Lexical Parameters, Based on Corpus Analysis of English and Swedish Cancer Data, of Relevance for NLG

Dimitrios Kokkinakis
Maria Toporowska Gronostaj
Göteborg University
Department of Swedish Language,
Språkdata
Sweden
{svedk,svemt}@svenska.gu.se

Catalina Hallett
David Hardcastle
Centre for Research in Computing
The Open University
Walton Hall
Milton Keynes MK7 6AA
{c.hallett,d.w.hardcastle}@open.ac.uk

Abstract

This paper reports on a corpus-based, contrastive study of the Swedish and English medical language in the cancer sub-domain. It is focused on the examination of a number of linguistic parameters differentiating two types of cancer-related textual material, one intended for medical experts and one for laymen. Language-dependent and language independent characteristics of the textual data between the two languages and the two registers are examined and compared. The aim of the work is to gain insights into the differences between lay and expert texts in order to support natural language generation (NLG) systems.

1 Introduction

Health care consumers are constantly exposed to the rapidly growing overload of medical information, e.g. general information on health and medication issues, 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. However, making the issue of empowerment of health care consumers (e.g. patients) is a challenging task because health care consumers make up a heterogeneous group of individuals with widely differing medical needs, educational background, medical literacy and age. In line with these challenges, the issue of patient empowerment, as well as the development and evaluation of 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. A 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 on an empirical basis, linguistic factors that involve contrastive characteristics of the medical sub-corpora. In our study it is assumed that effective lexical guidance is a prerequisite for consumers' access to medical information in these texts. Our study[1] is restricted to the subfield of *cancer* while our intended readership is the group of patients. The aim of our study is to gain insights into the differences between languages and registers for supporting systems that generate patient-friendly language.

Points of related work are given in (Section 2). In Section 3 we present a concise view of the findings in the corpora and in Section 4 we discuss the results. Finally, Section 5 summarizes the paper and provides some topics for future research.

[1] Our work belongs to the area of *consumer health informatics* which 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 (Eysenbach, 2000).

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. Campbell & Johnson (2001) investigated the syntactic differences between medical and non-medical corpora. Various experiments showed significant differences in syntactic content and complexity, for instance in the distribution of both simple part-of-speech and part-of-speech bigrams between discharge summaries and the Brown corpus. Cantalejo & Lorda (2003) analyzed the readability of health education materials and proposed improvements, emphasizing the issue of cooperation: “Invite target readers to help write and design the material”. Soergel et al. (2004) propose 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. Soergel et al. (2004) also propose that such a system needs a knowledge base for a consumer health ontology and relevant context-based usage information. Hsieh et al. (2004) explore the level of the appropriateness of MetaMap (part of the UMLS - nlm.nih.gov/pubs/factsheets/umls.html) 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, Ownby (2005) 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 Comparing Corpora

We have collected and analysed two different corpora in two registers (Section 3.1) along different, particular lexical, dimensions. These included loan/native words, lexical choice, periphrastic constructions, the use of pronouns and the use of a number of meta-markers as indicators for epexegesis.

3.1 Swedish and English Cancer Corpora

Two types of registers for each of the two languages are examined, namely expert-lay [lay] corpora and expert-expert [expert] corpora (Table 1). The English expert corpus consists of case studies and manual for medics, while the English lay corpus includes manuals for patients, patient information leaflets and patient testimonials. The Swedish expert corpus consists of internet-based material for experts while the lay part of the corpus is acquired from various news sites as well as patient-oriented sites (e.g. the Swedish Netdoktor).

