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

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: \[ FRI = \frac{206.835 - 1.015 \times (W/N) - 84.6 \times (L/W)}{ } \]. 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 \( Lo = (LongWords/W) \times 100 \). 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, \(LO = \text{num of unique tokens} \times 100/\text{total num of tokens}\), lexical density, \(LD = \text{num of lexical tokens} \times 100/\text{total num of tokens}\) and lexical sophistication, \(LS = \text{num of advanced tokens} \times 100/\text{num of lexical tokens}\). 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.

### 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:

<table>
<thead>
<tr>
<th>Measures</th>
<th>Expert</th>
<th>Non-Expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>#tokens/typess</td>
<td>75.699/11.765</td>
<td>70.491/8.554</td>
</tr>
<tr>
<td>#TTR</td>
<td>1.6.4</td>
<td>1.8.2</td>
</tr>
<tr>
<td>#TTR</td>
<td>1.3.5</td>
<td>1.4.5</td>
</tr>
<tr>
<td>#avg length of nouns (tokens)</td>
<td>9.42</td>
<td>8.68</td>
</tr>
<tr>
<td>#compound tokens</td>
<td>9003 (11.8%)</td>
<td>7649 (10.7%)</td>
</tr>
<tr>
<td>#unique compounds</td>
<td>4287 (33.6%)</td>
<td>2851 (30%)</td>
</tr>
<tr>
<td>%“pure” acronyms</td>
<td>681</td>
<td>184</td>
</tr>
<tr>
<td>acronyms in compounds</td>
<td>309</td>
<td>80</td>
</tr>
<tr>
<td>#part-of-speech</td>
<td></td>
<td></td>
</tr>
<tr>
<td>common nouns</td>
<td>18969 (25%)</td>
<td>16946 (24%)</td>
</tr>
<tr>
<td>proper nouns</td>
<td>2531 (3%)</td>
<td>1733 (2%)</td>
</tr>
<tr>
<td>main verbs</td>
<td>7595 (10%)</td>
<td>8958 (13%)</td>
</tr>
<tr>
<td>aux. verbs</td>
<td>862 (1%)</td>
<td>1363 (2%)</td>
</tr>
<tr>
<td>adj./participles</td>
<td>7293 (10%)</td>
<td>5763 (8%)</td>
</tr>
<tr>
<td>pers. pronouns</td>
<td>1829 (2%)</td>
<td>3082 (4%)</td>
</tr>
<tr>
<td>rest pronouns</td>
<td>1412 (2%)</td>
<td>1468 (2%)</td>
</tr>
<tr>
<td>rest</td>
<td>35528 (47%)</td>
<td>31749 (45%)</td>
</tr>
</tbody>
</table>

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">artery</mesh>”, i.e. artery) and 287 partial MeSH annotations (e.g. “<mesh tag="A08.186.566.166">arachnoid</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 matches indicate the occurrences of terms in the six hierarchies for the two types of text. The returned \(\chi^2\) figures (degree of freedom=5) indicate in all four cases that the difference is significant.

<table>
<thead>
<tr>
<th>Table 3. Distribution of MeSH annotations (A+B+C+D+E+F)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>complete match</strong></td>
</tr>
<tr>
<td>(\chi^2=350, p&lt;=0.001)</td>
</tr>
<tr>
<td><strong>unique compl. match</strong></td>
</tr>
<tr>
<td>(\chi^2=12.1, p&lt;=0.05)</td>
</tr>
<tr>
<td><strong>partial match</strong></td>
</tr>
<tr>
<td>(\chi^2=75.2, p&lt;=0.001)</td>
</tr>
<tr>
<td><strong>unique partial match</strong></td>
</tr>
<tr>
<td>(\chi^2=26.3, p&lt;=0.001)</td>
</tr>
</tbody>
</table>

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

- 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, v.f, 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.
Laypersons tend to focus on symptoms and professionals on diagnoses; hence the difference in the ranking of hjärta.

<table>
<thead>
<tr>
<th>Expert texts</th>
<th>Non expert texts</th>
</tr>
</thead>
<tbody>
<tr>
<td>stroke</td>
<td>hjärta</td>
</tr>
<tr>
<td>hjärtnfarkt</td>
<td>stroke</td>
</tr>
<tr>
<td>hjärta</td>
<td>blodpropp</td>
</tr>
<tr>
<td>hjärtvikt</td>
<td>hjärtnfarkt</td>
</tr>
<tr>
<td>hjärtsvikt</td>
<td>hjärta</td>
</tr>
<tr>
<td>hjärtsvikt</td>
<td>blodpropp</td>
</tr>
<tr>
<td>hjärtsvikt</td>
<td>hjärta</td>
</tr>
<tr>
<td>hjärstopp</td>
<td>propp</td>
</tr>
<tr>
<td>hjärtstopp</td>
<td>hjärta</td>
</tr>
</tbody>
</table>

Table 4. Most frequent keywords’ related terms

Diagnostic terms like stroke, hjärtnfarkt (heart attack) and hjärtsvikt (cardiac insufficiency) are ordered according to the same ranking sequence on both lists. For laypersons to have a minimal understanding of such terms, some knowledge about their place in the medical ontology, in other words knowledge of their hyperonym is required. Their hyperonym, hjärt-kärlsjukdom (cardiovascular diseases) happens to occupy the fifth position on the expert lists and the eleventh on the non-expert list, which means that the term can be considered familiar even to laypersons.

