Professional Language in Swedish Radiology Reports – Characterization for Patient-Adapted Text Simplification

Maria Kvist[ab], Sumithra Velupillai[a]

[a] *Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden*
[b] *Department of Learning, Informatics, Management and Ethics (LIME), Karolinska Institutet, Sweden*

Abstract

In health care, there is a need for patient adaption of clinical text, so that patients can understand their own health records. As a base for construction of automated text simplification tools, characterization of the clinical language is needed.

We describe a corpus of 0.43 mill. radiology reports from a University Hospital, characterize it quantitatively and perform a qualitative content analysis. The results show that a limited set of words and phrases are recurrent in the reports and can be used for exchange to more easy-to-read vocabulary. Semantic categories such as body parts, findings, procedures, and administrative information can be used in the simplification process.

This study investigates the potentials and the pitfalls for text simplification of medical Swedish into general Swedish for laymen.

Keywords: clinical text, health records, text simplification, natural language processing, patient empowerment

Introduction

Health records are accessible to read for patients in several countries via patient portals on the Internet, for example North America [1, 2] and the Scandinavian countries Denmark[1] and Norway[2]. Today, making record information available to the citizens via e-services is a highly prioritized goal also for Sweden [3, 4]. The roll-out of the e-service *Min journal* (My health record) is planned at the end of this year [5], using the patient portal *Mina Vårdkontakter*[3] (My Health Care Contacts) as presently used in Uppsala county. However, there are a number of issues being discussed involving technology, security, legal and ethical questions, but there are also language issues. The content of the health records will not be available to the patients just because the records are accessible on-line; it needs to be comprehensible for laymen readers. Many studies report problems for patients to understand health records [6, 7, 8, 9, 10]. There is a need for explicatory tools to decipher the content.

Clinical text

Patient records have a high content of medical terminology, necessary for exactness and medical safety. In health care, a precise language is needed to describe findings and events. The medical professional language is fact dense and also contains many abbreviations, neologisms and jargon, and is laced with words of foreign origin [7, 11, 12]. In Swedish medical records, foreign words are of Latin, Greek and English origin. The style is ungrammatical, telegraphic and informal, but also characterized by the fact that health records are legal documents, for example containing many passive verbs typical for official documents [12].

The text is not addressed to the patient, and this is a problem for patients who want to follow their own health care-process. Studies from several countries have shown that patients find it especially difficult to understand test results, radiology reports, and medication lists, and in all of these the main issue is medical terminology and abbreviations [6, 7, 8, 9, 13, 14].

Text simplification

Text simplification of medical records will require various, and probably simultaneous, approaches; e.g. lexical exchange for terminology, abbreviation expansion, compound splitting, and syntactic simplification. The level of simplification depends on the intended audience and the purpose.

There have been several efforts for making medical information more consumer friendly [15]. Kandula et al [16] developed a tool that addresses semantic difficulty by substituting difficult terms with easier synonyms, or hierarchically or semantically related terms, and syntactic complexity by splitting long sentences. Leroy et al [17] developed an algorithm for semi-automated simplification of medical text, using a measurement of term familiarity based on lexical and grammatical corpus analysis, to help estimate text difficulty.

Project: Automated text simplification of radiology reports

On-line health records can widen social gaps of health care usage if not taking into consideration the diverse language abilities of patient groups, as well as varying health literacy [15, 18]. If possible, the health records should be written in a language that is understandable for many patients. However, this is not always feasible. In radiologic reports, which convey communication from the radiologist to the treating physician, there is a need for precise descriptions, and it is not possible to compromise with words from layman terminology that can be less distinct. Thus they are not written with the aim of making the text comprehensible for the patient. Several studies have shown that radiology reports are among the most difficult form of clinical text to understand [e.g. 6].

To address the ongoing efforts in Sweden for making records available online, a pilot project on making medical records more readable for patients has been initiated. The long-term goal of this project is to construct a “translator” that can pro-

[1] www.sundhed.dk
[2] www.minjournal.no
[3] www.minavardkontakter.se

duce a simplified parallel text to the original medical text, thus respecting the needs of both the patients and the professionals.

The aim of this study is to describe a large corpus of Swedish radiology reports, made available for research by ethical approval. Our goals are to quantitatively analyze the content of these records, and qualitatively characterize the most frequent terms and sentences. The purpose is to understand the content of the radiology reports and to be able to use this corpus in the future development of a text simplification tool enabling patients to better comprehend medical text.

Materials and Methods

A corpus of Swedish radiology reports was characterized quantitatively in aspects of frequencies of words, bi- and trigrams, and sentences. A qualitative content analysis was performed for each of these aspects.

Materials

The radiology reports are part of the Stockholm EPR Corpus[4], a large corpus of health records containing more than 600 000 unique patients from the greater Stockholm area during the years 2006-2010 [19]. The records are de-identified with anonymized serial numbers for individual patients.

To create the Stockholm EPR X-ray corpus, we used radiology reports for examinations performed during the years 2009-2010 at Karolinska University Hospital. The size of the corpus is 434 427 reports, containing both the text of the referral as well as the result of the radiologic examination (i.e. questions and answers). Radiologic examinations were performed on in-house patients and patients referred from outpatient clinics to the radiology departments: general radiology as well as thoracic, neurologic and pediatric radiology. The reports are examinations of all patients for this period, i.e. both genders (50.34% females) and all ages from premature babies to a 108 year-old. Due to missing values in the database (1.6%), gender and birth year information is not available for all 152 170 unique patients. 19.41% are under the age of 18, and the majority of the patients are born 1931 - 1970 (51.97%). On average, there are 2.85 reports per patient (min = 1, max = 142), with 50% of the patients having 4 or less reports and 75% of the patients having 10 or less reports.

In this study, we examined only the texts originating from the radiology departments and not the text comprising the referrals.

Methods

For the quantitative corpus statistics analysis, we extracted a number of different corpus categories: all words, bi- and trigrams (sequences of two (bi-) and three (tri-) adjacent words in the corpus) and sentences. Moreover, we extracted all nouns, verbs and adjectives. For each category, we counted frequency information for types (unique occurrences) and tokens (all actual items). The Natural Language Toolkit (NLTK) [20] was used for extracting words, bi- and trigrams, and sentences. For nouns, verbs and adjectives, a Part-of-Speech (POS) tagger trained for general Swedish was used: Stagger [21]. To account for inflected forms of words, Stagger was also used to create lists of lemmatized words (inflections conflated to base form).

A qualitative content analysis was performed on the 100 most frequent items for each category (words, bi- and trigrams, sentences, nouns, verbs and adjectives). The method included a step of modification: the top 100-lists were edited in regard to names for de-identification and POS-tagging errors by simply removing these posts and replacing with next posts. The content analysis included classifying the content into new semantic categories. Verbs were also analyzed with regards to active and passive voice. The qualitative investigation was performed by a senior physician and a computer linguist.

Results

Corpus statistics

The statistics for words, bi-/trigrams and sentences in the Stockholm EPR X-ray corpus as well as the three word classes are shown in Tables 1 and 2.

Table 1 – Number of words, bigrams, trigrams and sentences in the Stockholm EPR X-ray corpus

	All words	bigram	trigram	sentences
types	200703	2534969	5357542	1874464
tokens	20290064	17728463	15276077	2567035
top100				
tokens	7150511	2759515	1403104	201074
%	35%	16%	9,2%	7,8%

The sum of the top 100 sentences represents 7.8 % of the total number of sentences in the corpus. However, the vocabulary is recurrent; the sum of the top 100 words represents 35 % of the total number of words used in the corpus, and 16 % of all bigrams are found among the top 100 most frequently occurring bigrams. When conflated to the base forms (lemmatized) in the three word classes, the proportions are higher (Table 2). More than half of the nouns are found among the top 100, as well as an astounding 79% and 74%, respectively of all the verbs and adjectives. This is in part due to the very frequent use of a few words (figure 2). Removing the 5 most common nouns, verbs and adjectives, shows that the top 100 still make up a sizable part of the total token (39 %, 35 % and 47 % respectively).

Table 2 – Number of nouns, verbs and adjectives in the Stockholm EPR X-ray corpus, lemmatized

	Noun	Verb	Adjective
types	111468	20351	25278
tokens	8254868	2079040	2951736
top100:			
tokens	4389486	1640789	2189288
%	53%	79%	74%

On average, a radiology report consisted of five sentences (min = 1, max = 66). The average length of the sentences (with word frequencies between 1 – 40) was 12 words (Figure 1). Many sentences were not full sentences; among the top 100 sentences only 23 contained both subject and predicate. Also, 7 of the top 100 sentences were composed of a single word, and 30 were composed of only two words. Among the

[4] This research was approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/2028-31/5

longest sentences of the 100 most frequent were standard phrases of administrative character.

A small number of words and expressions were found to be very frequent (Figure 2), which can be explained by the structure of the radiology reports. Headings and administrative phrases such as information about dates, names of radiologists, pagers and telephone numbers are included in the free text section of the reports. The peak at 22 words per sentence (Figure 1) reflects an example of this; these sentences contained a recurring administrative phrase.

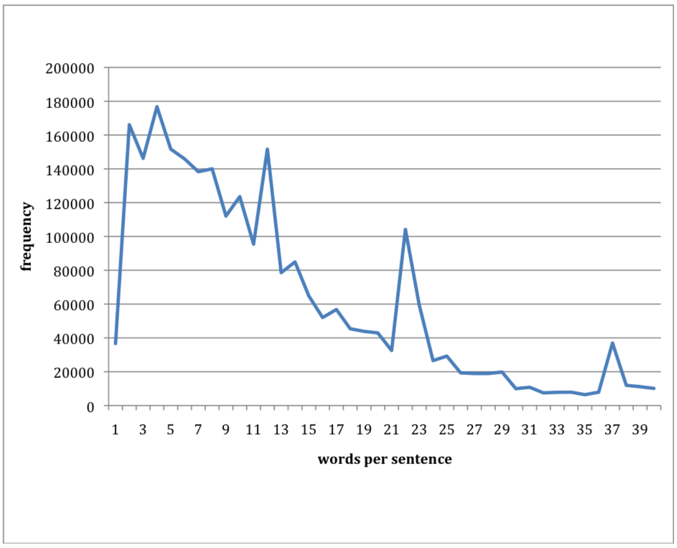

Figure 1 – Frequency curve for number of words per sentence, Stockholm EPR X-ray corpus

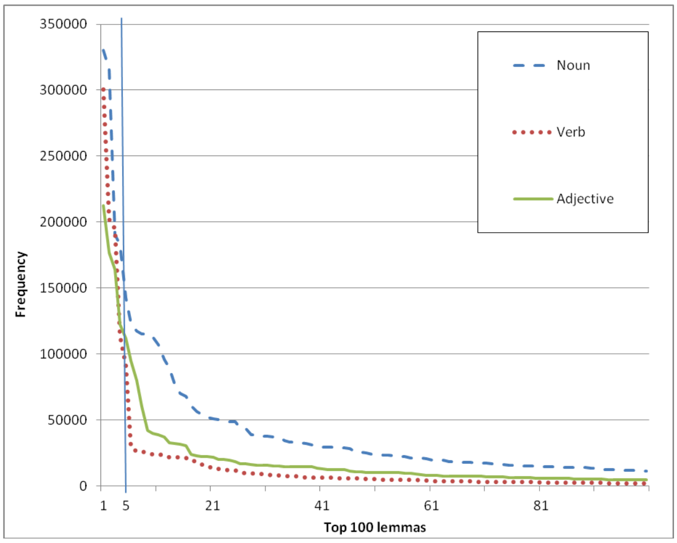

Figure 2- Frequency curves for the top 100 nouns, verbs and adjectives in the Stockholm EPR X-ray corpus

Content analysis

The structure of the radiology reports was usually coherent, providing a better orientation for the reader [22]: first a heading or first sentence describing the procedure and method used, and which body part was examined, thereafter a description of what is seen in the radiologic pictures produced during the examination, followed by an interpretation of the findings and their importance, with diagnostic reasoning. Administrative information, such as dates and names of physician(s) in charge would usually be at the end of the report.

The 100 most frequent sentences mainly contained information about findings, body parts, procedures and administration (Table 3). A vast majority of the mentioned findings were negated. In fact, the most common sentence in the corpus was *Ingen stas* (No stasis), from pulmonary X-rays. Negated findings were commonly expressed with simple negation cues: *ingen*, *inga*, *inget* (“no” in different inflections). Also, the most common sentences reporting on findings were about normal status, e.g. confirming normality of size for the ventricles of the brain.

Table 3- All words, sentences, bi-/trigrams; semantic categories from content analysis of top 100-lists. (nd= not determined)

	All words	**sentences**	**bigram**	**trigram**
abbrev.	18	12	38	29
admin.	**20**	16	**24**	**67**
definition	nd	4	1	4
method/ procedure	11	**25**	13	7
body part	8	**49**	**22**	10
position	7	0	11	1
finding	**18**	10	**21**	13
negated finding	nd	**61**	8	5
time	3	2	0	0
size	13	1	12	5

The top 100 bi- and trigrams were found to convey information about administrative phrases, body parts, findings, and procedures (Table 3). As all radiology reports contain the name of the examining radiologist as well as the senior radiologist contra-signing and taking the responsibility for the report, many bi- and trigrams contained person names (38 % and 70 % respectively). These were removed before the content analysis presented in Table 3.

