Karin Friberg Heppin

Resolving Power of Search Keys in MedEval a Swedish Medical Text Collection with User Groups: Doctors and Patients

Data linguistica

<http://hum.gu.se/institutioner/svenska-spraket/publ/datal/>

Editor: Lars Borin

Språkbanken • Språkdata
Department of Swedish Language
University of Gothenburg

22 • 2010

Karin Friberg Heppin

Resolving Power of Search Keys in MedEval a Swedish Medical Text Collection with User Groups: Doctors and Patients

In the quest for an ultimate query

Gothenburg 2010

ISBN 978-91-87850-41-7
ISSN 0347-948X

Printed in Sweden by
Intellecta Infolog Göteborg 2010

Typeset in LaTeX 2ε by the author

Cover design by Kjell Edgren, Informat.se

Front cover illustration:
Looking for me?
by Emil Werin ©

Author photo on back cover by Rudolf Rydstedt

ABSTRACT

This thesis describes the making of a Swedish medical test collection, unique in its kind in providing a possibility to choose user group: doctors or patients. The thesis also describes a series of pilot studies which demonstrate what kind of studies can be performed with such a collection. The pilot studies are focused on search key effectivity: What makes a search key good, and what makes a search key bad?

The need to bring linguistics and consideration of terminology into the information retrieval research field is demonstrated. Most information retrieval is about finding free text documents. Documents are built of terms, as are topics and search queries. It is important to understand the functions and features of these terms and not treat them like featureless objects. The thesis concludes that terms are not equal, but show very different behavior.

The thesis addresses the problem of compounds, which, if used as search keys, will not match corresponding simplex words in the documents, while simplex words as search keys will not match corresponding compounds in the documents. The thesis discusses how compounds can be split to obtain more matches, without lowering the quality of a search.

Another important aspect of the thesis is that it considers how different language registers, in this case those of doctors and patients, can be utilized to find documents written with one of the groups in mind. As the test collection contains a large set of documents marked for intended target group, doctors or patients, the language differences can be and are studied. The author comes up with suggestions of how to choose search keys if documents from one category or the other are desired.

Information retrieval is a multi-disciplinary research field. It involves computer science, information science, and natural language processing. There is a substantial amount of research behind the algorithms of modern search engines, but even with the best possible search algorithm the result of a search will not be successful without an effective query constructed with effective search keys.

SAMMANFATTNING

Denna avhandling beskriver arbetet med att bygga en svensk medicinsk testkollektion, unik i den meningen att det är möjligt att välja användargrupp: läkare eller patienter. Avhandlingen beskriver också en serie pilotundersökningar som visar på vilken typ av forskning som kan bedrivas med hjälp av en sådan kollektion. Pilotundersökningarna är inriktade på att förstå vad som gör sökord effektiva eller inte: Vilka egenskaper utmärker bra sökord och vilka egenskaper utmärker dåliga sökord?

Avhandlingen visar på nödvändigheten av att föra in språkvetenskap och beaktande av terminologi i forskningen om informationssökning. Informationssökning handlar främst om att söka fram information uttryckt i fri text. Text består av termer, de ämnen man söker efter beskrivs med termer och sökfrågor består av termer. Det är viktigt att förstå vilka egenskaper dessa termer har som är så viktiga inom informationssökning.

Avhandlingen beskriver de problemställningar som uppkommer i samband med sammansatta ord. Sammansatta ord som används som sökord matchar inte de enkla ord som sammansättningarna består av om det är dessa ord som står i ett dokument och inte sammansättningen. Det blir heller inte någon sökträff om sökordet är enkelt men det som står i dokumentet är en sammansättning. I avhandlingen diskuteras hur sammansättningar i sökord och text kan delas in i sina beståndsdelar utan att sökresultatet försämras.

En annan viktig aspekt i avhandlingen är att undersöka hur språket hos läkare och patienter skiljer sig åt för att se om det skulle gå att använda sig av dessa skillnader för att hitta dokument som riktar sig till den ena eller andra gruppen. Testkollektionens dokumentsamling innehåller en stor mängd texter som är uppmärkta med målgrupp: läkare eller patienter. Dessa dokument används för att undersöka språkskillnader mellan de båda grupperna och för att föreslå lämpliga metoder att rikta sökningar för att hitta dokument skrivna för respektive grupp.

Informationssökning är ett tvärvetenskapligt ämne som innefattar informationsvetenskap, datavetenskap och språkvetenskap. Det ligger mycket forskning bakom de sökalgoritmer som används idag. Men inte ens med den bästa sökalgoritm blir det ett lyckat resultat om inte den sökfråga som används är effektiv.

ACKNOWLEDGEMENTS

Thanks to Lars Borin, my advisor, who showed me the way
Jussi Karlgren, my co-advisor, who gave me confidence
Maria Toporowska-Gronostaj my first co-advisor
Martin Volk my second co-advisor
Dimitrios Kokkinakis, who proved to be an endless source of resources
Sofie Johansson Kokkinakis whose good advice always turned out to be so very good
Leif-Jöran Olsson and Rudolf Rydstedt for practical help and many laughs
Monica Lassi who made that one suggestion which made all the difference
Per Ahlgren who opened the door to information retrieval
Leif Grönqvist who made me believe it all was possible
Eero Sormunen, Eija Airo, Jaana Kekäläinen, Kalervo Järvelin, Heikki Keskustalo, and everyone else in the FIRE research group for sound advice, encouragement, practical help with the test collection and for making me feel at home at the Department of Information Studies and Interactive Media at the University of Tampere
Ari Pirkola who helped me find the last piece of the puzzle
Sanna Kumpulainen and Saila Huuskonen who became my friends in Tampere
Ritva Aydi for always loving me when I need it most
Martin Dackling for showing me how to take charge
Lilja Øvrelid for putting up with all my questions
Maia Andreasson who is always so very wise
Lena Ulrika Rudeke for brainstorming for theme
Tina and Henrik for helping me vizualize it all
Elias for artistic advise
Torbjörn, Helena, Stefan and Johan without whose judgments there would be nothing
The Graduate School of Language Technology that gave me a fantastic research environment
NGSLT for funding trips
My beloved husband Pär for endless love, endless support, even more endless love and endless proofreading
Emil & Anton, my number 42s, for all their love and hugs

Mammi & Pappi for encouragement, inspiration, support, sound advice, and endless proofreading
Håkan for challenges and support, and for finally letting me win a challenge
Monika for always being there
All my friends, colleagues, and close ones for everything
Douglas Adams for writing the trilogy in four parts *The Hitch Hiker's Guide to the Galaxy* (*The Hitch Hiker's Guide to the Galaxy, The Restaurant at the end of the Universe, Life, the Universe and Everything, So Long, and Thanks for all the Fish*) and the sequel *Mostly Harmless*, books from which I have borrowed all the intelligent little quotes dispersed here and there. He borrowed one from Bob Dylan. So thanks to Bob Dylan as well.

Share and Enjoy[1]

Karin Friberg Heppin
Gothenburg October 2010

[1] The Restaurant at the end of the Universe, chapter 2

CONTENTS

1 INTRODUCTION

DON'T PANIC

The Hitch Hiker's... chapter 3

What type of search keys are effective when searching for information in a collection of documents? What is the best way to treat compounds? When is it beneficial to use individual compound constituents as search keys and when does it ruin a search? Research on compounds in information retrieval has previously focused on whether or not to split compounds and whether or not to use the constituents as search keys alongside the original compound. My intention was to develop these ideas and not only split the compounds and use the parts without discrimination, but to study the parts to find features that distinguished constituents that yielded good search results from those that did not. This idea grew, and I found myself studying, not only compound parts, but search keys and terms in general. For should it not be the case that once a compound was split, the parts ought to behave as ordinary simplex words? Features that are good or bad for simplex words, ought they not be just as good for compound constituents, seeing that compound constituents basically are simplex words?

Most information retrieval resources are constructed for documents written in the English language. As I set out to study information retrieval in the Swedish language, and especially compounds in medical information retrieval, I needed Swedish resources. Swedish compounds are much more productive than English compounds. Swedish nouns and adjectives have a richer inflectional system than English ones. English resources would just not do.

During my time as a PhD candidate I was part of a research team at the Department of Swedish Language at the University of Gothenburg, doing research on medical information. This research includes studying the difference between information for doctors and information for patients. As I wanted my research to conform to other research at the department I decided to work with medical information retrieval.

For laboratory information retrieval research the **test collection** is essential. I needed not only a Swedish test collection, but also a collection that was domain specific, from the medical domain. To my knowledge, there existed no such collection, so initiating the creation of a Swedish medical test collection became part of my thesis work. The test collection is intended to **Eval**uate search strategies for **Med**ical documents, hence the name **MedEval**. As the department had worked with the difference in doctor and patient language we decided to build a test collection that regarded user groups, medical professionals and lay persons.

When working wiht the Swedish language the study of compounds is important, seeing that around 10%[2] of words in Swedish running text are compounds. To have a test collection which made it possible to perform research on compounds, it was decided that the MedEval test collection should not only enable the user to choose user group, but it would also have double indexes: one where the compounds were kept intact, and one where the compounds were split into constituents. This resulted in a test collection with unique possibilities for user choices.

This thesis describes the making of MedEval, a Swedish medical test collection. It also provides a series of pilot studies to show what kind of studies can be performed with such a collection. The pilot studies are focused on search key effectivity: What properties do effective and non-effective search keys have? Compounds and compound constituents used as search keys are given special attention. Another aspect in the pilot studies is how different language registers, in this case those of doctors and patients, can be utilized to find documents written with a specific group in mind.

1.1 Research questions

- What features do terms that are good search keys have? What features do terms that are bad search keys have? Can this knowledge be used to select compound constituents to use as search keys?

- Can specific features of professional language and of lay person language, respectively, be utilized when searching for medical documents for the two target groups?

- Can the questions above be answered using a medical test collection with two indexes containing different representations of compounds, split and

[2]The percentage depends on the genre of text: bureaucratic and scientific text have a higher percentage and fiction lower.

unsplit, and providing user group scenarios, professionals and lay persons? What other research questions can be answered with such a collection?

1.2 Outline of the thesis

Part I. Background The background chapters give a basic understanding of the research fields on which this work stands.

Information retrieval Chapter 2 gives an introduction to information retrieval.

Evaluation Chapter 3 gives a brief history and description of evaluation in information retrieval.

Linguistics Chapter 4 gives a linguistic background and describes linguistic issues that have impact on information retrieval. Special attention is given to compounds, which have an essential position in this thesis and in the Swedish language.

Medical information retrieval Chapter 5 presents previous work in medical information retrieval and in doctor/patient language.

Resolving power Finally, chapter 6 describes the development of views on the resolving power of terms.

Part II. The guide to the galaxy of the thesis The program declaration.

Drawing the road map Chapter 7 describes how the goals and intentions of the thesis came into being.

The travel instructions Chapter 8 contains a more practical description of the plan to reach the goal.

Part III. Resources In this part resources are described, first the preexisting ones, and then the building of the new resource: MedEval.

Tools Chapter 9 describes the tools and resources which have been used for the present research.

MedEval construction In chapter 10 the construction process of the MedEval test collection is described.

Part IV. Pilot studies This part presents a smorgasbord of experiments intended to give an idea of which kinds of experiments can be performed with the MedEval test collection.

Facets Chapter 11 describes some methods and obstacles in dividing topics into facets and in choosing search keys.

Survey Chapter 12 describes the initial survey of facets and search keys.

Search keys Chapter 13 describes the behavior and impact of different types of search keys.

Target and user groups Chapter 14 focuses on differences found in the language of documents assessed to have doctor target group, and of those assessed to have patient target group. It also suggests possible ways to utilize these differences when searching for documents for the two groups respectively.

Part V. Conclusions What it all leads up to.

1.3 Contributions

By writing this thesis and by the doing the work behind it I have contributed with the following:

- Creating a Swedish medical test collection containing 42 000 documents with:
  - documents assessed on a four-graded scale of relevance which allows for a fine-grained study of retrieval effectiveness.
  - documents assessed for target reader group: doctors or patients. This allows studies of retrieval of documents based not only on topic relevance but also on intended audience.
  - a collection of documents marked for target reader group: doctors or patients. This allows studies of differences in the language used when addressing these groups.
  - the potential of being a valuable resource in teaching. It could be used in language technology for studies in information retrieval and in linguistics for studies in genre specific language.
- Demonstrating the need to bring linguistics and terminology into the information retrieval research field, because:
  - most IR is done on text. It is important to understand the functions and features of the object of one's research, not just treat it like an object.

  - even with the best possible search algorithm the result of a search will not be successful without an effective query with effective search keys. The thesis shows that terms are not all equal, but show very different behavior. Therefore they should not be treated in batch mode.

- Describing research projects which can be done with a test collection such as the one described above, and performing pilot studies such as:

  - studying features of effective search terms and of ineffective search terms.
  - studying how one can utilize the difference in the language of documents written for professionals and for lay persons when retrieving documents for the different groups.

1.4 Notes on language and terms used in this thesis

An essential aspect of this thesis is to study the impact in information retrieval of Swedish terms with different properties, so it is important to convey to the reader the meaning and properties of the terms discussed. Therefore, when providing English equivalents, more attention has been paid to presenting literal translations than linguistically correct equivalents. When providing English equivalents, the literal translation is within quotation marks and the more linguistically accepted equivalent within parentheses, for instance: *kransartär* 'wreath artery' (coronary artery).

The terms **doctors**, (**medical**) **professionals**, and **experts** will be used more or less interchangeably in this thesis. The group of people that these terms refer to are all kinds of medical professionals, including nurses and other staff involved in the medical care of patients. The crucial feature is that the persons in this group have an education in the medical field.

The terms **patients**, **lay persons**, and **non-experts** will likewise be used more or less interchangeably. The crucial feature for this group is that they do not have a medical education.

Part I

Background

2 INFORMATION RETRIEVAL

There's a frood who really knows where his towel is

The Hitch Hiker's... chapter 3

Information retrieval, also called IR, is about representing, organizing and storing information items, or documents, so that they later can be located and retrieved if they contain information that satisfies a user's need of information (Baeza-Yates and Ribeiro-Neto 1999). In fact, information retrieval is not only about finding relevant documents, it is also about distinguishing these documents from the ones that are not relevant, the noise. This is becoming ever more important as the quantity of information available grows exponentially. The aim of an information retrieval system is to help the user to quickly sort out the documents most likely to contain the information which the user needs.

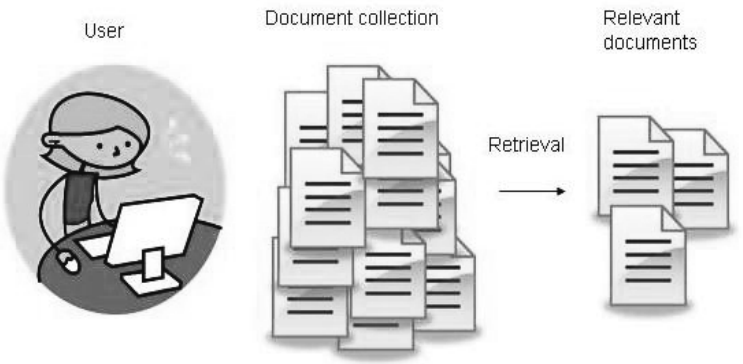


Figure 1: Information retrieval is about finding documents relevant to a user's need in a large collection of documents. The documents of the collection have been represented, organized and stored so that they can be found and retrieved when there is a need for the information in them.

Information retrieval is a field that spans from information science and computer science to linguistics. It covers the whole process from storing and indexing documents to retrieving them using search engines. In the early days of information retrieval, IR systems mostly concerned bibliographic information such as title, source and a few key words. The searches were typically not done on the full text of the documents, but on manually administered index terms which were chosen from a controlled vocabulary. Often it was not the person who needed information that performed the search, but professionally trained librarians. In the couple of last decades, we have seen a tremendous growth of computer power. Document databases have grown, not only in the number of documents stored, but also in content. Now the indexes, as a rule, contain the full text of the document resources.

Figure 2 shows the information retrieval process (Ingwersen and Järvelin 2005). This figure will recur, in a smaller scale, throughout the thesis when different parts of the process are discussed. The part concerned will in each case be in a darker color.

There are two prerequisites for information retrieval: documents to search for and reasons to search, the so called **information needs**. The documents have to be processed and stored in a way so that it is possible to find them again when needed. This is usually done by storing terms and features in indexes. A user has a situation where he or she needs information. The user must express this in some way and put a query to the search system. In order to get at relevant information the user's query is matched to the stored information in the index. The output of the system is a list of documents that have a high probability of answering the user's need. In order to evaluate this output there must be information available about which documents in the collection are relevant, that is, which documents the user would want to retrieve. For this reason preevaluations of documents are done, so called **relevance assessments**. Documents that have a high probability of being relevant to each information need are selected and judged in advance. The documents that are preassessed to be relevant are the **recall base**, that is the documents the user tries to get the system to retrieve. In the evaluation of the search result the list of retrieved documents is compared to stored information about the recall base. If the user is not satisfied with the query result he or she can use this result to again analyze the information need and make a new and better query.

Information retrieval has two main categories. **Ad hoc** retrieval where the queries vary but the collection of documents is more or less static, and **routing** or **filtering** where the queries remain more or less the same over time, but where the document collection is continuously changing. This is for example the retrieval approach employed in news wire services, where the main task is to describe the user profile and to implement this on a steady flow of new information.

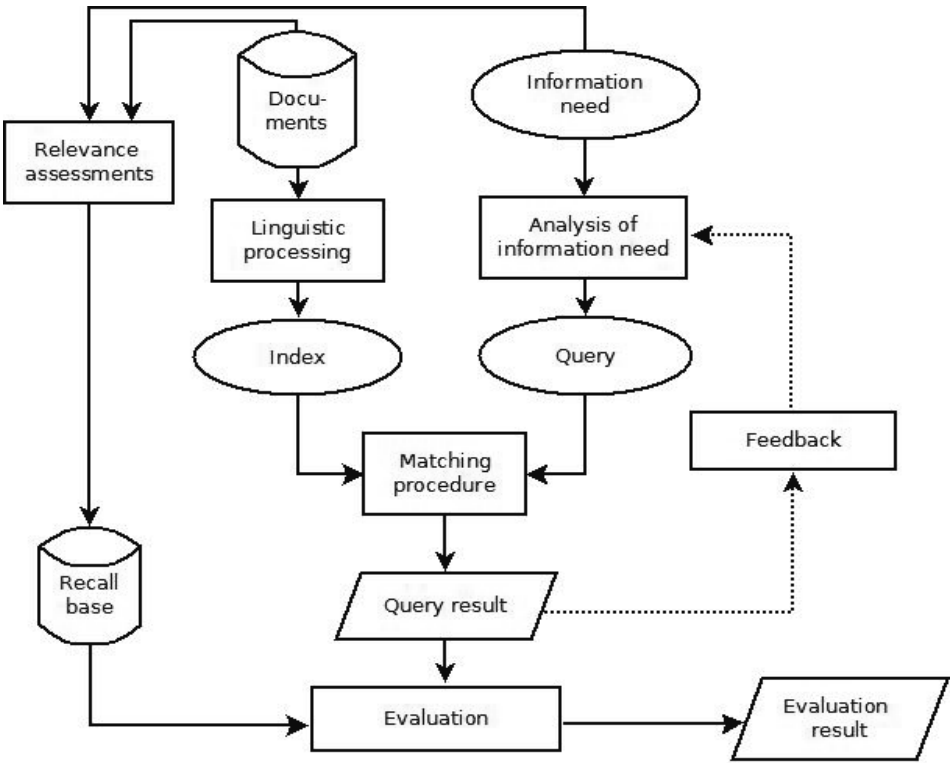


Figure 2: The information retrieval process. This figure will recur in a smaller scale at different points in this thesis with selected items in a darker color representing the part of the information retrieval process that is being discussed in that specific section.

Searching on the web can be referred to as ad hoc retrieval, although web data sets differ in many aspects from the standard ad hoc retrieval data sets (Metzler Jr. 2007). The web data sets are considerably larger, they are noisier, as there is no controll over what is published and they are less static. They also contain HTML markup and links which can be utilized in the search (see further section 2.5).

As the focus of this thesis is on search queries, and the MedEval document collection is static, information retrieval will henceforth refer to ad hoc retrieval, unless otherwise stated.

Information retrieval has a related research field in information extraction. Information extraction is about finding the information itself. In other words, the process of information retrieval returns documents, while the process of

information extraction returns information or facts found in the retrieved documents.

2.1 Basic terminology

As in any scientific field, there are a number of terms used in information retrieval. Below, the central terms in this thesis are given appropriate definitions.

Document A unit of information. The content of a document can be anything from pictures or music to newspaper articles and Internet pages. Most often it is text in some form. Documents are the items retrieved by the search system. In this thesis the term 'document' refers to text documents.

Index Storage of text and meta data about the documents of the collection, typically information about which terms the documents contain and the position of these terms in the documents. Additional information in the index can be titles, structures of the documents, dates of publication, authors, sources, types of publication and the like.

Index term A representation of a linguistic unit. It is used in the index to store information about the content of a document. Index terms do not have to be scientific expressions and they do not have to be of a certain form. They can represent simplex words, compounds, phrases or n-grams.

Information need A description of the subject that the user requires information about. In evaluation studies such as the present one, the information needs should be samples of realistic requests.

Topic Alternative expression for information need.

Query A formal representation of an information need, which is posed by the user to the retrieval system. The query is expressed by use of terms and operators in a formal language that the retrieval system can process. Put simply, it is the string that the user enters into the search field of the search engine.

Search key A single term in a query that is used to represent a concept of an information need. If it is to have any function it must be matched to an index term.

Query term A query term is a building block of a query. It can consist of a single search key, or several search keys that with the help of operators are organized as ordered or unordered phrases, as synonyms etc.

Collection frequency, cf The absolute frequency of a term in a collection of documents.

Term frequency, tf The absolute frequency of a term in a certain document. If t is a term and d a document, then $\mathrm{tf}_{t,d}$ is the term frequency for term t in document d.

Document frequency, df The number of documents which the term t appears in. If t is a term then df_t is the document frequency of term t.

Inverse document frequency, idf If N is the total number of documents in the corpus, then the inverse document frequency for the term t can be described as follows:

$$\mathrm{idf}_t = \log \frac{N}{\mathrm{df}_t} \tag{1}$$

tf∗idf factor The tf and idf factors combined into a single weight. It has high values for terms with high frequencies in few documents and can be described as follows:

$$\mathrm{tf}{*}\mathrm{idf}_{t,d} = \mathrm{tf}_{t,d} \times \mathrm{idf}_t \tag{2}$$

Facet A facet contains one or several terms and represents one concept of an information need. The terms in a facet play a similar semantic role. They can be, for example, near synonyms or more or less specific expressions for the same concept, but do not have to belong to the same part-of speech.[3]

Query operator Determines the relationship between query terms, for instance if terms should have different weights, if they should be treated as synonyms or if they must appear within a certain distance of each other.

Relevance The degree to which a document meets an information need. In this thesis it is mainly used in the sense of **topical relevance** which is the degree to which a document is *about* the information need, not to which degree it satisfies the user.

Test collection A laboratory environment that is used for research in information retrieval. It consists of a set of documents, a set of queries, and a set of known relevant documents for each information need.

[3]The term *facet* is polysemous. In information retrieval it is not used in the same meaning as in linguistics (see page 59).

Search engine The computer software that matches information in queries to information in indexes, then retrieves and, in most cases, ranks documents in order of calculated relevance.

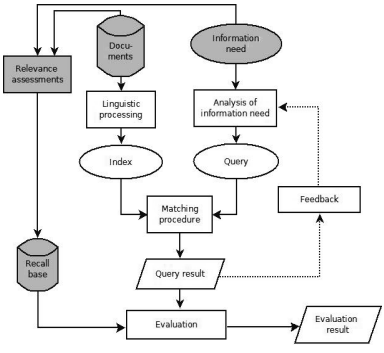
Document cut off value, DCV The position in the ranked list of retrieved documents at which a measure of effectiveness is calculated. For instance: 'Precision at DCV 10' means precision after 10 retrieved documents.

Run The result of a search. In an exact-match system the result is a set (of sets) of documents, while in a best-match system it is a (set of) ranked list(s) of documents.

Cross lingual information retrieval, CLIR Information retrieval where the query is put in one language and the documents retrieved are in another language. Machine translation or an interlingua can be used.

2.2 Test collection

To carry out research in information retrieval, a so called **test collection** is often used. A test collection is a test environment, in a laboratory setting, representing document collections and search environments in the real world.

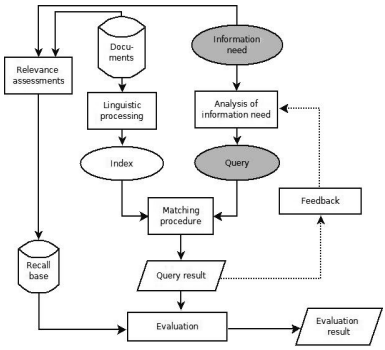


A test collection consists of three parts: (1) a static collection of **documents**, (2) a set of **information needs**, and (3) a set of known **relevant documents** for each information need. The sets of relevant documents in (3) are all subsets of the bigger set of documents in (2) (Baeza-Yates and Ribeiro-Neto 1999). To get a set of known relevant documents, documents are manually assessed and marked for relevance in advance.

The facts that the documents in the collection remain the same, and that the collection contains sets of documents which are known to be relevant to a certain set of information needs, make it possible to evaluate and compare search strategies. The laboratory setting allows the researcher to change one feature at a time in search algorithm, index or query. It is the relevance assessments that turn a collection of texts and topics into an information retrieval test collection.

2.3 Information needs and queries

Information needs are the subjects that the user want information about. In a test collection ther is a fixed set of subjects. These needs, or topics, are used as a base for posing queries to the system. Queries are created by the user searching for documents and are specific for each run. Information needs are static and do not change, while queries can be modified and expanded for new runs if the user is not satisfied with the results, that is, if the results of running a query does not meet the information need sufficiently.



Making a query entails transforming an information need into a formal request which can be put to the system. Query construction techniques can be either automatic or manual (Metzler Jr. 2007).

A query can be a simple list of terms, in which case all terms are given the same importance. The term list query allows for the user to write queries in natural language. However, such a query is not parsed or interpreted in any way. The terms are treated as a bag of words.

A query can contain operators that represent relations between terms or that give terms of varying importance different weights. These operators can be extended to specify that a term must occur in a certain document field, such as the title field or description field.

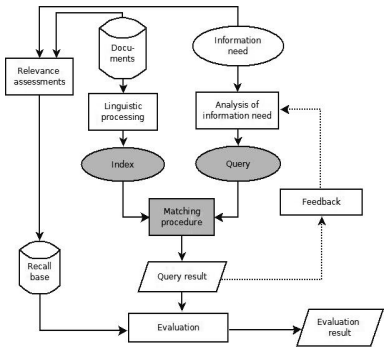
Information needs or topics vary greatly as to how difficult it is to retrieve documents which are relevant to them. It is therefore important to have a sufficient number of topics in the test collection used, making it probable that topics from varying levels of difficulty are represented.

In Carmel et al. 2006 the authors try to establish what it is that makes a topic difficult. They describe a topic as being dependent on the set of queries that can represent the topic and also on the document collection from which the set of relevant documents is retrieved. They claim that the topic difficulty is determined by how broad the topic is (how easy the topic is to describe) and how well the relevant documents are separated from the rest of the collection (how easy it is to differentiate the relevant documents from the non-relevant ones). This can be described by the following distances between components:

- The distance between the set of queries and the collection
- The distance among the queries

- The distance between the relevant documents and the rest of the collection
- The distance among the relevant documents
- The distance between the queries and the relevant documents

2.4 The classical models



A model of an information retrieval system contains four parts: (1) a set of logical representations of documents, (2) a set of logical representations of information needs, so called queries, (3) a framework for modeling the document representations, the queries and the relations between them, (4) a function which for a binary system determines if a query and a document match or not, and for a best-match system associates a query and a document representation with a real number. This number is used to sort the documents in order of similarity to the query and rank the documents. Different models employ different algorithms, but most use term frequencies in some way.

A binary, or exact-match, retrieval system returns documents that match the query fully. The retrieved documents are not ranked. A best-match retrieval system returns a list of documents where the documents are ranked in order of calculated relevance. The better the match between the query and a document, the higher the document is ranked. Relevant documents will hopefully be retrieved and ranked high, while non-relevant ones should not be retrieved, or should at least be ranked lower than more relevant ones.

2.4.1 The Boolean model

An early information retrieval model is the exact-match Boolean model. Its framework is based on set theory, and the matching function on Boolean algebra. Documents are retrieved if what is expressed in the query is true for the document. Boolean systems use binary weights. If an index term appears in a document, it is true for that document, if it does not, it is false (Baeza-Yates and Ribeiro-Neto 1999).

Boolean queries are logic expressions constructed of terms, the operators **AND**, **OR**, and **NOT** and brackets. The user divides the query into different concepts or facets. The terms within a facet are joined by disjunction, the operator **OR**. The facets are in turn joined by conjunction, the operator **AND**. If A, B, and C are the facets of a search query, then example 3 below is a query in the Boolean model.

$$(A_1 \textbf{ OR } A_2 \textbf{ OR } A_3) \textbf{ AND } (B_1 \textbf{ OR } B_2) \textbf{ AND } (C_1 \textbf{ OR } C_2) \tag{3}$$

The Boolean model is based on the assumption that if a document contains a term it is about this term. With this model a document has to satisfy the whole Boolean expression in order to be retrieved. At least one term per facet must be present in a document if there is to be a match. It also means that either the document satisfies the query or it does not. There is no partial matching. With no partial matching, there can be no ranking.

It is difficult to compose a Boolean query that gives a good result. Often the result is a retrieved set of documents which is either very large, if there are just a few facets, or empty if there are many.

2.4.2 The vector space model

With the vector space model the framework is a k-dimensional vectorial space and the ranking function is based on standard linear algebra operations on vectors. Here partial matching is possible. With partial matching it is not only the question if a document is retrieved or not, there is a computed degree of similarity between query and retrieved documents (Baeza-Yates and Ribeiro-Neto 1999). In the vector space model queries and documents are represented in a similar way by a k-dimensional vector, where k is the number of terms present in the document database. The document D_i can thus be represented by the vector in example 4, where d_{ij} is the weight of the jth term in the document (Croft, Metzler and Strohman 2010).

$$D_i = (d_{i1}, d_{i2}, \ldots, d_{ik}) \tag{4}$$

A document collection with n documents can be represented by a matrix of term weights. Each row in example 5 represents a document and the columns represent the terms of the collection.

$$\begin{array}{cccccc} & Term_1 & Term_2 & \dots & Term_k \\ D_1 & d_{11} & d_{12} & \dots & d_{1k} \\ D_2 & d_{21} & d_{22} & \dots & d_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ D_n & d_{n1} & d_{n2} & \dots & d_{nk} \end{array} \tag{5}$$

The query is regarded as a small document and represented in a vector similar to the document vectors, as in example 6.

$$Q = (q_1, q_2, \dots, q_k) \tag{6}$$

The document scores for the ranking is calculated by comparing the document vectors to the query vector by a similarity measure, for example the cosine measure. The idea of representing queries and text documents in the same manner and of comparing the contents for similarity in a statistical matching process was put forward by Luhn as early as in the late 1950s.

> This inquirer's document would then be coded in exactly the same manner as the documents of the collection have been encoded [...] Since an identical match is highly improbable, this process would be carried out on a statistical basis by asking for a given degree of similarity. (Luhn 1957: 316)

The assumption is that the more similar a document is to a query, the more relevant the document is to the information need that the query represents.

2.4.3 Probabilistic models

In probabilistic models the documents are ranked according to the calculated probabilities of their relevance to a query. They have a probabilistic framework based on sets and a ranking function based on standard probability operations and Bayes' theorem (Baeza-Yates and Ribeiro-Neto 1999).

2.4.3.1 The binary independence retrieval model

The classical probabilistic model, binary independence retrieval, was introduced by Robertson and Spärck Jones (1976). Here 'binary' implies Boolean, documents are either relevant or non-relevant, and 'independence' implies that the terms, in this model, are treated as if they occur in documents independently of each other, which of course is a simplification (Manning, Raghavan

and Schütze 2008). The idea is that for every topic there is an *ideal set*, R, of relevant documents. This set contains all the relevant documents and no others. The model assumes that the probability that a document will belong to R is dependent only on the topic and the document representation in the index. What the user has to do is find the ideal terms to describe R. The user can start by guessing the best terms. This guess generates a preliminary set of documents. The user can then give the system feedback by stating which of these documents truly were relevant. On the basis of this information the system refines the description of the ideal set. This process is repeated until a satisfying description of R is obtained (Baeza-Yates and Ribeiro-Neto 1999).

2.4.3.2 Language models

In the late 90s new IR models were proposed that were based on the idea of statistical language models. These models were not based on linguistic knowledge, but rather on the data of the collections themselves. The common approach was not to calculate the probability of relevance of a document to a query, but rather the probability that a query could be generated by a statistical language model built on the data of the document. The probability of a query being generated is calculated for each document in the collection, and the results are used for ranking the documents. This is called the 'query likelihood' approach.

The unigram language model has a 'bag of words' approach, where the position of a term relative to other terms is not considered (Metzler Jr. 2007). In such a model the phrase *The doctor contaminated the blood* is equivalent to *The blood contaminated the doctor*.

The *n*-gram language model takes the context of terms into consideration. The generation of a term is there dependent on the previous [n-1] terms. For example, the term *pressure* would have a higher likelihood of appearing after the term *blood* than after the term *tree* (Kraaij 2004; Metzler Jr. 2007).

The language models give a more formal framework than the models based on tf∗idf estimates for term probabilities. The drawback is that they do not support structured queries.

2.4.3.3 The inference network model

The inference network model is based on the formalism of Bayesian networks. These networks are directed acyclic graphs. Each node represents an event with a certain set of outcomes and the arcs denote relationships represented as

probabilistic dependencies. The calculation of the probability scores in the inference network model was long based on tf∗idf factors. This worked satisfactorily in many systems, however Metzler and Croft (2004) claimed that there was little formal justification for these calculations. The authors suggested that the reason that tf∗idf was so widely used was simply that it turned out to work well in spite of its heuristic nature.

The Indri/Lemur search engine[4] used in the research for this thesis is an inference network based on Bayesian networks combined with language models. This gives it formal justification as well as allowing the use of a complex structured query language. Since this model is the one that will be used henceforth, it will be given a more thorough presentation than the other models.

An inference network contains the node types listed below. They can be seen in figure 3.

- Document node (D)
- Prior or smoothing parameter nodes (Pr)
- Language model nodes (Mo)
- Representation nodes (r)
- Belief or query nodes (q)
- Information need node (I)

The document node corresponds to the event that a document is observed. It is assumed that only one document is observed at a time which means that there is one network for every document. A document may be represented in a number of ways in the document node, for example by a function which takes a sequence of features and tries it on every position in the document, indicating for each feature and each position if the feature is present or not. This can be done for any easily indexed feature of a document, for instance, a single term or a more complex proximity representation, such as a phrase.

The prior nodes give the probability that a document is relevant, based on query independent features, before matching the contents of the document to a query. The simplest prior is $\frac{1}{N}$, N being the number of documents in the collection. More advanced priors could, for example, give higher probability to documents written for doctors to doctor users and documents written for patients to patient users. For web searches the priors could give probability depending on the number of inlinks to the document, on click-through data or on PageRank (see section 2.5) (Metzler Jr. 2007).

[4] <www.lemurproject.org/>

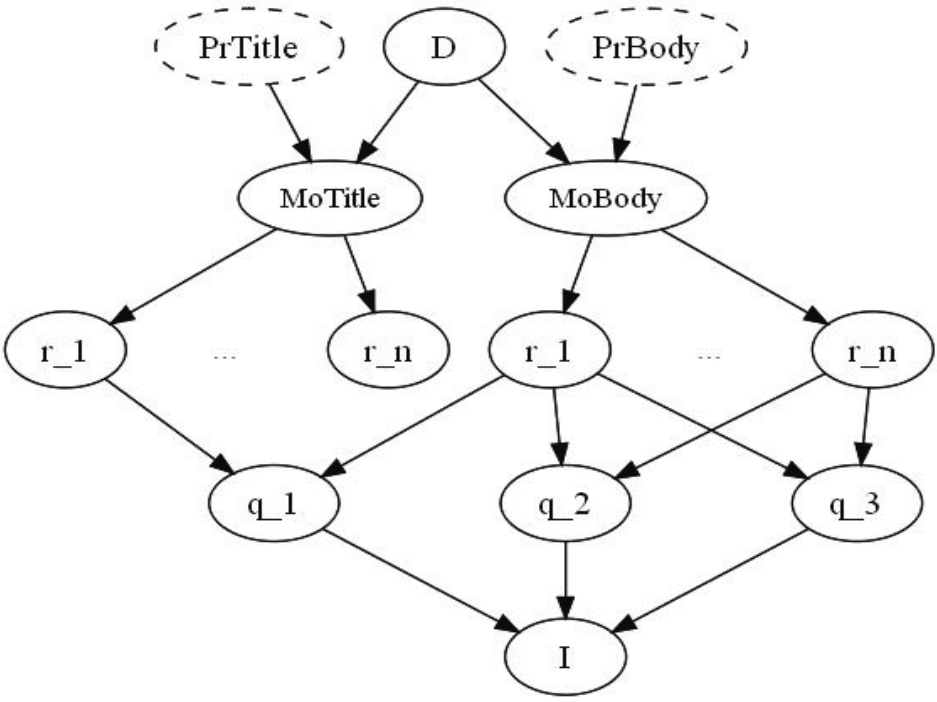


Figure 3: A simplified inference network model. The *D* node represents a document, the *r* nodes represent term occurrence or other features of the document. The *q* nodes represent the operators of a query and the *I* node represents an information need. *MoTitle* and *MoBody*, are language models of the document title and the document body which are estimated using data from the document and the prior nodes *PrTitle* and *PrBody*.

The language model nodes are estimated using the observed features in document *D* and the parameters in the prior nodes. Smoothing is used to overcome cases of zero probability. There may be more than one model node. If so, they correspond to different representations of the document, for example title, body or abstract (Croft, Metzler and Strohman 2010; Metzler 2005).

The representation nodes are binary random variables related to the features in the document representation. The information about the inference network as far down as the in the representation nodes can be precomputed and stored in the index for efficiency at runtime. Examples of events which can be represented in the representation nodes are shown below.

- The term ‘cancer’ occurred.
- The term ‘diabetes’ occurred in the title.
- The exact phrase ‘alzheimer’s disease’ occurred.

The belief or query nodes are also binary random variables. These are used to combine the probabilities of the representation nodes. They can also combine probabilities of other belief nodes, provided that the graph is still acyclic. The belief nodes are added to the inference network, as structured queries are put to the system. This means that the network structure, from the belief nodes down, changes for every query. An example of a network created for the query ‘#syn(lungcancer #uw3(lunga cancer))’ is shown in figure 4. The query indicates that *lungcancer* ‘lung cancer’ and the combination of *lunga* ‘lung’ and *cancer* ‘cancer’ appearing in any order within three tokens of each other, should be treated as instances of the same term.

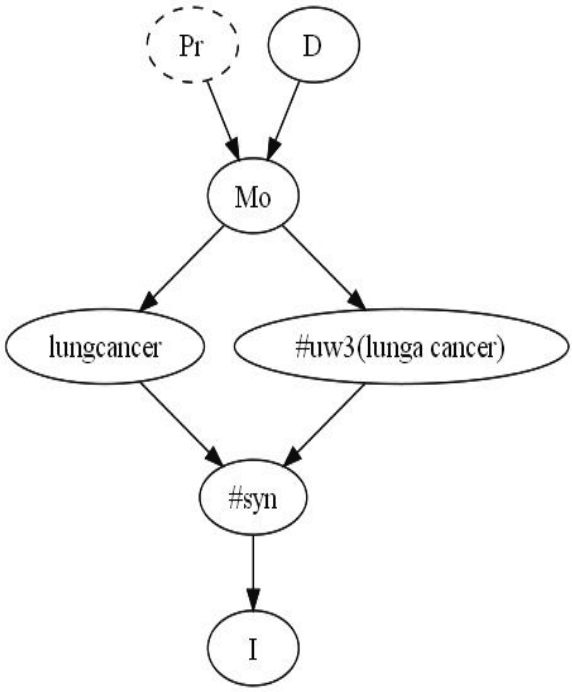


Figure 4: An example of a simple inference network model. This represents the probability that the information need *I* can be generated from the language model *Mo* based on document *D* and the priors *Pr* when the information need *I* is represented by the query #syn(lungcancer #uw3(lunga cancer)).

The information need node is actually a belief node which combines all information in the network into a single belief. The belief obtained is the belief that is used to rank the document comparing it to belief obtained in a similar manner, in other networks, for other documents.

In the inference network, given observed events make it possible to calculate the probability, or belief, that the different outcomes of the nodes will occur. The non-root nodes store conditional probability tables containing the probabilities of each outcome given the parent nodes. The root nodes, which do not have parent nodes, are assigned prior probabilities.

If we observe the document D, the node representing this document is instantiated with 'true'. The probability of observing the representation r_1 is determined by the conditional probability in the arcs between D and r_1. For each layer the probability is calculated in a similar way which finally leads to the belief $\mathrm{P}(I = \mathrm{true}|D = \mathrm{true}, Pr)$, which is the probability that the information need I is true given the document D and the assigned priors.

One can regard the basic inference network as consisting of two networks where the division goes between the representation nodes and the query nodes. The top network represents the document and consists of the document, priors, language models and representation nodes. This network can be estimated in advance and stored in the document index. The lower network represents the users request and consists of the information need node and the query nodes. This network is created anew every time a query is put to the retrieval system. The two networks are joined by links between the query nodes and the representation nodes (Turtle and Croft 1990).

It is the possibility to represent complex features in the representation nodes and relation features in the belief nodes that allows the user to search using structured queries, which is the strength of the inference network model (Croft, Metzler and Strohman 2010; Lemur nd).

2.5 Searching the web

There are similarities, but also major differences between web search and search in text collections, such as electronic libraries, or the laboratory counterparts, test collections. The similarities are that the user, having an information need, must convey this need to a search engine in a processable query, a query which the search engine matches to information about the document collection stored in an index. The matching results in a ranked list of documents, calculated to be relevant to the user's information need. The differences between web and text search are much about size, structure and user behavior, all issues

that add dimensions to web search not related to text content (Croft, Metzler and Strohman 2010).

The web contains billions of documents and grows exponentially, the web is constantly evolving, the contents changing. Web crawlers identify pages to index. However, before a crawler has covered any significant part of the web, the contents have already changed. The crawlers also do not have access to the whole web or are unable to index the contents for other reasons. Many sites are locked and require passwords, others are created dynamically as a user accesses the site and many do not contain text. It is simply not feasible to index and update more than a fraction of the web.

Not only the size of the document collection distinguishes web search from text document search. The number of queries put to the web systems are tens of millions a day. The search engines must have enough computer power to manage this. The type of queries also differ. Web search is employed by all types of users, and most of them are not professionals in either computer or information science. This fact, and also the number of queries, calls for the queries to be simple in structure. Most queries put to web search systems consist of a short list of words, in contrast to text search systems where the query language can be quite complicated.

A web page contains more information than the text itself, there is also structure, based on HTML markup. Web pages contain meta information which is information about the page, anchor texts and links to and from other pages. A web site consists of a home page with other pages in a hierarchical structure underneath. The structure is partly revealed by the URLs of the pages. The URLs of pages within a site begin with the URL of the home page, followed by a slash and a continuation. This information can be used in web search engines: pages higher in the hierarchy are ranked higher than pages further down. Many search engines list only a certain number of pages from an individual site, even if there are more pages that would come high in the ranking, The motivation is that pages on the same site can be reached from each other, thus it would gain the user more to list additional sites.

Web search engines include structure information in their search algorithms. PageRank, used by the search engine Google, takes into consideration the number of other web pages that link to the page and how important these pages are themselves (Brin and Page 1998; Page et al. 1999; Metzler Jr. 2007; Manning, Raghavan and Schütze 2008). Text in titles and in anchors is usually given more importance than plain text, and text at the top of the page is given more weight than text further down.

User behavior, such as clickthrough data, dwell time and search exit action are utilized in many web search algorithms. Clickthrough data are recordings of which documents users in previous searches have chosen to click on, dwell time is information about how long the user has remained on a certain page,

and search exit action how the user exits the search application, for example if the user goes to another URL, closes the browser window or prints the page (Croft, Metzler and Strohman 2010).

Web search engines have to try to detect and eliminate spam, spam being pages with little, if any, useful content. This can be web spam (information or sites published on the web with the purpose to annoy), link spam (where overuse of links artificially improves the ranking of a certain site), advertisement spam and other unwanted pages. This kind of filtering is not necessary in an organized text collection.

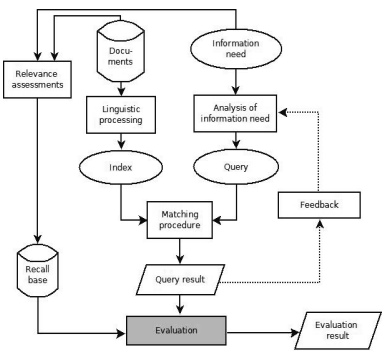
As anyone who likes can publish information on the web, there is a lot of misleading and false information, this is true within the medical domain as in all domains. At the same time more and more people, when feeling ill, consult the web before consulting medical professionals. As this is about peoples' health it is important that they do not get information that can harm them. Martin (2010) addresses the problem of reliability of consumers health documents on the web. The author presents an annotation study where the goal is to, through machine learning, create a system that automatically measures the reliability of web pages within the medical domain.

This is a thesis on the subject of natural language processing and the focus is on search in electronic text collections, focusing on text. Web search and search engine optimization, addressing the issues above, will be left to researchers closer to the areas of computer and information science.

3 EVALUATION

Mostly harmless

The Hitch Hiker's... chapter 6 and the title of the sequel to the trilogy in four parts



Information retrieval is the base for most intellectual activities. Learning and research is about finding out and finding what others have done earlier. Libraries and other information centers have always strived to deliver correct information to whomever requested it. Determining what someone needs, matching that need to what is in the collection, finding those items, delivering them in a timely manner, are all important challenges. Information retrieval as a technology and as a research field has been about innovation, but to a major extent research and development in the field has been powered by evaluation; comparing technologies and algorithms with each other in a systematic manner has provided the field with a non-controversial incremental evolution which has given us what we have today.

However, given this there are still many questions. What to measure? How to measure it? How to measure success when collections are so large that it is practically impossible to know how many relevant documents they contain? There is also the fact that some information needs are difficult to satisfy whatever system is used, or whatever search strategy, while others are easier. All this together makes it difficult to compare runs on different topics or different sets of topics with each other.

The goodness of an information retrieval system can be about effectiveness or about efficiency. Effectiveness is about how well the system manages to retrieve relevant documents and rank them. Effectiveness is ususally given in some kind of precision-recall measure (see section 3.5). Efficiency is about how quickly this is done and at what cost, measured in computer time and

space (van Rijsbergen 1979; Croft, Metzler and Strohman 2010). This thesis will only discuss the aspect of effectivity.

3.1 History

3.1.1 The Cranfield tests

Evaluation of information retrieval systems began with the Cranfield tests in the late 1950s. This was the first so called 'laboratory environment' for information retrieval experiments (Cleverdon 1967).

The first Cranfield test collection, constructed by Cyril Cleverdon and his colleagues at the Cranfield College of Aeronautics, was a collection of some 1 100 documents of research in metallurgy, and was small enough to have every document assessed for relevance to every query. In the first Cranfield experiments the evaluation measures of **precision** ratio and **recall** ratio were established. The approach to use relevance assessed document collections to measure relative search effectiveness is called the Cranfield paradigm, and is still a major influence in information retrieval research. The original Cranfield framework contained the following elements:

- A database containing a, for that time, large set of documents.
- A set of information needs that are expressed in plain text.
- A relevance judgment for every document in relation to every information need.

The second Cranfield project, in the mid-60s, used a test collection of 1 400 documents, mainly in the field of aerodynamics. In these experiments a set of 221 questions were used, and every document was assessed for relevance to each of the questions. Using this collection Cleverdon conducted contrastive experiments, testing single features against a baseline.

The scale of relevance used for the Cranfield tests was from 1 to 4 (see table 3.1). Note that all these levels are relevant to some degree, which means that, in practice, a five-level scale of relevance was used.

Cranfield used the different levels of relevance to study the effect of altering the breaking point when converting his scale to a binary scale of relevance. In different runs the sets of relevant documents had the scores of 1, 1-2, 1-3, and 1-4, respectively. After normalizing the results to the number of documents in each relevant set, he concluded that the relevant set only including documents of score 1 gave the best results.

Table 3.1: The four levels of relevance used in the Cranfield tests (Cleverdon 1967: 174 and 177).

Relevance score	Description
1	References which are a complete answer to the question.
2	References of a high degree of relevance, the lack of which either would have made the research impracticable or would have resulted in a considerable amount of extra work.
3	References which were useful, either as general background to the work or as suggesting methods of tackling certain aspects of the work.
4	References of minimum interest, for example, those that have been included from an historical viewpoint.

The second project was designed mainly to study the effects of using index languages with different features, such as single terms, controlled terms, simple concepts. These had variations such as synonyms, different word forms and hierarchical selections. The result showed clearly that single terms with different variations performed best. (Cleverdon 1967; Harter 1996; Kraaij 2004)

3.1.2 MEDLARS

One of the first evaluations of systems for searching in medical publications was the study of MEDLARS (Medical Literature Analysis and Retrieval System). It consisted of more than 800 000 short abstracts of articles from the medical domain. The articles were indexed manually using the MeSH thesaurus (see section 9.6) (Lancaster 1969).

The size of the MEDLARS collection made complete relevance assessments unrealistic. The solution was a procedure reminding of the pooling process which later would by elaborated by Spärck Jones and van Rijsbergen (1975). Lancaster describes it as gathering a **recall base**. The recall base consisted of relevant documents previously known by the requester, together with documents found by means other than MEDLARS and which were judged relevant by the requester. The assumption was that the recall ratio of the documents in the recall base would approximate the true recall of the whole collection, a **recall estimate**.

One of Lancaster's major contributions is his failure analysis, where he presents reasons why searches may be unsuccessful. He distinguishes between three major types of failure:

> 1. Recall failures due to the fact that the searcher did not cover all reasonable approaches to the retrieval of relevant articles.
> 2. Pure errors involving the use of inappropriate terms or the use of defective search logic.
> 3. Failures due to the levels of specificity and/or exhaustivity adopted in searching strategies. (Lancaster 1969: 131)

Lancaster goes on to comment on the error types and stresses the importance of the third type. He explains exhaustivity and specificity as terms that apply to both the indexing of documents and the formulation of search requests. He refers to exhaustivity in indexing as the extent to which items of subject matter in documents are represented in the index, and in request formulation the number of facets that are represented in the query. Lancaster describes specificity in indexing as the generic level at which a subject matter is indexed, and in searching the generic level that is chosen for the search key of a subject matter.

A more exhaustive index increases the chance to retrieve relevant documents since more aspects of the contents are represented. However, an exhaustive index also increases the risk of retrieving unwanted documents, or noise. A specific index increases the ability to reject non-relevant documents, but instead increases the risk of missing relevant ones.

> In fact, the central problem of searching is the decision as to the most appropriate level of specificity and exhaustivity to adopt for a particular request. The less specific and exhaustive the formulation, the more documents will be retrieved; recall will tend to increase and precision to decrease. The more specific and exhaustive the formulation, the fewer documents will be retrieved; recall will tend to deteriorate and precision to improve. For each particular request, we must decide in which direction to go. (Lancaster 1969: 132)

Lancaster illustrates exhaustivity and specificity with the example shown, somewhat modified, in figure 5. The figure shows two facets of a request on 'oximetry applied to patients with pulmonary emphysema'. An exhaustive query would employ search keys from both facets, a less exhaustive from only one facet. A specific query would have search keys from the lower levels of the hierarchies, a less specific query search keys from higher up in the hierarchies.

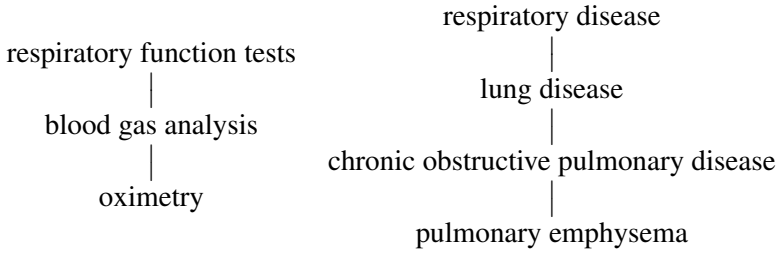


Figure 5: Two facets with terms arranged in a hierarchy of specificity. An exhaustive query would have search keys from both facets. A specific query would have search keys from low down in the hierarchy.

3.1.3 The ideal test collection

As document databases grew, it soon became obvious that complete relevance assessments would not be feasible. There was a limit to the scalability of test collections. At that time there existed some test collections which had been constructed with a specific experiment situation in mind. These collections had come to be reused in other experiment situations, something which was far from ideal. Karen Spärk Jones set out to plan and design the 'ideal' test collection (Robertsson 2008).

In Spärck Jones and van Rijsbergen (1975) the **pooling** method was proposed. With this method the collection is searched with a number of retrieval systems and only the top-ranked documents for each system and topic, the ones most likely to be relevant, will be assessed. For a more thorough description of pooling see section 10.5.

The research group behind the proposition of the ideal test collection never managed to collect enough funding to create their own ideal test collection.

3.1.4 The TREC experiments

In the beginning of the 1990s the need for larger test collections had become obvious. This motivated NIST, the U.S. National Institute of Standards and Technology, to organize TREC, the Text REtrieval Conference,[5] which is an annual event, from 1992 and forward (Voorhees and Harman 2005). Tasks and documents are provided to those who participate, and the participants, in turn, submit the runs of their systems to NIST. In this way, larger test collections have been built than otherwise would have been possible.

[5] <http://trec.nist.gov>

TREC is a conference and competition in information retrieval. The first TREC had two tasks for the participants, ad hoc and routing. This has grown and now each TREC conference is structured into a number of tracks. The tracks differ slightly over the years and represent research areas such as NLP, speech, web, and domain-specific research.

The main goals of the TREC conferences are to encourage IR research using large test collections, to increase communication between people who work with IR, and to research evaluation methods. The research of evaluation methods came about when it became obvious that there was a need to standardize the methods. Researchers had, up till then, used such varying methods that it was difficult to compare results.

As the TREC collection was much larger than earlier collections it was practically impossible to assess every document for relevance to all queries. Instead the pooling method became standard use. The TREC organizers construct the pools by letting participants contribute with lists of top ranked documents retrieved in runs on their individual systems. These lists are merged and duplicates removed. The TREC framework is based on the Cranfield framework, but with two major changes:

- The information needs are constructions with three distinct parts: title, description and narrative.
- The relevance assessments are not complete. The documents that are assessed are selected by the method of pooling.

The TREC collection is large and used by many researchers. To facilitate collaboration and comparison, a standard way of marking up the documents was developed, the trectext format. All documents are tagged with the XML tags. The <DOC> tags are used to delimit the individual documents, the <DOCNO> tags are used for the document identifier, and the <TEXT> tags for the textual content. These are common to all documents.

3.1.5 CLEF

The Cross Language Evaluation Forum, also known as CLEF,[6] is an evaluation series that is focused on cross-language information retrieval in European languages. The objective is to promote research of multilingual systems, including both cross-lingual systems, where the query is put in one language and

[6] <www.clef-campaign.org>

the retrieved documents are written in another language, and monolingual systems that can operate on several languages. CLEF organizes annual evaluation campaigns with a number of different tracks.

3.2 Relevance

The concept of relevance has been debated for more than four decades. The standard approach in laboratory based information retrieval research is that a document is considered relevant if it has a topic similarity to an information need and thus is desirable in the result set of a search. The relevance of a document is assumed to be independent of that of other documents.

There are several ways a document topically can meet an information need. Sormunen et al. (2001) list four features that relate to the degree of relevance of a document to an information need:

- The topic is discussed at length in the document.
- The document deals with several aspects of the topic.
- The document contains many words that relate to the topic.
- Several different expressions are used to describe the topic in order to avoid repetition.

A marginally relevant document may mention the topic in question only in few words, without much variation and maybe discusses the topic from one view point only.

The topical view of relevance is constantly challenged. The objections are mainly that the question if a document is desirable is not only dependent on the information need and the document, but also on who has expressed the need, for what purpose, in what situation, and at what time. A user is not necessarily satisfied with a document even if it is topically relevant. The user may already know about the document, the document may be outdated, or the contents can be too shallow or too complicated, or even simply not true.

However, there is no need to make an absolute choice between the different aspects of relevance, but rather to use them in different situations. The usual division made is between topical relevance and user relevance. Saracevic (1999) goes further and distinguishes between a number of types of relevance, among them: system or algorithmic relevance, topical or subject relevance, cognitive relevance or pertinence, situational relevance or utility, and motivational or affective relevance. Information retrieval systems consider the system or algorithmic relevance, since that is what they can measure, while a user can be interested in any kind of relevance. Saracevic 1999 adresses this issue.

> Difficulties arise when an object is of system relevance but not of cognitive relevance or utility, or conversely. If items are of cognitive relevance or utility, but were not reflected in the query, they are not and cannot be retrieved. At the bottom of IR research is a quest to align systems with other types or relevance. (Saracevic 1999: 1059)

There are several reasons why topical relevance is so widely used among information retrieval researchers. It is supposedly the most objective approach and an approach that makes it possible to reproduce experiments. It is relatively easy to measure and quantify with, for example, precision and recall methods. Not least important, it makes it possible to carry through large scale experiments with automated strategies, as long as there is a test collection with documents preassessed for relevance to a number of information needs.

It is, of course, important to study other aspects of relevance, not least user relevance. In the end, no one can dispute the importance of having satisfied users. For the types of relevance that are not topical or algorithmic, human users are important, and research is mostly done with user studies, not in a laboratory environment. There will be attempts to approach user relevance (without user studies) in later chapters of this thesis, where documents with different target audiences will be studied, as well as user scenarios in the test collection.

3.3 Reliability

A widely heard objection to relevance judgments is that they are subjective. They differ between judges and they differ for the same judge over time. This is a difficulty that is hard to get around. The purpose with information retrieval is to retrieve documents that a human user percieves as answering to his or her information need. As long as there is no algorithm that describes human reasoning satisfactorily, using human judges is the best we can do. However, in order to secure reliable results in information retrieval it is vital not to look at the absolute numbers of the results. Comparing effectiveness between systems or methods is done looking at relative differences.

If one would ask a group of assessors to rank all documents that they assess according to relevance into one ranked list for each topic, the sequential order of documents would not to differ much between assessors. However, the assessors may differ in where to draw the line between the degrees of relevance.[7] It is therefore important that the individual topics are not divided between assessors. In Voorhees 1998 the author stresses that it is not the measures themselves that are interesting, but the relative difference between systems and strategies,

[7] This is discussed in more detail in section 10.5.2.

which tend to be stable even when the absolute assessments differ. In other words, relevance assessments may vary, but the results based on them are sufficiently stable. Voorhees drew these conclusions after testing several different factors such as judgments made by authors versus non-authors, judgments made by single judges versus groups of judges and judgments made by judges from different organizations. She found differences in the absolute values of performance, but not significant difference in relative performance.

Harter (1996) has examined literature describing empirical studies on how variations in relevance judgments affect the measures of retrieval effectiveness. He states the following:

> All find significant variations in relevance assessments among judges. And all conclude that these variations have no appreciable effect on measures of retrieval effectiveness, that is, in the comparative ranking of different systems (indexing languages, document representations, etc.). (Harter 1996: 44)

3.4 Completeness

It is important to have a sufficiently complete judging. A more complete judgment gives a better model of recall performance. Harman 2005 describes completeness tests done on the TREC collection. These tests showed that there was no correlation between the number of documents that were judged, that is the depth of the pools, and the number of newly discovered relevant documents that had been missed in the original assessment pools. On the other hand, there was a relationship between the original number of relevant documents and the number of new relevant documents found. Put differently, topics with many relevant documents were more likely to have undiscovered relevant documents than topics with few relevant documents. Harman states:

> So having the exact number of relevant documents, or having an exact recall number, is not as crucial as knowing that the judgments are complete enough to ensure that comparisons of two methods using the test collections will be accurate. (Voorhees and Harman 2005: 43)

In most studies unjudged documents have been assumed to be non-relevant. This is questioned by several researchers, for example Zobel (1998), when commenting TREC results. Even though, in his opinion, the results are reasonably reliable and the relevance judgments provide a fair basis for measurement, he sees unjudged documents as a problem.

> . . . it is likely that at best 50%–70% of the relevant documents have been discovered, in particular because of the queries that have large numbers

> of answers; and we show that the measurement strategy of assuming unjudged documents to be non-relevant is questionable. (Zobel 1998: 307)

An approach to solve this problem is suggested by Eero Sormunen (p.c.). This is to judge unjudged documents as they are retrieved in new runs. This guarantees that all documents in the result sets are assessed. If it is not possible to employ the same assessor as for the previous judgments measures will have to be taken to see that the new assessments are consistent with the old ones.

3.5 Effectiveness

The output from an information retrieval search engine is a (ranked) list of documents. This list is compared to the recall base in order to obtain the relevance grades of the documents. The result is a (ranked) list of relevance values which needs to be translated to a small number of representative values or to a visual representation in order for the user to compare the results of different runs.

The standard way of measuring the effectiveness of information retrieval is by precision and recall expressed in a variety of different measures depending on what one wants to put focus on. The result can be presented in a table, in one or several numbers, or visualized in graphics. The concepts of precision and recall are described below, together with a number of commonly used precision and recall measures.

Precision A measure of the ratio between the number of relevant documents retrieved and the total number of documents retrieved. It is a measure of how much of what is retrieved that is relevant.

$$precision = \frac{|Ra|}{|A|} \tag{7}$$

where
$|Ra|$ = The number of relevant documents in the answer set (the retrieved documents)
$|A|$ = The number of documents in the answer set

Recall A measure of the ratio between the number of relevant documents retrieved and the total number of relevant documents. It is a measure of

how exhaustive the search is, that is, how much of what is relevant that is retrieved.

$$recall = \frac{|Ra|}{|R|} \tag{8}$$

where
$|Ra|$ = The number of relevant documents in the answer set
$|R|$ = The number of relevant documents in the recall base

Calculating absolute recall is not possible if the total number of relevant documents is not known. Instead recall can be calculated against the recall base, that is the number of known relevant documents.

Precision-recall curve For an easy-to-grasp visualization of the precision-recall result of a run, a precision-recall curve can be produced. In such a graph, as in figure 6, precision is plotted as a function of recall. The precision can be plotted for every position in the retrieved list down to a chosen rank. However, this makes it difficult to compare results for different topics as topics don't have a uniform number of relevant documents. Another approach is to compute interpolated precision values at fixed points of recall. This can be specified as the maximum of the precision at recall points greater than or equal to the recall value in question, as in formula 9.

$$pr(i) = max(pr(j)) \qquad where \quad j \geq i \tag{9}$$

Average precision A measure obtained by measuring precision at every ranking point where a relevant document is retrieved. The average of these precision measures is then calculated. A relevant document that is not retrieved receives precision zero. The average precision can be illustrated geometrically as the area underneath a non-interpolated precision-recall curve.

Mean average precision This is the mean of the average precision over all queries in one run or experiment situation.

11-point precision A measure obtained by measuring precision at recall levels of 0, 10, 20, ..., 100%. To get a single measure, the average of these measures is calculated for the 11pt precision. If desired, one can choose a different number of points, for instance 3pt precision.

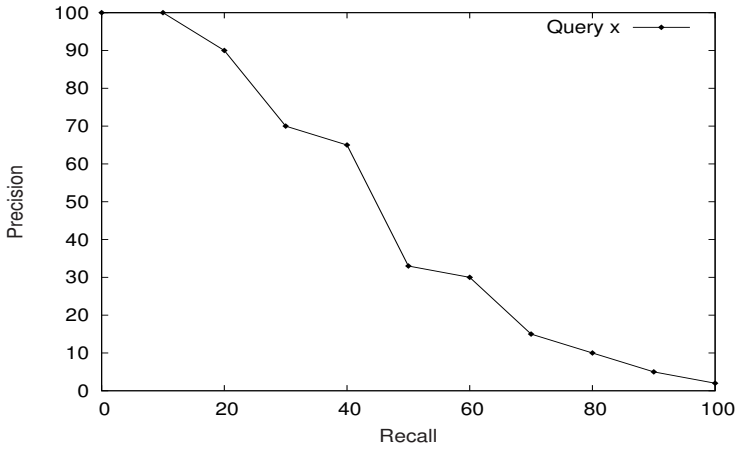


Figure 6: An example of an interpolated 11-point precision-recall curve for the fictive query x.

R-precision A measure obtained by measuring the precision after retrieval of R documents. R is here the number of known relevant documents for the topic in question.

bpref This measure is a function of the number of times judged non-relevant documents are retrieved before relevant documents. This measure addresses the issue of traditional pooling, where the system makes no difference between documents that have been assessed as non-relevant, and documents that have not been assessed at all and are assumed not to be relevant. With bpref only documents which have been assessed are taken into account (Buckley and Voorhees 2004).

$$\text{bpref} = \frac{1}{|R|}\sum_{r}(1 - \frac{|n \text{ ranked higher than } r|}{|R|}) \tag{10}$$

where

$|R|$ = The number of relevant documents in the recall base
r = A relevant document
n = A member of the first R judged non-relevant documents as retrieved by the system

Cumulated gain, CG A measure which makes use of relevance on a multi-level scale. The cumulated gain is calculated for each ranked position i by summing all relevance scores from 1 to i. In formula 11, $\mathrm{CG}(i)$ is the cumulated gain in position i of the ranked list, and $\mathrm{G}(i)$ is the degree of relevance for the document in that position.

$$\mathrm{CG}(i) = \mathrm{CG}(i-1) + \mathrm{G}(i) \tag{11}$$

A gain vector is shown in example 12. It represents the relevance grades of the retrieved documents. The numbers represent the relevance of the retrieved document in each position of ranking, with the document ranked to be most relevant furthest to the left.

$$\mathrm{G}' = \langle 3,2,3,0,0,1,2,2,3,0... \rangle \tag{12}$$

In the cumulated gain vector in example 13 the relevance values of the gain vector in example 12 are added and cumulated along the vector (Järvelin and Kekäläinen 2002).

$$\mathrm{CG}' = \langle 3,5,8,8,8,9,11,13,16,16... \rangle \tag{13}$$

Discounted cumulated gain, DCG This measure is the same as the measure for cumulated gain, with the addition that it uses a discounting factor that reduces the amount of the score added, for each step in the ranked list. The relevance score of a document is discounted by a function of its position in the gain vector. This function is a logarithm of the position number. The assumption is that the further down a document is found in the ranked list, the less it is worth to the user, and accordingly, a smaller portion of the relevance value is added to the value in the discounted cumulated gain vector. In formula 14, $\mathrm{DCG}(i)$ is the discounted cumulated gain in position i. $\mathrm{DCG}(1)$ is set to $\mathrm{G}(1)$ when $i = 1$ to avoid division by zero.

$$\mathrm{DCG}(i) = \begin{cases} \mathrm{G}(1) & \text{if} \quad i = 1 \\ \mathrm{DCG}(i-1) + \frac{\mathrm{G}(i)}{\log(i)} & \text{otherwise} \end{cases} \tag{14}$$

In example 15 we can see the relevance values from example 12 in a discounted cumulated gain vector. As we go along the list representing the ranked documents, for each position a smaller share of the documents' relevance value is added to the discounted cumulated gain.

$$\mathrm{DCG}' = \langle 3,5,6.89,6.89,6.89,7.28,7.99,8.66,9.61,9.61... \rangle \tag{15}$$

Normalized discounted cumulated gain, nDCG A measure which takes into consideration that the discounted cumulated gain described above is biased towards information needs with many relevant documents. The normalized cumulated gain relates the retrieved result to the ideal result, that is the maximum discounted cumulated gain for every position. With a scale of relevance of $0-3$ this would be achieved if all documents with relevance 3 were ranked first, followed by all documents given relevance 2, in turn followed by documents given relevance 1. The normalized discounted cumulated gain makes it possible to see, for every retrieved document, how well the system has performed compared to what would be maximally possible.

In example 16 the relevance values from example 12 are shown in a normalized discounted cumulated gain vector.[8] Since normalized discounted cumulated gain is a quotient between two gain values, the value does not necessarily increase as the vector is followed but can fluctuate up and down. It is always between 0 and 1.

$$\text{nDCG}' = \langle 1, 0.83, 0.89, 0.73, 0.6, 0.76, 0.89, 0.84... \rangle \tag{16}$$

The cumulated gain measures are based on the following two assumptions:

- Highly relevant documents are more valuable than marginally relevant documents.

- The later in the ranking a relevant document appears, the less valuable it is for the user as it is less likely that the user will examine the contents.

[8] In this example is it assumed that all relevant documents are among the 10 highest ranked which are shown in the vector.

4 LINGUISTICS AND IR

I am the main Dish of the Day. May I interest you in parts of my body?

The Restaurant... chapter 17

4.1 Language...

Most research in information retrieval has been in the field of information science or computer science. However, IR is about storing and finding information, mainly in the form of text, so there is an obvious connection to linguistics. As this thesis focuses on linguistic aspects of information retrieval, a description of linguistic issues is in place. Below are definitions of the central linguistic terms of this thesis.

Word The word 'word' itself can be used in several different senses. When it is important to distinguish between these senses, the terms described below will instead be used.

Lemma sign A linguistic sign which consists of a group of word forms that belong to the same part of speech, and that can be assigned to the same inflection series, or to more than one series if they converge and where the divergences show only facultative variation. In cases where the word is undeclineable, the inflection series consists of only one form (Svensén 2009; Allén 1970).[9]

Lemma form A certain form in the paradigm of a lemma sign.

[9]There are different schools as to which is the definition of the term *lemma* and of related terms. The question is if the term *lemma* denotes one word form or a group of word forms (the paradigm). I have here chosen the denotation *lemma sign*, from Svensén 2009 and the definition from Allén 1970. However, Allén, has used this definition for the term *lemma*. I have made the choice of defining *lemma* as one word form, and to use the term *lemma sign* for the group of words, in order to conform to the tradition of English speaking academia.

Lemma A lemma is the lemma form which is chosen to represent the lemma sign. It is usually the unmarked form of the lemma sign. In Swedish, the lemma form chosen as lemma is usually the following: the singular, nominative, indefinite form for nouns; the infinitive form for verbs; the singular, non-neuter, indefinite, positive form for adjectives.

Lexeme A lexeme consists of two parts: a group of closely related meanings (differing, for instance, in number or in tense) and a corresponding group of word forms, a lemma sign. These together form an abstract lexical unit. What word form is used to realize a certain occurrence of a lexeme, for instance in a text, depends on the syntactic context, and what aspect of the lexeme it is meant to denote, for example what number for nouns and what tense for verbs (Järborg 2003).

Word form A certain form of a certain lexeme.

Morpheme A morpheme is a minimal unit of meaning or of grammatical function. One word form can consist of one or several morphemes.

Stem A stem, in the linguistic sense, is the base of a lexeme, the form to which derivational and inflectional affixes are added. The stem is the form which is used as the modifier constituent in compounds.

A stem, in the information retrieval sense, is a substring that is common to all or several of the word forms of a lexeme. This stem is what is left when spurious affixes are removed in the stemming process. The purpose of such a stem is to serve as a representative of the lexeme.

Scientific term A lexical unit which has a stable and specialized meaning, representing a concept in a particular domain. A scientific term has a fixed meaning, in contrast to terms in general language where the meaning often is dependent on the context. A scientific term can be a simplex word, a compound or a lexicalized phrase. If a term is a lexicalized phrase, the individual words can usually not be rearranged or replaced by synonyms: for example, *Cesarean section* cannot be replaced by **Cesarean cut*.

Term In this thesis the scientific terms 'index term', 'query term' (for definitions see section 2.1 on page 12), and 'scientific term' (see above) will be used. If another sense of 'term' is intended, 'medical term', 'linguistic term' and so on will be used. When it is obvious (or not essential) which of these terms is intended, an unspecified 'term' can be used.

String A sequence of characters.

Text A sequence of words which stand in relation to each other and which convey some meaning. The order of the words is important.

Bag of words A collection of words where the order of the words is not considered, only the words themselves. A bag of words is like a set of words, with the exception that there is no restriction on the number of times that an entity can appear.

4.1.1 Morphology

Morphology concerns the internal structure of words, or word formation. A predictable process of word formation is **inflection**: the inflectional form is based on the syntax of the context where the lexeme occurs. In Swedish, the most common inflectional processes are performed by adding grammatical morphemes to a stem (or removing them), changing the stem vowel or both. In the languages of the world there are inflectional processes such as: change of vowels and of consonants, intonation, tone, reduplication, and others. Word formation can also be of the less predictable processes: **derivation** or **compounding**. In derivation the usual process is to add content morphemes (which are not lexemes) to a stem forming a new lexeme. New lexemes can also be derived in other ways, for instance by changing the intonation as in *import* (V)→*import* (N). In compounding two lexemes are combined into one new lexeme (Jensen 1990).

Knowledge about morphology can be of great use in information retrieval during the process of indexing and in query construction. Knowing which words have common semantic features can guide the process of conflation (see section 4.2.1) when indexing. When constructing queries the search terms must go through the same conflation process as the index terms. Alternatively, if conflation is not performed, knowledge of morphology can help determine which word forms to put within the synonym operator, **#syn()** (see figure 17 on page 96). In both cases it is about determining which word forms could be considered related enough to belong to the same concept or facet.

4.1.2 Inflection and derivation

Swedish nouns, adjectives and verbs are often inflected by adding suffixes to the stems. The inflectional suffixes add information without changing the lexeme. In Swedish, the inflectional categories of nouns are number, definiteness and case; the categories of adjectives are comparison, gender, number and def-

initeness; and the categories of verbs are tense, person, voice, mood and finiteness.

Inflectional suffixes are not used independently, they do not change the part of speech of the word they are attached to, and they do not form new lexemes. Table 4.1 shows non-neuter and neuter examples of Swedish inflected nouns and adjectives.

Table 4.1: An example of Swedish inflection of nouns and adjectives in the non-neuter and neuter genders, and their English equivalents. In addition to each of the noun forms in the table there can also be a suffix 's' indicating the genitive case. Thus, each Swedish noun has eight inflectional forms.

Indefinite singular	Definite singular	Indefinite plural	Definite plural
en bruten arm	den brutna armen	brutna armar	de brutna armarna
a broken arm	*the broken arm*	*broken arms*	*the broken arms*
ett brutet finger	det brutna fingret	brutna fingrar	de brutna fingrarna
a broken finger	*the broken finger*	*broken fingers*	*the broken fingers*

Derivations are, like compounds, lexemes formed from other lexemes. While a compound is formed from two lexemes, a derivation consists of one lexeme, the stem, and one or more derivational affixes. An affix can be a prefix placed before the stem, or a suffix placed after the stem. Unlike the bases of a compound, and the stem of a derivation, an affix cannot function as an independent word. There is not an absolute distinction between inflection and derivation, but some differences can be listed. These are summarized in table 4.2 (Källström 1999).

Table 4.2: A comparison between derivation and inflection.

	Inflection	Derivation
New lexeme	No	Yes
New part of speech	No	Often
General rules	Yes	Many exceptions
Change of sense	Modification	Significant
Semantic effect	Predictable	Not predictable

Inflection does not change a word into a new lexeme, as derivation does. As the lexeme is still the same, the part of speech remains unchanged. There is no significant change of meaning in inflection, rather a modification, such

as a change of number for nouns, a change of time for verbs, or a comparison change for adjectives.

Adding a derivational affix to a lexeme changes it into another lexeme. This new lexeme may or may not have the same part of speech as the stem, as in *grön+aktig* 'green+ish' (Adj→Adj) or as in *smuts+ig* 'dirt+y' (N→Adj), respectively. The meaning of the new lexeme derives, with some exceptions, from the meaning of the stem.

For inflection there are general rules to follow and general patterns, or paradigms. For derivation, though there are patterns and rules, for example, adding the suffix *-able* to a verb 'V' to produce an adjective with the sense 'able to V' (*changeable*, *noticeable*), there are many borderline cases and exceptions to when the patterns are applicable. Could you say ?*vomitable*? Certainly not **dieable*.

The semantic effect of inflection is essentially predictable. The semantic effect of derivation is not. The Swedish equivalent of the English prefix *un-* is *o-*. This prefix usually has the sense of 'not', as in *ofarlig* 'undangerous' and *okrossbar* 'uncrushable'. But not always. *o-* can take on a negative sense in other ways, such as in *oväder* 'unweather' (bad weather/storm) or *odjur* 'unanimal' (beast/monster).

Bybee (1985) sees the distinction between inflection and derivation as gradual, as a cline, rather than discrete. She explains the degree of semantic change as deciding where on the cline a word belongs. The greater the semantic difference between a stem alone and the stem plus an affix, the greater the likelihood that the affix is derivational and not inflectional. Bybee describes the degree of generality as also being important. The typical inflectional morpheme can combine with all lexical items of the appropriate class and with the appropriate semantic features. This generality puts constraints on the contents of inflectional affixes, which are less specific.

Bybee not only puts inflection and derivation along a cline, but presents a whole scale of ways to combine semantic elements into expression units, as seen in figure 7. At one end is expression by one lexical unit, for example *kill*, which is a combination of 'cause' and 'die', and *wade*, a combination of 'walk' and 'in water'. In order for two semantic elements to be combined into one lexical unit, they must be very relevant to each other. For instance, it is not probable that there would be a single lexical item for *walking in sunshine* (at least not in languages in countries with colder climate), as the sunshine does not affect the walking as much as the water does. At the other end of the cline Bybee puts syntactic expressions. Here the semantic units are expressed in separate and independent words, for example *come to know*, which is the syntactic expression of 'inchoative' and 'know'. Between syntactic and inflectional Bybee places the free grammatical expressions. These include morphemes, such

as clitics, particles and auxiliaries, which belong to closed classes and which occur in fixed positions, but are not bound to any lexical item.

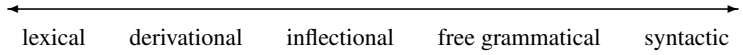


Figure 7: The various manners that semantic units combine into expression units can be ordered along a cline, from lexical expressions, the most highly fused, to syntactic expressions, the most loosely fused means of expression.

4.1.3 Word form frequencies – an example

In Friberg [Heppin] 2006 the present author describes a study about the choice of which inflectional forms that are used as lemma for different types of entries in different types of word lists, such as dictionaries, encyclopedic collections, thesauri and book indexes.[10]

Even though, for nouns, the word form of the entries is usually singular, indefinite, nominative, this is not always the case. The author listed lemmas containing the word *sjukdom* 'disease' in different inflectional forms in different word lists and compared them to the number of instances of the different inflectional forms of the lemma signs in the MedLex corpus.[11] The inflectional forms of the lemma sign *sjukdom* that were used are shown in the head of table 4.3. The rows that follow in the table contain frequencies in the MedLex corpus of the inflectional forms of different types of lemma signs. The types referred to are the following:

sjukdom The simplex word *sjukdom*.

modifier||*sjukdom* Compounds which are composed of a modifier and the head *sjukdom*, for example *bristsjukdom* 'deficiency disease'.

Proper name+GEN *sjukdom* Names of diseases which are constructed using a proper name in the genitive form, followed by *sjukdom*, for example *Alzheimers sjukdom* 'Alzheimer's disease'.

attribute *sjukdom* The word *sjukdom* preceded by an adjective phrase, for example *autoimmun sjukdom* 'autoimmune disease'.

[10] Recall that, in Swedish, the inflectional forms are constructed with suffixes (see section 4.1.2).

[11] As this study was performed in 2006, an earlier version of MedLex was used than the one used in MedEval.

Table 4.3: A comparison of inflectional word form frequencies of different types of phrases containing the word *sjukdom* 'disease'. The phrase types are: simplex word, compound with the head *sjukdom*, named entity of the form [Proper name+GEN *sjukdom*], and noun phrase of the form [attribute *sjukdom*].

	Indefinite singular *sjukdom*	Definite singular *sjukdomen*	Indefinite plural *sjukdomar*	Definite plural *sjukdomarna*
[*sjukdom*∗]	1 685 (42%)	1 057 (27%)	1 175 (29%)	66 (2%)
[modifier‖*sjukdom*∗	1 206 (50%)	178 (7%)	994 (41%)	42 (2%)
[Proper name+GEN *sjukdom*∗]	445 (99.6%)	0 (0%)	2 (0.4%)	0 (0%)
[attribute *sjukdom*∗]	656 (48%)	59 (4%)	609 (45%)	34 (3%)

There are two numbers in table 4.3 that stand out from the others. One is the simplex word in the singular, definite form. This pattern is the only one that has a substantial number of occurrences in this inflectional form. The reason given in Friberg [Heppin] 2006 is that the singular definite form is often used to refer back to an occurrence in indefinite form of any of the types. That is, whatever pattern was used the first time a certain disease was mentioned, the word later used to refer back to this occurrence was the simplex word in definite form. The other number to comment on is the singular indefinite form of the pattern [Proper name+GEN *sjukdom*], which is almost 100%. This pattern is used only for names of diseases. Names are seldom given in definite form in Swedish. These names of diseases seem to follow this pattern and stand in the indefinite form. Further, these phrases already have a determiner in the attribute, [Proper name+GEN], which does not allow a definite form to follow.

The conclusions concerning the form of entry terms in Friberg [Heppin] 2006 are that, even if the singular, indefinite, nominative form is most common as the choice of entry or lemma form, there are three things that seem to influence the decision: (1) If the word is a name lexicalized in definite form, the entry is also in definite form. In other cases the indefinite form is used. (2) If the word is a hyperonym which takes a list of other words as hyponyms, it may be in the plural form. (3) The type of work that the word is listed in influences the choice of word form for the entry.

A lexicon or dictionary seldom chooses anything but the most basic form of a lemma sign. The MedLex dictionary (Kokkinakis 2004) strictly uses this form for entries containing the word *sjukdom*.

Lindskog 2004 varies the form of the entries. This is a dictionary which is more of an encyclopedic type than lexicologic. Instead of giving information about the different word forms, it gives information about the meaning and content of the entries. The form of the entries, in most cases, mirrors the use of the word and its relations to other words.

An information thesaurus, such as the Swedish MeSH, is used for classification and uses tree structures to show relations between words. In cases where the leaves of the structure describe something uncountable, such as the name of a disease, they are in the singular form. In other cases, and higher up in the structure, the entry forms are in the plural. This is in a sense logical considering that the leaves only represent themselves but the other entries are hyperonyms with subordinate entries.

In the index of Gillå 2005, which is a course book used in nursing education, the form used as index entry, in the back of the book, mirrors how the words are used inside the book. The plural form is used when the lemma sign in question is a hyperonym and has several other words listed as hyponyms, which is consistent with the hyperonyms in the Swedish MeSH. Otherwise the singular form is used for lemma signs that have a general describing function.

4.1.4 Compounds

At first it may seem simple enough to define a compound: a semantic unit containing at least two lexemes which function as one new lexeme. But how does one define the lexemes that build up a compound, and what is meant by 'function as one lexeme'? There have been numerous attempts to find a definition that holds, and that is sufficient to discriminate compounds from derivations at one end and lexicalized phrases at the other. The attempts have been done both universally and for individual languages or language groups.

A solution supported by Lieber and Štekauer (2009), is to describe compounds, not as a well defined category, but as a cline. There are compounds that are comparatively easy to categorize, but there are others, rather fuzzy compounds, which could be argued to be defined as, for example, derivations or phrases. Instead of a definite categorization, one could make a list of criteria and define a compound as a unit which can be described with some amount of these criteria. Languages differ, so there could be one basic list of criteria which should apply universally, and, for each language or language group, the list could be completed with additional criteria. It is important to keep in mind that such a list will not be absolute. The criteria describe the compound prototype. Even within a specific language, there will not be any list of criteria that describes all compounds, and all compounds will not be described by any list.

A list of criteria for compounds could look like the one below.

- Two lexemes, or more, that function as one.
- The meaning is a single concept which is more precise than the sum of the meaning of the parts.
- A compound consists of head and modifier or of coordinated parts.
- An adjective or adverb will affect the compound as a whole, not only one constituent.
- It is not possible to insert another element between the compound constituents as they are syntactically inseparable.
- Compounds have specific stress patterns in the pronunciation.

A list for Swedish compounds could be completed with the additional criteria below.

- Productive.
- Right headed (or coordinated).
- Inflected as a whole, the inflectional morphemes are attached to the head, not the modifier.
- Written as one orthographic word.
- Can contain link elements.
- Primary stress on the first constituent, secondary stress on the syllable in the second constituent which would have the stress if that constituent was not part of a compound.

As stated above, none of the listed criteria is always true and the list is not sufficient to make a definite distinction between all compounds and non-compounds. Examples of exceptions to the criteria that Swedish compounds are written as one word are compounds where one constituent is a phrasal lexeme, *New York-börs* 'New York Stock Exchange' or an elliptic coordination, *hjärt- och kärlsjukdom* 'heart and vascular disease'. The latter example is also an example of an exception to the criterion that it is not possible to insert any element between the compound parts.

4.1.4.1 Two lexemes, or more, that function as one

A compound is said to be composed of two or more lexemes. However, it can be difficult to determine what should be counted as a lexeme, both regarding compound constituents and complete compounds, for instance when one compound constituent is void of meaning, while the meaning of the other constituent is obvious. In Swedish, the word for 'cherry' is *körs||bär*[12] 'körs berry'. The cherry is a berry, but there is no meaning for *körs*. Could *körs* still be considered a lexeme?

In addition, there is the case of loans. One English example of this is a loan from Swedish, i.e. *ombudsman*. In Swedish an 'ombud' is a representative, a person who speaks or acts on behalf of another person or group of persons, an *ombud+s||man* is a person who is employed to do this. In Swedish the string is a compound, but should it be categorized as a compound in English? Is *ombud* an English lexeme?

Related to the case of loan words is the case of neoclassical compounds. Neoclassical compounds are based on Greek and Latin roots. These are especially common in the language of medicine where terminology based on Greek and Latin has become a sort of lingua franca. The strings *cardiovascular* and *cardiogram* may be transparent for many English speaking persons, but could we call the elements English lexemes? What about the parts of *helicopter*? The number of English speaking persons who are aware that *pteron* is Greek for 'feather' or 'wing' is probably not overwhelming.

There is further a number of common affixes which have a clear meaning, but which hardly can be said to be lexemes. Many of these are neoclassical, but several have other origins. Some examples are *co-*, *mid-*, *bi-*, *-ment* and *-hood*.

Many words in the lexicon today were at one time compounds, but the constituents have with time more or less lost their content. This is an ongoing process and there are examples of morpheme combinations in all stages, beginning with one constituent losing its status as an independent lexeme and instead becoming a derivational suffix and possibly in the end becoming part of one morpheme. An example of a lexeme that is changing its status is 'man'. The basic meaning is still 'male person', but used in compounds the element is losing the sense of 'male' and can now be used meaning 'person' as in the example of *ombudsman* above.[13] For information retrieval this has a negative effect on the results when using 'man' as a search term.[14]

[12]The symbol '||' is used to mark the segmentation points of Swedish compounds and the '+' symbol to mark morpheme boundaries.

[13]Instead of using *man* 'man' or *kvinna* 'woman' in compounds *person* 'person' is often used to avoid including the aspect of gender.

[14]The term *man* is not an effective search key for other reasons. It is not specific. It is also a homograph to the indefinite pronoun meaning 'one/you'.

In his classic work on Swedish words, Teleman (1972) describes adjectival compounds as having a strong tendency of simplifying and changing the meaning of common constituents, either first or last, so that they resemble affixes than lexemes, both when it comes to sense and to distribution. He gives examples of first constituents of adjectival compounds that, even though they have varying meanings when used as free lexemes, are reduced to a sense of reinforcement when used in compounds: *blixt* 'lightning/flash', *jätte* 'giant', *sten* 'stone/rock', *toppen* 'the top', and *hel* 'whole'. An example of a constituent used in the same way, but in the opposite sense in the meaning of reduction, is *botten* 'bottom'. Teleman also gives examples of reduction of meaning in last constituents: *lös* 'loose' (lacking something desirable),[15] *fri* 'free' (lacking something undesirable), *full* 'full' (presence of something abstract), *fattig* 'poor' (containing very little of something), and *säker* 'safe' (protection for/protection against).

Blåberg (1988) describes this group of morphemes as being on the borderline between compounding and derivation. For instance, *trend+lös*, could be analyzed as 'trend loose' (loose as the trend suggests), even if the natural interpretation would be 'trendless' (lacking (desirable) trends). In his material he finds only one instance that would be ambiguous in pragmatic terms: *skandal‖rik* 'scandal rich' (rich in scandals) or *skandal+rik* 'scandalously rich'.

Dura (1998) calls morphemes such as these 'word-like affixes'. She describes them as originally free items that have become specialized in affixal use, and that are no longer free. This implies derivation and not compounding. The adjective *fri* 'free' is a free item, and can be used as a constituent in compounds such as *fri‖tid* 'free time', but in *svavel+fri* 'free of sulfur' *fri* would not be a free item, but a suffix which is limited to the meaning of 'not containing'. Another example from Dura is *blixt* 'lightning, flash'. The compound *blixt‖lampa* 'flash lamp' (flash light), is constructed of the free items *blixt* and *lampa*. However, *blixt* can also be used as a prefix. In this case it is a general magnifier that stresses a high tempo, as in: *blixt+snabb* 'lightning quick' (quick as lightning) and in *blixt+visit* 'lightning visit' (flying visit).

Researchers taking different stands on whether the above described words should be classified as compounds or as derivations, describing the morphemes in question as affix-like words or as word-like affixes can be seen as an illustration of the floating definition of compounds and other word formations. More important than to determine which definition is the correct one, is to state which definition is used in a certain situation. For practical information retrieval, the distinction itself is not very important. What is important is to realize that the morphemes which are the objects of this problematic classifi-

[15] This assessment of value is not always true, for example in *sladd‖lös* (wireless).

cation, if used as search keys, would probably not improve any results, seeing that they are so void of content.

Bybee (1985) does not make definite distinctions between manners of combining morphemes into lexical units. Just as she did describing the fusion of semantic elements (see figure 7), she claims that morpheme combination manners are on a continuum in which she includes compounding, incorporation, derivation and inflection. This is illustrated in figure 8 where lexicalized phrases have been added as having the most open class of items and the most specific meanings. Incorporation is the fusion of a verb's nominal patient with the verb itself. Incorporation is not common in the natural languages of the world, but occurs, for instance, in some native American and Australian languages. Transferring this phenomenon to English would yield sentences like: *I like to cakebake*. While compounding does not have many lexical restrictions, the nouns and verbs that can take part in incorporation are usually restricted. For instance in the native American language Pawnee, the nouns involved must refer to body parts, natural phenomena, food, or cultural products.

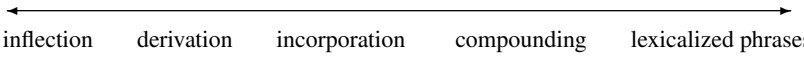


Figure 8: The manners possible when combining morphemes into lexical units can be ordered along a cline, where inflection is the most constrained, involving the smallest classes of items with the most abstract and general meanings while lexicalized phrases and compounding involve open classes of items and fuller and more specific meanings.

4.1.4.2 Separable and inseparable compound verbs

Many Swedish compound verbs are constructed by a verb prefixed by a particle. Often this particle is an inseparable part of the word, but for others the particle and the verb can be found in both a separated and an integral form Holmes and Hinchliffe (2003). The verbs that can be found in both forms belong to two different groups: one that is always separated except in the present participle and the past participle form, and one where there is a semantic or a stylistic difference between the forms. However, the verbs of the second group are always compounded in their participial forms, whatever the meaning. Table 4.4 shows an example of a compound verb with a semantic difference between the separable form, which has a literal meaning, and the inseparable form, which has a more figurative meaning.

Table 4.4: An example of a Swedish compound verb where there is a semantic difference between the separable and inseparable forms: *bryta av* (break off) and *avbryta* (interrupt). When the past participle is used, both variations are integrated.

Literal meaning	Figurative meaning
Han bryter av kvisten.	Han avbryter samtalet.
He breaks off the branch.	*He interrupts the conversation.*
Kvisten är avbruten	Samtalet är avbrutet.
The branch is broken off.	*The conversation is interrupted.*

4.1.4.3 One orthographic word.

Swedish compounds are, almost exclusively, written as one word. Usually the boundary between the parts is not marked in any way, but in some cases there is a link element (see section 4.1.4.4). In languages where compounds are written as one word, if one wants to analyze a compound, segmentation is necessary to extract the parts that the compound is composed of. This can be a problem since one and the same string may have a number of segmentation points. An example of a string with alternative segmentation points is the Swedish word *bildrulle* which can be decomposed into *bil*‖*drulle* 'car fool' (crazy driver) or *bild*‖*rulle* 'picture roll' (roll of film). The question can even be to determine if a word is a compound at all, or a simplex word. For example *vinglas*, which can be a compound: *vin*‖*glas* 'wine glass', or a simplex word: the passive or reflexive form of *vingla* 'wobble'. Another example is *finskor*, which could be either a compound: *fin*‖*sko+r* 'fine/elegant shoe PLUR' (party shoes), or the simplex word: *finsk+or* 'Finnish woman PLUR'.

In Swedish, compounding can be done recursively. Swedish has a strong tendency of producing long, complex compounds, stronger than, for example, English or the Romance languages. In table 4.5 shows an example of a Swedish compound where the English equivalent is constructed with three words and the French equivalent with eight words.

4.1.4.4 Link elements

The link element is a liaison item that is sometimes found between the parts of a compound. The Swedish link elements are shown in figure 9.

The link elements, except the hyphen, have developed from the genitive noun suffixes. The vowel forms were productive in older Swedish. Today it is

Table 4.5: A comparison of compositionality of equivalents in Swedish, English, and French. Swedish has a greater tendency to compose complex compounds written as one orthographic word than English and French.

Swedish	näthinnevensocklusion
English	retinal vein occlusion
French	occlusion de la veine centrale de la rétine

a e o u es s -

Figure 9: The Swedish link elements.

mainly the -s- form and the hyphen that are productive. There are no definite rules for when to use a link element and when not to, though there are many tendencies.

The link elements sometimes pose a problem to automatic compound splitters. It is not always obvious to which constituent letters at the boundary, most often the letter 's', belong. Splitters sometimes interpret a link element as the initial letter of the right component, or interprets the initial letter of the right component as a link element. This can be illustrated by the examples *vind+s||tak+et* 'attic ceiling DEF' and *vind||staket* 'wind fence'.[16] Note that, without context, one cannot say that one interpretation is correct and the other incorrect, unless one of them is ungrammatical. One should rather say that one version is more probable than the other.

The link elements can sometimes be of help. If the modifier, the left component, is itself a compound, the tendency is to use a link element between this modifier and the base. This can help the interpretation of compounds with more than two constituents. The Swedish compounds *skolbokhylla* and *skolbokshylla*, would both be translated to English as 'school book shelf'. However, they have different structures: *skol||bokhylla* (book shelf in a school), which has the simplex word *skola* as modifier and the compound *bok||hylla* as head, and *skolbok+s||hylla* (shelf for school books), which is a compound that has the compound *skol||bok* as modifier and the simplex word *hylla* as head. This difference becomes apparent through the link element. Both structures are seen in figure 10.

[16]The Swedish word *vind* has two senses: 'attic' and 'wind'. When used as modifier in a compound in the sense 'attic' it always takes the link element -s-, but it never takes the -s- in the sense 'wind'.

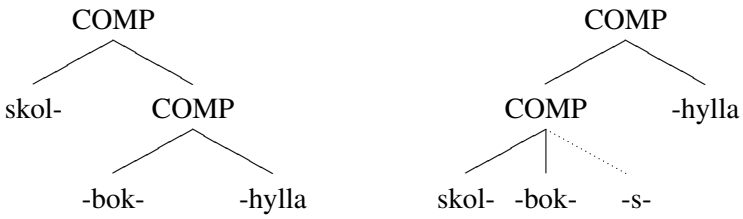


Figure 10: The link element -s- is often used in compounds where the modifier is a compound itself. This can help the interpretation of complex compounds, as in *skol‖bokhylla* (book shelf in a school) and *skolbok+s‖hylla* (shelf for school books). The structures of these compounds are shown above.

For some compounds, even when one of the constituents is a compound itself, the use of the link element varies. One example can be seen in table 12.1 on page 160, where the compound meaning 'blood lipid lowering' occurs in two variants: *blodfett+s‖sänkande* with 66 occurrences and *blodfett‖sänkande* with 37.

4.1.4.5 Elliptic coordination

In order to achieve economy in speech and writing, it is possible to use elliptic coordination, also called gapping, which is one of few situations when one can insert an element between compound constituents. To express two or more compounds which have either the modifier or the head in common, it is common to write the frequently occurring component only once. The element left out in one compound is replaced by a hyphen. Examples are: *hjärt- och kärl‖sjukdom* (i.e. *hjärt‖sjukdom och kärl‖sjukdom*) 'heart and vessel disease' (cardiovascular disease) and *binjure‖bark och -märg* (i.e. *binjure‖bark och binjure‖märg*) 'adrenal cortex and adrenal medulla'. In a search situation the query may give better result if not only *kärlsjukdom* and *binjurebark* but also *hjärtsjukdom* and *binjuremärg* were included as search keys.

Blåberg (1988) suggests that elliptic coordination can be used to test if a word is a derivation or a compound, as only compounds have elliptic coordination. He mentions one exception where a derivation has this kind of coordination, and this exception is actually quite common: *fri- och rättigheter* 'free- and right+doms' ((civil) freedoms and rights).

4.1.4.6 Compound types

Compounding is a productive way of constructing new lexemes in Swedish. New compounds are constructed when existing words do not cover a person's need to express her or himself. A compound that is coined on the fly, made up by a person who needs it, is called an **occasional compound**. Blåberg summarizes the usefulness of compounds, especially occasional compounds:

> Summing up, the function of compound formation is to provide transparent fillers for gaps in the inventory of lexical expressions of the language. Compounds are less explicit – less redundant – and shorter than full phrasal expressions. Name-worthy contents are also encountered language-internally, motivated by discourse factors. Then compounds fill typical functions of pronouns: they are less ambiguous and only slightly longer than pronouns. (Blåberg 1988: 43)

As they would otherwise be hard to interpret, occasional compounds usually have a transparent meaning, where the meaning of the whole can be derived from the meaning of the parts. Transparent compounds, both occasional and those more commonly used, are called **compositional compounds**. Compositional compounds are usually hyponyms of their head. But even if the meaning of a compositional compound is derived from the parts and the whole compound is a hyponym to the head of the compound, context or world knowledge is necessary to predict how the parts relate to each other. A *pocket knife*, for instance, is *carried in* the pocket, a *hunting knife* is *used when* hunting, and a *fruit knife* is *used for (peeling/cutting)* fruit. Often world knowledge is sufficient to determine which interpretation is the most probable, in other cases the meaning cannot be determined without context. In the example, *trä||låda* 'wood/wooden box' the modifier can either determine what the box is made of, or what is put in the box. Without context, both interpretations are quite probable.

There are compounds which are used frequently, and whose meaning has become fixed. Even if several interpretations are logically possible, only one is used, or at least, one meaning is used significantly more frequently than others. These compounds are called **lexicalized compounds** and they can often be found in dictionaries. Many lexicalized compounds have a meaning that has strayed far from the combined meaning of the components. This is especially true for frequent compounds used for common concepts. An example of a lexicalized, non-compositional compound is the Swedish word *träd||gård* 'tree yard' (garden). Even though the meaning of the first constituent means 'tree', a *trädgård* does not have to contain trees. Non-predictable compounds such as

these, are called **non-compositional compounds**. They are not necessarily hyponyms of their heads. By necessity, they have a fixed meaning as they would otherwise be difficult to understand.

Bauer (1983) and Blåberg (1988) make a three way distinction of compounds: compositional or rule-based, institutionalized and lexicalized. The institutionalized compounds can be seen as an intermediate stage which has features in common with both compositional compounds and lexicalized compounds. Like compositional/rule-based compounds, they have rule-governed form and content, but, like lexicalized compounds, they are bound to a fixed relation between the parts and potential ambiguity is ignored. One could call them lexicalized compositional compounds, as they have a fixed meaning but are transparent.

An example of an institutionalized compound could be *brand||bil* 'fire car' (fire engine). One could think of several relations between cars and fire, but the compound is, in practice, only used for the specially equipped vehicle used by professional fire fighters.

The lexicalized compounds in this three way distinction, are opaque the relation between the parts is fixed, and the typical lexicalized compound is not a hyponym of its head.

In natural language processing, lexicalized non-compositional compounds are often best treated as simplex words. Decomposing them is no help in determining the meaning or in finding conceptually related words. On the other hand, occasional compositional compounds are more interesting. If such a compound is decomposed, the parts could be found in dictionaries even if the compound as a whole is not, and the meaning may be derived. This is especially useful for a compound that is newly coined and therefore is not found in a dictionary. Decomposition could also be useful for institutionalized compounds, as the constituents have a semantic relationship with the whole compound. However, as they are lexicalized, the chance of finding the compound as a whole, in a lexicon or in a corpus, is probably greater than for occasional compositional compounds. If decomposition of institutionalized compounds is beneficial or not depends on how general and how frequent the compounds and their constituents are.

4.2 . . . in IR

4.2.1 Stemming and lemmatization

In the search process of information retrieval, identical strings in query and index are matched. This means that a match is missed if terms in index and

query are expressed in different inflectional or derivational forms. There are two approaches to dealing with this. One is indexing words in their original form, in a full form index which would contain all forms used in the document collection. In this case queries must contain all inflected forms the user deems could exist in relevant documents. In fact, this can sometimes be an advantage. For example, if a user wants to find documents about a specific instance of a concept, he or she can use only singular forms in the query, as plural forms would be unlikely in relevant documents. The second approach to dealing with different forms is conflation, to transform all inflectional variations of a lexeme, and sometimes also derivated forms, into one and the same string. Conflation can be either **stemming** or **lemmatization**. In stemming putative suffix sequences are stripped off the words, leaving a truncated word form. This string is not necessarily a grammatically meaningful form. It should be seen as a mere representative for the concept in question. Lemmatization, on the other hand, returns a grammatically correct form of the word, for instance the singular nominative for nouns or the infinitive for verbs. The chosen form represents the lemma sign.

Lemmatization in information retrieval has a practical purpose and is not as strict as lemmatization for linguistic purposes. In IR, what is insteresting is to obtain a representation for the term, in linguistics it is important to obtain a grammatical correct basic form to give a lexicologic description or to find the lemma that is used in dictionaries etc.

The use of conflation in English, which has few inflectional forms, is not obvious. Harman (1991) concluded that it was not useful. Hull (1996) studied the effects of stemming in a number of different circumstances. He found that stemming was most useful at high recall levels, and for short queries or short documents which usually contain only a small number of word forms each. Stemming showed to be less useful for long queries and long documents where it is more likely that at least some word form used has a match.

For the English language, the Porter algorithm is widely used (Porter 1980). The Porter stemmer successively removes suffixes or makes transformations of the stem in several steps. Another well known stemmer is the Lovins stemmer which removes suffixes by means of a longest-match algorithm (Lovins 1968).

The Scandinavian languages have a richer morphology than English, which leads to conflation yielding a more conclusive effect. Carlberger et al. (2001) developed a stemmer for Swedish which, in their experiments, resulted in a 15 percent increase in precision and an 18 percent improvement in relative recall. Their algorithm employs around 150 stemming rules. The rules consist of lexical patterns to match the suffixes of the words being stemmed, and a set of commands.

4.2.2 Homography, polysemy, and facets

A certain string does not necessarily carry the same meaning in all contexts. A string can be homonymic, polysemic or have several facets. A **homonym** is a string which represents different lexemes which, by chance, have the same representation. An example of a homonym is *page*, which can be a person, or a part of a book. **Polysemes** are words that have distinct meanings, but which originate from the same lexeme. For instance, a *leg* can be the leg of a person, but also the leg of a table. **Facets**,[17] on the other hand, are different aspects of the same concept. If you talk about a *person*, you could talk either about the body aspect or about the personality aspect of the same person (Croft and Cruse 2004).

In all cases, when using a string that is ambiguous in some way as a search key, there is a risk of precision deteriorating. Documents, where the string is used in a sense that does not agree with the information need in question, may be retrieved. The absolute recall will not deteriorate, as retrieving non-relevant documents does not hinder retrieval of relevant ones. However, both recall and precision at specific positions in the ranking will be lower if non-relevant documents are ranked higher than the relevant ones. This means that the relevant documents will be more difficult to locate.

There are several Swedish words from the medical domain that are homonyms, often to words that are quite common. A selection of these is shown in table 4.6. Having the index terms tagged for part-of-speech or domain could solve the problem of low precision somewhat.

4.2.3 Alternative search keys

When elaborating a topic there is often a variety of expressions and concepts that could describe it. It can be different expressions for the same phenomenon, a relation called **synonymy**, when different words have more or less the same meaning. There are other lexical relations: **hyponymy**, a hierarchical relation as in *thyroid gland – gland*, a thyroid gland is a kind of gland, **meronymy**, a part-whole relation, *heel – foot*, a heel is a part of a foot. There is also the **antonymy** relation which entails that one word is the opposite of the other, for example *left – right*. Antonyms are related in the sense that they can occupy the same slots in phrases, which implies that they are used in the same kind of situations.

[17] The term *facet* is polysemous. The linguistic term does not have the same sense as the information retrieval term (see page 13).

Table 4.6: A selection of Swedish medical terms which are homonyms to common words. The parts of speech are given for the English equivalents as several of these are homographs themselves.

Swedish	Equivalent 1	Equivalent 2	Equivalent 3	Equivalent 4
bak	behind *anat.*(N)	baking (N)	behind (Adv)	
ben	leg (N)	bone (N)		
bett	dentition *anat.* (N)	bite (N)	bit *equest.* (N)	asked (V)
blåsa	blister (N)	bladder (N)	blow (V)	
genom	genome (N)	through (Prep)		
hinna	membrane (N)	be in time (V)		
händer	hands (N)	happen(s) (V)		
led	joint (N)	suffered (V)	line (N)	track (N)
leder	joints (N)	lead(s) (V)	tracks (N)	
lever	liver (N)	live(s) (V)		
lår	thigh(s) (N)	chest/box (N)		
sena	tendon (Adj)	late (V)		
sår	wound(s) (N)	sow(s) (V)		
tunga	tongue (N)	heavy (Adj)		
tår	tear (N)	toes (N)	(small) drink (N)	
tänder	teeth (N)	light(s) (V)		
vad	calf *anat.*(N)	what (Pron)	bet (N)	how (Adv)
var	pus (N)	was/were (V)	where (Adv)	every (Adj)
ven	vein (N)	whined (V)		

When expanding queries it can be helpful to use alternative search keys that stand in relations such as the relations described above. Relevant documents may not contain the exact term you first chose for the query, but instead some term related to it in some way.

4.2.4 Decomposition of compounds

As a compound has at least two content-bearing morphemes and compounding is very productive in Swedish, a great part of information in Swedish text is contained in compounds. In the MedEval collection the ratio of compounds among the tokens is over 10%, which is shown in table 14.1. This agrees with what is reported in Hedlund 2002. Hedlund found in Swedish newspapers, after removal of stop words, that around 10% of the tokens were compounds.

The information in compounds is often essential to the contents of the documents where they are found. However, if a term occurs in a document only as a compound constituent and not as an independent word, there will be no match in the search process if only the independent term is used as search key. A corresponding situation will occur if a compound is used as search key, but only one or both constituents occur in the documents, and not the compound which was used in the query. One approach to finding information hidden in compounds is to do decomposition or segmentation of the compounds, that is split the compounds into their constituents. Thereafter the query can be expanded by adding the compound parts as search keys.

Research has been done on query expansion with compound components. This kind of manipulation of queries can increase recall, but can also result in lower precision. This may be the case if the original compound is non-compositional so that the parts are not relevant to the whole compound (see section 4.1.4.6) or if the parts have high frequency and/or low specificity. In such cases there is a risk that the use of compound constituents as search keys will result in a great deal of noise in the answer set.

Ahlgren (2004) gives examples of when decomposition of compounds can be useful and when it probably is not. For a compound such as *fot||boll* 'football' (soccer), expanding a query with *fot* and *boll* would in most cases result in lower precision since these words are used in many contexts, not only in texts concerning the ball game. On the other hand, Ahlgren points out, expanding a query containing the compound *narkotika||politik* 'drug politics', with *narkotika* and *politik*, would more likely be beneficial. Documents containing phrases like *politik mot narkotika* 'politics against drugs' would be retrieved in addition to documents containing the compound. Documents containing *narkotika* or *politik* alone would also be found. Here one can speculate that documents containing *narkotika* have a good chance of being relevant, while *politik* is a broad concept and could cause retrieval of non-relevant documents. Ahlgren did his experiments using the Swedish TREC collection containing newspaper articles from Göteborgs-Posten and Helsingborgs Dagblad.

Cöster, Sahlgren and Karlgren (2004) approach the fact that splitting compounds and using all parts in queries, often improve recall, but many times has a devastating result on precision. In their study, the queries are expanded with the leading constituent of the compounds, the modifier, to find other compounds with this as the first component. To obtain a balance between high recall and high precision, they use a Boolean quorum-level search method to rank documents. The documents are ranked both according to the tf$*$idf factor and to the number of matching Boolean combinations. Cöster et al. state that the results are encouraging, taking into consideration that the queries used were very short, with a maximum of 5 search terms.

4.2.5 Stop lists

There are words that do not contribute much to the information content of a document or query, such as articles, prepositions or conjunctions. These words are often very frequent and appear in more or less every document, something which is not good for discriminating documents from each other. They may be low in information content but they still account for a significant number of tokens in documents, making indexes larger and searches slower. A way to deal with these words is to put them in a so called **stop list**. The stop list is used as a filter when making the index and the words in the stop list will not be indexed. The stop list can also be used in automatic generation of queries. In this case the stop words will be excluded as search terms (Croft, Metzler and Strohman 2010).

The selection of stop words can be done either by using a functional approach, selecting function words, which have little topical content, or by using a frequency-based approach, selecting words with a document frequency above a certain threshold. A stop list can further be a subject stop list. If searches are done in a database of a certain domain, a stop list can be constructed containing words that appear in a significant part of the documents. In a subject stop list to be used on a medical database, the term *medical* could be included, whereas in a stop list for a linguistic database, the term *language* could instead be used. If these terms are common in the documents of the database, they will not be good at discriminating documents from each other.

The narratives of the information needs often contain specific phrases that do not contribute to the contents of the topics and would make queries less effective if they were used as query terms. These specific phrases can be put in a stop list to use when creating the queries. An example of such a stop phrase is: *Find documents that describe*.

Even if stop lists improve precision, they come with a downside. If stop words are removed, the possibility to do phrase searches on phrases containing words low in content, such as *To be or not to be*, is eliminated. For this reason, and as computers are becoming more powerful, the use of stop lists in indexing is becoming less common. The effect on precision will still remain if the stop list is used in query creation, but in that case the user has a choice as to whether stop words should be utilized or not.

4.2.6 Depending on tools

When manipulating text before indexing and when constructing queries, a number of linguistic tools can be used, for example tokenizers for splitting the

text into terms, stemmers or lemmatizers for conflating word forms into common representations, and compound splitters to decompose compounds. Later in the search process there can be a number of search engines with different search algorithms to choose from. The results the user gets when working with information retrieval must be interpreted bearing in mind that it is not always possible to have full control over the processes of these tools.

In this study the lemmatizer (see section 10.2.2) was applied to the documents before the compound splitter (see section 10.2.3) but not after. A consequence of this approach is that the modifier constituents, in the decomposed index, are listed in their stem forms, sometimes with an attached link morpheme. In order not to lose the information of the modifiers, it is important to include these forms in the search queries as synonyms to the lemmatized form. These synonym forms can also be used when searching in the non-decomposed index to catch stem forms that happen to stand in elliptic coordinations. It could also catch cases were the author of a document by mistake has written a compound with whitespace between the parts and the modifier is a stem.

The Indri/Lemur search engine, in the version used, treats the hyphen, '-', as whitespace. As many compounds, for example those with an acronym constituent, are constructed with hyphens, this affects the results. Such a compound will not be treated as a compound by the search engine, but as two simplex words, even before a compound splitter has been applied.

As the user is dependent on the functionality of the tools he or she uses, and it cannot always be expected that the user knows exactly how the tools work, it is a good idea to run the words used as search keys through the same tools as those used in the indexing process, if this is possible. This could also eliminate the effects of incorrect manipulation of strings by the tools. After all, the important factor for the results is that there is a match when there should be, and not that, for example, the suffix stripping has been done in a correct manner.

5 RESEARCH IN MEDICAL INFORMATION RETRIEVAL AND IN DOCTOR/PATIENT LANGUAGE

. . . almost, but not quite, entirely unlike . . .

The Hitch Hiker's... chapter 17

Research in natural language processing in the medical domain is mainly done in two categories of text: biomedical text and clinical text. Biomedical text is what is published about the domain in books, articles, reviews and so forth. Clinical text is what is written in the clinical setting, in the actual medical situation. This can be the text in patient health records, descriptions of patients, their medical history and present situation, findings from procedures or interviews, discharge summaries, consult reports and so forth (Meystre et al. 2008).

Clinical text differs from biomedical text in that it is produced with the purpose of communicating in the clinical situation. This text often contains a great deal of telegraphic text, shorthand phrases, abbreviations and acronyms. They also tend to contain more spelling mistakes than what is common in published texts (Dalianis, Hassel and Velupillai 2009; Velupillai 2009). These characteristics of clinical text make them an interesting challenge in natural language processing. However, the biggest challenge for research of clinical text may be a practical one. Since clinical texts, such as patient health records, contain confidential information it is very difficult to get these texts released for research. A great deal of effort is put into the research of deidentification of documents so that texts can become available to the research society without invasion of the patients' privacy (Velupillai et al. 2009).

The documents of MedEval are biomedical texts, texts written about medical issues. Before the creation of MedEval, there was no similar test collection. As far as the author of this thesis is aware, MedEval is the first Swedish medical test collection, it is the first test collection with documents assessed for target reader group and the first to have the recall bases adjusted to user groups, although Hahn, Honeck and Schulz (2002); Schulz, Honek and Hahn (2002)

address the issue of user groups by having two types of queries: professional and lay person queries. The author is also not aware of research connecting information retrieval and terminology. There is of course research relating to the different parts of the research behind this thesis: medical IR in different forms, IR and decomposition of (medical) compounds, and relations between expert and non-expert language. The rest of this chapter will give a brief overview of such research.

5.1 OHSUMED

OSHUMED is a medical test collection in English. It is built on nearly 350 000 references from MEDLINE, a large bibliographic database containing references to biomedical articles etc. from scientific journals (Hersh 2003). The references include title, abstract, MeSH indexing terms, author, source and publication type. The OSHUMED documents are assessed for topic relevance on a three-graded scale: definitely, possibly and not relevant. OSHUMED contains 106 topics generated by physicians from authentic situations and involve specific patients. The topics consist of both information about the patient in question and the information need related to that patient. In contrast to the topics in MedEval and in the TREC collection, the topics contain neither title nor narrative, as seen in figure 11.[18] (Hersh et al. 1994)

```
.I 1
.B
60 year old menopausal woman without hormone replacement therapy
.W
Are there adverse effects on lipids when progesterone is given
with estrogen replacement therapy
.I 2
.B
60 yo male with disseminated intravascular coagulation
.W
pathophysiology and treatment of disseminated intravascular
coagulation
```

Figure 11: Topics 1 and 2 from the OHSUMED test collection, which is based on clinical text.

The OHSUMED test collection is not profiled for lay person users. The information needs are written in telegraphic clinical style. It does not have

[18] <http://ir.oshu.edu/ohsumed/queries>

documents assessed for target group, there is no possibility to choose user group and the language of the topics is a professional language with many neoclassical terms.

5.2 The TREC genomics track

The genomics track was introduced in 2003 as the first domain-specific TREC track and ran annually until 2007. It had as a goal to study how domain-specific information can improve retrieval effectiveness and to provide information retrieval test collections in the genomics domain.

The track developed through the years and had ad hoc retrieval tasks, summarization tasks, text categorization tasks, and question-answering tasks.

The documents used in the genomics ad hoc task were based on the MEDLINE bibliographic database. It consisted of completed citations from 1994 to 2003. The topics were created by professional biologists. In the final years, the biologists were provided with generic templates and were asked to formulate needs that fitted the templates and that they themselves recently had experienced. An example of a generic template is given below.

> Find articles describing the role of a gene involved in a given disease.
> (Hersh and Voorhees 2009: 4)

The relevance judges generally had backgrounds in biology or medicine. In parallel with the templates provided for the topic creators, the judges were given explicit instructions on how to judge the documents. The documents were judged as 'definitely relevant', 'possibly relevant' or 'not relevant'.

Domain-specific techniques such as expanding queries with term synonyms and gene names, and non-domain-specific techniques such as advanced document weighting and query expansion were studied. There was some evidence that domain specific resources, such as controlled terminology lists, improved results somewhat, but not substantially, over using standard information retrieval resources (Hersh and Voorhees 2009).

5.3 Subword-based text retrieval

Hahn et al. (2001); Hahn, Honeck and Schulz (2002); Schulz, Honek and Hahn (2002); Hahn, Markó and Schulz (2005) propose an approach for the decomposition of medical compounds, in index and query. The compounds are split into **subwords** and affixes. What the authors call a subword is not a linguistically motivated morpheme, but rather medically motivated. In Schulz,

Honek and Hahn (2002) the authors phrase it: 'we trade linguistic atomicity against medical plausibility'. A subword is a semantically minimal unit which is motivated by its usefulness in document retrieval. What makes subwords different from compound constituents is that a subword is not split further if it is a lexicalized medical term. A linguistically motivated segmentation of the neoclassical term *diaphysis* (shaft of long bone) would be *dia+phys+is* while the subword segmentation is *diaphys+is* where the first part is a near synonym to 'shaft'.

To determine which subwords should not be decomposed further into morphemes, the research group constructed a subword dictionary containing terminology of clinical medicine, which includes scientific terms, clinicians' jargon and popular expressions. The morphological analyzer employs regular expressions and looks for the longest match from both left and right.

The research group compared the results of retrieval using several different matching methods: plain token match (with and without orthographic normalization and also with and without adjacency boost, assigning higher ranking to adjacent terms and lower to terms further from each other), trigram match, subword match and synonym-enhanced subword match where synonym class identifiers that represent sets of subwords were used as index terms.

The approach based on subword segmentation performed substantially better than the non-lexicon-based methods. However, the synonym-enhanced subword match performed worse than the subword approach. The best of the non-dictionary-based methods, plain token match with orthographic normalization and adjacency boost, had an average precision of 28.3% measured with 11pt precision. The subword match had 33.9% using the same measure and the synonym-enhanced subword match had 31.2%. This result suggests that the use of some form of dictionary is called for in retrieval of medical documents. The trigram approach gave poor results, lower than the plain token match baseline.

Hahn, Honeck and Schulz (2002); Schulz, Honek and Hahn (2002) used two sets of user queries for their study, namely expert queries, containing medical jargon, and layman queries. Some differences between the queries were observed. The adjacency criterion did not have effect on the lay person queries, which was explained by lay person queries containing fewer search terms. However, the subword approach especially, but also the synonym approach, gave a considerably higher gain for the non-expert queries than for the expert queries.

5.4 The Morphosaurus

The Morphosaurus project[19] is a development of the subword-based text retrieval described above. The Morphosaurus is a medical language tool which uses subword indexing to create an interlingua for cross language information retrieval. The Morphosaurus transforms both the queries and the documents into a language independent interlingua. The Morphosaurus uses the subword thesaurus to define interlingual semantic equivalence classes. In example 12 a German phrase is given together with an English and an interlingua equivalent.

- Erhöhte TSH-Werte erlauben die Diagnose einer primären Hypothyreose ...
- High TSH values suggest the diagnosis of primary hypothyroidism ...
- #up tsh #value #permit #diagnost #primar #small #thyre

Figure 12: German, English, and interlingua equivalents of the same phrase.

In Markó, Schulz and Hahn 2005 the Morphosaurus's interlingua approach and a direct query translation approach are compared to a baseline English monolingual retrieval using the OHSUMED collection. The Morphosaurus reaches 93% of the 11pt precision of the monolingual runs, while the query translation approach reached 62%.

5.5 MuchMore

The MuchMore[20] project is about concept-based cross-lingual information retrieval within the medical domain. The project aims to develop and evaluate methods for using multilingual thesauri when annotating German and English medical texts with semantic information. Concept-based methods are compared to corpus-based methods (Volk and Buitelaar 2002). In the experiments they used a parallel English-German corpus containing 9 000 scientific medical abstracts with 1 million tokens for each language.

Even though MuchMore is about cross-language retrieval, monolingual experiments were done in German and English to obtain baseline results for the cross-language runs.

In the MuchMore project semantic codes (MeSH, UMLS, and EuroWordNet) were assigned to both documents and queries. This semantic annotation

[19] <http://morphwww.medinf.uni-freiburg.de/index.html>
[20] <http://muchmore.dfki.de/>

was used as a sort of interlingua, as the annotation of the German queries are matched to the annotations in the English documents.

In the monolingual experiments the runs using semantic information from MeSH gave better results for precision and recall than runs using token (word forms) and lemma indexing. For the cross-language retrieval, combining MeSH annotation with lemma and token indexing outperformed runs using machine translation, while using EuroWordNet performed worse than baseline.

For a baseline in the cross-lingual experiments they used the tokens of the German queries directly on the English documents, something which retrieved 66 relevant documents. The idea was that the technical vocabulary has an overlap between the languages.

For the runs using the semantic coding they compared the semantic coding of German queries to the semantic coding of English documents. As in the monolingual runs, the MeSH coding had the best results, and EuroWordNet the worst.

However, using a similarity thesaurus built on the English-German parallel corpus gave a better result. The best result of all was obtained combining all methods. These runs came close to the monolingual runs in effectiveness.

5.6 Medical image retrieval

In 2003 the Cross Language Evaluation Forum, CLEF, started the ImageCLEF track with the goal of understanding and improving image retrieval in multilingual document collections. The following year, ImageCLEFmed, a medical image retrieval task was added.

Hersh et al. (2006) describe results of ImageCLEFmed 2005 where thirteen research groups participated in testing an image test collection. This first test collection was built from existing collections of images containing clinical case descriptions including radiographs, gross images, microscopic images and nuclear medicine images. The pictures were annotated in both English, French and German. 25 topics were developed for the test collection. These topics contained a textual information need statement and an index image. The topics were classified based on if they were more suited to be retrieved by a visual, a textual or a mixed algorithm. All topics were written in English, French, and German. In 2007 the number of topics had grown to 85 (Hersh, Müller and Kalpathy-Cramer 2008).

The relevance assessments were performed by physicians who were also graduate students in a biomedical informatics program. All images for a topic were assessed by a single assessor.

Show me images of right middle lobe pneumonia.
Zeige mir Bilder einer Lungenentzündung
des rechten mittleren Lungenlappens.
Montre-moi des images d'une pneumonie du lobe médial droit.

Figure 13: An example of an information need from the ImageCLEFmed 2005 test collection. This topic is classed as semantically oriented.

For 20 topics of the 2005 collection half of the images were judged by a second assessor. The images were judged on a three-graded scale: relevant, partially relevant and not relevant. As this collection partly contains duplicate relevance judgments, experiments were done to see how differences in judging affected the results. It turned out that the different judgments led to modest absolute changes in mean average precision, but the relative performance between runs was largely unchanged.

In 2005 there were 13 research groups which carried out runs using the same retrieval approach for all topics. These approaches included two categories of topic modification, automatic or manual, and three categories of retrieval systems, visual, textual, or mixed. The best results were achieved by the combination of automatic topic modification-mixed retrieval system.

The medical retrieval task of ImageCLEF 2008 had 66 662 images with annotations, 85 topics and 800 - 2 000 relevance judgments per topic (Hersh, Müller and Kalpathy-Cramer 2008). For the 2010 track the collection contains over 77 000 images.[21]

The image retrieval task has shown that a number of approaches can be used for image retrieval. Text based methods alone are more robust than visual techniques alone, but visual techniques work quite well on topics oriented towards visual retrieval especially in combination with text.

5.7 Expansion with professional and lay person language

Dioşan, Rogozan and Pècuchet (2009) address the fact that medical professionals and lay persons express themselves in different ways when discussing

[21] <http://www.imageclef.org/2010/medical>

medical issues. The authors discuss how to exchange information across user groups. The goal is that a search using non-expert terms should retrieve all types of documents written on the topic, no matter what style they are written in. The mapping of general language to professional language is motivated by the empowerment of lay persons.

The authors' approach is to see the problem as a question of automatic alignment between specialized terminology and general terminology. The objective is to enrich the information retrieval system with a set of links between corresponding concepts in the two sublanguages. The alignment is done by using different machine-learning techniques, such as k-nearest neighbor classifiers, evolutionary algorithms and support vector machines.

5.8 Building a lexicon of professional and lay person equivalents

Elhadad and Sutaria (2007) describe a corpus-driven method for building a medical lexicon of equivalents in professional and lay person language. They use a comparable corpus with abstracts of clinical studies on one side, and news stories written about these studies on the other side. Every news story includes a reference to the original scientific article.

The authors describe a method were they compute contingency tables based on co-occurrence of professional and lay person terms. Their definition of co-occurrence is that a term which occurs in professional documents must be present in at least one lay person document. The calculation is then done on the document frequencies in both groups of documents for the corresponding CUI (concept unique identifier).

The authors do not claim strict medical equivalence in the resulting pairs, as the professionals often express themselves more precisely than what can be understood, or needed, by a lay person. An example of such a pair is *diabetes mellitus*↔*diabetes*. The authors point out that the equivalence pairs may have any of a number of semantic links, such as a synonym relation, a hyponym relation or perhaps no semantic link at all.

5.9 Swedish expert and non-expert registers in the medical domain

Kokkinakis and Toporowska Gronostaj (2006) have made a corpus-based study of Swedish medical language, where they contrast the language, especially the vocabulary, of documents written for medical experts to the language of documents written for non-experts. They have selected documents from the

MedLex corpus within the domain of cardiovascular disorders. Thus the documents in their study are a subset of the documents of the MedEval collection.

5.10 Communication between doctors and patients

Berbyuk Lindström (2008) describes and analyzes communication between doctors and patients in Sweden. The thesis is mainly about the special situation of non-Swedish doctors and Swedish patients, and how they solve difficulties in communication. The study is done mainly on spoken language, as doctors and patients usually meet face to face.

Berbyuk Lindström focuses on the situation where the doctor has the advantage of his professional position, while the patient has the advantage of being more fluent in the language. The author compares this situation to the corresponding monocultural communication between Swedish doctors and Swedish patients.

6 RESOLVING POWER

Make it evil

The Restaurant... chapter 23

The resolving power of a term is about how effective the term is in describing the contents of a document and how good it is at differentiating documents described by the term from the rest of the collection. This chapter gives an overview of the ideas behind the resolving power of terms.

6.1 Significance within documents

Luhn 1958 is an early paper on automatic text summarization which came to have great importance in the upcoming field of information retrieval. The author presented the idea that the significance of a word in a text could be derived from the number of occurrences of that word in the text. He suggested that it is part of the writing process to repeat significant words when elaborating on a subject. He also pointed out that there is one group of words often repeated which are not significant. These are function words.

> The justification of measuring word significance by use-frequency is based on the fact that a writer normally repeats certain words as he advances or varies his arguments and as he elaborates on an aspect of a subject. This means of emphasis is taken as an indicator of significance. [. . .] certain other words must be present to serve the important function of tying these words together, the type of significance sought here does not reside in such words. (Luhn 1958: 160)

Luhn visualizes his idea in a word-frequency diagram reproduced in figure 14. The x-axis of the diagram represents the words of a typical text ranked in decreasing order of frequency, and the y-axis the frequencies of those words. The significant words are in the middle range, between the vertical lines. The

dashed line shows the degree of discrimination or resolving power of the terms. The words that are most likely to be significant to the contents of a text, and which should be chosen to represent the document, are in the frequencies between the vertical lines. Luhn describes the optimum placement of the vertical lines as a matter of practical experience and concludes that the lines can be adjusted to give the desired output.

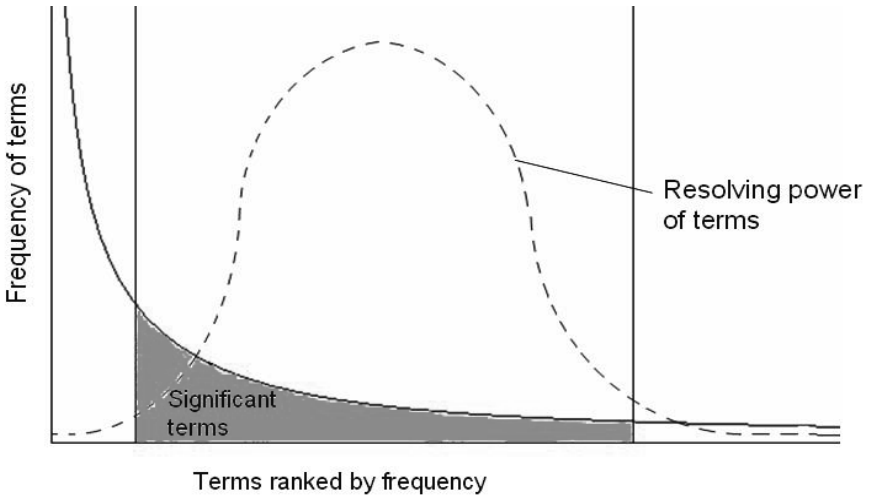


Figure 14: When the terms are ranked by frequency, the most significant terms tend to end up in the middle range, here in the colored area between the vertical lines (Luhn 1958).

Luhn ideas are based on earlier work by Zipf (1949). Zipf showed that if words in a text, or text collection, are ranked in order of decreasing frequency, they show a relationship now known as Zipf's law, shown in equation 17.

$$frequency \times rank \approx constant \tag{17}$$

This means that the frequency of a given word multiplied by the rank of that word will be approximately equal to the frequency of any other word multiplied by the rank of that word. Zipf's law implies that there are a few words that occur very frequently, and very many words that are very scarce. The frequency falls off very rapidly at first, but then more gradually as the words become less frequent. A curve, such as the sloping frequency curve in figure 14 is often referred to as a Zipf curve.

6.2 Significance in collection

Before giving her own definitions Spärck Jones summarizes the then current notion of the terms exhaustivity and specificity in the following way:

> We are familiar with the notions of exhaustivity and specificity: exhaustivity is a property of index descriptions, and specificity one of index terms [...] the exhaustivity of a document description is the coverage of its various topics given by the terms assigned to it; and the specificity of an individual term is the level of detail at which a given concept is represented. (Spärck Jones 1972: 11)

The author then goes on to propose an expanded view of the terms exhaustivity and specificity, for information retrieval purposes, moving away from the semantic view point. She suggests a statistical interpretation based not only on the term frequencies of individual documents, but on term frequencies of whole collections.

> We should think of specificity as a function of term use. It should be interpreted as a statistical rather than semantic property of index terms. [...] We can thus redefine exhaustivity and specificity for simple term systems: the exhaustivity of a document description is the number of terms it contains, and the specificity of a term is the number of documents to which it pertains. (Spärck Jones 1972: 12)

Spärck Jones points out the relationship between exhaustivity and specificity. The more exhaustive descriptions in a collection are, the less specific the terms become as they will be less discriminative. Extracting more key words per document would be more likely to increase the frequency of current keywords than to generate new ones. This would in turn make some terms less useful for distinguishing documents from each other. She points out that this can happen even to terms with a quite specific meaning if they are used frequently. A growing collection has the same effect on specificity of terms. Whatever the meaning, a term becomes less effective for retrieval if it is commonly used.

> A frequently used term thus functions in retrieval as a nonspecific term, even though its meaning may be quite specific in the ordinary sense. (Spärck Jones 1972: 12)

Spärck Jones (1972) discusses the implication of the different values of terms with different frequencies and suggests that terms should be weighted

in correlation with their collection frequency: a match of a non-frequent term should be of more value than a match of a frequent term. These ideas would later develop into what is now known as inverse document frequency or idf. Note that Spärk Jones in the citation below uses the term 'document frequency' in another sense than what is done in this thesis.

> Weighting by collection frequency as opposed to document frequency is quite different. It places greater emphasis on the value of a term as a means of distinguishing one document from another than on its value as an indication of the content of the document itself. The relation between the two forms of weighting is not obvious. In some cases a term may be common in a document and rare in the collection, so that it would be heavily weighted in both schemes. But the reverse may also apply. It is really that the emphasis is on different properties of terms. (Spärck Jones 1972: 16)

6.3 Discrimination value

Salton and McGill (1983) claim that term frequency alone is too crude to use for indexing for retrieval purposes. If the words were distributed randomly over a collection it would not be possible to distinguish between documents looking only at quantitative measures. Instead of talking of the resolving power of a term, as Luhn did, the authors speak of the discrimination value of terms. They also point out that the semantics of a term is important, not in itself as was the case in the early definitions of specificity, but in relation to the essence of the document.

> . . . a question of principle arises concerning the use of *absolute* frequency measures [. . .] The reason is that a useful index term must fulfill a dual function: on the one hand, it must be related to the information content of the document so as to render the item retrievable when it is wanted (the recall function); on the other hand, a good index term also distinguishes the documents to which it is assigned from the remainder to prevent the indiscriminate retrieval of all items, whether wanted or not (the precision function). Salton and McGill (1983: 62)

Salton (1981) and Salton and McGill (1983) describe the term discrimination value as the degree to which the use of a specific term is helpful in distinguishing some documents from the remainder of the collection. In other words, the discrimination value is the ability of a term to cluster documents that are simi-

lar and separate them from other documents. Salton illustrates document similarity graphically as an inverse function of document distance. He suggests that the space density should be measured as the average pair-wise similarity between all document pairs in the collection, which means that the more similar documents are, the closer they are in the graph. A good content term is one which decreases the space density when it is assigned to the documents of a collection. Documents are most easily retrieved when they are distinguishable from their neighbors. Salton (1981) divides discrimination values into three groups:

- Negative discrimination value for broad terms which make the document space density greater.
- Positive discrimination value for medium frequency terms that distinguishes a class of items from the remainder of the collection.
- Discrimination value close to 0 for specific terms which do not alter the document space.

Figure 15 shows three cases of what happens when terms with different degrees of discrimination value are added to an index or a query in indexing or retrieval respectively. The starting point is a document space within which documents are spread quite evenly. In the leftmost case a frequent/broad term is added. Most of the documents will be affected and the distance between the documents becomes smaller as the set of index/query terms describing the document/request become more alike. The document space is dense and it is difficult to discern the documents from each other. In the rightmost case a very rare/specific term is added. Here the relative distances between the documents remain more or less the same as without the term. A rare term will leave most documents unaffected, and there is not much of a clustering effect. In the middle case a term of medium frequency/specificity is added. Here the documents containing this term are clustered, appearing close in the document space, and, equally important, they are separated from the documents which do not contain the term and should be rejected. Thus this is a term with high discrimination value. Assuming that the term in question is relevant to a users information need, it would be useful as a search key.

In Salton 1981 and Salton and McGill 1983 the authors explain how an automatic indexing process could utilize term discrimination value to construct suitable index terms by phrase transformation or by thesaurus transformation. By phrase transformation they mean combining two or more terms into one index entry by requiring that they stand in a certain proximity of each other, and by thesaurus transformation they mean that one should regard related terms to be of the same concept and represent them all with a single representation.

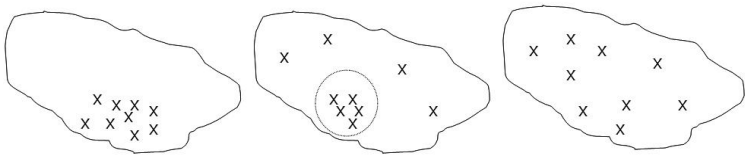


Figure 15: This figure shows the effects of adding terms to a document space where the documents are evenly distributed. In the document space to the left, a term with negative discrimination value is added to the index/query. This makes the space more dense and discrimination of documents becomes difficult. In the document space to the right, a term with discrimination value near zero is added. It does not affect the space density much and does not discriminate the documents. In the document space in the middle, a term with positive discrimination value is added. The documents which contain the term are clustered together.

The phrase transformation should be used on frequent terms with negative discrimination value to turn them into medium frequency terms with positive discrimination value. Salton and McGill suggest that the phrase construction should be very loose to avoid being over-specific, something which could lead to lower recall. It would be sufficient if the terms included in a so called phrase occurred in the same sentence.[22]

A thesaurus groups terms that are synonymous or semantically related into common classes. A thesaurus could therefore be used to transform very specific terms with near-zero discrimination value into terms with positive discrimination value by substituting rare or specific terms with the thesaurus class.

The authors emphasize that phrase transformation should only be used on terms with negative discrimination value and thesaurus transformation on terms with discrimination value near zero. This is illustrated in figure 16 where the arrows representing phrase transformation and thesaurus transformation both are directed towards the area with medium frequency terms and positive discrimination value. The phrase transformation is represented by the arrow in the left to right direction and the thesaurus transformation is represented by the arrow in the right to left direction. Any other use of these transformations would make the results deteriorate.

[22] In Salton 1981 the recommendation is even less strict: the terms of the phrase were required only to co-occur in the same document, possibly in the same sentence.

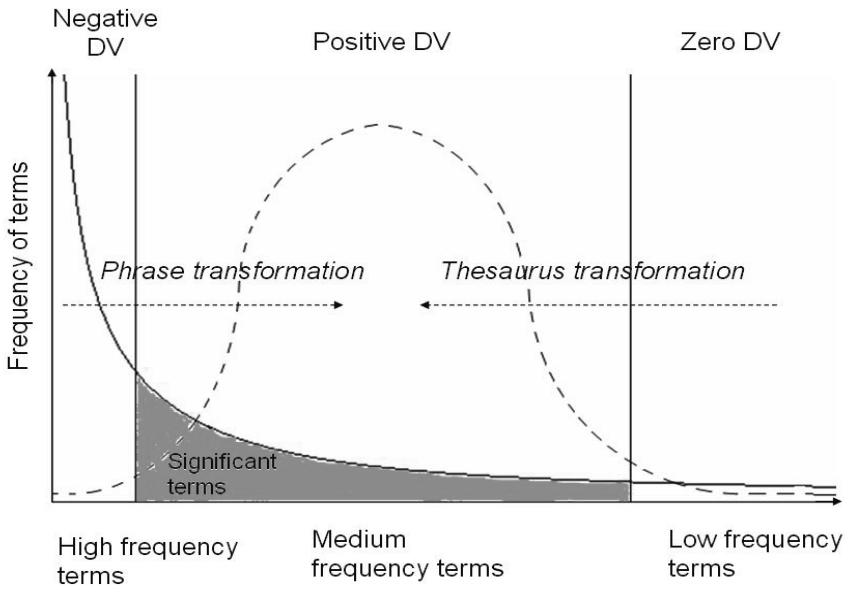


Figure 16: Phrase transformation can be performed on terms with negative discrimination value and thesaurus transformation can be performed on terms with discrimination value near zero. This would bring both categories of terms closer to a positive term discrimination value.

6.4 Significance within query – key goodness

While Luhn (1958) defined the resolving or resolution power of terms within documents, and Spärck Jones (1972) defined it in the context of collections, Pirkola and Järvelin (2001) describe term resolution power in the context of other keys, for a given request.

> The *resolution power* of a key is its ability, *in the context of other keys for a given request,* to increase query performance. (Pirkola and Järvelin 2001: 575)

The authors made an a posteriori analysis and ranked the query keys in order of resolution power, dividing them into three groups:

Best key Key which, in combination with other keys, increases effectiveness more than all other keys.

Good keys Keys which, in combination with other keys increase effectiveness.

Bad Keys Keys which, in combination with other keys, lower or have no impact on effectiveness.

Pirkola and Järvelin found that one search key often has much higher resolution power than the other keys and could be selected automatically. Giving this high resolution power (HRP) key more weight than the other keys improved the search results significantly

In most cases the HRP keys were parts of noun phrases, often proper names or components of proper names. Proper names were more often best keys than good keys or bad keys. The HRP words have a tendency to cluster, that is, they appear in fewer documents and with more occurrences in each document than good or bad keys with the same frequency in the collection. Therefore Pirkola and Järvelin conclude that one can use the statistical properties of words to automatically determine their semantic significance for a given topic.

While it often was possible to find statistical differences between the best and the good keys, the good keys were often difficult to statistically differentiate from bad keys, even though there were some tendencies. The bad keys were heterogeneous and tended to be bad for many reasons. They had a high standard deviation in their statistical values, such as document frequency and collection frequency. A number of bad keys were semantically vague, such as *approach* and *consequence*. A higher percentage of the bad keys than that of the good keys had a document frequency of over 5 000.

Pirkola and Järvelin conclude that there is a difference between the **logically** most important keys, and the keys which **in practice** discriminate relevant documents from non-relevant documents. While it is not difficult for a user to pick out logically important keys, the ones that are most about the topic, knowing which one in practice is the best one is more difficult as it depends on factors the user is not aware of, such as the contents of the database and the statistical properties of the keys in it. Therefore an automatic calculation of the HRP key can be valuable.

Part II

The reader's guide to the galaxy of this thesis

7 DRAWING THE ROAD MAP

Deep Thought

The supercomputer[23]

7.1 Where to go

The work on this thesis started with an ambition to investigate compounds and compound constituents in a Swedish document collection of the medical domain. This demanded an appropriate test collection. Since test collections are scarce, domain specific collections even more so, and Swedish medical test collections non-existent prior to MedEval, the first step along the way had to be to build a suitable collection. Further on, it became apparent that studying the behavior of compounds as search keys in a medical collection should not be done in isolation but rather in context with and in comparison to other types of search keys. As always in information retrieval, it is the relative results that are interesting. The directions had to be turned from studying compounds as search keys to studying search keys in general, albeit with a special lookout for compounds.

A goal was set: to study which features terms had that worked well as search keys and which features terms had that did not work well. The plan was to study both simplex words, compounds and compound constituents. The knowledge of term effectivity could then be used, for instance, to select which compound constituents that would be beneficial to use as search keys in queries. So far along, the candidate research questions were:

[23] Deep Thought is the name of the computer in Douglas Adams's trilogy in four parts. The computer was built to learn *The ultimate answer to the ultimate question of Life, the Universe and Everything*. It took Deep Thought 7.5 million years to compute the answer. The answer was: **forty-two**.

- What features do terms that are good search keys have?
- What features do terms that are bad search keys have?

It was soon clear that these questions were premature. They could not be answered without going back along the road for the more basic questions:

- What is a good search key?
- What is a bad search key?

Obvious answers to these questions were: 'A good search key is a term that gives a good search result' and 'A bad search key is a term that gives a bad search result'. This, however, demanded going back even further for even more basic questions:

- What is a good search result?
- What is a bad search result?

What do we mean by search result goodness? Good precision? Good recall? These measures exist in a trade-off relationship, and as satisfaction with the result is very much dependent on what a particular user needs for the moment, it is difficult, if not impossible, to specify, once and for all, what is good and what is bad. However, what we can do is look at behavior patterns of search terms with different features and to discuss when the various behaviors are beneficial and when they are not.

As could be seen in the previous chapter, a great deal has been written on term effectivity since the birth of the research field of information retrieval in the 1950s. The earlier works focused on index terms, not on terms used as search keys. The question back then was which terms should be used to index a document in order to facilitate retrieval of that specific document when searching for information on the subject of the document.

Research conditions have changed considerably since the 1950s. Most important is that the power of computers has increased by many orders. Now the use of stop lists and controlled vocabulary in indexing is diminishing. Documents are instead indexed in full text, possibly manipulated in some way, for example by lemmatization or by decomposing compounds. With indexes now usually containing all document terms in some form, the question is not what to index, but rather which terms should be used as search keys to find documents on the subject that you need information about.

The questions of index terms and search key terms are related, like two sides of the same coin. An index term is not effective if you do not search for

it. A search key cannot be effective if it does not match a term in the index. A term must both be in the index and be used as a search key in order to be effective. This is why what is written on the goodness of index terms can be applied to the goodness of search terms. However, the focus has changed from the document in the collection to the query in the search. From 'How can I most efficiently find this document again when I need the information?' to 'How can I most efficiently find documents containing the information that I need?'

Even though index terms and search key terms are two sides of the same coin, the now common full text indexing, and the overwhelming and growing amount of documents, has changed the conditions. With all words of the stored texts indexed in some form, the tf∗idf values are affected as the overall number of terms is higher. A fair part of the terms will be indexed, not only representing documents where they are salient to the content, but also representing other documents, where they have a more peripheral use. Full text indexing improves recall potential, at least for documents where the desired topic is present but where it is not the main subject. However, there is a risk of loss of precision. For instance, documents which explicitly state that they are *not* about a certain subject may be retrieved, seeing that negative statements contain the terms that they negate, and full text indexing indexes all terms. This was not a problem when documents were manually indexed.

Manual indexation has not entirely disappeared. One situtation where it is still very important is keywords in library cataloging systems or in abstracts where they often have a separate heading under which the author (or someone else) chooses the terms that best describe the content. Here again, the purpose of the indexing terms is to facilitate the retrieval of a specific document.

A significant part of this thesis is about the building of the MedEval test collection. That, as such, is not about term goodness, but creating a collection like MedEval has made it possible to perform experiments on term goodness based on domain and on user and target groups. Building a test collection gave good insight into the areas of information retrieval most relevant to a linguist such as indexing, assessment of relevance and of target groups and, not least, query construction.

The assessment of documents for target groups can be utilized in two ways. One direction is to study all assessed documents and compare the language of the two groups. The results of this comparison can be used as a guideline in constructing suitable queries. The other direction is to put queries to the search system and compare the results of the different user scenarios, for example running synonyms separately as search keys to see if they retrieve documents predominantly for one user group or the other.

The construction of the MedEval test collection was completed by a series of pilot studies with focus on term goodness and on patient/doctor language. These pilot studies are intended to show the way and to illustrate what kind of studies can be done with such a collection.

