markers are those referring to sentence structures (declarative, interrogative, hypothetical and so on), verbal tenses and moods.

was therefore to create a corpus of 80 biomedical scientific full texts spanning over a period of more than 160 years, classified according to a cognitively and linguistically based theory. We have assessed the quality of the manual annotation of Uncertainty markers calculating the inter-annotator agreement. Moreover, we have validated the annotation performing experiments using Yamcha, a text chunker, to train a classifier for recognizing the Uncertainty markers.

2. Theoretical background

Our theoretical linguistic frame of reference while analyzing the 80 articles in our corpus is formed by the previous work by J.S. Petöfi's (1973, 1980, 1981, 2004). First, Petofi's *Theory of Text Structure and World Structure* provides us with a model of deep structure. This model proposed that texts are structured as *Atomic Texts*, ultimately constituted by three types of propositions: descriptive, world-constitutive, and performative. These, according to Petöfi, constitute the basic structure of any communication. Descriptive propositions represent the pieces of information communicated by writers. World-constitutive propositions represent writers' evidential and epistemic attitudes in relation to the content of the descriptive proposition. Finally, performative propositions specify the particular illocutionary act (Austin 1962) performed by writers. The world-constitutive proposition is the one that interests us most here, since the topic of Certainty/Uncertainty in communication is related, more or less directly, to what in linguistic literature is called *epistemicity* and *evidentiality*.

2.1 Certainty and Uncertainty, Epistemicity and Evidentiality

As for *epistemicity*, which refers to linguistic markers such as, for example, the adverbs *sure, undoubtedly, certainly, perhaps, probably*…, in literature there are a few slightly different definitions of it: with this term some authors refer to the *speaker's attitude* regarding the *reliability* of the information (e.g. Dendale and Tasmowski 2001, González 2005), others to the *judgment of the likelihood* of the proposition (e.g. Nuyts 2001b, Plungian 2001, Cornillie 2007), others to the *commitment to the truth* of the message (e.g. Sanders and Spooren 1996, De Haan 1999, González 2005).

According to our theoretical point of view, the above mentioned definitions can all be re-conceptualized in terms of the labels "Certainty" and "Uncertainty", in the sense that *at the communicative level*, i.e. when communication occurs, writer's *attitude* (regarding the reliability of the information) or *judgment* (of the likelihood of the proposition) or *commitment* (to the truth of the message) can only be one of Certainty or Uncertainty. When I say, for example, *Certainly Peter is at home*, I communicate that it is certain for me, i.e. I am certain, that the piece of information p (= *Peter is at home*) is true, i. e. I'm saying that I evaluate p as true. *Uncertainty* means that, when I tell you, for example, *Perhaps Peter is at home*, I am saying that I do not know whether p is true or false, therefore I communicate p as uncertain, i.e. I tell you that I am not certain towards the truth of p. In this sense, the relationship between epistemicity and Certainty/Uncertainty is direct.

With the term *evidentiality*, scholars (see for example, Van Der Auwera and Plungian 1998, De Haan 1999, Nuyts 2001a, 2001b, Plungian 2001, Cornillie 2007, Papafragou et al. 2007) usually refer to the linguistic markers that reveal the *source of information* communicated by writers, namely how they gain access to that information.

According to our psychological approach to the study of the relationship between language and mental processes, this access can only be *perceptual* or *cognitive*, since human beings get information only through perception and cognition. The first term refers to the five senses and proprioception; the second one refers to thought, memory, imagination etc.

If I say, for example, *I see that Peter is at home*, I explicitly communicate the information source (*I see*); though in the utterance there is no epistemic marker, the evidential (perceptual) verb *I see* is enough to indirectly communicate Certainty. In this sense, the relationship between evidentiality and Certainty is indirect. It would be the same if in the above example, instead of *I see ,* there were an evidential (cognitive) verb such as *I remember*.

2.2 The relationship between evidentiality and epistemicity

The relationship between evidentiality and epistemicity is highly debated in literature and is considered to be of three main types (Dendale and Tasmowski 2001, González 2005, Cornillie 2007): disjunction (De Haan 1999, Aikhenvald 2003, 2004); inclusion (Givón 1982, Chafe 1986, Palmer 1986, Willett 1988, Papafragou 2000, Mushin 2001, Ifantidou 2001) and overlap (Van Der Auwera and Plungian 1998, Plungian 2001).

According to the results of our previous studies on written and oral corpora (Bongelli and Zuczkowski 2008; Zuczkowski, Bongelli, Riccioni 2011; Riccioni, Bongelli, Zuczkowski forthcoming; Bongelli, Riccioni, Zuczkowski submitted), evidentiality and epistemicity seem to be two sides of the same coin: not only epistemic but also evidential markers communicate Certainty and Uncertainty.

2.3 Known/Certain and Believed/Uncertain Theory

In our previous studies we found that, normally, when a piece of information is communicated as *certain* (epistemicity) by writer, at the same time it is also communicated as *known* (evidentiality) to her/him (and vice versa). On the contrary, when a piece of information is communicated as *uncertain* (epistemicity), at the same time it is also communicated as *believed* (evidentiality) by her/him (and vice versa). *Believed* is a general term we chose to refer to what, in our definition, includes not only beliefs but also opinions, impressions, suppositions,

assumptions, conjectures etc., i.e. briefly all that is not communicated as known, as knowledge.
In other words, it is as if there were two different "Territories of Information", to use Kamio's (1994, 1997) terminology: a piece of information is communicated as belonging either to the (epistemic) territory of the Certainty, which overlaps the (evidential) territory of the Known (and vice versa), or to the (epistemic) territory of the Uncertainty, which overlaps the (evidential) territory of the Believed (and vice versa).
This means that the diverse and numerous (lexical and morphosyntactic) evidential and epistemic markers can be led back to two main macro-markers, each of them has two faces, one evidential and the other epistemic: *I know / I am certain; I believe / I am uncertain.*

2.3.1. Markers of the Known/Certainty

The *Known/Certainty*, at the communicative level, is all that writer says s/he perceives, remembers and knows, in a broad sense.
In English the main lexical markers of the Known /Certainty are the following:

- evidential verbs in the first person singular of the simple present tense such as the verbs *I know, I remember, I see* …or in the third person singular or plural …*it reminds me, they recall me*…;
- epistemic adverbs, such as *undoubtedly, surely, certainly…*, adjectives *sure…*, verbal expressions such as *I am sure; I have no doubt; I am convinced*….

Through morphosyntactic markers, the Known /Certainty is normally communicated by:

- sentences in the present, past and future indicative with no lexical evidential or epistemic marker (neither of the Known nor of the Believed: for example *Yesterday Peter was at home; Peter will be at home tomorrow*).

2.3.2. Markers of the Believed/Uncertainty

The *Believed/Uncertainty* is very differentiated in its internal structure, inasmuch as it includes different cognitive processes such as: *to have the impression, to be of the opinion, to suppose, to doubt* etc. At the communicative level, opinions, suppositions, impressions etc. have in common that they do not communicate Certainty, but Uncertainty, Possibility and Hedging in variable degrees; they do not communicate what writer knows, but what s/he believes, in the broadest sense.
In English, the Believed/Uncertainty is normally communicated by the following lexical markers:

- epistemic verbs like *I suppose, I think, I believe, I imagine, I doubt, it seems to me…* and also through
- verbal epistemic expressions like *it is probable, it is possible, I am not certain, I am uncertain, I am not sure…*, adverbs like *probably, perhaps…*, adjectives like *likely, possible…*;
- modal verbs like *can* and *must* when used in an epistemic sense.

Through morphosyntactic markers, the Believed/Uncertainty is normally communicated by:

- modal verbs in conditional and subjunctive moods;
- if clauses (we did not take into account the zero conditional - simple present in the protasis as well as simple present in apodosis - since "if" can be paraphrased by a temporal conjunction);
- epistemic future (i.e. the conjectural use of future paraphrasable with the believed expressions).