Administrative words and phrases were common, present in 20 of top 100 words and 16 of top 100 sentences, and dominated the trigrams (67 of 100). Body parts were mentioned in about half of the top 100 sentences, and in 22 and 10 of the top 100 bigrams and trigrams, respectively.

Foreign words were of Latin, Greek and English origin. They were most commonly words for body parts, positions and procedures. Generally, words originating from Latin were used for body parts while English words were used for the names of methods and for radiologic equipment.

Miscellaneous information such as administrative routines, definitions of various grading scales, or technical descriptions of examinations or procedures, was present in 45 sentences of the top 100 most common.

Abbreviations were common, 18 of the top 100 words were abbreviated. Of these, 7 were common abbreviations (e.g. *tel* = telephone, *cm* =centimeter*)*, 10 were domain specific (e.g. *iv* = intravenous) and one ambiguous (*ca* for cancer or circa).

In Figure 3, the semantic categories are shown for different word classes. Most commonly, nouns were words for findings (*pleuravätska, fraktur*; pleural effusion, fracture), body parts (*hjärna, mjälte*; brain, spleen) and administrative words (*dokumentdatum, preliminärsvar*; date of document, preliminary report). The adjectives most often concerned descriptions about findings, such as positions (*vänster, dorsal*; left, dorsal) and size (*liten, lång*; small, long). Verbs belonged to different semantic categories, with verbs about findings (e.g. *bukta, påvisa*; bulge, detect) dominating. Of the top 100 verbs, 70 were found to be in active and 30 in passive voice.

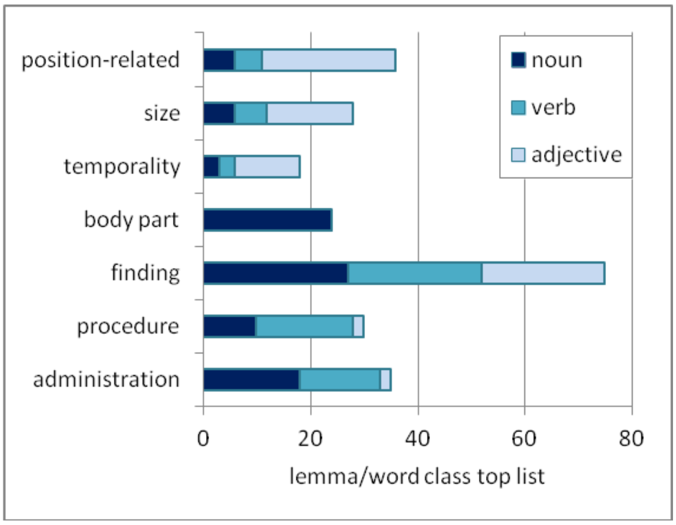

Figure 3 – Number of nouns, verbs, and adjectives in top 100-lists, according to semantic categories.

Discussion

We describe a large corpus of Swedish radiology reports. This corpus can be used to develop a text simplification tool for Swedish clinical text.

Historically, medical terminology origins from both Latin and Greek. The Latin terms generally denote body parts while Greek, the language of pathology, give rise to diagnostic terms as well as names for different medical specialties [12]. Nowadays, English expressions are gaining influence on Swedish health records since this is the vocabulary used in textbooks and medical journals. Many foreign words have been assimilated to the Swedish medical language by the addition of Swedish inflections, but will seem “Latin” to the layman reader and hence incomprehensible. Previous studies have shown differences in professional and layman vocabulary in the Swedish medical domain [23]. Also, the close resemblance of expressions in different languages, combined with Swedish inflections, results in diverging spellings which complicates automatic processing of clinical text.

Findings reported among the top 100 sentences in the Stockholm EPR X-ray corpus were dominated by reports of normal conditions and exclusions. Many negations were probably due to mandatory reporting of certain aspects, e.g., for X-ray of lungs to negate pulmonary infiltrates, for CT of brain to negate tumors or bleeding. This reflects something important: if the report conveys a new finding, this is not written in a standardized way. Instead, for each such situation, it is described in more varied ways. This has implications for a future text simplification system, as these more varied formulations probably convey more details that affect the patient directly.

