SemanticMining

NoE 507505

Semantic Interoperability and Data Mining in Biomedicine

Deliverable D27.2

Empowering the patient with language technology

Delivery date: month 38

Report Version: 1

Report Preparation Date: 2007.02.28

Dissemination level: POOH

Associated work package: WP27

Lead contractor: UGOT

Information Society Technologies

Project funded by the European Community under the FP6 Programme "Integrating and Strengthening the European Research Area" (2002-2006)

Administrative information

Lead contractor/partner for WP/Deliverable
Göteborg University, Sweden

Assisting partners for WP/Deliverable
Université Paris Descartes, France
The Open University, UK

Author list
Lars Borin
Natalia Grabar
Maria Toporowska Gronostaj
Catalina Hallett
David Hardcastle
Dimitrios Kokkinakis
Sandra Williams
Alistair Willis

Semantic Mining Deliverable D27.2: Empowering the patient with language technology

Lars Borin (1), Natalia Grabar (2),
Maria Toporowska Gronostaj (1), Catalina Hallett (3),
David Hardcastle (3), Dimitrios Kokkinakis (1),
Sandra Williams (3), Alistair Willis (3)
(1) Göteborg University, (2) Université Paris Descartes, (3) The Open University

Contents

Part I
Introduction

1 Empowering the patient with language technology

This report forms the second deliverable of Work Package 27 of the EC Network of Excellence 507505 *Semantic Interoperabilty and Data Mining in Biomedicine* (Semantic Mining). The purpose of the work presented here has been to develop well-founded and coherent arguments that language technology can be put to good use in designing IT solutions with the aim of empowering patients and other non-professionals who wish to access medical information, e.g. health record contents, and also, more concretely, to propose a research program with this goal.

We laid the groundwork in the first WP 27 deliverable (Åhlfeldt et al. 2006), a literature survey aspiring to present a picture of the state of the art in patient-friendly information systems. In this second deliverable we report on our own contrastive corpus-based investigations of the differences between professional and non-professional medical language in several languages (part II). The differences between language registers is well-studied *per se*, but the multilingual aspect that our corpus studies now bring to this field turns out to be fertile new ground which we have only begun to explore.

On the basis of the earlier literature survey and the results of the corpus studies, we then in part III go on to draw some tentative conclusions about how "less patient-friendly documents" could be turned into more patient-friendly ones using language technology, with a particular emphasis on natural language generation (NLG) techniques.

Part II
Corpus study

2 Introduction

The corpus studies presented here represent a first attempt at characterizing in concrete terms the differences between (medical) professional and lay language cross-linguistically. We review and correlate the findings of three different studies of this topic, made by the present authors in

various combinations (Hallett, Hardcastle & Willis 2006; Kokkinakis & Toporowska Gronostaj 2006; Krivine et al. 2006; since the present paper constitutes a summary and extension of the three previous works, we will not as a rule refer to them explicitly in what follows). The investigated languages are English (based on the data in Hallett, Hardcastle & Willis 2006), Japanese (Krivine et al. 2006), Russian (Krivine et al. 2006) and Swedish (Kokkinakis & Toporowska Gronostaj 2006).

The studies are only partially overlapping in the textual and linguistic phenomena investigated, hence the cross-linguistic generalizations will concern a subset of the investigated characteristics. On the other hand, the fact that we had three investigations of different languages made from slightly different points of departure, has actually allowed us to reinterpret retrospectively some of the data in each of the individual studies in the light of the other two.

Health care consumers are a heterogeneous group of individuals with widely differing medical needs, backgrounds, levels of medical literacy and ages. In recent years, they have been exposed to a rapid growth in the amount of medical information available, 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 laypeople. The language of these texts covers 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 maintained by health care consumers at the other. To make information accessible to 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 Eysenbach (2000), 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.

The aim of the corpus analysis exercise was to provide us with basic information on the lexical and syntactical features of medical texts, with a view of producing medical reports easily understandable by patients.

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 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 the provision of open access to a non-expert medical vocabulary; see, for instance, Tse & Soergel 2003. The development of a lexical database, Medical WordNet, consisting of medically relevant terms intended for non-experts, is discussed in Smith & Fellbaum 2004. 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 continuously evolving standardization, while the highly contextually dependent usage of medical terms on the part of laypeople is much more difficult to capture in applications.[1]

Brown, Price & Cox (1997) acknowledge 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. Cantalejo & Lorda (2003) 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". Soergel, Tse & Slaughter (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, Tse & Slaughter (2004) also propose that such a system needs a knowledge base for a consumer health ontology and relevant context-based usage information. Hsieh, Hardardottir & Brennan (2004) explore the level of the appropriateness of MetaMap (part of the Unified

[1]The Consumer Health Vocabulary is an open source collaborative initiative in which technical terms used by health care professionals are linked to consumer health vocabularies <http://www.consumerhealthvocab.org/>.

Medical Language System, 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, 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.

In more general terms, Kittredge (2003) discusses that sub-languages can deviate from a standard language lexically, syntactically and semantically. Among properties of a sublanguage being of relevance for the NLP applications and in particular in the design of the descriptive grammar, lexicon and the various stages of the processing algorithms, Kittredge (2003: 437) names the following:

- restricted lexicon (and possibly including special words not used elsewhere in the language);
- relatively small number of distinct lexical classes;
- restricted sentence syntax;
- deviant sentence syntax;
- restricted word co-occurrence patterns which reflect domain semantics;
- restricted text grammar;
- different frequency of occurrence of words and syntax patterns from the norm for the whole language – each sublanguage has its own profile, which can be used to help set up preferred interpretations for new texts.

3 Corpora and genres

3.1 Languages, corpora and texts

The languages investigated here are (British) English and Swedish in some detail, and, more superficially, Japanese and Russian.

3.1.1 English corpora

We have collected four small-size corpora (total word count approximately 280,000 words) containing texts in the domain of cancer, which cover three communication procedures and four discourse genres. In the Expert-Expert category we have two corpora: case studies written for the benefit of students and clinicians (collected online)and extracts from the *Merck Manual for Medics* (Beers & Berkow 2006). The Expert-Lay corpus contains cancer-related texts from the *Merck Manual for Patients* (Beers 2006). The Lay-Lay corpus consists of online patient testimonials relating their cancer experience (referred to as "stories" in the tables below). A description of the four corpora is presented in table 1. The online materials contain texts about a variety of types of cancer. Although we have taken reasonable care that no one type of cancer dominates the corpus, these texts are not representative for the domain of cancer, nor are the various cancer types equally represented. The Merck manuals are the closest to each other in overall content, although even in their case some types of cancer are over-represented in one manual as opposed to the other.

Corpus	Communication type	Discourse genre	Size
case studies	Expert-Expert	teaching/research	86908
Merck medics	Expert-Expert	manual	61032
Merck patients	Expert-Lay	manual	55154
stories	Lay-Lay	blog	78668

Table 1: English corpus description

Whilst some of the texts we have used come from internet sites that encourage the distribution of their materials, others either have no specific copyright information available or require approval. In particular, the Merck manuals are copyrighted materials which require written approval from the copyright holder for any kind of personal, research or commercial use.

Additionally, we performed a selection of files from the British National Corpus (BNC) and split them into four categories according to a list of keywords found in their header. The resulting subcorpora are:

- **Expert-Expert**: Academic seminars on cancer and also the GUT journal portion of the BNC, totalling 745k words over 3,395 documents.
- **Expert-Lay**: Pamphlets about cancer, AIDS and general healthcare issues, totalling 150k words over 927 documents.
- **Lay-Lay**: A mixture of magazine, newsletter and journalist pieces

about healthcare in quite general terms, totalling 130k words over 38 documents.

- **GP Consultations**: 119 GP consultations comprising 85k words from the spoken-language part of the BNC. Because it is from the spoken part it is very different from the other subcorpora, so for high level analysis it is something of an outlier in many respects.

Since the BNC subcorpora are much more varied in terms of both domain and register, we intend to use the BNC material as a reference corpus only, for testing our findings on the main corpus.

3.1.2 Swedish corpora

The lay versus professional sub-classification of medical texts is a very rough, pragmatic, addressee-focused classification of corpora, which requires a more fine-grained sub-categorization based on form and content. As far as form is concerned, there is no doubt that specific sub- genres need to be recognised and that this information is important for contrastive linguistic studies of medical sub-languages. Reaching consensus among the research community on relevant sub-genre categories is an important step towards evolving standards for their encoding and in consequence for account of linguistic contrasts.

Within NLP there exist a number of approaches to automatic text and genre classification, some more sophisticated than others. For instance, Karlgren & Cutting (1994) apply statistical discriminant analysis; Stamatatos, Kokkinakis & Fakotakis (2000) apply stylistically homogeneous categories such statistical measures of vocabulary richness using frequency counts; while Hahn & Wermter (2004) apply n-gram character statistics.

The MEDLEX Corpus comprises Swedish textual material assembled from the internet, consisting of approximately 10 million words; for details, see Kokkinakis 2006. Out of this corpus, we selected two sub-corpora (roughly 85,000 tokens each), using a predefined set of ten keywords relevant to the cardiovascular disorders' subdomain. Since the MEDLEX-Corpus has been already annotated with meta-descriptors such as "<title>" we decided for simplicity reasons to only search on the title descriptor of each document in the corpus. The list of keywords consisted of the following words and word fragments (including compounds containing these words):

fragmin 'an anticoagulant'
heparin
hemostas 'hemostasis'
hjärt(a) 'heart'
koagulantia 'coagulants'
propp 'thrombus, thrombosis'
stenos 'stenosis'
stroke
trombos 'thrombus, thrombosis'
waran 'an anticoagulant'

The only prerequisite has been that at least one of the keywords be present 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.

To maximally maintain the inherent linguistic homogeneity of expert and lay corpora, we decided not to include the texts from the ask-the-net-doctor sites nor documents covering electronic patient records, because these represent rather specific subgenres of medical texts as compared to the main body of our medical corpora. They also represent divergent communicative settings which might have an effect on the account of linguistic generalizations concerning the main core of investigated sub- corpora with a clear educational profile. Thus our generalizations might be less to the point, or not to be valid, for the mentioned subgenres, which deserve a prior, separate study of their own corpora before the contrasts between the potential sublanguages can be fully elucidated.

3.1.3 Japanese and Russian texts

The Japanese and Russian corpora were collected from the internet. They consist of texts dealing with diabetes and nutrition, especially obesity in connection with diabetes. The corpora were collected with the help of keyword lists, a manual seed set of keywords to which were added terms from UMLS (NLM 2005) and later further equivalent terms from the initial set of retrieved documents.

The document sets were divided into scientific and lay/popularized by native speakers of the languages using their intuition. Table 2 gives a quantitative overview of the texts. For each language the total word count is approximately 100,000 words of scientific texts and 200,000 words of popularized texts.

language	text type	no of texts
Japanese	scientific	199
Japanese	popularized	426
Japanese	total	625
Russian	scientific	45
Russian	popularized	150
Russian	total	195

Table 2: Japanese and Russian texts

4 Investigated variables

4.1 Readability

Readability is an objective, but rather crude, measure that estimates the difficulty in reading text (without considering layout, familiarity of the subject or subject complexity).

There are a number of readability indices available in the literature. The Flesch Reading Ease test (FLESCH) scores documents according to the following formula (higher scores indicate documents that are harder to read):

$$206.835 - 1.015 \times \frac{total\ words}{total\ sentences} - 84.6 \times \frac{total\ syllables}{total\ words}$$

The basic idea is that each group of contiguous 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. FLESCH estimates the reading comprehension level necessary to understand a written document. For a given document, FLESCH 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.

The Flesch-Kincaid Grade Level score transforms the Flesch score into years of education necessary to understand a text. The same meaning applies to the Fog-Gunning index (FOG).

The LIX index is a popular readability index in Scandinavia developed by Björnsson (1968). LIX is defined as $Lm + Lo$ 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 and 50 usually indicates newspaper language and 50–60 professional language.

4.1.1 English

We computed the FOG, FLESCH and FLESCH-KINCAID indices, alongside a number of other measures (detailed in table 3).

	case studies	Merck medics	Merck patients	stories
word count	86908	61032	55154	78668
clean word count	81771	58311	54268	77543
types	8.80%	10.68%	7.49%	8.61%
% complex words	30.17%	30.68%	19.97%	11.27%
avg syllables/word	2.08	2.12	1.79	1.52
avg words/sentence	17.02	18.68	20.89	18.17
FOG	18.87	19.74	16.34	11.77
FLESCH	13.67	8.46	33.97	59.76
FLESCH-KINCAID	15.58	16.72	13.71	9.43

Table 3: Corpus complexity indices for English subcorpora

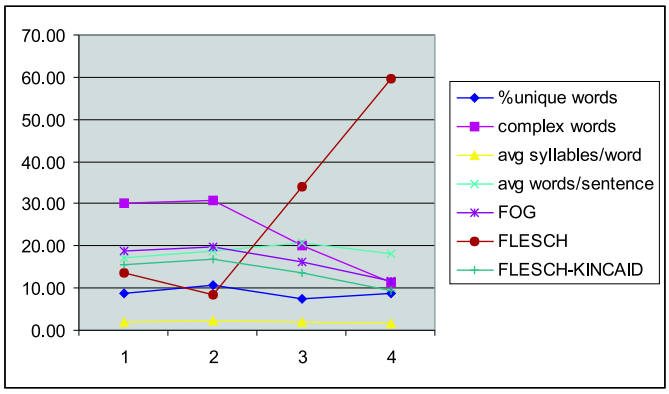


Figure 1: Readability measures for English subcorpora (1: case studies; 2: Merck medics; 3: Merck patients; 4: stories)

The FLESCH and FOG readability indices confirm the intuition that Expert-Expert texts are more difficult to understand than both Expert-Lay

and Lay-Lay texts. The most difficult to read appear to be the texts in the Merck-medics corpus (the FLESCH-KINCAID score indicates 16 years of education necessary to understand them), whilst the easiest are the patient testimonials (can be understood by an average person with 9 years of education). This fact is supported by both the distribution of complex words (i.e., words with more than 3 syllables) and the average number of words per sentence. If we perform the same type of analysis on the BNC subcorpora we notice a similar trend in difficulty, with the Expert-Expert texts being the most difficult to understand and the GP consults the easisest (table 4). The only surprising result is the fact that Lay-Lay texts appear to be more difficult to understand than the Expert-Lay ones.

	Expert-Expert	Expert-Lay	Lay-Lay	GP consults
% complex words	13.77%	6.82%	6.64%	1.58%
avg syllables/word	1.97	1.69	1.63	1.31
avg words/sentence	21.42	14.65	20.30	5.93
FLESCH	18.19	49.41	48.24	90.16
FLESCH-KINCAID	16.04	10.00	11.57	2.16

Table 4: Complexity indices of the BNC subcorpora

4.1.2 Swedish

Two readability tests were applied on the texts in order to determine the difficulty level of the writing style, the FLESCH and LIX indices; see table 5. 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.

For a description of the vocabulary knowledge, we used the notions of lexical originality, LO $= no\ of\ unique\ tokens \times 100/total\ no\ of\ tokens$, lexical density, LD $= no\ of\ lexical\ tokens \times 100/total\ no\ of\ tokens$ and lexical sophistication, LS $= no\ of\ advanced\ tokens \times 100/no\ of\ lexical\ tokens$. 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 Laufer & Nation 1995. Although LD exhibits similar results in the two texts, the LO and LS figures were clearly higher in the expert texts.

index	expert	non-expert
FRI	19.89	42.43
LIX	47.60	37.90
LO	19.09	15.60
LD	50.62	51.60
LS	17.69	4.60

Table 5: Readability indices for Swedish subcorpora

4.2 Lexicon

4.2.1 English

Word statistics Firstly, we constructed frequency lists for content words and lemmas (content words were selected by using the Cornell list of stop words) and calculated the percentage of word/lemma types in each of the corpora (see appendix 1 for a list of the top 100 most frequent content words in the four corpora). In order to assess the over- or under-usage of the content words in one corpus compared to all the others we applied a log-likelihood measure (see appendix 2 for the top 50 words with significant differences for each of the 6 pairs of corpora).

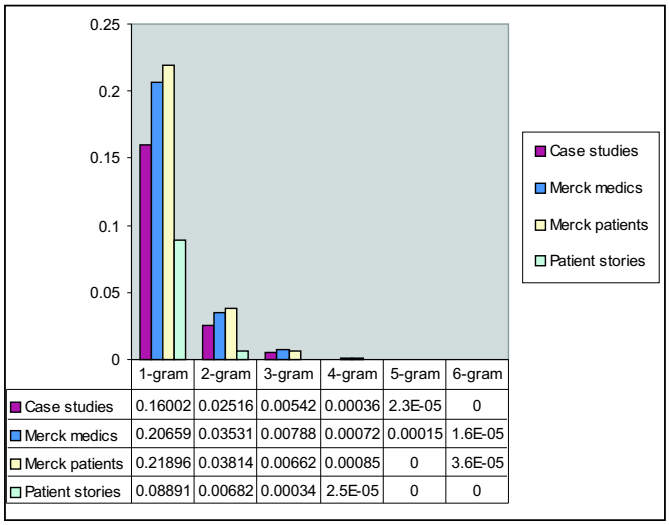
We also computed the frequency of "outsiders" by performing a BNC look-up and identifying words that do not appear in the BNC. This experiment was intended to give us uncommon words, which would presumably have a highly technical content. However, since most of our texts were written using American spelling, the results obtained were less than reliable. A tf.idf-based ordering of the word types based on the BNC proved equally unsatisfactory, since it returned in the top 20 words that were mis-spelled and thus not present in the BNC.

