Towards A Better Understanding of Uncertainties and Speculations in Swedish Clinical Text – Analysis of an Initial Annotation Trial

Sumithra Velupillai
Department of Computer and Systems Sciences (DSV)
Stockholm University
Forum 100
SE-164 40 Kista, Sweden
sumithra@dsv.su.se

Abstract

Electronic Health Records (EHRs) contain a large amount of free text documentation which is potentially very useful for Information Retrieval and Text Mining applications. We have, in an initial annotation trial, annotated 6 739 sentences randomly extracted from a corpus of Swedish EHRs for sentence level (un)certainty, and token level speculative keywords and negations. This set is split into different clinical practices and analyzed by means of descriptive statistics and pairwise Inter-Annotator Agreement (IAA) measured by F_1-score. We identify *geriatrics* as a clinical practice with a low average amount of uncertain sentences and a high average IAA, and *neurology* with a high average amount of uncertain sentences. Speculative words are often n-grams, and uncertain sentences longer than average. The results of this analysis is to be used in the creation of a new annotated corpus where we will refine and further develop the initial annotation guidelines and introduce more levels of dimensionality. Once we have finalized our guidelines and refined the annotations we plan to release the corpus for further research, after ensuring that no identifiable information is included.

1 Introduction

Electronic Health Records (EHRs) contain a large amount of free text documentation which is potentially very useful for Information Retrieval and Text Mining applications. Clinical documentation is specific in many ways; there are many authors in a document (e.g. physicians, nurses), there are different situations that are documented (e.g. admission, current status). Moreover, they may often be written under time pressure, resulting in fragmented, brief texts often containing spelling errors and abbreviations. With access to EHR data, many possibilities to exploit documented clinical knowledge and experience arise.

One of the properties of EHRs is that they contain reasoning about the status and diagnoses of patients. Gathering such information for the use in e.g. medical research in order to find relationships between diagnoses, treatments etc. has great potential. However, in many situations, clinicians might describe uncertain or negated findings, which is crucial to distinguish from positive or asserted findings. Potential future applications include search engines where medical researchers can search for particular diseases where negated or speculative contexts are separated from asserted contexts, or text mining systems where e.g. diseases that seem to occur often in speculative contexts are presented to the user, indicating that more research is needed. Moreover, laymen may also benefit from information retrieval systems that distinguish diseases or symptoms that are more or less certain given current medical expertise and knowledge.

We have, in an initial annotation trial, annotated 6 739 sentences randomly extracted from a corpus of Swedish EHRs for sentence level (un)certainty, and token level speculative keywords and negations[1]. In this paper, a deeper analysis of the resulting annotations is performed. The aims are to analyze the results *split into different clinical practices* by means of descriptive statistics and pairwise Inter-Annotator Agreement (IAA) measured by F_1-score, with the goal of identifying a) whether specific clinical practices contain higher or lower amounts of uncertain expressions, b)

[1]This research has been carried out after approval from the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2009/1742-31/5

whether specific clinical practices result in higher or lower IAA - indicating a less or more difficult clinical practice for judging uncertainties, and c) identifying the characteristics of the entities annotated as speculative words, are they highly lexical or is a deeper syntactic and/or semantic analysis required for modeling? From this analysis, we plan to conduct a new annotation trial where we will refine and further develop the annotation guidelines and use domain experts for annotations in order to be able to create a useful annotated corpus modeling uncertainties, negations and speculations in Swedish clinical text, which can be used to develop tools for the automatic identification of these phenomena in, for instance, Text Mining applications.

2 Related Research

In recent years, the interest for identifying and modeling speculative language in natural language text has grown. In particular, biomedical scientific articles and abstracts have been the object of several experiments. In Light et al. (2004), four annotators annotated 891 sentences each as either highly speculative, low speculative, or definite, in biomedical scientific abstracts extracted from Medline. In total, they found 11 percent speculative sentences, resulting in IAA results, measured with kappa, between 0.54 and 0.68. One of their main findings was that the majority of the speculative sentences appeared towards the end of the abstract.