Observations in this study confirm earlier findings that different medical terms can be used for the same pathology when excluding and reporting normality, or reporting specific findings [24]. For example, the word *fraktur* (fracture) was used to describe a pathological finding, but the more general term *skelettskada* (skeletal injury) was used when negating a fracture.

The content of the radiologic reports could to a large extent be classified into semantic categories, similarly to what has been found in other studies on clinical text [11]. There are several studies on automated tools for entity recognition for some of these categories, e.g. findings and body parts [25, 26]. Such tools could be used in conjunction with lexical exchange for text simplification.

Only 23 of the 100 most frequent sentences were complete, containing both a subject and a predicate. However, the majority of the short sentences contain an implicit subject and predicate, e.g. *Ingen stas* (No stasis) could be rewritten as a sentence such as “The radiology image (*subject*) shows (*predicate*) no stasis”.

Abbreviations were found to be of two kinds; abbreviations from general language and abbreviated medical terminology. The expansion of clinical abbreviations is not a trivial task and will require domain-adapted Natural Language Processing (NLP) tools, preferable context aware for disambiguation.

An important part of the radiology report is the concluding remarks with diagnostic speculation and reasoning, often intertwined with expressions for hedging and uncertainty. For text simplification, these parts need to be considered with great care since this poses special problems for layman comprehension [27].

Current state-of-the-art NLP tools are not tailored for this type of fragmented and information dense language that requires a lot of implicit knowledge. However, as has been shown in this study, the majority of the words convey a limited vocabulary and a large amount of recurrence, e.g. a small set of frequent adjectives. This has important implications: if the most frequent words, phrases and sentences are converted to more easy-to-read variants, a large proportion of the content is captured. Previous approaches for this problem include use of related expressions from hierarchical terminologies [16] or hyperlinked explanations from dictionaries [28]. The aim of the present project is to approach this not by introducing distracting fact boxes or choice of multiple lexical suggestions, but instead to produce a complementary simplified and coherent text adjacent to the original text.

We have limited this study to a description of the Stockholm EPR X-ray corpus. To deepen the understanding of these texts, we will conduct comparison studies with other text types. Furthermore, we are not reporting any conclusions on readability. This, and the level of required simplification, should be investigated with user studies.

Conclusion

We present a large resource of Swedish radiology reports. The qualitative analysis reveals that the most common words and expressions are about body parts, procedures, and findings, but also administrative issues. The study disclosed a set of recurring sentences and expressions, making up a considerate part of the corpus, implying that standard phrases can be identified and exchanged as part of a text simplification process.

Less easy to automatically process are the more varied descriptions of unique pathological findings.

The Stockholm EPR X-ray corpus can be used for development of text simplification and other NLP tools for health informatics purposes.

Acknowledgments

This work was supported by a grant from the Vårdal Foundation. The authors thank Martin Duneld for fruitful discussions .

References

[1] Archer N, Fevrier-Thomas U, Lokker C, McKibbon KA, and Straus SE. Personal health records: a scoping review. J Am Med Inform Assoc. 2011; 18(4): 515–522.

[2] Emont S. Measuring the impact of Patient Portals: What the literature tells us. Oakland: California HealthCare Foundation, 2011; pp. 1-19.

[3] Agerberg M. Journal på nätet ska bli del i hälsopaket. Läkartidningen. 2012; 109: 266-71.

[4] National eHealth – the strategy for accessible and secure information in health and social care. Ministry of Health and Social Affairs, 2010. www.sweden.gov.se/ehealth

[5] Centrum för eHälsa i samverkan (CeHis). Uppsalas e-tjänst Min journal återanvänds för hela landet. Nyhetsarkiv 2013-03-07. http://www.cehis.se/nyhetsarkiv/uppsalas_etjanst_min_journal_ateranvands_for_hela_landet/ (senast inloggad 2013-03-18)

[6] Keselman A, Slaughter L, Arnott Smith C, Kim H, Divita G, Browne A, Tsai C, and Zeng-Treitler Q. Towards Consumer-Friendly PHRs: Patients' Experience with Reviewing Their Health Records. In: AMIA Symposium Proc., 2007; pp. 399-403.

[7] Aantaa K. Mot patientvänligare epikriser - En kontrastiv undersökning. Dept. of Nordic languages, Turku University, Finland, 2012.

[8] Segall N, Saville JG, L'Engle P, Carlson B, Wright MC, Schulman K, and Tcheng JE. Usability Evaluation of a Personal Health Record. Proc. AMIA Annu Symp 2011; pp 1233–1242.