Normally, lexical and morphosyntactic markers interact; however, while the latter are always present in a text, the former can be absent.

2.3.3. Writer's Uncertainty

In a text, Uncertainty markers can refer either to the author's Uncertainty or to somebody else's Uncertainty. Both types of Uncertainty can refer to the present or past or future. Only the Uncertainty markers expressed by the author in the present, i.e. when s/he is writing her/his paper, were taken into account and tagged, since only in this case the author is communicating her/his own current (= present) Uncertainty towards the piece of information *p* that s/he is giving readers (for example, *I doubt that p*). According to our theoretical point of view, the other cases (*I doubted/ I will doubt/ John doubts/ John doubted /John will doubt that p*), were not taken in to account. In the same way, an adjective such as "possible" was not tagged when it was used in a paste tense sentence (e.g. *The adhesions over the front of the bladder were also broken down, so that it was possible to pass a rubber tube across and flush out the left iliac fossa*) or when it was referred to anyone else point of view (e.g. *It is possible to dr. Johns that …*).
So, for the purpose of the present study, we focus specifically on the detection of writers' Uncertainty markers.

3. Biouncertainty corpus

3.1 British Medical Journal corpus and sampling structure

We performed a linguistic analysis of 80 papers from the British Medical Journal available from PubMed Central (http://www.ncbi.nlm.nih.gov/pmc/journals/3/, last accessed March 2011) and randomly selected from four distinct time periods (1840-1880, 1881-1920, 1921-1960, and 1961-2007). The choice of four period categories was based both on statistical considerations, namely that we could have the ability to detect trends over time, but also on a theoretical and historical basis. Specifically, the 1840-1880 was chosen given that this represents a period for medicine where physiology and experimentation were still being introduced in European medicine, possibly affecting the degree of Uncertainty language; the 1881-1920 period represented the turn of the century, with an overall excitement about the apparent unlimited ability of science in predicting facts and being certain; the 1921-1960 period represent a time with sequential world

wars along with their direct impact on research; finally, the 1961-2007 period represents our current language patterns.

Our random sampling strategy followed a structured sequence, including (1) random selection of journal volumes, (2) random selection of pages within the specific volume, and (3) when two or more articles were present on the same page, then their choice was once again made using a random number. Random number generation was made in real time by one of the authors during article sampling using the R language (http://www.r-project.org/, last accessed February 2011).

Our corpus of 80 papers consists of approximately 225,000 tokens.

3.2 Corpus availability

The corpus will be shortly available at a dedicated Web site at https://sites.google.com/site/biouncertainty/corpus (last accessed March, 2012) under a Creative Commons License (http://europe.creativecommons.org/, last accessed March, 2012).

4. Manual annotation

In the first step of the analysis, 10 language specialists from the Research Center for Psychology of Communication at the University of Macerata, all fluent in English, analyzed four full text articles using the same reading-grid but also their linguistic skills. In the second step, 80 full text papers were analyzed in pairs by the same 10 language specialists, so that consensus could be achieved when questions arose. A final analysis was conducted, now involving 5 of the 10 previous linguists to ensure that all markers had been appropriately classified. In this phase, linguists used the WordSmith software in order to improve the quantitative analysis, i.e. to check that all occurrences of Uncertainty markers had been detected. All analysis results were placed in a cloud environment so that it could be accessible by all researchers in real time.

4.1 Inter-annotators agreement

Prior to the initiation of this study, all language specialists gathered to discuss the Uncertainty classification in detail to ensure that they all agreed on each of its components. Each sentence in the full text of each article was evaluated by two independent language specialists. Observer agreement was then evaluated through non-weighted kappa coefficients, measuring overall agreement (0.89).

4.2 Results

A total of 2,758 Uncertainty markers were manually annotated in the corpus. Specifically, there were:

- 722 *modal verbs in conditional and subjunctive moods*;
- 235 *if clauses*;
- 896 *modal verbs*;
- 417 *uncertainty non verbs*;
- 388 *verbs.*

The occurrences of *epistemic future* were numerically low and their epistemic interpretation was often controversial. For these reasons, they were excluded from the analysis.