High ranked on the non-expert list, the polysemous word blodpropp and its short form propp belong to two different conceptual categories, “organic object” vs. “cardiovascular disease”. The former gets a concrete reading, ‘a blood clot’, and the latter an abstract one, referring to ‘health condition caused by a blood clot, thrombus’. Since the abstract reading hints at the concrete reading, the correct disambiguation load often lies on the word’s lexical and/or syntactic context. Unfortunately, it is often the case that general dictionaries explain only the word’s concrete meaning, leaving a layperson in the lurch. More exhaustive information can be obtained from MeSH, even if its definition explicates only the concrete reading of blodpropp ‘thrombus’. The meaning referring to the health condition thrombosis can be obtained from the MeSH ontological hierarchy, in which the node blodproppssjukdom (thromboembolism) is a hyponym of blodpropp (thrombosis) whose top hyperonym is hjärt-kärlsjukdom (cardiovascular disease). Thus the second, disease reading is mediated in MeSH via the ontology. The ambiguity factor of the word blodpropp can possibly serve as an explanation why its frequency is low in the expert texts (40 occurrences) and, in tandem, why it ranks as the second word on the non-expert list (221 occurrences). This observation conforms to [5]. Clinical terms are by necessity complex and not easily amenable to being represented in client language without a full definition; a hierarchical placement of terms has proved beneficial in orienting a client in the meaning of the term. This strategy can contribute to the fact that client language has greater variability in meaning, which results in choosing superordinate, more general terms instead of subordinate, narrower ones.

Listing of synonyms with comments on their register is another step that might not only be layperson friendly but could also contribute to bridging the communication gap between laypersons and professionals. Since the task of manual extraction of semantic information from corpora is both time- and cost-consuming, further elaboration of semantic acquisition approaches needs to be investigated (cf. [17]).

5.4 Pragmatic relevance

The binary division of the examined medical corpora alludes to the potential groups of recipients represented by health care workers and laypersons. The two recipient groups have partly a common pragmatic purpose, namely communication, which needs to be based on maximum mutual understanding to ensure the best care, partly a separate one, i.e. communication within their own groups, focused on sharing medical knowledge.

From the analysis of the vocabulary follows that the stock of medical terms in the non-expert texts constantly grows. According to [19:185], “the vocabulary expansion leads to a greater differentiation of conceptual categories and to a more precise ‘articulation of reality’ and therefore to the perfection of inter-human communication”. The relevance of this observation for the issue of patient empowerment is obvious. The dynamic vision of vocabulary expansion also entails an increase in medical literacy, but at the same time it contributes to segregation of health consumers with respect to their initiation into medical knowledge. To compensate for this knowledge segregation, lexical resources need to integrate lexical and medical knowledge in a user-friendly and flexible way to suit the actual needs of particular users.

The number and the diversity of the types of lexical data gathered by the contrastive analysis of the corpora are also of pragmatic importance for the construction of an open-source, multidimensional, on-line lexical resource providing selective and dynamic lexical assistance for health care consumers. The information included there on
words’ morphological, syntactic and semantic behaviour should be integrated with some basic and/or advanced encyclopaedic medical information and information on English equivalents to support information extraction from other lexical and textual sources. The fusion of all information is a key issue for the empowerment of health care consumers as well as for a number of NLP applications (e.g. generation).

6 Conclusion
In this study, we have compared the language in two types of register, i.e. expert and non-expert Swedish texts in the domain of cardiovascular disorders. The main question that arises from this work is: what are the practical benefits, if any, brought about by this study? We believe that our work provides some guidance for those interested in improving the readability of health-related information material. It attempts to integrate a language-independent approach (statistics and frequency criteria) with a language-dependent approach (vocabulary and its linguistic properties). We hope that our work provides some insights and relevant pragmatic implications on how to bridge the language barrier between health care consumers and professionals.

We are fully aware that this vocabulary study is just a beginning and is to be complemented with an extensive analysis of syntactic relations on sentence level as well as phrase level. Types of sentences/clauses, their coordination and subordination structures, and the distribution of passive constructions are some of the issues for future work, which will be carried out by statistic means as well as available parsing tools. Another relevant research topic is the extraction of the types of syntactic or/and semantic patterns characteristic of the laycorpora in order to re-formulate the content of expert documents in a user-friendly way. Such patterns can also generate new information for enriching the lexical resource with semantic relations. In the near future, we also intend to continue our efforts to improve readability measures and investigate how these indexes are related to how consumers use and benefit from material on health care information websites.

Acknowledgements
This work was partially supported by the Semantic mining - Semantic Interoperability and Data Mining in Biomedicine - NoE 507505. The tool used for the frequency band analysis has been developed by M. Stissing (Århus komm. Sprogcenter).

References: