Lay Language versus Professional Language within the Cardiovascular Subdomain - a Contrastive Study

DIMITRIOS KOKKINAKIS, MARIA TOPOROWSKA GRONOSTAJ
Department of Swedish Language, Språkdata
Göteborg University
Box 200, SE-405 30, Göteborg
SWEDEN
{dimitrios.kokkinakis; maria.gronostaj}@svenska.gu.se   http://www.gu.se/

Abstract: - This paper reports on a corpus-based, contrastive study of Swedish medical language. It is focused on the vocabulary used in two types of medical textual material: professional portals and web-based consumer sites within the domain of cardiovascular disorders. Linguistic, statistical and quantitatively based readability studies are considered in order to find the typical language-dependent and, possibly, language independent characteristics of the material examined and suggest concrete measures that might bridge the gap in medical vocabulary as used by laypersons/consumers and professionals.

Keywords: - contrastive analysis, health literacy, patient terminology, natural language processing, Swedish

1 Introduction

Health care consumers make up a heterogeneous group of individuals with widely differing medical needs, educational background, medical literacy and age. In recent years, they have been exposed to the rapidly growing overload of medical information, e.g. general information on health and medication issues, patients' electronic health records written by and for health care providers, individual advisory information given by net doctors for laypersons. The language of these texts manifests a variety of levels of difficulty, with e-health records and research-oriented texts at one end and ask-the-doctor texts and web portals driven by health care consumers at the other. To make information accessible to the health care consumers, it has to be tailored to their individual needs. Thus the issue of empowerment of health care consumers (e.g. patients) is in accordance with the European Union's data protection directive, in effect since 1998, requiring that all member countries enact legislation enabling patients to have access to their medical records.

In line with this recommendation, the issue of patient empowerment, as well as the development and evaluation of generic methods and tools for assisting patients to better understand their health and health care, has been one of the many goals of the EU-funded "Semantic Mining" network. One strand of this research is developing means for generating patient-friendly, readable texts that paraphrase the content of the electronic health records and other types of health-related information. There are several ways to approach the task and our study focuses on examining linguistic factors that involve contrastive characteristics of the medical sub-corpora, in combination with the results provided by readability tests and other statistical means. Therefore, in our study it is assumed that effective lexical guidance is a prerequisite for consumers' access to medical information in these texts. This pilot study, restricted to the subfield of *cardiovascular disorders*, is an in-depth method study of vocabulary rather than a broad corpus examination. The work presented belongs to the area of *consumer health informatics* which, according to [1], is the branch of medical informatics that analyses consumers' needs for information; it studies and implements methods of making information accessible to consumers, and also models and integrates consumers' preferences into medical information systems.

2 Background

The assessment of reading comprehension, on one hand, and the discrepancy between reading abilities of patients and written patient information, on the other, have been in the focus of a number of studies in the past. However, very few consumer-level vocabularies have been explored so far, in spite of a growing need for providing open access to a non-expert medical vocabulary; see, for instance, [2] and [3]. The development of a lexical database, *Medical WordNet*, consisting of medically relevant terms intended for non-experts, is discussed in [4]. Such a database can be a valuable lexical resource for consumer health information systems that need to comprehend both expert and non-expert medical vocabulary and to map between the two. One motivation for such work is the fact that medical terms, as used by professionals, are subject to control by standardization efforts, continuously evolving, while the highly contextually dependent usage of medical terms on the part of laypersons is much more difficult to capture in applications.

[5] acknowledges that a terminology designed to support clinical records can only accurately account for the patient's problems if the patient's natural language is supported, since patients have a need to understand and validate their records. [6] analysed the readability of health education materials and proposed improvements, emphasizing the issue of cooperation: "Invite target readers to help write and design the material". [7] proposes an interpretive layer framework for helping consumers "find, understand and use medical information when and where it is needed". The authors claim that this is something that can be accomplished by bridging mismatches in knowledge representation between the professional's perspective and the lay perspective and by filling in gaps in consumer knowledge. [7] also proposes that such a system needs a knowledge base for a consumer health ontology and relevant context-based usage information. [8] explores the level of the appropriateness of MetaMap (part of the UMLS) in capturing linguistic meaning of the terms used by patients in free text. In 53% of the cases MetaMap captured the linguistic meaning of the parsed terms used by the patients participating in the study, which is regarded by the authors as a very encouraging figure that demonstrates the possibility of using natural language processing (NLP) tools to automatically extract and capture the linguistic meaning of the terms patients used in their e-mail messages. Finally, [9] investigated the influence of several aspects of the readability (e.g. use of passive voice) of health care information from websites intended for the elderly. His results show that easier-to-read sites could be differentiated most consistently from more difficult ones by vocabulary complexity.

3 Material and Corpus Selection

The MEDLEX Corpus is a Swedish textual material assembled from the Internet, consisting of approximately 10 million words; for details, see [10]. Out of this corpus, we selected two sub-corpora, using a predefined set of ten keywords relevant to the cardiovascular disorder subdomain. The list of keywords consisted of the following words and word fragments: *fragmin, heparin, hemostas* (hemo-stasis), *hjärt* (heart), *koagulantia* (coagulants), *propp* (thrombus, thrombosis), *stenos* (stenosis), *stroke, trombos* (thrombus, thrombosis), *waran* (also including compounds with these strings). The only prerequisite has been that at least one of the keywords could be found in the main title heading of an article. In this way, we could ensure that the two sampled sub-corpora were highly correlated with the cardiovascular sublanguage. The first sub-corpus, *the non-expert corpus*, derives from a number of Swedish daily newspapers and other online health information sources targeted to consumers (e.g. the Swedish NetDoktor). The second sub-corpus, *the expert corpus,* derives from two Swedish medical resources intended for professionals and specialists across a broad spectrum of medical professions: *Läkartidningen*, published weekly by the Swedish Medical Association**,** and *Dagens Medicin*, a news site for medical professionals.

4 Method

4.1 Readability indexes and frequency bands

Readability is an objective, but rather crude, measure that estimates the difficulty in reading text (without considering layout or subject complexity). In addition, a few simple metrics provide a description of the vocabulary as well as a rough indication of lexical richness. Frequency bands were also examined. Two readability tests were applied on the texts in order to determine the difficulty level of the writing style, the *Flesch readability index (FRI)* and the *LIX index*; Table 1.

FRI estimates the reading comprehension level necessary to understand a written document. For a given document, FRI is an integer (0-100). Lower numbers indicate greater difficulty; scores of 0-30 are college graduate level, scores of 50-60 are high-school level and 90-100 should be readable for fourth-graders. FRI is calculated from three statistics: the total number of sentences (N), words (W) and syllables (L); by the formula: $FRI = 206.835 - 1.015 \times (W/N) - 84.6 \times (L/W)$. The basic idea is that each group of continuous non-blank characters counts as a word and each vowel in a word counts as one syllable. To this basic rule there are a number of sub-rules, e.g. words of ≤3 letters count as one syllable. The *LIX index* is a popular readability index in Scandinavia developed by [11]. LIX is defined as the sum of $Lm+Lo$ where $Lm=W/N$ and $Lo=(LongWords/W)*100$. Here *LongWords* are tokens longer than 6 characters. A higher LIX indicates greater difficulty: a LIX between 40-50 usually indicates newspaper language and 50-60 professional language. For a description of the vocabulary knowledge, we used the notions of lexical originality, *LO = num of unique tokens x 100/total num of tokens*, lexical density, *LD = num of lexical tokens x 100/total num of tokens* and lexical sophistication, *LS = num of advanced tokens x 100/num of lexical tokens*. LO measures the learner's/reader's performance relative to the group in which the composition was written. LD is defined as the percentage of lexical words (nouns, verbs, adjectives/participles and adverbs) in a text and LS is the percentage of 'advanced' words in a text (here tokens not included in the frequency bands 0-8; see further below); for a discussion of the weaknesses of these metrics see [12]. Although LD exhibits similar results in the two texts, the LO and LS figures were clearly higher in the expert texts.

	Expert	Non-Expert
FRI	20.75	38.94
LIX	50.42	42.17
LO	16.5	13.1
LD	49.3	50.2
LS	9.18	5.21
Frequency bands	Group 0: 73.87 % Group 1: 7.18 % Group 2: 3.56 % Group 3: 2.17 % Group 4: 1.29 % Group 5: 1.1 % Group 6: 0.71 % Group 7: 0.74 % Group 8: 0.58 % Outsiders: 8.8 %	Group 0: 80.21 % Group 1: 6.48 % Group 2: 3.02 % Group 3: 1.81 % Group 4: 1.06 % Group 5: 0.94 % Group 6: 0.75 % Group 7: 0.66 % Group 8: 0.47 % Outsiders: 4.6 %

Table 1. Readability indexes and frequency-band profile

Finally, using frequency bands, we calculated the percentage of word-overlap (punctuation and names of persons, locations and organizations were automatically identified and filtered out from the texts) in the sub-corpora with the 10,000 most frequent lemmas in the whole MEDLEX-Corpus, a method similar to the lexical frequency profile for assessing vocabulary knowledge, discussed in [12]. The results (Table 1) show that the *outsiders*, word types not in the 10,000 most frequent lemmas, are almost twice as many in the expert texts as in the non-expert texts, while *group 0*, the 1000 most frequent lemmas, corresponds to as much as 73.87% of all lemmas in the expert and 80.21% in the non-expert texts. The overlap of *outsiders* between expert/non-expert texts is very low, only 225 types or 4.2% (see Section 5 for a discussion).

4.2 Quantitative profile

Using a number of different counts, the quantitative characteristics of the selected textual material are summarized in Table 2, showing that:

Measures	Expert	Non-Expert
#tokens/#types	75.699/11.765	70.491/8.554
#TTR	1:6.4	1:8.2
#sTTR	1:3.5	1:4.5
#avg length of nouns (tokens)	9.42	8.68
#compound tokens	9003 (11.8%)	7649 (10.7%)
#unique compounds	4287 (33.6%)	2851 (30%)
#TSR	18.2	14.7
"pure" acronyms	681	184
acronyms in compounds	309	80
#part-of-speech		
common nouns	18969 (25%)	16946 (24%)
proper nouns	2531 (3%)	1733 (2%)
main verbs	7595 (10%)	8958 (13%)
aux. verbs	862 (1%)	1363 (2%)
adj./participles	7293 (10%)	5763 (8%)
pers. pronouns	1829 (2%)	3082 (4%)
rest pronouns	1412 (2%)	1468 (2%)
rest	35528 (47%)	31749 (45%)

Table 2. Quantitative profile of the two corpora

- there is a considerable difference in the number of types in the expert texts as compared to the non-expert ones, which indicates the more scientific profile of the former. Therefore, the type/token ratio (TTR) reflects the fact that the non-expert texts are composed of fewer word forms but are repeated more often (lexical variation). In the non-expert texts, on average, any word form is repeated nearly 8.2 times, as opposed to 6.4 times in the expert texts. Since TTR is a crude measure of lexical variation, the standardized TTR (sTTR) has been proposed as a better alternative. sTTR is computed every *n* words (here every 10,000 tokens) and a running average is computed, which means that an average TTR is based on consecutive 10,000-word chunks of text (*cf.* [13]). sTTR shows a similar picture to TTR, but with lower figures;
- the average length of nouns is greater in the expert texts; longer words are an indication of technical terminology (*cf.* [15]);
- there is a small difference in the number of compound forms 11.8 (33.6% unique) in the expert texts compared to 10.7 (30% unique) in non-expert ones. The high percentage in both cases can be explained by the fact that Swedish is a compounding language;

- the token/sentence ratio (TSR) is higher in expert texts (18.2 compared to 14.7 for non-expert texts);
- there is a significant difference in the number of "pure" acronyms, such as *NSAID*, and also acronyms in compound forms, such as *TNF-alfa*, with a predominance in the expert texts, indicating a clear overuse in these texts;
- the most important differences between the part-of-speech classes are observed for main verbs, auxiliary verbs and personal pronouns. In non-expert texts, there are 13% main verbs compared to 10% in expert texts, and 2% auxiliaries compared to 1%; 4% personal pronouns in non-experts compared to 2% in expert text.