5. Automatic annotation

In order to validate the usefulness of the resource, we have performed experiments using YamCha[3], an open source text chunker, to train a classifier for recognizing the Uncertainty markers. YamCha is a generic and customizable tool applied in different NLP tasks, such as POS tagging, Named Entity Recognition, base NP chunking, and Text Chunking. It exploits Support Vector Machines (SVMs).

The documents have been processed using TreeTagger[4], a language-independent PoS tagger and the corpus annotation converted to IOB format.

5.1 Experiments and Results

Experiments were performed adopting a 10-fold cross validation experimental setup with a document-level split. To classify the current token, we used features such as token, lemma and PoS tag contained in a window of tokens before and after the token. Within this framework, we could also explore the tag of previous tokens. We adopted a one-vs-all classification strategy. Table 1 reports the results obtained using a window of 2 tokens before and 2 tokens after the current token, in addition to the tag of the 2 previous tokens.

UNCERTAINTY MARKERS	Precision	Recall	F_1
MODAL_IN_COND/SUB	92.74	94.29	93.51
IF	53.90	35.62	42.89
MODALVERB	95.07	95.71	95.39
NONVERB	83.02	59.00	68.98
VERB	84.41	64.68	73.24
Overall	88.66	78.90	83.49

Table 1. Results of 10-fold cross validation using a window of 2 tokens before and after the current token.

Extending the window around the token to be classified (3 tokens before and 3 tokens after the current token, plus the tag of the 3 previous tokens), results were worse (Table 2), with a general increase in precision but a drop in recall.

UNCERTAINTY MARKERS	Precision	Recall	F_1
MODAL_IN_COND/SUB	93.12	94.29	93.70
IF	54.05	55.75	34.88
MODALVERB	94.41	95.37	94.89
NONVERB	89.17	47.32	61.83
VERB	90.04	56.36	69.33

[3] http://chasen.org/~taku/software/yamcha/

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Overall	90.89	74.56	81.92

Table 2. Results of 10-fold cross validation using a window of 3 tokens before and after the current token.

The results obtained using 10-fold cross validation can be compared with those obtained with a simple baseline based on manually written patterns extracted from the annotation guidelines (Table 3).

UNCERTAINTY MARKERS	Precision	Recall	F_1
MODAL_IN_COND/SUB	84.22	98.89	90.97
IF	26.42	93.99	41.24
MODALVERB	78.98	99.66	88.12
NONVERB	77.09	79.31	78.19
VERB	89.16	66.23	76.01
Overall	68.67	90.42	78.06

Table 3. Results of baseline based on manually written patterns.

Comparison between baseline and the 10-fold cross validation demonstrates that for all markers there was a decrease in recall along with an increase in precision. This increase in precision also led to a consequence F_1 increase compared to the baseline results.
A clear trend in all the experiments is that the "if" label obtains results substantially lower than those for the other labels. This finding is associated with the fact that "if" labels are not expressed through lexical markers.