MeSH terminology In order to assess the medical content of the corpora we performed a series of experiments by identifying MeSH terms in the corpora and computing a series of parameters:

- MeSH frequency and types count (see appendix 4 for the top 100 most frequent MeSH terms)
- log-likelihood values to compare the overuse of MeSH terms across corpora (see appendix 5)
- Frequency of 1-, 2-, 3-, 4-, 5- and 6-gram MeSH terms
- distribution of MeSH terms in the 16 top-level MeSH categories

As it can be seen in Figure 2, our experiments confirmed the intuition that texts written by medical experts contain a significantly higher number of medical terms than texts written by non-experts, as well as significantly higher number of complex MeSH terms (i.e., consisting of 3 or more words). However, we have found that an unexpectedly high number of MeSH terms are used in the texts written by medics for the benefit of patients. At this point, we do not know if this is a characteristic of the Merck manual alone (since this is our only source of Expert-Lay texts) or if it reflects a common approach to producing texts for patients. Although the patient testimonials contain an unexpectedly large number of MeSH terms, it can be noticed that these are mainly one-word terms and the frequency of terms drops sharply for terms longer than 2 words.

A closer look at the categories to which the MeSH terms belong shows an uneven distribution of categories across corpora (Fig.3). Patient testimonials contain a higher number of MeSH terms in the less technical MeSH categories, such as *Geographical location*, *Education, sociology and social phaenomena*, *Persons*, *Technology of food and beverages*, *Psychiatry and Psychology*. In the expert-written texts, a higher number of MeSH terms come from the more technical *Diseases* and *Chemicals and drugs* categories. This difference is increased even further if we look only at MeSH terms longer than 2 words, with patient testimonials containing almost exclusively medical terms in more common use.



	1-gram	2-gram	3-gram	4-gram	5-gram	6-gram
Case studies	0.16002	0.02516	0.00542	0.00036	2.3E-05	0
Merck medics	0.20659	0.03531	0.00788	0.00072	0.00015	1.6E-05
Merck patients	0.21896	0.03814	0.00662	0.00085	0	3.6E-05
Patient stories	0.08891	0.00682	0.00034	2.5E-05	0	0

Figure 2: Distribution of MeSH terms into categories in English subcorpora

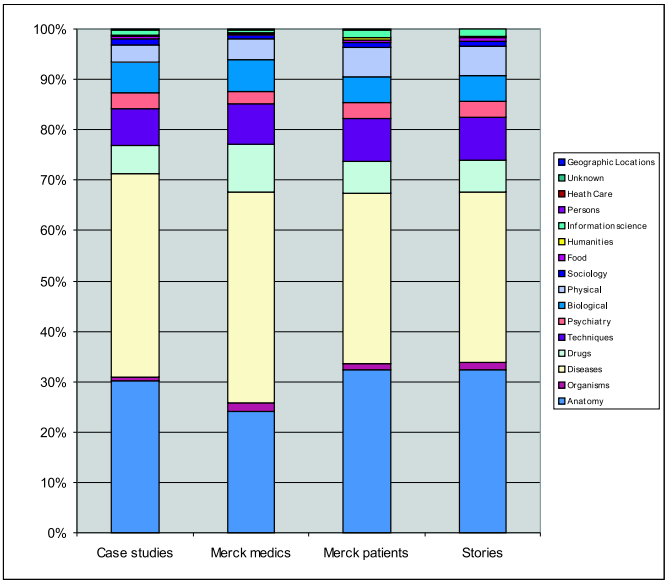


Figure 3: Distribution of MeSH terms into categories in English subcorpora

4.2.2 Russian

Personal pronoun usage We expect to find more personal pronouns in the popularized texts than in the scientific documents. In the case of the Russian material, a relatively high frequency of occurrence of the first person singular subject pronoun *ja* is a strong indicator that we are dealing with a popularized text. In the Cyrillic alphabet, this is one letter (the last letter in the alphabet), which means that there may be some cases of some homonymy, mainly with the initial letter 'Ja' (or Ja.) as used in names. Still, the difference is great: 7.12 occurrences per document in the popularized subcorpus versus 0.29 occurrences in the scientific texts.

The second person singular subject pronoun *ty* is likewise an indicator of popularized text, although less strong, with 0.92 against 0.37 occurrences, but noteworthy is that all 14 occurrences in the scientific subcorpus are found in one document only.

The first person plural subject pronoun *my* again is more frequent in the popularized subcorpus than in the scientific texts: 2.25 and 1.00 occurrences, respectively.

The second person plural subject pronoun is no different: 4.45 versus 2.55 occurrences.

In the case of the third person subject pronouns, there is no difference

between the two subcorpora.[2]

4.2.3 Swedish

Quantitative data Finally, using frequency bands,[3] 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 subcorpora 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 Laufer & Nation 1995. The results (table 6) 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 73.35% of all lemmas in the expert and 80.24% in the non-expert texts. The overlap of outsiders between expert/non-expert texts is very low, only 225 types or $\approx 4\%$ (see section 5 for a discussion).

frequency band	expert	non-expert
group 0	73.35%	80.24%
group 1	7.47%	6.56%
group 2	3.59%	2.97%
group 3	2.17%	1.84%
group 4	1.24%	1.05%
group 5	1.11%	0.92%
group 6	0.71%	0.68%
group 7	0.87%	0.57%
group 8	0.53%	0.46%
outsiders	8.96%	4.71%

Table 6: Frequency-band profiles for Swedish subcorpora

Using a number of different counts, the quantitative characteristics of the selected textual material are summarized in table 7, showing that:

- 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

[2]But these pronouns are not restricted to human or even animate referents, rather they are used according to the grammatical gender of their antecedent (or referent), which means that they are not "personal" in the literal sense of this word.

[3]The tool used for the frequency band analysis of the Swedish subcorpora was developed by M. Stissing (Århus komm. Sprogcenter).

fewer word forms but are repeated more often (lexical variation). In the non-expert texts, on average, any word form is repeated nearly 8.8 times, as opposed to 6.9 times in the expert texts. Since TTR is a crude measure of lexical variation, which furthermore decreases systematically, but nonlinearly, as the text length increases (i.e., the sample mean – or average number of instances of each type – grows larger with increasing text length; see, e.g. Baayen 2001), 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. Lebart, Salem & Berry 1998). sTTR shows a similar picture to TTR, but with lower figures overall;

- the average length of nouns is greater in the expert texts; longer words are an indication of technical terminology (cf. Bodenreider & Pakhomov 2003);

- there is a small difference in the number of compound forms: 12.1 (36.1% unique) in the expert texts compared to 10.4 (31.6% unique) in non-expert ones. The high percentage in both cases can be explained by the fact that Swedish is a compounding language. The most common compounds in the expert texts have been the terms: *hjärt∼svikt*, *hjärt∼infarkt*, *hjärt-∼kärl∼sjukdom*, while the most common compound terms in the non-expert texts have been: *blod∼propp*, *hjärt∼infarkt*, *hjärt∼svikt*;

- the token/sentence ratio (TSR) is higher in expert texts (18.7 compared to 14.8 for non-expert texts). The TSR value for sentences that include at least one verb, TSRverb, increases 1.7 points to 20.4 for expert and 1.3 points to 16.1 for non-expert texts;

- there is a significant difference in the number of "pure" acronyms, such as *NSAID*, *ASA* and *PCI* and also acronyms in compound forms, such as *TNF-alfa*, *WPW-syndrom* and *BNP-test* 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 12.7% main verbs compared to 10% in expert texts, and 2.9% auxiliaries compared to 1%; 4.3% personal pronouns in non-experts compared to 2.1% in expert text.

We now turn to a contrastive linguistic overview of the vocabulary which complements its quantitative profile as just presented. The point of

measure	expert	non-expert
tokens/types	84,787/12,270	84,915/9,554
TTR	1:6.9	1:8.8
sTTR	1:3.4	1:3.9
nouns (avg len of nouns)	21,381 (9.5)	20,223 (8.67)
compounds	10,252 (12.1%)	8,857 (10.4%)
unique compounds	4,435 (36.1%)	3,027 (31.6%)
TSR	18.7	14.8
TSRverbs	20.4	16.1
"pure" acronyms	847	224
acronyms in compounds	423	84
parts of speech		
common nouns	21,381 (25.2%)	20,223 (23.8%)
proper nouns	2,575 (3%)	2,094 (2.4%)
main verbs	8,403 (10%)	10,841 (12.7%)
aux. verbs	8,371 (9.9%)	6,778 (7.9%)
adj./participles	8,371 (9.9%)	6,778 (7.9%)
pers. pronouns	1,823 (2.1%)	3,708 (4.3%)
other pronouns	1,538 (1.8%)	1,842 (2.1%)
others	38,918 (47%)	36,909 (43.9%)

Table 7: Quantitative profiles of Swedish subcorpora

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*;
- 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 pronounced 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 (Nordman 1992), 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, etc.

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 include:

- 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

MeSH terminology The analysis discussed so far was extended by the use of a Swedish MeSH (Medical Subject Headings)[4] 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 4.1.2) 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. Zeng et al. 2005). Our findings revealed that in the expert texts, there were 4,620 complete MeSH matches (e.g. "<mesh tag="A07.231.114">artär</mesh>", i.e. artery) and 409 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 6,144 complete MeSH matches and 277 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

[4]MeSH is the controlled vocabulary thesaurus of the U.S. National Library of Medicine (NLM). 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/>.

order to see whether the distribution of these figures is significant, we applied the χ^2 statistic 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). χ^2 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 8) indicate the occurrences of terms in the six hierarchies for the two types of text. The returned χ^2 figures (degree of freedom=5) indicate in all four cases that the difference is significant.

	expert	non-expert
complete match χ^2=819, $p \leq 0.001$	4,620 (516+5+1,914+754+1405+26)	6,144 (1,937+13+2576+635+971+12)
unique compl. m. χ^2=33.7, $p \leq 0.001$	941 (171+5+327+154+270+14)	847 (240+7+272+143+178+7)
partial match χ^2=101.5, $p \leq 0.001$	409 (11+2+153+42+201+0)	277 (37+10+97+52+63+18)
unique partial m. χ^2=38.7, $p \leq 0.001$	241 (10+2+80+23+126+0)	129 (23+7+41+19+38+1)

Table 8: Distribution of MeSH annotations (A+B+C+D+E+F) in Swedish subcorpora

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 laypeople in understanding medical language, if the information is made accessible to them via an online dictionary (cf. Smith & Fellbaum 2004). 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 9).

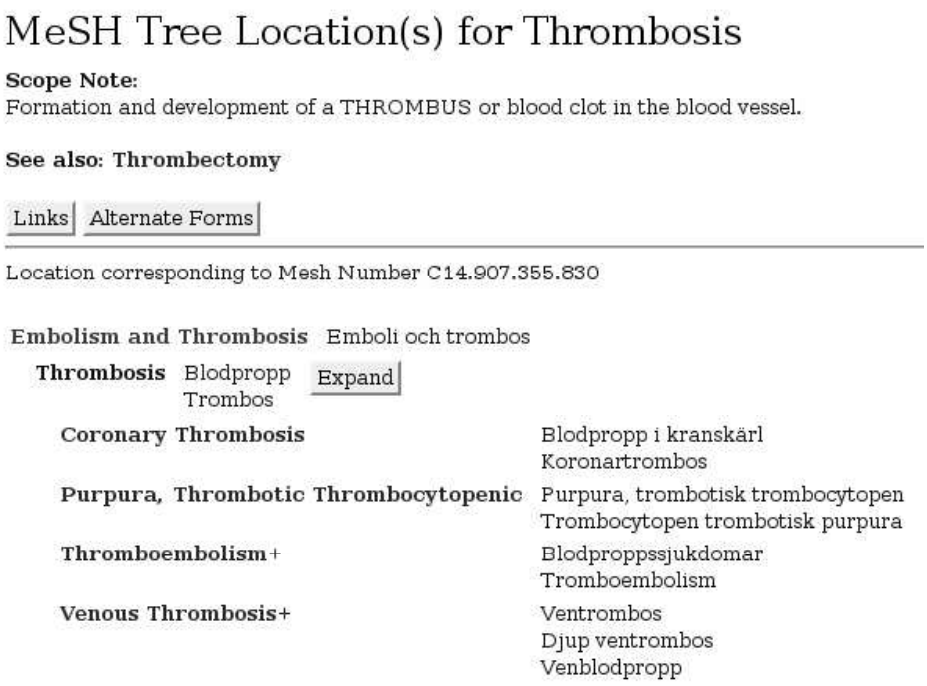
It is of importance to note that these two lists share four out of the six words, which means that both of them capture the most central terms in the cardiovascular domain. The partially different ranking of these four words can be explained partly by differences in the textual material, partly by different preferences of laypeople and professionals for describing health conditions. Laypeople tend to focus on symptoms and professionals on diagnoses; hence the difference in the ranking of *hjärta*.

expert texts		non-expert texts	
182	*hjärtsvikt* 'cardiac insufficiency'	368	*hjärta* 'heart'
162	*stroke* 'stroke'	347	*blodpropp* 'thrombus/thrombosis'
156	*hjärtinfarkt* 'heart attack'	253	*stroke* 'stroke'
104	*hjärta* 'heart'	207	*hjärtinfarkt* 'heart attack'
87	*hjärt-kärlsjukdom* 'cardiovascular disease'	156	*propp* (lit.) 'clot' (short for *blodpropp*)
62	*hjärtstopp* 'cardiac arrest'	147	*hjärtsvikt* 'cardiac insufficiency'

Table 9: Most frequent keywords' related terms in Swedish subcorpora

Diagnostic terms like *stroke* and *hjärtinfarkt* 'heart attack' are ordered according to the same ranking sequence on both lists. For laypeople 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 disease' happens to occupy the fifth position on the expert keylists and the eleventh on the non-expert keylist, which means that the term can be considered familiar even to laypeople. 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 key to correct disambiguation often lies in 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 hierarchy, in which the node *blodproppssjukdom* (thromboembolism) is a hyponym of blodpropp (thrombosis) whose top hyperonym is the node *hjärt-kärlsjukdomar* (cardiovascular diseases); see figure 4. Thus the second, disease reading is mediated in MeSH via the thesaurus structure. 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 (347 occurrences). This observation confirms those of Brown, Price & Cox (1997). Clinical terms are by necessity complex and not easily amenable to being represented in patient language without a full definition; a hierarchical placement of terms has proved beneficial in orienting a patient in the meaning of the term. This strategy can contribute to the fact that non-

professional language has greater variability in meaning, which results in choosing a superordinate, more general term, instead of a subordinate one.


MeSH Tree Location(s) for Thrombosis

Scope Note:
Formation and development of a THROMBUS or blood clot in the blood vessel.

See also: Thrombectomy

Links | Alternate Forms

Location corresponding to Mesh Number C14.907.355.830

Embolism and Thrombosis	Emboli och trombos
Thrombosis Expand	Blodpropp Trombos
Coronary Thrombosis	Blodpropp i kranskärl Koronartrombos
Purpura, Thrombotic Thrombocytopenic	Purpura, trombotisk trombocytopen Trombocytopen trombotisk purpura
Thromboembolism+	Blodproppssjukdomar Tromboembolism
Venous Thrombosis+	Ventrombos Djup ventrombos Venblodpropp

Figure 4: Part of the MeSH hierarchy (from <http://mesh.kib.ki.se/>)

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 laypeople 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. Kokkinakis, Toporowska Gronostaj & Warmenius 2000).

4.3 Grammar

4.3.1 English

Part-of-speech and syntactic role statistics A set of similar experiments to the ones described in section 4.2.1 were performed on the parsed texts in order to obtain frequency counts for various morphological and syntactical categories. The corpora were tagged with the CLAWS5 tagset[5] using a Brill tagger and parsed with the RASP parser (Briscoe, Carroll & Watson

[5]See <http://www.comp.lancs.ac.uk/ucrel/claws5tags.html>.

2006). See tables 10 and 11 for a summary of the results (complete listings of the various subcategories of part-of-speech can be found in appendix 3).

Again, this analysis did not return any surprise findings. The larger proportion of personal pronouns and past tense verbs in the patient testimonials is an indication of the more personal nature of the discourse and of the fact that they relate past histories, and provides no useful information about the style of discourse which we can extrapolate on in generating patient reports.