4.3 MeSH terminology

The analysis discussed so far was extended by the use of a Swedish MeSH[1] (Medical Subject Headings) annotator on the texts, in order to find out the distribution of the number of medical terms in the two corpora. In this way, we were also able to account for text characteristics that extend beyond simple surface counts (e.g. tokens and types) and thus complement the quantitative analyses with more qualitative data. Assessing term difficulty is clearly a shortcoming of applying general readability measures (Section 4.1) to health-related content. It has been argued that simple techniques, such as counting the number of syllables in words or appearance on frequency lists, often do not apply to health-related contexts, which typically contain a large number of technical terms *cf.* [14].

Our findings revealed that in the expert texts, there were 3537 complete MeSH matches (e.g. "<mesh tag=" A07.231.114">artär</mesh>", i.e. artery) and 287 partial MeSH annotations (e.g. "sub<mesh tag="A08.186.566.166">araknoid</mesh>", i.e. sub-arachnoid), while in the non-expert texts there were 5062 complete MeSH matches and 204 partial ones. The non-unique figures for complete match clearly show that in non-expert texts there is more use of terminology; however, the figures based on unique occurrences indicate a higher number of *different* terms in the expert texts, which means that in non-expert texts there is a clear indication of repetitive use of the terms, while in the expert texts there is a richer use of terminology. In order to see whether the distribution of these figures is significant, we applied the χ2 (chi square) as guidance, calculated on the six most important hierarchies of MeSH, namely: A (Anatomy), B (Organisms), C (Diseases), D (Chemicals and Drugs), E (Analytical, Diagnostic and Therapeutic Techniques and Equipment), and F (Psychiatry and Psychology). χ2 measures the similarity of one sub-corpus to another with respect to frequencies of individual words or other linguistic features. The figures in parentheses (Table 3) indicate the occurrences of terms in the six hierarchies for the two types of text. The returned χ2 figures (degree of freedom=5) indicate in all four cases that the difference *is* significant.

	Expert	**Non-Expert**
complete match χ2=350, p<=0.001	3537 (454+13+ 1554+404+ 1097+15)	5062 (1441+11+ 2131+447+ 1021+11)
unique compl. match χ2=12.1, p<=0.05	601 (105+4+ 216+97+170+9)	491 (108+5+ 185+89+97+7)
partial match χ2=75.2, p<=0.001	287 (16+4+ 136+9+122+0)	204 (34+10+ 90+ 17+35+18)
unique partial match χ2=26.3, p<=0.001	160 (13+3+ 64+8+72+0)	91 (18+6+ 30+14+22+1)

Table 3. Distribution of MeSH annotations (A+B+C+D+E+F)

5 Discussion

5.1 Lexical profile of the vocabulary

The contrastive linguistic overview of the vocabulary undertaken below complements its quantitative profile presented in 4.2. The point of departure is lexical, which means that morphological, morpho-syntactic and semantic properties of the vocabulary in the texts are brought into focus. Special attention is paid to a subset of lexical properties which captures the types of linguistic contrasts of relevance for the analysis, e.g. distribution of parts of speech, compositionality of word forms, form familiarity and also the manifested semantic relations between words. In this study we also take advantage of information on the distribution of words on frequency bands. Linguistic contrasts are manifested most clearly at the extreme ends of the frequency bands, namely by words listed within the groups 0 and so-called *outsiders*. Group 0 is a local, common core vocabulary representative of the expert and non-expert texts examined here. Its vocabulary shares the following characteristics:

- all parts of speech are represented in this group, including a few very common abbreviations, such as: *ca*, *m.* and *mg*;

[1] MeSH® is the controlled vocabulary thesaurus of the NLM, U.S. National Library of Medicine. The original data from NLM have been supplemented with Swedish translations made by staff at the Karolinska Institute Library (http://mesh.kib.ki.se/swemesh/).