6. Discussion and Future Work

The manual annotation of the 80 papers from the BMJ, conducted using the reading grid described in section 2.3.2 and supported by a high agreement index (0.89), led us to the following results. Out of the 2,758 Uncertainty detected markers, 1,801 were lexical (*modal verbs*, *verbs*, and *non verbs*) and 957 were morphosyntactic (*if clauses* and *modal verbs in conjunctive and subjunctive mood*). Consistently with the results of a previous analysis of a corpus of Italian written texts (Bongelli and Zuczkowski 2008) when a writer communicates something as uncertain s/he uses more lexical than morphosyntactic markers.
We have created and validated a corpus of scientific, historical texts spanning over 160 years with extensive validation from language specialists. We have also applied Machine-Learning techniques to the recognition of Uncertainty markers obtaining encouraging results for most of the elements of our classification. Future research should apply this new classification beyond the realm of scientific article, possibly applying it to electronic health record texts toward the creation of decision support systems.
By making the corpus available under a Creative Commons license, we expect that this corpus can be useful for other natural language processing projects evaluating Uncertainty and hedging using a different set of algorithms. We also hope that the open access of our classification will provide an incentive toward the expansion through additional, randomly chosen articles where Uncertainty is tagged using semi-automated methods (Kiyavitskaya et. al 2005; Rosenthal et. al 2010).
In this study we limit ourselves to identify Uncertainty markers. At the moment, we are mainly working to a) the identification of the scope related to these markers and b) to assess the existence of qualitative and quantitative differences in the distribution of Uncertainty markers during the four historical time periods (see section 3.1).
Although this is a random sample, at this point the corpus is restricted to 80 articles and therefore should be expanded through semi-automated annotation methods, with iterative cycles between NLP algorithms and manual annotation by language experts.
At the end of this project, we will have made a significant improvement in our knowledge about the historical evolution of Certainty/Uncertainty language patterns in the writing of scientific papers within a 168-year span. This knowledge will guide us not only in understanding where the future of scientific communication will be, but will also assist us in training the next generation of biomedical researchers in the USA, the European Union and the rest of the world.

Acknowledgements

The work by FBK was carried out in the context of the project "eOnco - Pervasive knowledge and data management in cancer care". We would like to thank Taku Kudo and Yuji Matsumoto for making YamCha available.

References

Agarwal, S., Yu, H. (2010). Detecting hedge cues and their scope in biomedical literature with conditional random fields. *J. Biomedical Inform* 43 (6): 953-961.

Aikhenvald, A. (2003). Evidentiality in typological perspective. In A. Aikhenvald, R. M. W. Dixon (Eds), *Studies in Evidentiality* (pp. 1-31). Amsterdam/Philadelphia: John Benjamins.

Aikhenvald, A. (2004). *Evidentiality*. Oxford: Oxford University Press.

Austin, J. S., (1962). *How to do things with words*. Oxford: Oxford University Press.

Bongelli, R., Zuczkowski, A. (2008). *Indicatori linguistici percettivi e cognitivi*. Roma: Aracne.

Bongelli, R., Riccioni, I., Zuczkowski, A. (submitted). Certain-Uncertain, True-False, Good-Evil in Italian Political Speeches.

Chafe, W. (1986). Evidentiality in English conversation and academic writing. In W. Chafe, J. Nichols (Eds), *Evidentiality. The Linguistic Coding of Epistemology* (pp. 261-272). Norwood, NJ: Ablex Publishing Corp.

Cornillie, B. (2007). *Evidentiality and Epistemic Modality in Spanish (Semi-) Auxiliaries. A Cognitive–Functional Approach*. Berlin: Mouton de