The road map is now nearly finished. What remains is describing the goal, i.e. the final research questions of the thesis. The questions were mentioned in the introduction chapter, but for clarity they are repeated here:

- What features do terms that are good search keys have? What features do terms that are bad search keys have? Can this knowledge be used to select compound constituents to use as search keys?
- Can specific features of professional language and of lay person language, respectively, be utilized when searching for medical documents for the two target groups?
- Can the questions above be answered using a medical test collection with two indexes containing different representations of compounds, split and unsplit, and providing user group scenarios, professionals and lay persons? What other research questions can be answered with such a collection?

8 TRAVEL INSTRUCTIONS

For heaven's sake, mankind, it's only four light years away you know.

The Hitch Hiker's... chapter 3

8.1 The means to get there

Now that it was clear where we wanted to go we needed the means to get there. The tools and resources used, partly for the construction of MedEval, but mainly for the experiments, are described in chapter 9. The main resource for the experiments was of course MedEval itself. The steps along the way of constructing this collection are described in chapter 10.

8.2 Survey of the landscape

Having come this far the question was: How can we reach the goal of discovering what features good search key terms have and what features bad search key terms have?

The first stretch was finding search keys to examine, and to construct baseline queries to compare against. There were many forks in the road: Deciding which concepts of the topics should be represented, deciding how many synonyms or near synonyms should be used and which synonyms.

It was all done by dividing the information needs into facets containing near synonyms found in the MedEval topics (see appendix A) and in the Swedish MeSH.[24] Lexicalized phrases were treated as terms. For each topic the facets were combined into a baseline query. The choices that were made in this stretch of the road are described in chapter 11.

The first survey of the landscape of facets and search keys was performed in user scenario 'None', which does not differentiate between expert and non-

[24] <http://mesh.kib.ki.se/>

expert documents. For directions on how to study the effectivity of facets and search keys some of the ideas described in chapter 6 were implemented on MedEval. Results consistent with these early ideas would indicate that the ideas are applicable also to Swedish medical texts.

Traveling down the linguistic road included examining simplex words and compounds, compound heads and modifiers. Another aspect could be to examine if the domain of the search term played any role in search term goodness, but this has been left to later studies.

To get an overview of the landscape, runs were made using the baseline queries constructed earlier. To conclude which were the best facets for each topic, the facets were tested in two opposite ways. The first was to test each facet separately, to see if it was effective on its own. The second was to run the query with one facet at a time removed, to see how this removal affected the results. The same method was used on the facet level to examine how well the terms of the facets worked as search keys. Here each term was run separately as a one search key query and then, one term at a time was removed from the facet. This overview resulted in a number of effective and ineffective terms being identified. Thus it was possible to study the features of keys giving good results and of keys giving bad results.

Some aspects studied were:

- The tf∗idf factor: frequency and clustering of terms.
- Term goodness in the context of of the database or in the context of other terms.
- Specificity, exhaustivity and their relationship to the topic.
- Whole compounds vs. compound constituents, compositional vs. non-compositional compounds and simplex words vs. lexicalized phrases.

To summarize all of these aspects: When your task involves choosing between a whole compound and one constituent or the other, or between simplex words and lexicalized phrases, the map seems to suggest that you treat them all equally and that you treat them as terms whatever the form. It seems to suggest that you look at the features in the context of your topic (is this a term that is about your topic?) and in the context of the database (is this a term in the middle range of frequencies and clustered in not too many documents?).

8.3 Experts and non-experts

The last area to be examined was differences in the language of documents written for the two target groups doctors and patients. The idea was to see if

there were specific features in expert and in non-expert language that could be utilized when searching for medical documents for the two groups of users.

To examine the difference between professional language and lay person language, sets of documents were created in order to compare documents from the two target groups, doctors and patients, to each other and to the whole collection. Different type and token frequencies were calculated as well as average word length, ratio of compounds, and multiword units. Is there, for example, a difference in the types of multiword units in documents for the two target groups? Is there a difference in frequencies for the two user groups?

8.4 Zooming in

As the result of this thesis is an overview of how linguistics and a test collection such as MedEval can contribute to research in medical information retrieval, it is also an invitation to continue the research, to zoom in on the details. Hopefully, in the future we will see more detailed maps of the treatment of compounds and of retrival of documents targeted at certain user groups. This is just the beginning.

Part III

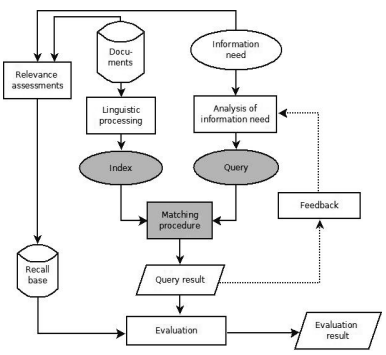
Resources and test environment

9 TOOLS AND RESOURCES

It is a mistake to think you can solve
any major problems just with potatoes.

Life... chapter 24

9.1 The Indri/Lemur retrieval system



The Indri search engine is part of the Lemur Toolkit,[25] an open-source toolkit, developed at the University of Massachusetts, Amherst. It is a development of the Lemur's InQuery search engine which was based on the inference network model with ranking based on $tf*idf$ estimates. Indri is also based on the inference network model but now combined with language modeling. The inference network makes complex structured queries possible and the language model gives the belief values formal justification. The documents are retrieved and ranked according to $P(I|D,\alpha,\beta)$, which is the belief that the information need I is met given the document D and the priors α and β. For more information on the retrieval models behind the Indri search engine see sections 2.4.3.2 and 2.4.3.3.

The basic building blocks of the Indri Query Language are terms. The atomic term consists of one search key. Terms can be simple, atomic terms or proximity terms. Proximity terms are constructed of terms and of proximity operators which define, for instance, ordered or unordered phrases and synonyms. Belief operators are used on terms to define how to combine the evidence of the terms, for example by weighting, filtering and more. The belief operators, #combine(), #weight() etc., are used to calculate, for each document, the belief

[25] <www.lemurproject.org/>

that the query is relevant to that document. The documents are ranked according to this belief score (see section 2.4.3.3) (Strohman 2007; Lemur nd). The most important operators of the Indri query language are shown in figure 17.

- Proximity operators

  #syn($q_1\ q_2 \ldots q_n$**)** Synonym operator. The terms in the expression are treated as instances of the same term.

  #odN($q_1\ q_2 \ldots q_n$**)** Ordered window. The terms must appear ordered with not more than N-1 terms between them in order to match.

  #N($q_1\ q_2 \ldots q_n$**)** Ordered window. Equivalent to #odN($q_1\ q_2 \ldots q_n$).

  #uwN($q_1\ q_2 \ldots q_n$**)** Unordered window. The terms must appear, in any order, within a window of length N.

- Belief operators

  #combine($q_1\ q_2 \ldots q_n$**)** Combine operator. The belief of all terms, are combined and treated equally.

  #weight($q_1 w_1\ q_2 w_2 \ldots q_n w_n$ **)** Weight operator. The beliefs of the terms are given weights, w_i, according to how important the user believes they are, thereafter they are combined.

  #band($q_1\ q_2 \ldots q_n$**)** Boolean AND. Matches only if all terms match.

  #filreq($t\ q$**)** Filter require operator. Only documents that match term t are considered and they are ranked according to term q.

- Context restrictor

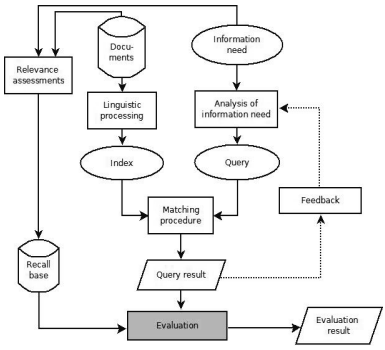
  term.field Field restriction. There is a match if **term** appears in a field named **field**. The field names are determined by the tagging of the documents in the database: TITLE, DESC etc. (see section 10.1).

Figure 17: Operators in the Indri Query Language, where q_i is a term.

The parts of the Lemur Toolkit used for the work behind this thesis was the Indri index builder (see section 10.3) and retriever. They can both be run by a graphical user interface or a command-line interface. The graphical interface was sufficient for indexing, For retrieval, both interfaces were used. The graphical interface gave easy access to the retrieved documents which could be reached by a simple click on a document ID in the result list. Looking at retrieved documents gave quick feedback as to the kind of documents that had been retrieved. The command line interface, on the other hand, had to be used for batch retrieval and to get the results as a ranked list of document IDs in text format. Such a list could be manipulated into the input formats of the evaluation applications such as trec_eval and VisualVectora (see sections 9.2 and 9.4).

9.2 trec_eval

The evaluation tool trec_eval[26] was developed by Chris Buckley for the TREC community and is the standard tool used by the community for ad hoc retrieval (Buckley and Voorhees 2005). It takes, as input, a file with relevance assessments for a number of topics, and a file with query results, that is one or several ranked lists of document IDs and the corresponding topic IDs. It returns values for a set of standard evaluation measures: precision and recall at different document cut off values, mean average precision, number of retrieved relevant documents, and many more. Figure 18 shows the first few lines of the file that trec_eval returns. The measures used are based on binary relevance, but trec_eval supports multi-level relevance scales. The default minimum value for a document to be considered relevant is 1, but this can be changed to the desired level.




```
num_q                    all   1
num_ret                  all   179
num_rel                  all   33
num_rel_ret              all   29
map                      all   0.4061
gm_ap                    all   0.4061
R-prec                   all   0.4242
bpref                    all   0.5859
recip_rank               all   1.0000
num_nonrel_ judged_ret   all   56
exact_prec               all   0.1620
exact_recall             all   0.8788
11-pt_avg                all   0.4070
3-pt_avg                 all   0.4011
```

Figure 18: The first few lines of a trec_eval result. Some examples of values given are: number of returned documents, size of recall base, number of relevant retrieved documents, mean average precision, number of non relevance judged documents retrieved and 11 point average.

[26] <http://trec.nist.gov/trec_eval/>

9.3 The Query Performance Analyser

The Query Performance Analyser (QPA) is a web based tool developed by the Finnish Information Retrieval Expert group (FIRE) at the Department of Information Studies[27] at the University of Tampere in Finland (Sormunen, Halttunen and Keskustalo 2002). QPA is used to visualize and compare the effectivity of individual queries. The user chooses which database to search in, search system and topic, and then enters a query in the formal language of the search engine chosen. The results are shown in a number of different visualizations.

The result page first generated for each query presents a histogram, a pie chart and a document list as shown in figure 19. The histogram represents the retrieved list of documents, 100 documents at a time. The top ranked document is presented at the far left. Following the histogram to the right, each bar represents the next document in the list. Documents that are assessed to have relevance grade 3 are fully colored, documents with assessed relevance grade 2 are colored half way up, and documents of assessed relevance grade 1 are one fourth colored. Finally documents that are assessed to have relevance 0, or have not been assessed for the query, are uncolored.

The QPA supports a 4-graded scale of relevance, but the precision-recall measures are based on a binary scale. As default, relevance degrees 0 and 1 are considered non-relevant and degrees 2 and 3 relevant. The default DCV is 200. Both of these settings can be adjusted to the desired level.

The recall level at DCV 200 is shown in the pie chart.

Under the histogram, the document list contains links to a span of ten retrieved documents. Clicking on a link opens a page with the chosen document. To the left of each document link there is a bar demonstrating the assessed relevance degree. If a user wishes to look at another span of ten documents he or she can click on the horizontal bar of the histogram and the corresponding document span. There is also a link to a randomly chosen missed relevant document. Looking at this document can give hints as to which additional search keys may be used.

The QPA has a search history page, where previous searches are shown for comparison. It also has visualization pages, one which shows histograms from the latest searches, and one which shows precision-recall curves for the results of these searches (see figures 20 and 21).

[27] Now Department of Information Studies and Interactive Media.

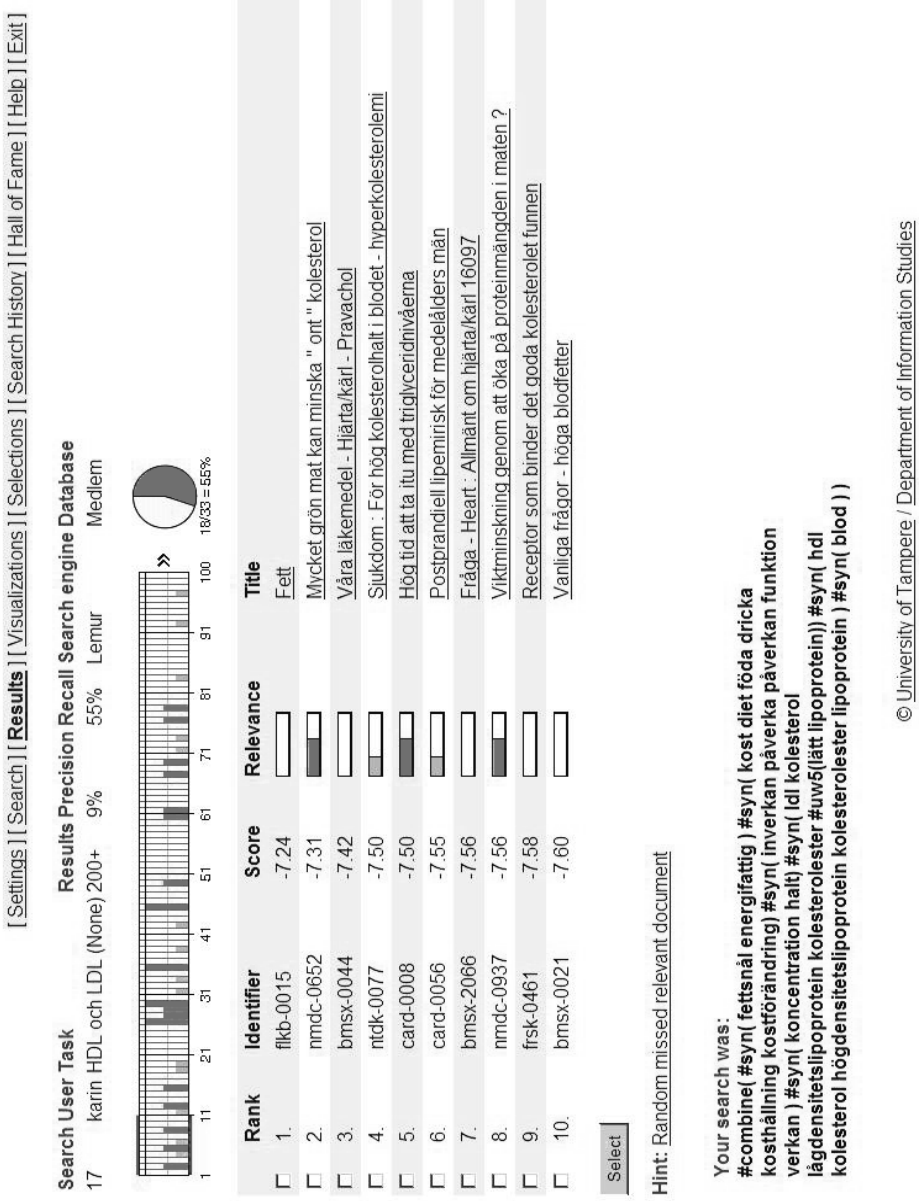


Figure 19: The QPA result page with histogram, here showing the relevance for the 100 highest ranked documents, a pie chart showing recall at DCV 200, and a list of the first 10 documents with relevance bar and link to the document. The query used for the search is also shown.

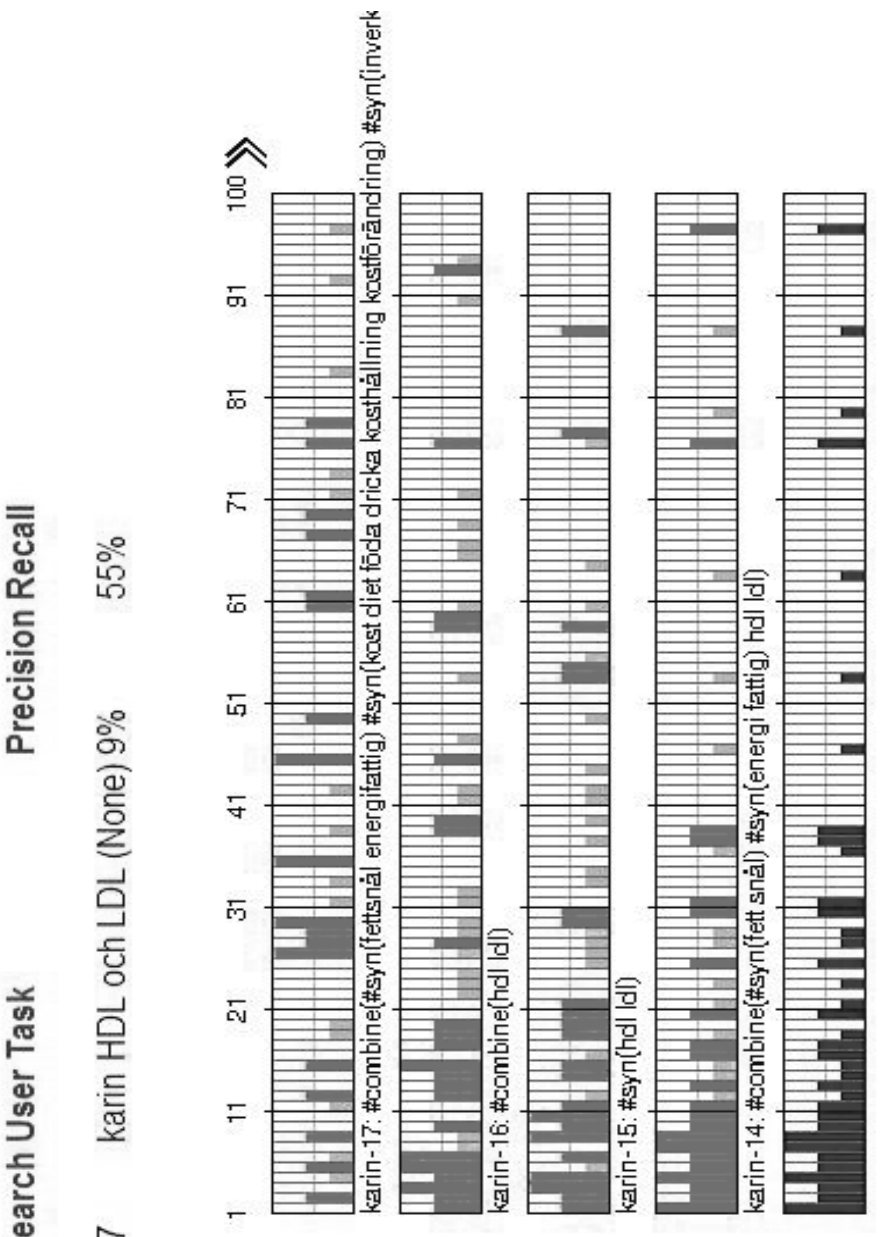


Figure 20: The QPA visualization page with histograms from the four latest runs. The darker histogram at the bottom is an all time high, showing the user's best run so far.

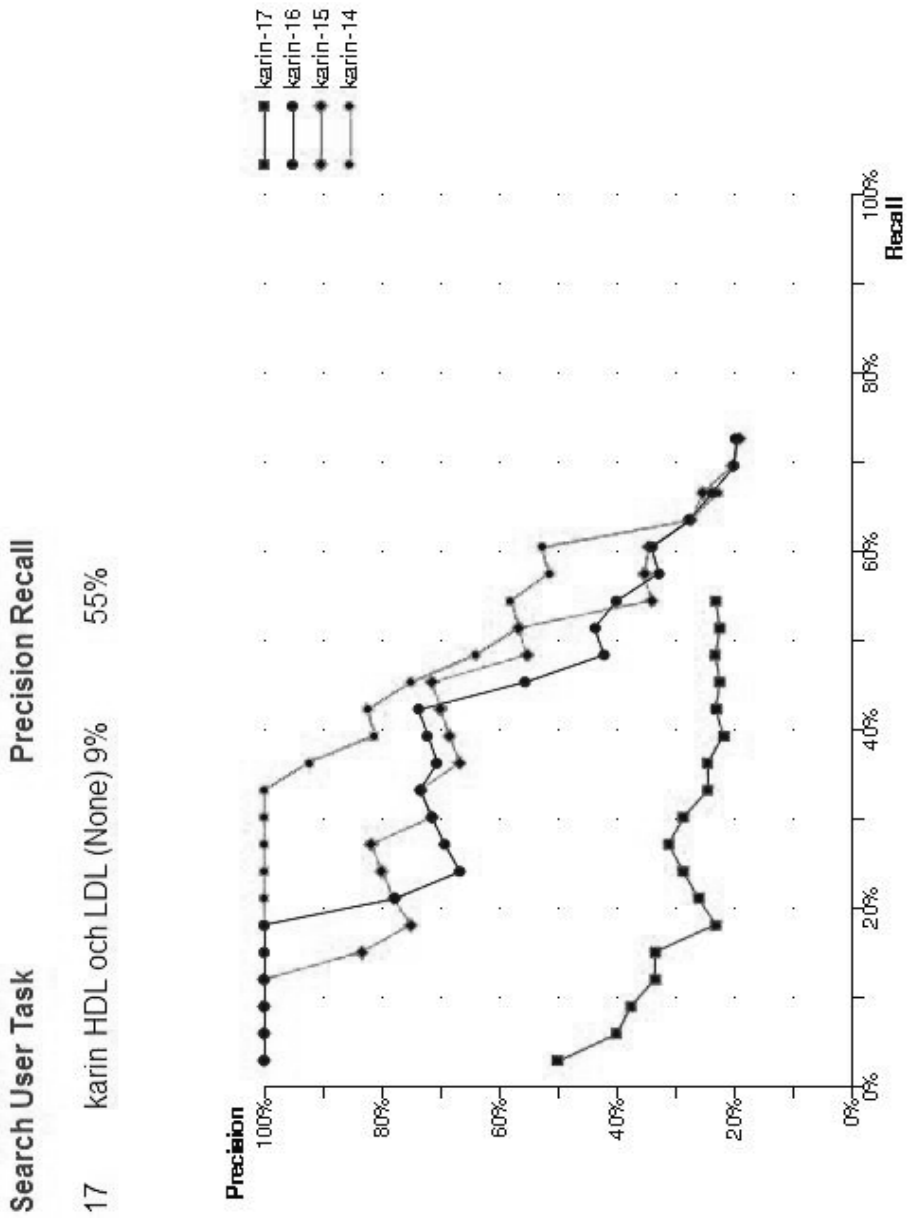


Figure 21: The QPA visualization page with precision-recall curves for the four runs shown in figure 20. On the QPA page the curves are all in different colors and are easy to discern.

9.4 VisualVectora

VisualVectora is a visualization tool, also developed by the FIRE research group at the University of Tampere. It allows the user to visualize retrieval results, calculated by the cumulated gain based evaluation algorithms (see section 3.5). These results can be from runs of individual queries and runs across topics. VisualVectora allows the user to compare runs with several queries for each topic. VisualVectora visualizes results in cumulated gain (CG), discounted cumulated gain (DCG), and their normalized variants, normalized cumulated gain (nCG) and, normalized discounted cumulated gain (nDCG) (Järvelin et al. 2008).

```
1  1  ntdk-0077   1
1  2  nmdc-0652   2
1  2  ltxx-1297   3
1  1  ntdk-0074   4
1  2  ltxx-2525   5
1  2  svrd-1865   6
1  2  mdll-0226   7
1  3  pfzr-0047   8
1  2  nmdc-0937   9
1  2  vrdg-0434  10
```

Figure 22: The first lines of a result file in the format used as input to VisualVectora. The columns contain topic numbers, relevance grades, document IDs, and ranks.

The VisualVectora system takes as input the ranked lists of document IDs, with corresponding topic numbers, relevance grades and ranks (see figure 22). It also takes a file with data about the ideal cumulated gains for each topic. This data is necessary to produce the normalized gain curves. There is one ideal data file for every user scenario, None, Doctors, and Patients, since the relevance grading differs. Each ideal data file contains, for every topic, information about the number of documents that have been assessed to have each relevance grade. The number of documents to have relevance grade 0 is set to 'unlimited'. This is because not only documents assessed to have relevance 0 are counted as 0, but also the unassessed ones. Thus it is not necessary to have a count of the unassessed documents. An excerpt of the ideal data file for user scenario 'None' is shown in figure 23.

Examples of the VisualVectora visualization of the cumulated gain and discounted cumulated gain curves for the baseline query of topic 1 are shown in

```
1 (0 unlimited) (1 53) (2 29) (3 4)
2 (0 unlimited) (1 70) (2 3) (3 1)
4 (0 unlimited) (1 40) (2 26) (3 4)
5 (0 unlimited) (1 5) (2 9) (3 0)
7 (0 unlimited) (1 66) (2 58) (3 8)
```

Figure 23: Data for the ideal cumulated gain in the user scenario 'None' for the first five topics, with topic ID 1, 2, 4, 5, and 7. The data contains the number of documents at each relevance level.

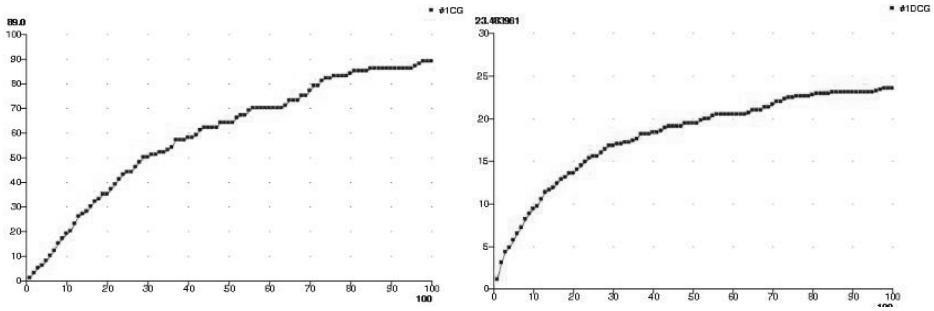


Figure 24: The results for the baseline query of topic 1 shown in cumulated gain (CG) to the left, and in discounted cumulated gain (DCG) to the right, for the 100 top ranked documents. Both curves grow, never turning downwards. The cumulated gain grows faster and reaches cumulated gain levels in the 80s. The discounted cumulated gain levels off and does not grow past the low 20s.

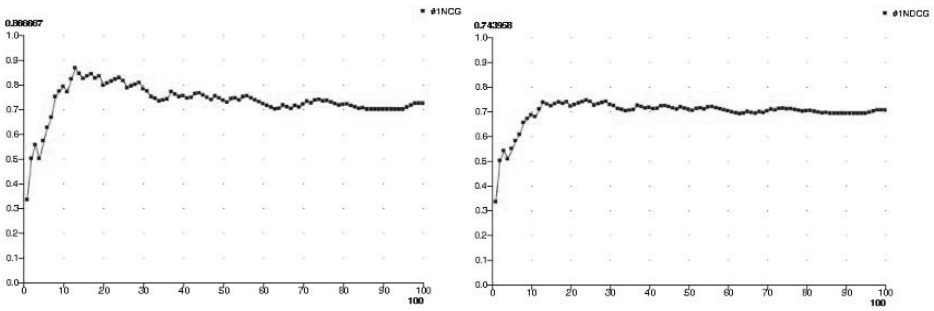


Figure 25: The results for the baseline query of topic 1 shown in normalized cumulated gain (nCG) to the left, and in normalized discounted cumulated gain (nDCG) to the right, for the 100 top ranked documents. These curves are always between 0 and 1 and they fluctuate up and down.

figure 24. The CG and DCG curves keep growing as long as relevant documents are found. The DCG curve levels off earlier than the CG curve, as a smaller and smaller portion of the relevance scores are added to the gain. In the normalized curves, which are shown in figure 25, the cumulated gain and the discounted cumulated gain are compared to the gains of the ideal curves. The values are therefore always between 0 and 1 and the curves fluctuate up and down. Note that all four graphs in figures 24 and 25 represent the same result, although with different evaluation measures.

9.5 MedLex

MedLex is a workbench for lexicographic work, which consists of two main parts: The MedLex lexicon, a lexical database containing 4 500 medical lemmas, and a medical corpus (Kokkinakis 2004). Since 2004, it is continuously being built at the NLP research unit at the Department of Swedish Language, University of Gothenburg.[28]

The MedLex corpus consists of scientific articles from medical journals, teaching material, guidelines, patient FAQs, blogs, health care on-line information etc. In july 2010 the MedLex corpus contained 76 000 documents or 47.7 million tokens,[29] and it is still growing. However, a test collection must have a static set of documents in order to allow re-usability and comparisons of different search strategies. The MedEval test collection is therefore built on a snapshot of the documents of MedLex in October 2007 when MedLex contained 42 000 documents or 15.2 million tokens.[29]

The MedLex corpus is lemmatized and searchable in the KWIC-format (KeyWord In Context). Searches can be conducted using different parameters such as strings, lemmas, part-of-speech tags or semantic tags. The semantic tags carry information such as 'disease', 'chemical substance', 'anatomical term', 'person' or 'measure'.

9.6 MeSH – A medical thesaurus

MeSH (Medical Subject Headings)[30] is a medical thesaurus produced by the National Library of Medicine in the United States. A thesaurus is a type of

[28]Originally MedLex was supported by Vocab AB and Västra Götalandsregionen. Later MedLex was supported by the Semantic Mining Network of Excellence – Semantic Interoperability and Data Mining in Bio-medicine, an EU project with 25 partners in 11 countries of the European Union.

[29]Dimitrios Kokkinakis p.c.

[30]<http://www.nlm.nih.gov/mesh/>

dictionary where words are not organized alphabetically, but according to conceptual or thematic relations. It shows relations between terms, if they are synonyms or a broader or a narrower term for the same concept, or if they are related in some other way. A thesaurus is a controlled vocabulary which can be used as a resource in both the indexing and the retrieval process in information retrieval.

MeSH is organized in a hierarchical structure with eleven levels. The top 16 categories are very broad, for example : **A**: Anatomy, **B**: Organisms, **C**: Diseases, **D**: Chemicals and Drugs, **E**: Analytical, Diagnostic, and Therapeutic Techniques and Equipment, and **F**: Psychiatry and Psychology. Further down in the hierarchy are more specific headings such as 'Muscle Cramp' or 'Heel'. The MeSH hierarchy contains more than 22 000 headings. MeSH does not only have a hierarchical structure but also contains many cross-references that map headings to each other.

The Swedish MeSH[31] is based on a translation of the American MeSH. The original data from U.S. National Library of Medicine has been completed with Swedish translations performed by staff at the Karolinska Institutet University Library.

[31] <http://mesh.kib.ki.se/>

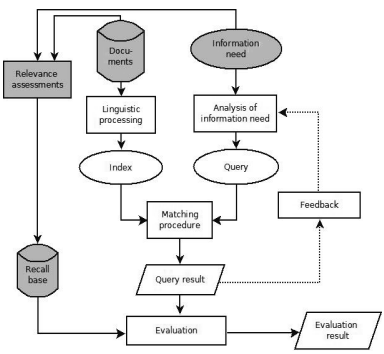
10 CREATING THE MEDEVAL TEST COLLECTION

The Guide *is definitive. Reality is frequently inaccurate.*

The Restaurant... chapter 6

Creating a new test collection is a major undertaking involving a team of people and many hours, or rather months, of work. This is one reason why the number of test collections available is very modest, especially for languages other than English. Before MedEval, there was no medical test collection in Swedish, only a collection created for the TREC conferences containing articles from the daily newspapers Göteborgs-Posten and Helsingborgs Dagblad from the year 1994.

The NLP unit at the Department of Swedish Language at the University of Gothenburg has, through a number of years, been involved in several projects of research in medical language processing. Therefore, as studies in information retrieval were to be initiated, it seemed evident that the research at the department was best served if these studies were also done within the medical domain.



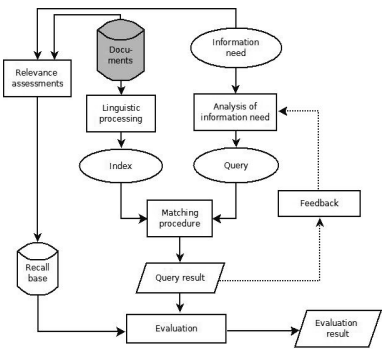
As no Swedish test collection built on medical documents existed, a decision was made to build a new test collection. With a new collection the NLP unit could take control over the architecture of the collection, and make decisions such as using a finer grained scale for relevance judging, which made it possible to employ a wide variety of evaluation tools. The most important decision was to include user groups in the collection, and assess documents, not only for relevance to topics, but also for intended groups of readers or target groups (Friberg [Heppin] 2007; Friberg Heppin 2008, 2009, 2010).

Information on health and medical issues differs in one important aspect from information in other specialized domains. It concerns the whole general public. It is in the interest of all citizens to be able to gather information that is of importance to their health and that enables them to take an active role in their own health care. This is in contrast to information in many other specialized domains, for example the construction domain or the computer domain. You can live a full life without knowing much about the construction techniques of the houses you reside in, or how the insides of your computer work. One of the research questions in this thesis is how documents aimed at experts and at non-experts differ, and if it is possible for a user, by a certain query formulation, to influence the retrieval and ranking of documents so that the documents retrieved are written for the desired target group. If this research can empower people and help them find information written in a way that is suited for them, an important goal has been reached.

As stated above, it takes many person hours to create a test collection, but it is not only time demanding, it is also resource demanding. It would not have been possible to build a collection in the scope of a doctoral thesis, had there not been a set of resources ready and available. The NLP unit has built, and is still building, the MedLex corpus which contains the documents of the collection. They have also created the lemmatization and compound decomposition tools used in the indexing process. These resources will be described later in this chapter.

10.1 The documents

The documents in the MedEval test collection are from the MedLex corpus, which was described in section 9.5. The documents in the collection are of all types of available biomedical information that can be found in an electronic format. This does not include clinical text. The composition of different genres can be seen in table 10.1.



The document collection was tagged in the trectext format where the <DOC>, <DOCNO>, and <TEXT> tags are required. Each individual document is surrounded by <DOC> tags, which delimit the documents from each other. The document ID is surrounded by <DOCNO> tags. The ID consists of four characters identifying the source, a hyphen, and a four-digit running number, one

number series for each source. The <TEXT> tags surround the body of text. If the document is from the internet, the web address is supplied immediately after the text tag. The trectext format also allows for optional tags, such as the <DATE> tag which is used in the MedEval documents around the date of the publication, if it is known. If the date, or part of the date, is not known, it is replaced by Xs or zeros. Including the date of publication makes it possible to search for documents written in a certain period of time, using the field restriction operator (see figure 17 on page 96). The trectext tags used in MedEval are shown in figure 26.

Conforming to a standard format such as trectext facilitates use of publicly available tools such as the Indri/Lemur Toolkit.

Table 10.1: The genres and sizes of the MedEval document sources. The document collection is a snapshot of the MedLex corpus in October 2007 (D. Kokkinakis, p.c.).

Type of source	Number of documents	Percent of documents	Number of tokens	Percent of tokens
Journals and periodicals	8 453	20.0	5.3 mil.	34.6
Specialized sites	14 631	34.6	2.9 mil.	19.1
Pharmaceutical companies	9 200	21.8	2.3 mil.	14.8
Government, faculties, institutes and hospitals	2 955	7.0	2.0 mil.	13.3
Health-care communication companies	4 036	9.6	1.7 mil.	11.3
Media (TV, daily newspapers)	2 980	7.1	1.0 mil.	6.9
Total	42 255	100	15.2 mil.	100

<**DOC**> </**DOC**> The document tags delimit the documents from one another.

<**DOCNO**> </**DOCNO**> The document ID-tags surround the document ID, which is a unique identifier for each document.

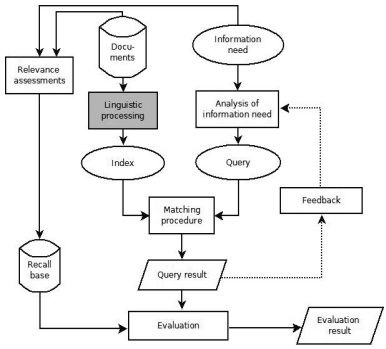
<**TITLE**> </**TITLE**> The title tags surround the title, which is written by the author of the document.

<**DATE**> </**DATE**> The date tags surround the date of publication.

<**TEXT**> </**TEXT**> The text tags surround the main text of the document.

Figure 26: The trectext XML tags used to mark up the fields of the MedEval documents.

10.2 Linguistic processing



Before the documents were indexed, they went through linguistic processing in order to obtain the word forms that should be in the index entries. As information retrieval is about matching document terms in the index with topic terms in the queries, it is a good idea to let the topic terms go through the same linguistic processes so that they will match the index terms.

In the linguistic processing, the MedEval documents were tokenized, converted to lower case and lemmatized. MedEval has two separate indexes. For one of these, the compounds were also decomposed into constituents.

To demonstrate the linguistic and indexing processes, we will follow the fictive documents 1 through 4. These documents are shown in figure 27. The fictive documents are taken through the linguistic and indexing processes in the figures that follow. Comments as to how the different steps in the processes relate to the indexing of MedEval are made. The fictive documents represent the documents of the MedEval collection.

1	Information retrieval is about finding documents that satisfy information needs.
2	Terms are stored in indexes pointing to their position in the documents.
3	The user poses a query describing her information need.
4	Terms in the queries are matched against the indexed terms of the documents.

Figure 27: The fictive documents 1-4 represent the documents in the MedEval collection. They will be taken through the indexing process in figures throughout the following sections.

10.2.1 Tokenization

Tokenization is the process of recognizing the terms in a text. This is done by delimiting terms from each other and separating sentence delimiters, quotation marks, etc. from the terms. Characters, such as quotation marks, which are not necessary for the application in question, in this case indexing, can be removed.

Tokenization is performed to obtain suitable units to process. The MedEval tokens were to be turned into lists of terms to be used as index entries. Tokenization also implies deciding what is a term. For example, multiword terms *yellow fever* can be tokenized as either one term or two: yellow fever or yellow fever. However, there are no multiword tokens in the MedEval index. If a user would like to search for a multiword concept, such as *yellow fever*, this is still possible with the use of a proximity operator.

```
<DOC>
<DOCNO> FLKB-0004 </DOCNO>
<TITLE> Cell , vävnad , kroppens organisation </TITLE>
<DATE> 2006-04-XX </DATE>
<TEXT> http://www.folkbildning.net/ [...]
Någon gång för drygt tre miljarder år sedan föddes den första
cellen . Den uppstod under speciella betingelser i urhavet .
Troligen bildades först s.k. smarta molekyler i form av RNA (
ribonukleinsyra ) . Senare uppstod DNA ( deoxiribonukleinsyra ) ,
en spiralformad molekyl uppbyggd av kolhydrat , fosfat och
kvävebaser . Det är också möjligt att de första DNA-molekylerna
spreds som ett smittämne från någon annan plats i rymden där
levande organismer redan fanns . För att cellen skulle överleva ,
och dessutom trivas , var det viktigt att miljön , d.v.s. det
urhav som den skapades i , kunde bibehållas . Det var viktigt att
temperatur , saltkoncentration och ph-värde var konstant [...]
</TEXT>
</DOC>
```

Figure 28: The figure shows an example of a MedEval document which is tokenized and tagged in the trectext format. The ID, title, date of publication and text body are marked with XML tags.

Without tokenization, the last term in the first proper sentence in the TEXT field of the document of figure 28 would be '*cellen.*', that is, the last character of the term would be a period. This would not be an appropriate unit to use neither in the index nor when counting word length or word form frequencies.

Tokenization is often a complicated undertaking (He and Kayaalp 2006). Many decisions that are made, for instance how to treat hyphens, periods, slashes, and apostrophes, result in undesired side effects. The punctuation in the MedEval documents was preserved at the sentence boundaries, but were separated from the terms by whitespace. The Indri Indexer ignored these characters so they did not have to be explicitly removed. In cases where a period, a hyphen, etc. was an integral part of a term, it was not separated from the term, for instance in common abbreviations such as *d.v.s* 'i.e.' and *s.k.* 'so-called' in figure 28. These terms were kept intact.

Slashes immediately surrounded by characters were kept, leaving strings like *och/eller* ‘and/or’. This type of string would later be interpreted by the compound splitter as a compound and thus included in the compound frequency count (see table 14.1). The MedEval collection contains compounds where one constituent is a multiword unit, for example *New York-börs* ‘New York Stock Exchange’. These are compounds that, contrary to the criteria of Swedish compounds, are not orthographically written as one word. The tokenizer interprets *New* as independent, but *York-börs* as a unit. This interpretation will not affect the indexing as the Indri indexer, in the version used, treats non-alphanumerical characters as whitespace. *new york-börs*[32] will thus be indexed as three separate terms. However, the string *york-börs* will be counted as a compound in the compound frequency counts.

Figure 29 illustrates the fictive documents 1-4 after tokenization. The periods are here separate tokens, but they will be ignored by the Indri indexer.

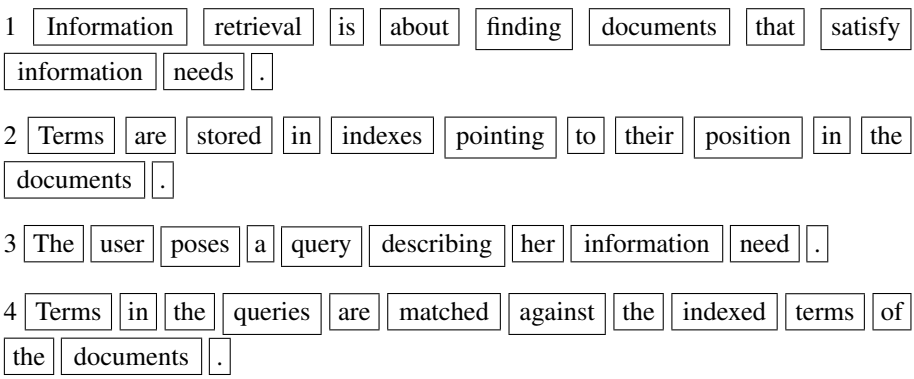


Figure 29: The fictive documents are tokenized. Each token represents one indexing term.

10.2.2 Lemmatization

Lemmatization in information retrieval is the process of finding a lemma to use as a common representation for all forms of a lemma sign which occur in the documents of a collection. A more thorough description of lemmatization is found in section 4.2.1.

The lemmatizer used on the MedEval documents was developed by Kokkinakis and Johansson-Kokkinakis at the NLP unit, University of Gothenburg, and is described in Kokkinakis 2001. The lemmatizer is tag-dependent, and

[32] All characters were converted to lowercase before indexing.

lemmatizes words first annotated by a part-of-speech tagger. If a verb, noun, or adjective tag is found, the lemmatizer goes on to match the characters of the word to strings described by regular expressions in the lemmatization rules. Examples of these rules are shown in figure 30. According to these rules the verb *antändas* 'ignite PASSIVE' would be mapped to the infinitive *antända*, the noun *dj:arna* 'dj PLUR DEF' would be mapped to the singular nominative form *dj* and the adjective *vuxnas* 'adult POS PLUR GEN' to the positive, singular, nominative form *vuxen*.

```
UPPERCASE [\101-\132\300-\335]
lowerCASE [\141-\172\340-\375]
ANY {UPPERCASE}|{lowerCASE}

<verb>{ANY}*[svtl]ändas     {stem(1,‘‘’’);}
<noun>{ANY}+:(a|e)rna       {stem(5,‘‘’’);}
<adjective>{ANY}+uxn(a|e)s  {stem(3,‘‘en’’);}
```

Figure 30: Examples of lemmatization rules from the lemmatizer built by Kokkinakis and Johansson-Kokkinakis.

In the rules of figure 30 <verb>, <noun>, and <adjective> are start conditions. A string is lemmatized only when it is part-of-speech tagged and the tag matches one of the start conditions. Furthermore, the string itself must match one of the corresponding regular expressions. The variable ANY stands for any combination of upper- and lowercase characters according to conditions stated in the first lines. If a match is found, the function stem reads the string, removes, from the end of the string, the number of characters denoted by the first argument and adds the string of the second argument.

A successful tagging can disambiguate homographs of different parts-of-speech. However, for homographs with the same part-of-speech, specific rules are necessary for good performance. For instance, the form *satt* is the imperfect tense of the verb *sitta* 'sit' and the perfect participle and supine form of the verb *sätta* 'put'. For compounds, where a certain modifier can only be combined with one of these lexemes, rules are composed to point the lemmatizer to the correct lemma. For example, *iscensatt* 'in scene put' (produced), *ifrågasatt* 'in question put' (questioned), and *insatt* 'in put' (initiated) should all have the form *sätta* as the lemma of the base constituent.

```
<verb>{ANY}*(iscen|ifråga|in|...)satt  {stem(3,‘‘ätta’’);
```

Figure 31: The rule which maps compounds with the head *satt* to the correct lemma: **sätta*.

There are a handful of Swedish inflectional endings which are problematic since they are common character combinations or even homographs with short words, for example *-en* 'singular definite' or 'plural definite' noun ending, *-ar* 'plural indefinite' noun ending or 'present tense' verb ending, and *-or* 'plural indefinite' noun ending. As independent lexemes they mean 'juniper', 'are' (119.6 square yards), and 'mite', respectively. Without context, it is not possible to tell if *häcken* is the equivalent of 'the hedge' or 'hedge juniper' or if *stenar* is the equivalent of 'rocks' or if it is a measure of land covered with rocks. In the present study, a couple of lemmatization mistakes were encountered. For instance *lår||ben* 'thigh bone' was interpreted as *lårb+en*, the definite form of the nonsense word *lårb* and *pollen* 'pollen' as *poll+en* the definite form of the nonsense word 'poll'.

Kokkinakis reports results for the lemmatizer where the F-value is 96.6% for nouns, 99.3% for verbs, and 97.6% for adjectives and participles.

Figure 32 shows the fictive documents 1-4 now also converted to lower case and lemmatized, while figure 33 shows an example of a MedEval document which has been tokenized, converted to lower case and lemmatized.

1 information retrieval be about find document that satisfy information need .

2 term be store in index point to their position in the document .

3 the user pose a query describe her information need .

4 term in the query be match against the index term of the document .

Figure 32: All letters in the tokenized terms are converted to lower case and the terms are lemmatized.

10.2.3 Decomposition

As MedEval has two indexes, where one contains split compounds, decomposition of compounds is part of the linguistic process of the test collection.

The decomposition of compounds for the segmented index was done with a heuristic domain-independent finite state based compound splitter developed by Kokkinakis (2001, 2009). The splitting algorithm is based on the ideas of Brodda (1979). Brodda states that, in compounding languages such as Swedish, when two words are combined in a compound, the result is often

```
<DOC>
<DOCNO> FLKB-0004 </DOCNO>
<TITLE> cell vävnad kropp organisation </TITLE>
<DATE> 2006-04-xx  </DATE>
<TEXT> http://www.folkbildning.net ...
någon gång för drygt tre miljard år sedan föda den första cell
den uppstå under speciell betingelse i urhav troligen bilda först
s.k. smart molekyl i form av rna ribonukleinsyra . senare uppstå
dna deoxiribonukleinsyra , en spiralformad molekyl uppbyggd av
kolhydrat fosfat och kvävebas det vara också möjlig att de första
dna-molekyl sprida som ett smittämne från någon annan plats i
rymd där levande organism redan finna för att cell skola överleva
och dessutom trivas vara det viktig att miljö d.v.s. det urhav
som den skapa i kunna bibehålla det vara viktig att temperatur
saltkoncentration och ph-värde vara konstant [...]
</TEXT>
</DOC>
```

Figure 33: An example of a pseudotext put to the indexer for the non-decomposed index. It has been tokenized, converted to lower case, and lemmatized.

a cluster of letters, mostly consonants, in the segmentation point, that are of a kind which is not allowed internally in simplex words, for example the cluster ***rkskr*** in *kor**k**||**skr**uv* 'cork screw'. Brodda describes a six level hierarchy of consonant clusters, from those that say nothing about the probability of a segmentation point, like ***ll***, to those that always signal a segmentation point, like ***ntst*** as in *ka**nt**||**st**en* 'edge stone' (curbstone) or ***gkn*** as in *stu**g**||**kn**ut* 'cottage knot' (cottage corner).

Kokkinakis' compound splitter is a non-lexical, qualitative splitter based on the distributional properties of graphemes, identifying grapheme combinations, and through them possible compound boundaries. The splitter is based on a list of bi-, tri- and fourgram character sequences from several hundreds of lemmatized simplex words. This list should contain most grapheme combinations which are allowed in non-compound words. Kokkinakis used this list to generate a new list of bi-, tri- and fourgrams, this time of combinations which were not in the original list of allowed combinations. The new list thus contained combinations which were not allowed in simplex words, and therefore indicated possible compound boundaries.