Vincze et al. (2008) describe the creation of the BioScope corpus, where more than 20 000 sentences from both medical (clinical) free texts (radiology reports), biological full papers and biological scientific abstracts have been annotated with speculative and negation keywords along with their scope. Over 10 percent of the sentences were either speculative or negated. In the clinical sub-corpus, 14 percent contained speculative keywords. Three annotators annotated the corpus, and the guidelines were modified several times during the annotation process, in order to resolve problematic issues and refine definitions. The IAA results, measured with F_1-score, in the clinical sub-corpus for negation keywords ranged between 0.91 and 0.96, and for speculative keywords between 0.84 and 0.92. The BioScope corpus has been used to train and evaluate automatic classifiers (e.g. Özgür and Radev (2009) and Morante and Daelemans (2009)) with promising results.

Five qualitative dimensions for characterizing scientific sentences are defined in Wilbur et al. (2006), including levels of certainty. Here, guidelines are also developed over a long period of time (more than a year), testing and revising the guidelines consecutively. Their final IAA results, measured with F_1-score, range between 0.70 and 0.80. Different levels of dimensionality for categorizing certainty (in newspaper articles) is also presented in Rubin et al. (2006).

Expressions for communicating probabilities or levels of certainty in clinical care may be inherently difficult to judge. Eleven observers were asked to indicate the level of probability of a disease implied by eighteen expressions in the work presented by Hobby et al. (2000). They found that expressions indicating intermediate probabilities were much less consistently rated than those indicating very high or low probabilities. Similarly, Khorasani et al. (2003) performed a survey analyzing agreement between radiologists and non-radiologists regarding phrases used to convey degrees of certainty. In this study, they found little or no agreement among the survey participants regarding the diagnostic certainty associated with these phrases. Although we do not have access to radiology reports in our corpus, these findings indicate that it is not trivial to classify uncertain language in clinical documentation, even for domain experts.

3 Method

The annotation trial is based on sentences randomly extracted from a corpus of Swedish EHRs (see Dalianis and Velupillai (2010) for an initial description and analysis). These records contain both structured (e.g. measure values, gender information) and unstructured information (i.e. free text). Each free text entry is written under a specific heading, e.g. *Status*, *Current medication*, *Social Background*. For this corpus, sentences were extracted only from the free text entry *Assessment* (Bedömning), with the assumption that these entries contain a substantial amount of reasoning regarding a patient's diagnosis and situation. A simple sentence tokenizing strategy was employed, based on heuristic regular expressions[2]. We have used Knowtator (Ogren, 2006) for the annotation

[2]The performance of the sentence tokenizer has not been evaluated in this work.

work.

One senior level student (SLS), one undergraduate computer scientist (UCS), and one undergraduate language consultant (ULC) annotated the sentences into the following classes; on a sentence level: *certain*, *uncertain* or *undefined*, and on a token level: *speculative words*, *negations*, and *undefined words*.

The annotators are to be considered *naive* coders, as they had no prior knowledge of the task, nor any clinical background. The annotation guidelines were inspired by those created for the BioScope corpus (Vincze et al., 2008), with some modifications (see Dalianis and Velupillai (2010)). The annotators were allowed to break a sentence into subclauses if they found that a sentence contained conflicting levels of certainty, and they were allowed to mark question marks as speculative words. They did not annotate the linguistic scopes of each token level instance. The annotators worked independently, and met for discussions in even intervals (in total seven), in order to resolve problematic issues. No information about the clinic, patient gender, etc. was shown. The annotation trial is considered as a first step in further work of annotating Swedish clinical text for speculative language.

Clinical practice	**# sentences**	**# tokens**
hematology	140	1 494
surgery	295	3 269
neurology	351	4 098
geriatrics	142	1 568
orthopaedics	245	2 541
rheumatology	384	3 348
urology	120	1 393
cardiology	128	1 242
oncology	550	5 262
ENT	224	2 120
infection	107	1 228
emergency	717	6 755
paediatrics	935	8 926
total, clinical practice	4 338	43 244
total, full corpus	6 739	69 495

Table 1: Number of sentences and tokens per clinical practice (#sentences $>$ 100), and in total. ENT = Ear, Nose and Throat.

3.1 Annotations and clinical practices

The resulting corpus consists of 6 739 sentences, extracted from 485 unique clinics. In order to be able to analyze possible similarities and differences across clinical practices, sentences from clinics belonging to a specific practice type were grouped together. In Table 1, the resulting groups, along with the total amount of sentences and tokens, are presented[3]. Only groups with a total amount of sentences $>$ 100 were used in the analysis, resulting in 13 groups. A clinic was included in a clinical practice group based on a priority heuristics, e.g. the clinic "Barnakuten-kir" (*Paediatric emergency surgery*) was grouped into paediatrics.

The average length (in tokens) per clinical practice and in total are given in Table 2. Clinical documentation is often very brief and fragmented, for most clinical practices (except urology and cardiology) the minimum sentence length (in tokens) was one, e.g. "basal", "terapisvikt" (*therapy failure*), "lymfödem" (*lymphedema*), "viros" (*virosis*), "opanmäles" (*reported to surgery*, compound with abbreviation). We see that the average sentence length is around ten for all practices, where the shortest are found in rheumatology and the longest in infection.

As the annotators were allowed to break up sentences into subclauses, but not required to, this led to a considerable difference in the total amount of annotations per annotator. In order to be able to analyze similarities and differences between the resulting annotations, all sentence level annotations were converted into *one* sentence class only, the primary class (defined as the first sentence level annotation class, i.e. if a sentence was broken into two clauses by an annotator, the first being *certain* and the second being *uncertain*, the final sentence level annotation class will be *certain*). The sentence level annotation class *certain* was in clear majority among all three annotators. On both sentence and token level, the class *undefined* (a sentence that could not be classified as *certain* or *uncertain*, or a token which was not clearly speculative) was rarely used. Therefore, all sentence level annotations marked as *undefined* are converted to the majority class, *certain*, resulting in two sentence level annotation classes (*certain* and *uncertain*) and two token level annotation classes (*speculative words* and *negations*, i.e. to-

[3]White space tokenization.

kens annotated as *undefined* are ignored).

For the remaining analysis, we focus on the distributions of the annotation classes *uncertain* and *speculative words*, per annotator and annotator pair, and per clinical practice.

Clinical practice	Max	Avg	Stddev
hematology	40	10.67	7.97
surgery	57	11.08	8.29
neurology	105	11.67	10.30
geriatrics	58	11.04	9.29
orthopaedics	40	10.37	6.88
rheumatology	59	8.72	7.99
urology	46	11.61	7.86
cardiology	50	9.70	7.46
oncology	54	9.57	7.75
ENT	54	9.46	7.53
infection	37	11.48	7.76
emergency	55	9.42	6.88
paediatrics	68	9.55	7.24
total, full corpus	120	10.31	8.53

Table 2: Token statistics per sentence and clinical practice. All clinic groups except urology (min = 2) and cardiology (min = 2) have a minimum sentence length of one token.

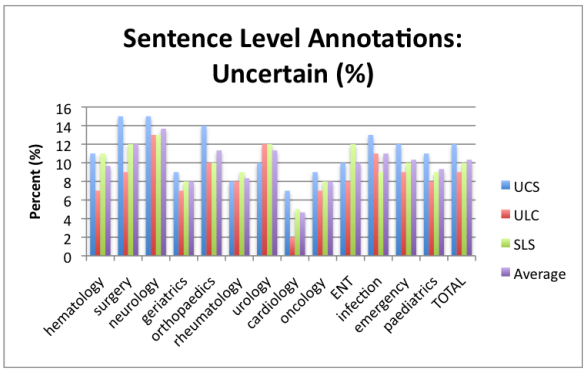

Figure 1: Sentence level annotation: *uncertain*, percentage per annotator and clinical practice.

4 Results

We have measured the proportions (in percent) per annotator for each clinical practice and in total. This enables an analysis of whether there are substantial individual differences in the distributions, indicating that this annotation task is highly subjective and/or difficult. Moreover, we measure IAA by pairwise F_1-score. From this, we may draw conclusions whether specific clinical practices are harder or easier to judge *reliably* (i.e. by high IAA results).

