Uncertainty Detection in Natural Language Texts

SUMMARY OF PHD THESIS

Veronika Vincze

Research Group on Artificial Intelligence

and

University of Szeged

July 2014

Supervisor: János Csirik, DSc

University of Szeged

Doctoral School in Computer Science

1 Motivation

Uncertainty is an important linguistic phenomenon that is relevant in many fields of language processing. In its most general sense, it can be interpreted as lack of information: the receiver of the information (i.e. the hearer or the reader) cannot be certain about some pieces of information. Thus, uncertain propositions are those whose truth value or reliability cannot be determined due to lack of information. Distinguishing between factual (i.e. true or false) and uncertain propositions is of primary importance both in linguistics and natural language processing applications. For instance, in information extraction, many applications seek to extract factual information from text. Applications should handle detected modified parts in a different manner. A typical example is protein-protein interaction extraction from biological texts, where the aim is to mine text evidence for biological entities that are in a particular relation with each other. Here, while an uncertain relation might be of some interest for an end-user as well, such information must not be confused with factual textual evidence (reliable information). In machine translation, it is also necessary to identify linguistic cues of uncertainty since the source and the target language may differ in their toolkit to express uncertainty (one language employs an auxiliary, the other employs just a morpheme). To cite another example, in clinical document classification, medical reports can be grouped according to whether the patient definitely suffers, probably suffers or does not suffer from an illness.

Although uncertainty has been widely investigated in the literature, there are different terms in usage that denote slightly different linguistic phenomena. Modality is usually associated with uncertainty (Palmer, 1986), but the terms factuality (Saurí and Pustejovsky, 2012), veridicality (de Marneffe et al., 2012), evidentiality (Aikhenvald, 2004) and commitment (Diab et al., 2009) are also used. On the other hand, there are several NLP applications that seek to detect uncertainty in natural language texts in a couple of domains (e.g. biomedical texts (Velldal et al., 2012) or news (Saurí and Pustejovsky, 2012)), however, the variability in the terminology makes it difficult to compare these approaches and evaluate them on a uniform basis. Most of these approaches use annotated databases for evaluation: several uncertainty corpora like BioScope (Vincze et al., 2008), Genia (Kim et al., 2008), FactBank (Saurí and Pustejovsky, 2009), the CoNLL-2010 Shared Task corpora (Farkas et al., 2010) etc. have been constructed in the last few years. Nevertheless, the lack of unified annotation principles leads to the impossibility of direct comparison of these

corpora, which means that each of the above uncertainty detectors is optimized for the corpus or domain on which it was trained. In other words, existing uncertainty detectors can hardly be used across domains, and creating new resources and tools for each domain is time consuming and costly. Instead, a unified comprehensive approach would be optimal, which can be adapted to the specific needs of each domain without extra efforts, and the language independence of the model would also be optimal.

In this thesis, we aim at detecting uncertainty in English and Hungarian natural language texts. This research question can be investigated from a dual perspective since it is situated in the field of natural language processing, i.e. in the intersection of linguistics and computer science. Thus, in our investigations, we also make use of linguistic background but the emphasis is put on computer science. As opposed to earlier studies that focused on specific domains and were English-oriented, we offer here a comprehensive approach to uncertainty detection, which can be easily adapted to the specific needs of many domains and languages. In our investigations, we pay attention to create linguistically plausible models of uncertainty that are exploited in the implementation of our uncertainty detectors for several domains, with the help of supervised machine learning techniques.

2 Aims of the Thesis

The main aims of this thesis can be set up as follows. We first list our aims that are related to the field of computer science and then we list those related to linguistics. First, we will detect semantic uncertainty cues in English and Hungarian texts, we will examine the domain specificity of uncertainty cue distribution in the domains of biological texts, Wikipedia articles and news media and we will show how domain adaptation techniques may be exploited to across these domains. Second, we will detect discourse-level uncertainty in English and Hungarian, thus testing the language independence of uncertainty categories. Third, we demonstrate how uncertainty detection may prove useful in a real-world application, namely, information extraction from clinical discharge summaries, where the main task is to classify documents according to whether the patient suffers, probably suffers or does not suffer from a specific illness. Fourth, for these aims, it is essential to offer a unified framework in which all kinds of linguistic uncertainty may be easily placed, forming two main groups: semantic uncertainty and discourse-level uncertainty. This classification is based on considerations from both linguistics and computer science. Fifth, we will present our

own annotated corpora that conform to the principles of the above-mentioned unified framework of uncertainty. Sixth, we will compare existing corpora and annotation schemes, with special emphasis on scope-based and event-based annotation schemes and we will argue that these differences may have different implications in practical NLP applications.

3 Structure of the Thesis

This thesis can be divided into three main parts, which are briefly summarized below. The first part of the thesis introduced the background of uncertainty detection and the basics of machine learning. In the second part of the thesis, we presented uncertainty phenomena as they occur in language and annotated corpora, whereas in the third part of the thesis, we demonstrated how linguistic uncertainty can be detected in natural language texts by automatic methods.

3.1 Part I: Background

In the first part of the thesis, a general background to uncertainty detection was offered and basic concepts of machine learning were introduced, which was essential for our investigations and experiments on uncertainty detection.

3.2 Part II: Uncertainty Phenomena in Language

In the second part of the thesis, uncertainty phenomena were presented as they occur in natural language. Semantics in itself is not sufficient to give account of every uncertainty related linguistic phenomenon. There are syntactic tools to express uncertainty, for instance, passive constructions without explicitly marking the agent usually evoke the concept of weasel, i.e. sourceless propositions (Ganter and Strube, 2009). Furthermore, there are other uncertainty phenomena which cannot be described with the help of semantic or syntactic tools. They are usually related to the pragmatic factors of the discourse or world knowledge. For instance, the reliability of the information may be undermined by extralinguistic facts (e.g. the speaker is known to be a liar or no evidence is given for a statement in a scientific paper) hence it cannot be determined whether the proposition is true or false. In this way, pragmatic aspects of uncertainty and the progressive nature of discourse should be

also taken into account (see e.g. de Marneffe et al. (2012), Saurí and Pustejovsky (2012)). Thus, uncertainty is a complex phenomenon that can be understood in depth only if syntactic, semantic and pragmatic aspects are considered at the same time. We proposed an interdisciplinary unified framework for all phenomena related to uncertainty, having taken into account semantic, syntactic, pragmatic and computational linguistic considerations and we described semantic and discourse-level uncertainty phenomena in detail.

Determining whether a given proposition is uncertain or not may involve using a finite dictionary of linguistic devices, i.e. cues. On the other hand, in order to develop a supervised machine learning framework for uncertainty detection, manually annotated corpora are necessary. We presented our English and Hungarian corpora annotated on the basis of the above-mentioned framework and we also showed the specialties of cue distribution across languages, corpora, genres and domains.

It is not only the concept of uncertainty that might differ from corpus to corpus but the linguistic unit that is marked as uncertain or not can also be different. Moreover, some corpora distinguish levels of uncertainty, i.e. more or less probable statements are separately annotated. We compared the event-based and scope-based methods of uncertainty annotation by contrasting the Genia Event and BioScope 1.0 corpora and we also touched upon the question how levels of uncertainty are distinguished in several corpora.

3.3 Part III: Uncertainty Detection

In the third part of the thesis, uncertainty phenomena were detected in natural language texts. Uncertainty cue candidates do not display uncertainty in all of their occurrences. For instance, the mathematical sense of *probable* is dominant in mathematical texts while its ratio can be relatively low in papers in the humanities. The frequency of the two distinct meanings of the verb *evaluate* (which can be a synonym of *judge* /an uncertain meaning/ and *calculate*) /which is not uncertain/ is also different in the bioinformatics and cell biology domains. Compare:

(1) To **evaluate**$_{CUE}$ the PML/RARalpha role in myelopoiesis, transgenic mice expressing PML/RARalpha were engineered.

(2) Our method was **evaluated** on the Lindahl benchmark for fold recognition.

In order to differentiate between cue and non-cue uses of the same lexical items, we developed a machine learning algorithm. There, we also focused on the domain-dependent aspects of semantic uncertainty detection in English and we examined the recognition of uncertainty cues in context.