The grapheme combinations that signaled potential compound boundaries were arranged into groups of two to eight characters. An example of a two-character cluster was *sg* which can be found in compounds such as *viru**s**||**g**enom* 'virus genome' and *fibrinoly**s**||**g**rupp* 'fibrinolysis group'. Four-character clus-

ters were, for example, *ngss* and *gssp* which can be found in *sväljn**ings*‖*s**vårighet* ‘swallowing difficulty’ and *mässlin**gs*‖*sp**ecifik* ‘measles specific’, respectively.

Compounds with very short constituents, two or three characters long, were given special attention by adding characteristic contexts, usually 4-6 characters long. These short constituents were for example *tå* ‘toe’, *yt* ‘surface’, *lår* ‘thigh’, *hår* ‘hair’, *sår* ‘wound’, *tum* ‘thumb’, *syn* ‘sight’, *hud* ‘skin’, *gen* ‘gene’.

With some manual adjustments, such as checking that certain patterns are followed by non-empty characters, or that specific characters do not follow other patterns, the splitter achieved over 96% accuracy. An example of a segmentation rule with manual adjustment is seen figure 34. It is the case of the combination `spla` where it is required that this string should not be followed by `r`, `t`, `d`, or the empty character. This restriction is there to avoid splitting inflected forms of the verb *haspla* ‘reel/coil’. In the example the function `segment` puts the segmentation point after position 1 of the matched string.

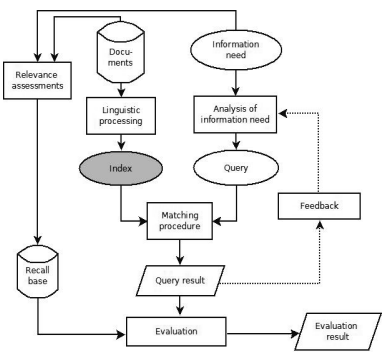
```
spla[^rdt ]   {segment(1);}
```

Figure 34: The rule which puts a segmentation marker after position 1 of the matched string `spla` with the restriction that the matched string is not followed by `r`, `t`, `d`, or the empty character.

When processing the input to the splitter to find candidate compounds, the words are scanned from left to right in order to identify grapheme combinations that signal potential compounds.

10.3 Indexing

The data structure used to represent the documents in MedEval is the inverted index, which is common in modern search engines. The idea behind the inverted index is the same as for the index of a textbook. The index is arranged by terms, and each index term entry is followed by a list of references to where that term can be found. Each entry is called a **posting** and each reference to where a term is found is called a **pointer**. Apart from terms and pointers, the postings may contain a variety of optional information, such as collection frequency or term frequency in individual doc-



```
<DOC>
<DOCNO> FLKB-0004 </DOCNO>
<TITLE> cell vävnad kropp organisation </TITLE>
<DATE> 2006-04-xx 2006-04- xx  </DATE>
<TEXT> http://www.folkbildning.net ...
någon gång för drygt tre miljard år sedan föda den första cell
den uppstå under speciell betingelse i urhav troligen bilda först
s.k. smart molekyl i form av rna ribonukleinsyra ribo nuklein
syra . senare uppstå dna deoxiribonukleinsyra deoxiribo nuklein
syra , en spiralformad spiral formad molekyl uppbyggd upp byggd
av kolhydrat kol hydrat fosfat och kvävebas det vara också möjlig
att de första dna-molekyl dna- molekyl sprida som ett smittämne
smitt ämne från någon annan plats i rymd där levande organism
redan finna för att cell skola överleva över leva och dessutom
trivas vara det viktig att miljö d.v.s. det urhav som den skapa i
kunna bibehålla det vara viktig att temperatur saltkoncentration
salt koncentration och ph-värde ph- värde vara konstant [...]
</TEXT>
</DOC>
```

Figure 35: An example of a pseudotext put to the indexer for the decomposed index. It has been tokenized, converted to lower case and lemmatized, just as for the first index, but the compounds were also decomposed. In this document the date, containing the character sequence 'xx' representing an unknown day, and not numerals as in the date format, has been treated as a compound and been decomposed.

uments. Having this information in the index facilitates the calculation of tf∗idf values at query time. The postings often contain the positions of the terms in the documents, something which makes it possible to search for phrases or to make conditions such as that terms must appear within a certain distance of each other or in a certain order. The postings sometimes also contain formatting information, for example if the term was emphasized in some way.

The process of constructing an inverted index depends on what information one wishes the index to contain. In a full form index, the entries represent the actual word forms written in the documents, in contrast to a lemmatized index where the entries represent the corresponding lemmas.

Figure 36 illustrates how the tokens of the fictive documents 1-4 have been converted into a list of index entries with pointers to the documents where they occur. Note that the noun 'index' and the verb 'index' are represented by the same index term. In figure 37 the entries have been completed with information about the term frequencies in the different documents. The index in figure 38 has pointers with information about the positions of the terms in the documents, information which is necessary when using proximity operators. In

this index the term frequency in the documents is implicitly given by the length of the list of pointers for each term and document.

In the segmented index of MedEval, the compounds are indexed as one orthographic word and also by the compound constituents. For instance, *saltkoncentration* 'salt concentration' which is indexed as *saltkoncentration*, *salt*, and *koncentration*. To see the difference, compare figures 33 and 35 which contain part of a document in the pseudotext form sent to the indexer.

A consequence of adding the compound constituents as new tokens is that a compound formed from two lexemes has three postings in the index, and a compound formed from three lexemes has four postings, each with a separate position indicator. This will affect the term weighting since the term count in the individual documents and in the collection will be higher than in the index without segmentation. It will also affect the impact of queries constructed with proximity operators which the user uses to state that terms must appear in a document within a certain distance of each other. However, with the compound parts occupying token positions of their own, the distances according to the indexes will be greater than what they are in the original documents. A better solution could have been to give the postings for a compound and the corresponding compound parts one and the same token position, so that *saltkoncentration*, *salt*, and *koncentration* all would have the same position number. This was not possible with the Indri Index Builder.

After the linguistic preprocessing, the documents were sent to the Indri Index Builder, shown in figure 39. The input to the indexer consisted of the name and the path to the new index, the paths to the processed document files, the names of the fields in the documents that should be indexed: TITLE and TEXT, and the format of the documents: trectext. There is an option to supply one's own stop word list by indicating the path to the stop list document, and there is also the option to choose the Krovetz or Porter stemmer. However, no stop word list was used for MedEval, as this would hamper the possibilities to do phrase searches. The stemmers, supplied by the indexer, were not used either, as these are built for the English language. The documents were instead, as described before, lemmatized beforehand.

The Indri index builder constructs indexes with information about the position of the terms in the individual documents in a similar manner as the index in figure 38. This allows proximity searches and phrase searches.

The index in figure 38 has information about term positions. This makes it possible to make a query using the proximity operator **#od1()** to make the restriction that the term *information* must be in a position immediately before the term *retrieval*: #od1(information retrieval). This would retrieve document 1, where this combination is found, but not document 3 where the term immediately following *information* is *need*.

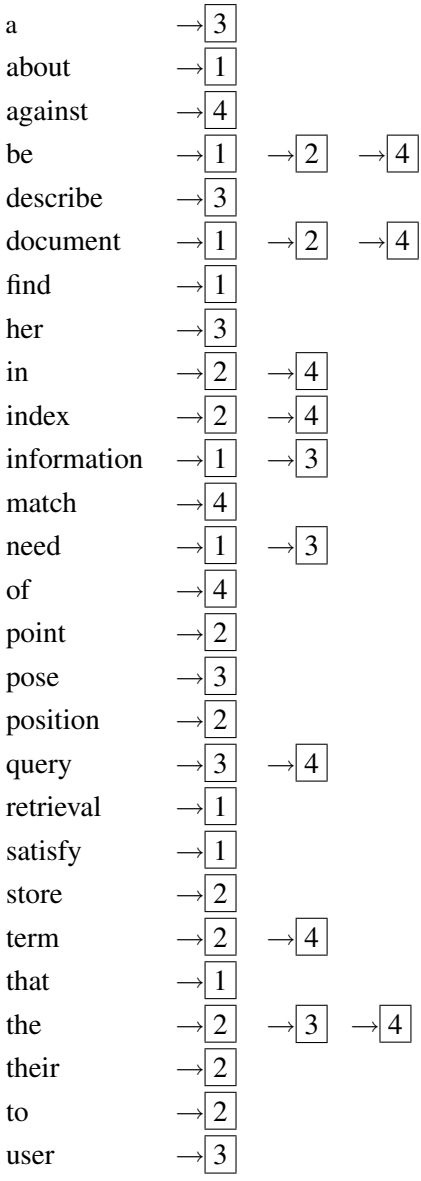


Figure 36: An inverted index with pointers to the documents where the terms occur. Note that the noun 'index' and the verb 'index' are represented by the same entry.

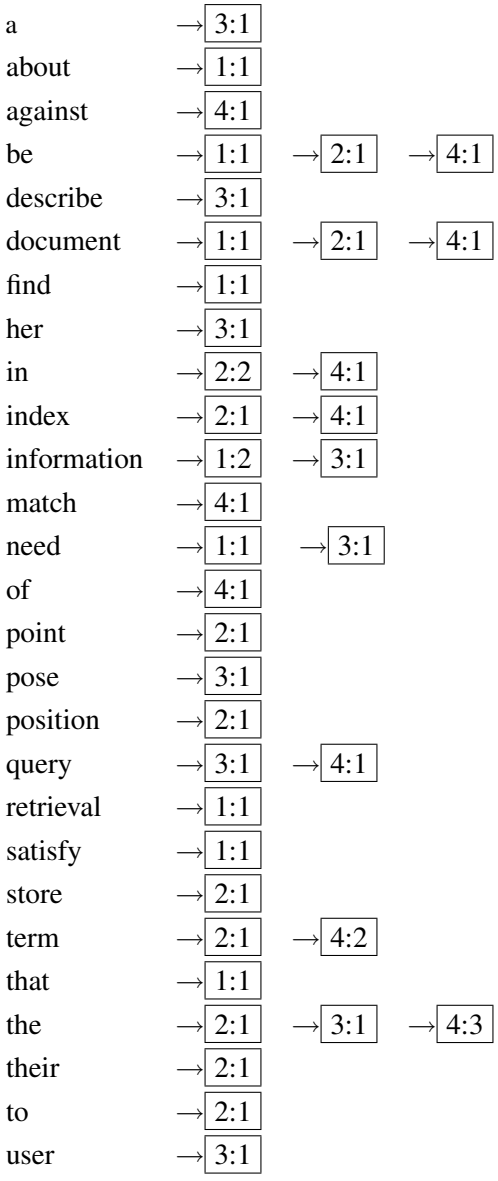


Figure 37: An inverted index where the pointers contain counts for the term frequency in each document.

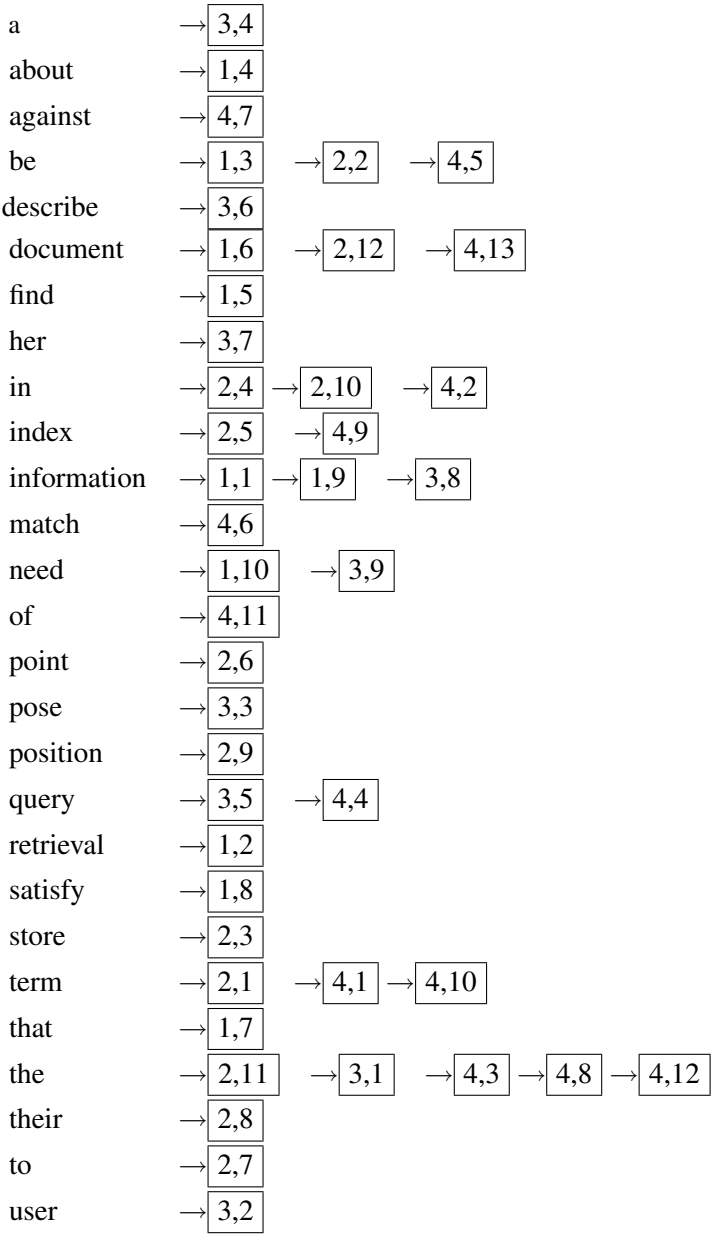


Figure 38: An inverted index where the pointers have information about the positions of the terms in the documents. The information about term frequency in the individual documents is here given implicitly by the length of the lists of pointers for each term and document.

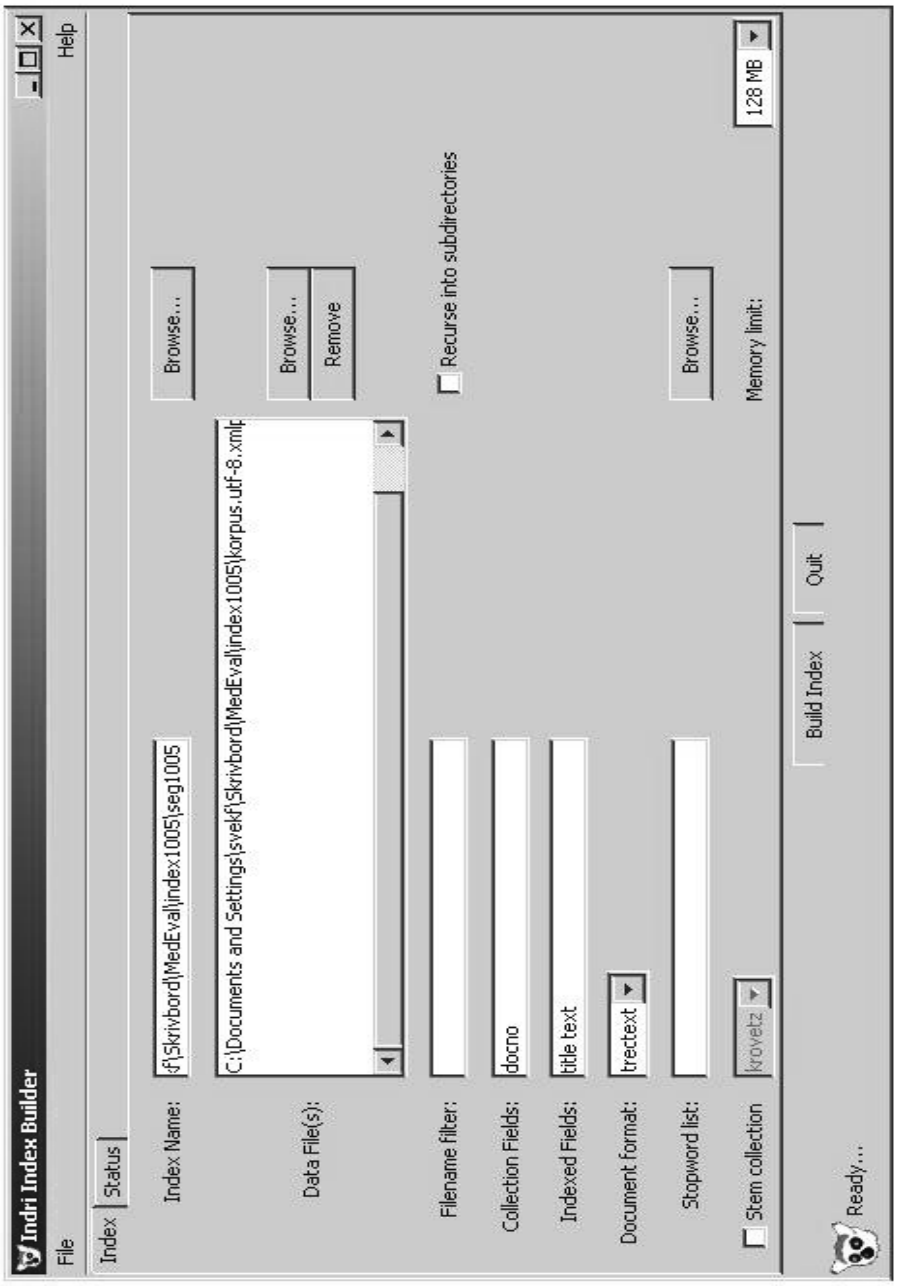
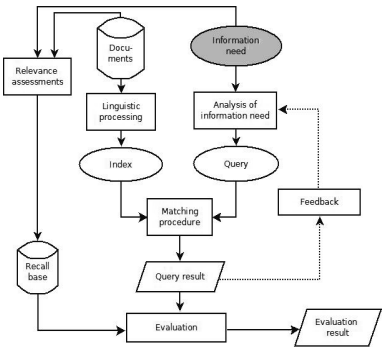


Figure 39: The graphic user interface of the Indri Index Builder.

10.4 Topic development

As MedEval is a medical test collection, all information needs are from this domain. To come up with realistic needs that could be asked for in real medical situations, topic creators with medical knowledge were employed. Two medical students, in their fourth year of studies, were hired to perform this task. A hundred topics were created in the first stage. Some were not used, as a sufficient number of relevant documents could not be found. Others were merged, for example if they differed only by a patient or doctor point of view. In the end, 62 topics were used in the MedEval test collection.



The process of developing information needs or topics is carried through in several steps. The procedure used for MedEval, which is described below, is inspired by INEX 2006 Guidelines for Topic Development (Larsen and Trotman 2006) and descriptions of the TREC collections (Voorhees and Harman 2005).

The topic creators were asked to familiarize themselves with the collection and the documents in it. After getting an overview of the collection, the creators wrote a list of short phrases describing possible information needs. These phrases, if used, would eventually be the topic titles.

Having candidate title phrases, the next step was to explore the collection again, now more thoroughly, to see whether the topics written were suitable to enable future assessors to consistently judge and grade documents for relevance. The topic creators performed trial runs with the Indri/Lemur search engine. These trial runs helped them to decide the exhaustivity of the information needs. There had to be a suitable number of relevant documents for each need. The estimated final number could vary, but should not be lower than five, preferably over ten and up to fifty or more.

Topic creators cannot determine for certain the number of existing relevant documents for the needs, and this is not the point. Test searches were conducted to get an indication of the amount of existing relevant documents. A simple search key query to the Indri/Lemur system should not give more than in the order of thirty hits among the top hundred. Too many hits would be an indication that the query was too general. If a basic query for a topic results in very high precision and recall, there is no room to test possible improvements with different search algorithms or search queries. Likewise, it is important for creators to check that relevant documents for every need actually exist. If a

topic had fewer than two or more than thirty relevant documents among the top hundred retrieved, the creators were instructed to abandon or revise the topic.

After assessing the top documents among the ones retrieved with the test queries, the creators could use the contents of these documents as feedback to modify the topic if the number or retrieved relevant documents was not in the suitable range. They could, for example, add parameters to make the topic more concise, or remove or modify parameters to make the topic more general.

When the titles and the main ideas of the topics were ready, it was time to write the narratives. A narrative is several sentences long and stipulates what makes a document relevant or not. The information here should be clear and concise. There must be no room for misunderstanding. The narrative can explain in what context the information should be searched for and why. The narrative contains the only guidelines in judging whether a document is relevant or not.

The last part of the topic to be written was the description. The description is a natural language interpretation of the information need, written in one or two sentences, usually a question or a request. The description can be regarded, in essence, as the information need itself.

Finally, all parts of the topics were revised. Title, description and narrative must all describe the same need. There must be no information in the description that is not in the narrative, and no information in the title that is not in the description. This does not entail that the wording must be the same. On the contrary, correctly used synonyms can be of help for a user when constructing queries. In the end it is always the narrative that determines when a document is relevant and when it is not (Larsen and Trotman 2006).

The topics were converted to XML format in the TREC style, just as the documents were. The topic tags can be seen in figure 40. Each topic was provided with tags for topic number, title, description, and narrative. This makes it easy to use the different parts of the topics separately, in queries, to test the impact of having different query lengths. The topics tags are shown in figure 40. Figure 41 shows an example of a MedEval information need, with translation into English. All topics of the MedEval test collection can be found in appendix A.

10.5 Relevance assessments

The thing that makes a test collection different from an ordinary text corpus is not only that it contains specified information needs. More important is that it contains a set of known relevant documents for each information need.

<**TOP**> </**TOP**> The topic tags delimit the topics from one another.

<**TOPNO**> </**TOPNO**> The topic ID tags surround the topic number. This is a unique identifier for each topic.

<**TITLE**> </**TITLE**> The title tags surround the title which usually is a phrase summarizing the information need.

<**DESC**> </**DESC**> The description tags surround the description which contains concise information about the topic, described in natural language. It is usually in the form of a question or a request.

<**NARR**> </**NARR**> The narrative tags surround the narrative which explains in detail what makes a document relevant or not relevant to the topic. The narrative alone should give all information necessary for the relevance assessments.

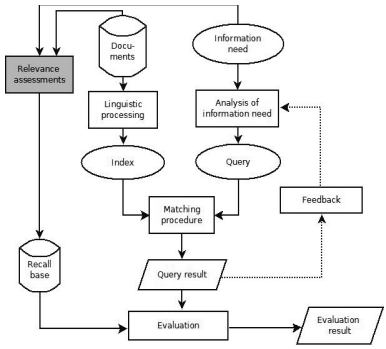
Figure 40: The TREC style XML-tags used to mark up the MedEval topics.

```
<TOP>
<TOPNO>7</TOPNO>
<TITLE> Biverkningar vid cellgiftsbehandling av cancer </TITLE>
<DESC> Vilka biverkningar kan man räkna med vid behandling av
cancer med cellgifter? </DESC>
<NARR> Relevanta dokument innehåller information om
cellgifter/cytostatika, deras biverkningar och vilka typer av
cancer som bör behandlas med dessa. Beskrivning av strategin vid
cellgiftsbehandling är relevant. </NARR>
</TOP>


<TOP>
<TOPNO>7</TOPNO>
<TITLE> Side effects of cancer chemotherapy </TITLE>
<DESC> Which side effects can be expected when treating cancer
with chemotherapy? </DESC>
<NARR> Relevant documents contain information about
chemotherapy, its side effects and which types of cancer should
be treated this way. Description of the strategy in treatment
with chemotherapy is relevant. </NARR>
</TOP>
```

Figure 41: An example of an information need, topic 7, where the whole topic and the ID, title, description, and narrative are tagged. The information need is first given in Swedish, thereafter in an English translation.

These sets of documents have become known through relevance assessments made by humans. It is because of these known relevant documents that we can perform laboratory research in information retrieval as we know it: the Cranfield paradigm. Having information about which documents should be retrieved makes studies efficient in time. It is not necessary to read through the retrieved documents for each run in order to know whether the run was effective or not. Instead it is sufficient to look at the relevance scores of the retrieved documents. This also makes documentation of the success of the runs consistent, as the relevance score for a certain document always remains the same.

Unfortunately, making relevance assessments is very time consuming, and as it is done by humans who do not give their time away for free, it also comes with high economic costs. Therefore test collections rarely contain as complete judgments as could be desired.



An ideal test collection would have a complete set of relevance judgments with every document assessed for relevance to every information need. With a collection of over 42 000 documents and with 62 information needs, as there is in MedEval, taking an estimated average of 8 minutes to assess each document, working 40 hours a week, it would take four persons over 42 years to finish the assessments. Already during the MEDLARS experiments, in the 1960s, and with the elaboration of the ideal test collection in the 1970s (see sections 3.1.2 and 3.1.3), researchers realized that complete relevance assessments were not feasible. During the years since then, methods have been developed to extract subsets of documents for each topic, containing the documents which are believed most likely to be relevant to that particular topic.

10.5.1 Pooling

In the TREC experiments, and for other large test collections, the subsets of documents that were assessed were selected by a method called **pooling**, which implies extracting documents with a high probability of being relevant to each topic in a series of different runs. The term 'pooling' is most often used when documents are collected using different search engines. However, one can discern **system pooling** from **query pooling**, the latter being pooling done with one search engine, but with several queries. In the pooling process, the *k* top-

ranked documents are pooled for the topic in question from each run. The name for the measure of k is **pool depth**.

Extracting subsets of documents to assess by some kind of pooling makes the work load reasonable. However, one must remember that all relevant documents may not have been assessed. Documents can exist which are relevant, but which have not been marked as such. To catch as many of the relevant documents as possible it is important to use different strategies in the extraction of the documents.

Throughout the history of information retrieval the amount of manual work required to create reliable test collections has been an issue under discussion. Pooling was the first step in reducing the amount of work. TREC, having many participants using a number of different search engines, managed to have large pools. For smaller research groups this is not an option. Cormack, Palmer and Clarke (1998) present two successful methods for reducing the number of documents in the pool, while still obtaining a reliable result. 'Move-to-front' pooling still relies on using a number of systems. The number of assessments is reduced by adding more documents to the pool from successful performances and fewer from those that were not as successful. The other method they call 'interactive searching and judging'. No formal strategy is used. Each assessor is asked to continue judging the results of a query as long as it seems fruitful. When the assessor deems that no more relevant documents will be found, a new query is posed. This is repeated as long as the assessor thinks it is meaningful.

Sanderson and Joho (2004) take the method of interactive searching and judging a step further. They noticed how few of the individual runs in the experiments performed poorly. In their study they compare individual runs, both manual and automatic, with the official sets of relevant documents from TREC. In comparison with TREC8 both manual and automatic runs have only 15% of the runs showing considerable differences, while 77% of the manual runs and 49% of the automatic runs show correlations with TREC8 which suggests that they should be treated as effectively equivalent according to measures presented in Voorhees 2001. Sanderson and Joho (2004) do not want to dispute the standard pooling process such as the TREC pooling. Their aim is to open up possibilities for researchers with limited resources and limited time to construct test collections. The approaches that they comment on in the quotation below are, in addition to the one described, 'move-to-front' pooling and 'interactive searching and judging'.

> Third, a new approach is explored where the ranked output of a single automatic search on a single retrieval system is assessed for relevance: no pooling whatsoever. Using established techniques for evaluating the quality of relevance judgments, in all three cases, test collections are formed that are as good as TREC. (Sanderson and Joho 2004: 282)

Since there was limited time and economic resources creating MedEval, the extraction of documents was done on a smaller scale with only one search engine, namely Indri/Lemur. Query pooling, with pool depth 100, was employed when extracting the documents to assess.

Four different search methods were used in the MedEval extraction, that is, four runs for every information need. For each run, the 100 documents ranked most likely to be relevant were extracted, if in fact so many were retrieved. This means that for every need there should be between 100 and 400 documents to assess. 100 if every search method has exactly the same 100 documents ranked highest, and 400 if the four search methods each retrieve at least 100 documents and also do not have any documents in common in the top 100 positions.

For about half of the information needs, a number of documents were assessed in addition to the ones collected by the four searches mentioned above. The reason was technical problems during a first round of searches. The process of collecting and assessing documents had to start over. The new extraction process rendered ranking lists slightly different from the first round of lists. Following expert advice from the FIRE research group at the University of Tampere, it was decided to keep documents retrieved and assessed in the first searches in the pool, with the motivation that it is always better to have as complete a set of assessed documents as possible. In the second round, the same assessor as in the first round judged the documents in each topic pool.

As can be seen in table 10.3 on page 136, the number of documents assessed for each information need was between 115 and 358. A low number such as 115 suggests that more varying search strategies could be in order. Hopefully, in the future, there will be opportunity to complete the present searches with additional ones done with different search engines and with even more differentiated searches. A way to keep consistency in the assessments but still not waste the work already done by the current assessors, could be to set apart the documents graded with relevance score 0 and not reassess these. The documents with some degree of relevance should be reassessed along with new documents in the pools. Making a new round of relevance assessments would also give a good opportunity to compare the judgments of different assessors.

The documents in the query pools of the MedEval test collection were retrieved using the two indexes described in section 10.3, one with and one without decomposed compounds. Two searches were done in each of these indexes. For each index, one search was intended to be broad and one more specific, with the requirement that at least one instance of each facet had to be found. This requirement restriction was implemented with the filter requirement operator, **#filreq()**, together with the **#band()** and **#combine()** operators. The first term in these queries, constructed with the proximity operator Boolean AND,

makes the restriction, and the second part, constructed with the proximity operator combine, determines how the ranking is calculated. The queries put to the index with decomposed compounds contained decomposed terms as search keys, and also more synonyms than the queries put to the non-decomposed index. Examples of queries from the four different runs are shown in figure 42. These queries were used to select documents to assess for topic 7. The first two queries were put to the non-decomposed index, and the last two to the decomposed index.

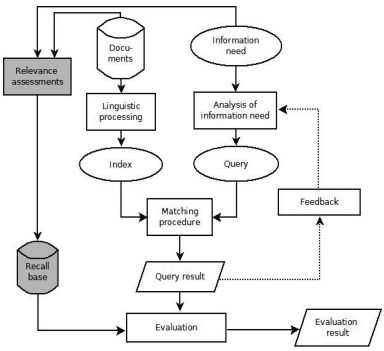
```
#combine( #syn(cellgift cellgiftsbehandling cytostatika)
biverkning cancer )

#filreq( #band( #syn(cellgift cellgiftsbehandling cytostatika)
biverkning cancer ) #combine( #syn(cellgift cellgiftsbehandling
cytostatika) biverkning cancer ))

#combine( #syn(cellgift cell gift cellgiftsbehandling cellgifts
behandling cytostatika cytostatikum cytostaticum) #syn(biverkning
risk) #syn(cancer tumör) )

#filreq( #band( #syn(cellgift cell gift cellgifts behandling
cellgiftsbehandling cytostatika cytostatikum cytostaticum)
#syn(biverkning risk) #syn(cancer tumör) )#combine( #syn(cellgift
cell gift cellgifts behandling cellgiftsbehandling cytostatika
cytostatikum cytostaticum) #syn(biverkning risk) #syn(cancer
tumör) ))
```

Figure 42: The queries used to extract documents for the pool of topic 7. The first two queries were used in the non-decomposed index, and the last two in the decomposed index.

For each topic, the result of the pooling was four lists of document IDs. These were merged in one file per topic. The IDs were sorted in alphanumerical order and duplicates were removed. This is important to avoid bias, as the assessors must not know how the documents were ranked in the initial runs or in how many searches each document was retrieved. The documents corresponding to the IDs in the pools were printed on paper and fixed in bundles, one or, if needed, several for each topic. The papers were only printed on one side to avoid negative bias caused by documents ending up on the left page of a spread.

10.5.2 Judging

The documents in the pools were assessed for relevance according to the corresponding information needs. Four medical students were hired to do the assessments, but not the same students as the creators of the MedEval topics. Domain knowledge is essential to truly understand the topics and the contents of the documents to be assessed. Medical knowledge allows an assessor to recognize, for example, Latin and Swedish equivalents or the correspondence between the commercial names of drugs and their chemical components.



As early as in the 70s, Saracevic (1975) studied what influenced relevance judgments. He concluded that the judges' expert knowledge was essential for consistency in judging.

> The more judges know about a query, the higher is the agreement among judges on relevance judgments and the more stringent judgments become. (Saracevic 1975: 340)

> It may be expected that the greater the judges' subject knowledge, the higher will be their agreement on relevance judgments. Subject knowledge seems to be the most important factor affecting the relevance judgment as far as human characteristics are concerned. (Saracevic 1975: 341)

The assessment procedure that was employed for the MedEval judging was inspired by the INEX Relevance Assessment Guide (INEX nd) and Ahlgren (2004) and is described below.

All the documents extracted for each need were assessed by one and the same assessor for reasons of consistency. Different assessors may have different opinions on which exact relevance grade they give a certain document, but given two documents they tend to agree which is the more relevant. Assessors may have different opinions on where to draw the line between different grades of relevancy, but there is a significant consensus in relative relevancy. This is sufficient for research according to the Cranfield paradigm since it is the relative results that are interesting, not the absolute numbers. This has been concluded in several studies, and already in Saracevic (1975).

> It is most significant to note that the relative relevance score of documents in a group, expecially [sic] among the documents with high relevance, may be expected to be remarkably consistent even when judges with differing backgrounds make the relevance judgments. Thus, it may be more profitable to compare the relative position of documents in a set than to compare the relevance ratings assigned to individual documents. (Saracevic 1975: 341)

The assessors working with MedEval were instructed to begin by studying an information need, so that they became familiar with it. To keep a current view of the topic they should always keep a written copy of the need at hand when reading through the documents, to be able to refresh their memory at any time. The assessors were asked to read every document to be assessed carefully, marking, in the margins, paragraphs that contributed to the topic. A paragraph was considered relevant if the assessors judged that they would use the information if writing an article or report on the topic. After reading a document through, the assessors were instructed to go over the document again, read through the marked paragraphs, and in accordance with them decide what degree of relevance the document should be assigned. To further assure the consistency of the judging, the assessors were instructed to return to the topic text every now and then and review it.

Each document was judged on its own merits. That is, seeing a piece of information for the umpteenth time did not make it any less relevant than it was the first time. Even if information repeated is already known to the assessor, a user may not retrieve the documents in the same order. Information seen at an early stage in one situation may not be familiar in another situation when a different set of documents have been retrieved, or simply read in a different order.

The TREC test collections, which have set the standard for many test collections, use a binary scale of relevance (Voorhees and Harman 2005). A document is considered either relevant or non-relevant. In the MedEval test collection the relevance assessments were made on a four graded scale, $0-3$ as recommended by Sormunen (2002) (see table 10.2). Four levels, instead of two, allow for a finer differentiation in the evaluation of search strategies, especially when it comes to retrieval of highly relevant documents compared to moderately relevant documents. These subtle differences can be visualized, for example by using gain vector evaluation (see section 3.5). Assessors also tend to find it easier to judge documents if there are at least three grades of relevance (Croft, Metzler and Strohman 2010). If later, an evaluation tool is used that only supports a binary scale of relevance, the four graded scale can easily be turned into a binary scale, if one, for example, regards documents graded

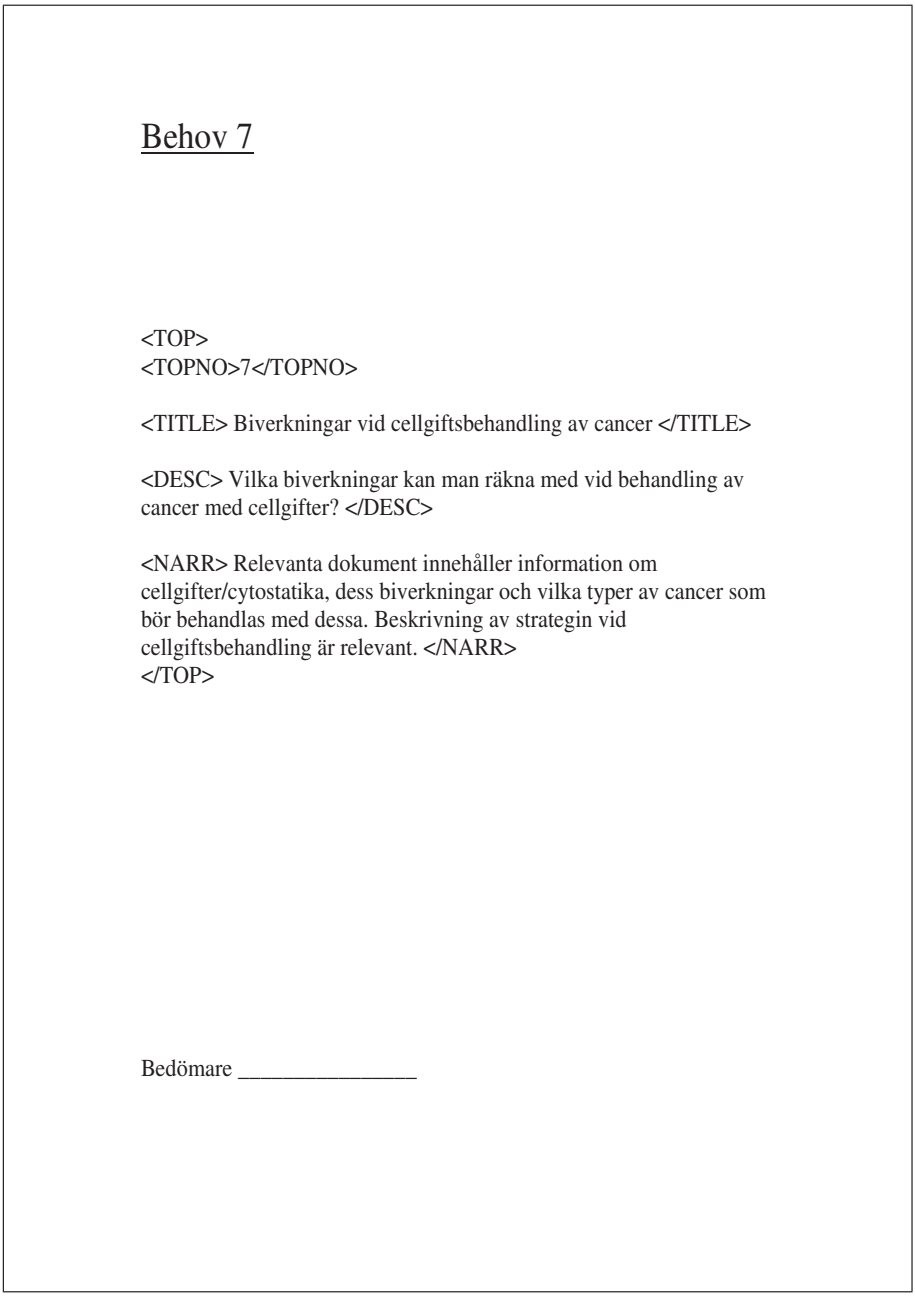
with 0 or 1, as well as unassessed documents, as non-relevant and documents graded 2 or 3 as relevant. This was the case when the QPA visualization tool was used in this work (see section 9.3). Another option is to vary the binary division by treating everything but relevance 3 as non-relevant, or by all documents with any relevance, 1-3, as relevant, and to study how this affects the results. These two last choices would represent the very selective user, who only wants very informational documents, or the user who wants to be sure to cover as much existing information as possible.

Table 10.2: The four-level scale of relevance according to Sormunen (2002) which is used in the MedEval relevance assessments.

Score	Label	Description
0	Non-Relevant	The document does not contain any information about the topic.
1	Marginally relevant	The document does not contain any other information about the topic than what is included in the description of the topic.
2	Fairly relevant	The document contains more information about the topic than the description, but the presentation is not exhaustive. If it is a topic with several aspects, only some of the aspects are covered.
3	Highly relevant	The document discusses all themes of the information need. If it is a topic with several aspects, all or most of the aspects are covered.

The relevance judged in MedEval is the **topical relevance**, how well the document corresponds to the topic or how much it is about the topic. The assessors were instructed not to involve **user relevance** in the relevance grade, that is how relevant a document is to a certain user at a certain point of time, for example, if the document contains information that is new to the user, if the author is believed to be reliable, or if the document is outdated.

In the MedEval collection, assessments of the documents were made, not only for relevance, but also for intended readers, or **target groups**. The assessors had to decide, for each document regardless of relevance grade, whether the document was written for lay persons or for medical professionals. Having assessment of target group in the test collection is taking a step towards measuring user relevance and utility without doing user studies involving informants.



Behov 7

<TOP>
<TOPNO>7</TOPNO>

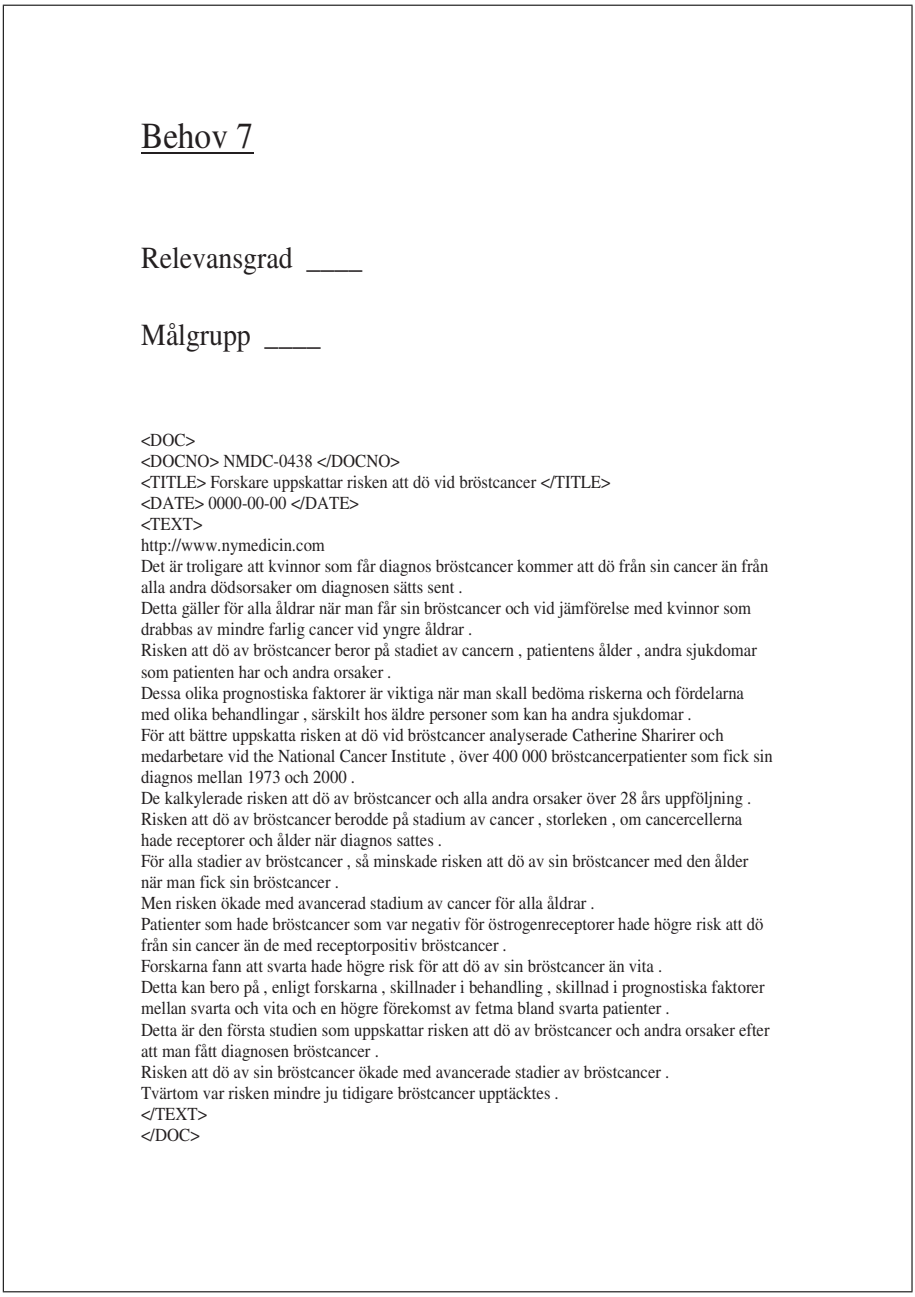
<TITLE> Biverkningar vid cellgiftsbehandling av cancer </TITLE>

<DESC> Vilka biverkningar kan man räkna med vid behandling av cancer med cellgifter? </DESC>

<NARR> Relevanta dokument innehåller information om cellgifter/cytostatika, dess biverkningar och vilka typer av cancer som bör behandlas med dessa. Beskrivning av strategin vid cellgiftsbehandling är relevant. </NARR>
</TOP>

Bedömare _____

Figure 43: An example of an information need as given to the assessor. *Behov 7* is Swedish for 'Need 7', and *Bedömare* is Swedish for 'Assessor'.



Behov 7

Relevansgrad ____

Målgrupp ____

<DOC>
<DOCNO> NMDC-0438 </DOCNO>
<TITLE> Forskare uppskattar risken att dö vid bröstcancer </TITLE>
<DATE> 0000-00-00 </DATE>
<TEXT>
http://www.nymedicin.com
Det är troligare att kvinnor som får diagnos bröstcancer kommer att dö från sin cancer än från alla andra dödsorsaker om diagnosen sätts sent .
Detta gäller för alla åldrar när man får sin bröstcancer och vid jämförelse med kvinnor som drabbas av mindre farlig cancer vid yngre åldrar .
Risken att dö av bröstcancer beror på stadiet av cancern , patientens ålder , andra sjukdomar som patienten har och andra orsaker .
Dessa olika prognostiska faktorer är viktiga när man skall bedöma riskerna och fördelarna med olika behandlingar , särskilt hos äldre personer som kan ha andra sjukdomar .
För att bättre uppskatta risken at dö vid bröstcancer analyserade Catherine Sharirer och medarbetare vid the National Cancer Institute , över 400 000 bröstcancerpatienter som fick sin diagnos mellan 1973 och 2000 .
De kalkylerade risken att dö av bröstcancer och alla andra orsaker över 28 års uppföljning .
Risken att dö av bröstcancer berodde på stadium av cancer , storleken , om cancercellerna hade receptorer och ålder när diagnos sattes .
För alla stadier av bröstcancer , så minskade risken att dö av sin bröstcancer med den ålder när man fick sin bröstcancer .
Men risken ökade med avancerad stadium av cancer för alla åldrar .
Patienter som hade bröstcancer som var negativ för östrogenreceptorer hade högre risk att dö från sin cancer än de med receptorpositiv bröstcancer .
Forskarna fann att svarta hade högre risk för att dö av sin bröstcancer än vita .
Detta kan bero på , enligt forskarna , skillnader i behandling , skillnad i prognostiska faktorer mellan svarta och vita och en högre förekomst av fetma bland svarta patienter .
Detta är den första studien som uppskattar risken att dö av bröstcancer och andra orsaker efter att man fått diagnosen bröstcancer .
Risken att dö av sin bröstcancer ökade med avancerade stadier av bröstcancer .
Tvärtom var risken mindre ju tidigare bröstcancer upptäcktes .
</TEXT>
</DOC>

Figure 44: An example of a document given to an assessor to judge in relevance to topic 7. *Relevansgrad* is Swedish for ‘Relevance grade’, and *Målgrupp* is Swedish for ‘Target group’.

When assessing the documents for a target group, the assessors decided for each document which group of readers was the intended one and marked the documents with a **P**, for *patienter* 'patients', if a document was written for lay persons, or with an **L**, for *läkare* 'doctors', if it was written for medical professionals. The assessors were forced to mark either a **P** or an **L**. The assumption is that doctors and patients could both have a certain, although not equal, interest in most documents. A third category including both doctors and patients would open up for the risk of having the majority of the documents marked for this group.

The number of documents that were assessed to be relevant to some degree, varies significantly between topics, as can be seen in table 10.3. In general, there are very few highly relevant documents (in total 207), several more fairly relevant documents (in total 1052), and significantly more marginally relevant documents (in total 2380). When looking at the set of all documents judged relevant, the numbers of documents judged to have doctor target group and patient target group are fairly evenly distributed. The biggest relative difference is for the marginally relevant documents, where 984 have target group doctors and as many as 1 396 have target group patients. If one instead looks at the individual topics there are quite big differences. For topic 28, most of the highly relevant and fairly relevant documents have target group doctors. Topic 36 has an almost equal number of doctor and patient documents, while topic 92 had no documents of any relevance grade marked with target group doctors. If one would construct ideal gain vectors, the None and the Patients ideal gain vectors would coincide fully for topic 92, while the cumulated gain for the Doctors scenario would be very low originating from downgraded patient documents. The differences in distribution of doctor and patient documents are illustrated in figure 45. These graphs show the ideal cumulated gain curves for all user groups for topics 28, 36, and 92. Ideal cumulated gain graphs for the 30 MedEval topics with the biggest recall bases are found in appendix B.

Table 10.3: Relevance assessments made by the assessors for the MedEval test collection.