POS	Case studies	Merck patients	Merck medics	Stories
common noun	30.85	31.51	32.46	19.76
proper noun	3.17	0.92	1.67	2.05
main verb	6.69	8.91	7.43	11.48
auxiliary verb	6.08	7.25	7.32	8.84
adj/participle	17.16	10.42	16.00	6.09
personal pronoun	0.67	0.75	0.39	8.66
other pronoun	0.05	0.32	0.08	0.36
rest	35.32	39.92	34.65	42.76

Table 10: Part-of-speech distribution in English subcorpora

Noun phrase length We measured noun phrase length in the four subcorpora, using roughly the same notion of "noun phrase" as in the Swedish investigation (see section 4.3.2). The figures are presented in table 12.

The picture that emerges conforms partly to expectations, as the patient testimonials show the shortest NPs on average and also the shortest maximal NP length, but now together with the other material directed at lay people, namely Merck patients.

4.3.2 Swedish

Phrase and clause statistics Not only frequency analysis of vocabulary but also measures related to syntactic annotation, manual or automatic, can be a criterion for differentiation of scientific from general texts and style in general (cf. Biber 1995). High occurrences of noun phrases as well as use of infinitival and passive constructions are considered to be representative of scientific texts (Sager, Dungworth & McDonald 1980). 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

function	case studies		Merck medics		Merck patients		stories	
	freq	frac	freq	frac	freq	frac	freq	frac
ncmod	27318	.3356	17514	.2991	13308	.2550	15526	.2117
dobj	11632	.1429	8639	.1475	8338	.1598	11341	.1546
conj	7453	.0916	7846	.1340	4972	.0953	5771	.0787
ncsubj	7716	.0948	5476	.0935	5228	.1002	9855	.1344
det	7973	.0979	4061	.0694	5816	.1114	9112	.1243
iobj	5345	.0657	3769	.0644	3422	.0656	3557	.0485
aux	2359	.0290	2672	.0456	2594	.0497	3806	.0519
xcomp	3481	.0428	2440	.0417	2236	.0428	3841	.0524
passive	1990	.0244	1753	.0299	1491	.0286	1263	.0172
ta	2500	.0307	1556	.0266	1215	.0233	1417	.0193
ccomp	1381	.0170	1195	.0204	1460	.0280	4397	.0600
xmod	998	.0123	652	.0111	683	.0131	1035	.0141
cmod	552	.0068	515	.0088	938	.0180	1106	.0151
obj	305	.0037	198	.0034	158	.0030	364	.0050
pcomp	106	.0013	93	.0016	95	.0018	351	.0048
pmod	87	.0011	68	.0012	102	.0020	104	.0014
obj2	114	.0014	45	.0008	28	.0005	259	.0035
argmod	22	.0003	25	.0004	59	.0011	111	.0015
csubj	16	.0002	13	.0002	4	.0001	47	.0006
xsubj	17	.0002	13	.0002	40	.0008	44	.0006
comp	36	.0004	11	.0002	4	.0001	0	.0000
arg	4	.0000	4	.0001	0	.0000	28	.0004

Table 11: Distribution of syntactic functions in English subcorpora

domains, which do not include medical language, are given in Nordman 1992.

To our knowledge, the syntactic properties of Swedish medical language have not been examined previously. Therefore, we applied a Swedish parser to the part- of-speech annotated version of the corpus; for a description of the parser, see Kokkinakis & Johansson Kokkinakis 1999. Table 13 summarizes the findings regarding the syntactic analysis.[6] The number of "medical" noun phrases, i.e., where the head is a medical term, the active main clauses and all types of subordinate clauses is higher in the non-expert texts. The difference between the passive constructions is small. Extensive lists, of for instance symptoms and drugs, in the expert

[6]Note that the "noun phrases" referred to in table 13 are "simple", or "basic", noun phrases, which do not contain prepositions, relative pronouns or conjunctions (unless the NP or part of it has been analyzed as a named entity in the named entity recognition step which precedes the syntactic parsing proper, in which case the NP may contain prepositions, etc.).

measure	Case studies	Merck patients	Merck medics	Stories
no of NPs	21443	13631	16026	21764
avg length (no of tokens)	2.14	1.90	1.95	1.69
max length (no of tokens)	24	13	20	13
std dev	1.186	0.855	0.974	0.854

Table 12: Noun phrase length in English subcorpora

syntactic construction	expert	non-expert
noun phrases	17,392	13,841
noun phrases ≥ 5 tokens	439	185
"medical" NPs	3,683	5,420
prep. phrases	9,854	7,951
passive main clauses	1,178	1,093
active main clauses	4,348	5,479
infinitive-adverbial-relative clauses	655-967-811	967-1,511-1,079
questions	316	517

Table 13: Syntactic constructions in Swedish subcorpora

texts, might be an explanation of these results: i.e., a relatively smaller number of clauses, a high number of noun and prepositional phrases. An interesting feature of the explanative profile of the lay corpus is the use of the syntactic bigram *om man* 'if one', consisting of a conditional subjunction, followed by an indefinite pronoun which functions as a subject of a subordinate clause. The use of this bigram is ten times more frequent in the lay corpora (240 occurrences), as compared to the professional texts (21 occurrences), which is a relevant contrastive property of the two subcorpora.

This particular syntactic construction is also semantically interesting. The chameleon-like referential meaning of the indefinite pronoun varies with the type of the corpus. In the texts written by medical experts, it usually carries reference to medical professionals (in 19 out of the 21 uses), while in the lay corpora it refers mostly to health care consumers. This observation has, in turn, further consequences for the choice of co-occurring semantic types of the predicates and their semantic restrictions. In case of reference to professionals, the predicates often turn to be agentive verbs performing some medical actions, in contrast to laypeople that are stricken with illness or seen as potential patients (also in the grammatical sense). The high frequency of the suppositive structures in the lay sub-corpora contributes to their readability, which is no doubt appreciated by laypeople.

4.4 Semantics and pragmatics

4.4.1 Japanese

Politeness In Japanese, the politeness value of certain linguistic items can be used to as a criterion for distinguishing the two text categories. Politeness values are expressed on the predicate (verb, adjective or noun) standing at the end of the sentence. Distinguishing two politeness levels, polite and neutral, we find marked differences in their use:

The polite style is much more frequent in the popularized texts (found in 32.52 sentences per document on average) than in the scientific subcorpus (an average of 13.62 sentences per document), whereas the reverse holds for the neutral style (23.46 and 1.98 sentences per document on average).

4.4.2 Russian

Conditional mood We would expect that uncertainty would be marked more often in the popularized subcorpus. In Russian, the invariable particle *by* – marker of the conditional mood – is a common way of expressing uncertainty, hence its frequency of occurrence should correlate well with the occurrence of uncertainty in the discourse. In particular, many occurrences of *by* will be in conditional clauses (where it occurs together with the conditional subjunction *esli*). Cf. the discussion earlier (in section 4.3.2) about Swedish conditional clauses as a discriminating feature for nonprofessional texts.

The particle *by* turns out to be a very good discriminating feature for the distinction between scientific and popularized texts, with an average of 0.74 occurrences per document in the former and 1.44 in the popularized subcorpus of Russian.

4.4.3 Swedish

Adjective use In the context of biomedicine, adjectival modification has been studied both for the identification of hierarchical relations among biomedical terms and also in applications such as automatic construction of terminologies and ontologies (cf. Bodenreider & Pakhomov 2003). For the contrastive characteristics of adjectival modification (including participles) across the two corpora, we made a comparison using the frequency of occurrence of each string found in either of the two corpora and the likelihood ratio test (which is also known as the G-statistic; Dunning 1993). All adjectives and participles were extracted from the two corpora, lemmatised and compared. There are more adjective/participle tokens and

types in expert texts (8,371 tokens, 2,892 types, 1,669 lemmas) than in non-expert texts (6,778 tokens and 1,786 types, 1,182 lemmas). Table 14 shows the adjectives with the highest log-likelihood values.

expert texts		non-expert texts	
119.906	klinisk	110.645	kallad
47.570	randomiserad	71.919	hämtad
39.769	diagnostisk	67.345	mycken
30.901	prediktiv	51.190	vanlig
30.901	asymtomatisk	43.463	läckande
30.68	skild	39.932	blodförtunnande
28.521	malign	34.874	viss
26.142	kateterburen	30.768	lätt
24.953	neuropsykiatrisk	28.363	oregelbunden
22.847	tillgänglig	27.352	förträngd
22.574	optimal	26.690	bra
22.377	natriuretisk	25.792	svår
21.385	postoperativ	25.020	hög
21.349	ischemisk	22.522	vätskedrivande
20.196	relaterad	22.522	inre
19.007	epidemiologisk	21.276	flest
18.422	cerebral	20.912	smärtstillande
17.818	koronar	19.506	förebyggande
16.630	associerad	19.102	stark
16.630	adekvat	17.887	gammal

Table 14: Adjectives most characteristic of each text type in Swedish sub-corpora

The higher the log-likelihood figure is in expert texts the greater the deviation from the non-expert texts and vice versa. The respective lists are also semantically heterogeneous; the adjectives in the expert texts are medically significant in contrast to those in the non-expert texts, which are generally descriptive. The lack of overlap between the two lists indicates different collocational preferences. Enhancement of lexical resources with such information seems to be supportive for the users.

4.5 Other variables

4.5.1 Japanese and Russian: Document layout and typography

Images In documents collected from the internet, particularly web pages (HTML documents), the use of images could provide a clue as to the genre

of the documents. Images are easy to find, as they are normally introduced by the HTML element <img>.

The number of images turns out to be discriminating in Russian – with a mean of 44.0 images per document in the popularized subcorpus against 15.5 in the scientific texts – whereas the situation in Japanese is that both text types display exactly the same figure, 21.54 images per document on the average.

Simple counting of images is easy but unfortunately does not tell us anything about the types of image used and their function.

Tables Tables should appear more often in scientific prose than in popularized texts. Again, there are a couple of HTML elements that can be counted directly and mechanically, i.e., <table> and possibly <caption>. Unfortunately, the table facility of HTML is often used by webpage authors purely for purposes of layout, e.g. as an alternative to frames or lists.

In any case, there turns out to be a difference between the two text types in Russian with respect to this criterion. However, it goes against expectations: There are more tables (HTML <table> elements) in the popularized texts (13 per document on average) than in the scientific texts (8.5 per document). This could reflect the fact that the information in the popularized texts is more "digested", turned into a more palatable format for the average citizen, whereas the scientific texts are "cruder" and for this reason as well demand more from their reader.

Lists In the case of Russian, the HTML list elements <ul> and <ol> serve as good indicators of the genre of the text. Lists are used more often in the scientific subcorpus – with 1.55 occurrences of <ol> and 2.26 occurrences of <ul> on the average per document – than in the popularized texts, with only 0.40 occurrences of <ol> and 0.50 occurrences of <ul> per document.

Typographical marking of emphasis Emphasis can be marked in HTML documents by the use of the elements <i> (italics), <b> (bold) and (strong emphasis, normally displayed as bold). There are differences in how frequently text emphasis is used between the two text types, but they point in different directions in Russian and Japanese.

In Russian, the emphasis markers are used more in scientific texts (<i> appears 13.18 times and <b> 30.97 times per document on average) than in the popularized subcorpus (2.61 <i> and 8.01 <b>).

In Japanese, <b> and appear together 13.05 times per document on average in the popularized texts, as against 8.87 in the scientific

subcorpus. The <i> element is too little used overall in the Japanese material for any conclusions.

Punctuation Punctuation can reflect the complexity of the text in terms of how complex its sentences are (comma, semicolon, colon, parentheses, etc.), but also convey information about emotivity (question and exclamation marks), and finally show references to other sources of information (quotation marks).

In the Russian corpora, emotion-conveying punctuation seems to be underused in the scientific subcorpus compared to the popularized texts. In the scientific subcorpus there were 4.0 question marks and 1.76 exclamation marks per document, whereas the corresponding figures for the popularized texts were 6.25 question marks and 5.63 exclamation marks.

Unexpectedly, quotation marks are used much more often in the popularized subcorpus in Russian than in the scientific texts: 17.32 occurrences per document in the former and only 3.76 in the scientific subcorpus. Still, this seems to be a strong distinguishing characteristic for the two subcorpora. Apparently most instances of quotation marks in both subcorpora are not quotations in the narrower sense, but the quotation marks are rather employed as a general distancing mechanism (meaning, roughly: "Someone else is responsible for (the formulation of) this word/phrase"), in the scientific texts for referring to laymen's terms

5 Per-language conclusions

5.1 English

The quantitative analysis of the corpus produced few surprise findings. The distribution of morphological and syntactical constructions is indicative of the general nature of discourse (technical vs non-technical), and does not present any obvious additional features that could separate the medical domain from any other technical domain. The lexical analysis shows insignificant differences in lexical variety between subcorpora, although with a slight increase in the percentage of lexical types in texts written by experts for experts. Interesting results were found in the analysis of MeSH terms, with the highest distribution found in the texts written by experts for patients. On closer look, we have found that this higher frequency is not due to the more technical nature of Expert-Lay texts, but to the fact that medical concepts are often accompanied by explanations of terms using other (sometimes more technical) medical concepts, or simply by synonyms (as in alternative trade names for drugs). Further analysis of MeSH terms found that the medical terms in patient testimonials are of

a less technical nature than in texts written by experts, and this becomes especially obvious when looking at longer MeSH terms (2 or more words).

5.2 Japanese and Russian

Since Japanese and Russian were investigated contrastively in the original article, we here note such findings that are common to the two languages, while those traits that are characteristic of one of the languages only, are discussed below, as we discuss which crosslinguistic generalizations can be made on the basis of the corpus study (section 6).

For Japanese and Russian, mostly other variables were investigated than in the case of the other two languages of this corpus study. It seems, however, that Russian sides with English and Swedish in using less personal pronouns in lay text than in professional text. Arguably, pronoun usage differences in these three languages serve the same purpose as the different frequencies of polite and neutral forms in the Japanese texts, where polite forms are a characteristic of the lay register and neutral forms predominate in the professional texts.

There are common differences in the use of some (HTML) typographical elements, supporting the notion that the professional texts are more of an "information conduit" (mainly one-way) than a communication channel (for interaction).

5.3 Swedish

The binary division of the examined medical corpora alludes to the potential target groups represented by health care workers and laypeople. The two target 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 our frequency based and lexical analysis of the vocabulary it is also clear that the stock of medical terms in the non-expert texts grows. According to Grabarczyk (1987: 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 Grabarczyk's (1987) observation for the issue of patient empowerment is obvious. The dynamic vision of vocabulary expansion entails also an increase in medical literacy, but at the same time it contributes to segregation of health consumers with respect to their initiation in the 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 the refinement a number of NLP applications (e.g. generation and health information retrieval and understanding; Zeng & Tse 2006).

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 will provide 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 by an extensive analysis of deeper syntactic relations on sentence level as well as phrase level. An in-depth study of the co-ordinated and subordinated structures in different sentences/clauses will be the issues for future work. As Bodenreider & Pakhomov (2003) note, adjectives may be useful to characterize corpora into genres, and thus, adjectival modification can be exploited in applications such information retrieval of biomedical documents. Another relevant research topic is the extraction of the types of syntactic or/and semantic patterns characteristic of the non-professional corpora 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 investigate how readability measures are related to how consumers use and benefit from material on health care information websites.

6 Cross-linguistic generalizations

Looking at the individual studies, we are struck by some results that they have in common. By and large, generally held assumptions about the differences between more formal (professional) and more informal (lay) written registers are confirmed by these studies. Hence, in this respect, at least, they present no real surprises. Their value lies, firstly, in their cross-linguistic focus and, secondly, in the hard data they present in support of these assumptions, and in the directions they give us for moving into hitherto uncharted territory, e.g. more thorough and rigorous investigations of the syntax of the two kinds of text, particularly with a view to propose syntactical transformations that are within the reach of today's language technology.

One intriguing piece of information that emerges from this comparison of the English and Swedish corpus investigations is that for the English "X-Lay" subcorpora, the character of "X" seems to matter. On a number of variables, the Expert-Lay subcorpus (Merck patients) patterns with the two English Expert-Expert subcorpora (case studies and Merck medics) – and with the Swedish expert subcorpus – rather than with the English Lay-Lay subcorpus (patient testimonials – "stories") and the Swedish non-expert subcorpus, or in some cases in-between the two extremes. This seems to be true at least for the following variables:

- percentage of complex words
- medical (MeSH) term distribution
- word length
- sentence length
- percentage of common nouns
- percentage of verbs
- percentage of personal pronouns (also Russian)
- noun phrase length

This could mean that the simple dichotomy expert – non-expert is actually too crude, as discussed briefly in section 3.1.2, and should be replaced by a more many-faceted notion of the kinds of texts involved (see also below). It could also mean, however, that the authors of the Merck patients material have failed to tune their text to the envisaged readership. Only further research can clarify this matter.

6.1 Readability

Readability varies as predicted in the English and Swedish studies, with professional language being more demanding on the reader than lay language.

Readability is an indirect measure of complexity of vocabulary and syntax. Indirect, because length (in words and sentences) can be assumed to correlate with complexity, but of course we cannot make texts easier to read simply by mechanically shortening words and sentences, e.g. by inserting spaces in the middle of words and sentence punctuation in the middle of sentences. Readability measures are a symptom of some underlying linguistic factor(s), much in the same way that temperature as displayed by a thermometer is a symptom; we cannot make our environment warmer or colder by manipulating the thermometer.