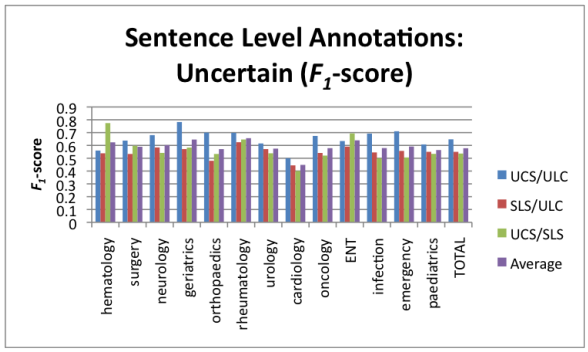

Figure 2: Pairwise F_1-score, sentence level annotation class *uncertain*.

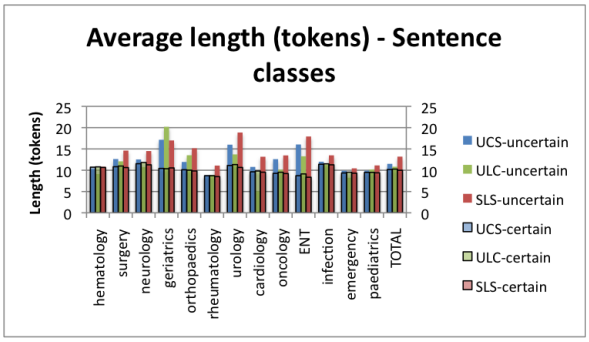

Figure 3: Average length in tokens, per annotator and sentence class.

In Figure 1, we see that the average amount of uncertain sentences lies between 9 and 12 percent for each annotator in the full corpus. In general, UCS has annotated a larger proportion of uncertain sentences compared to ULC and SLS.

The clinical discipline with the highest average amount of uncertain sentences is *neurology* (13.7 percent), the lowest average amount is found in *cardiology* (4.7 percent). Surgery and cardiology show the largest individual differences in proportions (from 9 percent (ULC) to 15 percent (UCS), and from 2 percent (ULC) to 7 percent (UCS), respectively).

However, in Figure 2, we see that the pairwise IAA, measured by F_1-score, is relatively low, with an average IAA of 0.58, ranging between 0.54 (UCS/SLS) and 0.65 (UCS/ULC), for the entire corpus. In general, the annotator pair UCS/ULC have higher IAA results, with the highest for *geriatrics* (0.78). The individual proportions for un-

certain sentences in *geriatrics* is also lower for all annotators (see Figure 1), indicating a clinical practice with a low amount of uncertain sentences, and a slightly higher average IAA (0.64 F_1-score).

4.1 Sentence lengths

As the focus lies on analyzing sentences annotated as *uncertain*, one interesting property is to look at sentence lengths (measured in tokens). One hypothesis is that uncertain sentences are in general longer. In Figure 3 we see that in general, for all three annotators, uncertain sentences are longer than certain sentences. This result is, of course, highly influenced by the skewness of the data (i.e. uncertain sentences are in minority), but it is clear that uncertain sentences, in general, are longer on average. It is interesting to note that the annotator SLS has, in most cases, annotated longer sentences as uncertain, compared to UCS and ULC. Moreover, *geriatrics*, with relatively high IAA but relatively low amounts of uncertain sentences, has well above average sentence lengths in the *uncertain* class.

4.2 Token level annotations

When it comes to the token level annotations, *speculative words* and *negations*, we observed very high IAA for *negations* (0.95 F_1-score (exact match) on average in the full corpus, the lowest for *neurology*, 0.94). These annotations were highly lexical (13 unique tokens) and unambiguous, and spread evenly across the two sentence level annotation classes (ranging between 1 and 3 percent of the total amount of tokens per class). Moreover, all negations were unigrams.