	Relevant documents					Non-relevant documents					
Topic	Rel=3		Rel=2		Rels	Rel=1		Rel=0		Irrels	Total
	Doc	Pat	Doc	Pat		Doc	Pat	Doc	Pat		
1	1	3	17	12	**33**	35	18	42	50	145	178
2	0	1	2	1	**4**	19	51	12	37	119	123
4	3	1	9	17	**30**	13	27	63	110	213	243
5	0	0	1	8	**9**	0	5	35	195	235	244
7	5	3	19	39	**66**	29	37	43	47	156	222
9	0	0	1	1	**2**	3	1	73	161	238	240
10	0	0	0	6	**6**	0	11	35	208	254	260
11	0	2	6	5	**13**	6	12	73	120	211	224
12	0	0	5	5	**10**	7	7	51	76	141	151
13	1	0	0	2	**3**	3	1	95	67	166	169
16	2	1	5	12	**20**	7	30	31	81	149	169
18	1	4	2	15	**22**	14	9	48	100	171	193
19	2	1	5	5	**13**	6	13	161	114	294	307
20	1	1	2	19	**23**	3	72	57	51	183	206
21	0	1	4	5	**10**	4	12	95	142	253	263
23	7	3	35	14	**59**	46	21	122	110	299	358
25	2	1	22	12	**37**	32	12	138	59	241	278
26	3	3	6	8	**20**	24	13	72	45	154	174
27	0	1	2	0	**3**	10	7	139	103	259	262
28	3	0	33	7	**43**	45	13	58	56	172	215
31	1	2	5	8	**16**	1	7	112	74	194	210
32	0	2	4	7	**13**	12	19	108	51	190	203
36	3	1	27	25	**56**	45	40	49	52	186	242
37	3	2	21	6	**32**	8	11	51	138	208	240
38	1	0	3	4	**8**	15	21	34	50	120	128
39	4	1	12	7	**24**	22	35	138	103	298	322
41	2	0	5	10	**17**	19	25	67	85	196	213
42	6	5	7	29	**47**	8	76	65	31	180	227
43	2	1	2	3	**8**	14	16	89	98	217	225
44	1	2	3	5	**11**	19	23	94	130	266	277
46	4	2	15	4	**25**	18	28	65	75	186	211
48	1	1	14	5	**21**	3	5	97	85	190	211
49	1	0	9	20	**30**	33	48	49	92	222	252
50	2	0	5	0	**7**	35	24	112	150	321	328
51	2	1	5	15	**23**	26	42	68	97	233	256
53	0	0	12	4	**16**	18	20	131	38	207	223
54	0	1	6	2	**9**	5	10	68	23	106	115
55	2	3	0	1	**6**	13	7	103	38	161	167
56	1	8	8	7	**24**	1	8	93	55	157	181
57	3	3	8	2	**16**	12	6	90	72	180	196
58	0	0	5	8	**13**	8	21	49	49	127	140
62	2	0	12	5	**19**	41	39	65	60	205	224
63	0	1	4	5	**10**	3	5	70	125	203	213

Continued on next page...

Table 10.3- continued from previous page.

	Relevant documents					Non-relevant documents					
Topic	Rel=3		Rel=2		Rels	Rel=1		Rel=0		Irrels	Total
	Doc	Pat	Doc	Pat		Doc	Pat	Doc	Pat		
65	3	1	4	12	**20**	19	71	61	60	211	231
66	1	0	6	2	**9**	3	8	69	105	185	194
67	3	5	2	3	**13**	4	23	82	157	266	279
68	5	1	28	11	**45**	12	19	39	74	144	189
69	6	2	19	7	**34**	35	30	127	72	264	298
73	1	0	5	3	**9**	20	6	111	129	266	275
75	2	0	11	14	**27**	6	36	50	111	203	230
76	1	0	1	1	**3**	8	16	81	48	153	156
77	1	1	11	5	**18**	26	24	69	132	251	269
82	6	2	14	29	**51**	21	43	63	95	222	273
83	3	0	4	7	**14**	38	60	85	97	280	294
85	0	0	4	3	**7**	11	6	63	74	154	161
90	1	0	4	4	**9**	2	3	71	43	119	128
91	0	2	3	0	**5**	6	6	46	96	154	159
92	0	2	0	34	**36**	0	44	62	137	243	279
94	1	1	2	9	**13**	34	41	96	103	274	287
96	3	5	3	3	**14**	14	10	66	147	237	251
97	8	6	23	9	**46**	29	29	110	64	232	278
100	0	0	3	6	**9**	11	13	48	68	140	149
Sum	117	90	515	537	**1 259**	984	1 396	4 709	5 515	12 604	13 873
Min	0	0	0	0	**2**	0	1	12	23	106	115
Max	8	8	35	39	**66**	46	76	161	208	321	358
Mean	1.9	1.5	8.3	8.7	**20.3**	15.9	22.5	76.0	89.0	203.3	223.8
Med	1	1	5	6	**16**	12.5	18.5	68.5	83	203	224.5
StDev	1.9	1.7	8.3	8.3	**15.2**	12.9	18.0	31.4	40.6	51.6	54.7

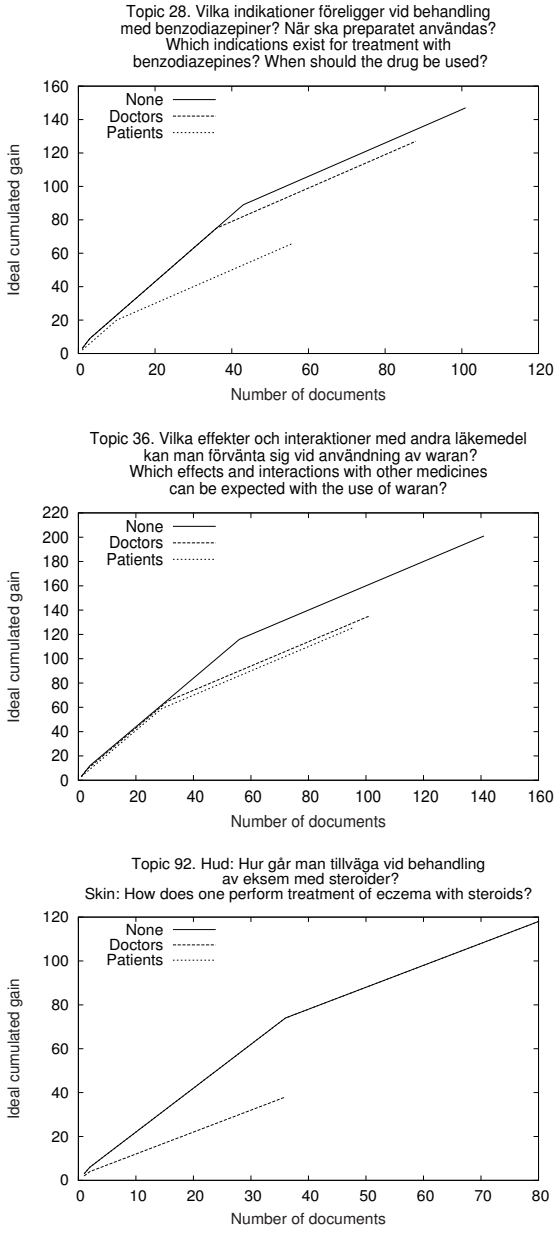
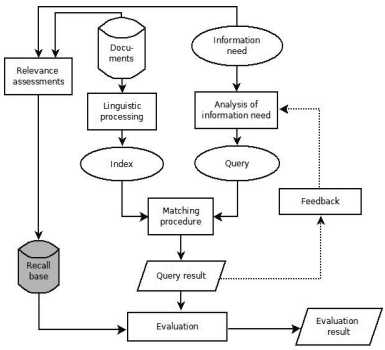


Figure 45: The recall bases of topics 28, 36, and 92 represented in ideal cumulated gain for all scenarios. For topic 28 most of the relevant documents had target group Doctors. Topic 36 had the relevant documents spread fairly evenly between the target groups. Topic 92 showed no documents of any relevance for documents marked Doctors. Thus, the None and the Patients curves coincide fully, while the cumulated gain for Doctors is very low.

10.6 Six collections in one



The MedEval test collection allows the user to state **user group**: 'None' (No specified group), 'Doctors' or 'Patients'. This choice directs the user to one of three scenarios. The None scenario contains the original relevance grades as made by the assessors. The Doctors scenario contains the same grades with the exception that the grades of the documents marked for the patient target group are downgraded by one. In the same way the Patients scenario has the documents marked for the doctor target group downgraded by one. This means that, for a doctor user, patient documents originally given relevance 3, are graded with 2, documents given relevance 2 are graded 1 and documents given relevance 1 are graded 0, and likewise in the patient scenario for the doctor documents. The idea is that a document that was written for a reader from one user group but retrieved for a user from the other group will not be non-relevant, but indeed less useful than a document from the correct target group. In the case of a document intended for the patient target group, this would (hopefully) contain background facts that most doctors already know. On the other hand, when it comes to documents intended for the doctor target group, even though they may be topically relevant for a patient's need, the risk is that they are written in such a way that the patient has difficulty grasping the whole content.

Creating user scenarios in this way with crudely adjusted relevance scores is done solely for research purposes. In a real life situation this would not be realistic as documents are not marked with relevance to different topics, and usually not with target audience group. The purpose is to give a possibility to study whether different types of queries tend to favor documents of either kind. The pilot studies in chapter 14 show examples of how such studies can be done.

The downgrading of document relevance for doctor and patient users results in smaller recall bases for the different user scenarios. One must bear this in mind when comparing the results for the different user groups. The same number of retrieved relevant documents at a certain document cutoff value will have the same precision, but not the same recall values as the recall base is smaller. For some topics with a very small recall base already in the None scenario, such as topics 9, 13, and 76, the recall bases in the downgraded scenarios may be too small to be meaningful.

Table 10.4: A subset of the assessments of topic 1, as given by the assessor. The first column contains IDs of assessed documents, the second column the relevance grade of that document to topic 1, and the third column the assessed target group of the document: L for läkare (doctors) or P for patienter (patients).

Document ID	Relevance grade	Target group
NTDK-2081	0	P
NTDK-2164	1	P
NTDK-2604	0	P
NTDK-2710	2	P
PFZR-0047	3	P
PFZR-0054	1	L
PRKT-0032	0	L
PRKT-0035	0	L
PRKT-0044	2	L
PRKT-0048	0	L
PRKT-0054	1	L
PRKT-0663	1	L
PRKT-0729	1	L
PRKT-0738	0	L
SCHR-0043	2	L
SCHR-0061	1	L
SOSX-0002	2	L
SVDX-0087	3	P
SVDX-0089	1	P
SVDX-0123	0	P
SVDX-0125	2	P
SVRD-0032	1	P
SVRD-0113	0	P
SVRD-0614	3	P

Table 10.5: The same subset of relevance assessments for topic 1 as in table 10.4, but here the grades are adjusted to target groups. The first relevance column leaves the grades as the assessors made them, without adjustment. In the second column, which shows relevance grades for doctor users, the relevance for the patient documents have been downgraded. In the third column the same is done for patient users and doctor documents.

Document ID	Relevance None	Relevance Doctors	Relevance Patients
NTDK-2081	0	0	0
NTDK-2164	1	0	1
NTDK-2604	0	0	0
NTDK-2710	2	1	2
PFZR-0047	3	2	3
PFZR-0054	1	1	0
PRKT-0032	0	0	0
PRKT-0035	0	0	0
PRKT-0044	2	2	1
PRKT-0048	0	0	0
PRKT-0054	1	1	0
PRKT-0663	1	1	0
PRKT-0729	1	1	0
PRKT-0738	0	0	0
SCHR-0043	2	2	1
SCHR-0061	1	1	0
SOSX-0002	2	2	1
SVDX-0087	3	2	3
SVDX-0089	1	0	1
SVDX-0123	0	0	0
SVDX-0125	2	1	2
SVRD-0032	1	0	1
SVRD-0113	0	0	0
SVRD-0614	3	2	3

When searching in the MedEval test collection, in addition to indicating user group, the user must choose which index to search in, with or without split compounds. This choice is present in all three user scenarios. This means that the same query in connection with the same topic will give six different results depending on which user scenario and which index are chosen. Searching with one and the same query, keeping the index constant but varying the user group, will give you identical lists of ranked documents, but as you have different recall bases you will get varying results for whatever type of measurement you choose. Keeping the user group constant, but varying the index will give you different lists of documents to compare to one and the same recall base. An example of results varying in this manner for one and the same query can be seen in figure 46.

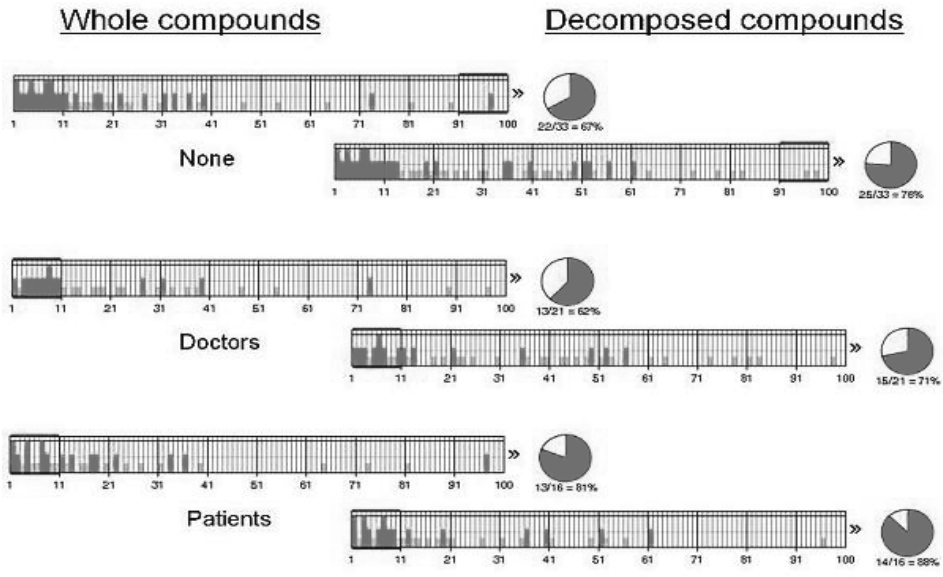


Figure 46: Six different results for the same query for the six possible combinations of index and user group in MedEval.

Part IV

Pilot studies

11 CONSTRUCTING FACETS

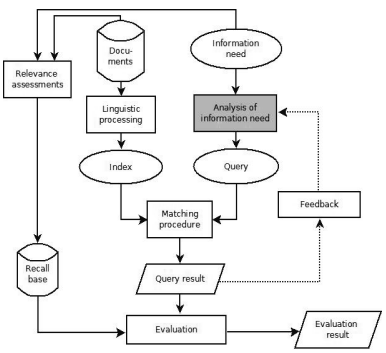
How many roads must a man walk down?

The Hitch Hiker's... chapter 23[33]

Information retrieval queries are often divided into facets which represent the main concepts of the corresponding information need. Each facet consists of a group of terms that play a similar semantic role. The terms within a facet are often, but not necessarily, synonyms or near synonyms. The part of speech of a term is not essential, as long as the terms within a facet evokes similar notions. It is a more fuzzy similarity than between near synonyms. The two main points are **family of concepts** and **aspect** as suggested by Sormunen et al.

> A facet is a concept (or a family of concepts) identified from, and defining one exclusive aspect of a request or a search topic. (Sormunen et al. 2001: 361)

To evaluate queries or the impact of facets and search keys, it is common to have a baseline query to compare against to see if the item examined is beneficial for the result or not. For the initial experiments with MedEval baseline queries, containing all facets, were constructed for each topic. To study the impact of individual search keys within a facet, baseline queries were constructed which contained all terms within the facet.



[33] Borrowed from Bob Dylan.

11.1 Choice of operators

When creating the baseline queries of MedEval, four operators of the Indri query language were used: **#combine()**, **#syn()**, **#od1()**, and **#uw5()**, combine, synonym, ordered window and unordered window.

The belief operator #combine() was used to join all terms into one query where they were all treated equally. The proximity operator #syn() was used to delimit the facets from each other. All terms within this operator were treated as instances of the same term.

The proximity operator, #od1(), was used to deal with the problem that the hyphen was ignored by the Indri search engine, in the version used, and thus constituents of compounds constructed with the hyphen were treated as separate words. The #od1() represents an ordered window where the terms must appear in the order they are listed, without intervening terms. The compound *ssri-preparat* would thus be represented by the term `#od1(ssri preparat)`. A side effect of this is that a search term constructed in this way with the proximity operator also corresponds to instances where the terms within the operator appear separately in immediate sequence in documents. This will affect searches with the decomposed index. The whole and the split compound in the pseudotext will match the same queries: *ssri-preparat* and *ssri preparat*.

The #uw5() operator was used for phrases or multiword units conveying a specific concept. These could, for instance, be adjective-noun phrases, *medicinsk behandling* 'medical treatment' or named entities such as names of diseases, *Alzheimers sjukdom* 'Alzheimer's disease'. The #uw5() operator gives an unordered window, which allows the terms within in any order, but with the restriction that they must appear within a window of 5 terms. The words in phrases or multiword units often appear in paraphrases or with intervening adjectives, articles or prepositions. Relaxing the proximity requirements, by allowing intervening terms, allowed matching of such cases.

Figure 47 shows an example of a baseline query where all four operators mentioned are used.

11.2 Selection of terms

To divide the terms of a query into facets one must of course have terms. The starting point in building the baseline queries was to use the terms that occurred in the topics, that is the content terms of the titles, the descriptions and the narratives, such as they are listed in appendix A. In order not to let a random choice of words from the topic creators cause differences in the results, it was decided to use the same sets of words in all topics with similar facets. Finally

```
#combine(#syn(behandla behandling strategi
behandlingsstrategi behandlingsmetod
behandlingsalternativ tillvägagångssätt genomföra)
#syn(använda användning användande administrationssätt insättande
nyttjande nyttja tillämpa)
#syn(dos dosering)
#syn(försiktighet biverkning komplikation risk följd riskfaktor)
#syn(terapiintervall)
#syn(serotoninupptagshämmare ssri #od1(ssri preparat)
#uw5(medicinsk behandling) agens preparat preparattyp läkemedel)
#syn(depression nedstämdhet #uw5(depressiv sjukdom))
#syn(gravid havande graviditet havandeskap))
```

Figure 47: The baseline query for topic 31. The contents of the #syn() operators each represent one facet.

the Swedish MeSH[34] was consulted to find synonyms for expansion of the medical terms. The MeSH lexical entries and terms from the explanations in the thesaurus were used.

The set of terms used to create the facets were thus the content terms that occurred in all parts of the topics and their near synonyms from other topics, and, for medical concepts, terms describing them in MeSH. From this set of terms, all words with similar meaning were grouped together. A consequence of grouping terms in this fashion is that some facets that cover general concepts contain a very large set of terms, such as the first facet in figure 47. This is the facet for the concept [behandla, behandling] ‘treat, treatment’.

```
#syn(behandla behandling strategi behandlingsstrategi
behandlingsmetod behandlingsalternativ tillvägagångssätt
genomföra)
```

‘treat’ ‘treatment’ ‘strategy’ ‘treatment strategy’ ‘treatment method’ ‘treatment alternative’ ‘procedure’ ‘perform’

Constructing facets is not as straight forward as one may think. The rest of this chapter will explain the reasoning behind the construction of the facets for the baseline queries. The facets in the examples belong to different topics and are chosen to illustrate different kinds of problems.

[34] <http://mesh.kib.ki.se/>

The facet for the concept [pregnancy] shows terms with varying parts of speech. The first two terms are adjectives while the last two are nouns. All evoke the same concept, and are thus placed in the same facet.

```
#syn(gravid havande graviditet havandeskap)
```
'pregnant' 'pregnant' 'pregnancy' 'pregnancy'

Style, approach of aspect and, to a certain degree, specificity, have not motivated separate facets. A distinction is here made between aspect and approach. While the aspects of a request, the facets, are the main concepts, the approach is how one looks at one of these concepts.

In the facet below there are different terms in different styles for the concept [menopause]. The most scientific style is *menopaus* and the most casual is *övergångsålder* 'transition age' (change of life).

```
#syn(menopaus klimakterie övergångsålder)
```
'menopause' 'menopause' 'transition age' (change of life)

In the facet below are several terms which can be perceived to have different meanings, but they are all approaches of the concept [something going wrong in medical treatment].

```
#syn(försiktighet biverkning komplikation risk följd riskfaktor)
```
'caution' 'side-effect' 'complication' 'risk' 'consequence' 'risk factor'

Concepts are not always named when discussed, instead paraphrases can be used, such as 'high temperature' for the concept [fever]. This relationship between a term and the term paraphrase allows, for example 'fever' and 'body temperature' to be put in one facet.

```
#syn(feber kroppstemperatur)
```
'fever' 'body temperature'

Terms that are more or less interchangeable in a certain context can be put in the same facet. But sometimes drawing the line between facets is difficult. It is a balance between being consistent and keeping the characteristics of the information needs. For some topics, a certain distinction may be important, for others not. When, for example, writing an article about a certain disease it may not be important if the patient is a man or a woman. If this is the case, the author could use the term *patient* 'patient' or *person* 'person'. However, writing about a specific case, the terms *man* 'man' or *kvinna* 'woman' may

very well be chosen instead. Using these terms does not necessarily imply that what is written is applicable only to men or only to women. Therefore it can be motivated to put all these terms in the same facet. In other cases it is essential to know if the patient is male or female. In the MedEval baseline queries, terms for 'person', 'patient', 'man', and 'woman' are put in the same facet representing the [patient] concept, except in cases where the information need specifically addresses persons one gender.

```
#syn(person patient man kvinna)
```
'person' 'patient' 'man' 'woman'

A facet can contain terms with a varying degree of specificity or terms that are similar in some other aspect. In a strict medical sense *allergy* and *hypersensitivity* are not the same disorders, but they are both about a person not tolerating a certain substance. The choice to put them in the same facet or not depends on how specific an answer the user wants. For the baseline queries in this study they were put in the same facet.

```
#syn(allergi allergisk överkänslighet överkänslig)
```
'allergy' 'allergic' 'hypersensitivity' 'hypersensitive'

In the case of compounds, different degrees of specificity can sometimes motivate the creation of different facets. The choise is not only if the difference between the terms is big enough to motivate more than one facet, but also in which facet the compound should be. The one to which the first, or the one to which the second constituent belong? Even though a compound often is a hyponym of the second constituent, it is not certain that the best choice is to group the compound with the head. In the [cancer] facet, we have *cancer* 'cancer' and *cancertyp* 'cancer type'. The head of the compound, *typ* 'type', is almost void of content. It is not so that 'cancer type' is a kind of 'type'. Instead the terms *cancer* and *cancertyp* are near synonyms. In the same topic we also find the terms *matstrupscancer* 'food throat cancer' (esophageal cancer) and *matstrupe* 'food throat' (esophagus). Here the distinction was made between the disease and the location. Thus these terms were separated into different facets.

```
#syn(cancer tumör neoplasm svulst cancertyp
matstrupscancer)
```
'cancer' 'tumor' 'neoplasm' 'growth' 'cancer type' 'food throat cancer' (esophageal cancer)

```
#syn(matstrupe matstrup esofagus)
```
'food throat' (esophagus) 'food throat' (esophagus) 'esophagus'

The [esophagus] facet in the example contains the term *matstrup*. This is not a free word form, but the stem of the compound *mat||strupe*. Stem forms like these occur in both indexes in the case where a documents contains an elliptic coordination. In the decomposed index the stems also appear when there has been a compound segmentation and the modifier is the stem of the lexeme, not the complete lemma. Stems have been included in the baseline facets to reduce the dependency on the decomposition tool and the lemmatizer for the result of the searches.

Two recurring concepts in the topics were [medicine/drug] and [treatment]. When specific drugs or specific treatments were named in a topic, the name was added to the basic facet of [medicine/drug] or of [treatment]. In the example below terms denoting 'chemotherapy' have been added to the [medicine/drug] facet.

```
#syn(cellgiftsbehandling cellgift cytostatika
#uw5(medicinsk behandling) agens preparat
preparattyp läkemedel)
```
'cell poison treatment' (chemotherapy treatment) 'cell poison' (chemotherapy) 'chemotherapy' 'medical treatment' 'agent' 'preparation' 'preparation type' 'medicine'

The phrases *medicinsk behandling* 'medical treatment' and *cellgiftsbehandling* 'cell poison treatment' (chemotherapy) were sorted under the medicine/drug facet even though the head of the phrases, in both cases, is *behandling* 'treatment'. This was done because the modifier carried more information than the head.

In the experiments that followed, which are described in chapters 12 and 13, the construction of baseline facets such as [medicine/drug] and [treatment], with a large number of low specificity terms, turned out not to be ideal. These facets had very low effectiveness when used as queries.

12 LOOKING AT FACETS AND TERMS

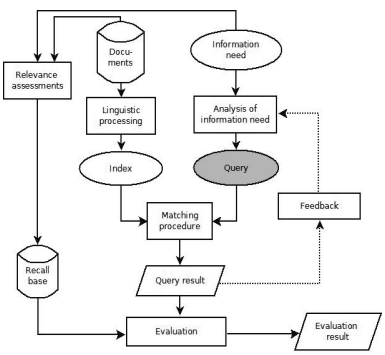
Once you know what it is you want to be true,
instinct is a very useful device for enabling you to know that it is.

So Long... chapter 7

In the first tests with the MedEval collection, searches were made using the individual terms of the topic descriptions as single search key queries. Thereafter terms from the baseline queries, described in chapter 11, were tested in the context of other terms. The intention was to get a broad look at how the terms behaved as search keys. Which concepts or search keys worked well? Which did not work well? Which facets and which search keys gave high recall early in the ranked list? Which search keys did not reach high recall until many documents were retrieved? This preliminary survey was intended as a starting point and inspiration for later experiments.

12.1 Term survey

In the initial tests on individual terms in relation to the information needs, the topic descriptions were used as a source to choose terms. (All topic descriptions are found in appendix A.) The stop words were ignored, and only the content words were used. The compounds in the descriptions were decomposed and the compound constituents were used as additional terms. The queries were run through the Indri search engine for the two indexes, with and without decomposed compounds. Both recall and nDCG values were extracted at document cut off value 10, 20, and 1 000,



that is after 10, 20, and 1 000 retrieved documents. DCV 10 and 20 were chosen to represent the impatient, and the slightly less impatient, user. This reflects the fact that most information seekers will look only at the first 10 documents, and seldom past 20 documents in a ranked list of retrieved documents. DCV 1 000 was chosen to see if the query in question was effective at all. If documents were retrieved at higher ranking levels this could be an indication that the terms used could be effective in combination with other terms. Alternatively, if a query did not retrieve any relevant documents, even at DCV 1 000, the terms used were surely not effective, at least not for the topic in question. Gain curves showing results as far as DCV 1 000 also indicate if the query posed to the system only retrieves relevant documents early in the ranked list, or if the user may benefit from retrieval of many documents.

The information extracted in this survey was put into histograms with bars grouped for each term. The bars in each term group were, from left to right: Non-decomposed index at DCV 10, 20, and 1 000, decomposed index at DCV 10, 20, and 1 000. The histogram for topic 1 is shown in figure 48.

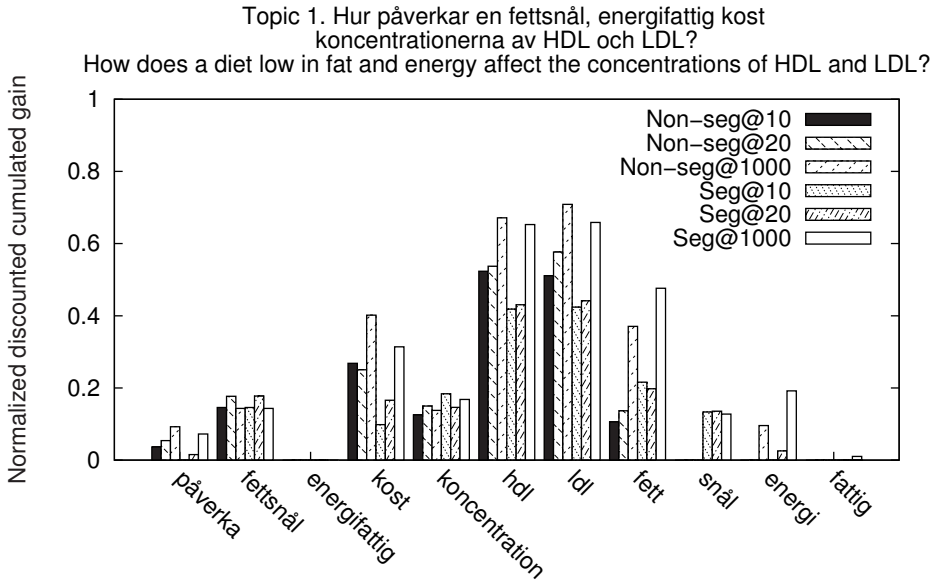


Figure 48: The nDCG values at 10, 20, and 1 000 documents for the content words of the description of topic 1. For each term, the three leftmost bars are for the index without decomposed compounds, and the three rightmost for the index with decomposed compounds. The term equivalents are, from left to right: 'affect 'fat stingy' (low-fat) 'energy poor' (low in energy) 'diet' 'concentration' 'HDL' 'LDL' 'fat' 'stingy' 'energy' 'poor'.

The histograms give a visualization of how the search keys behave in the MedEval collection and for the information need in question. It is important to remember that, even though the properties of a term affect the result, a search key is not good in itself, only in relation to a certain information need in a certain collection. A search key such as 'diet', for example, would give different kinds of hits in a collection of medical documents and in a collection built on articles from weekly magazines. A search also has to describe the information need well in order to be effective. A search key can be very good for one need but only moderately good or not good at all for another. This would be the case for a search key with high tf∗idf factor which describes one topic well but is not relevant for another topic.

Below are some reflections of the histogram in figure 48. The bars in the figure shows the results for topic 1 using the content words one at a time, as search keys. There are two compounds in the topic description. These compounds have been segmented and the constituents used as individual search keys.

The histogram is completed with figures 49 through 54 where nDCG curves up to DCV 1 000 are shown for selected terms of topic 1. These figures give a more complete overview of the nDCG values throughout the ranked list of retrieved documents. It is from these curves that the values in the histograms are picked. These nDCG curves are screen dumps from the VisualVectora visualization tool.

påverka 'affect' – Low effectiveness. A term with a general meaning, not specific for this topic. It has slightly better results in the non-decomposed index.

fettsnål 'fat stingy' (low-fat) – Somewhat effective. The meaning is essential for the topic. Figure 49 shows nDCG curves for the search key *fettsnål* for both the non-decomposed and the decomposed index. The curves are very similar. The result is moderately successful for the very first documents, then the curve flattens out. As can be seen in table 12.1 on page 160, there are only 37 occurrences of *fettsnål* in the collection, which entails that a maximum of 37 documents containing the term can be found. Even though *fettsnål* has been decomposed by the compound splitter for the index, it is more like a derivation than a compound (see section 4.1.4.1). It is not common as a compound constituent, therefore it gives only slight differences between the indexes.

energifattig 'energy poor' (low in energy) – The meaning of the term is essential for the topic but the term has low frequency and renders no hits.

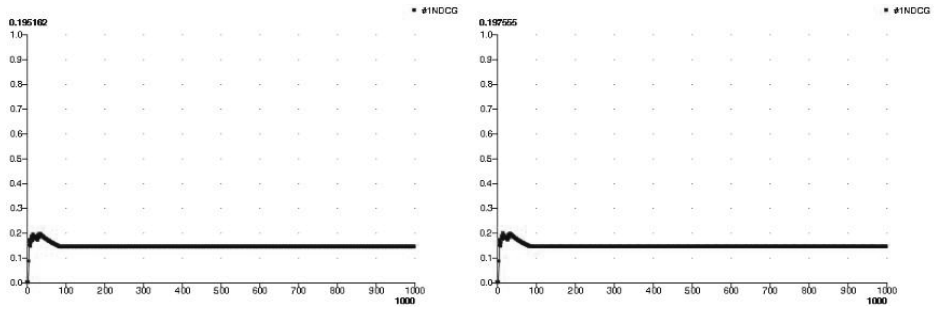


Figure 49: For topic 1, the search term *fettsnål* 'fat stingy' (low-fat) shows, moderately successful results for the first 100 documents. After that, the curve flattens out. No more documents containing the term are found. The results are similar for the two indexes. The result for the non-decomposed index is shown to the left and for the decomposed index to the right.

Only two documents in the collection contain the term, and these documents were not relevant to the topic. Also this term can be seen rather as a derivation than a compound.

kost 'diet' – Quite effective in the non-decomposed index, not as effective in the decomposed index. The meaning is essential, but not exclusive, for the topic. As can be seen in figure 50 the form of the nDCG curve is similar in both indexes, with an initial peak and then a slowly, but constantly, growing curve. For the non-decomposed index the initial peak is higher and the curve is throughout on a higher level than for the decomposed index. A slowly growing nDCG curve, as obtained by both indexes, indicates that the search term has high frequency and occurs in many documents, both relevant and non-relevant.

koncentration 'concentration' – Somewhat effective. It is a term that is essential, but not exclusive, to this topic. The results are similar for both indexes, but slightly better for the decomposed index.

fett 'fat' – Somewhat effective in the non-decomposed index but quite effective in the decomposed index. This is a term that is essential, but not exclusive, to this topic. For both indexes the curve is slowly rising. For the non-decomposed index the curve flattens out at about DCV 700, indicating that no more documents with the term are found. As we can see in table 12.1, the lemma *fett* (all inflectional forms) has about 1 700 occurrences. This seems reasonable seeing that words that are significant for a document tend to occur more than once in that document, and be-

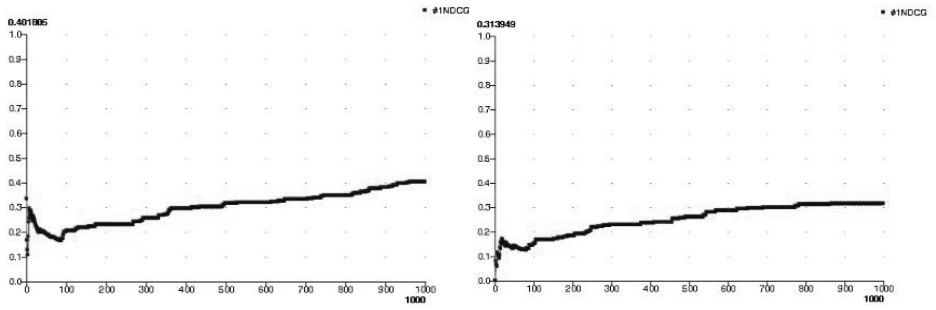


Figure 50: After an initial peak, the search term *kost* shows, for topic 1 in the non-decomposed index to the left, a slowly but constantly growing nDCG value. The nDCG curve for the decomposed index, to the right, has a similar shape, but is throughout on a lower level.

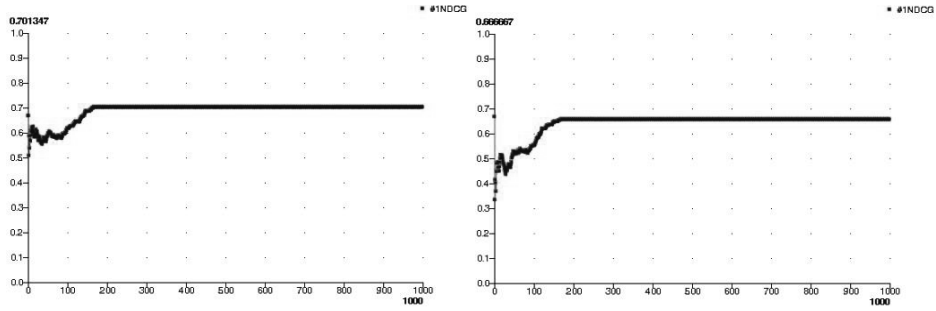


Figure 51: The search term *ldl* shows, for topic 1, very early a high nDCG value, in both indexes the non-decomposed index to the left and the decompose to the right. After 200 documents all documents containing the term *ldl* have been retrieved. The curves turn flat as no more documents are retrieved. The curve in the decomposed index is throughout on a lower level.

cause the term is general enough to occur also in documents that are not relevant to this topic.

hdl 'hdl' – Effective. Very significant to the topic. The runs in the non-decomposed index give the best results.[35]

[35]Swedish compounds composed with acronyms such as *hdl* and *ldl* take the '-' as a link element. The results for *hdl* and *ldl* in these runs are affected by the fact that the Indri search engine, in the version used, treats hyphens as whitespace. This means that the search engine matches *hdl* and *ldl*, not only to the acronyms themselves when they are used a simplex terms, but also to the acronyms as constituents in non-decomposed compounds constructed with the hyphen. This makes the impact of decomposition smaller.

ldl 'ldl' – Effective. Very significant to the topic. The runs in the non-decomposed index give the best results.[35]

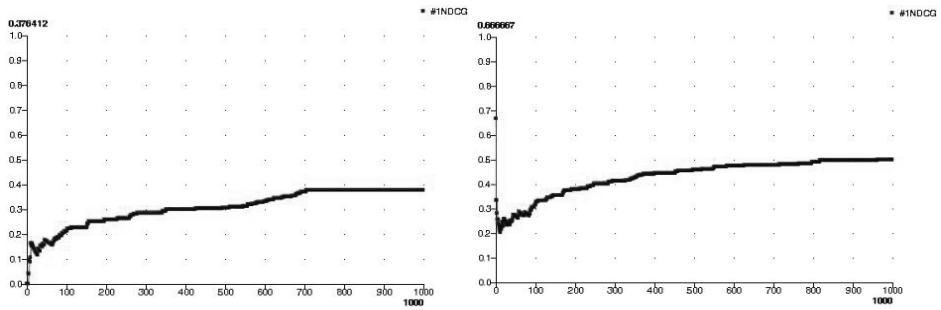


Figure 52: The search term *fett* 'fat' shows, for topic 1 in the non-decomposed index, to the left, fairly successful results. The curve grows continuously for the first 800 documents and then flattens out. The results for the same search term and topic in the decomposed index, to the right, shows a similar curve as for the non-decomposed index, but the curve is throughout on a higher level, and levels out later.

snål 'stingy' – This word-like affix is not relevant to the topic when transformed to an independent word. There are no hits in the non-decomposed index but some hits in the decomposed index. The nDCG values are lower than when using the original term *fettsnål*, from which it is derived, indicating that the hits are mixed up with noise not ranked so high before. Among the relevant documents that are retrieved are the ones that contain the compound *fettsnål* where the matches are from this decomposed compound.

energi 'energy' – Low effectiveness. Relevant to the topic, but not exclusively. It has 575 occurrences in the collection, a medium frequency, but it is very likely common in a variety of topics. The low or nonexistent bars at DCV 10 and 20 indicate that there is a significant amount of noise. The term *energi* has 1 423 occurrences. This gives effect in the decomposed index, something which can be seen in figure 54, where the rightmost curve keeps growing much longer than the leftmost one.

fattig 'poor' – This word-like affix has extremely low effectiveness. The content of the corresponding term is not relevant to the topic. The existing hits are not from the decomposed original word as this did not occur in any relevant documents. They could be from semantically similar words such as *fettfattig* 'fat poor' (low in fat), which has a collection frequency of 22 (see table 12.1).

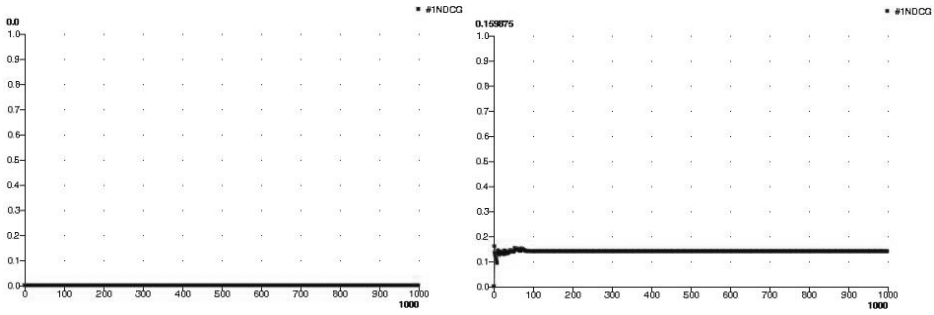


Figure 53: For topic 1 the search term *snål* 'stingy' does not retrieve any relevant documents at all using the non-decomposed index, the result to the left. Hence this curve is totally flat. The result for the same topic and search key in the decomposed index, to the right, shows similarities with the curve for the full compound *fettsnål*. The documents retrieved with the search key *snål* are the same as the ones retrieved with the search key *fettsnål* but now mixed with additional noise. This curve is flatter than in the curve for *fettsnål* in figure 49. This indicates more noise.

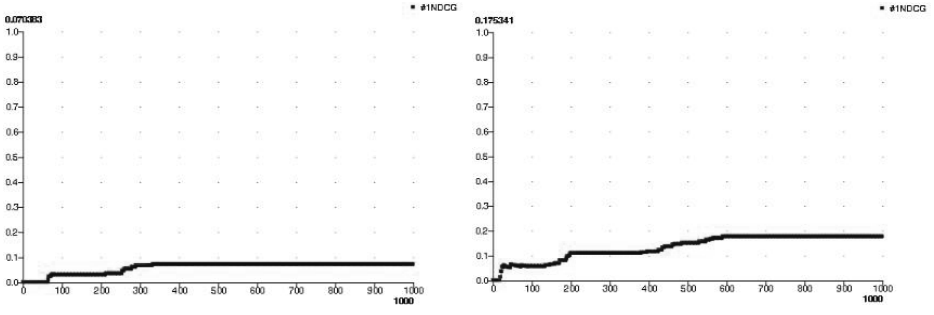


Figure 54: The search key *energi* 'energy' shows, for topic 1 better results in the decomposed index. The decomposition increases the number of matching tokens, which makes the nDCG curve grow faster and for a longer time for the decomposed index.

Studying the histograms for the individual terms in relation to information needs confirmed some beliefs and raised some questions. It was not surprising that general and common words such as *påverka* 'affect' were not effective as search keys. But why does the term *kost* 'diet' have better results in the non-decomposed index while the term *fett* 'fat' has better success in the decomposed index? The fact that a term with few occurrences does not give many hits is obvious, but sometimes words have surprisingly low frequency. For instance, the term *energifattig*, a perfectly normal compound, occurred in

only two documents. These two documents had relevance grade 2 for one topic each, although none of these was topic 1. For these two topics the search key *energifattig* (low in energy) could contribute with at least one relevant document. One can speculate that in another collection, for example about diets or nutrition, the term may have had a bigger impact. Terms with very low frequency will of course never hamper the results much. The worst case scenario is a few documents of noise.

12.2 Effects of decomposition

To get an idea of why using *fett* as a search key gave better results when changing from the non-decomposed to the decomposed index, while using *kost* gave worse results, all terms in the document collection containing the simplex terms *fett* and *kost* were extracted. The 35 most frequent terms, in both cases, are listed in tables 12.1 and 12.2 together with their frequencies. Anders Thurin, MD at Sahlgrenska University Hospital, was asked to judge how relevant these terms were to topic 1. His assessments are seen in the rightmost columns of the tables. The '∗' sign marks relevance for terms found in the non-decomposed index, that is the lemma forms of the lemma sign itself, and the '+' sign marks term relevance for the terms that would be the additional matches using the decomposed index, that is when *fett*/*kost* are constituents of compound terms. A higher number of '∗' or '+' signs signifies a higher level of term relevance to the topic.

Overall, the compounds containing *fett* are slightly more relevant and more frequent than the compounds containing *kost*. The compounds containing *fett* are also more specific and relate to fat and contents of the blood, which is essential to the subject of the topic. The compounds containing *kost*, on the other hand, are less specific and concern diets in general.

A crucial fact concerns the most frequent compounds in the two tables, that is, the most common words which would be matches in the decomposed index but not in the non-decomposed index. In table 12.1 we find *fettsyror* 'fatty acids' and *blodfetter* 'blood lipids' with frequencies 563 and 476. Both of these compounds are quite relevant to topic 1, marked with relevance ++. In table 12.2 we find, as the most frequent additional matches, the compounds *kosttillskott* 'dietary supplement' and *hälsokostpreparat* 'health food preparation', which are relatively general and not relevant to topic 1. They have frequencies 219 and 178. This means that a significant part of the additional matches that will come with the decomposition of compounds will have a good probability of being relevant for the search term *fett*, but not for the search term *kost*. As nothing is removed neither from the query nor the index, the relevant doc-

uments that were found with the search key *kost* using the non-decomposed index are still retrieved with the decomposed index. However, several of them will come later in the ranked list of documents as the list now contains more noise. This noise is what makes the result deteriorate.

One important difference between the terms *fett* and *kost*, as used in the description of topic 1, is that *fett* occurs as a constituent of another word while *kost* occurs as an independent simplex word. When decomposing all compounds in the index the search key *kost* will match all occurrences of compounds in the index which have *kost* as a constituent. These compounds will mostly not be on the same level of specificity. As for *fett* we have a different situation. The corresponding term in the topic description is *fettsnål* 'fat stingy' (low-fat). This is a word constructed with *snål*, a word-like affix (see section 4.1.4.1), making the word *fettsnål* rather a derivation than a compound. The purpose of using this affix is to transform the noun *fett* into an adjective and diminishing it. The core meaning of this adjective is *fett*, but now with a component of degree.

What happens after decompostion is that we match the search key *kost* with compounds in the index which are on a different level of specificity than the term in the topic. The term *kost* is a broad concept. When it is made into a compound by adding a modifier it is often to narrow the concept. As for the search key *fett*, we first of all now match it to the instances of the decomposed *fettsnål*, which was the term in the description. These matches were not successful with the non-decomposed index. We also match the term to several compounds which refer to the degree of fat, just as the description term *fettsnål*, such as *fettintag* 'fat intake', *fettrik* 'rich in fat', *fettinnehåll* 'fat content' and several others. The term *fett* describes a substance, and looking at table 12.1 it seems that the compounds using this term do not so much narrow the concept as describe different degrees of fat or describe situations were fat occurs in the body.

What is described above seems a probable reason why the impact of the search keys in question differs so much in the two indexes. It is clear that search terms, simplex, compounds or compound constituents, behave differently and that the impact of compound decomposition differs widely for different terms. Investigating this effect of decomposition could be an interesting future project using the MedEval test collection.

Table 12.1: The 35 most frequent terms in the corpus containing the lexeme *fett* 'fat', and their relevance to topic 1. The '∗' sign is used to mark relevance for lemma forms that would match *fett* in the non-decomposed index, and the '+' sign for the additional matches in the decomposed index. The more '∗' or '+' signs, the more relevant to topic 1.

Frequency in collection	Swedish word	English equivalent	Relevance to topic 1
1416	fett	fat	∗∗
563	fettsyror	fatty acids	++
476	blodfetter	blood lipids	++
192	fettet	the fat	∗∗
188	fetter	fats	∗∗
112	transfett	trans fatty acid	+
95	transfetter	trans fatty acids	+
88	fettsyrorna	the fatty acids	+
80	blodfetterna	the blood lipids	++
70	åderförfettning	vessel fattening (arteriosclerosis)	
66	blodfettssänkande	blood lipid lowering	+++
65	fettväv	fat tissue (adipose tissue)	
65	fettceller	fat cells (adipocytes)	
63	fettvävnad	fat tissue (adipose tissue)	
57	kroppsfett	body fat	
56	fettväven	the adipose tissue	
52	fettrik	rich in fat (ketogenic)	++
49	fettsyran	the fatty acid	++
44	fettlösliga	fat soluble	
41	underhudsfett	subcutaneous fat	
41	fettintag	fat intake	+++
40	transfettsyror	trans fatty acids	+
40	fettsugning	liposuction	
37	blodfettsänkande	blood lipid lowering	+++
37	fettsnål	low-fat	+++
35	fettsyra	fatty acid	+
31	fettlever	fatty liver	
30	fettinnehåll	fat content	+++
29	fetthalt	fat content	+++
27	fettvävnaden	the adipose tissue	
26	fettdepåer	fat depots	
25	fettcellerna	the adipocytes	
25	fettknöl	fat lump (lipoma)	
24	bukfett	abdomen fat (intra-abdominal fa)t	
22	fettfattig	fat poor (low in fat)	+++

Table 12.2: The 35 most frequent terms in the corpus containing the lexeme *kost* 'diet/food', and their relevance to topic 1. The '∗' sign is used to mark relevance for lemma forms that would match *kost* in the non-decomposed index, and the '+' sign for the additional matches in the decomposed index. The more '∗' or '+' signs, the more relevant to topic 1.