In the context of this work, however, the role of readability could be to give us a first quick and dirty indication that something needs to be done with a text in order to make it palatable to a particular group of intended readers. E.g., even if we cannot of course physically shorten the words in a text, we might well consider replacing some words with shorter (near) synonyms. Similarly, on the syntactic level, complex sentences could be transformed into sequences of main clauses.

6.2 Special terminology

Generally, special terms are more frequent in professional texts, i.e., medical terms as found using MeSH. Nonprofessional texts also contain such terms, but in this case the terms are more likely to be:

(a) more general in the sense that they coincide with words in general language (e.g. for body parts);

(b) less specialized in some other way, e.g., they designate a larger anatomical structure or a class of ailments rather than an anatomical detail or a specific disease;

(c) not really medical in the narrower sense, i.e., they belong to such subsections of MeSH as, e.g., *Geographical location*.

6.3 An inordinate fondness for nouns?

The professional texts in this study generally conform to the oft-noted tendency of using more nouns than everyday (written) language, and, correspondingly, of using less verbs. Medical terminology – like specialized

terminologies generally, at least in the languages that we have been investigating – are "noun-heavy"; the majority of the terms are nouns (or the corresponding relational adjectives) or noun phrases. Verbs tend to be semantically empty. On the other hand, this is also a characteristic of formal or bureaucratic language in general, so it is not immediately obvious whether it is the medical character of the texts which is responsible for the preponderance of nouns, or simply their formal nature. This has some implications for how the language of the texts could be made more accessible to non-specialists. In the case of general formal-bureaucratic language, there are normally ways of reducing the share of nouns (generally nominalizations) of the texts, by introducing constructions with semantically non-empty verbs instead.

6.4 Grammar matters

From the corpus studies, we get both direct and indirect information about grammar – morphology, morphosyntax and syntax – in the investigated language varieties/registers. This is an area where we have only begun to scratch the surface, however, and we expect to return to the issue of how grammatical differences reflect differences in register, or put in another way, differences in authorship and (perceived) readership of texts.

Here are some grammatical differences that we have found in our corpus study:

- Sentences are generally longer in professional language than in the lay variety. This statistic was explicitly calculated for the Swedish corpora, and is indirectly available for the English material, since one parameter in the calculation of readability indices (Flesch, Lix) is normally sentence length (in words). Pending a more detailed syntactic analysis, we cannot know if this is due to a greater syntactic complexity on the sentence level (more subordination) or simply because NPs tend to be longer in professional texts (see the next item). We suspect that both factors are present, however.

- Noun phrases tend to be longer in professional texts. This is certainly connected with the general facts that special terminology is made almost exclusively of nouns (see section 6.2), and that professional (and formal, bureaucratic) language favors nouns in general to a greater degree than everyday language.

6.5 Pragmatic features

There seem to be some pragmatic features which distinguish the two main kinds of registers investigated, professional and lay texts. The evidence

for this is mainly indirect, as with some of the grammatical features (section 6.4).

The higher frequency of personal pronouns in the lay texts point to a more personal style in this register. In the case of Japanese, this is also indicated by the different politeness markers found in the two registers. Lay texts also favor past tense verbs, indicating a narrative style: an unfolding of events, rather than a statement of timeless facts.

The usage of punctuation marks in Russian supports the notion that the lay texts are less formal, more "intimate" than the professional texts.

Finally, the conditional mood – expressing (among other things) uncertainty – seems to be a characteristic of lay texts rather than professional texts. How this is to be interpreted remains to be investigated in more detail.

7 Future work

The picture presented here of documents on medical matters falling into one of two categories is of course a grossly oversimplified one. On the one hand, the 'lay' population is quite diversified in its background knowledge, educational level, etc. On the other, 'healthcare professionals' also make up a heterogeneous body of individuals with different educational backgrounds and differing communicative needs. What we are dealing with is a spectrum of texts and a number of communication needs, between doctors and laymen (the case considered here), but also between the various professions within the healthcare system. This means that we need to conduct further investigations of the differential linguistic characteristics of the various communicative settings involved. However, we are even now in a position to formulate some tentative requirements on patient-friendly documentation systems, and to advance some recommendations for the creation of such systems.

Part III
Using language technology for the creation of patient-friendly documents

8 Introduction

The generation of patient-friendly documents will become necessary because of new laws in several European countries. For instance:

- Law on public health no 2002-303 adopted on 4 March 2002 in France;
- Social Services Act in Sweden, Data Protection Act 1998;
- Access to Health Records Act 1990 and Data Protection Act 1998 in UK.

According to these laws and acts, hospitals and medical institutions must be able to provide patients with their clinical documents and, moreover, these documents should be understandable for patients. As many researchers have observed that there are a number of differences between expert and non-expert language, the aim of this report is to propose some recommendations in order to overcome some of these differences and to create patient-friendly documents.

The recommendations can address different kinds of criteria according to the areas concerned. For instance, for computer science they would concern interfaces, data structures and algorithms; for language technology they would describe the choices made when non-linguistic content is transformed into language (words and terms used, syntactic structures, document layout, etc) or when linguistic content is modified to better suit a particular category of reader or listener; for the psychologic area they would specify the ergonomic characteristics of the interface, etc. The purpose of this report is to specify the recommendations as they can be stated from the point of view of language technology in order to be used by a natural language generation system.

In the following, we first define the purpose of creating patient-friendly documents (section 9). The bulk of this part of the report is devoted to the description of recommendations (section 10). We then describe the context in which such documents will be generated, i.e., the Natural Language

Generation demonstrator (section 11) and general principles of the evaluation of the demonstrator and of the proposed recommendations (section 12).

9 Purpose of patient-friendly documents

The task of generating patient-friendly documents can be thought of as fulfilling at least translational (section 9.1) or educational (section 9.2) purposes. Note that these are not mutuallyexclusive, and possibly not exhaustive either.

9.1 Translational purpose

From the translational point of view, the generation of patient-friendly documents is conceived as a translational problem. In this case, expert jargon, and especially the terms used, are translated into patient language. Thus, the main resources needed are two-fold lexicons and terminologies which link controlled terminologies, such as MeSH, Snomed or ICD, to patient vocabularies.

If a more detailed comparison between expert and non-expert documents is performed, it appears that at other linguistic levels (morphology, syntax, etc) more differences can be observed. Thus, the translation can also be performed at these additional levels. In this case, the system should aim at transforming the morphological, grammatical, syntactic etc. structures as well. Translation at these additional levels requires additional resources and databases.

In section 10, the recommendations devoted to the translational purpose are marked *(t)*.

9.2 Educational purpose

The educational purpose of generation of patient-friendly documents goes beyond "simple" translation. The aim is then not only to adapt the content for patients but also to explain to them the objective of treatments and procedures, the meaning of diseases, the anatomy of an organ and its surrounding tissue, etc. In this case, the aim is to help patients to understand their illness and the usefulness of medical treatment, and, in this way, to make the interaction between medical staff and patients more efficient. The resources needed for this purpose are multi-fold. According to the solutions chosen, they and can be provided by different media (i.e., text, image, video).

In the section 10, the recommendations devoted to the educational purpose are marked *(e)*.

10 Recommendations

In this section, we address the recommendations which can be used for the creation and detection of patient-friendly documents and their evaluation. These recommendations are formed of a set of criteria from different levels of the content and structure of documents: morphology (section 10.1), lexicon and terminology (section 10.2), syntax (section 10.3), personalisation (section 10.4) and document layout and presentation (section 10.5). These criteria define a preliminary coarse grouping of problem areas, which however turns out to correlate well with a grouping of the recommendations, according to the specific solutions and resources involved. It is important to note that these different levels and criteria participate all together in the creation and description of expert and non-expert documents, and that they are interrelated inside the discourses which are specific to experts and to non-experts.

At each level of criteria:

- we first present observations of differences between expert and patient languages as they emerged from the research studies,
- and then propose solutions suitable for the adapting of medical documents to patients' needs.

These recommendations have mainly been compiled from the previous literature survey (Åhlfeldt et al. 2006) and from the corpus study described in part II of this report. The legal aspect, that is to whom the generated documents can be distributed, is not addressed here.

10.1 Morphology

The morphology level addresses word formation processes: Definition of the morphological components used for the creation of "new" words and their analysis. Note that this level is related to the level of lexicon and terminology (sec. 10.2), as words formed at the morphological level will further be used for the creation of terms.

10.1.1 Observed problems

It has been observed that medical jargon has a tendency to use Latin terms (Surján & Héja 2003; Krivine 2005; Kokkinakis & Toporowska Gronostaj 2006):

{*axilla*, *armpit*}, {*derm*, *skin*}, {*adip-*, *fat*}

as well as large number of abbreviations and acronyms. Additionally, Bodenreider & Pakhomov (2003) show that longer words are an indication of technical terminology. But according to the Swedish corpus study (Kokkinakis & Toporowska Gronostaj 2006 and section 4.2.3 above), in Swedish documents, there is no important difference of the use of compound forms in expert and non-expert documents, which is due probably specifically to this language.[7]

10.1.2 Solutions

For the generation of patient-friendly documents at the morphological level:

(t) terms and words with local (English, French, Swedish, etc) roots should be preferred. The paraphrasing with the MorphoSaurus (Markó, Schulz & Hahn 2005; Schulz 2007) tool or with synonyms recorded in various biomedical terminologies, such as Snomed, MeSH, can be helpful for this purpose.

(e) anyway, if used, Greek and Latin roots should be explained.

10.2 Lexicon and terminology

The level of lexicon and terminology addresses the creation and especially usage of medical terms in order to introduce and describe medical concepts.

10.2.1 Observed problems

At the lexical and terminological levels, many previous studies have observed differences between expert and non-expert documents. The results indicate that a doctor's choice of vocabulary affects patient satisfaction immediately after a general practice consultation and that using the same vocabulary as the patient can improve patient outcomes (Williams & Ogden 2004). Thus, the common finding states that the terminology used by medical doctors should be adapted to patient knowledge (Bouhaddou & Warner 1995; Waisman et al. 2003; White, Singleton & Jones 2004).

[7] But note that "compounds" and "long words" are two largely independent parameters. True, a compound will on the average be longer than a simplex word, but there are longer and shorter compounds, just as there are longer and shorter simplex words, and it is a fair assumption that the longer compounds will be found more often in the professional part of the corpus.

For instance, in previous research it has been noted that patients could not recognise the equivalence between several synonymous terms, as in the following examples:

- {*bleeding*, *hemorrhage*}
- {*broken*, *fractured*}, {*break*, *fracture*}
- {*heart attack*, *myocardial infarction*}
- {*stitches*, *sutures*}
- {*diarrhoea*, *loose stools*}
- {*cast*, *splint*}

While with other terms, patients have difficulties in defining them, e.g.:

> *metastasis*, *meningitis*, *lethargy*, *virus*, *hypertension*, *strep throat*, *herpes*, *tumor*, *Pap smear*, *uterus*, *fever* (Thompson 2005), *rheumatism*, ...

In sum, any technical medical term can potentially present understanding problems for patients.

10.2.2 Solutions

For the generation of patient-friendly documents at the terminological and lexical level, several solutions are possible:

(e) Use of explanations through a special discharge nurse as well as use of written information (Waisman et al. 2003);

(e) Use of multimedia which is supposed to provide a comfortable environment to learn about medical problems (Miyawaki et al. 1995). Notice that previous projects (Magic (McKeown et al. 1997), MedView (Torgersson & Falkman 2002), Persival (Elhadad & McKeown 2001)) attempted to do so, but often different media were not integrated;

(e) Use of pictures and graphical material for the explanations;

(e) Systematic use of definitions of technical terms as corpora analysis has suggested. In order to collect definitions of medical terms, it is possible to use medical web sites and portals which exist in several countries and languages, and which provide (directly, or link to) publicly available resources:

- the MedLinePlus portal[8] of American governmental health web sites;
- the Health on the Net Foundation[9] portal lists over 1,200,000 accredited web pages in numerious languages;
- the Swedish web site of the National Board of Health and Welfare[10] covers about 150 general terms with their definitions and comments;
- the Cancer Research portal[11] in the UK covers various terms related to this area;
- the French portal CISMeF[12] indexes over 12,000 web pages in French.

(t) Use of patient terms, compiled into "problem lists" (Lauteslager et al. 2002; Miller et al. 2003; Bui et al. 2004). Often, these patient "problem lists" are bootstrapped manually from medical consultation notes (from patient emails or other) and then matched to existing controlled terminologies (Campbell & Payne 1994; Hales, Schoeffler & Kessler 1998; Smith, Stavri & Chapman 2002; Brenna & Aronson 2003; Tse & Soergel 2003; Plovnick & Zeng 2004; Fabry et al. 2005). Notice that only a few resources with patient terms are available for wide usage and that only the *MedlinePlus* resource, included in the UMLS, seems to provide the actual alignment between expert and patient terms. Some of these resources are:

- the UMLS (NLM 2003) resource includes such a layman terminology, *MedlinePlus* (Zeng & Tse 2006), compiled from the MedlinePlus portal. The *MedlinePlus* covers over 1,400 terms;
- the Consumer Health Vocabulary Initiative[13] partners gather patient oriented resources: i.e., *MedlinePlus*, *ClinicalTrials.gov*, *Centers for Disease Control: Topic Index* and *Food & Drug Administration: Information for Consumers*;
- the Medical WordNet (Smith & Fellbaum 2004) initiative, which seems to be currently under development, would propose a resource of layman oriented medical terms aligned with the WordNet network of synsets (Fellbaum 1998);
- the Wikipedia[14] resources offer explanations and encyclopeadic

[8] <http://medlineplus.gov>
[9] <http://www.hon.ch/>
[10] <http://app.socialstyrelsen.se/termbank>
[11] <http://www.cancerresearchuk.org/>
[12] <http://www.chu-rouen.fr/cismef>
[13] <http://www.consumerhealthvocab.org>
[14] <http://www.wikipedia.org>

information on various terms in several languages;

- the Swedish terminology bank of the National Board of Health and Welfare[15] currently covers about 600 search terms recommended for use in communication within health care services and in communication with patients;
- the terminological bank of the Swedish Council on Technology Assessment in Health Care[16] currently covers about 200 terms.

(t) Use of terms paraphrased with the Morphosaurus (Markó, Schulz & Hahn 2005; Schulz 2007) tool or the ones which can be obtained from series of synonymes as recorded in various biomedical terminologies, such as Snomed, MeSH, etc:

myocardial infarction: *cardiac infarction*, *heart attack*, *infarction of heart*.

It seems that patients find technical terms reassuring and respect doctors who use them (Ogden et al. 2003). In this respect, the educational aspect of patient documents should be preferred. In this case, expert documents will remain nearly intact, which means that the information conveyed would not be affected. But, at the same time, it will be necessary to add explanations and definitions for patients could understand these documents.

In order to make decisions automatically about which terms should be explained the findings about empirical evidence about difference in usages of terms by doctors and patients is helpful.

10.3 Syntax

The syntax level deals with sentence structure and use of part-of-speech categories when constructing sentences. There are very few investigations focusing on the syntactical level of medical documents.

10.3.1 Observed problems

As for sentence structure and complexity it is correlated with the readability of documents, i.e., the fact whether is it easy or not to understand the document for a patient.

For instance, Ownby (2005) investigates several aspects of readability, especially sentence complexity and the use of passive voice, and shows

[15] <http://app.socialstyrelsen.se/termbank>
[16] <http://www.sbu.se/ordlista/list.asp>

that expert and non-expert documents are different in these respects. Passive constructions and non-finite clauses are thus known to be representative of scientific texts (Sager, Dungworth & McDonald 1980). Williams (2003) also shows that sentence and word length influence the readability of documents. Sentence complexity can also be observed in the use of punctuation (Krivine et al. 2006) (i.e., the use of commas, semi-colons, colons, parentheses, exclamation marks, question marks, quotation marks, etc.). Thus, as reported in part II of this report, Krivine et al. (2006), Kokkinakis & Toporowska Gronostaj (2006) and the OU NLG group thus have observed that sentences are longer when they are created by experts. E.g. for Swedish, the token/sentence ratio is 18.7 in expert texts and 14.8 within non-expert texts. The Flesch readability index (Flesch 1948) computed by Kokkinakis & Toporowska Gronostaj (2006) and the OU NLG group confirms that the readability of expert documents is lower, compared to non-expert documents, as the length and complexity of sentences and words in them are higher.

The use of conditional structures (i.e., introduced by *if*) in patient documents has been observed in Swedish (Kokkinakis & Toporowska Gronostaj 2006), Russian and French (Krivine et al. 2006). But this seems to affect particularly the content of documents and less the way the content is presented.

As for the use and distribution of part-of-speech categories, Richardson (1996) observed that expert language uses nouns instead of verbs and adjectives instead of nouns. Indeed, a high frequency of nominalisations is characteristic of scientific texts (Nordman 1992).