	English	Swedish
Expert (size #words)	140 000	190 000
Lay (size #words)	155 000	170 000
Expert (words/sentence)	17.67	17.03
Lay (words/sentence)	18.95	15.18
Expert (complex words)	30.38%*	11.45%**
Lay (complex words)	14.61%*	8.94%**

Table 1. Qualitative profile of the corpora (*>3 syllables; **solid compounds)

3.2 Loan/Native words

We considered 30 suffixes, prefixes and infixes that are indicative of medical words of Greek or Latin origin. We specifically selected those affixes that are, at the same time, representative for the cancer domain and are less likely to appear in general purpose vocabularies. We associated with each affix one or more English/Swedish words that correspond to the loan suffix, e.g.

- #mammo#/breast-/bröst-
- #nephr/nefr#-/kidney/njur-
- #angio#/artery-/artär-/-åder

- #hepat#/liver-/lever-
- #oesophag/esofag#/throat-/matstrup-

Whilst the English expert corpus exhibited an almost equal distribution of loan and English equivalents, the lay corpus contained predominantly English equivalents, with a much smaller number of loan terms. The Swedish data exhibits similar results, although the Swedish expert texts show rather less uniformity in the distribution of loan/native words (Table 2).

	English		Swedish	
	Lay	Expert	Lay	Expert
Loan	0.76%	3.65%	0,67%	1,04%
Native	3.99%	3.42%	3,27%	2,95%

Table 2. Distribution of loan/native words

3.3 Lexical Choice

For English and Swedish, there is a pretty clear tendency of using lay terms in lay corpora, instead of more specific terms, as compared to the expert corpus. We examined constructions containing a number of cancer-related terms in the subdomain. The sample of examples listed in Table 3 is indicative of the preferable vocabulary choice for the two types of corpora.

	English		Swedish	
	Lay	Expert	Lay	Expert
tumour/tumor	697	1532	578	710
cancer*	2690	548	1172	997
carcinoma	177	773	18	83
malignancy	1	98	5	56
neoplasm	-	165	-	7
metastasis	14	114	78	524

Table 3. Lexical choice (*including “cancersjukdom” – cancer disease)

3.4 Periphrastic Constructions

Many medical terms have a lot of justifiable alternate forms with several orthographic and lexical variants. The analysis of the English corpora showed a clear preference of compounds constructions in both lay and expert texts (Table 4). The analysis of the Swedish showed a similar tendency, with a few exceptions, e.g. ‘tjocktarmscancer’ (cancer of the colon) occurred 69 times in the expert and 38 in the laytexts, while its periphrastic “cancer i tjocktarmen” occurred 44 times in the expert and 18 times in the lay texts.

We could not draw any clear conclusions from this exercise, apart from the fact that the compound forms (e.g. “breast cancer”, “bröstcancer”) are the preferred expressions in both corpora.

	English		Swedish	
	Lay	Expert	Lay	Expert
breast cancer	221	37	471*	460*
cancer of/in the breast	2	0	9	2
lung cancer	75	44	108*	64*
cancer of/in the lung	4	0	8	1

Table 4. Periphrastic writing (*solid compounds, “bröstcancer”, “lungcancer”)

3.5 Pronouns

There is a clear preference of using the pronouns “you/your/yours/yourself/yourselves” in the lay texts in both languages (Table 5). Also, in the Swedish data the use of the pronoun “man” (one) is also very common, with 1707 occurrences in the lay and 671 in the expert texts.

	English	Swedish
Lay	757	511
Expert	9	15

Table 5. Distribution of 2nd person pronoun

3.6 Epexegesis

We investigated the use of connective phrases, denoted by a handful words and punctuation marks, which signal the presence of synonyms, paraphrases, or substitution (*cf.* Pearson, 1998). We considered for instance the following words and expressions that may indicate explanations: *call(ed), known as, aka, layman’s terms, mean(s), in other words, what is.* We found 378 occurrences in the English lay corpus (0.24%) and 76 occurrences in the English expert corpus (0.05%). Corresponding expressions in Swedish, such as “så kallade” (so called) were four times more frequent in the lay texts.