Frequency in collection	Swedish word	English equivalent	Relevance to topic 1
1395	kost	diet	∗∗
547	kosten	the diet	∗∗
219	kosttillskott	dietary supplement	
178	hälsokostpreparat	health food peparation	
126	kostvanor	food habits	++
97	kostråd	dietary advice	++
86	kostens	the diet's	++
52	kosthållning	diet holding (diet)	++
39	kostråden	the dietary advices	++
33	kostbehandling	dietary treatment	++
32	kostfiber	dietary fiber	
30	kostfibrer	dietary fibers	
28	hälsokost	health food	+
25	kosttillägg	dietary supplement	
25	medelhavskost	mediterranean diet	+
24	kostrådgivning	dietary advice service	+
23	kostförändringar	dietary changes	++
23	kostomläggning	dietary change	++
22	husmanskost	plain [Swedish] cooking	
22	kostintag	food intake	
21	normalkost	normal diet	
21	hälsokostaffärer	health food stores	
21	lågkolhydratkost	carbohydrate-restricted diet	
21	kostrekommendationer	diet recommendations	+
14	hälsokostaffär	health food store	
14	lågfettkost	low fat diet	+++
14	kostintaget	diet intake	+
13	hälsokosten	the health food	
13	koster	diets	∗
13	hälsokostbutiker	health food stores	
12	kostfaktorer	diet factors	+
11	kosttillskottet	the dietary supplement	
10	kostvanorna	the food habits	+
10	vegankost	vegan diet	
10	kostintervention	diet intervention	+

12.3 Remove one and keep one

After the initial survey was done, where the effectiveness of individual description terms was investigated in the context of the collection, search keys were instead examined in the context of other search keys, and the facets were investigated in the context of other facets as suggested by Pirkola and Järvelin (2001). These experiments were done using the baseline queries described in chapter 11.

Two approaches were used: (1) Which facets/search keys have an impact on the result? Is the result better/worse if they are removed from the query? (2) Which facets/search keys perform well on their own?

The first approach began by doing baseline runs. When the facets were investigated, the baseline query included all facets of a topic, and when the search keys were investigated, the baseline query included all terms of a facet. The idea was to remove the facet or search key which was being examined, then make a new run and see how the removal affected the result. If the result had deteriorated, compared to baseline, this implied that the facet or search key contributed to a good result. If the result was better without including the facet or search key, this implied that it did more harm than good, retrieving noise. If removing a facet or key did not affect the result at all, this was an indication that the concept or term had low frequency in the collection and did not retrieve any significant number of relevant or non-relevant documents. A small change in the results could be an indication of a number of things. The facet or search key could have such a low frequency that the impact was minor, resulting in some hits, but also noise, so that the new hits and the new noise, more or less, cancelled each other out. It could also be the case that the facet or term made a good query, and performed well on its own, but that, in the documents, it often co-occured with another facet or search key that performed equally well on its own, retrieving the same documents. If that second facet or key was still used in the query, the impact of removing the first one would not be as great as could have been expected considering the good result of using the search it on its own.

Figures 55 and 56 show the results of this type of runs for the facets in topics 4 and 23, and figures 57 and 58 show similar results for the terms within one chosen facet for each of these topics. The results are shown in normalized discounted cumulated gain at document cut off value 20. Only one DCV point is used in order to make a more distinct visualization. As it is both a cumulated and a discounted value, DCV 20 comprises the results at lower DCV points.

In the histograms, the leftmost dark bars represent the runs of the baseline queries, with all facets or all terms respectively. To the right of the baseline bar are pairs of bars corresponding to the different facets, or the different terms

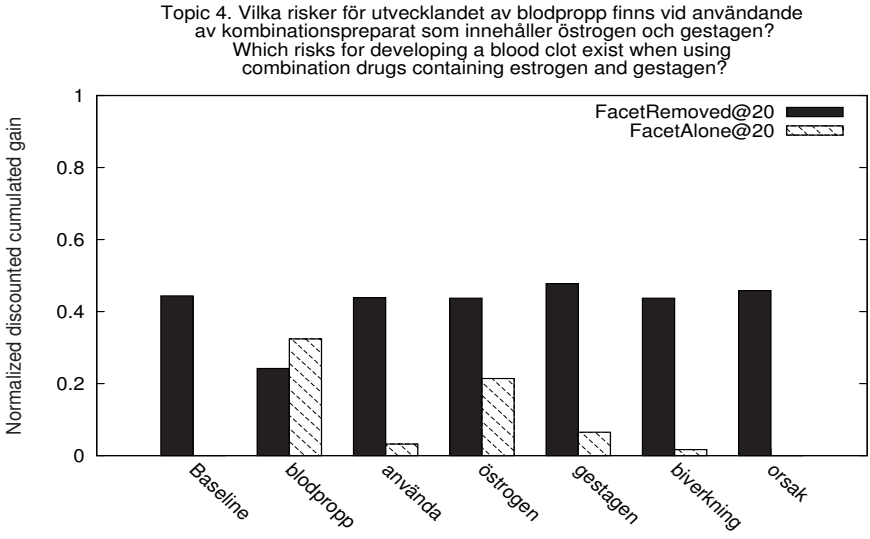


Figure 55: The facets of topic 4. The first dark bar represents the baseline query. The following dark bars are the nDCG results at DCV 20 of running queries where the facet examined is removed from the baseline query. The light bars are the results of runs using the facets alone. The facet equivalents are from the left: [blood clot], [use], [estrogen], [gestagen], [side effect], and [cause].

of a facet. The dark bars are the results of running a query where the facet investigated, or the term investigated, is removed. The light bars are the results of runs using only the facet in question or the term in question.

The facets that were chosen for the term investigations in figures 57 and 58 were [estrogen] and [neuroleptics]. These two facets were among the best keys for their topics, they also have terms in common and therefore can demonstrate that the same search key can behave differently for different topics.

In figures 55 to 58 an effective query has a low dark bar, indicating that the result deteriorated when the facet or search key was removed from the query, or a tall light bar, which indicates that the search key could retrieve relevant documents used alone in a query. It is not necessary to have both a low dark bar and a tall light bar.

A non-effective query has a tall dark bar or a low or empty light bar. A dark bar that is taller than the baseline bar indicates that the facet or search key in question retrieved additional non-relevant documents making the nDCG value lower. The query was more effective without the term or facet. For several facets and search keys there is no light bar. This means that the corresponding query did not retrieve any relevant documents at DCV 20. It does not neces-

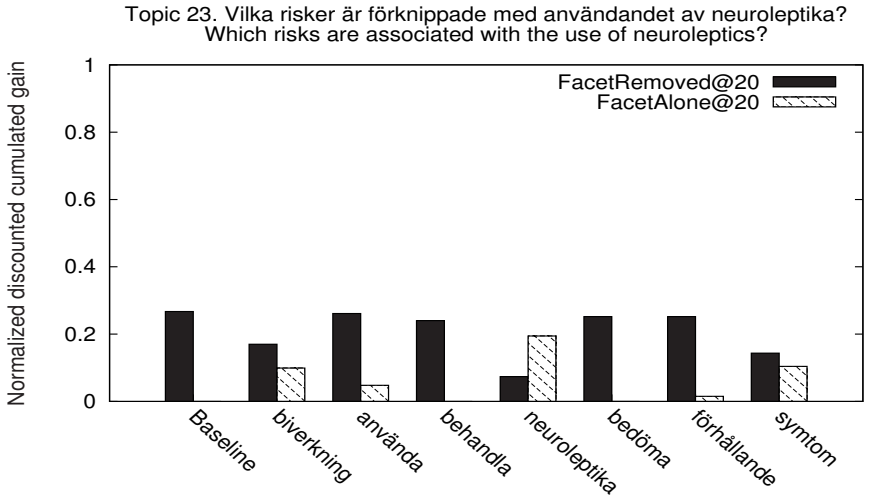


Figure 56: The facets of topic 23. The first dark bar represents the baseline query. The following dark bars are the nDCG results at DCV 20 of running queries where the facet examined is removed. The light bars are the results of runs using the facets alone. The facet equivalents are from the left: [side effect], [use], [treat], [neuroleptica] [assess], [relation], and [symptom].

sarily mean that the facet or search key is bad. It may very well be effective in combination with other facets or search keys, improving the result when added to the query. In fact, the facet or search key may very well have retrieved relevant documents when used alone, but at higher DCV than 20.

As indicated by the comments above, there is not a direct relationship between the heights of the individual bars and the effectiveness of the corresponding query. The dark and the light bars have to be considered together and in the context of, not only the baseline query, but also the other facets or terms. As usual, it is the relative result that is interesting, not the absolute result.

Some reflections can be made studying the facet histograms. The facets that clearly turned out not to be effective are [cause], [treat], and [assess]. These all represent general concepts which could apply to a very large number of topics. They are not good at discriminating documents on different subjects from each other. None of them retrieved any relevant documents at DCV 20. For topic 4, the result improved slightly when the [cause] facet was removed, implying that this facet contributes to retrieving more noise than relevant documents. For topic 23, the result deteriorated slightly when the facets [treat] and [assess] were removed. This means that the facets actually did contribute slightly, but there are definitely other candidates that perform better.

For topic 4, the most effective facets were [blood clot] and [estrogen], which are concepts very relevant to the topic. Surprising is that [gestagen] does not get particularly good results. Removing that facet gives better results than keeping it. This indicates that this facet retrieves noise when used and does not retrieve enough hits relevant to this facet to compensate. However, the facet [gestagen] does retrieve relevant documents when used on its own. One reason that the result does not deteriorate when the facet [gestagen] is removed is probably that this facet has a big overlap with the [estrogen] facet. These facets are identical except for the terms *östrogen* and *gestagen* themselves. These terms also have related meanings, both denotating female hormones. Not least important is that the information need explicitly expresses that relevant documents discuss *östrogen* and *gestagen* in combination. Thus many of the relevant documents that contain terms from the [gestagen] facet are very likely also retrieved by the [estrogen] facet.

For topic 23 the baseline itself is quite low, probably because all facets, except [neuroleptics], are very general. There are two facets besides [neuroleptics] that give somewhat good results: [side effect] and [symptom]. Even though these two facets are general, they are less general than other facets, but most important, they are the facets that, besides [neuroleptics], are the most salient for the topic. They are also on the same level of specificity as the topic.

In figures 57 and 58 the behavior of terms in similar facets are compared. The facets examined are both concepts of [medicine/drugs]. However, the specific drugs are different, and the drugs are named on different levels of specificity. In topic 4, there is a specific name of a hormone substance, *östrogen*, while, for topic 23, the terms *neuroleptika* or *antipsykosmedel* 'anti psychosis substance' are expressions for a family of medicines.

For both facets the most specific terms, the ones denoting the kind of medicine, give the best results. There are three terms, *agens* 'agent', *medicinsk behandling* 'medical treatment' and *preparattyp* 'preparation type', which, for both facets, give results similar to baseline when removed and no hits at all when used alone. This is an indication that the terms have low frequency in the collection alternatively very high. The terms *preparat* and *läkemedel* 'medicine' give bad results for topic 4 but relatively good results for topic 23. These terms have very high frequencies in the MedEval collection, which explains why they bring noise to topic 4. It is interesting to see that they are quite effective for topic 23. A hypothesis is that it is because this topic is on a more general level of specificity. No specific substance is asked for in topic 23, only a family of drugs. Topic 4, on the other hand, names a specific substance: *östrogen*.

These results are in line with what was stated already in the MEDLARS experiments, that for good results it is important that information needs and queries are on the same level of specificity (see section 3.1.2).

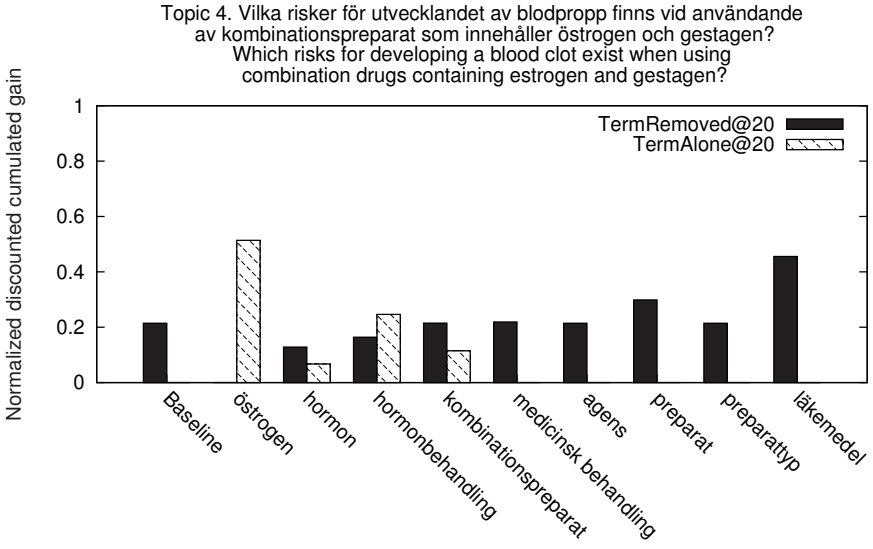


Figure 57: Examining terms of the drug facet of topic 4. The terms are in order: 'estrogen', 'hormone', 'hormone treatment', 'combination drug', 'medical treatment', 'agent', 'preparation', preparation type', and 'medicine'.

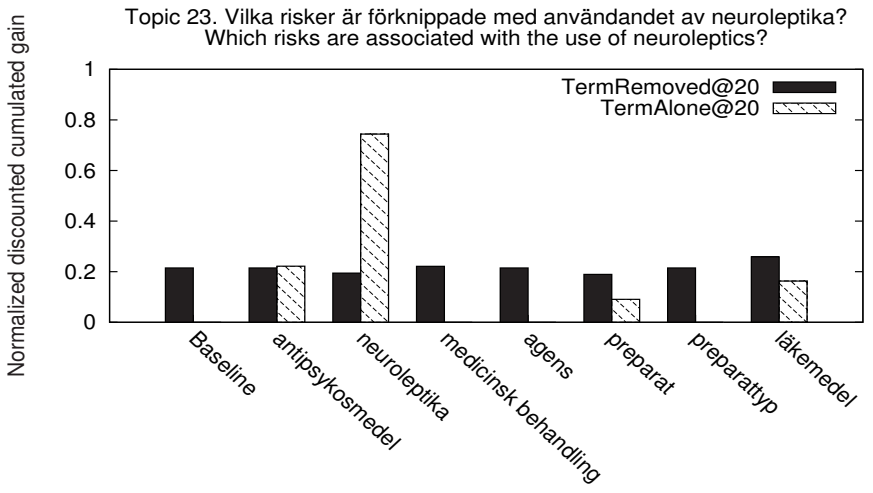


Figure 58: Examining terms of the drug facet of topic 23. The terms are in order: 'antipsycotic agent', neuroleptics', 'medical treatment', 'agent', 'preparation', preparation type', and 'medicine'.

12.4 Merging facets

Considering the facts that the facets [estrogen] and [gestagen] are so similar, and that the information need describes a situation were the two substances are used in combination, the question arises if these two terms in fact should be put in the same facet. Allotting similar terms to different facets, while at the same time expanding both these facets with all present synonyms and near synonyms (as described in chapter 11) give such synonym terms double (or more) impact when there are two (or more) similar terms in one topic. To see what the effect would be of treating the two female hormones, in topic 4, as synonyms, the facets for these two concepts were merged into one. In this new facet the synonyms and near synonyms only occur once each, giving them a more reasonable impact. The separate facets and the merged facet are shown in figures 59 and 60.

```
#syn(östrogen hormon hormonbehandling kombinationspreparat
#uw5(medicinsk behandling) agens preparat preparattyp läkemedel)

#syn(gestagen hormon hormonbehandling kombinationspreparat
#uw5(medicinsk behandling) agens preparat preparattyp läkemedel)
```

Figure 59: The original two facets for the named hormones in topic 4. Several near synonyms appear in both facets, and thus have double impact.

```
 #syn(östrogen gestagen hormon hormonbehandling
kombinationspreparat #uw5(medicinsk behandling) agens preparat
preparattyp läkemedel)
```

Figure 60: The merged facet for the two named hormones in topic 4. Now each near synonym occurs only once.

The effect of having a combined facet for the two terms are shown in the histogram of figure 61, which corresponds to the histogram in figure 55, but now with [estrogen] and [gestagen] merged into one facet. There are two main effects: the baseline is on a higher level and it is clear that the query is less effective when the combined drug facet is removed.

New runs were also done examining the impact of the terms in the new [estrogen/gestagen] facet of topic 4. The results can be seen in figure 62. The named entities *östrogen* and *gestagen* were run separately as one search key queries, but also in a query were they were combined by use of the synonym operator. All of these three runs were very effective. Most effective was the

run were the terms were treated as synonyms. Interesting to see is that the search key *gestagen* gave better results than the search key *östrogen*. In the runs with separate facets [gestagen] facet performed very poorly (see figure 55), noteably worse than the [estrogen] facet. This happens in spite of the fact that the only difference between the facets is the search key with the name of the hormone. What this shows is that there are many factors that affect the results. In this case, one important factor is the document frequencies of the two terms. The term *gestagen* occurs in only 68 documents, the term *östrogen* in 435 documents. Bearing in mind that topic 4 only has 132 documents of relevance 1, 2 or 3, the number of non-relevant documents containing the term *östrogen* is significantly larger than the number of non-relevant documents containing the term *gestagen*. When *gestagen* is used alone, as a single search key query, only 68 documents are retrieved, but the precision is high. When *östrogen* is used alone, six times as many documents are retrieved. Since there is nothing that says that documents containing both terms should be ranked in better positions, the precision is, with high probability, considerably lower. On the other hand, when the complete facets are used, it is a different situation. Now the low number of *gestagen* terms gets mixed with the high frequency terms such as *läkemedel*, and the documents that contain *gestagen* are to few to maintain the high precision. The *östrogen* terms, with their higher frequency succeed better.

Additional runs were done using the #synonym() and the #combine() operators on the *östrogen* and *gestagen* search keys and facets. The results are shown in figure 63. Six different runs were made, two with single search key queries, one with the search keys united with the synonym proximity operator, treating them as instances of the same term, and one with the combine belief operator, treating them as two different terms. Finally, runs were made using the complete facets. In one run *östrogen* and *gestagen* were regarded as synonyms and used in one common facet, as shown in figure 60, and in the other they were put in the two separate facets, as shown in figure 59. These two facets were combined with the combine operator. In figure 63 the results of these six runs are shown up to DCV 40, after which the values did not change much.

In figure 63 we can see that the results using the *östrogen* and *gestagen* search keys individually gave quite good results, using them together, even better. However, something that gave poor results was using the complete facets. Using the combine operator on the two facets was not effective at all, the first document with any relevance was found at DCV 7. Treating the terms *östrogen* and *gestagen* as synonyms and putting them in the same facet, gave considerably better results. In this case the near synonyms with high frequency and low specificity did not have as much impact as they did when using the combine operator on the queries. This is consistent with Strzalkowski's statement:

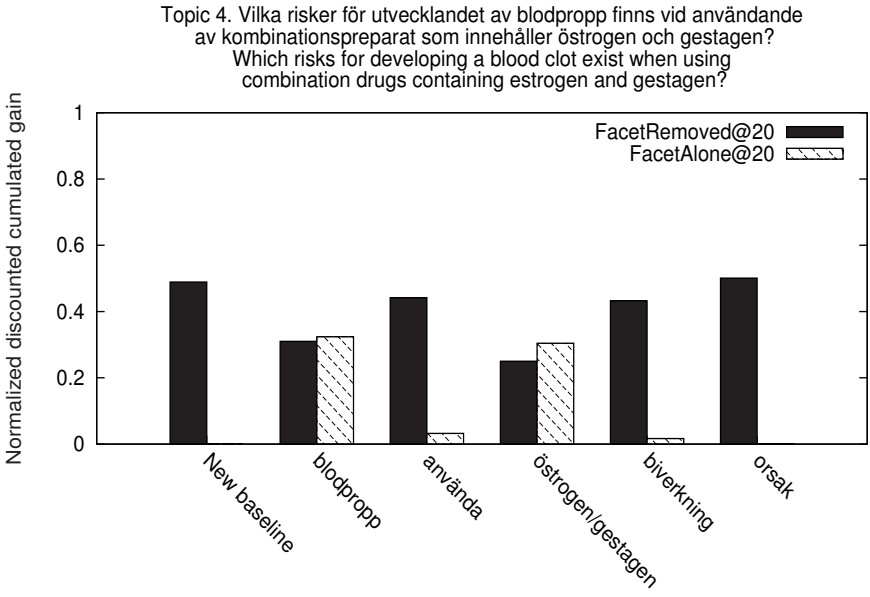


Figure 61: A new baseline merging the facets [estrogen] and [gestagen]. The facet equivalents are from the left: [blood clot], [use], [estrogen/gestagen], [side effect], and [cause].

> [R]emoving low-quality terms from the queries is at least as important (and often more so) as adding synonyms and specializations. (Strzalkowski 1994: 319)

As for the single search key queries *östrogen* has very high nDCG values at low DCV levels, *gestagen* not so good. At DCV 14 the situation changes and at higher DCV levels it is *gestagen* that has the better nDCG values. The best results at low DCV values are obtained by using the synonym operator on the two single search keys. The run using the combine operator on the two search keys has a weaker start, but catches up. Between DCV 14 and 25 the curves are quite even, but after DCV 25 it is the query which uses the combine operator that has the best result. To decide if the likeness between these two pairs of queries is a coincidence or not, more investigation is needed.

It would be interesting to research further on how to best treat facets which are as similar as [östrogen] and [gestagen] described here. There are other ways to go than the ones described here. Using the weight operator, **#weight()**, is one example. One could then give high weight to specific or low frequency terms and low weight to less specific or more frequent terms. In the example illustrated here *östrogen* and *gestagen* would have high weights, while the near syn-

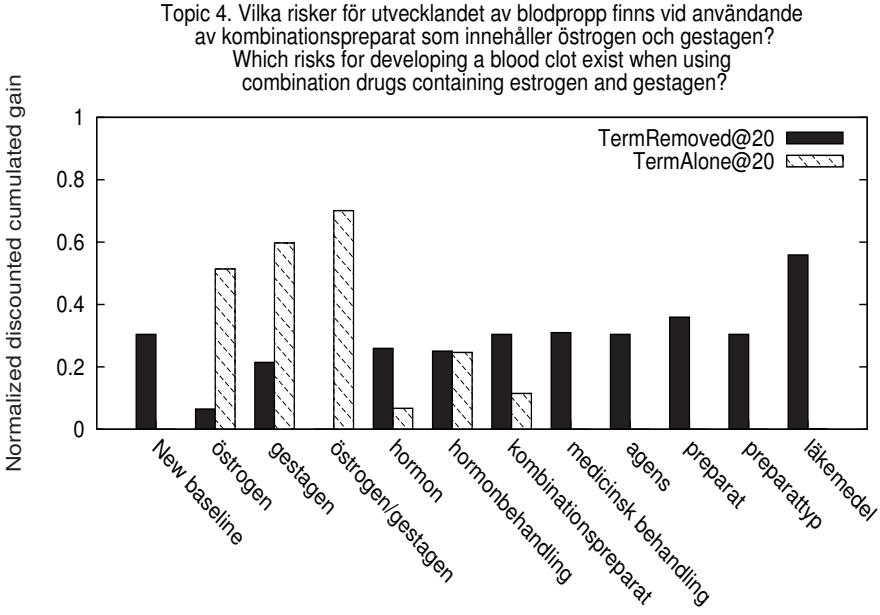


Figure 62: A new baseline merging the facets [estrogen] and [gestagen]. The term equivalents are from the left: 'estrogen', 'gestagen', 'estrogen/gestagen', 'hormone', 'hormone treatment', 'combination drug', 'medical treatment', 'agent', 'preparation', 'preparation type', 'medicine'.

onyms for 'medicine/drugs' would get low weights. This may make it possibile to keep two facets, but not give to much impact to the high frequency/non-specific terms.

12.5 Term variation

When documents are indexed manually by a professional indexer who is choosing index terms from a controlled vocabulary, the actual wording which the author employs when writing a document is not as important as when the full texts of the documents are indexed automatically. Authors not only vary their choice of terms or ways to phrase concepts, they also choose different spellings. In Swedish documents, these differences can be due to accepted variants, influences from foreign languages, often English, abbreviations or spelling mistakes. For example, use of English words instead of Swedish words, or use of English spelling in Swedish words: 'ph' instead of 'f': *lymfo→lympho*, 'th' instead of 't': *torax→thorax*, 'c' instead of 'k': *bradykardi→bradycardi*

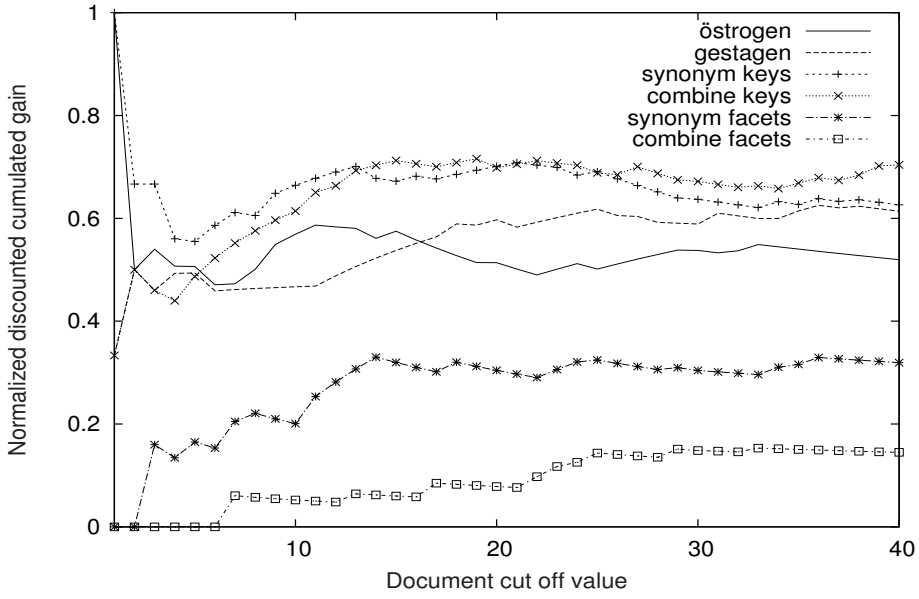


Figure 63: Comparing nDCG curves for topic 4 using synonym and combine operators. The terms *östrogen* and *gestagen* are used as single search key queries. They are also used together, in one query with the synonym operator, and in another with the combine operator. Two queries were constructed using complete facets. First *östrogen* and *gestagen* were regarded as synonyms and used in one common facet, as in figure 60. Finally in two separate facets as in figure 59.

or an overuse of hyphens: *tetracyklin→tetra-cyklin* (Kokkinakis 2009). Kokkinakis illustrates the variety that can occur for a single term with eleven different spellings found in MedLex for diarré 'diarrhea': *diarr*, *diarre*, *diarre´*, *diaré*, *diarrè*, *diarree*, *diarrée*, *diarrhea*, *diarrhoea*, *diarrhorea*, and *diarrre*.

In medical texts abbreviations are quite common. This is often the case for Latin or Greek anatomical terms and Latin or Greek names of organisms. Kokkinakis (2009) gives some examples, shown in table 12.3, from the MedLex corpus. With this in mind it can, in some cases, be a good idea to relax the queries using different spelling variations within the #syn() operator. Abbreviations commonly used in phrases motivate search on individual phrase constituents in combination with the complete phrases. Caution should be taken not to do this if the individual constituent has a very high collection frequency.

Table 12.3: Abbreviational varieties found in the MedLex corpus and their counterparts in the Swedish MeSH.

MedLex text examples	Swedish MeSH terms
n. abducens	Nervus abducens
v. cava inferior, v.cava inf.	Vena cava inferior
a. iliaca, art iliaca	Arteria iliaca
staph aureus, s. aureus	Staphylococcus aureus
h pylori, h. pylori	Helicobacter pylori
c. pneumoniae	Chlamydia pneumoniae

12.6 Reflections

Treating compounds in information retrieval in a uniform manner, for example with decomposition, does not always give the desired results. Seeing compounds as the diverse entities that they are, and treating them accordingly should give better results. Even though this thesis does not contain thorough investigations of the different situations, the pilot studies suggest that there are several methods that probably would give good results, and that should be investigated further.

- Before deciding to split a compound, in index or query, check the tf$*$idf values of the constituents to see if the constituents have greater resolving power than the compound as a whole. Use only the constituents that do, that is, use only constituents that are closer to the middle range of collection frequencies than the original compound. If a constituent is salient to the topic and has a very low frequency, there is little risk in using it, even if it is further away from the middle range than the orignal compound.

- When decomposing a compound/derivation on the form [N Adj] usually only one of the constituents is useful. The original word could be of the type *blixt*||*snabb* 'lightning fast' where the noun has the function of reinforcing the adjective which here is the useful constituent, or of the type *fett*||*snål* where the adjective has the function of quantification and making an adjective of the noun, and here it is the noun that is useful. If adjectives of this form are decomposed, only one constituent should be used further.

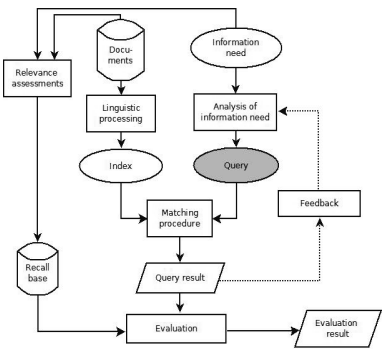
13 SEARCH KEY BEHAVIOR

It is very easy to be blinded to the essential uselessness of them by the sense of achievement you get from getting them to work at all.

So Long... chapter 35

13.1 Basic behavior

The overall research question for this thesis is: what features do good and what features do bad search terms have? To answer that question we must first of all know what we mean by good and bad respectively. The answer may at first seem obvious. But there are many ways in which a run can be successful or unsuccessful and several ways in which a search key can improve a result, or make it worse. It is quite safe to say that a good result is one with high precision and high recall. The problem is that these two measures stand in a trade-off relationship to each other. An increase of recall often entails a decrease in precision, and vice versa. Combining a baseline query with new search keys that retrieve additional relevant documents often results in more noise as well.



The figures 64 through 70 demonstrate some typical search key behaviors. The ellipse in figure 64 represents a certain recall base, the documents which the user aims to retrieve. This recall base recurs in the figures that follow. The ellipses added in figures 65 through 70 represent the result sets of queries consisting of one search key each, the search keys Key 1 through 6.

Key 1, in figure 65 represents a typical query, retrieving a fair amount of relevant documents, but also some noise. Key 2, in figure 66, is less successful,

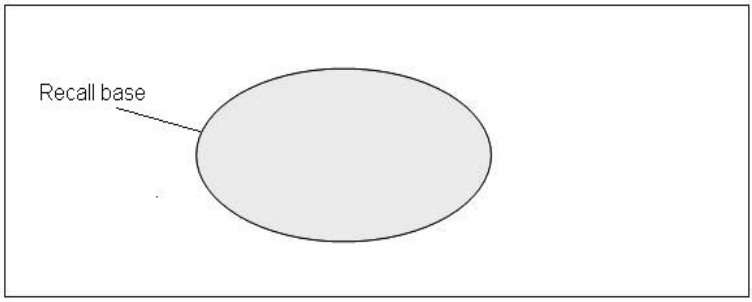


Figure 64: The rectangle above represents the universe of documents. Referring to MedEval it would be the 42 000 documents of the MedEval collection. The ellipse represents the recall base, the documents, relevant to a certain topic, that the user wants to retrieve. This recall base recurs in figures 65 through 70.

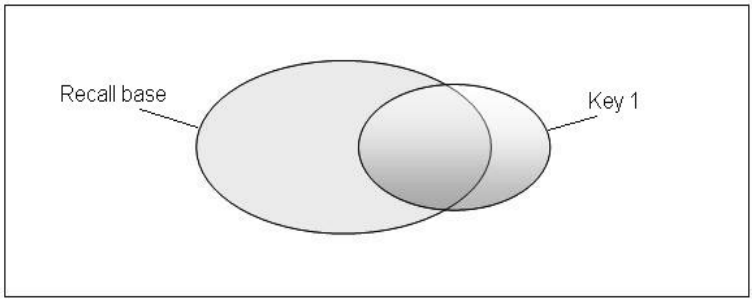


Figure 65: The big ellipse represents the recall base, and the small ellipse the result of a query containing one search key, Key 1. In this example Key 1 retrieves a fair amount of relevant documents, but also some noise.

retrieving fewer relevant documents and more noise. Key 3, in figure 67, does not retrieve many documents, but the ones retrieved are all relevant. Comparing Key 1 and Key 3, which one is best? The one with higher recall, but lower precision? Or the one with very good precision, but not good recall? The answer depends on the needs of the user. In some cases the user only wants an answer to a question, and one relevant document is enough. Other times it is essential to find as many relevant documents as possible.

Expanding a query by combining additional search keys can bring on different effects. 'Combining' here implies using the #combine() or the #syn() operators. These function somewhat like disjunctions and make the result sets bigger. The difference between the operators is that #combine() treats the search

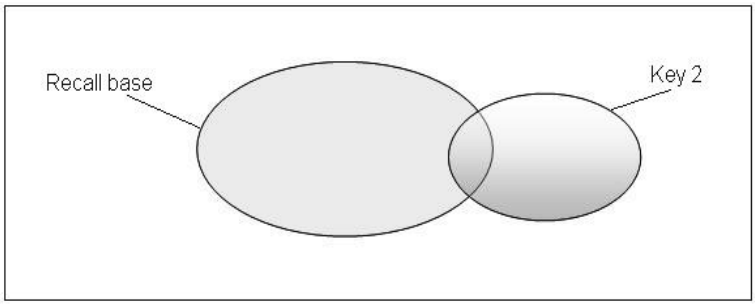


Figure 66: In this example, the search key Key 2, is not as successful as Key 1. There are fewer relevant documents retrieved and more noise.

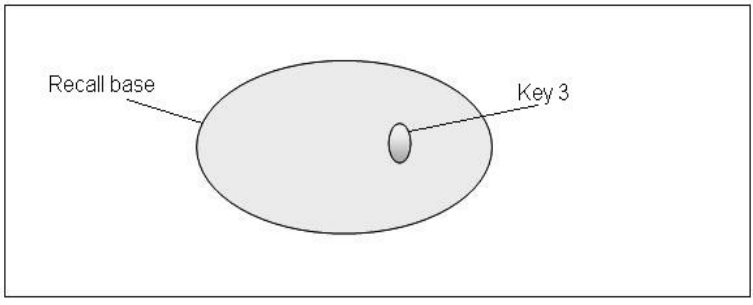


Figure 67: In the third example, Key 3 retrieves only a few relevant document, but no noise. This could be a good search key to use as a one search key query for a user who is satisfied with few documents, and who does not have the patience, or the time, to search through noise, a user who wants a quick answer to a certain question.

keys as two different terms, while #syn() treats them as the same term, which then would have a larger term frequency (providing that the terms exist in the collection), and also a larger document frequency (provided that the terms do not always co-occur in the same documents). This affects the tf∗idf factor.

Some effects are demonstrated in figures 68 through 70. In figure 68 Key 4 and Key 1 have similar result sets, when used separately as one search key queries. Combining these keys will not have much effect on the result. It would retrieve some additional relevant documents and some more noise. However, it would rearrange the ranking of the documents, as documents containing both keys would be moved up in the ranking, while documents containing only one would be moved down.

In figure 69 Key 5 has a small result set. This set is completely separated from the result set of Key 1, but most importantly, it retrieves no noise. Both

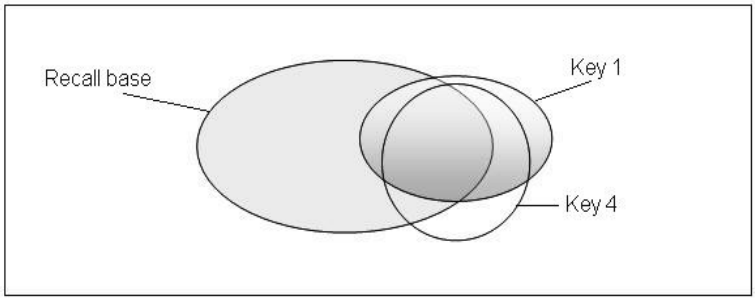


Figure 68: In this example we go back to using Key 1 and then expand the query with Key 4. Key 4 alone would retrieve more or less the same documents as Key 1 alone, and so would the combination. However, the order in the ranked result list would be affected.

these properties makes Key 5 a good key for expanding the query, at least it is harmless, as it improves the result without cost, cost being noise. If Key 5 would be good to use on its own in a query depends on the need of the user, just as for Key 3 above. The behavior of Keys 3 and 5 is typical of very specific search keys that match the information need but have low frequency in the collection and therefore only retrieves a small number of documents.

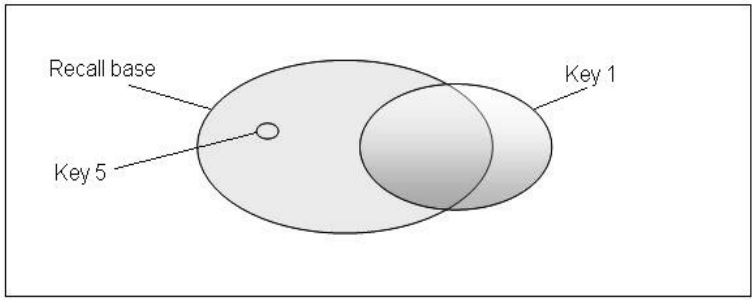


Figure 69: Combining instead Key 5 with Key 1 retrieves relevant documents not in the first result set, and adds no noise. This would improve the results. Maybe not by much, but without cost, as there will not be additional non-relevant documents retrieved.

If we instead expand Key 1 with Key 6 as in figure 70, more relevant documents would be retrieved than with Key 5, but again, a considerable amount of noise. Again, the documents where both search keys are found would move up in the ranking. In this case, there are both relevant and non-relevant documents

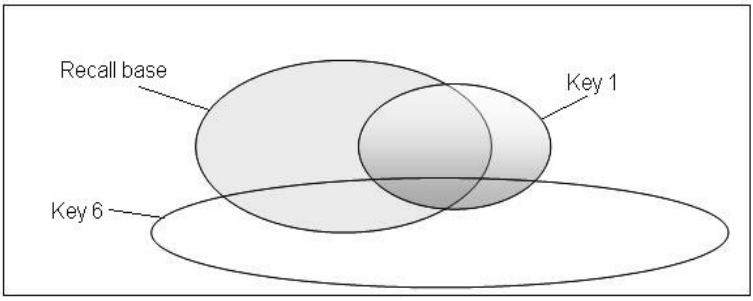


Figure 70: Key 6 adds new relevant documents to the result set, but also a considerable deal of noise. If adding Key 6 is beneficial or not depends on the needs of the user: if it is important to retrieve many relevant documents, or if one or a few documents are sufficient and the user does not want to or have time to search through noise.

where both keys occur. As before, it is difficult to determine which result is the better. It depends on how the ranking turns out and it also depends on the user's needs, and the user's patience.

13.2 Effective vs. ineffective search keys from the topic descriptions

Chapter 12 describes a survey done on the content words of the topic descriptions. The the best and the worst search keys of these runs were gathered in two tables. Table 13.1 contains information about the terms that gave no recall whatsoever at DCV 1 000 and table 13.2 lists the terms that were most effective, the high resolution power keys (see section 6.4). For topics where no search key stood out as rendering significantly better results than other terms, nothing is listed in the table. The term *läkemedel* 'drug/medicine' occurs in both tables, illustrating that a search key can be effective for one topic while giving few or no hits for another, as shown in figures 57 and 58. In the tables the terms are presented in combination with the corresponding topics, as the effectiveness of a term always is in relation to a topic. There are terms which are listed for more than one topic in one table. These entries contain different numbers for the recall base, but otherwise the numbers are identical. Below are explanations for the labels of tables 13.1 and 13.2.

Collection frequency The number of occurrences of the term in the MedEval collection such as it is in the lemmatized non-decomposed index.

Document frequency The number of documents in which the term occurs, counting the number of documents retrieved by the Indri search engine.

tf in docs with term 'Collection frequency'/'Document frequency'. The average number of term occurrences in the documents that actually contain the term. This indicates the clustering tendencies of the term.[36]

idf Inverse document frequency. This corresponds to how much weight each term occurrence is given by the ranking algorithm. The weight is in a reverse relationship to the number of documents in which the term occurs.[36]

Average tf∗idf 'tf in docs with term' × 'idf'. Corresponds to the average total sum of the weight the term is given in each document that contains the term.[36]

Recall base The size of the set of documents assessed to be relevant for the topic in question.

POS Part-of-speech of the term: Adj – adjective, N – noun, V – verb, No – number, NE – named entity, Adv – adverb.

When two POSes are stated, the term is ambiguous. The first POS is the one that applies to the term as it stands in the current information need. The MedEval indexes do not contain lexical information so the search engine will match occurrences of both kinds.

Complexity The structure of the term: C – compound, S – simplex word, FC – first constituent of a compound or phrase, LC – last constituent of a compound or phrase, CC – complex compound, where one constituent is a compound itself, CV – separable compound verb, Phr – fixed phrase, Acr – acronym.

There are some differences between the effective and the ineffective terms that are easy to see. In the table with effective terms all terms except two are nouns, several of these are named entities. The two terms which are not nouns are, in fact, the adjective part, *instabil* 'unstable' and *transitorisk* 'transient', of phrasal names of illnesses, *instabil angina pectoris* 'unstable angina pectoris' and *transitorisk ischemisk attack* 'transient ischemic attack'. The variation of

[36]The Indri/Lemur search engine does not use the tf∗idf factor directly, but the numbers are interesting as an indication of the clustering tendencies. In fact, Metzler Jr. (2007) describes the Indri/Lemur #combine() operator as producing an idf effect. As there are no phrasal index terms in the MedEval indexes the tf and idf based values for the phrases in the tables reflect the values that a single term with the same number and clustering of occurrences would have obtained.

part of speech in the other table is noticeably greater. Here are adjectives, adverbs and even a number. There is only one ineffective term labeled as a named entity: *johannes*. Even though *Johannes* is a man's name this occurrence is a special case as it is a constituent of a lexicalized compound *johannes||ört* 'John's herb' (St. John's wort), and does not refer to anyone called Johannes.

The frequency of the terms vary considerably in the table with ineffective search keys, from 0 to 32 917 tokens. In the other table the variation is more moderate, from 8 to 15 258. The latter number is for the term *läkemedel* 'medication/drug', which occurs in both tables. The second highest frequency count is less than half of this, 7 064 for the term *smärta* 'pain'.

The bottom rows of the tables give an overview. The average frequency count for the effective search keys is 29%, less than a third, of the average of that for the ineffective keys. The median numbers show a smaller, but still significant, difference. The median frequency of the good keys is 50% of that of the bad keys. The standard deviation is significantly higher for the poor search keys, which implies a big variation in frequencies, which is expected as terms with very high or with very low frequencies seldom have much resolving power.

The frequency numbers agree with rather obvious presuppositions about what can make a search key ineffective even if it is relevant to a topic. If there are no occurrences of a term in the collection, the term as a search key cannot render any hits. A single term can't match more documents than there are occurrences of the term. Few occurrences means few hits. On the other hand, if there are significantly more occurrences of a term in the collection than there are relevant documents, the risk is high that the relevant documents, even if they are retrieved, will come mixed with non-relevant documents and end up too far down in the ranking for a good result.

The relation between the document frequency and the collection frequency, mirrors the likelihood of a term, having appeared once in a document, appearing again in the same document. A clustering of a term in a document is, according to Luhn (1958), an indication that the term is important for the contents of that document. The clustering tendency is directly represented in the column with the average term frequency of the term in the documents that actually contain the term. The average term frequency in these documents for the ineffective terms is 1.5, while it is 2.4 for the effective terms. This means that the poor search keys do not have the same tendency to cluster as the effective ones.

Inverse document frequency is an indication to give more weight to terms that have a tendency to cluster. The tables show that the effective terms have an average idf of 2.5 and the ineffective terms have an average idf of 2.0.

The column showing average tf∗idf reflects the average total weight that is given to the term in question for each document where the term occurs. The value is high for terms that occur many times in few documents. A high value indicates that the term will have more impact in the queries where it is used.

The numbers in the column, showing the recall base for the topics to which the terms belong, are rather an indication of how difficult or specific the topics are, not how good or bad the search key is. Remember that the table of poor terms contain search keys that did not give any recall at all for the first 1 000 documents in any index. The outcome of a search is complex, dependent on many factors. Whatever the reason, the average and median numbers for the recall base for the topics corresponding to the poor search keys were only half the size of the recall bases corresponding to the effective terms.

Table 13.1: Ineffective search keys among terms and compound constituents found in the topic descriptions.

Topic	Term	Collection frequency	Document frequency	tf in docs with term	idf	Average tf∗idf	Recall base	POS	Complexity
1	energifattig	3	2	1.5	4.3	6.5	33	Adj	C
2	komplikation	2 799	1 311	2.1	1.5	3.2	4	N	S
2	uppstå	3 481	2 418	1.4	1.2	1.8	4	V	CV
2	behandling	32 917	10 690	3.1	0.6	1.8	4	N	S
2	infektion	6 700	3 619	1.9	1.1	2.0	4	N	S
4	utvecklande	120	101	1.2	2.6	3.1	30	N/V	CV
5	indicerad	305	164	1.9	2.4	4.5	9	Adj/V	S
5	börja	9 115	6 148	1.5	0.8	1.2	9	V	S
9	fetma	1 190	463	2.6	2.0	5.0	2	N	S
9	blodtryck	4 102	1 463	2.8	1.5	4.1	2	N	C
9	blod	6 408	2 926	2.2	1.2	2.5	2	N	FC
9	tryck	1 673	990	1.7	1.6	2.8	2	N	LC
10	väga	915	734	1.2	1.8	2.2	6	N/V	S
10	ii	1 032	561	1.8	1.9	3.5	6	No	LC
11	viktminskningspreparat	3	3	1.0	4.1	4.1	13	N	CC
12	verkningsmekanism	255	193	1.3	2.3	3.1	10	N	C
12	verkning	43	37	1.2	3.1	3.6	10	N	FC
12	mekanism	1 398	804	1.7	1.7	3.0	10	N	LC
12	fettupptag	2	1	2.0	4.6	9.3	10	N	C

Continued on next page...

Table13.1- continued from previous page.

Topic	Term	Collection frequency	Document frequency	tf in docs with term	idf	Average tf∗idf	Recall base	POS	Complexity
13	indikation	1 960	840	2.3	1.7	4.0	3	N	S
13	använda	21 734	9 222	2.4	0.7	1.6	3	V	S
13	läkemedel	15 258	4 755	3.2	0.9	3.0	3	N	C
13	läke	9	21	0.4	3.3	1.4	3	V	FC
13	medel	3 635	2 045	1.8	1.3	2.3	3	N	LC
13	adhd-behandling	2	2	1.0	4.3	4.3	3	N	C
13	behandling	32 917	10 690	3.1	0.6	1.8	3	N	LC
13	kriterium	812	445	1.8	2.0	3.6	3	N	S
16	klimakterium	39	26	1.5	3.2	4.8	20	N	S
18	klimakterium	39	26	1.5	3.2	4.8	22	N	S
19	johannes	32	26	1.2	3.2	4.0	13	NE	FC
20	tillgå	143	129	1.1	2.5	2.8	23	V	CV
20	exempelvis	2 958	2 012	1.5	1.3	1.9	23	Adv	S
21	typ	8 544	4 723	1.8	1.0	1.7	10	N	S
26	använda	21 734	9 222	2.4	0.7	1.6	20	V	S
26	diagnostisering	2	2	1.0	4.3	4.3	20	N	S
27	alkoholavgiftning	0	0	0	0	0	3	N	C
32	indikation	1 960	840	2.3	1.7	4.0	13	N	S
37	yttra	512	468	1.1	2.0	2.1	32	V	S
37	karaktär	483	391	1.2	2.0	2.5	32	N	S
37	nyttjande	9	9	1.0	3.7	3.7	32	N/V	S

Continued on next page...

Table13.1- continued from previous page.

Topic	Term	Collection frequency	Document frequency	tf in docs with term	idf	Average tf∗idf	Recall base	POS	Complexity
38	lämpligt	10	9	1.1	3.7	4.1	8	Adv	S
38	reguljär	24	22	1.1	3.3	3.6	8	Adj	S
41	cancerrelaterad	26	21	1.2	3.3	4.1	17	Adj	C
42	virushämmare	0	0	0	0	0	47	N	C
44	uppvisa	861	720	1.2	1.8	2.1	11	V	CV
49	utvecklande	120	101	1.2	2.6	3.1	30	N/V	CV
50	typ	8 544	4 723	1.8	1.0	1.7	7	N	S
51	drabba	9 245	5 286	1.7	0.9	1.6	23	V	S
54	utsätta	2 138	1 656	1.3	1.4	1.8	9	V	CV
54	lämplig	1 592	1 137	1.4	1.6	2.2	9	Adj	S
56	tillämpa	657	48	7 1.3	1.9	2.6	24	V	S
63	teknik	1 993	1 074	1.9	1.6	3.0	10	N	S
63	redskap	104	89	1.2	2.7	3.1	10	N	S
63	använda	21 734	9 222	2.4	0.7	1.6	10	V	S
63	cancermisstanke	3	3	1.0	4.1	4.1	10	N	C
66	behandla	9 760	5 594	1.7	0.9	1.5	9	V	S
66	följd	1 662	1 906	0.9	1.3	1.2	9	N	S
73	diagnos	7 533	4 066	1.9	1.0	1.9	9	N	S
76	recidiv	557	178	3.1	2.4	7.4	3	N	S
77	sepsissjukdom	0	0	0	0	0	18	N	C
77	blodtrycksförändring	6	5	1.2	3.9	4.7	18	N	CC

Continued on next page. . .