Moreover, Bodenreider & Pakhomov (2003) have explored the behaviour of adjectival modifiers across the two written genres using texts from Medline[17] and the Mayo clinic[18]. They found that a much greater range of adjectives was used for the wider audience. Grabar & Zweigenbaum (2003) observed the productivity of denominal adjectives when they occur in the same document with their base nouns:

> {*stomach/N*, *stomachal/A*}, {*diabetes/N*, *diabetical/A*}, {*asthma/N*, *asthmatic/A*}

The study has been conducted in French on a general language newspaper (*Le Monde*) corpus and a medical corpus collected from the medical portal CISMeF. They found that such adjectives are more productive and frequent in the medical corpus. Their use is then correlated with the use of their base nouns. This observation can be explained by the fact that such adjectives are components of medical terms, which are widely used

[17] <http://www.ncbi.nlm.nih.giv/entrez>
[18] <http://www.mayoclinic.com>

in specialised texts. For instance, in such terms, nouns and their adjectives, which refer to the human anatomy, can be used in order to localise diseases, procedures, etc:

> *aortic valve prosthesis*, *injury of subclavian vein*, *intravenous injection*

Kokkinakis & Toporowska Gronostaj (2006) and the OU NLG group investigated the use of verbs in expert and non-expert documents, finding e.g., that Swedish expert documents contain respectively 10% and 1% of main and auxiliary verbs, while the non-expert documents in the Swedish corpus contain 12.7% and 2.9% of these categories (see section 4.2.3 above). The use of verbs is thus more common in patient documents. As correlative to this fact, the use of pronouns, addressed here as "personalisation" (section 10.4), is also more frequent in non-patient documents. As more correlation to this fact, the use of nouns is less frequent. Indeed, in this case, verbs such as (*investigate*, *inform*, *observe*) will be preferred to corresponding nouns (*investigation*, *information*, *observation*).

10.3.2 Solutions

For the generation of patient-friendly documents at the syntactical level, several solutions are possible:

(t) Content can be organised and formulated using simple and short sentences (Williams 2003).

(t) The sentences should not contain passive, infinitival structures.

(t) Noun phrases which encode terms can be syntactically transformed through the POS tagging, shallow parsing and then transformation rules, for instance with Faster (Jacquemin 1999)

> *tracheal stenosis* $\Rightarrow$ *stenosis of the trachea*
> *catheter ablation of tissue of heart* $\Rightarrow$ *excision with catheter of tissue of the heart*

(e) Anyway, if the use of long sentences is preferred, for instance in order to not affect the meaning conveyed, these sentences can be illustrated with additional information, such as definitions and paraphrases (sec. 10.2), graphical material (sec. 10.5), etc. The access to the information encoded by such sentences can be interactive so that the user could read or listen to them again and again.

10.4 Personalisation or use of personal pronouns

What we call *personalisation* is the presence of the reader or audience in the document, often through the use of personal pronouns like *you*, *he*, *they*, etc. This criterion corresponds to the level at which the sentences are included in the conversational situations and at which the interaction between speakers and the discourse are formed.

10.4.1 Observed problems

Krivine et al. (2006), Kokkinakis & Toporowska Gronostaj (2006) and the OU NLG group have observed that pronouns are more frequent in non-expert documents compared to expert documents. Indeed, the use of pronouns allows the creator of the documnet to make the content more personal, while the scientific and expert literature remains "abstract".

As we noticed in section 10.3, the use of pronouns is correlated with the frequent use of verbs and infrequent use of nouns. Moreover, several personal pronouns (1^{st} singular and plural, and 2^{nd} singular and plural) seem to be one of the most efficient criteria for the automatic discrimination between expert and patient documents (Krivine et al. 2006). Notice that this observation will not be suitable for the discrimination, as expert literature, of discharge letters written by general practitioners to specialists and vice versa. Actually, these letters show high use of personal pronouns.

Notice additionally that Japanese patient-oriented documents show the frequent use of "forms of address" (Tomimitsu 2005; Krivine et al. 2006).

10.4.2 Solutions

For the generation of patient-friendly documents at the personalisation level, the following solutions can be adapted:

(t) Sentences should be generated so as to contain personal pronouns and corresponding syntactic structures. In this way, patients should feel more directly addressed by the documents, and may be more involved in the process of communication with experts. If so, they can feel more involved in the caregiving process as well.

(t) The sentences can contain structures which would reflect the spoken language and address patients in a more informal or "polite" way:

well, *let's say*, *actually*, ...

10.5 Document layout and presentation

"Document layout and presentation" refer to the graphical and logical organisation of documents. In the case of html documents, layout can be organised through the usage of the html tags and css files.

10.5.1 Observed problems

Krivine et al. (2006) observed that patient documents show a more complex and sophisticated presentation of information. This observation relies on several html tags: image capturing <img>, table capturing <table>, lists <ul> and enumerations <ol>, hypertext tags <a>, and tags for putting strings into bold <b> and italic <i> characters. Although these tags are not always used in the expected way, they are much more frequent in patient-oriented documents, as it has been observed in Russian and French corpora.

10.5.2 Solutions

For the generation of patient-friendly documents at the level of document layout and presentation, the following solutions can be adapted:

(t) Generation of text with at least minimal structure and itemisation which would help the reading of the document;

(t) Generation of dialogues, thereby adding a complementary modality, which could potentially help patients to understand the logical sequence of events and relationships between the medical concepts involved;

(e) Use of graphical material (i.e., images and pictures) in order to illustrate the concepts involved (medical devices, body structure and position of injuries, diseases, procedures, etc), which adds an additional modality. But notice that, according to Hameen-Anttila et al. (2004), pictograms did not help children understand patient information. Possibly the results would have been different if the pictures had been better and they had been used in a better context, i.e. in real information leaflets;

(e) Use of video material in order to better describe and illustrate medical procedures;

(e) Propose supplementary (hyper)links to more related information.

10.6 Summary

At the different levels of document content and structure, observed through the contrastive analysis of expert and patient documents, we propose ways for the adaptation of expert content for patients. Often several ways are possible and not mutually exclusive. We distinguish especially the translational and educational purposes of patient-oriented documents which both allow adaptation of the content of documents for patients.

It seems that the educational purpose is more suitable that the translation as it would not affect the content but propose additional information necessary for the understanding of expert statements.

It is remarkable, however, how conflicting the evidence is in the area of adapting medical documents to patient (Åhlfeldt et al. 2006: sec. 7.2.6). Thus, several medical informatics studies (Skinner, Strecher & Hospers 1994; Strecher et al. 1994; Campbell et al. 1994; Osman et al. 1994) found positive effects of tailoring documents to patients' needs. On the other hand, studies involving the use of NLG systems (Jones et al. 1994; Lennox et al. 2001; Reiter, Robertson & Osman 2003; Jones et al. 2006) failed to demonstrate significant effects of tailoring. Note that there were a number of differences between the studies (outcomes, medical areas, information supports involved). Other studies should be conducted. Additionaly, it seems that in the cancer area, which is characterized by more complex treatments and surgery and where there may not be such direct ways in which education can help patients, it is difficult to demonstrate the benefits of tailored information (Jones et al. 1994; Jones et al. 2006).

11 "Proof-of-concept" NLG demonstrator

This section describes requirements for a proof-of-concept demonstration system for generating patient-friendly summaries and scripted dialogues from simulated breast cancer patient chronicles. These will be read out by autonomous agent characters and rendered as short movie clips to be watched on a computer. This demonstrator uses clinical material prepared during the ongoing UK CLEF project (Hallett & Scott 2005).

We do not describe "requirements" in the normal computer science sense but rather research requirements for evaluation of the demonstrator; i.e., how will we measure whether the system achieves its research objectives. During the short time span for development (i.e., 10 months), the NLG group from the Open University, UK, has started to address some of the relevant research issues. The recommendations proposed in the section 10 are addressed where possible.

11.1 Context of Use

We believe that watching short movie clips describing medical case histories very similar to their own will involve patients in a vicarious learning experience. That is, an experience where they will benefit and learn from watching autonomous agent characters discuss a case history. If a patient were to watch one of these just before her next consultation with a doctor, it could potentially help her in a number of ways:

- by reminding her of her own case history;
- by giving her practical examples of the meaning and usage of medical terms relating to her case;
- and (in the case of the scripted dialogues) by demonstrating how to ask practical medical questions relating to her case.

11.2 Input

The OU NLG group already uses a simulator developed for the ongoing project CLEF (Hallett & Scott 2005) that simulates "chronicles" of treatment for breast cancer patients in a relational database and generates summaries for doctors. The WP27 NLG demonstrator uses the same chronicle simulator database as input from which it generates summaries for patients.

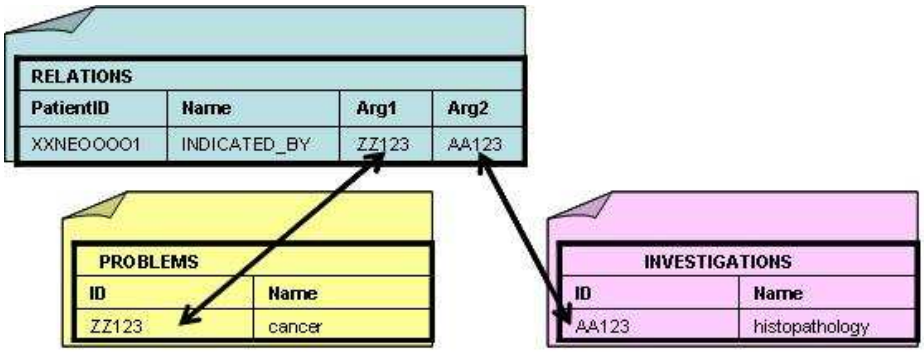


Figure 5: An example of chronicle database tables and semantic links

The chronicle database has a number of tables describing: medical interventions, investigations, problems, etc. These are indexed by simulated patient IDs. It also contains a relations table that semantically relates entities in the other tables to one another, e.g. an investigation entry in the *investigations* table can be related to a problem entry in the *problems* table

by an entry of the *relations* table that links them in an *INDICATED_BY* relation (see figure 5). This means that an investigation took place and found the patient was suffering from a problem such as cancer (provided that additional fields in the investigation and problem records show that the investigation was completed and the problem exists). Our generator uses the links in the *relations* table to search the other tables and choose interesting content for the patient summaries.

11.3 Output mode 1: Monologue Summaries

Patient summaries are generated and read out by an autonomous "newsreader" agent (see figure 6). Commercial text-to-speech software is employed to make the agent 'speak' the generated text. The Loquendo text-to-speech system, used in our demonstrator, <http://:>www.loquendo.com is available in many languages, e.g. British English, German, French and Swedish.

The summaries can be saved as movie files that can be played on a computer and viewed on the screen. The summaries will describe episodes in a simulated patient's chronicle. For instance, the simulated patient might have had some tests done and results of the tests may have indicated that the patient had cancer. The patient may then have had some surgery to remove the cancer and further treatment, such as chemotherapy to kill remaining cancer cells. The summary would describe these events and explain the medical terms involved.

Figure 6: Autonomous "newsreader" agent for reading patient summaries

11.4 Output mode 2: Scripted Dialogues

Monologue summaries are transformed by rules in the NLG system into scripted dialogues (i.e., similar to scripts for theatrical plays). These are acted out by autonomous agent characters (see figure 7) and saved, played and viewed as before.

Figure 7: Autonomous agents “acting” a scripted dialogue

11.5 NLG Technology

For the proof-of-concept demonstrator, the architecture is relatively simple. Content selection consists of SQL queries based on our investigations of the semantic relations in the chronicle database. A document structuring algorithm constructs semantic structures, each of which is represented by a “medical episode” template with fixed semantic relations between “medical entities” objects which are built from the results of the SQL queries and fill slots in the episode template. Medical terms are looked up in a table of term definitions which are built into “gloss” objects which also fill slots in the episode structures where they are linked by "explanation" relations. Rules map monologue semantic structures to dialogue structures (essentially these assign existing parts of the semantic structure to dialogue turns and add some additional “question” turns). Realisation is achieved through simple string processing with reference to a simple discourse history which maintains a list of entities mentioned so far in the document. The output is formatted as web pages with embedded Active X commands to control the autonomous agent characters. Movies are produced from these using screen-capture software.

The system outputs consist of short sentences and short words. Where medical terms are used, explanations are included. Term definitions are taken from the Cancer Research UK patient information website and modified slightly to agree with tense and to use simpler language, where appropriate. This and the two alternative presentation modes should fit the requirements we specified in section 10.

12 Objectives for the "proof-of-concept" NLG demonstrator

Watching the monologue and dialogue patient chronicle descriptions could help patients in a number of ways. Our evaluation will compare the effectiveness of the two NLG outputs on each of the following objectives:

- communicate descriptions of case histories to patients;
- explain medical terms to patients;
- encourage patients to ask questions during consultations;
- generate outputs that patients enjoy and find helpful;
- and generate outputs that patients find easy to use.

Our hope is that this combination of objectives will have the benefit of increasing accessibility, especially for patients with little medical knowledge, or poor literacy and numeracy. It will also save doctors' time by providing an additional backup source of explanation for patients.

The first two objectives concern effective communication; i.e. the generated outputs should describe and convey the information in the simulated patient chronicles and the relevant medical terms in a manner that patients can understand. Obvious issues here are how to express medical information in an understandable way, how much medical detail to include and to what scientific depth medical concepts should be described. On one hand, we require that the system should not patronise adults by describing concepts in a childish way, nor should it repeat information that they already know. On the other hand, it can be reassuring for patients to have their medical knowledge confirmed and repeating information by summarising it is promoted as good practice in doctor-patient communication (Silverman, Kurtz & Draper 1998). In an evaluation, we could determine whether patients and doctors like or dislike the generated descriptions, and why, by asking them. We could also evaluate whether the descriptions are understandable by asking comprehension questions.

In their guide for communicating with patients, Silverman, Kurtz & Draper (1998) advocate that before explaining medical information to a patient, a doctor should assess a patient's prior knowledge and the extent of her/his wish for information. In an ideal world, our system would elicit information about how much medical knowledge patients possess, and want to acquire, before generating explanations. However, obtaining this information is problematic. Obviously we cannot expect patients

to fill in lengthy questionnaires about their medical knowledge and how much they want to know each time we generate descriptions. Furthermore, asking patients whether they understand terms would give inaccurate results because people overestimate their own medical knowledge; Chapman et al. (2003) asked 150 members of the public questions about cancer terms and found that only 52% understood that the phrase "the tumour is progressing" was bad news, nevertheless the participants were fairly confident that they understood it. Our own corpus comparison of doctor-authored vs. patient-authored documents found that some terms are commonly used by both doctors and patients, some terms are commonly used by doctors but not patients, and some are commonly used by patients but not doctors. We should make use of these statistics, as well as data from studies such as Chapman et al. 2003 in an algorithm which the NLG system would use to make a decision about which terms require explanations to be generated, and which do not.

Our second and third objectives have an educational aspect, i.e. to teach patients new medical terms, to equip them with language relevant for their consultations and (in the case of the dialogue output) to show them examples of how to ask questions during consultations. We hope that the latter will involve patients in a vicarious learning experience and encourage them to ask more questions, as demonstrated by Craig et al. (2000). In our evaluation, we could question patients to determine whether they would ask more questions in a consultation before and after they view the generated material.

Our final two objectives concern usability. During evaluation, we can also determine whether patients find the system usable via a questionnaire.

13 Conclusion

On the basis of the previous literature survey and corpus study, we have formulated a set of recommendations for adapting expert clinical documents for patients. The proposed recommendations can fit to translational or educational purpose. These are not mutually exclusive and can be complementary, but in order to not alter the original expert medical information the educational purpose should be preferred. The proposed recommendations concern several levels of the content and structure of documents: Morphology, lexicon and terminology, syntax, personalisation and document layout and presentation. At each level, specific solutions and resources are necessary.

We described as well a prototype Natural Language Generation demonstrator through which the recommendations can be implemented,

and gave main the principles of the evaluation of this demonstrator. The demonstrator will first be applied in the area of breast cancer for the generation of documents in order to help patient to understand their clinical case and to improve the communication between patients and physicians.

Part IV
Perspectives

Several perspectives emerge from the work presented in this deliverable.

If, in the future, a more detailed corpus analysis is conducted then other criteria may emerge from this study, enabling us to make further contributions to the task of adapting expert documents for patients.

Detection and creation of specific resources (vocabularies, terminologies, database of definitions, etc.) is necessary for the generation of patient-friendly documents. Creation of such resources is a specific and time-consuming task. The importance of this task should not be underestimated.

The effective implementation of several or all of the recommendations within the NLG demonstrator is another perspective. It will allow us to evaluate the efficency of the presented criteria.

The NLG demonstrator under development currently generates patient documents in English. As an interesting future development, we intend to to adapt this demonstrator to other languages, such as French, Swedish or German. This research will involve, at least: the detection and selection of the resources needed for such a demonstrator; the translation and adaptation of content within the languages in this group; the selection of and translation of the medical terms involved; and the ability to recover definitions for terms.