Gruyter.
Crompton, P. (1997). Hedging in Academic writing: Some Theoretical Problems. *English for Specific Purposes* 16 (4): 271-287.
De Haan, F. (1999). Evidentiality and epistemic modality: setting boundaries. *Southwest Journal of Linguistics* 18 (1): 83-101.
Dendale, P., Tasmowski, L. (2001). Introduction. Evidentiality and related notions. *Journal of Pragmatics* 33 (3): 349-357.
Farkas, R., Vincze, V., Móra, G., Csirik, J., Szarvas, G. (2010). The CoNLL-2010 shared task: learning to detect hedges and their scope in natural language text. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning - Shared Task. Uppsala: Sweden.
Givón, T. (1982). Evidentiality and epistemic space. *Studies in Language* 6 (1): 23-49.
González, M. (2005). An approach to Catalan evidentiality. *Intercultural Pragmatics* 2 (4): 515-540.
Hyland, K. (1994). Hedging in academic writing and EAP textbooks. *English for Specific Purposes* 13: 239-256.
Hyland, K. (1998) Boosting, hedging and the negotiation of academic knowledge. *Text* 18 (3): 349-382.
Ifantidou, E. (2001). *Evidentials and Relevance*. Amsterdam/Philadelphia: John Benjamins.
Kamio, A. (1994). The theory of territory of information. The case of Japanese. *Journal of Pragmatics* 21 (1): 67-100.
Kamio, A. (1997). *Territory of information*. Amsterdam/Philadelphia: John Benjamins.
Kim, J., Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J. (2009). Overview of BioNLP'09 Shared Task on Event Extraction. In Proceedings of the BioNLP 2009 Workshop. Boulder: Colorado.
Kiyavitskaya, N., Zeni, N., Cordy, J. R., Mich, L., Mylopoulos, J. (2005). *Semi-Automatic Semantic Annotations for Web Documents*. In Proceedings of SWAP 2005, Vol-166, online http://ceur-ws.org/Vol-166/27.pdf
Mushin, I. (2001). *Evidentiality and Epistemological Stance*. Amsterdam/Philadelphia: John Benjamins.
Nuyts, J. (2001a). *Epistemic Modality, Language and Conceptualization*. Amsterdam/Philadelphia: John Benjamins.
Nuyts, J. (2001b). Subjectivity as an evidential dimension in epistemic modal expressions. *Journal of Pragmatics* 33 (3): 383-400.
Palmer, F. (1986). *Mood and Modality*. Cambridge: Cambridge University Press.
Papafragou, A. (2000). *Modality: Issues in the Semantics-Pragmatics Interface*. Oxford: Elsevier Science.
Papafragou, A., Li, P., Choi, Y., Han, C. (2007). Evidentiality in language and cognition. *Cognition* 103 (2): 253-299.
Petöfi, J. S. (1973). Towards an empirically motivated grammar theory of verbal texts. In J. S. Petöfi & H. Rieser (Eds.), *Studies in Text Grammar* (pp. 205-275). Dordrecht: Reidel.
Petöfi, J. S. (1980). Interpretazione e teoria del testo. In G. Galli (Ed.), *Interpretazione e contesto* (pp. 21–43). Torino: Marietti.
Petöfi, J. S. (1981). La struttura della comunicazione. In G. Galli (Ed.), *Interpretazione e strutture* (pp. 101–157). Torino: Marietti.
Petöfi, J. S. (2004). *Scrittura e interpretazione. Introduzione alla testologia semiotica nei testi verbali*. Roma: Carocci.
Plungian, V. A. (2001). The place of evidentiality within the universal grammatical space. *Journal of Pragmatics* 33 (3): 349-357.
Rosenthal, S., William J. Lipovsky, W.J., McKeown, K., Thadani, K, Andreas, J. (2010). *Towards Semi-Automated Annotation for Prepositional Phrase Attachment*. In Proceedings of the Seventh conference on International Language Resources and Evaluation LREC 2010, www.lrec-conf.org.
Riccioni, I., Bongelli, R., Zuczkowski, A. (forthcoming). The Communication of Certainty and Uncertainty in Italian Political Media Discourses.
Salager-Meyer, F. (1994). Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes* 13 (2): 149-170.
Sanders J., Spooren W. (1996). Subjectivity and certainty in epistemic modality: A study of Dutch epistemic modifiers. *Cognitive Linguistics* 7 (3) : 241-264.
Van Der Auwera, J., Plungian, V. A. (1998). Modality's semantic map. *Linguistic Typology* 2 (1): 79-124.
Vincze, V. Szarvas, G., Farkas, R., Móra, G., Csirik, J. (2008). The Bio-Scope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes. *BMC Bioinformatics* 9 (Suppl. 11): S9.
Willett, T. (1988). A cross-linguistic survey of the grammaticalization of evidentiality. *Studies in Language* 12 (1): 51-97.
Zuczkowski A., Bongelli R., Riccioni I. (2011). Proposizione costitutiva di mondo e indicatori linguistici percettivi e cognitivi nella lingua italiana. In K. Hoelker, C. Marello (Eds.), *Dimensionen der Analyse von Texten und Diskursen - Dimensioni dell'analisi di testi e discorsi* (pp. 41-61). Lit Verlag: Münster.