We also addressed the problem of identifying uncertainty cues in Hungarian texts and we presented our machine-learning based methods developed for solving the task.

Uncertainty detection proves to be useful in real-world applications as well. We illustrated this with the example of identifying obesity and related morbidities in the flow-text parts of clinical discharge summaries.

4 Results of the Thesis

The main results achieved in this thesis will be summarized in the next sections, listed in the order of relevance for computer science. The papers in which these results have been published are also listed, together with the author's main contributions, which were seen and approved by the co-authors of the papers.

4.1 Detecting Semantic Uncertainty

We carried out experiments on detecting semantic uncertainty in English and Hungarian – for the latter task, to the best of our knowledge, we reported the first published results. We implemented an accurate semantic uncertainty detector that distinguishes four fine-grained categories of semantic uncertainty (epistemic, doxastic, investigation and condition types) and our experiments revealed that shallow features provide good results in recognizing semantic uncertainty for both English and Hungarian. We also applied domain adaptation techniques and achieved successful results for uncertainty detection across various domains and genres in English, and we extended the feature set with semantic and pragmatic features for Hungarian (**Thesis 1**).

The main results include:

- an accurate semantic uncertainty detector that distinguishes four fine-grained categories of semantic uncertainty (epistemic, doxastic, investigation and condition types);
- the first results on semantic uncertainty detection in Hungarian texts were reported;

- our experiments revealed that shallow features provide good results in recognizing semantic uncertainty both in English and Hungarian;
- new features were introduced in the Hungarian machine learning setting for semantic uncertainty detection like semantic and pragmatic features;
- we achieved successful results for domain adaptation in English across various domains and genres by applying domain adaptation techniques to fully exploit out-of-domain data and minimize annotation costs to adapt to a new domain;
- we proved that domain specificities have a considerable effect on the efficiency of machine learning in Hungarian semantic uncertainty detection.

In Szarvas et al. (2012), semantic uncertainty phenomena are identified by a cross-domain uncertainty detector. The author participated in the data preparation and corpus annotation, she designed the uncertainty categories to be identified, she defined some of the features implemented in the machine learning algorithm, she compared the domain- and genre-specific characteristics of the texts concerning uncertainty detection and she carried out the error analysis of the experiments. The co-authors implemented the machine-learning based uncertainty detector and carried out the experiments for English, however, experimental results are considered as a shared contribution of all authors. Vincze (2014) introduces machine learning methods for identifying semantic uncertainty in Hungarian texts, based on a rich feature set that includes semantic and pragmatic features as well. Experiments on Hungarian are exclusively the author's own work.

4.2 Detecting Discourse-level Uncertainty

We implemented systems for detecting three types of uncertainty at the discourse level (weasels, hedges and peacocks). We introduced a baseline method for detecting discourse-level uncertainty in English Wikipedia texts and we applied a supervised machine learning approach to do the same in Hungarian, which was based on sequence labeling and exploited a rich feature set. We achieved reasonable results for both languages (**Thesis 2**).

The main results are the following:

- the first results on discourse-level uncertainty detection in Hungarian texts were reported;

- a baseline method for detecting discourse-level uncertainty in English Wikipedia texts;
- the first machine learning results on discourse-level uncertainty detection in Hungarian were reported;
- new features were introduced in the machine learning setting for discourse-level uncertainty detection like semantic and pragmatic features;
- we proved that domain specificities have a considerable effect on the efficiency of machine learning in Hungarian discourse-level uncertainty detection.

In Vincze (2013), the author presents some baseline experiments on identifying discourse-level uncertainty phenomena in English and she also compares her results with those of previous studies. Vincze (2014) introduces machine learning methods for identifying discourse-level uncertainty in Hungarian texts, based on a rich feature set that includes semantic and pragmatic features as well. All of the results described in these papers are the author's own work.

4.3 Uncertainty Detection in the Medical Domain

We presented a real-world application of uncertainty detection: we introduced our approach to determine the status of the patient concerning obesity and 15 related diseases from clinical discharge summaries. Our uncertainty detector had a significant role in labeling questionable cases, which proves that an uncertainty detector can be adequately applied in a real-world information extraction task (**Thesis 3**).