References

Åhlfeldt, H., L. Borin, P. Daumke, N. Grabar, C. Hallett, D. Hardcastle, D. Kokkinakis, C. Mancini, K. Markó, M. Merkel, C. Pietsch, R. Power, D. Scott, M.T. Gronostaj, S. Williams & A. Willis. 2006. "Literature review on patient-friendly documentation systems." Technical Report 2006/04, Department of Computing, Faculty of Mathematics and Computing, The Open University, Milton Keynes, UK.

Baayen, R. Harald. 2001. *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publishers.

Beers, Mark H., ed. 2006. *The Merck Manual – Second Home Edition, Online Version*.

Beers, M.H. & R. Berkow, eds. 2006. *The Merck Manual of Diagnosis and Therapy, Online Version*.

Biber, Douglas. 1995. *Dimensions of Register Variation. A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Björnsson, Carl-Hugo. 1968. *Läsbarhet*. Stockholm: Liber.

Bodenreider, Olivier & Serguei V. Pakhomov. 2003. "Exploring adjectival modification in biomedical discourse across two genres." *Proc. of the NLP in Biomedicine ACL-03 Workshop*. ACL, 105–112.

Bouhaddou, O. & H. Warner. 1995. "An interactive patient information and education system (Medical HouseCall) based on a physician expert system (Iliad)." *Medinfo*. 1181–1185.

Brenna, P.F. & A.R. Aronson. 2003. "Towards linking patient and clinical information: detecting UMLS concepts in e-mail." *J Biomed Inform* 36 (4-5): 334–341.

Briscoe, Ted, John Carroll & Rebecca Watson. 2006. "The second release of the RASP system." *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. Sydney: ACL, 77–80.

Brown, P., C. Price & Y.M. Cox. 1997. "Patient language – evaluating its relationship to a clinical thesaurus." *Proc. of the AMIA Annual Fall Symposium*. Nashville, USA: AMIA.

Bui, A.A., R.K. Taira, S. El-Saden, A. Dordoni & D.R. Aberle. 2004. "Automated medical problem list generation: towards a patient timeline." *Medinfo*. 587–591.

Campbell, J.R. & T.H. Payne. 1994. "A comparison of four schemes for codification of problem lists." *Proc Annu Symp Comput Appl Med Care*. 201–205.

Campbell, MK, BM DeVellis, VJ Strecher, AS Ammerman, RF DeVellis & RS Sandler. 1994. "Improving dietary behavior. the effectiveness of tailored messages in primary care settings." *American Journal of Public Health* 84 (5): 783–787.

Cantalejo, I.M. Barrio & Simón P. Lorda. 2003. "Can patients read what we want them to read? analysis of the readability of printed materials for health education." *Atención Primaria* 31 (7): 409–414.

Chapman, Kristina, Charles Abraham, Valerie Jenkins & Lesley Fallowfield. 2003. "Lay understanding of terms used in cancer consultations." *Psycho-Oncology* 12: 557–566.

Craig, S. D.., B. Gholson, M. Ventura & A. C. Graesser. 2000. "Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning." *International Journal of Artificial*

Intelligence in Education 11: 242–25.

Dunning, Ted. 1993. "Accurate methods for the statistics of surprise and coincidence." *Computational Linguistics* 19 (1): 61–74.

Elhadad, N. & K.R. McKeown. 2001. "Towards generating patient specific summaries of medical articles." *roc of NAACL WS on automatic summarization*. Pittsburg, 31–39.

Fabry, P., R. Baud, P. Ruch, C. Despont-Gros & C. Lovis. 2005. "Methodology to ease the construction of a terminology of problems." *Int J Med Inform*.

Fellbaum, Christiane. 1998. "A semantic network of English: the mother of all WordNets." *Computers and Humanities. EuroWordNet: A multilingual database with lexical semantic network* 32 (2–3): 209–220.

Flesch, R. 1948. "A new readability yardstick." *Journal of Applied Psychology* 23: 221–233.

Grabar, Natalia & Pierre Zweigenbaum. 2003. "Productivité à travers domaines et genres : dérivés adjectivaux et langue médicale." *Langue française* 140: 102–125.

Grabarczyk, Zenon. 1987. "Scientific discource against the background of standard language." In *Special Language. From Human Thinking to Thinking Machines*, edited by Christer Laurén & Marianne Nordman. Clevedon: Multilingual Matters.

Hahn, Udo & Joachim Wermter. 2004. "Pumping documents through a domain and genre classification pipeline." *Proceedings of 4th International Conference on Language Resources and Evaluation*. Lisbon: ELRA.

Hales, J.W., K.M. Schoeffler & D.P. Kessler. 1998. "Extracting medical knowledge for a coded problem list vocabulary from the UMLS knowledge sources." *Annual Symposium of the American Medical Informatics Association (AMIA)*. 275–279.

Hallett, Catalina, David Hardcastle & Alistair Willis. 2006. "Corpus analysis progress report." Technical Report, EC Semantic Mining NoE WP27. WP27 Internal deliverable, September 2006.

Hallett, Catalina & Donia Scott. 2005. "Structural variation in generated health reports." *Proceedings of the 3rd International Workshop on Paraphrasing*. Jeju Island, Korea.

Hameen-Anttila, K, K Kemppainen, H Enlund, PJ Bush & A Marja. 2004. "Do pictograms improve children's understanding of medicine leaflet information?" *Patient Education and Counseling* 55 (3): 371–378.

Hsieh, Y., G. A. Hardardottir & P. F. Brennan. 2004. "Linguistic analysis: Terms and phrases used by patients in e-mail messages to nurses." *MEDINFO*. IOS Press.

Jacquemin, Christian. 1999. "Syntagmatic and paradigmatic representations of term variation." *37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*. University of Maryland, 341–348.

Jones, R.D., J. Pearson, S. McGregor, A.J. Cawsey, A. Barrett & N. Craig. 1994. "Randomised trial of personalised computer based information for cancer patients." *British Medical Journal* 319 (7219): 1241–1247.

Jones, R.D., .J Pearson, A.J. Cawsey, D. Bental, A. Barrett & J. White. 2006. "Effect of different forms of information produced for cancer patients on their use of the information, social support, and anxiety: Randomised trial." *British Medical Journal* 332: 942–948.

Karlgren, Jussi & Douglass Cutting. 1994. "Recognizing text genres with simple metrics using discriminant analysis." *Proceedings of the 15th. International Conference on Computational Linguistics (COLING)*. Kyoto: ACL, 1071–1075.

Kittredge, Richard I. 2003. "Sublanguages and controlled languages." In *The Oxford Handbook of Computational Linguistics*, edited by Ruslan Mitkov, 430–447. Oxford: Oxford University Press.

Kokkinakis, Dimitrios. 2006. "Collection, encoding and linguistic processing of a Swedish medical corpus." *Proc. of the 5th LREC*. Genoa: ELRA.

Kokkinakis, Dimitrios & Sofie Johansson Kokkinakis. 1999. "A cascaded finite-state parser for syntactic analysis of Swedish." *Proc. of the 9th European Chapter of the Association of Computational Linguistics (EACL)*. Bergen: ACL.

Kokkinakis, Dimitrios & Maria Toporowska Gronostaj. 2006. "Lay language versus professional language within the cardiovascular subdomain – a contrastive study." *Proc. of BIO'06*. Athens.

Kokkinakis, Dimitrios, Maria Toporowska Gronostaj & Karin Warmenius. 2000. "Annotating, disambiguating & automatically extending the coverage of the Swedish SIMPLE lexicon." *Proc. of the 2nd LREC*. Athens: ELRA.

Krivine, Sonia. 2005. "Critères pour la catégorisation automatique des documents numériques. application au discours scientifique russe du web dans le cadre du projet deco." Technical Report, INaLCO. MSc report.

Krivine, Sonia, Masaru Tomimitsu, Natalia Grabar & Monique Slodzian. 2006. "Relever des critères pour la distinction automatique entre les douments médicaux scientifiques et vulgarisés en russe et en japonais." In *Verbum ex machina (TALN vol. 1). Actes de la 13e conférence sur le traitement automatique des langues naturelles*, edited by Piet Mertens, Cédrick Fairon, Anne Dister & Patrick Watrin, 522–531. Louvain-la-Neuve: Presses universitaires de Louvain.

Laufer, Batia & Paul Nation. 1995. "Vocabulary size and use: Lexical richness in L2 written production." *Applied Linguistics* 16 (3): 307–329.

Lauteslager, M., H.J. Brouwer, J. Mohrs, P.J. Bindels & H.G. Groundmeijer. 2002. "The patient as a source to improve the medical record."

Fam Med 19 (2): 167–171.

Lebart, Ludovic, André Salem & Lisette Berry. 1998. *Exploring Textual Data*. Dordrecht: Kluwer Academic Publishers.

Lennox, A.S., L.M. Osman, E. Reiter, R. Robertson, J.A. Friend & I. McCann. 2001. "The cost-effectiveness of computer-tailored and non-tailored smoking cessation letters in general practice. a randomised controlled trial." *British Medical Journal* 322: 1396–1400.

Markó, Kornél, Stefan Schulz & Udo Hahn. 2005. "MorphoSaurus – design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain." *Methods of Information in Medicine* 44 (4): 537–545.

McKeown, K.R., S. Pan, J. Shaw, D. Jordan & B. Allen. 1997. "Language generation for multimedia healthcare briefings." *Proc of Fifth Conference on Applied Natural Language Processing*. ACL, 277–282.

Miller, J., C. Driscoll, S. Kilpatrick & E. Quillen Jr. 2003. "Management of prenatal care information: integration of the problem list and clinical comments." *Top Health Inf Manage* 24 (1): 42–49.

Miyawaki, S., K. Takada, M. Furukawa & S. Adachi. 1995. "An interactive consultation multimedia software for orthodontic patients." *Medinfo*. 1308.

NLM. 2003. *UMLS Knowledge Sources Manual*. Bethesda, Maryland: National Library of Medicine. `www.nlm.nih.gov/research/umls/`.

Nordman, Marianne. 1992. *Svenskt fackspråk*. Lund: Studentlitteratur.

Ogden, J., R. Branson, A. Bryett, A. Campbell, A. Febles, I. Ferguson, H. Lavender, J. Mizan, R. Simpson & M. Tayler. 2003. "What's in a name? an experimental study of patients' views of the impact and function of a diagnosis." *Fam Pract* 20 (3): 248–253.

Osman, L.M., M.I. Abdalla, J.A.G. Beattie, S.J. Ross, I.T. Russell, J.A. Friend, J.S. Legge & J.G. Douglas. 1994. "Reducing hospital admission through computer supported education for asthma patients." *British Medical Journal* 308: 568–571.

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

Plovnick, R.M. & Q.T. Zeng. 2004. "Reformulation of consumer health queries with professional terminology: as pilot study." *J Med Internet Res* 6 (3): e27.

Reiter, E, R Robertson & L Osman. 2003. "Lessons from a failure: Generating tailored smoking cessation letters." *Artificial Intelligence* 144: 41–58.

Richardson, T. 1996. "The terminology of patient-focused care nouns as verbs, adjectives as nouns." *Revolution* 6 (1): 31–34.

Sager, J.C., D. Dungworth & P.F McDonald. 1980. *English Special Languages*. Wiesbaden.

Schulz, Stefan. 2007. "Evaluation of the multilingual medical dictionary." Wp20 deliverable, EC NoE 507505 Semantic Interoperability and Data Mining in Biomedicine.

Silverman, J., S. Kurtz & J. Draper. 1998. "Skills for communicating with patients." ISBN 1857751892. Radcliffe Medical Press.

Skinner, C.S., V.J. Strecher & H. Hospers. 1994. "Physicians' recommendations for mammography: Do tailored messages make a difference?" *American Journal of Public Health* 84 (1): 43–49.

Smith, Barry & Christiane Fellbaum. 2004. "Medical Wordnet: A new methodology for the construction and validation of information." *Proc. of the 20th Conf. on Comp. Ling.* Geneva: ACL, 371—382.

Smith, C.A., P.Z. Stavri & W.W. Chapman. 2002. "In their own words? a terminological analysis of e-mails to a cancer information service." *Annual Symposium of the American Medical Informatics Association (AMIA)*. 697–701.

Soergel, D., T. Tse & L Slaughter. 2004. "Helping healthcare consumers understand: An "interpretive layer" for finding and making sense of medical information." *MEDINFO*. IOS Press.

Stamatatos, Efstathios, George Kokkinakis & Nikos Fakotakis. 2000. "Automatic text categorization in terms of genre and author." *Computational Linguistics* 26 (4): 471–495.

Strecher, V.J., M. Kreuter, D.J.D. Boer, S. Kobrin, H. Hospers & C.S. Skinner. 1994. "The effects of computer-tailored smoking cessation messages in family practice settings." *J Fam Pract.* 39 (3): 262–270.

Surján, G. & G. Héja. 2003. "About the language of Hungarian discharge reports." *Stud Health Technol Inform* 95: 869–873.

Thompson, H.J. 2005. "Fever: A concept analysis." *J Adv Nurs* 51 (5): 484–492.

Tomimitsu, Masaru. 2005. "Exploitation de critères de distinction automatique des textes scientifiques et vulgarisés autour des notions "diabète / régime alimentaire"." Technical Report, INaLCO. MSc report.

Torgersson, O. & G. Falkman. 2002. "Using text generation to access clinical data in a variety of contexts." Edited by G. Surján, R. Engelbrecht & P. McNair, *MIE*. 460–465.

Tse, T. & D. Soergel. 2003. "Exploring medical expressions used by consumers and the media: An emerging view of consumer health vocabularies." *Proc. AMIA Symp.* AMIA, 674–678.

Waisman, Y., N. Siegal, M. Chemo, G. Siegal, L. Amir, Y. Blachar & M. Mimouni. 2003. "Do parents understand emergency department discharge instructions? a survey analysis." *Isr Med Assoc J* 5 (8): 567–570.

White, P., A. Singleton & R. Jones. 2004. "Copying referral letters to

patients: The views of patients, patient representatives and doctors." *Patient Educ Couns* 55 (1): 94–98.

Williams, N. & J. Ogden. 2004. "The impact of matching the patient's vocabulary: A randomized control trial." *Fam Pract* 21 (6): 630–635.

Williams, Sandra. 2003. "Language choice models for microplanning and readability." *Proceedings of the Student Workshop of the Human Language Technology and North American Chapter of the Association for Computational Linguistics Conference (HLT-NAACL03 Student Workshop)*. Edmonton: ACL, 13–18.

Zeng, Qing, Eunjung Kim, Jon Crowell & Tony Tse. 2005. "A text corpora-based estimation of the familiarity of health terminology." *ISBMDA*.

Zeng, Qing T. & Tony Tse. 2006. "Exploring and developing consumer health vocabularies." *Am Med Inform Assoc* 13: 24–29.