[9] Wibe T, Hellesø R, Slaughter L, and Ekstedt M. Lay people's experiences with reading their medical record. Soc Sci Med 2011; 72: 1570–3.

[10] Zeng-Treitler Q, Kim H, Goryachev S, Keselman A, Slaughter L, and Smith C. Text characteristics of clinical reports and their implications for the readability of personal health records. Studies in Health Technology & Informatics 2007: 129(Pt 2): 1117-21.

[11] Friedman C, Kra P, and Rzhetsky A. Two biomedical sublanguages: a description based on theories of Zellig Harris. J of Biomed Informatics 2002: 35: 222–235.

[12] Fogelberg M, Petersson G, and Nyman H. Medicinens språk. 2nd ed. Stockholm: Liber. 2006.

[13] Pyper C, Amery J, Watson M, and Crook C. Patient's Experiences when accessing their online electronic patient records in primary Care. Br J Gen Pract 2004; 498: 38–43.

[14] Keselman A, and Arnott Smith C. A classification of errors in lay comprehension of medical documents. J of Biomedical Informatics 2012; 45: 1151–1163.

[15] Keselman A, Logan R, Arnott Smith C, Leroy G, and Zeng-Treitler Q. Developing Informatics Tools and Strategies for Consumer-centered Health Communication. J Am Med Inform Assoc 2008: 15:4, 473-483.

[16] Kandula S, Curtis D, and Zeng-Treitler Q. A Semantic and Syntactic Text Simplification Tool for Health Content. In: Proc AMIA 2010: pp. 366-370.

[17] Leroy G, Endicott JE, Mouradi O, Kauchak D, and Just ML. Improving Percieved and Actual Text Difficulty for Health Information Consumers using Semi-Automated Methods. In: AMIA Annu Symp Proc, 2012; pp 522-31.

[18] Sarkar U, Karter AJ, Liu JY, Adler NE, Nguyen R, Lopéz A, and Schillinger D. The Literacy Divide: Health Literacy and the Use of an Internet-Based Patient Portal in an Integrated Health System—Results from the Diabetes Study of Northern California. J Health Com: Internat. Perspec. 15: 183–96. 2010.

[19] Dalianis H, Hassel M, Henriksson A, and Skeppstedt M. Stockholm EPR Corpus: A Clinical Database Used to Improve Health Care. Proc 4th Swedish Language Technol Conf, (SLTC-2012), 2012; pp. 17-18.

[20] Bird, S, Loper, E and Klein, E. Natural Language Processing with Python. O'Reilly Media Inc, 2009.

[21] Östling R. Stagger: A modern POS tagger for Swedish. In: Proc. 4th Swedish Language Technol Conf, 2012.

[22] von Heine A, and Wirell S. Röntgenremissen – dialog i flera dimensioner. Lund: Studentlitteratur, 2012.

[23] Kokkinakis D, and Toporowska Gronostaj M. Lay Language versus Professional Language within the Cardiovascular Subdomain - a Contrastive Study. In: Proc 2006 WSEAS Int. Conf. on Cellular & Molecular Biology, Biophysics & Bioengineering, Athens, Greece, 2006; pp1-7.

[24] Velupillai S, Dalianis H, and Kvist M. Factuality levels of diagnoses in Swedish medical text. In: Proc. MIE 2011. Moen A et al, eds. IOS Press, 2011; pp 559-563.

[25] Skeppstedt M, Kvist M and Dalianis H. Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text. In: Proc 8th LREC2012. 2012; pp 1250-1257.

[26] Wang Y, and Patrick J. Cascading classifiers for named entity recognition in clinical notes. In: Proc. Workshop on Biomed Info Extraction, 2009; pp. 42–49.

[27] Hanauer DA, Liu Y, Mei Q, Manion FJ, Balis UJ, and Zheng K. Hedging their Mets: The Use of Uncertainty Terms in Clinical Documents and its Potential Implications when Sharing the Documents with Patients. AMIA Annu Symp Proc, 2012; 321-330.

[28] Slaughter L, Oyri K, and Fosse E. Evaluation of a Hyperlinked Consumer Health Dictionary for reading EHR notes. Stud Health Technol Inform 2011; 169: 38-42.

Address for correspondence

Maria Kvist, MD, PhD, Dept. of Computer and Systems Sciences, Stockholm University, Forum 100, 164 40 Kista, Sweden. maria.kvist@karolinska.se