Nor Jnl Ling 37.2, 297–323 © Nordic Association of Linguists 2014 doi:10.1017/S0332586514000213

Smith, Kelly, Beata Megyesi, Sumithra Velupillai & Maria Kvist. 2014.
Professional language in Swedish clinical text: Linguistic characterization and comparative studies. *Nordic Journal of Linguistics* 37(2), 297–323.

Professional language in Swedish clinical text: Linguistic characterization and comparative studies

Kelly Smith, Beata Megyesi, Sumithra Velupillai & Maria Kvist

This study investigates the linguistic characteristics of Swedish clinical text in radiology reports and doctor's daily notes from electronic health records (EHRs) in comparison to general Swedish and biomedical journal text. We quantify linguistic features through a comparative register analysis to determine how the free text of EHRs differ from general and biomedical Swedish text in terms of lexical complexity, word and sentence composition, and common sentence structures. The linguistic features are extracted using state-of-the-art computational tools: a tokenizer, a part-of-speech tagger, and scripts for statistical analysis. Results show that technical terms and abbreviations are more frequent in clinical text, and lexical variance is low. Moreover, clinical text frequently omit subjects, verbs, and function words resulting in shorter sentences. Clinical text not only differs from general Swedish, but also internally, across its sub-domains, e.g. sentences lacking verbs are significantly more frequent in radiology reports. These results provide a foundation for future development of automatic methods for EHR simplification or clarification.

Keywords clinical text, comparative register analysis, doctor's daily notes, electronic health records, medical terminology, part-of-speech tagging, radiology reports

Kelly Smith, Department of Computer and Systems Sciences, Stockholm University, Postbox 7003, 164 07 Kista, Sweden. kellys@dsv.su.se
Beáta Megyesi, Department of Linguistics and Philology, Uppsala University, Postbox 635, S-751 26 Uppsala, Sweden. beata.megyesi@lingfil.uu.se
Sumithra Velupillai, Department of Computer and Systems Sciences, Stockholm University, Postbox 7003, 164 07 Kista, Sweden. sumithra@dsv.su.se
Maria Kvist, Department of Computer and Systems Sciences, Stockholm University, Postbox 7003, 164 07 Kista, Sweden & Department of Learning, Informatics, Management and Ethics (LIME), Karolinska Institutet, Sweden. maria.kvist@karolinska.se

1. INTRODUCTION

Patients want greater insight into their own healthcare process as well as the ability to influence it. While governments respond to the citizens' desire to have access to their own medical records, previous research has shown that patients

often have difficulties comprehending their electronic health records (EHRs). Health care personnel can even have difficulties understanding records written by other professions or specialists. Therefore, the need for automatic methods to simplify and/or clarify EHRs in order to aid layman comprehension becomes apparent. However, before automatic tools for simplification or clarification of a particular domain can be developed, it is essential that the linguistic features that differ from everyday Swedish language are identified and quantified through a linguistic analysis. By using off-the-shelf natural language processing (NLP) tools developed for general Swedish, we can quickly and consistently identify the linguistic characteristics of large texts. Moreover, by performing an error analysis of the tools, we can identify the specific characteristics of a particular domain that differ from general Swedish as well as make suggestions for the adaptation of tools for more accurate processing.

The purpose of this study is to identify and quantify the linguistic features that characterize Swedish clinical text, and compare them to biomedical journal text and general Swedish. Large data sets for each of these text types exist, and by analyzing them using automatic computational tools followed by an error analysis the linguistic characteristics which differ from general Swedish can be identified.

The linguistic characteristics investigated in our study are based in part on the findings of previous research, but also on the situational characteristics of the texts. We perform a register analysis based on the idea that the linguistic features of a text are linked to its situational characteristics and thus serve important communicative purposes (Biber & Conrad 2009). The features investigated include:

- word and sentence composition, e.g. length, type–token ratio, and hapax, dis, and tris legomena (hapax, dis and tris legomena are words that occur once, twice or thrice in a corpus)
- lexical complexity, e.g. vocabulary differences between corpora, and the use of technical terms and abbreviations
- sentence structures, e.g. parts of speech (POS) and POS sequences
- complexity as a combination of the above features, as expressed by some readability measures for general Swedish

On the basis of the situational characteristics of EHRs (which are more thoroughly presented in Section 2.1), e.g. their production circumstances, communicative purpose, and the active participants in their production and their relationships to one another, we hypothesize that EHRs:

- exhibit a greater amount of technical terms compared to the reference corpora due in part to the shared expert knowledge of medical personnel, as well as their need to be precise
- have a more similar vocabulary to the biomedical text than general Swedish

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

- are more telegraphic, containing more abbreviations, less function words, and shorter sentences due to being produced under time pressures

In addition, we use off-the-shelf state-of-the-art computational tools, e.g. a tokenizer to segment words and punctuation, and a part-of-speech tagger for the annotation of each token with their part of speech, in order to aid the text analysis. The long-term goal is that the findings can provide a basis for the ongoing and future development of automatic methods for simplifying or clarifying the free text of Swedish EHRs for laypeople.

2. BACKGROUND

2.1 Analyzing medical texts

Through the process of text analysis, many different linguistic characteristics can be identified, spanning from those which serve important communicative purposes, those that are conventionally associated with a given genre, to those solely serving an aesthetic purpose. Of interest in the present study is the first, which is known as register analysis (Biber & Conrad 2009). According to Biber & Conrad, the basis of a register analysis is a description of the situational use of a text or the purpose of the text variety, the linguistic features of the text, and the link between the two in the assumption that the specific linguistic features exist due to the situational circumstances and thus serve important communicative functions. For a register analysis to be effective different types of registers must be compared and by doing so an individual register's features are more apparent.

Medical text for professionals can be found either as biomedical text in scientific journals or as clinical text in patient records (Friedman, Kra & Rzhetsky 2002). They share a common vocabulary of medical terminology to convey information as precisely as possible, thus complex concepts can be described in professional technical terms without long explanations. However, the two text types vary as they are written for different purposes and under different conditions. While biomedical text is well phrased and formal with a scientific purpose of conveying knowledge, clinical text is often produced under time pressure as memory notes or as a means to provide information to other healthcare professionals. Medical terminology is in this case used for speed, as well as to make documentation precise and medically safe. The purpose of the EHR is to document and communicate a patient's health care over time concisely and effectively, and to ensure proper and secure patient care. For a physician, the use of a telegraphic style with many abbreviations is not only more rapid to write, but also presents a quicker overview on a shorter span of text.

The intended addressee of the EHR, while written about the patient, is primarily other health care personnel (Allvin 2010). In line with this, patients have found it

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

hard to understand health records, partly due to insufficient conceptual knowledge of the professional medical language (Pyper et al. 2004, Keselman et al. 2007, Aanta 2012).

There are differences between subdomains of clinical text, e.g. progression notes written on a daily basis for internal use and more formal parts of the EHR such as radiology reports, which are addressed to other physicians in different departments or hospitals.

2.2 Linguistic features found in clinical text

Previous researchers in Swedish and other languages have studied several linguistic features that characterize clinical text. The following EHR text example includes misspellings (underlined), abbreviations (bold), and words of foreign origin (italics) (Grigonyté et al. 2014:78).

> **Cirk** och **resp** stabil, ***pulm*** **ausk** något nedsatt **a-ljud** ***bilat***, *cor* **RR HF** 72, **sat** 91% på 4 **l O2**. Följer Miktionslissta. I samråd med <title> bakjour <First name> <Second name>, som bedömmer **pat** som komplicerad sjukdomsbild, så följer vi vitala parametrar, samt svara han ej på smärtlindring, så går vi vidare med ***CT*** **BÖS**.
>
> [**Circ** and **resp** stable, **pulm ausc** somewhat weak **resp** sound **bilat**, *cor* **RR HF** 72, **sat** 91% on 4 **l O2**. Following list for micturation. Consulting <title> senior **dr** on call <First name> <Second name>, who aseses **pat** as complicated condition, so we follow vital parameters, and anwers he not to pain-relief, so we go on to **CT ABD**.]

Several studies have shown that the heavy use of professional TECHNICAL TERMINOLOGY is prevalent in the free text of EHRs written in a number of languages. For example, a British study found that almost half of the patients who were presented with their own health records needed to consult a glossary to clarify terms (Pyper et al. 2004). Previous studies have found that between 6% and 10% of the Swedish EHRs studied consisted of technical terminology or jargon (Allvin 2010, Olsson 2011). In addition, terms can also be expressed through the use of synonyms (Allvin et al. 2011) and are frequently abbreviated or expressed as acronyms (Liu, Lussier & Friedman 2001, Skeppstedt, Kvist & Dalianis 2012).

ABBREVIATIONS AND ACRONYMS in EHRs are often ad hoc, created on the fly, and can be domain dependent (Liu et al. 2001, Xu, Stetson & Friedman 2007, Dalianis, Hassel & Velupillai 2009). Between 3% and 14% of words were found to be abbreviations in clinical text (Aramaki et al. 2009, Adnan, Warren & Orr 2010, Allvin 2010, Aantaa 2012, Skeppstedt et al. 2012). Abbreviations vary greatly; they do not necessarily follow any standard format and are often ambiguous as the same abbreviation or acronym can be used for a number of different words depending on

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

the context (Liu et al. 2001, Pakhomov, Pedersen & Chute 2005). An American study found that 33% of abbreviations in a medical terminology had multiple meanings (Liu et al. 2001). As an example, the abbreviation RA can express 25 different concepts, e.g. renal artery, right atrium, refractory anemia, radioactive, right arm, and rheumatoid arthritis (Pakhomov et al. 2005). Further, a word can be abbreviated in different ways, and an abbreviation or acronym can also have the same spelling as general words (Liu et al. 2001).

It is said that EHRs generally make use of PASSIVE VERBS to a greater extent than regular text (Kvist et al. 2011) and that these constructions present challenges to readers (Ownby 2005, Borin et al. 2009). Allvin (2010) found that the use of passive verbs was common in Swedish clinical text due to the fact that the patient often was not referred to explicitly, thus omitting the subject and using a passive verb. Passive verbs are common in Swedish discharge letters as physicians may write about procedures or exams that they are not performing themselves, or it is unclear or irrelevant to state who did, so they simply omit this information (Aantaa 2012).

The OMISSION OF SUBJECTS as well as the use of unclear subjects also has been found to be a common element of EHRs (Friedman et al. 2002, Aantaa 2012). The lack of subjects can lead to difficulties in interpreting who performs or experiences the actions of a given sentence. The patient is the main subject of EHRs, however, medical personnel can use the same type of subjectless sentences to describe their own actions, as in 'plans for invasive electrophysiology study', where plans is used as a verb. From a sample of ten Swedish EHRs, only 10% of the sentences had an explicit subject (Allvin 2010). Aantaa (2012) hypothesized that this could lead to problems in the comprehension of EHRs, but was not found to be a problem for Swedish and Finnish patients reading discharge letters.

The OMISSION OF VERBS is another linguistic phenomenon that has been found in English, Swedish, Finnish and German EHRs (Friedman et al. 2002, Aantaa 2012, Bretschneider, Zillner & Hammon 2013). Clinical documents generally describe a patient's condition with frequent use of nouns and adjectives, thereby omitting verbs with assumed low information content (Friedman et al. 2002). Also, sentences can be observed which only consist of one noun, for example 'Chills' in which it is assumed that the omitted verb is 'had'. These findings were confirmed in a study on German radiology reports (Bretschneider et al. 2013) where verbs were frequently omitted, as in the example 'In the lung, there are no effusions found' could typically be expressed as a subjectless and verbless expression 'Lung, no effusions'. The same trend was also observed in Swedish and Finnish EHRs (Aantaa 2012).

The use of FOREIGN WORDS is a prominent feature of both biomedical and clinical text, and words originating from Greek and Latin are prevalent in English, Swedish, Finnish, and German EHRs (Allvin et al. 2011, Bretschneider et al. 2013, Fan et al. 2013, Kvist et al. 2011). A study on German EHRs found that many Greek- and Latin-rooted words introduced unusual inflection forms, often used interchangeably

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

with the corresponding German word (Bretschneider et al. 2013) and the same is true for Swedish EHRs with many incorrect combinations of spellings as a result of the swedification of diagnosis expressions in 1987 (Smedby 1991).

It is important to note that the linguistic features that have been presented can all co-occur in clinical text. Unusual inflections and misspellings of words can be combined with less known abbreviations and expressions. For example, the word Noradrenalin was expressed in 60 different ways in Swedish intensive care unit records, and as much as 350 ways in Finnish (Allvin et al. 2011). All these features and the combinations of them can be assumed to make clinical text divergent from standard language.

2.3 Automatic analysis of clinical text

By using large amounts of textual data, empirically-based results can be obtained. NLP tools can be used to process these large amounts of text. Pre-processing steps for text analysis usually include sentence splitting and tokenization, which is the process of segmenting running text into words and sentences. Subsequent steps then depend on pre-processing being as accurate as possible. These can include spell-checking, POS tagging, i.e., the process of annotating words with their part of speech with or without morphological features, and parsing to produce a syntactic structure for a given input. Tomanek, Wermter & Hahn (2007) experimented with SENTENCE SPLITTING and TOKENIZATION on German clinical data and studied whether an in-domain training corpus improved results or not. They found that it is not critical for sentence splitting but that tokenization performance can be significantly improved.

Another NLP tool that achieves high accuracy on general language tasks is that of POS TAGGING. Accuracy for state-of-the-art POS taggers is around 97%. However when applying standard POS taggers to the clinical domain, this accuracy drops in part due to the high amount of specialized vocabulary. Coden et al. (2005) achieved a POS tagging accuracy of 87% when applying a general English language tagger to clinical data, and a more recent study reported a highest accuracy of 88.6% when applied to clinical data (Ferraro et al. 2013). On the other hand, Hahn & Wermter (2004) achieved surprisingly high accuracy rates thereby refuting previous claims by Campbell & Johnson (2001) that general language off-the-shelf taggers cannot be used on medical text without adaptation. They found that the statistical tagger TnT, when only trained on a German language newspaper corpus (NEGRA) and subsequently applied to a clinical corpus achieved an accuracy of 95.2%. When adapting the tagger by providing a small medical lexicon, the accuracy was raised to 96.7%, which is comparable to state-of-the-art taggers on newspaper text. When the tagger was trained only on clinical (in-domain) data, i.e. an annotated medical corpus of 90,000 words and extended tags for common lexical properties of the medical domain (e.g. enumerations, Latin forms in technical terms, and reference

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

patterns related to formal document structures), accuracy increased to 98.0%. The authors claim that the sublanguage of clinical data is simpler than that of newspaper language and is the most likely reason for the high accuracy.

There are other computational tools, such as syntactic parsers, which could be used for the automatic analysis of clinical text in order to reconstruct the syntactic structure of the sentences, named entity recognition for detection of technical terminology (Krauthammer & Nenadic 2004, Skeppstedt et al. 2012), or other strategies for identifying and expanding abbreviations (Xu et al. 2007, Isenius, Velupillai & Kvist 2012). However, such tools need to be thoroughly adapted to clinical text in order to produce reliable output, and were therefore not used in this study.

In conclusion, previous studies have discussed a number of linguistic features that permeate clinical text. These include the use of technical terms, abbreviations/acronyms, foreign words, omission of verbs and subjects, as well as the use of passive verbs. These features have in some cases been found to decrease the readability of EHRs for laymen readers and can also be problematic for NLP tools used for processing clinical text (Meystre et al. 2008). The same can be said for lower level pre-processing tasks, such as tokenization and POS tagging, which in general offer favorable results on standard language but fall short in the clinical domain.

3. METHODS

In order to examine the linguistic characteristics of Swedish clinical and biomedical text, and compare them to general Swedish, five corpora were compiled, automatically processed and analyzed. The tools for automatic analysis were evaluated and adapted. Linguistic features were quantified and compared across the data sets.

3.1 Data

Five different corpora were used. Two corpora of modern written Swedish were selected as reference sets for general Swedish. For formal, biomedical text, a portion of the Journal of the Swedish Medical Association was used. Two corpora provided the basis for the analysis of the clinical data: radiology reports and doctor's daily notes, as these two portions of English language EHRs are viewed as some of the most difficult for patients to comprehend (Keselman et al. 2007).

The Stockholm Umeå Corpus (SUC) is a balanced corpus consisting of Swedish text written in the early 1990s, totaling 1.1 million words. The first version of SUC was constructed in the 1990s, updated to version 2.0 in 2006, and to version 3.0 in 2012. The corpus is tokenized, each token is annotated with its lemma and annotated with POS and morphological features. The SUC 3.0 tag set consists of

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

153 tags, out of which 23 are POS tags and three delimiter tags (Östling 2013). The size of SUC 3.0 is appropriately large to use in corpus linguistics studies, the annotation is validated, and it is balanced and consists of modern Swedish, hence we chose SUC 3.0 as a reference corpus to compare the EHR corpora to.

PAROLE is a corpus totaling 19.4 million words (see http//spraakbanken.gu.se/parole/). The corpus consists of fiction, newspaper text, periodicals and web text produced between 1976 and1997. The corpus is annotated automatically with POS tags and morphological features following the SUC tag set. However, as the annotation of the corpus is not human validated, it will only be used for lexical comparisons.

In order to compare EHRs to formal and structured biomedical text, a portion of the SWEDISH BIOMEDICAL CORPUS (LTK) based on the electronic editions of the *Journal of the Swedish Medical Association* (*Läkartidningen*) was used (Kokkinakis 2012) as another reference corpus. Articles produced in 1996 were extracted consisting of 2,345 journal articles, 2,025,714 tokens, and 117,081 types.

The Stockholm Electronic Patient Record (EPR) Corpus (Dalianis et al. 2009, Dalianis et al. 2012) was developed at the Department of Computer and System Sciences at Stockholm University in collaboration with Karolinska University Hospital as a resource for clinical data research. It consists of over one million de-identified patient records from 2006–2010 for more than 600,000 patients within the greater Stockholm metropolitan area. The data was collected from over 500 healthcare units using the EHR system Take Care. The corpus consists of both structured information, e.g. gender and laboratory results, and unstructured information in free-text sections. Ethical approval was granted by the Regional Ethical Review Board in Stockholm, permission number 2012/2028-31/5.

Two subsets of the Stockholm EPR corpus were used in the current study, radiology reports (SEPR-X), and daily notes written by physicians (SEPR-DAY). The data, when obtained, had not been linguistically pre-processed in any way.

SEPR-X contains 434,427 radiology reports written in 2009–2010 for 152,170 unique patients and is comprised of referrals and the subsequent results of the radiology department's examinations (Kvist & Velupillai 2013). The present study only used the radiology results, which are written by radiologists, totaling 10,482,271 tokens and 118,980 types.

SEPR-DAY consists of 100,000 doctor's daily notes. For the present study, the free-text section called 'daganteckning' [daily notes] was used. This portion consists of 4,994,376 tokens and 130,107 types.

3.2 Tokenization and preprocessing

SUC, PAROLE, and LTK were already sentence segmented and tokenized when obtained. The two Stockholm EPR corpora were preprocessed with The Stockholm

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

Tagger (Stagger) (Östling 2013), a general Swedish POS tagger with built in sentence and word tokenization. SEPR-X contained a large amount of administrative sentences; these were removed prior to tokenization as they did not contribute to the medical content. Moreover, abbreviations containing punctuation resulted in incorrect word tokenization. Therefore, abbreviation definitions in Stagger's built in tokenizer were modified and extended in order to handle these domain-specific abbreviations.

3.3 POS tagging

In this study, Stagger was also used for POS tagging. It is based on the Averaged Perceptron algorithm with a reported accuracy of 96.6% when trained and tested on SUC 3.0 (Östling 2013). The tag-set used is shown in Table 1.

Tag	Part of speech	Example	Translation
AB	Adverb	inte	*not*
DT	Determiner	denna	*this*
HA	Interrogative/Relative adverb	när	*when*
HD	Interrogative/Relative determiner	vilken	*which*
HP	Interrogative/Relative pronoun	som	*who*
HS	Interrogative/Relative possessive	vars	*whose*
IE	Infinitive marker	att	*to*
IN	Interjection	ja	*yes*
JJ	Adjective	glad	*happy*
KN	Conjunction	och	*and*
NN	Noun	pudding	*pudding*
PC	Participle	dansande	*dancing*
PL	Particle	ut	*out*
PM	Proper noun	Mats	*Mats*
PN	Pronoun	hon	*she*
PP	Preposition	av	*of*
PS	Possessive pronoun	min	*mine*
RG	Cardinal number	tre	*three*
RO	Ordinal number	tredje	*third*
SN	Subjunction	att	*to*
UO	Foreign word	the	*the*
VA	Active verb	kasta	*throw*
VP	Passive or Deponential verb	kastas	(*is*) *thrown*
VB	Verb compound or Abbreviation	obs	*obs*(*erve*)
MID	Minor delimiter	.	
MAD	Major (sentence) delimiter	–	
PAD	Pairwise delimiter	(	

Table 1. The POS tag set used in the present study.

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

	Analysis1	Analysis2
SEPR-X	.830	.868
SEPR-DAY	.853	.873

Table 2. POS tagging error analyses 1 (original POS tag set) and 2 (adapted POS tag set) of the clinical corpora.

When applying Stagger to SEPR-X and SEPR-DAY, an error analysis was performed to determine how the tagger performed in general as well as on specific POS categories. The analysis was carried out on randomly selected record entries and consisted of 4,585 words (415 sentences) from the SEPR-DAY data and 4,610 words (397 sentences) from SEPR-X. The accuracy is shown in the column 'Analysis 1' for the original tag set and in the column 'Analysis 2' for the adapted tag set in Table 2, and the F-scores for each POS of the original and the adapted tag sets are shown in Table 3 below.

The tag set developed for general Swedish is not necessarily optimal for clinical text. Two particular tags were often found to be erroneous: proper nouns and foreign words, as these tags were frequently applied to technical terms. The distinction between some POS categories and the handling of multi-word expressions was also problematic. In some cases more than one POS category was applicable, for example the EHR system's name 'Take Care' which is both a foreign word and a proper noun.

Following the first error analysis, two POS categories were changed:

- Foreign words (UO) were retagged as nouns (NN), aside from a small stop list, as the detection and labeling of the foreign words in the clinical data sets was often erroneous (low F-scores in Table 3), and most of the foreign words were nouns.
- Proper nouns (PM) were changed to nouns (NN) for two reasons: some medical expressions, such as medications, could be either proper nouns or nouns, and inconsistent tagging of the different portions of multi-word proper nouns was common.

The training corpus (SUC) was adjusted and Stagger retrained on it. This new model was used to retag the clinical corpora, as well as LTK. Results from the final error analysis are shown in the column labeled 'Analysis 2' in Table 3. Not surprisingly, accuracy increased from 83.0% to 86.8% (SEPR-X) and from 85.3% to 87.3% (SEPR-DAY) after the adjustments were made to the tag set (Table 2). This can very likely be attributed to making the tag set smaller by removing tags. As can be seen in Table 2 some F-scores decreased in the second error analysis. There are three main reasons for these results. First, in a number of these situations the decrease is very

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

	SEPR-X				SEPR-DAY			
	Analysis 1		Analysis 2		Analysis 1		Analysis 2	
POS tag	Total	F-score	Total	F-score	Total	F-score	Total	F-score
NN	952	.91	1425	.97	1148	.91	1414	.95
VA	134	.96	151	.96	467	.95	418	.94
VP	57	1.0	52	.97	85	.94	79	.94
PP	308	.99	293	.98	491	.98	505	.98
AB	148	.9	150	.89	349	.93	310	.94
PN	34	.97	30	.94	100	1.0	72	1.0
JJ	360	.83	335	.89	284	.82	270	.84
MAD	442	1.0	429	1.0	432	1.0	433	1.0
MID	469	.99	481	1.0	259	.99	340	.99
KN	98	.99	97	.98	176	.99	149	.99
DT	124	.99	129	.98	86	.98	91	.99
PM	349	.93	0	NA	179	.79	0	NA
SN	6	1.0	5	1.0	37	.94	37	1.0
PC	69	.86	77	.94	124	.96	83	.93
HP	11	.96	16	.96	41	.99	27	.93
PL	5	.89	8	.89	30	1.0	35	.94
RG	759	.99	756	1.0	166	.97	231	.98
PS	5	.91	3	.91	9	.95	3	.85
IE	5	.50	4	.50	30	.98	18	1.0
HA	8	1.0	2	1.0	16	.97	11	.96
PAD	86	.74	50	1.0	28	1.0	36	.56
UO	26	.47	0	NA	8	.40	0	NA
IN	1	1.0	0	NA	3	.75	2	.57
RO	1	1.0	2	1.0	7	.83	3	.86
HD	0	NA	0	NA	0	NA	0	NA
HS	0	NA	0	NA	0	NA	0	NA

Table 3. POS tagging error analyses 1 (original POS tag set) and 2 (adapted POS tag set) of each clinical corpus (SEPR-X, SEPR-DAY) showing the number of each POS and the achieved F-score.

small and most likely insignificant (e.g. PP). Second, in some instances the amount of a given POS tag is less than in the first analysis in which case an F-score decrease is not strange. Lastly, the error analysis was only performed by one person, thus there might be errors in the human judgment.

3.4 Identifying technical terms and abbreviations

In order to approximate the prevalence of abbreviations and technical terms in the clinical corpora compared to the reference corpora, a small portion of randomly selected sentences from each of the four corpora was manually annotated. Technical

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

terms were those words that were considered to have a specific meaning within a specific field of expertise. Many technical terms were abbreviated, and in this case were annotated only as abbreviations. A computational linguist performed the annotation.

3.5 POS sequence analysis

Since the corpora were not syntactically parsed, the prevalence of missing subjects was determined using a different approach. The 100 most common sentential part-of-speech sequences were automatically extracted from each corpus and thereafter analyzed manually. For each sequence, ten examples from each corpus were extracted in order to simplify the comprehension of these abstract POS sequences. If the majority of the examples for a given POS sequence were found to be incorrectly tokenized or tagged, the sequence was classified as incorrect. Some sequences were found to only consist of metadata such as information on the author and publication date, e.g. 'Stockholm: Carlsson Bokförlag, 1995' (an example from LTK). Sequences not found to adhere to any of these categories were assigned to the category with subject or without subject.

3.6 Feature combinations

Given the frequency of words and certain POS tags as well as word and sentence lengths, we can measure the complexity of clinical data by using some common readability metrics for Swedish. In particular we use those that measure the relationship between word and sentence lengths (LIX), nominal versus verbal style (NR), and lexical variation (OVIX) (Smith, Danielsson & Jönsson 2012).

LIX, which stands for *läsbarhetsindex* [readability index], is the Swedish standard for approximating how difficult a text is (Melin 2004) and is presented in Equation (i). The difficulty levels and examples of text genres for the range of LIX values are shown in Table 4 (Mühlenbock & Johansson Kokkinakis 2009).

$$\mathrm{LIX} = \frac{n(\text{words})}{n(\text{sentences})} + \frac{n(\text{words} > 6 \text{ characters})}{n(\text{words})} \times 100 \qquad \text{(i)}$$

LIX value	Difficulty level and genre
< 30	Very easy, children's books
30–40	Easy, normal text/fiction
40–50	Medium-difficult, informative text/newspapers
50–60	Difficult, specialist literature
> 60	Very difficult, research

Table 4. LIX values and their interpretation (from Mühlenbock & Johansson Kokkinakis 2009).

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

OVIX, which stands for *ordvariationsindex* [word variation index], is a popular metric for measuring lexical variation. It is expressed as the ratio of the number of types, i.e. unique words, and the total amount of words (Mühlenbock & Johansson Kokkinakis 2009). Equation 2 is used to calculate OVIX values. For reference, Mühlenbock & Johansson Kokkinakis (2009) present OVIX scores ranging from 60 to 69 for various texts from LäsBarT, an easy-to-read corpus.

$$\text{OVIX} = \frac{\log n(\text{words})}{\log\left(\frac{\log n(\text{types})}{\log n(\text{words})}\right)} \tag{ii}$$

NOMINAL RATIO (NR) is the quotient of the total amount of nouns, prepositions, and participles by the total amount of verbs, adverbs, and pronouns (Equation 3), and measures a text's nominal style, which usually implies that a text is more dense in information (Mühlenbock & Johansson Kokkinakis 2009). For a 'normal' text the quotient is 1. The more information rich a text is, the higher the NR tends to be, and the more sparse, the lower the NR.

$$\text{NR} = \frac{n(\text{nouns}) + n(\text{prepositions}) + n(\text{participles})}{n(\text{pronouns}) + n(\text{adverbs}) + n(\text{verbs})} \tag{iii}$$

4. RESULTS

Different genres and subgenres of medical text were characterized. SUC and PAROLE represent standard Swedish, LTK represents biomedical text, and SEPR-X and SEPR-DAY, consisting of radiology reports and doctor's daily notes, represent subgenres of clinical text. It was found that the clinical data exhibits a lower lexical variance than the reference corpora. Short words, short sentences, technical terms, and abbreviations were also found to be more common. The clinical corpora also make less frequent use of verbs, function words and subjects. Differences between the two clinical corpora were also found.

4.1 General corpus statistics

SUC consists of just over one million tokens, LTK approximately two million tokens, SEPR-DAY almost five million tokens, and SEPR-X is double this at ten million (Table 5). SEPR-X has the lowest type-token ratio. SEPR-DAY had only a slightly higher ratio followed by LTK with twice as high as SEPR-X and SUC the highest. SUC has the highest amount of hapax, dis, and tris legomena, followed by LTK, then SEPR-DAY, and finally SEPR-X. The amount of punctuation is higher in the clinical data, compared to both SUC and LTK.

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

	SUC	LTK	SEPR-X	SEPR-DAY
No. tokens	1,166,121	2,025,714	9,527,807	4,754,437
No. types	97,124	117,081	203,611	154,205
Type–token ratio	8.33%	5.77%	2.13%	3.24%
No. sentences	74,163	118,542	1,112,581	471,290
% long words	3.0%	5.1%	3.2%	2.8%
% words	88.51%	89.29%	85%	84.78%
% punctuation	11.49%	10.71%	15%	15.22%
Average token length	4.79	5.31	4.80	4.61
Average sentence length	15.72	17.89	8.56	10.09
Hapax legomena	4.7%	3.1%	1.2%	1.9%
Dis legomena	1.2%	0.9%	0.3%	0.4%
Tris legomena	0.6%	0.4%	0.1%	0.2%

Table 5. General corpus statistics. Long words are all words longer than 13 characters. % words is percentage of all tokens not tagged with a delimiter tag. Hapax, dis, and tris legomena are represented as percent of word tokens that occur only once, twice, and thrice, respectively.

4.2 Differences in token and sentence length

Table 5 also presents the average token versus sentence length of the corpora. LTK has the highest average token length, followed by SEPR-X and SUC, and finally SEPR-DAY. However, these averages are very close. Figure 1 more explicitly shows the differences in token length. Tokens consisting of two characters are more infrequent than those of one or three characters.

The amount of words longer than 13 characters were measured. Surprisingly, SEPR-X contained a slightly greater number than SUC and SEPR-DAY. LTK on the other hand contained the most.

The average tokens per sentence vary more between the four corpora (Table 5). Using SUC as a reference point, it is not surprising that LTK, with more formal and technical text, has a slightly higher average sentence length. Both clinical corpora have lower average sentence lengths than SUC. The frequency distribution of tokens per sentence for each of the corpora shows that SUC and LTK follow the same type of pattern, while the two Stockholm EPR corpora are to a great extent concentrated at the lower end of the spectrum (Figure 2). Approximately 70% of SEPR-X and 63% of SEPR-DAY consist of sentences shorter than 11 words. This is compared to approximately 35% for SUC and 29% for LTK.

As with short words, it is not a surprise that short sentences are more common in the clinical data, since they are known to often omit subjects, verbs, and words that can be considered low in information content. The use of closed word classes, such as pronouns (PN), determiners (DT), and conjunctions (KN), in the Stockholm EPR corpora is much lower than in SUC and LTK (Table 8 below).

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

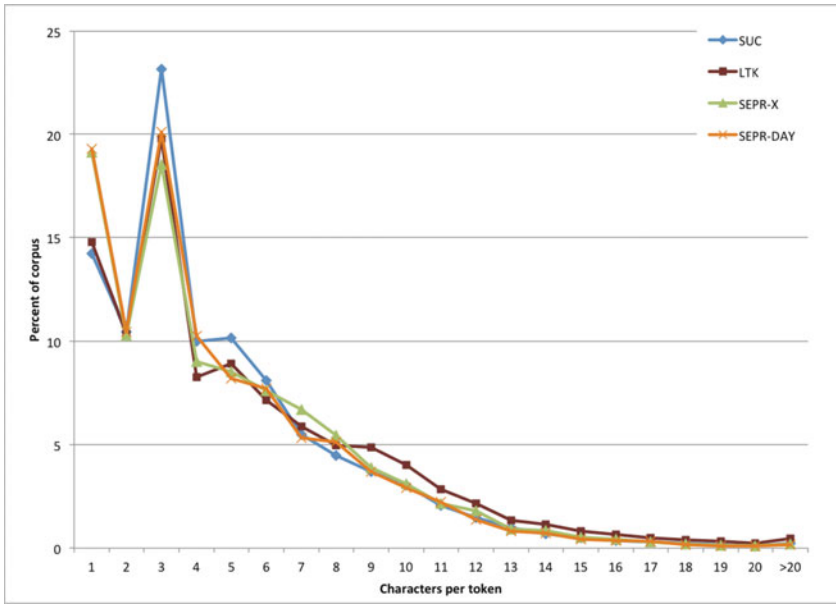

Figure 1. (Colour online) The frequency distribution of the number of characters per token in each corpus (SUC, LTK, SEPR-X, SEPR-DAY).

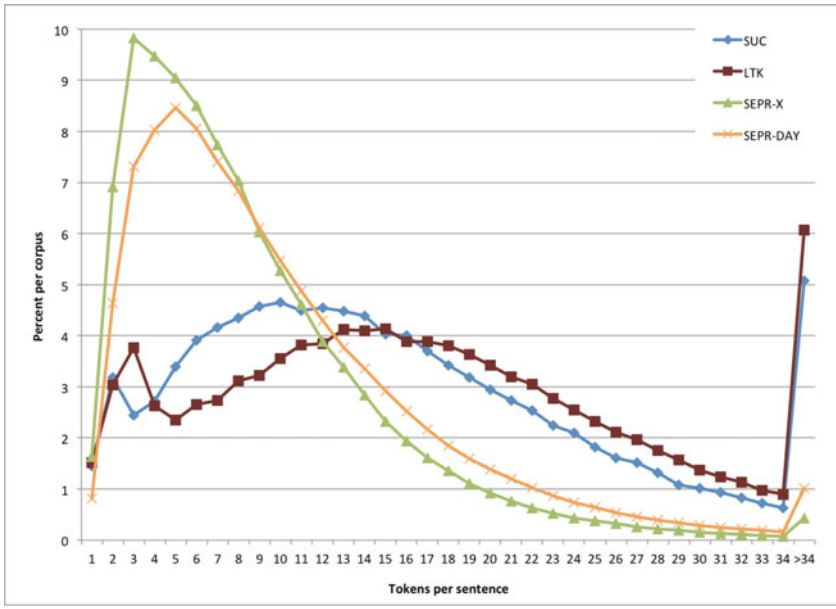

Figure 2. (Colour online) The frequency distribution of the number of tokens per sentence in each corpus (SUC, LTK, SEPR-X, SEPR-DAY).

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

	SUC	LTK	SEPR-X	SEPR-DAY
% in Parole	74.3	53.1	20.2	26.8
% in SUC	NA	29.2	10.6	14.4
% in LTK	35.2	NA	16.3	22.4

Table 6. Percentage of word types from each corpus found in Parole, SUC, and LTK, respectively.

4.3 Vocabulary

A comparison of the amount of word types in LTK, SEPR-X, and SEPR-DAY, which are also found in SUC or PAROLE, approximates how much non-standard vocabulary is found in each type of text (Table 6). Only 30% of the word types in LTK are also found in SUC. However, in the clinical data the proportion is even smaller. Not surprisingly, when comparing to PAROLE, which is approximately 18 times larger than SUC, these numbers increase. Around half of the vocabulary of LTK matches that of PAROLE, compared to only about a fifth and a fourth of SEPR-X and SEPR-DAY, respectively. When comparing SUC, SEPR-X, and SEPR-DAY to LTK on the other hand, 35%, 16%, and 22% for each respective corpus match.

A majority of the word types in the Stockholm EPR corpora were not found in SUC or PAROLE. Furthermore, the results also indicate that the lexical variance of these two corpora is lower than that of SUC.

The type–token ratio of SUC is the highest (Table 5 above), as expected since SUC includes different text types, e.g. news or novels. The type–token ratio is lower for the more concentrated type of media, the biomedical journal LTK. The clinical text exhibit a very low type–token ratio, especially SEPR-X with only a fourth of that of SUC.

The same example of low versus high lexical variance can be seen in the amount of hapax legomena, i.e. the number of words that occur only one time in a corpus. SUC has a higher number of hapax legomena than in any of the other corpora (Table 5). LTK is in second place, and then SEPR-DAY followed by SEPR-X.

4.4 Technical terms and abbreviations

A small portion of each corpus was manually annotated for abbreviations and technical terms during the error analysis (Table 7). As expected, SUC contains the least amount of abbreviations. LTK, as expected for biomedical text, contained a slightly higher number of abbreviations. However, LTK was surpassed by the clinical corpora, especially by SEPR-DAY.

During the annotation of abbreviations, technical terms were also annotated. The SUC and LTK samples both contained a very low number of technical terms

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

	No. words	Technical terms	Abbreviations
SUC	5,565	0.4%	1.0%
LTK	5,074	0.9%	3.0%
SEPR-X	4,673	2.3%	8.1%
SEPR-DAY	4,593	7.7%	7.8%

Table 7. Number of words from each corpus analyzed in the annotation of technical terms and abbreviations, and the percentage found.

(Table 7). SUC was made up of only 1% technical terms and LTK only had 2% more. On the other hand, the clinical corpora made much more frequent use of technical terms, approximately 8%.

4.5 Part of speech frequencies

Figure 3 shows the percentages of each POS in the four corpora that make up more than 1% of the corpora (see the tag set interpretation in Table 1 above). Nouns rank as number one in terms of frequency among all four corpora (Figure 3). They are however more prevalent in LTK, SEPR-DAY, and especially in SEPR-X, than in SUC. The second most common POS in SUC is that of active verbs. In this case the amount is much lower in the clinical corpora and LTK, compared with SUC. At the top of Figure 3 the total amount of verbs (passive plus active) are shown, and SUC contains nearly double the amount of verbs than the other three corpora do. However, while LTK may contain the same percentage of verbs as the clinical corpora (Table 5 above), the clinical corpora exhibit a much higher prevalence of verbless sentences (Figure 4). Between 11% and 16% of the sentences in SUC and LTK lack verbs, compared to the high frequency in the clinical text, at 43% in SEPR-DAY and 63% in SEPR-X. The amount of passive verbs is nearly the same in SUC and SEPR-DAY (Figure 5), and passive verbs are slightly more common in the more formal texts of LTK and SEPR-X. Other noticeable differences in POS frequencies between the corpora are the low number of pronouns and determiners in the clinical corpora and LTK when compared to SUC. On the other hand, SUC exhibits a much lower number of cardinal numbers and participles than the other corpora. The clinical corpora differ slightly from SUC and LTK in terms of making more frequent use of adjectives, and making less frequent use of functions words such as conjunctions, relative pronouns, and relative adverbs.

In summary, the POS statistics for the clinical corpora and LTK are similar to one another and differ most from SUC in a number of categories. The most noticeable differences are nouns, verbs, cardinal numbers, participles, determiners, and pronouns.