Appendix 1: Top 100 most frequent words

	Case studies			Merck medics			Merck patients			Stories		
	Word	Freq	Rel freq	Word	Freq	Rel freq	Word	Freq	Rel freq	Word	Freq	Rel freq
1	mass	569	0.011	patients	498	0.014	cancer	1204	0.043	cancer	337	0.012
2	tumor	513	0.010	tumor	401	0.011	treatment	323	0.012	treatment	260	0.009
3	diagnosis	429	0.008	cancer	345	0.010	cells	313	0.011	time	244	0.009
4	md	347	0.007	cell	337	0.010	people	290	0.010	back	189	0.007
5	cells	332	0.007	disease	293	0.008	spread	287	0.010	hospital	154	0.005
6	discussion	322	0.006	tumors	278	0.008	breast	256	0.009	told	152	0.005
7	carcinoma	319	0.006	cells	236	0.007	symptoms	238	0.009	breast	138	0.005
8	cell	319	0.006	treatment	218	0.006	cancers	223	0.008	day	136	0.005
9	tumors	288	0.006	therapy	201	0.006	radiation	220	0.008	doctor	113	0.004
10	patient	241	0.005	diagnosis	193	0.006	blood	211	0.008	good	112	0.004
11	common	230	0.005	carcinoma	179	0.005	tumor	207	0.007	life	110	0.004
12	ct	219	0.004	chemotherapy	168	0.005	chemotherapy	206	0.007	chemotherapy	109	0.004
13	patients	209	0.004	bone	165	0.005	therapy	199	0.007	thought	107	0.004
14	left	205	0.004	common	156	0.004	lymph	177	0.006	weeks	97	0.003
15	imaging	200	0.004	malignant	146	0.004	surgery	175	0.006	months	96	0.003
16	year	193	0.004	symptoms	137	0.004	women	162	0.006	years	91	0.003
17	cm	189	0.004	primary	136	0.004	cell	161	0.006	people	90	0.003
18	lesion	187	0.004	occur	131	0.004	diagnosis	152	0.005	feel	90	0.003
19	images	187	0.004	yr	129	0.004	risk	143	0.005	found	90	0.003
20	differential	184	0.004	lesions	113	0.003	disease	139	0.005	days	89	0.003
21	disease	176	0.003	radiation	109	0.003	skin	137	0.005	dr	82	0.003
22	pathology	171	0.003	lymph	103	0.003	nodes	137	0.005	started	82	0.003
23	cases	170	0.003	small	102	0.003	tumors	132	0.005	home	81	0.003
24	tissue	166	0.003	survival	102	0.003	bone	131	0.005	ann	80	0.003
25	high	163	0.003	lymphoma	101	0.003	common	127	0.005	asked	78	0.003
26	findings	163	0.003	normal	101	0.003	develop	125	0.004	make	78	0.003
27	case	162	0.003	metastases	98	0.003	removed	122	0.004	made	78	0.003
28	malignant	162	0.003	stage	97	0.003	tissue	121	0.004	week	75	0.003
29	cystic	158	0.003	marrow	97	0.003	lung	120	0.004	knew	75	0.003
30	include	157	0.003	surgery	96	0.003	names	119	0.004	felt	74	0.003
31	small	151	0.003	include	94	0.003	person	118	0.004	family	71	0.002
32	history	147	0.003	radiotherapy	93	0.003	trade	118	0.004	surgery	70	0.002
33	bone	145	0.003	patient	90	0.003	liver	114	0.004	doctors	69	0.002
34	lesions	142	0.003	syndrome	90	0.003	drugs	112	0.004	hair	68	0.002
35	breast	141	0.003	leukemia	89	0.003	years	112	0.004	work	67	0.002
36	presentation	135	0.003	lung	88	0.003	cancerous	107	0.004	called	64	0.002
37	appearance	133	0.003	risk	85	0.002	stage	103	0.004	scan	63	0.002
38	low	131	0.003	biopsy	84	0.002	prostate	100	0.004	things	63	0.002
39	years	130	0.003	prognosis	83	0.002	biopsy	100	0.004	radiotherapy	63	0.002
40	large	128	0.003	cases	83	0.002	body	99	0.004	year	63	0.002
41	normal	127	0.003	occurs	79	0.002	small	99	0.004	support	62	0.002
42	primary	127	0.003	high	79	0.002	doctor	98	0.004	lymphoma	61	0.002
43	areas	123	0.002	pain	79	0.002	leukemia	96	0.003	side	61	0.002
44	typically	122	0.002	present	79	0.002	type	96	0.003	chemo	60	0.002
45	show	122	0.002	clinical	78	0.002	early	89	0.003	story	59	0.002
46	thyroid	122	0.002	signs	77	0.002	doctors	87	0.003	long	59	0.002
47	shows	120	0.002	ch	77	0.002	large	86	0.003	operation	59	0.002
48	contrast	119	0.002	increased	74	0.002	pain	85	0.003	lump	58	0.002
49	lymph	119	0.002	tissue	72	0.002	heart	84	0.003	bit	57	0.002
50	weighted	119	0.002	cancers	71	0.002	normal	82	0.003	thing	57	0.002
51	present	118	0.002	liver	69	0.002	carcinoma	78	0.003	pain	57	0.002
52	cancer	115	0.002	metastatic	68	0.002	men	78	0.003	blood	56	0.002

Appendix 1: Top 100 most frequent words

53	radiology	114	0.002	drugs	68	0.002	tests	78	0.003	normal	56	0.002		
54	positive	112	0.002	early	66	0.002	performed	78	0.003	left	56	0.002		
55	mr	112	0.002	serum	65	0.002	called	78	0.003	night	55	0.002		
56	benign	111	0.002	nodes	65	0.002	examination	77	0.003	lymph	55	0.002		
57	presented	109	0.002	mg	62	0.002	intestine	76	0.003	results	53	0.002		
58	solid	108	0.002	mass	62	0.002	children	73	0.003	prostate	53	0.002		
59	ultrasound	108	0.002	chronic	61	0.002	lymphoma	71	0.003	appointment	53	0.002		
60	lung	105	0.002	large	60	0.002	types	66	0.002	disease	53	0.002		
61	metastases	104	0.002	age	60	0.002	high	66	0.002	feeling	52	0.002		
62	gross	102	0.002	low	60	0.002	occur	64	0.002	check	51	0.002		
63	metastatic	101	0.002	incidence	59	0.002	prognosis	64	0.002	give	51	0.002		
64	age	100	0.002	occasionally	59	0.002	time	63	0.002	hope	51	0.002		
65	image	100	0.002	levels	58	0.002	cure	63	0.002	diagnosed	50	0.002		
66	visible	99	0.002	surgical	58	0.002	hodgkin	62	0.002	radiation	49	0.002		
67	lymphoma	98	0.002	staging	58	0.002	marrow	61	0.002	small	49	0.002		
68	pain	98	0.002	ct	58	0.002	treated	58	0.002	medical	49	0.002		
69	masses	96	0.002	lesion	57	0.002	system	57	0.002	bone	49	0.002		
70	chest	96	0.002	results	56	0.002	area	56	0.002	test	49	0.002		
71	rare	95	0.002	blood	56	0.002	age	55	0.002	treatments	49	0.002		
72	soft	93	0.002	anemia	56	0.002	effects	55	0.002	decided	49	0.002		
73	wall	93	0.002	lymphomas	55	0.002	removal	55	0.002	find	48	0.002		
74	necrosis	92	0.002	loss	55	0.002	ct	54	0.002	gave	48	0.002		
75	radiologist	91	0.002	cure	55	0.002	procedure	54	0.002	end	48	0.002		
76	attending	91	0.002	studies	55	0.002	abnormal	54	0.002	therapy	47	0.002		
77	renal	89	0.002	rate	53	0.002	survival	53	0.002	patients	47	0.002		
78	showed	88	0.002	rare	53	0.002	bleeding	52	0.002	point	46	0.002		
79	pulmonary	88	0.002	skin	52	0.001	parts	52	0.002	start	46	0.002		
80	nodes	88	0.002	women	52	0.001	bladder	51	0.002	high	46	0.002		
81	signal	86	0.002	response	52	0.001	needed	51	0.002	hours	45	0.002		
82	bowel	86	0.002	bleeding	51	0.001	part	50	0.002	due	45	0.002		
83	arrow	86	0.002	renal	51	0.001	tissues	50	0.002	put	45	0.002		
84	adenocarcinoı	86	0.002	acute	51	0.001	screening	50	0.002	news	44	0.002		
85	specimen	85	0.002	multiple	51	0.001	produce	49	0.002	result	44	0.002		
86	occur	85	0.002	involvement	50	0.001	chest	49	0.002	showed	44	0.002		
87	multiple	83	0.002	including	49	0.001	include	48	0.002	effects	44	0.002		
88	mri	82	0.002	node	49	0.001	number	48	0.002	surgeon	42	0.001		
89	melanoma	82	0.002	malignancy	49	0.001	stomach	47	0.002	call	42	0.001		
90	enhancement	81	0.002	rarely	49	0.001	surrounding	47	0.002	oncologist	42	0.001		
91	clear	81	0.002	examination	49	0.001	kidney	47	0.002	diagnosis	42	0.001		
92	bladder	80	0.002	abdominal	49	0.001	red	46	0.002	information	42	0.001		
93	pancreas	78	0.002	metastasis	49	0.001	year	46	0.002	biopsy	41	0.001		
94	abdominal	78	0.002	commonly	48	0.001	developing	45	0.002	dose	41	0.001		
95	survival	77	0.002	elevated	48	0.001	abdomen	45	0.002	problem	41	0.001		
96	biopsy	76	0.002	site	48	0.001	grow	44	0.002	wanted	41	0.001		
97	scan	76	0.002	benign	48	0.001	uterus	44	0.002	tests	41	0.001		
98	cd	76	0.002	cervical	48	0.001	enlarged	44	0.002	taking	40	0.001		
99	type	76	0.002	specific	47	0.001	loss	43	0.002	stage	40	0.001		
100	pancreatic	75	0.001	colon	47	0.001	life	43	0.002	nurse	40	0.001		

Appendix 2: Log-Likelihood comparison (top 50 words)

	Case studies-Merck medics			Case studies- Merck patients			Case studies - Stories			Merck medics - Merck patients			Merck medics - stories			Merck-patients-stories		
	word	**L-L**	**use**	**word**	**L-L**	**use**	**word**	**L-L**	**use**	**word**	**L-L**	**use**	**word**	**L-L**	**use**	**word**	**L-L**	**use**
1	md	364.2	+	cancer	1808.5	-	mass	448.7	+	cancer	714.6	-	patients	351.0	+	cancer	538.0	+
2	discussion	326.2	+	people	568.5	-	back	330.8	-	patients	585.8	+	cell	318.1	+	cells	271.8	+
3	mass	302.8	+	spread	424.8	-	md	311.1	+	people	440.2	-	tumor	314.0	+	cancers	261.3	+
4	patients	253.1	-	mass	381.8	+	told	309.5	-	spread	251.9	-	tumors	278.3	+	spread	224.9	+
5	cancer	221.4	-	treatment	366.0	-	time	308.0	-	breast	211.4	-	back	261.0	-	back	212.5	-
6	yr	205.6	-	cancers	362.1	-	tumor	292.7	+	names	192.9	-	time	249.2	-	told	207.2	-
7	images	177.7	+	md	305.2	+	carcinoma	286.0	+	person	191.3	-	hospital	245.6	-	hospital	172.5	-
8	left	166.2	+	discussion	283.2	+	cancer	276.7	-	trade	191.3	-	told	242.5	-	develop	165.9	+
9	year	164.8	+	radiation	278.4	-	discussion	260.7	+	doctor	158.9	-	carcinoma	214.3	+	symptoms	163.6	+
10	imaging	161.3	+	chemotherapy	262.9	-	treatment	255.6	-	cancerous	155.8	-	doctor	180.2	-	trade	156.2	+
11	chemotherapy	156.1	-	symptoms	245.3	-	doctor	211.7	-	yr	151.7	+	yr	154.4	+	cell	156.2	+
12	treatment	148.2	-	trade	243.9	-	cell	207.2	+	doctors	141.0	-	common	150.2	+	tumor	155.5	+
13	ch	138.0	-	names	227.3	-	tumors	206.2	+	blood	133.6	-	disease	139.0	+	names	150.1	+
14	therapy	137.9	-	cancerous	221.1	-	diagnosis	186.9	+	lesions	123.1	+	primary	138.7	+	day	149.2	-
15	case	134.0	+	person	218.4	-	day	182.0	-	cancers	120.0	-	cells	137.6	+	common	143.3	+
16	weighted	124.9	+	patient	212.0	+	imaging	179.3	+	removed	114.5	-	feel	133.7	-	tumors	140.2	+
17	cystic	120.4	+	drugs	200.2	-	feel	173.2	-	patient	105.9	+	dr	130.8	-	leukemia	135.4	+
18	radiology	119.6	+	blood	196.0	-	common	168.0	+	malignant	96.3	+	occur	127.2	+	skin	129.8	+
19	mr	117.5	+	surgery	190.3	-	good	167.9	-	intestine	96.1	-	malignant	124.7	+	risk	122.9	+
20	presentation	107.5	+	doctor	184.7	-	differential	164.9	+	body	95.4	-	asked	124.4	-	thought	122.5	-
21	differential	107.1	+	patients	183.8	+	hospital	163.9	-	ch	90.6	+	people	120.6	-	radiation	121.6	+
22	image	104.9	+	doctors	179.8	-	people	159.5	-	radiotherapy	87.6	+	knew	119.6	-	people	115.5	+
23	presented	104.8	+	therapy	179.6	-	asked	158.8	-	women	86.4	-	lesions	118.3	+	women	115.4	+
24	radiologist	95.5	+	skin	174.6	-	images	157.2	+	called	86.3	-	thought	118.2	-	dr	111.8	-
25	attending	95.5	+	women	173.7	-	knew	152.7	-	years	85.4	-	felt	118.0	-	carcinoma	110.0	+
26	drugs	95.1	-	develop	172.7	-	life	152.7	-	serum	76.5	+	good	117.7	-	time	109.8	-
27	pathology	92.8	+	removed	167.2	-	cm	151.3	+	parts	75.5	-	home	112.7	-	good	109.3	-
28	showed	92.4	+	lesion	164.5	+	lesion	149.5	+	mg	72.9	+	weeks	111.8	-	ann	109.0	-
29	signal	90.3	+	differential	151.5	+	thought	147.2	-	heart	72.3	-	ann	111.2	-	lung	108.6	+
30	arrow	90.3	+	pathology	150.4	+	findings	146.1	+	metastases	71.5	+	week	110.2	-	cancerous	107.8	+
31	disease	89.0	-	cm	148.2	+	cells	146.1	+	clinical	70.9	+	doctors	110.1	-	intestine	107.2	+
32	leukemia	86.6	-	breast	136.5	-	cystic	141.6	+	lesion	67.0	+	metastases	107.7	+	therapy	104.7	+
33	typically	86.2	+	images	133.9	+	doctors	140.5	-	radiation	66.9	-	leukemia	106.6	+	knew	102.2	-

34	cure	83.3	-	findings	133.2	+	started	138.2	-	noncancerous	63.2	-	started	103.2	-	tissue	99.7	+
35	cm	81.6	+	cystic	128.9	+	ann	134.3	-	lump	63.2	-	things	100.5	-	blood	99.5	+
36	specimen	80.1	+	intestine	128.8	-	felt	134.0	-	develop	62.1	-	make	97.2	-	asked	97.0	-
37	radiotherapy	75.2	-	leukemia	120.6	-	pathology	123.3	+	nearby	61.3	-	chemo	95.7	-	drugs	92.2	+
38	arrows	74.5	+	risk	120.5	-	story	120.2	-	skin	60.9	-	story	94.1	-	liver	88.4	+
39	visible	72.7	+	presentation	118.7	+	things	118.9	-	developing	58.8	-	hair	92.7	-	things	85.9	-
40	gross	71.4	+	case	118.7	+	cases	117.2	+	tomography	58.4	-	lump	92.5	-	started	85.0	-
41	radiation	70.5	-	lesions	115.0	+	include	117.0	+	malignancy	57.6	+	ch	92.2	+	chemo	81.8	-
42	enhancement	69.7	+	cure	114.1	-	make	116.8	-	symptoms	54.6	-	work	91.2	-	story	80.4	-
43	solid	69.5	+	weighted	104.7	+	bit	116.1	-	resection	54.1	+	bit	90.9	-	person	79.4	+
44	computed	69.3	+	called	104.3	-	thing	116.1	-	tests	52.8	-	thing	90.9	-	home	79.3	-
45	demonstrates	68.8	+	radiology	100.3	+	home	114.5	-	involvement	50.6	+	include	90.6	+	performed	79.0	+
46	power	68.2	+	left	98.5	+	chemo	112.9	-	table	50.6	+	life	90.4	-	thing	77.7	-
47	anemia	67.2	-	mr	98.5	+	dr	112.7	-	treatment	50.5	-	months	89.0	-	week	76.2	-
48	syndrome	67.1	-	parts	98.4	-	operation	110.9	-	computed	50.3	-	syndrome	86.0	+	nodes	72.7	+
49	signs	66.9	-	tests	98.0	-	lesions	110.3	+	incidence	50.2	+	appointment	84.5	-	types	72.5	+
50	symptoms	66.2	-	presented	95.9	+	appearance	109.5	+	primary	50.0	+	feeling	82.9	-	appointment	72.2	-