The main results include:

- our automatic system for identifying morbidities in the flow-text parts of clinical discharge summaries;
- we showed how uncertainty detection may enhance information extraction tasks;
- an uncertainty detector integrated into the system;
- the results demonstrate that a simple approach based on dictionary lookup and uncertainty/negation detection may be successfully applied for the task.

In Farkas et al. (2009), it is empirically shown how uncertainty detection can be fruitfully applied in a real-world task, namely, predicting morbidities from clinical texts. The author's main contributions to the paper were offering linguistics-based rules for uncertainty and negation detection, collecting uncertainty cues typical of the medical domain, determining the linguistic scope of such cues and collating dictionaries of relevant medical terms and morbidity names. The latter is a shared contribution with another co-author and statistical methods for term identification and context detection and the application of biomarkers in the system were the contributions of other co-authors. Again, the final results of the system are considered as a shared contribution of all authors.

4.4 Classification of Uncertainty Phenomena

We offered a language-independent classification of uncertainty phenomena on the basis of theoretical linguistic and computational linguistic background. We paid attention to both semantic and discourse-level uncertainty, we compared the annotation principles of existing corpora annotated for uncertainty and we also provided a unified framework in which all the uncertainty phenomena touched upon in earlier studies can be adequately placed, which served as a base for manually annotating corpora for linguistic uncertainty cues (**Thesis 4**).

The main results include:

- a language-independent classification of semantic uncertainty;
- a language-independent classification of discourse-level uncertainty;
- a comparison of the annotation principles of existing corpora annotated for uncertainty;
- a unified framework in which all the uncertainty phenomena touched upon in earlier studies can be adequately placed.

In Szarvas et al. (2012) and Vincze (2013), the classification of semantic and discourse-level uncertainty phenomena is presented in detail, which is solely the author's contribution.

4.5 Creating Corpora Annotated for Uncertainty

We created several corpora (BioScope, FactBank, WikiWeasel, hUnCertainty) and annotated them for uncertainty cues, based on the above-mentioned unified framework for un-

certainty phenomena. We also presented statistical data on cue distribution in the corpora, which revealed the domain- and genre-dependence of uncertainty detection. These corpora were used in our machine learning experiments on uncertainty detection (**Thesis 5**).

The main results include:

- the English corpora BioScope, FactBank and WikiWeasel were annotated for semantic uncertainty cues;
- WikiWeasel was also annotated for discourse-level uncertainty cues;
- the Hungarian corpus hUnCertainty was annotated for semantic and discourse-level uncertainty cues;
- hUnCertainty and WikiWeasel 3.0 are annotated on the basis of the same principles, and their cue distribution exhibit similarities, which proves the language independence of our classification of uncertainty phenomena;
- statistical data were presented on the frequency of uncertainty cues;
- based on corpus data, the distribution of semantic uncertainty cues was compared across genres and domains, which revealed the domain- and genre-dependency of uncertainty detection.

Vincze et al. (2008), Vincze (2010), Farkas et al. (2010), Szarvas et al. (2012), Vincze (2013) and Vincze (2014) introduce these corpora. The author was responsible for designing the methodology of corpus building, preparing the annotation guidelines, supervising the annotation process, moreover, she also participated in annotating and checking the data. She also carried out a statistical analysis of cue distribution in each corpus, thus proving the domain specificity of uncertainty phenomena. The co-authors of the above papers made only marginal contributions to these results like statistical analysis of data in BioScope 1.0 and defining some of the general annotation principles in BioScope 1.0.

4.6 Scope-based and Event-based Annotations

We categorized the differences between the linguistic-based and event-oriented annotation of negation and speculation in the intersection of the BioScope 1.0 and Genia Event corpora. We concluded that the scope-oriented annotation system is more adaptable to non-biomedical applications because of the high level of domain specificity in the event-oriented

annotation system. We also argued that the strength of uncertainty can manifest at the levels of both semantic and discourse-level uncertainty (**Thesis 6**).