4 Discussion

Target text analysis is the very first step in the design and development of Natural Language Generation systems. Moreover, several researchers have emphasised the fact that corpus analysis is instrumental in reducing the effort involved in constructing the complex knowledge bases generally required by NLG systems (Knight & Hatzivassiloglou 1995, Langkilde & Knight 1998, Pan & Shaw 2004).

Since our intended target texts emulate the style and lexical content of the analysed corpora (Hallett & Scott, 2005), we are able to offer several recommendations and scoring mechanisms for bilingual English-Swedish NLG systems. More specifically, we are able to:

- informing an NLG system with regard to the appropriate lexical choices and syntactic constructions
- assess whether an automatically generated text is appropriate as patient information material, by analysing its readability level, lexical composition and syntactic complexity and comparing with the reference lay corpus. Similarly, for NLG systems that generate multiple variants, our analysis can help score the alternatives in order to make the best choice

5 Conclusions

In this study, we have compared the language in two types of register, i.e. expert and non-expert English and Swedish texts in the domain of cancer. A series of corpus-based experiments were conducted in order to assess the lexical variety of the corpora. The main question that arises from this work is: *what are the practical benefits, if any, brought about by this study particularly for the field of natural language generation.*

We hope that our work provides some insights and relevant pragmatic implications on how to support the generation of patient-friendly documents (particularly electronic health records and discharge letters) at the lexical and terminological level (use of explanations and definitions, use of paraphrased terms, use of “patient” terms etc.).

In the near future, we are planning to extend the analysis to discover discourse-related features, such as rhetorical relations, and actually look into more detail into semantic features. Moreover, we are currently in the process of adapting an existing NLG-system (Hallett et al., 2007) to both Swedish as well as to the cancer subdomain for English.

Acknowledgements

This work has been supported by the “Semantic Interoperability and Data Mining in Biomedicine”, a Network of Excellence (NoE) funded by the European Commission under Framework 6.

References

Campbell DA and Johnson SB. 2001. *Comparing Syntactic Complexity in Medical and Non-Medical Corpora*. Proc AMIA Symp. 2001;90-4

Cantalejo B. IM. and Lorda S. P. 2003. *Can Patients Read What We Want Them to Read? Analysis of The Readability of Printed Materials for Health Education.* Aten Primaria. 30;31(7):409-14.

Eysenbach G. 2000. Consumer Health Informatics. *British Med. J.* 320:1713-1716

Hallett C. and Scott D. 2005. *Structural Variation In Generated Health Reports*. Proc. of the 3rd International Workshop on Paraphrasing (IWP2005). Korea.

Hallett C., Scott D. and Power R. 2007. *Composing Questions through Conceptual Authoring.* J. of Computational Linguistics. (to appear).

Hsieh Y., Hardardottir G. A. and Brennan P. F. 2004. *Linguistic Analysis: Terms and Phrases Used by Patients in E-mail Messages to Nurses*. MEDINFO. Amsterdam: IOS Press.

Knight K. and Hatzivassiloglou V. (1995). *Two-Level, Many-Paths Generation*. Proc. of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95), Cambridge, MA.

Langkilde I. and Knight, K. (1998). *Generation that Exploits Corpus-Based Statistical Kn*owledge. Proc. of Coling-ACL'98. Quebec, Canada.

Ownby R.L. 2005. *Influence of Vocabulary and Sentence Complexity and Passive Voice on the Readability of Consumer-Oriented Mental Health Information on the Internet.* Proc. AMIA, USA.

Pearson J. 1998. *Terms in Context*. Amsterdam: John Benjamins Publishing Company

Shimei Pan and James Shaw. 2004. *SEGUE: A Hybrid Case-Based Surface Natural Language Generator*. Proc. of ICNLG, Brockenhurst, U.K.

Soergel D., Tse T. and Slaughter L. 2004. *Helping Healthcare Consumers Understand: an “Interpretive Layer” for Finding and Making Sense of Medical Information*. MEDINFO. IOS Press.