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

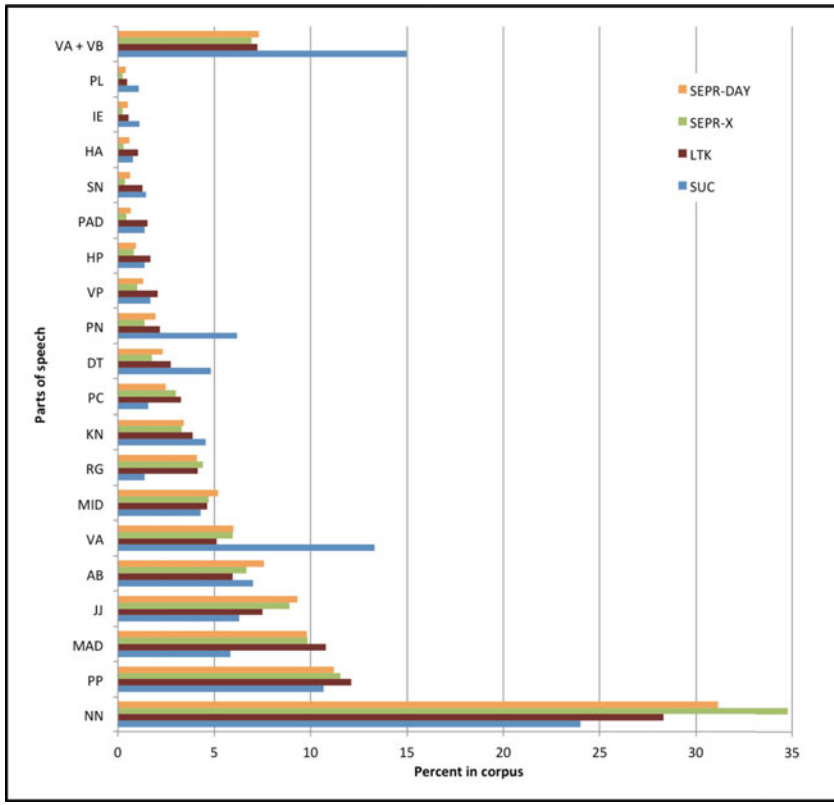

Figure 3. (Colour online) The percentage of each part of speech (only those over 1%) in the four corpora (SUC, LTK, SEPR-X, SEPR-DAY). At the top, the total number of verbs is calculated as the sum of active (VA) and passive (VP) verbs.

4.6 POS sequence analysis

In order to gain an insight into the most common sentence structures of the clinical data, as well as to give an approximation of how prevalent missing subjects were, the 100 most common POS sequences in the corpora were automatically extracted and then manually analyzed.

The top five most common POS sequences for each corpus are shown in Table 8 below. Table 9 shows the proportion of each corpus that the sequences made up. This table shows that the two clinical corpora are more repetitive in the use of the same sequence types. By only investigating the top 100 most common sequences, 19% of SEPR-DAY and 26% of SEPR-X, versus only 8% and 9% in SUC and LTK respectively was observed.

By referring to examples of the POS sequences from the data it was found that some sequences were incorrectly tagged or tokenized. A number of sequences from

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

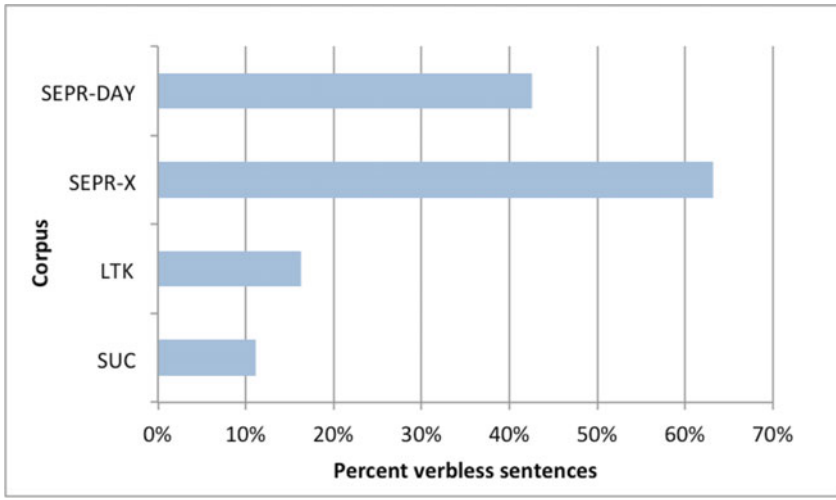

Figure 4. (Colour online) The percentage of verbless sentences in each corpus (SUC, LTK, SEPR-X, SEPR-DAY).

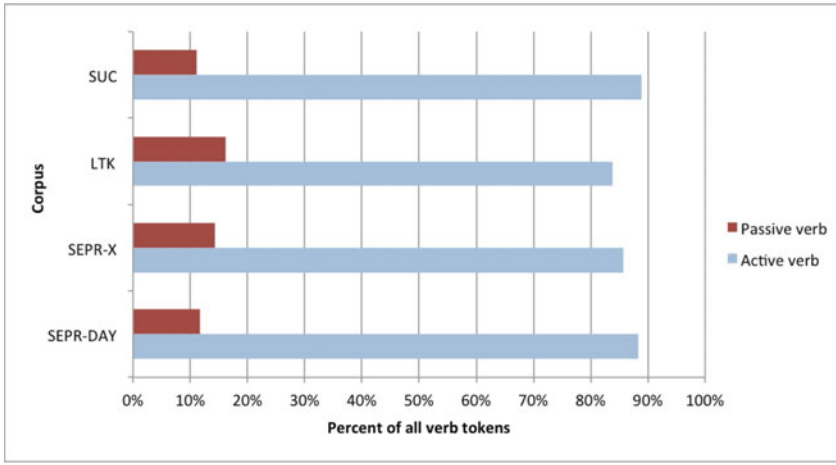

Figure 5. (Colour online) The ratio of passive to active verbs in each corpus (SUC, LTK, SEPR-X, SEPR-DAY).

LTK and SUC consisted of article metadata, such as the name and profession of the article's author, figure references, or citations. Thus, these types of sequences were not analyzed with regards to whether or not they contained or omitted subjects, and are represented in Table 8 as *meta* and *incorrect*.

SEPR-DAY and SEPR-X contained the highest number of sequences without a subject (Table 9), approximately four times the amount in SUC and LTK. Since only

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

	POS sequence	Category
SUC	NN	Subject
	NN,NN	Meta
	RG,NN	Subject
	RG	Meta
	JJ,NN	Subject
LTK	NN	Meta
	NN,RG,MAD	Meta
	JJ,NN	Subject
	RG,MAD	Meta
	NN,PP,NN	Subject
SEPR-X	NN,MAD	Subject
	JJ,NN,MAD	Subject
	NN	Incorrect
	NN,NN,NN	Meta
	NN,NN	Meta
SEPR-DAY	NN,MAD	Subject
	JJ,NN,MAD	Subject
	NN,PP,NN,MAD	Subject
	DT,NN,MAD	No subject
	NN,RG,MAD	Subject

Table 8. The five most common POS sequences from each corpus (SUC, LTK, SEPR-X, SEPR-DAY) along with sequence categories: Subject = contained a subject, Incorrect = incorrectly tokenized, Meta = consisted of metadata.

		Subject	No subject	Meta	Incorrect	Total
SUC	No. seqs	71	12	13	4	100
	% of corpus	4.94%	0.43%	2.41%	0	7.78%
LTK	No. seqs	53	9	29	9	100
	% of corpus	3.90%	0.27%	4.79%	0.36%	9.33%
SEPR-X	No. seqs	37	42	17	4	100
	% of corpus	11.22%	7.32%	5.09%	2.45%	26.05%
SEPR-DAY	No. seqs	51	41	3	5	100
	% of corpus	9.77%	7.7%	0.64%	1.29%	19.40%

Subject = contained a subject, Incorrect = incorrectly tokenized, Meta = consisted of metadata

Table 9. *No. seqs* shows the total number of sequences out of the top 100 POS sequences from each corpus (SUC, LTK, SEPR-X, SEPR-DAY) containing or omitting subjects, as well as those consisting of meta data or incorrectly tagged/tokenized. *% of corpus* shows the proportion of these categories, as the top 100 sequences of each entire corpus.

the top 100 most common sequences of the corpora were analyzed, we may conclude that between 7% and 8% of the clinical corpora consist of subjectless sentences. The subjectless constructions were often statements of negative findings, the most simple being <DT,NN,MAD>, for example 'Inga infektionstecken' [No infection signs].

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

	SUC	LTK	SEPR-X	SEPR-DAY
OVIX	88	83	76	78
LIX	41	50	39	37
NR	1.29	1.93	3.28	2.13

Table 10. The three readability metrics (OVIX, LIX, NR) for each corpus (SUC, LTK, SEPR-X, SEPR-DAY).

Other common types of subjectless sequences were those indicating an action, but not stating who performed it, for example 'Dubblerar Oxycontin' ['Doubles Oxycontin], i.e. the doctor doubles the dose of the drug Oxycontin.

4.8 Feature combinations

As for the combination of various linguistic features, the readability metrics LIX, OVIX, and NR for each of the corpora are reported in Table 10. SUC has a LIX value of 41 and can be used as a reference point. It has a LIX value in between those that are common for text types considered 'easy' and 'medium-difficult'. LTK has a value of 50, the rank common for 'difficult or specialist literature'. However, the LIX values for the clinical corpora are lower than SUC (35 and 38 for SEPR-DAY and SEPR-X respectively). These results are in line with the results on word and sentence length (short words and sentences are more common in the clinical corpora).

The OVIX values indicate that SUC has the highest amount of lexical variation, followed by LTK, SEPR-DAY and finally SEPR-X. This correlates with other findings on lexical variation, i.e. a low type–token ratio and a small number of hapax, dis, and tris legomena in the clinical corpora compared to LTK and SUC (Table 5 above).

NR measures nominal versus verbal style of a text. A nominal style can indicate that a text is denser in information. Using SUC, with an NR value of 1.29 as a reference point, the nominal ratio of each text subsequently rises. LTK and SEPR-DAY have a slightly higher NR, while SEPR-X is around 2.5 times higher than SUC. This is due to the higher amount of nouns, participles, and prepositions and a lower amount of verbs and pronouns (the POS proportions of each corpora can be found in Figure 3 above).

5. DISCUSSION

The purpose of this study was to characterize the Swedish clinical language in EHRs as compared to standard Swedish and formal biomedical text. Are the clinical corpora different from SUC and LTK in terms of lexical complexity, word and sentence

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

composition and sentence structures? The results show that in practically all of the linguistic features studied, this is the case.

In terms of word and sentence composition, the results indicate that the clinical corpora make more frequent use of short words and sentences, which was expected. The fact that clinical data makes frequent use of short sentences is in line with previous research on EHRs in a number of languages which have discussed their telegraphic nature (Friedman et al. 2002, Adnan et al. 2010, Kvist et al. 2011, Aantaa 2012). This can be due to the omission of information that can be considered unnecessary, such as verbs and subjects and especially low use of function words.

Lexical variance was low for the clinical corpora, as exhibited by a much lower type–token ratio; hapax, dis, and tris legomena; as well as having a low OVIX score. SEPR-X showed the lowest amount of word types and the lowest variance in all these linguistic features. A different study of Swedish radiology reports found the same type of results in a set of recurring sentences and expressions which made up a generous portion of the corpus (Kvist & Velupillai 2013).

Unknown words have repeatedly been named as a prominent characteristic of the free text in EHRs and can consist of technical terms, misspellings, and abbreviations (Liu et al. 2001, Adnan et al. 2010, Allvin 2010, Patrick et al. 2010, Kvist et al. 2011, Bretschneider et al. 2013,). This was also found to be true in the present study. Technical terms and abbreviations were more prevalent in the clinical text than the other two text types. We also found that the EHRs shared less than a fourth of the words in Parole. This is in line with English language EHRs where about 30% of the words were unknown (Patrick et al. 2010). More interesting was the finding that the clinical corpora did not share the vocabulary of LTK to any greater extent, although this could in part depend on the smaller size of the corpus. This is also observed in the comparison of the clinical corpora with SUC.

Compared to SUC, the clinical text made less use of verbs and function words, and slightly more frequent use of passive verbs, which is in line with previous research (Friedman et al. 2002, Bretschneider et al. 2013). The POS sequence analysis, giving an approximation of the prevalence of omitted subjects, indicated that this is more common in the EHRs. The finding that the clinical text, especially radiology reports, have more verbless sentences than general Swedish is in line with findings of German radiology reports (Bretschneider et al. 2013).

Also of interest was how the corpora differed in complexity as measured by some readability metrics often applied to Swedish. We have only applied basic, existing measures for readability expressed as a combination of linguistic features observed in this study. The results indicated that the clinical corpora used a more nominal style than SUC and LTK as indicated by nominal ratio (NR), and that the lexical variance of the clinical corpora was lower than SUC and LTK (OVIX). The results for LIX are not consistent with the results of OVIX and NR. Word and sentence length cannot solely be used to measure the complexity of a text. In order to investigate the readability of a

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

text, further studies are necessary to identify what people find difficult to understand and how the text might be simplified or clarified for better layman understanding.

An 'off-the-shelf' tokenizer and POS tagger could be used without adaptation on the clinical data with an accuracy of 87%, compared to 97% on general Swedish. Thus domain adaptation should be performed in order to obtain higher accuracy. Previous research has shown that by adding a small in-domain training corpus, tagging accuracy can be greatly increased (Coden et al. 2005). In addition, by slightly modifying the tag set, e.g. tagging foreign words and proper nouns as nouns, tagging accuracy was improved in the present study. As a gold standard of approximately 10,000 words from clinical text was created, it would be interesting to use this in order to adapt Stagger and further improve results. Spell-checking and syntactic parsing would also be beneficial for future studies on clinical text, but must first be adapted to the clinical domain.

6. SUMMARY AND CONCLUSION

The ability to access your own EHR online is part of a growing trend in Sweden and abroad and is already in place for millions of patients. Easy access does not, however, ensure that the patient will understand what has been written about their health care. Previous research has found that EHRs are filled with linguistic features that are difficult to comprehend without background knowledge. In order to develop automatic methods for EHR simplification, details on the characteristic linguistic features that deviate from general Swedish must be available for this to be successful.

The present study used a comparative register analysis in order to determine the prevalence of a number of linguistic features. Five corpora were studied, two of which represented clinical text in the form of radiology results and doctor's daily notes. The other corpora represented formal, biomedical journal text and general Swedish respectively. By comparing the results of the characterization of each corpus, it was found that clinical text differed in all aspects to varying degrees in terms of word and sentence composition, lexical complexity, and common sentence structures. Compared to standard Swedish, free text in EHRs make more frequent use of short words and sentences, abbreviations, and technical terms, and make less frequent use of subjects, verbs, and function words. The lexical variation of the clinical text, especially for radiology reports, is significantly lower than standard Swedish as well as the biomedical domain. These linguistic features can be directly linked to the situational circumstances of the text. In addition, various standard Swedish measurements of readability (LIX, OVIX, and NR) were applied to the free text in EHRs. The OVIX metric indicates that the clinical text has a lower lexical variance than SUC and LTK. The clinical text also has a more pronounced nominal style than SUC and LTK as indicated by NR. Considering only the length of words and

https://doi.org/10.1017/S0332586514000213 Published online by Cambridge University Press

sentences as measured by LIX does not seem to be a sufficient indicator of complexity and needs further investigation.

In order to be able to automatically process large amounts of clinical data for quantitative, reliable linguistic analysis, we explored the use of 'off-the-shelf', state-of-the-art computational tools to tokenize the texts and annotate the words with their appropriate POS without adapting them to the specific clinical domains. The tokenizer and POS tagger developed for standard Swedish was applied to the clinical data and was partly found to be insufficient for handling their complexities. A small-scale tagger adaptation slightly improved tagging results, as well as produced a small gold standard for future work on POS tagging of clinical texts. The results clearly show that, although the available computational tools for language processing can be used, adaptation to the specific domain is necessary for better performance.

By quantifying the linguistic features of the EHRs and comparing them to standard Swedish, a foundation has been laid for various types of future work on clinical data. The long-term goal is that our findings can provide a basis for the ongoing and future development of automatic methods for simplifying or clarifying Swedish EHRs. Lastly, our belief is that the methods and tools applied to analyze the domain of clinical data can be directly applicable to other genre or domain analysis and be helpful for linguists interested in text analysis.

ACKNOWLEDGEMENTS

The authors wish to thank the anonymous *Nordic Journal of Linguistics* reviewers for valuable feedback. This work was supported by funding from the Vårdal Foundation. We are grateful to Hercules Dalianis for the initiative of Stockholm EPR Corpus, as well as Martin Duneld and Maria Skeppstedt for fruitful discussions and technical assistance.

REFERENCES

Aantaa, Kirsi. 2012. Mot patientvänligare epikriser. En kontrastiv undersökning [Towards more patient friendly discharge letters: A contrastive study]. MA thesis, Department of Nordic Languages, University of Turku.

Adnan, Mehnaz, Jim Warren & Martin Orr. 2010. Assessing text characteristics of electronic discharge summaries and their implications for patient readability. *Proceedings of the Fourth Australasian Workshop on Health Informatics and Knowledge Management* 108, 77–84.

Allvin, Helen. 2010. Patientjournalen som genre. En text- och genreanalys om patientjournalers relation till patientdatalagen [The patient record as genre: A text and genre analysis of the relationship of patient records and the Patient Data Act]. MA thesis, Department of Nordic Languages, Stockholm University.

Allvin, Helen, Elin Carlsson, Hercules Dalianis, Riitta Danielsson-Ojala, Vidas Daudaravicius, Martin Hassel, Dimitrios Kokkinakis, Heljö Lundgren-Laine, Gunnar H Nilsson, Øystein Nytrø, Sanna Salanterä, Maria Skeppstedt, Hanna Suominen & Sumithra Velupillai. 2011. Characteristics of Finnish and Swedish intensive care nursing narratives: A comparative analysis to support the development of clinical language technologies. *Journal of Biomedical Semantics* 2(Suppl. 3):S1.

Aramaki, Eiji, Yasuhide Miura, Masatsugu Tonoike, Tomoko Ohkuma, Hiroshi Mashuichi & Kazuhiko Ohe. 2009. TEXT2TABLE: Medical Text summarization system based on named entity recognition and modality identification. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing* (BioNLP '09), 185–192.

Biber, Douglas & Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Borin, Lars, Natalia Grabar, Maria Toporowska Gronostaj, Catalina Hallett, David Hardcastle, Dimitrios Kokkinakis, Sandra Williams & Alistair Willis. 2009. *Semantic Mining Deliverable D27.2: Empowering the Patient with Language Technology* (Technical Report Semantic Mining, NOE 507505), 1–75. Göteborg: Göteborg University.

Bretschneider, Claudia, Sonja Zillner & Matthias Hammon. 2013. Identifying pathological findings in German radiology reports using a syntacto-semantic parsing approach. *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing* (BioNLP '13), 27–35.

Campbell, David A. & Stephen B. Johnson. 2001. Comparing syntactic complexity in medical and non-medical corpora. *AMIA Annual Symposium Proceedings 2001*, 90–94.

Coden, Anni R., Serguei V. Pakhomov, Rie K. Ando, Patrick H. Duffy & Christopher G. Chute. 2005. Domain-specific language models and lexicons for tagging. *Journal of Biomedical Informatics* 38(6), 422–430.

Dalianis, Hercules, Martin Hassel, Aron Henriksson & Maria Skeppstedt. 2012. Stockholm EPR Corpus: A clinical database used to improve health care. *Proceedings of Fourth Swedish Language Technology Conference*, 17–18.

Dalianis, Hercules, Martin Hassel & Sumithra Velupillai. 2009. The Stockholm EPR Corpus – characteristics and some initial findings. *Proceedings of the 14th International Symposium on Health Information Management Research – ISHIMR 2009*, 243–249.

Fan, Jung Wei, Elly W. Yang, Min Jiang, Rashmi Prasad, Richard M. Loomis, Daniel S. Zisook, Josh C. Denny, Hua Xu & Yang Huang. 2013. Syntactic parsing of clinical text: Guideline and corpus development with handling ill-formed sentences. *Journal of the American Medical Informatics Association* 20, 1–10.

Ferraro, Jeffrey P., Hal Daumé III, Scott L. DuVall, Wendy Webber Chapman, Henk Harkema & Peter J. Haug. 2013. Improving performance of natural language processing part-of-speech tagging on clinical narratives through domain adaptation. *Journal of the American Medical Informatics Association* 20(5), 931–939.

Friedman, Carol, Pauline Kra & Andrey Rzhetsky. 2002. Two biomedical sublanguages: A description based on the theories of Zellig Harris. *Journal of Biomedical Informatics* 35(4), 222–235.

Grigonyté, Gintaré, Maria Kvist, Sumithra Velupillai & Mats Wirén. 2014. Improving readability of Swedish electronic health records through lexical simplification: First results. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@EACL*, 74–83.

Hahn, Udo & Joachim Wermter. 2004. High-performance tagging on medical texts. *Proceedings of the 20th International Conference on Computational Linguistics* (COLING '04), 973–979.

Isenius, Niklas, Sumithra Velupillai & Maria Kvist. 2012. Initial results in the development of SCAN: A Swedish clinical abbreviation normalizer. *Proceedings of the CLEF 2012 Workshop on Cross-language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis*.

Keselman, Alla, Laura Slaughter, Catherine Arnott Smith, Hyeoneui Kim, Guy Divita, Allen Browne, Christopher Tsai & Qing Zeng-Treitler. 2007. Towards consumer-friendly PHRs: Patients' experience with reviewing their health records. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 399–403.

Kokkinakis, Dimitrios. 2012. The journal of the Swedish Medical Association – a corpus resource for biomedical text mining in Swedish. *Proceedings of the 3rd Workshop on Building and Evaluating for Biomedical Text Mining (BioTxtM), LREC 2012 Workshop*, 40–44.

Krauthammer, Michael & Goran Nenadic. 2004. Term identification in the biomedical literature. *Journal of Biomedical Informatics* 37(6), 512–526.

Kvist, Maria, Maria Skeppstedt, Sumithra Velupillai & Hercules Dalianis. 2011. Modeling human comprehension of Swedish medical records for intelligent access and summarization systems – future vision, a physician's perspective. *Proceedings of Scandinavian Health Informatics Meeting*, 31–35.

Kvist, Maria & Sumithra Velupillai. 2013. Professional language in Swedish radiology reports – characterization for patient-adapted text simplification. *Proceedings of Scandinavian Conference on Health Informatics*, 55–60.

Liu, Hongfang, Yves A. Lussier & Carol Friedman. 2001. A study of abbreviations in the UMLS. *AMIA Annual Symposium Proceedings 2001*, 393–397.

Melin, Lars. 2004. *Fattaru?!* [Do ya get it?!]. *Forskning och Framsteg* 3.

Meystre, Stephane M., Guergana K. Savova, Karin C. Kipper-Schuler & John F Hurdle. 2008. Extracting information from textual documents in the electronic health record: A review of recent research. *IMIA Yearbook of Medical Informatics* 47(S1), 128–144.

Mühlenbock, Katarina & Sofie Johansson Kokkinakis. 2009. LIX 68 revisited – an extended readability measure. *Proceedings of Corpus Linguistics 2009*, http://ucrel.lancs.ac.uk/publications/cl2009/364_FullPaper.doc.

Olsson, May. 2011. Vem begriper patientjournalen? [Who comprehends the patient record?]. BA thesis, Department of Language and Literature, Linné University.

Östling, Robert. 2013. Stagger: An open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology* 3, 1–18.

Ownby, Raymond. 2005. Influence of vocabulary and sentence complexity and passive voice on the readability of consumer-oriented mental health information on the Internet. *AMIA Annual Symposium Proceedings 2005*, 585–588.

Pakhomov, Serguei, Ted Pedersen & Christopher G. Chute. 2005. Abbreviation and acronym disambiguation in clinical discourse. *AMIA Annual Symposium Proceedings 2005*, 589–593.

Patrick, Jon, Mojtaba Sabbagh, Suvir Jain & Haifeng Zheng. 2010. Spelling correction in clinical notes with emphasis on first suggestion accuracy. *Proceedings of the 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM)*, 2–8.

Pyper, Cecilia, Justin Amery, Marion Watson & Claire Crook. 2004. Patients' experiences when accessing their on-line electronic patient records in primary care. *The British Journal of General Practice* 54, 38–43.

Skeppstedt, Maria, Maria Kvist & Hercules Dalianis. 2012. Rule-based entity recognition and coverage of SNOMED CT in Swedish clinical text. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12), 1250–1257.

Smedby, Björn. 1991. Medicinens Språk: språket i sjukdomsklassifikationen – mer konsekvent försvenskning eftersträvas [Language of medicine: The language of diagnose classification – more consistent Swedification sought]. *Läkartidningen* 88, 1519–1520.

Smith, Christian, Henrik Danielsson & Arne Jönsson. 2012. A good space: Lexical predictors in word space evaluation. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12), 2530–2535.

Tomanek, Katrin, Joachim Wermter & Udo Hahn. 2007. A reappraisal of sentence and token splitting for life sciences documents. *Proceedings of 12th World Congress on Health (Medical) Informatics – Building Sustainable Health Systems*, 524–528.

Xu, Hua, Peter Stetson & Carol Friedman. 2007. A study of abbreviations in clinical notes. *AMIA Annual Symposium Proceedings 2007*, 821–825.