The main results include:

- categorizing the differences between the linguistic-based and event-oriented annotation of negation and speculation in the intersection of the BioScope 1.0 and Genia Event corpora;
- estimating the frequency of mismatch categories;
- resolution strategies were offered for mismatch categories: syntactic mismatches can be solved by methods based on dependency parsing and the unified annotation scheme can offer solutions for some semantic issues;
- the scope-oriented annotation system is more adaptable to non-biomedical applications because of the high level of domain specificity in the event-oriented annotation system;
- we argued that the strength of uncertainty can manifest at the levels of both semantic and discourse-level uncertainty.

In Vincze et al. (2011), the principles behind scope-based and event-based uncertainty detection are compared on the basis of two corpora. The author's main contributions to this paper were categorizing and analyzing the mismatches between the corpora, providing the principles behind scope-based annotation, offering resolution strategies for mismatches and discussing some of the practical implications of the annotation methodology on uncertainty detection. The co-authors of the paper were responsible for principles behind event-based annotation and statistical analysis of the mismatches.

The relationship of the publications and the above listed theses is visually represented in Table 1 and the interrelationship of thesis topics, chapters and theses is visualized in Figure 1.

5 Conclusions and Future Work

In this thesis, we focused on uncertainty detection in natural language texts. On the basis of the main contributions, we can argue that:

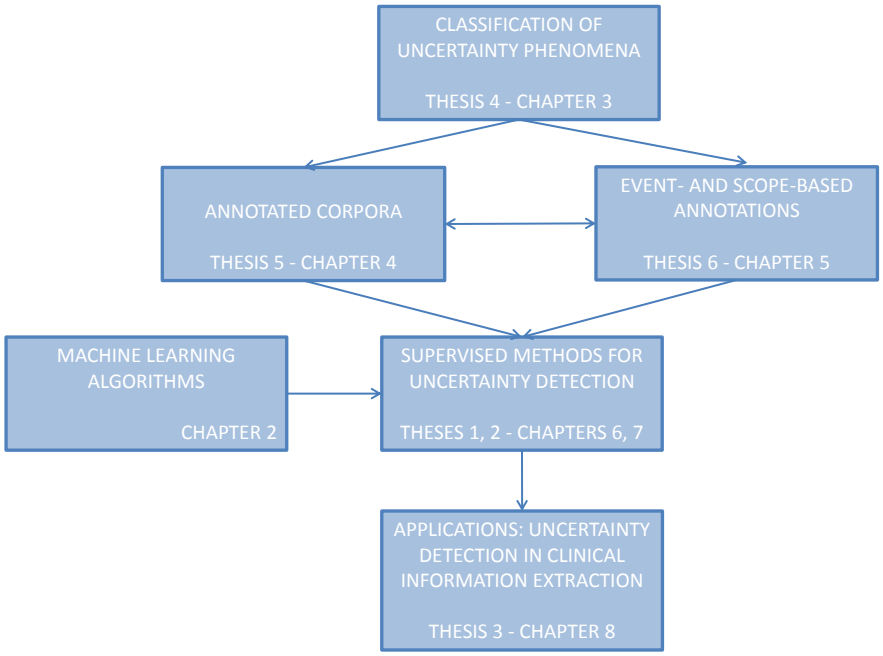

Figure 1: Thesis topics, chapters and theses.

	Thesis					
	1	2	3	4	5	6
BMC 2008 (Vincze et al., 2008)					•	
JAMIA 2009 (Farkas et al., 2009)			•			
NESP 2010 (Vincze, 2010)					•	
CoNLL 2010 (Farkas et al., 2010)					•	
JBMS 2011 (Vincze et al., 2011)						•
CL 2012 (Szarvas et al., 2012)	•			•	•	
IJCNLP 2013 (Vincze, 2013)		•		•	•	
COLING 2014 (Vincze, 2014)	•	•			•	

Table 1: The author's most important publications and the theses.

- linguistic uncertainty can be modeled in a language- and domain-independent way;
- there are domain specificities of uncertainty cue distribution;
- supervised machine learning methods can be successfully applied for uncertainty detection;
- machine learning-based uncertainty detection can be successfully carried out for English as well as for Hungarian;
- domain adaptation techniques may help diminish the distance between domains in uncertainty detection;
- the annotation scheme may determine the field of usage of the corpora, e.g. corpora with event-based annotation are mostly used in biological information extraction;
- uncertainty detectors can enhance the performance of information extraction systems as illustrated by the example of identifying morbidities in the flow-text parts of clinical discharge summaries.

Besides the main points described above, the results of the thesis may be applicable in other fields of NLP research such as information extraction, information retrieval, document classification and machine translation, as well as in other disciplines such as theoretical and contrastive linguistics.

As future work, we would like to detect uncertainty in other types of texts (for a pilot study on detecting uncertainty in Hungarian webtext, see Vincze et al. (2014)) as well as in texts written in other languages. For that purpose, we would like to annotate some data in new domains and languages and we would like to extend our tools to those areas as well. Later on, we would like to integrate our uncertainty detectors into some IE or IR applications. We believe that our research on uncertainty detection can be successfully exploited in the solution of several NLP tasks and so, it will contribute to develop novel approaches in many fields of natural language processing.

References

Aikhenvald, Alexandra Y. 2004. *Evidentiality*. Oxford University Press, Oxford.

de Marneffe, Marie-Catherine; Manning, Christopher D.; Potts, Christopher. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38:301–333, June.

Diab, Mona; Levin, Lori; Mitamura, Teruko; Rambow, Owen; Prabhakaran, Vinodkumar; Guo, Weiwei. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop*, pp. 68–73, Suntec, Singapore, August. Association for Computational Linguistics.

Farkas, Richárd; Szarvas, György; Hegedűs, István; Almási, Attila; Vincze, Veronika; Ormándi, Róbert; Busa-Fekete, Róbert. 2009. Semi-automated construction of decision rules to predict morbidities from clinical texts. *Journal of the American Medical Informatics Association*, 16:601–605.

Farkas, Richárd; Vincze, Veronika; Móra, György; Csirik, János; Szarvas, György. 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 (CoNLL-2010): Shared Task*, pp. 1–12, Uppsala, Sweden, July. Association for Computational Linguistics.

Ganter, Viola; Strube, Michael. 2009. Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia Tags and Shallow Linguistic Features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 173–176, Suntec, Singapore, August. Association for Computational Linguistics.

Kim, Jin-Dong; Ohta, Tomoko; Tsujii, Jun'ichi. 2008. Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics*, 9(Suppl 10).

Palmer, Frank Robert. 1986. *Mood and Modality*. Cambridge University Press, Cambridge.

Saurí, Roser; Pustejovsky, James. 2009. FactBank: a corpus annotated with event factuality. *Language Resources and Evaluation*, 43:227–268.

Saurí, Roser; Pustejovsky, James. 2012. Are you sure that this happened? Assessing the factuality degree of events in text. *Computational Linguistics*, 38:261–299, June.

Szarvas, György; Vincze, Veronika; Farkas, Richárd; Móra, György; Gurevych, Iryna. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics*, 38:335–367, June.

Velldal, Erik; Øvrelid, Lilja; Read, Jonathon; Oepen, Stephan. 2012. Speculation and negation: Rules, rankers, and the role of syntax. *Computational Linguistics*, 38:369–410, June.

Vincze, Veronika; Szarvas, György; Farkas, Richárd; Móra, György; Csirik, János. 2008. The BioScope Corpus: Biomedical Texts Annotated for Uncertainty, Negation and their Scopes. *BMC Bioinformatics*, 9(Suppl 11):S9.

Vincze, Veronika; Szarvas, György; Móra, György; Ohta, Tomoko; Farkas, Richárd. 2011. Linguistic scope-based and biological event-based speculation and negation annotations in the BioScope and Genia Event corpora. *Journal of Biomedical Semantics*, 2(Suppl 5):S8.

Vincze, Veronika; Simkó, Katalin Ilona; Varga, Viktor. 2014. Annotating Uncertainty in Hungarian Webtext. In *Proceedings of LAW VIII*.

Vincze, Veronika. 2010. Speculation and negation annotation in natural language texts: what the case of BioScope might (not) reveal. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pp. 28–31, Uppsala, Sweden, July. University of Antwerp.

Vincze, Veronika. 2013. Weasels, Hedges and Peacocks: Discourse-level Uncertainty in Wikipedia Articles. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 383–391, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Vincze, Veronika. 2014. Uncertainty detection in Hungarian texts. In *Proceedings of Coling 2014*.