Appendix 3: Distribution of part-of-speech tags

POS	Case studies		Merck patients		Merck medics		Stories	
	Abs freq	Rel freq	Abs freq	Rel freq	Abs freq	Rel freq	Abs freq	Rel freq
NN1	14219	16.81	5392	9.93	9548	15.69	4562	5.74
AJ0	268	0.32	226	0.42	152	0.25	195	0.25
AT0	35	0.04	43	0.08	37	0.06	89	0.11
PRP	7186	8.49	5063	9.32	3641	5.98	6295	7.92
NN2	3118	3.69	2683	4.94	2595	4.27	3951	4.97
CJC	121	0.14	93	0.17	72	0.12	814	1.02
AV0	21	0.02	79	0.15	13	0.02	223	0.28
PRF	3200	3.78	2311	4.25	3345	5.50	2963	3.73
NP0	838	0.99	1092	2.01	783	1.29	1269	1.60
CRD	343	0.41	414	0.76	206	0.34	1103	1.39
VVN	2113	2.50	655	1.21	1171	1.92	1149	1.44
VBZ	230	0.27	105	0.19	88	0.14	1985	2.50
DT0	1332	1.57	1308	2.41	828	1.36	2063	2.59
VVZ	184	0.22	200	0.37	137	0.23	440	0.55
CJS	215	0.25	29	0.05	39	0.06	230	0.29
VVB	10	0.01	2	0.00	1	0.00	46	0.06
VVI	671	0.79	488	0.90	601	0.99	196	0.25
VVG	20012	23.65	12417	22.86	14245	23.41	12468	15.68
VBB	5416	6.40	4212	7.75	4902	8.06	3048	3.83
VM0	2680	3.17	502	0.92	1017	1.67	1631	2.05
VBD	182	0.22	104	0.19	80	0.13	423	0.53
NN0	5	0.01	13	0.02	0	0.00	274	0.34
PNP	570	0.67	406	0.75	235	0.39	6890	8.66
TO0	29	0.03	159	0.29	41	0.07	174	0.22
VBI	16	0.02	13	0.02	5	0.01	114	0.14
UNC	139	0.16	199	0.37	171	0.28	242	0.30
VVD	3062	3.62	2013	3.71	2036	3.35	1630	2.05
ZZ0	6068	7.17	4377	8.06	4998	8.22	6182	7.77
CJT	550	0.65	635	1.17	407	0.67	1854	2.33
AJC	369	0.44	20	0.04	132	0.22	46	0.06
XX0	770	0.91	612	1.13	651	1.07	381	0.48
DPS	721	0.85	20	0.04	21	0.03	1916	2.41
EX0	21	0.02	23	0.04	27	0.04	89	0.11
DTQ	481	0.57	489	0.90	670	1.10	440	0.55
ORD	120	0.14	51	0.09	123	0.20	198	0.25
VHZ	1684	1.99	1086	2.00	1326	2.18	537	0.68
POS	26	0.03	56	0.10	20	0.03	160	0.20
VHB	19	0.02	2	0.00	3	0.00	233	0.29
AVP	0	0.00	1	0.00	1	0.00	38	0.05
VBN	0	0.00	0	0.00	0	0.00	64	0.08
VHD	8	0.01	5	0.01	7	0.01	53	0.07
VHI	44	0.05	41	0.08	33	0.05	42	0.05
VDZ	129	0.15	189	0.35	186	0.31	449	0.56
AJS	109	0.13	33	0.06	10	0.02	874	1.10
PNQ	10	0.01	15	0.03	3	0.00	68	0.09
VDB	77	0.09	36	0.07	27	0.04	231	0.29
AVQ	178	0.21	268	0.49	169	0.28	142	0.18
VBG	746	0.88	1010	1.86	1177	1.93	1114	1.40
VDD	819	0.97	793	1.46	783	1.29	1020	1.28
PNX	353	0.42	51	0.09	35	0.06	2352	2.96
ITJ	801	0.95	597	1.10	548	0.90	1195	1.50
VHG	814	0.96	1355	2.49	1025	1.68	2407	3.03
VDN	2013	2.38	1248	2.30	1408	2.31	1821	2.29
PNI	862	1.02	796	1.47	723	1.19	334	0.42
VDG	250	0.30	253	0.47	195	0.32	792	1.00
VDI	350	0.41	36	0.07	142	0.23	29	0.04

Appendix 4: Most frequent MeSH terms

	CASE STUDIES		MERCK MEDICS		MERCK PATIENTS		PATIENT TESTIMONIALS	
	MESH term	Freq	MESH term	Freq	MESH term	Freq	MESH term	Freq
1	tumor	513	patients	498	cancer	1209	cancer	337
2	diagnosis	429	tumor	401	treatment	324	treatment	260
3	cells	332	cancer	345	cells	315	time	244
4	carcinoma	319	cell	337	breast	256	back	189
5	cell	315	disease	293	symptoms	239	who	168
6	tumors	288	tumors	278	cancers	224	hospital	154
7	patient	235	cells	235	radiation	220	breast	138
8	patients	209	treatment	218	blood	211	will	117
9	differential diagnosis	173	therapy	201	tumor	208	life	110
10	disease	176	diagnosis	193	chemotherapy	206	chemotherapy	109
11	pathology	171	carcinoma	179	therapy	199	family	71
12	tissue	167	chemotherapy	168	lymph	177	surgery	70
13	findings	163	bone	165	surgery	175	hair	68
14	history	147	symptoms	137	women	161	work	67
15	bone	145	radiation	109	cell	161	breast cancer	65
16	breast	138	lymph	103	who	152	came	64
17	thyroid	122	survival	102	diagnosis	152	radiotherapy	63
18	lymph	119	lymphoma	100	radiation therapy	148	lymphoma	61
19	cancer	115	metastases	98	risk	143	pain	57
20	radiology	111	marrow	97	disease	139	bit	57
21	ultrasound	108	surgery	96	skin	138	lymph	55
22	lung	105	radiotherapy	93	lymph nodes	132	blood	55
23	metastases	104	syndrome	90	tumors	133	appointment	53
24	metastatic	101	patient	90	bone	131	prostate	53
25	pain	100	lung	88	tissue	122	disease	53
26	lymphoma	97	leukemia	88	lung	120	feeling	52
27	chest	96	risk	85	names	119	treatments	49
28	necrosis	93	biopsy	84	breast cancer	116	bone	49
29	adenocarcinoma	86	prognosis	83	person	118	radiation	49
30	melanoma	82	pain	79	liver	114	therapy	47
31	bladder	80	signs	77	drugs	113	patients	47
32	lymph nodes	76	bone marrow	72	biopsy	100	news	44
33	pancreas	78	tissue	72	prostate	100	diagnosis	42
34	survival	77	cancers	71	leukemia	97	biopsy	41
35	biopsy	76	liver	69	pain	89	nurse	40
36	sarcoma	72	metastatic	68	heart	84	friends	39
37	neoplasm	72	drugs	68	men	78	symptoms	37
38	treatment	68	serum	65	carcinoma	78	mother	36
39	liver	68	radiation therapy	59	intestine	76	insurance	36
40	woman	66	lymph nodes	59	children	73	husband	36
41	power	65	incidence	59	lymphoma	71	hormone	35
42	metastasis	65	blood	56	prognosis	64	marrow	34
43	cyst	65	anemia	56	time	64	use	34
44	hemorrhage	63	lymphomas	55	lung cancer	61	bone marrow	32
45	blood	63	women	52	bone marrow	60	cells	32
46	therapy	62	skin	52	marrow	61	side effects	30
47	white	61	bleeding	51	blood cells	56	control	30
48	tomography	61	metastasis	49	procedure	54	consultant	29
49	symptoms	60	colon	47	causes	53	tumor	29
50	histology	54	breast	46	survival	52	lymph nodes	27
51	neoplasms	53	lymph node	44	bleeding	52	future	26
52	will	52	growth	43	prostate cancer	50	self	25
53	chondrosarcoma	51	brain	43	bladder	51	head	25
54	risk	49	human	40	screening	51	friend	25
55	prognosis	47	prostate	39	tissues	50	women	25
56	adenoma	47	platelet	39	chest	49	mastectomy	24
57	lymph node	43	melanoma	39	kidney	48	reading	24
58	head	44	carcinomas	39	stomach	47	letter	24
59	carcinomas	44	function	38	abdomen	45	bed	24
60	serum	43	symptoms and signs	36	uterus	44	let	24
61	microscopy	43	adenocarcinoma	37	rectum	43	hand	24
62	brain	43	who	36	life	43	wife	24

63	secondary	42		infection	36		lymphomas	42		arm	24
64	men	42		children	36		estrogen	41		research	23
65	kidney	42		chest	36		death	41		patient	23
66	infection	41		antigen	35		lymph node	39		neck	23
67	staining	40		lung cancer	34		use	39		son	23
68	prostate	40		secondary	34		brain	39		prostate cancer	22
69	muscle	40		plasma	34		woman	39		interferon	21
70	cytoplasm	40		neck	33		colon	38		health	21
71	thyroid cancer	38		antigens	33		growth	38		chest	21
72	surgery	39		stomach	32		esophagus	37		sleep	21
73	pheochromocytoma	39		pressure	32		tomography	36		person	20
74	children	39		bladder	32		adenocarcinoma	35		catheter	20
75	magnetic resonance	37		esophagus	31		lungs	34		tablets	20
76	women	38		polyps	29		white	34		specialist	20
77	time	38		men	29		urine	33		transplant	19
78	mph	38		fever	29		treatments	33		x rays	18
79	carcinoid	38		use	28		neck	33		computer	18
80	neck	37		chromosome	28		infection	32		procedure	18
81	magnetic	37		production	27		large intestine	31		rest	18
82	recurrence	36		neoplasms	27		immune system	31		love	17
83	radiation	36		infections	27		side effects	31		internet	17
84	gallbladder	36		findings	27		ovaries	31		bladder	17
85	pelvis	35		death	27		pancreas	30		thinking	17
86	origin	35		squamous cell carcinoma	26		transplantation	30		booklet	17
87	incidence	35		x rays	26		united states	28		name	17
88	hospital	35		urine	26		polyps	28		brain	16
89	cysts	34		time	26		metastatic	28		set	16
90	marrow	33		control	26		ultrasound	28		woman	16
91	chemotherapy	33		complications	26		vagina	28		appointments	16
92	perfusion	32		tissues	25		mole	28		exercises	15
93	lungs	32		stem	25		melanoma	27		risk	15
94	growth	32		light	25		lymphocytes	27		nurses	15
95	female	32		erythrocytosis	25		mastectomy	26		attitude	15
96	who	31		diarrhea	25		stem	26		mail	15
97	evaluation	31		antibodies	25		needle	26		daughter	15
98	edema	31		syndromes	24		small intestine	25		fear	15
99	adults	31		sarcoma	24		mouth	25		book	15
100	thymoma	30		causes	24		cervix	25		appetite	15
101	squamous cell carcinoma	27		will	23		development	25		liver	15
102	mesoblastic nephroma	28		thrombocytopenia	23		white blood cells	23		urine	14

Appendix 5: Log-Likelihood comparison of MeSH terms

	Case studies- Merck medics			Case studies - Merck patients			Case studies - Stories			Merck medics-Merck patients			Merck medics - Stories			Merck patients- Stories		
	Term	LL	use	Term	LL	use	Term	LL	use	Term	LL	use	Term	LL	use	Term	LL	use
1	patients	146.3	-	cancer	1196.9	-	back	403.1	-	patients	668.3	+	time	413.1	-	back	370.9	-
2	radiology	145.1	+	patient	297.1	+	time	395.9	-	cancer	548.4	-	back	394.6	-	hospital	306.4	-
3	differential diagnosis	143.5	+	patients	264.2	+	cancer	386.2	-	breast	170.5	-	hospital	356.3	-	time	285.0	-
4	cancer	139.8	-	cancers	246.5	-	treatment	344.0	-	names	170.4	-	who	225.8	-	will	181.9	-
5	pathology	131.0	+	pathology	216.2	+	who	257.3	-	person	168.9	-	cell	177.8	+	came	144.2	-
6	chemotherapy	110.4	-	treatment	215.5	-	carcinoma	225.0	+	patient	120.8	+	patients	164.6	+	cancers	128.6	+
7	findings	97.3	+	differential diagnosis	207.9	+	hospital	217.2	-	blood	103.0	-	will	163.0	-	radiotherapy	119.9	-
8	treatment	95.4	-	findings	195.4	+	tumor	206.0	+	radiotherapy	102.4	+	life	161.0	-	appointment	119.4	-
9	history	91.1	+	radiation	171.2	-	life	189.4	-	cancers	90.4	-	tumors	160.4	+	bit	112.5	-
10	therapy	89.0	-	names	162.6	-	tumors	154.0	+	serum	87.2	+	tumor	157.5	+	work	110.1	-
11	necrosis	85.1	+	chemotherapy	162.2	-	cell	148.5	+	breast cancer	87.2	-	came	148.1	-	radiation therapy	106.2	+
12	power	84.9	+	person	154.6	-	came	138.8	-	metastases	86.2	+	breast	147.1	-	patients	105.9	-
13	thyroid	81.1	+	radiation therapy	150.5	-	bit	138.3	-	intestine	82.4	-	hair	140.7	-	cells	105.5	+
14	ultrasound	80.7	+	drugs	141.5	-	chemotherapy	133.8	-	who	82.3	-	work	138.4	-	hair	104.7	-
15	tomography	79.7	+	radiology	140.3	+	appointment	128.6	-	blood cells	71.4	-	carcinoma	135.2	+	life	99.8	-
16	drugs	74.2	-	symptoms	138.5	-	feeling	126.2	-	women	63.5	-	bit	131.9	-	news	99.1	-
17	diagnosis	73.7	+	carcinoma	128.2	+	differential diagnos	122.0	+	incidence	59.4	+	appointment	122.6	-	nurse	90.1	-
18	leukemia	68.7	-	history	122.7	+	diagnosis	121.4	+	heart	58.7	-	feeling	120.3	-	cancer	85.5	+
19	woman	61.5	+	surgery	111.5	-	hair	115.8	-	cell	55.9	+	treatment	107.2	-	feeling	83.3	-
20	cyst	56.3	+	skin	108.9	-	findings	115.0	+	tumor	53.9	+	family	104.8	-	husband	81.1	-
21	patient	55.8	+	necrosis	108.0	+	work	113.6	-	tomography	51.5	-	news	101.8	-	insurance	81.1	-
22	symptoms and signs	52.9	-	diagnosis	105.0	+	will	112.0	-	disease	49.4	+	cancer	94.9	-	friends	79.3	-
23	chondrosarcoma	52.6	+	blood	104.1	-	treatments	109.8	-	tumors	46.0	+	nurse	92.6	-	names	77.4	+
24	cytoplasm	52.3	+	who	103.2	-	news	106.8	-	large intestine	44.4	-	friends	90.2	-	leukemia	76.0	+
25	radiotherapy	51.1	-	thyroid	102.3	+	nurse	97.1	-	skin	44.3	-	treatments	88.0	-	family	75.4	-
26	anemia	50.7	-	women	98.1	-	family	95.4	-	radiation therapy	43.6	-	mother	83.3	-	mother	72.7	-
27	mph	49.7	+	tumor	97.6	+	cells	95.1	+	radiation	43.3	-	husband	83.3	-	cell	66.0	+
28	platelet	49.2	-	therapy	93.9	-	friends	94.6	-	metastasis	43.1	+	insurance	83.3	-	consultant	65.3	-
29	pancreas	48.9	+	neoplasm	91.0	+	pathology	92.2	+	signs	43.0	+	breast cancer	79.5	-	tumors	64.4	+
30	magnetic	48.4	+	intestine	88.5	-	husband	87.4	-	uterus	40.8	-	consultant	67.1	-	carcinoma	61.1	+
31	magnetic resonance	48.4	+	breast cancer	87.6	-	breast cancer	79.1	-	symptoms and signs	40.5	+	leukemia	66.4	+	intestine	59.6	+
32	microscopy	48.1	+	metastases	86.0	+	radiotherapy	69.2	-	united states	40.1	-	metastases	65.1	+	future	58.6	-
33	gallbladder	47.0	+	blood cells	84.9	-	insurance	68.3	-	syndrome	39.5	+	friend	57.8	-	friend	56.3	-
34	signs	46.5	-	leukemia	83.6	-	patient	66.5	+	carcinoma	36.4	+	cells	55.7	+	who	54.8	-

Appendix 5: Log-Likelihood comparison of MeSH terms

35	lymphomas	46.2	-	hemorrhage	79.6	+	necrosis	65.6	+	neoplasms	36.2	+	let	55.5	-	skin	54.4	+
36	hospital	45.7	+	power	73.3	+	metastases	64.5	+	woman	34.6	-	letter	55.5	-	letter	54.1	-
37	white	44.9	+	histology	68.3	+	consultant	62.3	-	erythrocytosis	33.5	+	reading	55.5	-	reading	54.1	-
38	radiation	44.7	-	lung cancer	67.5	-	adenocarcinoma	60.7	+	plasma	33.0	+	wife	55.5	-	bed	54.1	-
39	syndrome	44.5	-	neoplasms	67.0	+	friend	60.7	-	white blood cells	32.9	-	son	53.2	-	wife	54.1	-
40	cancers	44.3	-	procedure	62.2	-	let	58.2	-	symptoms	32.8	-	future	52.4	-	symptoms	53.2	+
41	breast	41.0	+	risk	60.7	-	reading	58.2	-	lymph nodes	32.0	-	serum	49.1	+	tumor	52.2	+
42	antigens	40.8	-	adenoma	59.4	+	radiology	57.8	+	red blood cells	31.5	-	sleep	48.6	-	son	51.8	-
43	neoplasm	40.5	+	metastasis	57.7	+	future	55.2	-	magnetic	31.5	-	syndrome	48.4	+	patient	51.8	-
44	disease	39.8	-	microscopy	54.4	+	surgery	55.2	-	etiology	30.9	+	mastectomy	47.9	-	sleep	47.3	-
45	marrow	38.3	-	serum	54.4	+	pancreas	55.0	+	procedure	30.8	-	hand	47.9	-	risk	46.7	+
46	x rays	38.2	-	heart	52.9	-	mother	54.1	-	prostate	30.5	-	disease	47.6	+	let	46.5	-
47	symptoms	37.5	-	breast	52.3	-	history	53.5	+	magnetic resonance in	30.1	-	person	46.3	-	specialist	45.1	-
48	histology	37.4	+	cytoplasm	50.6	+	tissue	53.2	+	magnetic resonance	30.1	-	tablets	46.3	-	tablets	45.1	-
49	bone marrow	37.4	-	staining	50.6	+	thyroid	52.2	+	mastectomy	30.0	-	research	45.7	-	lung	43.3	+
50	human	37.2	-	chondrosarcoma	50.5	+	control	51.0	-	gliomas	29.5	+	incidence	44.6	+	computer	40.5	-