On the other hand, we observed large variations in IAA results for *speculative words*. In Figure 4, we see that there are considerable differences between exact and partial matches[4] between all annotator pairs, indicating individual differences in the interpretations of what constitutes a speculative word and how many tokens they cover, and the lexicality is not as evident as for negations. The highest level of agreement we find between UCS/ULC in *orthopaedics* (0.65 F_1-score, partial match) and *neurology* (0.64 F_1-score, partial match), and the lowest in *infection* (UCS/SLS, 0.31 F_1-score).

[4]Partial matches are measured on a character level.

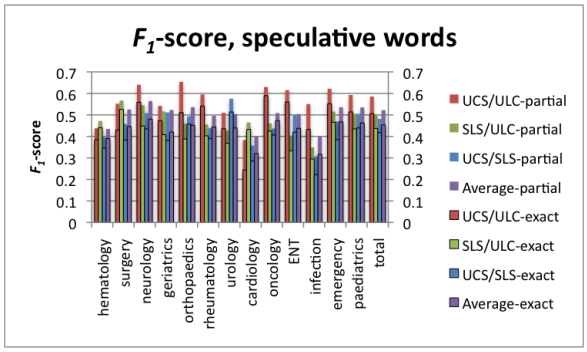

Figure 4: F_1-score, speculative words, exact and partial match.

4.2.1 Speculative words – most common

The low IAA results for *speculative words* invites a deeper analysis for this class. How is this interpreted by the individual annotators? First, we look at the most common tokens annotated as *speculative words*, shared by the three annotators: ”?”, ”sannolikt” (*likely*), ”ev” (*possibly*, abbreviated), ”om” (*if*). The most common speculative words are all unigrams, for all three annotators. These tokens are similar to the most common speculative words in the clinical BioScope subcorpus, where *if*, *may* and *likely* are among the top five most common. Those tokens that are most common per annotator and not shared by the other two (among the five most frequent) include ”bedöms” (*judged*), ”kan” (*could*), ”helt” (*completely*) and ”ställningstagande” (*standpoint*).

Looking at *neurology* and *urology*, with a higher overall average amount of uncertain sentences, we find that the most common words for *neurology* are similar to those most common in total, while for *urology* we find more n-grams. In Table 3, the five most common speculative words per annotator for neurology and urology are presented.

When it comes to the unigrams, many of these are also *not* annotated as speculative words. For instance, ”om” (*if*), is annotated as speculative in only 9 percent on average of its occurrence in the neurological data (the same distribution holds, on average, in the total set). In Morante and Daelemans (2009), *if* is also one of the words that are subject to the majority of false positives in their automatic classifier. On the other hand, ”sannolikt” (*likely*) is almost always annotated as a speculative word (over 90 percent of the time).

	UCS	ULC	SLS
neurology	? sannolikt (*likely*) kan (*could*) om (*if*) pröva (*try*) ter (*seem*)	? kan (*could*) sannolikt (*likely*) om (*if*) verkar (*seems*) ev (*possibly*, abbr)	? sannolikt (*likely*) ev (*possibly*, abbr) om (*if*) ställningstagande (*standpoint*) möjligen (*possibly*)
urology	kan vara (*could be*) tyder på(*indicates*) ev (*possibly*, abbr) misstänkt (*suspected*) kanske (*perhaps*) planeras tydligen (*apparently planned*)	mycket (*very*) inga tecken (*no signs*) kan vara (*could be*) kan (*could*) tyder (*indicates*) misstänkt (*suspected*)	tyder på(*indicates*) i första hand (*primarily*) misstänkt (*suspected*) kanske (*perhaps*) skall vi försöka (*should we try*) kan vara (*could be*)

Table 3: Most common speculative words per annotator for *neurology* and *urology*.

4.2.2 Speculative words – n-grams

Speculative words are, in Swedish clinical text, clearly not simple lexical unigrams. In Figure 5 we see that the average length of tokens annotated as *speculative words* is, on average, 1.34, with the longest in *orthopaedics* (1.49) and *urology* (1.46). We also see that SLS has, on average, annotated longer sequences of tokens as *speculative words* compared to UCS and ULC. The longest n-grams range between three and six tokens, e.g. ”kan inte se några tydliga” (*can't see any clear*), ”kan röra sig om” (*could be about*), ”inte helt har kunnat uteslutas” (*has not been able to completely exclude*), ”i första hand” (*primarily*). In many of these cases, the strongest indicator is actually a unigram (”kan” (*could*)), within a verb phrase. Moreover, negations inside a *speculative word* annotation, such as ”inga tecken” (*no signs*) are annotated differently among the individual annotators.

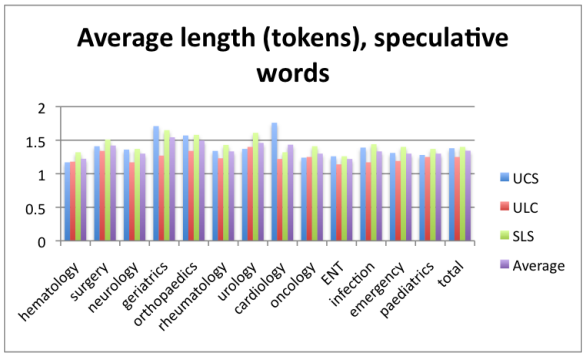

Figure 5: Average length, speculative words.

4.3 Examples

We have observed low average pairwise IAA for sentence level annotations in the *uncertain* class, with more or less large differences between the annotator pairs. Moreover, at the token level and for the class *speculative words*, we also see low average agreement, and indications that *speculative words* often are n-grams. We focus on the clinical practices *neurology*, because of its average large proportion of uncertain sentences, *geriatrics* for its high IAA results for UCS/ULC and low average proportion of uncertain sentences, and finally *surgery*, for its large discrepancy in proportions and low average IAA results.

In Example 1 we see a sentence where two annotators (ULC, SLS) have marked the sentence as *uncertain*, also marking a unigram (”ospecifik” (*unspecific*) as a *speculative word*. This example is interesting since the utterance is ambiguous, it can be judged as certain as in *the dizziness is confirmed to be of an unspecific type* or uncertain as in *the type of dizziness is unclear*, a type of utterance which should be clearly addressed in the guidelines.

<C> Yrsel av ospecifik typ. </C>

<U> Yrsel av <S> ospecifik </S> typ. </U>

<U> Yrsel av <S> ospecifik </S> typ. </U>

Dizziness of unspecific type

Example 1: Annotation example, *neurology*. Ambiguous sentence, *unspecific* as a possible speculation cue. C = Certain, U = Uncertain, S = Speculative words.

An example of different interpretations of the minimum span a *speculative word* covers is given in Example 2. Here, we see that ”inga egentliga märkbara” (*no real apparent*) has been annotated in three different ways. It is also interesting to

note the role of the negation as part of amplifying speculation. Several such instances were marked by the annotators (for further examples, see Dalianis and Velupillai (2010)), which conforms well with the findings reported in Kilicoglu and Bergler (2008), where it is showed that explicit certainty markers together with negation are indicators of speculative language. In the BioScope corpus (Vincze et al., 2008), such instances are marked as speculation cues. This example, as well as Example 1, is also interesting as they both clearly are part of a longer passage of reasoning of a patient, with no particular diagnosis mentioned in the current sentence. Instead of randomly extracting sentences from the free text entry *Assessment*, one possibility would be to let the annotators judge all sentences in an entry (or a full EHR). Doing this, differences in where speculative language often occur in an EHR (entry) might become evident, as for scientific writings, where it has been showed that speculative sentences occur towards the end of abstracts (Light et al., 2004).

<U> <S><N> Inga </N> egentliga </S> <S> märkbara</S> minnessvårigheter under samtal. </U>.

<U> <N> Inga </N> <S> egentliga </S> märkbara minnessvårigheter under samtal. </U>.

<U> <S><N> Inga </N> egentliga märkbara </S> minnessvårigheter under samtal. </U>.

No real apparent memory difficulties during conversation

Example 2: Annotation example, *neurology*. Different annotation coverage over negation and speculation. C = Certain, U = Uncertain, S = Speculative words, N = Negation

In *geriatrics*, we have observed a lower than average amount of uncertain sentences, and high IAA between UCS and ULC. In Example 3 we see a sentence where UCS and ULC have matching annotations, whereas SLS has judged this sentence as certain. This example shows the difficulty of interpreting expressions indicating possible speculation – is ”ganska” (*relatively*) used here as a marker of certainty (as certain as one gets when diagnosing this type of illness)?

The word ”sannolikt” (*likely*) is one of the most common words annotated as a speculative word in the total corpus. In Example 4, we see a sentence where the annotators UCS and SLS have judged it to be *uncertain*, while UCS and ULC have marked the word ”sannolikt” (*likely*) as a *speculative word*. This is an interesting example, through informal discussions with clinicians we were informed that this word might as well be used as a marker of high certainty. Such instances show the need for using domain experts in future annotations of similar corpora.

<U> Både anamnestiskt och testmässigt <S> ganska </S> stabil vad det gäller Alzheimer sjukdom. </U>.

<U> Både anamnestiskt och testmässigt <S> ganska </S> stabil vad det gl̈ler Alzheimer sjukdom. </U>.

<C> Både anamnestiskt och testmässigt ganska stabil vad det gäller Alzheimer sjukdom. </C>.

Both anamnesis and tests relatively stabile when it comes to Alzheimer's disease.

Example 3: Annotation example, *geriatrics*. Different judgements for the word ”ganska” (*relatively*). C = Certain, U = Uncertain, S = Speculative words.

<C>En 66-årig kvinna med <S>sannolikt</S> 2 synkrona tumörer vänster colon/sigmoideum och där till levermetastaser.</C>.

<U>En 66-årig kvinna med <S>sannolikt</S> 2 synkrona tumörer vänster colon/sigmoideum och där till levermetastaser.</U>.

<C>En 66-årig kvinna med sannolikt 2 synkrona tumörer vänster colon/sigmoideum och där till levermetastaser.</C>.

A 66 year old woman likely with 2 synchronous tumours left colon/sigmoideum in addition to liver metastasis.

Example 4: Annotation example, *surgery*. Different judgements for the word ”sannolikt” (*likely*). C = Certain, U = Uncertain, S = Speculative words.

5 Discussion

We have presented an analysis of an initial annotation trial for the identification of uncertain sentences as well as for token level cues (*speculative words*) across different clinical practices. Our main findings are that IAA results for both sentence level annotations of uncertainty and token level annotations for speculative words are, on av-

erage, fairly low, with higher average agreement in *geriatrics* and *rheumatology* (see Figures 1 and 2). Moreover, by analyzing the individual distributions for the classes *uncertain* and *speculative words*, we find that *neurology* has the highest average amount of uncertain sentences, and *cardiology* the lowest. On average, the amount of uncertain sentences ranges between 9 and 12 percent, which is in line with previous work on sentence level annotations of uncertainty (see Section 2).

We have also showed that the most common *speculative words* are unigrams, but that a substantial amount are n-grams. The n-grams are, however, often part of verb phrases, where the head is often the speculation cue. However, it is evident that speculative words are not always simple lexical units, i.e. syntactic information is potentially very useful. Question marks are the most common entities annotated as *speculative words*. Although these are not interesting indicators in themselves, it is interesting to note that they are very common in clinical documentation.

From the relatively low IAA results we draw the conclusion that this task is difficult and requires more clearly defined guidelines. Moreover, using *naive* coders on clinical documentation is possibly not very useful if the resulting annotations are to be used in, e.g. a Text Mining application for medical researchers. Clinical documentation is highly domain-specific and contains a large amount of internal jargon, which requires judgements from clinicians. However, we find it interesting to note that we have identified differences between different clinical practices. A consensus corpus has been created from the resulting annotations, which has been used in an experiment for automatic classification, see Dalianis and Skeppstedt (2010) for initial results and evaluation.

During discussions among the annotators, some specific problems were noted. For instance, the extracted sentences were not always about the patient or the current status or diagnosis, and in many cases an expression could describe (un)certainty of someone other than the author (e.g. another physician or a family member), introducing aspects of perspective. The sentences annotated as *certain*, are difficult to interpret, as they are simply *not uncertain*. We believe that it is important to introduce further dimensions, e.g. explicit certainty, and focus (*what* is (un)certain?), as well as time (e.g. *current* or *past*).

6 Conclusions

To our knowledge, there is no previous research on annotating Swedish clinical text for sentence and token level uncertainty together with an analysis of the differences between different clinical practices. Although the initial IAA results are in general relatively low for all clinical practice groups, we have identified indications that *neurology* is a practice which has an above average amount of uncertain elements, and that *geriatrics* has a below average amount, as well as higher IAA. Both these disciplines would be interesting to continue the work on identifying speculative language.

It is evident that clinical language contains a relatively high amount of uncertain elements, but it is also clear that naive coders are not optimal to use for interpreting the contents of EHRs. Moreover, more care needs to be taken in the extraction of sentences to be annotated, in order to ensure that the sentences actually describe reasoning about the patient status and diagnosis. For instance, instead of randomly extracting sentences from within a free text entry, it might be better to let the annotators judge all sentences within an entry. This would also enable an analysis of whether speculative language is more or less frequent in specific parts of EHRs.

From our findings, we plan to further develop the guidelines and particularly focus on specifying the minimal entities that should be annotated as *speculative words* (e.g. ”kan” (*could*)). We also plan to introduce further levels of dimensionality in the annotation task, e.g. cues that indicate a high level of certainty, and to use domain experts as annotators. Although there are problematic issues regarding the use of *naive* coders for this task, we believe that our analysis has revealed some properties of speculative language in clinical text which enables us to develop a useful resource for further research in the area of speculative language. Judging an instance as being certain or uncertain is, perhaps, a task which can never exclude subjective interpretations. One interesting way of exploiting this fact would be to exploit individual annotations similar to the work presented in Reidsma and op den Akker (2008). Once we have finalized the annotated set, and ensured that no identifiable information is included, we plan to make this resource available for further research.

References

Hercules Dalianis and Maria Skeppstedt. 2010. Creating and Evaluating a Consensus for Negated and Speculative Words in a Swedish Clinical Corpus. To be published in the proceedings of the Negation and Speculation in Natural Language Processing Workshop, July 10, Uppsala, Sweden.

Hercules Dalianis and Sumithra Velupillai. 2010. How Certain are Clinical Assessments? Annotating Swedish Clinical Text for (Un)certainties, Speculations and Negations. In *Proceedings of the of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta, May 19-21.

J. L. Hobby, B. D. M. Tom, C. Todd, P. W. P. Bearcroft, and A. K. Dixon. 2000. Communication of doubt and certainty in radiological reports. *The British Journal of Radiology*, 73:999–1001, September.

R. Khorasani, D. W. Bates, S. Teeger, J. M. Rotschild, D. F. Adams, and S. E. Seltzer. 2003. Is terminology used effectively to convey diagnostic certainty in radiology reports? *Academic Radiology*, 10:685–688.

Halil Kilicoglu and Sabine Bergler. 2008. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. *BMC Bioinformatics*, 9(S-11).

Marc Light, Xin Ying Qiu, and Padmini Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In Lynette Hirschman and James Pustejovsky, editors, *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*, pages 17–24, Boston, Massachusetts, USA, May 6. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2009. Learning the scope of hedge cues in biomedical texts. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 28–36, Morristown, NJ, USA. Association for Computational Linguistics.

Philip V. Ogren. 2006. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–275, Morristown, NJ, USA. Association for Computational Linguistics.

Arzucan Özgür and Dragomir R. Radev. 2009. Detecting speculations and their scopes in scientific text. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1398–1407, Singapore, August. Association for Computational Linguistics.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *HumanJudge '08: Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 8–16, Morristown, NJ, USA. Association for Computational Linguistics.

Victoria L. Rubin, Elizabeth D. Liddy, and Noriko Kando. 2006. Certainty identification in texts: Categorization model and manual tagging results. In *Computing Affect and Attitutde in Text: Theory and Applications*. Springer.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S-11).

J. W. Wilbur, A. Rzhetsky, and H. Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7:356+, July.