- the majority of words are simplex;
- occurrences of compounds (e.g. *hjärt-kärlsjukdom* 'cardiovascular disease') are exceptional;
- native vocabulary dominates; loan words from Latin and Greek are very few (e.g. *antibiotikum* 'antibiotic');
- general vocabulary dominates, but it has a touch of medical profile due to the sub-domain selected; hence a certain prevalence of words referring to anatomy, diseases, symptoms and treatment, as well as to the medical staff and organization of health care.

In the subsequent groups, 1 to 8, we observed that:

- in both the expert and non-expert texts, only a subset of parts of speech is represented, since most of the so-called stop words (e.g. conjunctions) belong to the core vocabulary;
- the number of compounds grows rapidly in both expert and non-expert texts, but the increase of unique compound forms is more conspicuous in the expert texts;
- occurrence of medical terms in the expert texts is higher; for instance, within group 8 of the expert texts the number of medical terms is about twice as large as for to the corresponding group of the non-expert texts;
- Latin and Greek loan words are more frequent in the expert texts;
- the prevalence of nominalizations, a characteristic of scientific texts [15], is reaffirmed in our study. There are five times more nouns than verbs in group 8 of the expert texts. The corresponding factor for the non-expert text is four;
- the number of medically relevant abbreviations and acronyms increases across groups 1 to 8. The total number of self-contained acronyms is four times larger in the expert texts than the non-expert texts. Abbreviations usually refer to dosage of drugs, types of medical examinations and their measures, specification of anatomic locations, and others.

The tendencies described for groups 1 to 8 are also valid for the group of outsiders. However, the differences become more evident as medical terminology is gaining ground. The group of outsiders is also more heterogeneous in its internal composition because it also includes new elements, namely occurrences of foreign words, particularly English ones. Thus, an array of word forms that need to be handled when processing medical text and designing lexical guidance includes:

- unique or less frequent acronyms and abbreviations in the expert texts like: *ESCS, TMR, vf, vka, bitr*
- medical compounds words with or without acronyms: *ICD-grupp, fotopletysmograf*
- medical simplex words: *tromb, torsion*
- text unique or less frequent medical non-compound words: *ögon, oxytocin*
- foreign words: *grown-up, serious*
- misspellings (medical and general language)
- general language compounds
- general language simplex words

5.2 Syntactic parameters

Not only frequency analysis of vocabulary but also syntax can be a criterion for differentiation of scientific from general texts. High occurrences of noun phrases as well as use of infinitival and passive constructions are considered to be representative of scientific texts in general ([16], [17:14]). Syntactic analysis focused on sentence length and structure brings out further parameters which can support the contrastive analysis of the texts. Some insights concerning comparison of Swedish expert texts in various domains, which do not include medical language, are given in [15]. Syntactic properties of Swedish medical language showing contrasts between expert/non-expert corpora have not been examined yet, to our knowledge.

5.3 Lexico-semantic parameters

The meanings of medical word forms can be studied with respect to semantic relations like synonymy, antonymy, hyperonymy, hyponymy and meronymy. Here, we argue that explicit information on these relations can *not* only support the contrastive analysis of the medical sub-corpora but can also offer significant help for laypersons in understanding medical language, if the information is made accessible to them via an online dictionary (*cf.* [4]). To illustrate the issue, we take a closer look at the six most frequent keywords' related terms in these sub-corpora and reflect on the correlations between meanings, semantic relations and their frequencies (Table 4).

It is of importance to note that these two lists share four out of the six words, which means that both of them capture the most central terms in the cardiovascular domain. The partially different ranking of these four words can be explained partly by differences in the textual material, partly by different preferences of laypersons and professionals for describing health conditions.