Table13.1- continued from previous page.

Topic	Term	Collection frequency	Document frequency	tf in docs with term	idf	Average tf∗idf	Recall base	POS	Complexity
85	överdosering	148	86	1.7	2.7	4.6	7	N	C
90	behandla	9 760	5 594	1.7	0.9	1.5	9	V	S
91	hemolysinducerad	0	0	0	0	0	5	Adj	C
96	innebära	7 674	4 768	1.6	0.9	1.5	14	V	S
100	störd	382	261	1.5	2.2	3.2	9	Adj/V	S
100	tsh-insöndring	0	0	0	0	0	9	N	C
Average		4 027	1 874	1.5	2.0	2.9	12		
Median		812	468	1.5	1.7	2.8	9.0		
StDev		7 312	2 853	0.7	1.2	1.7	9.7		

Table 13.2: Effective search keys among terms and compound constituents found in the topic descriptions.

Topic	Term	Collection frequency	Document frequency	tf in docs with term	idf	Average tf∗idf	Recall base	POS	Complexity
1	kolesterol	923	368	2.5	2.1	5.2	33	N	S
1	lipoprotein	108	26	4.2	3.2	13.3	33	N	C
2	erytomycin	125	66	1.9	2.8	5.3	4	NE	C
4	blodpropp	1 315	480	2.7	1.9	5.3	30	N	C
4	östrogen	1 256	433	2.9	2.0	5.8	30	NE	S
5	reductil	60	20	3.0	3.3	10.0	9	NE	S
11	reductil	60	20	3.0	3.3	10.0	13	NE	S
11	viktminskning	372	243	1.5	2.2	3.4	13	N	C
13	concerta	38	17	2.2	3.4	7.6	3	NE	S
16	hormonbehandling	407	181	2.2	2.4	5.3	20	N	C
16	klimakterie	435	241	1.8	2.2	4.1	20	N	S
19	interaktion	439	229	1.9	2.3	4.3	13	N	S
19	johannesört	81	35	2.3	3.1	7.1	13	NE	C
19	läkemedel	15 258	4 755	3.2	0.9	3.0	13	N	C
20	akne/acne	2 362	1 054	2.2	1.6	3.6	23	N	S
23	antipsykosmedel	8	3	2.7	4.1	11.1	59	N	CC
23	neuroleptikum	192	96	2.0	2.6	5.3	59	N	C
25	transitorisk ischemisk attack	35	28	1.3	3.2	4.0	37	NE	Phr

Continued on next page...

Table 13.2- continued from previous page.

Topic	Term	Collection frequency	Document frequency	tf in docs with term	idf	Average tf∗idf	Recall base	POS	Complexity
25	tia	137	68	2.0	2.8	5.6	37	NE	Acr
25	transitorisk	46	33	1.4	3.1	4.3	37	Adj	FC
26	kranskärlssjukdom	555	180	3.1	2.4	7.3	20	N	CC
28	bensodiazepin/ benzodiazepin	230	87	2.6	2.7	7.1	43	NE	C
31	ssri	268	170	1.6	2.4	3.8	16	NE	Acr
31	ssri-preparat	141	93	1.5	2.7	4.0	16	N	C
32	depression	2 428	944	2.6	1.7	4.2	13	N	S
36	waran	102	82	1.2	2.7	3.4	56	NE	S
38	viagra	181	81	2.2	2.7	6.1	8	NE	S
39	alzheimer/alzhiemers	1 036	322	3.2	2.1	6.8	24	NE	S
41	anemi	537	222	2.4	2.3	5.5	17	N	S
42	herpes	821	306	2.7	2.1	5.7	47	NE	S
44	urinvägsinfektion	610	287	2.1	2.2	4.6	11	NE	CC
46	vaccination	1 279	416	3.1	2.0	6.2	25	N	S
50	nsaid	567	300	1.9	2.1	4.1	7	NE	Acr
50	ulcus	138	64	2.2	2.8	6.1	7	N	S
53	alzheimer/alzhiemers	1 036	322	3.2	2.1	6.8	16	NE	S
62	smärta	7 064	2 706	2.6	1.2	3.1	19	N	S
63	magsäck	655	287	2.3	2.2	4.9	10	N	C
65	pollenallergi	198	76	2.6	2.7	7.2	20	N	C

Continued on next page...

Table 13.2- continued from previous page.

Topic	Term	Collection frequency	Document frequency	tf in docs with term	idf	Average tf∗idf	Recall base	POS	Complexity
68	trombos	395	130	3.0	2.5	7.6	45	N	S
73	appendicit	48	33	1.5	3.1	4.5	9	N	S
77	sepsis	227	108	2.1	2.6	5.4	18	N	S
82	kramp	528	319	1.7	2.1	3.5	51	N	S
92	eksem	6 712	2 198	3.1	1.3	3.9	36	N	S
96	instabil	423	162	2.6	2.4	6.3	14	Adj	FC
97	hjärtsvikt	2444	523	4.7	1.9	8.9	46	N	C
Average		1 162	418	2.4	2.4	5.8	24		
Median		407	180	2.3	2.4	5.3	20		
StDev		2 596	839	0.7	0.6	2.2	16		

13.3 Luhn revisited

Luhn (1958) claims that the words that best identify the contents of a document are in the middle frequencies. He illustrates this with a figure containing a Zipf curve combined with a curve representing resolution power (see figure 14 on page 76). To investigate how well the MedEval corpus correlates to Luhn's idea, the word frequencies of the collection were calculated and represented in a Zipf curve for MedEval. This is shown in figure 71 where the term representations are ranked according to frequency and the axes of the graph represent frequency and rank.

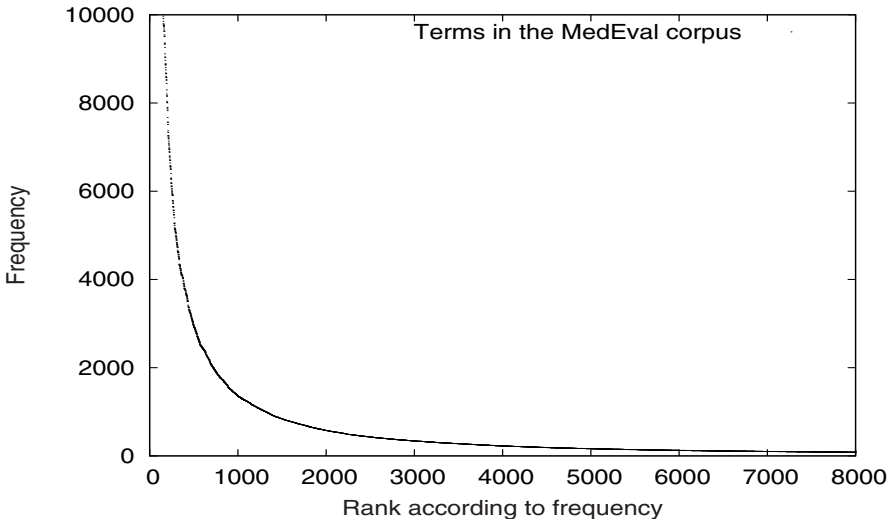


Figure 71: The Zipf curve resulting from frequency and rank of all terms in the MedEval corpus. The curve is cut off at frequency 10 000 and rank 8 000.

If Luhn's idea is applicable, the ineffective terms would be either at the leftmost end, where the curve is nearly vertical, or at the rightmost end where the curve is nearly horizontal. The effective terms would be in the middle range, where the curve changes direction from vertical to horizontal. In figure 72 the ineffective and the effective terms have been extracted from the rank-frequency table and inserted in separate graphs. The positions of the representations of ineffective and effective terms in the MedEval Zipf curve turn out to be very consistent with Luhn's ideas.

There are four terms among the effective ones which have such a high frequency that they are positioned on the nearly vertical part of the curve, where the ineffective words would be expected. These are seen in table 13.3. Note

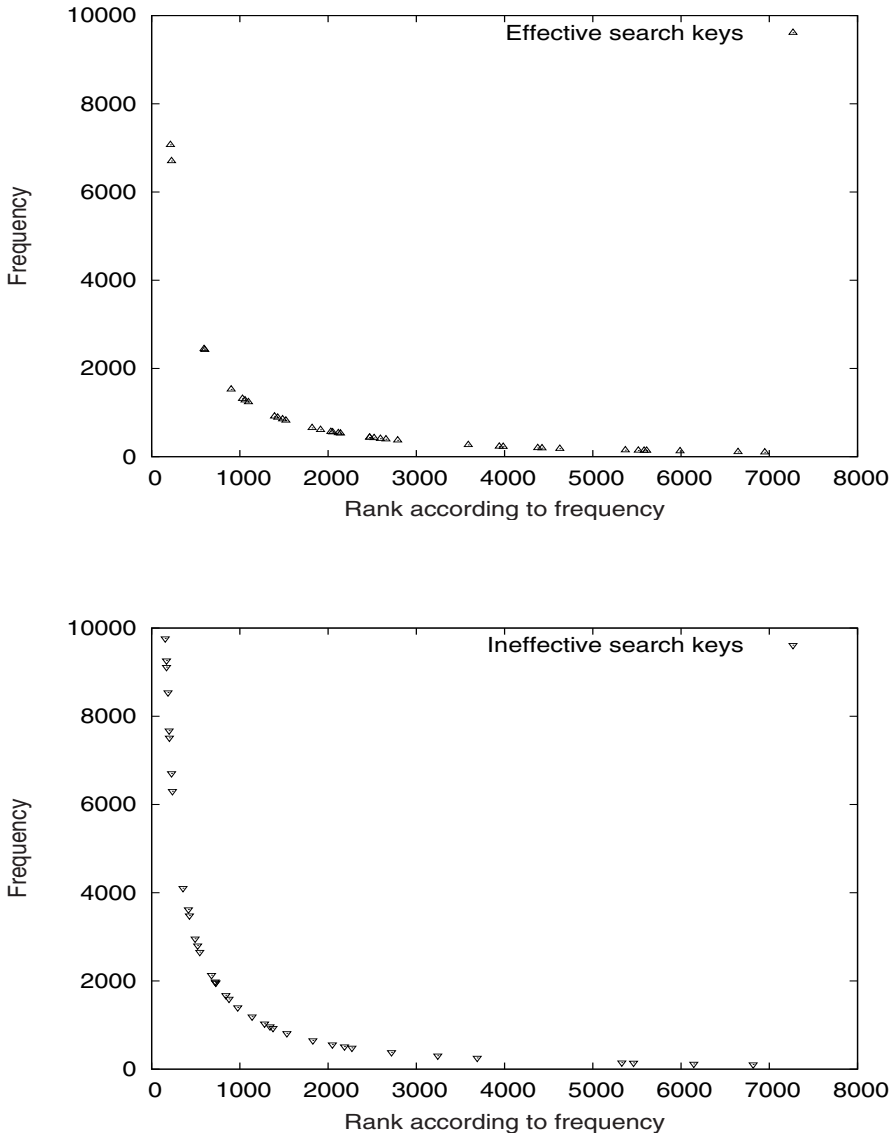


Figure 72: The graphs show representations of the effective and the ineffective search keys extracted from the Zipf curve resulting from all terms in the MedEval corpus. The top graph shows the positions of the terms which were effective as search keys, listed in table 13.2, and the bottom graph the positions of the terms which were ineffective as search keys, listed in table 13.1.

that the terms in question are in all cases one of the two main concepts of the corresponding topic. This illustrates that the resolution power of a search key is not only dependent of the collection frequency of the term, but also very much dependent on the topic for which it is used. In fact the term *läkemedel* 'medicine' is present also in the list of ineffective search keys, but this time for another topic (see table 13.4). In this case the specificity of the search key does not match the specificity of the [drug] concept in the topic which is on a higher level as it contains the name of a drug. The behavior of the search key *läkemedel* in these two examples is an indication that the level of specificity of the search key should match the specificity of the topic for best results. We saw similar results when comparing the runs for topics 4 and 23 using this same search key in section 12.3.

Table 13.3: Effective search keys with high frequency. Here shown together with the topics for which they were effective.

Term	Equivalent	Frequency	Topic and equivalent
läkemedel	drug/ medicine	15 258	19. Hur påverkas läkemedel av samtidigt intag av johannesört? 19. How are medicines affected by simultaneous intake of St John's wort?
smärta	pain	7 064	62. Hur bedömer man att en smärta kräver behandling med opioider? 62. How does one decide that a pain requires treatment with opioids?
eksem	eczema	6 712	92. Hud: Hur går man till väga vid behandling av eksem med steroider? 92. Skin: How does one perform treatment of eczema with steroids?
depression	depression	2 428	32. Indikationer för behandling av depression hos barn med SSRI. 32. Indications for treatment of depression in children with SSRI.

The curves in figure 72 extend far at both ends. In the figure they have been cut off at frequency 10 000 and rank 8 000 for clarity. Out of the 43 effective terms 9 have positions outside the cut off points, 1 with high frequency, *läkemedel* which has been commented above, and 8 with low frequencies. Of the 58 ineffective terms, 19 have positions beyond the cut off points, 3 with high frequencies and 16 with low frequencies. Also the fact that more ineffective terms than effective ones are at the far ends of the Zipf curve is consistent with Luhn's ideas.

Table 13.4: Ineffective search keys with high frequency. Here shown together with the topics for which they were ineffective.

Term	Equivalent	Frequency	Topic and equivalent
behandling	treatment	32 917	2. Vilka komplikationer kan vara gällande vid behandling av infektion hos gravida med erytromycin? 2. Which complications can come in question in treatment of infection in pregnant [women] with erythromycin?
läkemedel	drug/ medicine	15 258	13. På vilka indikationer används läkemedlet concerta vid ADHD-behandling? 13. On which indications is the drug concerta used in ADHD treatment?
använda	use (V)	21 734	26. Vilka kriterier och tekniker används för diagnosticering av akuta koronara syndrom? 26. What criteria and which techniques are used for diagnosing acute coronary syndromes?

Terms with low frequency can be good search keys if they are one of the primary keys. Table 13.5 shows the eight efective terms that had such low frequencies that they were outside the graph in figure 72. They are all names of drugs, diseases, families of drugs or, as in the case of *transitorisk*, part of such a name. They are also all very salient to the topic and on the same level of specificity.

13.4 Test of significance

To test the statistical significance of the clustering of effective terms, in the middle range of terms ranked by collection frequencies, as represented in figure 72, the Mann Whitney U test[37] was used. The Mann Whitney U test takes a list of items and calculates the rank sum of a chosen category. It then checks if the chosen category tends to occur at either end of the list. If the variation is more than by chance the items are assumed to be in a non-random order. The assumption is that, if there was no difference between the categories, for a large population, the number of instances where an item of category A is higher in the list than an item of category B would be equal to the number of instances

[37] Supplied by Jussi Karlgren: <http://www.sics.se/node/800>.

Table 13.5: Effective search keys with low frequency. Here shown together with the topics for which they were effective.

Term	Equivalent	Frequency	Topic and equivalent
reductil	Reductil	60	5. När är det indicerat att börja behandla fetma med preparatet reductil? 5. When is it indicated to start using the drug Reductil to treat obesity? 11. Vilka biverkningar har de två viktminskningspreparaten reductil och xenical? 11. Which side effects do the two weight loss preparations Reductil and Xenical have?
concerta	Concerta	38	13. På vilka indikationer används läkemedlet concerta vid ADHD-behandling? 13. On what evidence is the drug Concerta used in treatment of ADHD?
johannesört	St. John's wort	81	Hur påverkas läkemedel av samtidigt intag av johannesört? 19. How are medicines affected by simultaneous intake of St John's wort?
antipsykosmedel	antipsychotic agents	8	23. Vilka risker är förknippade med användandet av neuroleptika? 23. Which risks are associated with the use of neuroleptics?
transitorisk ischemisk attack	transient ischemic attack	35	25. Vilka symtom och vilken efterföljande behandling får man vid TIA? Which symptoms and what subsequent treatment does one get with TIA?
transitorisk	transient	35	25. Vilka symtom och vilken efterföljande behandling får man vid TIA? Which symptoms and what subsequent treatment does one get with TIA?
appencicit	appencicitis	48	73. Hur ställs diagnosen appendicit vid buksmärta? 73. How is appendicitis diagnosed when there is abdominal pain?

where an item of B is higher than an item of A. To be more precise, the null hypothesis is that, for all instances of pairs with one item of category A and one item of category B, the probability that B is higher in the list than A would be the same as the probability that A would be higher than B (see formula 18). This test is equivalent to the Wilcoxon test (Karlgren 2000; Croft, Metzler and Strohman 2010).

$$P(B > A) = P(A > B) = \frac{1}{2} \tag{18}$$

Before the test, all terms and their frequencies were listed in order of frequency and categorized: 'E' for effective, and 'N' for non-effective. The Mann Whitney U test does not consider how big the difference of frequency is between the items in the list, only the ranking. Rank numbers can be used, but if they are it is important to consider ties. If several items have the same frequency, they should be given the average ranking of all items of the tie. For instance, if the terms in positions 6 and 7 have the same frequency, they should both be given the ranking 6.5.

When the frequency ranked list of effective and ineffective terms from MedEval was first tested, the result was negative. There was no tendency for either category to cluster either at the beginning of the list or at the end of the list. However, this was not surprising. The representations in figure 72 show that the 'E' category is clustered in the middle of the range, and the 'N' category at both ends. In other words, there was not a statistically significant difference in how often one category was ranked higher than the other. The high and the low instances of the 'N' category canceled each other out.

In order to be able to use the Mann Whitney U test to catch the clustering of the 'E' category in the middle range, the list was split in two, one list with the first 52 items, and one with the last 51 items. The two list halves were run separately through the Mann Whitney U test. This time the test showed a significant difference of the clustering of the two categories. In both runs the result was: YES, the result is non-random. For the lower frequency half of the list the result also stated that the effective, 'E', category was in the high end of the list, and for the higher frequencies the 'E' category was in the low end of the list. These results together conclude that the category of effective search terms was clustered in the middle of the list. This is consistent with Luhn (1958) and the representations in figure 72.

13.5 Reflections

The results of the runs with the content words of the topic descriptions support the claims of Luhn, Salton and Pirkola, Järvelin described in chapter 6. The

```
Rank sum: 375.5 Criterion (95%) < 416.939335286214
Result: YES: low.

Rank sum: 860 Criterion (95%) > 740.055634826555
Result: YES: high.
```

Figure 73: The results of running the list of search keys ranked by frequency and marked for effectiveness, divided in two parts, through the Mann Whitney U test. For both halves the result is: YES the result is non-random. For the half with the higher frequencies the 'E' category is clustered in the lower half, and for the half with lower frequencies the effective, 'E', category is clustered in the higher half.

experiments support the conclusion the terms in the middle range of frequencies that have the highest probability of discriminating documents relevant to the topic.

High frequency terms can sometimes be effective as search keys if they are on the same level of specificity as the topic and very salient to the topic. However, one has to be careful with high frequency search keys. With growing size of document collection the problem is not so much retrieving the relevant documents, as not to retrieve noise.

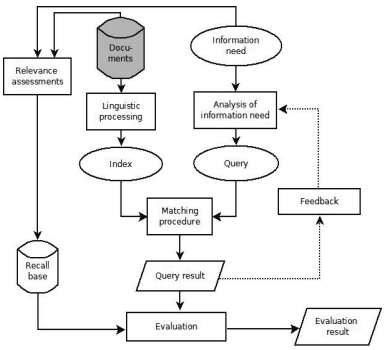
Low frequency terms can be effective if they are specific and very salient to the topic. This seems to be especially true of named entities and components of named entities. The risk of using low frequency terms as search keys is not very big. They can never retrieve more non-relevant documents than the number of documents they appear in.

Low frequency compounds can benefit from decomposition. The individual constituents can be used as search keys if they are relevant to the topic and if they are closer to the middle frequency range of the collection terms, than the original compound.

14 LOOKING AT DOCTOR AND PATIENT DOCUMENTS

To summarize the summary of the summary: people are a problem.

The Restaurant... chapter 28



With the increasing amount of information around us, especially electronic information, people have begun to demand to be informed. It is becoming more and more common, for patients and their close ones, not to rely solely on the physician. They search for information both before consulting a medical professional in order to be able to ask the right questions, and after, to find out more. It is also important for physicians to keep up with new research and to find information about cases similar to the ones they encounter in their work situations. There is so much to know in the medical field that it is impossible for anyone to know it all. The lay person and the professional both need information, maybe about the same topic, but usually not in the same form. The professional is hopefully equipped with consierably more background information.

This chapter brings up differences in the language of medical information intended for lay persons and for professionals and it discusses possible ways of using that information to retrieve documents intended for each group respectively.

14.1 Type/token variations

For a classification of documents according to the intended reader group to be useful, there must be a measurable difference between the document classes.

To examine which differences could be found in the MedEval collection different subsets were extracted, based on the target groups and on the relevance.

To contrast expert and non-expert language all assessed documents, even the ones with relevance grade 0 were used. This was done to achieve as large a collection of documents marked for target group as possible. A number of documents were extracted for the pools of more than one topic, and more than 500 of these were marked with different target groups for different topics. These documents are represented in both the Doctors assessed set and the Patients assessed set. There is also a set which only contains these documents that were marked for both target groups.

Table 14.1 shows a number of type/token frequencies in the different subsets of the collection. In each set duplicates were removed when a document had been assessed for more than one topic. The subsets considered are described below.

Entire collection All documents in the MedEval collection.

Assessed documents All documents that have been assessed for any topic.

Doctors All documents that for at least one topic have been assessed to have target group Doctors.

Patients All documents that for at least one topic have been assessed to have target group Patients.

Common files All documents that for at least one topic have been assessed to have target group Doctors and for another to have target group Patients.

Doctors relevant All documents that for at least one topic have been assessed to have at least relevance grade 1 and to have target group Doctors.

Patients relevant All documents that for at least one topic have been assessed to have at least relevance grade 1 and to have target group Patients.

Before counting frequencies, the files were cleaned from tags, IDs, dates (in the date tag, not in the actual text), web information, punctuation marks and numerical expressions. For this reason the number of tokens stated for the entire collection is smaller in table 14.1 than in table 10.1 on page 109. Full form types are the original word forms in the documents before lemmatization. Lemma types are the same terms after lemmatization.

Some observations are readily made by studying table 14.1. The number of tokens per document is significantly smaller for the entire collection than in any subset. This means that there is a large number of short documents that were not retrieved by any query in the pooling process. Maybe not surprising, since short documents contain few terms which can match the queries.

Table 14.1: Type and token frequencies of the terms in the documents of the whole MedEval collection and of different subsets of the collection.

	Entire collection	Assessed documents	Doctors Assessed	Patients Assessed	Common files	Doctors relevant	Patients relevant
Number of documents	42 250	7 044	3 272	4 334	562	1 233	1 654
Tokens	12 991 157	5 034 323	3 232 772	2 431 160	629 609	1 361 700	988 236
Tokens/document	307	715	988	561	1 120	1 104	596
Average word length	5.75	6.04	6.29	5.73	6.16	6.33	5.63
Full form types	334 559	181 354	154 901	92 803	50 961	87 814	43 825
Lemma types	267 892	146 631	126 217	73 121	40 857	71 974	34 263
Lemma type token ratio	48.5	34.3	25.6	33.2	15.4	18.9	28.8
Compound tokens	1 273 874	573 625	412 475	237 267	76 117	179 580	92 420
Full compound types	187 904	99 614	83 846	47 387	24 083	45 257	20 157
Lemma compound types	144 159	78 508	66 907	37 151	19 685	36 867	16 006
Ratio of compounds	0.098	0.114	0.128	0.098	0.120	0.132	0.094

The finding that unjudged documents on average tend to be shorter than judged documents, both relevant and non-relevant, is consistent with the results of experiments described in Karlgren 2000. One reason, according to Karlgren, is that non-retrieved items often contain tables and numerical information. A longer document also has a bigger chance of touching relevant subjects, and not only relevant subjects, but also confusingly similar subjects which are non-relevant. Karlgren not only found that stylistic measures differed between non-retrieved documents and retrieved documents, but also between relevant documents and non-relevant documents. He suggests that stylistic measures could be used to distinguish *interesting* texts from *less interesting* ones.

The documents in the patients set had only 57% of the doctors' number of tokens per document. Even though there were over 1 000 more patient documents than doctor documents, there were over 50 000 more lemma types in the doctor documents and almost 30 000 more lemma compound types. The average word length in the doctor documents was 6.29 compared to 5.73 for the patient documents. The ratio of compound tokens was also higher in the doctor documents, 0.128 compared to 0.098.

Tables 14.2 and 14.3 illustrate the fact that the doctor documents contain more and longer terms and more compounds than patient documents. The tables show frequencies of all full form types of strings beginning with the randomly chosen term *förmak* 'atrium' in doctor and patient documents respectively. The doctor documents have 75 full form types that begin with *förmak* while patient documents have 18. That is more than four times more types for the doctor documents.

In the MedEval collection the type-token ratio in the patient documents is higher than in the doctor documents, which means that the number of tokens per type is higher. This difference is consistent with the findings of Kokkinakis and Toporowska Gronostaj (2006). They conclude that the more scientific profile of the expert documents entails a larger number of types. In the non-expert documents they find fewer word forms, but these word forms are repeated more often.

In table 14.1 we can see that there is a clear difference in the type-token ratio in the different subsets of the MedEval collection. However, one has to be careful when comparing the type-token figures, as they are very much dependent of the size of the collection described. As a collection grows, most of the new tokens are of types already present in the collection. The number of new types decreases with the size of the growing collection. Bearing this in mind it is even more noteworthy that the type-token ratio in the Patients assessed set is 33.2 compared to 25.6 in Doctors assessed, even though there are 800 000 more tokens in the Doctors set.

Table 14.2: The types and frequencies of all terms in the expert documents containing the string förmak∗ ‘atrium’. Compare to the much lower number of types in the non-expert documents in table 14.3.

Type	Token	Type	Token
förmak	93	förmaksmuskelns	1
förmaken	21	förmaksmuskulaturen	2
förmakens	1	förmaksmyocyterna	2
förmaket	11	förmaksmyokard	3
förmakets	1	förmaksmyokardiet	1
förmaks	21	förmaksmyxom	2
förmaksaktivering	1	förmaksnivå	2
förmaksaktivitet	1	förmaksnära	1
förmaksaktiviteten	2	förmaksoch	1
förmaksanatomi	1	förmakspacing	7
förmaksarytmi	2	förmakspeptider	1
förmaksarytmier	9	förmaksrytmer	1
förmaksbidraget	1	förmaksseptostomi	1
förmaksbradyarytmi	1	förmaksseptum	2
förmaksdefibrillator	2	förmaksseptumaneurysm	10
förmakseffekt	2	förmaksseptumdefekt	5
förmaksfladder	57	förmaksseptumdefekten	1
förmaksfladdret	2	förmaksseptumdefekter	1
förmaksflimmer	544	förmaksseptums	1
förmaksflimmerablationer	2	förmaksstimulerat	1
förmaksflimmerattacker	1	förmaksstimulering	5
förmaksflimmerduration	2	förmaksstorlek	2
förmaksflimmerepisoder	4	förmaksstorleken	1
förmaksflimmerfladder	2	förmakssynkron	1
förmaksflimmerpatienter	4	förmakssystole	1
förmaksflimmerrecidiv	1	förmakstaket	1
förmaksflimmertendensen	1	förmakstakykardi	11
förmaksflimmerunderhållande	1	förmakstakykardie	8
förmaksflimret	16	förmakstromb	2
förmaksflimrets	4	förmakstryck	1
förmaksfrekvenser	1	förmakstrycket	1
förmaksfunktion	1	förmaksvolym	2
förmaksförstoring	1	förmaksvägg	1
förmaksimpuls	1	förmaksväggarna	2
förmaksinhiberad	1	förmaksväggen	6
förmakskontraktion	4	förmaksvävnaden	2
förmakskontraktionen	6	förmaksöra	9
förmakskontraktionens	1	förmaksöronen	2
förmaksmuskeln	1		

Table 14.3: The types and frequencies of all terms in the non-expert documents containing the string förmak∗ 'atrium'. Compare to the higher number of types in the expert documents in table 14.2.

Type	Token	Type	Token
förmak	73	förmaksflimmer	219
förmaken	21	förmaksflimmerattacker	1
formakens	2	förmaksflimmerpatienter	1
förmaket	14	förmaksflimret	28
förmaks	1	förmakslimmer	1
förmaksarytmier	2	förmaksmyocyterna	1
förmakseffekt	1	förmakstakykardi	1
förmaksfladder	2	förmaksutlösta	2
förmaksflimer	1	förmaksöra	1

14.2 Synonyms

For many medical concepts there are terms of different registers. Professionals often use neoclassical terms originating from Latin or Greek while the ordinary person often uses simpler Swedish terms.

An example of how retrieval results can differ with search terms of different registers is shown in figure 74. Two synonyms, one used in professional language and one in lay person language, *anemi* 'anemia' and *blodbrist* 'blood deficiency' (anemia), were run through the non-decomposed index. For all user groups *anemi* showed higher effectiveness than *blodbrist*. Most interesting is that for the doctor user group *anemi* had retrieved all relevant documents at DCV 200[38] while the term *blodbrist* had not retrieved any. For patients the neoclassical term did not perform quite as well as it did for doctors, and the Swedish term did not perform as bad as it did for doctors.

It is for the doctor documents that the difference in expert and non-expert terms occurring is most apparent. One plausible reason is that professionals usually do not use lay terms. These are often imprecise and can even be misleading, for instance *blodbrist* 'blood deficiency' (anemia), does not refer to a deficiency of blood, but rather a deficiency of red blood cells or of hemoglobin. In contrast, texts written for lay persons often contain both lay terms and professional terms. Sometimes a professional term is used, and the lay term is added as an explanation, or a lay term is used and the professional term is presented as additional information. Examples of both situations are shown in figure 75. It is interesting that the patient documents often contain medical

[38]Default document cut off value for the QPA.

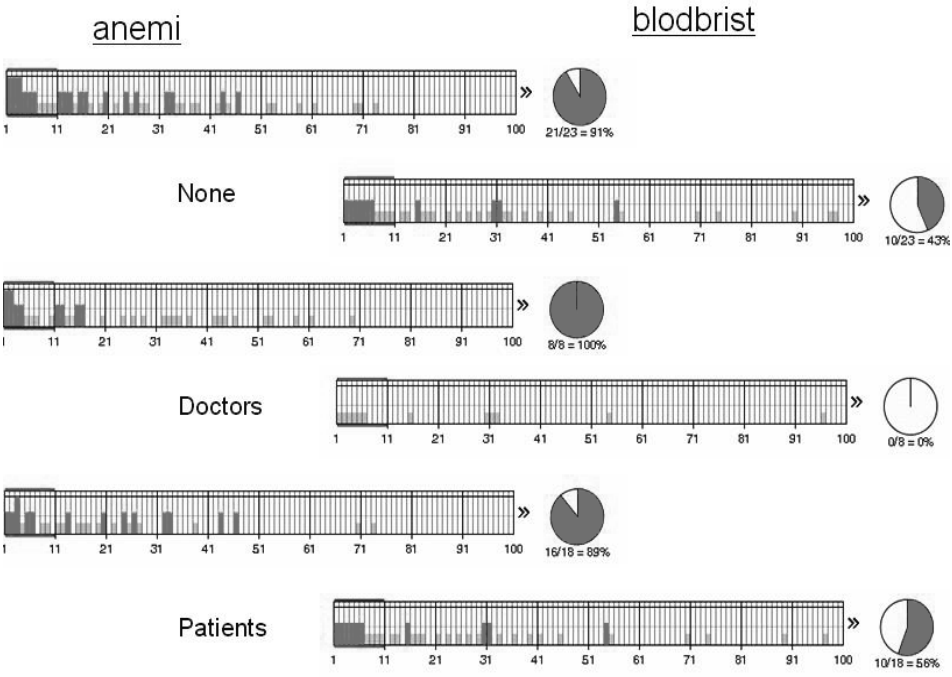


Figure 74: The near synonyms *anemi* 'anemia' and *blodbrist* 'blood deficiency' (anemia) were run through all three user scenarios in the non-decomposed index in QPA. All user groups had better results with the neoclassical term *anemi* than with the lay term *blodbrist*. Most striking was that the Doctors user group had full recall for the neoclassical term and no recall at all for the lay term at default DCV 200.

terms from both registers bearing in mind that the doctor documents contain significantly more types.

14.3 Multiword units

To study the difference between the doctor and the patient documents, on phrase level, the documents assessed to have doctors and patients target groups were run separately through MWT (multi-word term), a perl program which counts the frequencies of occurring multiword units of different lengths. The program uses both lexical and statistical calculation. It is based on ideas presented by John Justeson and Slava Katz in an IBM Research report in 1993, later published as Justeson and Katz 1995, and in code by Katz at New York University in 1995. It was later again developed further by Jussi Karlgren,

B12 är ett vitamin som är nödvändigt för bildningen av röda blodkroppar, brist kan då ge det vi kallar perniciös anemi (anemi betyder just blodbrist).
B12 is a vitamin that is necessary for the production of red blood cells, deficiency can cause what we call pernicious anemia (anemia means precisely blood deficiency).

...t.ex. fel på sköldkörteln, diabetes eller en speciell form av blodbrist , s.k. perniciös anemi.
...e.g. failure of the thyroid gland, diabetes or a special form of anemia, known as pernicious anemia.

Figure 75: Two examples of synonyms of different styles being used in one sentence. In the first example the lay term is used to explain the professional term, and in the second the professional term is supplied as additional information to the lay term.

originally for teaching purposes.[39] The MWT program does not consider any linguistic factors, it only counts the frequencies of tokens occurring in different combinations. Not all of these combinations are grammatical units.

In the runs presented here the units had a length of between 3 and 6 tokens separated by whitespace. Not only the longest found unit in a certain position was counted, but also the shorter ones which the longer unit contained. This is evident in the beginning of table14.6 which shows the multiword units in the patient documents. These documents contain a substantial number of documents of the kind: 'Ask the doctor', which are written using a standard format. In all these documents the service and the answering doctors are introduced with identical phrases. The doctors themselves also reuse greetings. The consequence of this is that a substantial part of the most frequent multiword units are these recurring phrases and their substrings. The number of entries in table 14.6, which contains the patient multiword units, is chosen so that the number of entries following the 'Ask the doctor' units will correspond to the number of entries in table 14.5 with the doctor multiword units.

The doctor and patient documents show clear differences in both the frequencies of phrases and the types of phrases extracted. Obvious in the table representing patient documents are the phrases originating from the 'Ask the doctor' documents. This kind of document is typical of patient documents, but the exact phrases will be very dependent on the source.

If one disregards the 'Ask the doctor' phrases, the most frequent multiword units in the patient documents contain mostly phrases belonging to general language, not typical for medical documents. The most frequent unit which

[39] Jussi Karlgren p.c.

could be said to be medical, *att drabbas av* 'to be afflicted by', occurs only 95 times in the patient documents.

Table 14.4: The most frequent patient multiword units, disregarding the fixed phrases from 'ask the doctor' documents.

Frequency	Swedish multiword unit	English equivalent
199	att det är	that it is
120	med hjälp av	using
104	på grund av	because
102	för att få	to get
102	när det gäller	when it comes to
101	det är inte	it is not
95	att drabbas av	to be afflicted by
95	det är en	it is a
94	om du har	if you have
88	men det är	but it is

Just as the phrases from 'Ask the doctor' stand out among the multiword units from patient documents there are phrases that stand out among the multiword units from the doctor documents: phrases concerning medication. The majority of the most frequent multiword units denote the form or dosage of medicine. Throughout the table there are several names of drug companies. In table 14.5, from the doctor documents, there is a large number of general language phrases, but it is the medical phrases that dominate.

A striking difference between the doctor and patient multiword units is that the frequencies are much higher in the doctor documents. The most frequent multiword unit for the doctor documents occurred 5 648 times, while the most frequent multiword unit in the patient documents only occurred 316 times. At the end of the tables the doctors have frequency 175 at rank 103, while the patients have frequency 53 at the same rank.

The fact that the doctor documents were found to have a higher number of multiword units than the patient documents is consistent with the results presented in Rosell and Velupillai 2005. The authors present a hypothesis that phrases could be a better representation unit for medical texts than words. The results were not encouraging, but they did show that phrase representation had better results applied on a medical corpus than phrase representation applied on a corpus of Swedish newspaper articles. Bearing in mind that the MedLex corpus, which MedEval is based on, contains a large number of newspaper articles written for a lay audience, and that the medical texts used in Rosell and Velupillai 2005 were collected from the medical journal Läkartidningen[40] which is one of the sources of MedLex, a parallel can be drawn.

[40] <http//www.lakartidningen.se>

Table 14.5: Multiword units in doctor documents. The first two columns show the rank and the frequencies of the most frequent units. The following columns show the units and their English equivalents.

Rank	Freq	Swedish multiword unit	English equivalent
1	5648	mg - tillståndsinnehavare	mg - permit holder(s)
2	2044	tablett - styrka	tablet - strength
3	1802	filmdragerad tablett -	film-coated tablet -
4	1802	filmdragerad tablett - styrka	film-coated tablet - strength
5	1661	lösning - styrka	solution - strength
6	1497	på grund av	due to 'on ground of'
7	1080	mg ml -	mg ml -
8	1069	mg ml - tillståndsinnehavare	mg ml permit holder(s)
9	987	– - tillståndsinnehavare	– - permit holder(s)
10	970	orifarm ab produkt	orifarm ltd product
11	844	hos patienter med	in patients with
12	752	hård - styrka	hard - strength
13	641	med hjälp av	using 'with help of'
14	575	när det gäller	when it comes to
15	500	ökad risk för	increased risk of
16	481	www.lakartidningen.se nr -	www.lakartidningen.se no -
17	461	paranova läkemedel ab	paranova drug ltd
18	459	av patienter med	of patients with
19	453	paranova läkemedel ab produkt	paranova drug ltd product
20	446	pulver och vätska	powder and solvent
21	444	pulver och vätska till	powder and solvent for
22	440	för behandling av	for treatment of
23	439	att det är	that it is
24	419	www.lakartidningen.se nr - volym	www.lakartidningen.se no - volume
25	411	pm senast uppdaterad	pm last updated
26	410	kan leda till	can lead to
27	401	har visat att	has/have shown that
28	390	pulver och vätska till injektionsvätska	powder and solvent for injection solution
29	386	astrazeneca ab produkt	astrazeneca ltd product
30	371	det vill säga	i.e. 'that will say'
31	367	suspension - styrka	suspension - strength

Continued on next page...

Table 14.5- continued from previous page.

Rank	Freq	Swedish multiword unit	English equivalent
32	367	depottablett - styrka	prolonged release tablet - strength
33	357	pfizer ab produkt	pfizer ltd product
34	352	vid behandling av	in treatment of
35	348	novartis sverige ab	novartis sweden ltd
36	346	för att få	[for] to get
37	346	för patienter med	for patients with
38	344	de patienter som	those patients that
39	339	att det finns	that there is
40	337	till följd av	as a result of
41	322	det är viktigt	it is important
42	314	för att kunna	[for] to be able
43	314	talar för att	suggest(s) that 'speaks for that'
44	313	en av de	one of the
45	312	är viktigt att	is important that
46	308	procent av patienterna	percent of the patients
47	299	det är viktigt att	it is important that
48	288	att det inte	that it not
49	275	under de senaste	during the last
50	270	till och med	even/up to 'to and with'
51	261	medartuum ab produkt	medartuum ltd product
52	257	potentiella bindningar eller	potential bonds or
53	257	potentiella bindningar eller jävsförhållanden	potential bonds or conflicts of interest
54	257	ökar risken för	increase(s) the risk of
55	256	merck sharp dohme	merck sharp dohme
56	255	har visat sig	has/have proved 'has/have shown itself'
57	255	att drabbas av	to be afflicted by
58	250	en del av	a part of
59	249	för att minska	[for] to reduce
60	246	de senaste åren	the last years
61	244	glaxosmithkline ab produkt	glaxosmithkline ltd product
62	237	procent av alla	percent of all
63	234	är det viktigt	is it important
64	232	det finns en	there is a
65	227	det är en	it is a
66	224	det viktigt att	it important that

Continued on next page...

Table 14.5- continued from previous page.

Rank	Freq	Swedish multiword unit	English equivalent
67	217	minska risken för	reduce the risk of
68	217	är det viktigt att	is it important that
69	215	det är inte	it is not
70	212	hos patienter som	in patients that
71	210	och det är	and it is
72	207	stöd för att	support for [to]
73	203	men det är	but it is
74	202	merck nm ab	merck nm ltd
75	201	novartis sverige ab produkt	novartis sweden ltd product
76	198	studier har visat	studies have shown
77	193	olika typer av	different types of
78	191	det finns ett	there is a
79	190	ge upphov till	cause ‘give source to’
80	190	procent av de	percent of the
81	189	är en av	is one of
82	189	under senare år	during later years
83	188	efter det att	after which that
84	187	en kombination av	a combination of
85	186	stada arzneimittel ag	stada arzneimittel ag
86	186	innehavare av godkännande	holder of authorization
87	186	innehavare av godkännande för försäljning	holder of authorization for marketing
89	186	innehavare av godkännande för	holder of authorization for
90	186	en ökad risk	an increased risk
91	185	män och kvinnor	men and women
92	184	merck nm ab produkt	merck nm ltd product
93	183	aventis pharma ab	aventis pharma ltd
94	182	är den vanligaste	is the most common
95	181	cross pharma ab	cross pharma ltd
96	180	cross pharma ab produkt	cross pharma ltd product
97	179	tyder på att	suggest(s) [on] that
98	178	ett stort antal	a great number
99	176	hexal a s	hexal a s
100	175	inhalationspulver - styrka	inhalation powder - strength
101	175	ger upphov till	cause(s) ‘give(s) source to’
102	175	av dem som	of those that
103	175	kan vara en	can be a

Table 14.6: Multiword units in patient documents. The first two columns show the rank and the frequencies of the most frequent units. The following units show the units and their English equivalents.

Rank	Freq	Swedish multiword unit	English equivalent
1	316	läs mer om	read more about
2	312	mer om tjänsten fråga doktorn	more about the service ask the doctor
3	312	mer om tjänsten	more about the service
4	312	läs mer om tjänsten fråga	read more about the service ask
5	312	tjänsten fråga doktorn	the service ask the doctor
6	312	mer om tjänsten fråga	more about the service ask
7	312	om tjänsten fråga	about the service ask
8	312	läs mer om tjänsten fråga doktorn	read more about the service ask the doctor
9	312	läs mer om tjänsten	read more about the service
10	312	om tjänsten fråga doktorn	about the service ask the doctor
11	218	ställa din fråga till	put your question to
12	218	välkommen att ställa din fråga till	welcome to put your question to
13	218	att ställa din fråga	to put your question
14	218	att ställa din fråga till	to put your question to
15	218	ställa din fråga till mig	put your question to me
16	218	att ställa din fråga till mig	to put your question to me
17	218	välkommen att ställa din fråga	welcome to put your question
18	218	att ställa din	to put your
19	218	välkommen att ställa din	welcome to put your
20	218	ställa din fråga	put your question
21	218	välkommen att ställa	welcome to put
22	199	att det är	that it is
23	121	med hjälp av	using ‘with help of’
24	120	är överläkare på	is senior doctor in
25	116	söderberg är överläkare på onkologiska	söderberg is senior doctor in the oncological
26	116	martin söderberg doktor martin söderberg är	martin söderberg doctor martin söderberg is
27	116	hälsningar martin söderberg doktor	regards martin söderberg doctor

Continued on next page...

Table 14.6- continued from previous page.

Rank	Freq	Swedish multiword unit	English equivalent
28	116	martin söderberg är	martin söderberg is
29	116	martin söderberg doktor	martin söderberg doctor
30	116	söderberg doktor martin söderberg är	söderberg doctor martin söderberg is
31	116	martin söderberg är överläkare på onkologiska	martin söderberg is senior doctor in the oncological
32	116	martin söderberg doktor martin söderberg	martin söderberg doctor martin söderberg
33	116	doktor martin söderberg är	doctor martin söderberg is
34	116	söderberg är överläkare på	söderberg is senior doctor in
35	116	söderberg doktor martin söderberg är överläkare	söderberg doctor martin söderberg is senior doctor
36	116	söderberg är överläkare på onkologiska kliniken	söderberg is senior doctor at the oncological clinic
37	116	martin söderberg är överläkare på	martin söderberg is senior doctor at
38	116	söderberg doktor martin söderberg	söderberg doctor martin söderberg
39	116	hälsningar martin söderberg	regards martin söderberg
40	116	är överläkare på onkologiska	is senior doctor at the oncological
41	116	doktor martin söderberg är överläkare på	doctor martin söderberg is senior doctor at
42	116	hälsningar martin söderberg doktor martin söderberg	regards martin söderberg doctor martin söderberg
43	116	hälsningar martin söderberg doktor martin	regards martin söderberg doctor martin
44	116	söderberg är överläkare	söderberg is senior doctor
45	116	är överläkare på onkologiska kliniken	is senior doctor at the oncological clinic
46	116	martin söderberg doktor martin	martin söderberg doctor martin
47	116	martin söderberg är överläkare	martin söderberg is senior doctor
48	116	doktor martin söderberg är överläkare	doctor martin söderberg is senior doctor
49	116	söderberg doktor martin	söderberg doctor martin
50	116	doktor martin söderberg	doctor martin söderberg
51	104	på grund av	because ‘on ground of’

Continued on next page. . .

Table 14.6- continued from previous page.

Rank	Freq	Swedish multiword unit	English equivalent
52	103	vid medicinkliniken på	at the medical clinic at
53	102	hälsningar anders dahlqvist anders	regards anders dahlqvist anders
54	102	hälsningar anders dahlqvist anders dahlqvist	regards anders dahlqvist anders dahlqvist
55	102	för att få	[for] to get
56	102	arbetar vid medicinkliniken på länsjukhuset gävle-sandviken	works at the medical clinic at the county hospital gävle-sandviken
57	102	arbetar vid medicinkliniken på länsjukhuset	works at the medical clinic at the county hospital
58	102	vid medicinkliniken på länsjukhuset gävle-sandviken	at the medical clinic at the county hospital gävle-sandviken
59	102	arbetar vid medicinkliniken	works at the medical clinic
60	102	vid medicinkliniken på länsjukhuset	at the medical clinic at the county hospital
61	102	när det gäller	when it comes to
62	102	hälsningar anders dahlqvist	regards anders dahlqvist
63	102	arbetar vid medicinkliniken på	works at the medical clinic at
64	101	det är inte	it is not
65	95	överläkare på infektionskliniken	senior doctor at the infection clinic
66	95	att drabbas av	to be afflicted by
67	95	det är en	it is a
68	94	om du har	if you have
69	88	men det är	but it is
70	87	är det viktigt	is it important
71	85	det kan vara	it can be
72	75	ökar risken för	increase(s) the risk of
73	74	att det inte	that it not
74	71	att det finns	that there is
75	71	en del av	a part of
76	70	när behandlingen är	when the treatment is
77	70	därför är det	therefore it is
78	69	behandlingen är avslutad	the treatment is completed
79	69	när behandlingen är avslutad	when the treatment is completed
80	68	om man har	if one has

Continued on next page...

Table 14.6- continued from previous page.

Rank	Freq	Swedish multiword unit	English equivalent
81	67	om det finns	if there is
82	65	är en av	is one of
83	64	ökad risk för	increased risk of
84	61	att man har	that one has
85	61	om det är	if there is
86	60	är det viktigt att	is it important to
87	60	det viktigt att	it important to
88	59	kan leda till	can lead to
89	59	kan det vara	can it be
90	58	på så sätt	in that way
91	57	beror på att	depend(s) on [that]
92	57	det finns en	there is a
93	55	om hjärta kärl – http	on heart vessels - http
94	55	allmänt om hjärta kärl	general on heart vessels
95	55	allmänt om hjärta kärl – http	general on heart vessels – http
96	55	om hjärta kärl	on heart vessels
97	55	allmänt om hjärta	general on heart
98	55	om hjärta kärl –	on heart vessels –
99	55	allmänt om hjärta kärl –	general on heart vessels –
100	54	och det är	and it is
101	54	en av de	one of the
102	53	det bra att	it good that
103	53	att man kan	that one can
104	53	för de flesta	for [the] most
105	52	det vill säga	i.e. ‘that will say’
106	51	är svårt att	is difficult to
107	50	för att se	[for] to see
108	50	är det bra att	is it good that
109	50	men det finns	but there is
110	50	är det bra	is it good
111	49	för behandling av	for treatment of
112	48	det är ett	it is a
113	48	det är mycket	it is very
114	48	är det också	is it also
115	48	det är svårt	it is difficult
116	47	att du har	that you have
117	47	om man är	if one is

Continued on next page...

Table 14.6- continued from previous page.

Rank	Freq	Swedish multiword unit	English equivalent
118	47	det är svårt att	it is difficult to
119	46	risk att drabbas	risk to be afflicted
120	45	som har en	that has/have a
121	45	man har en	one has a
122	45	vara svårt att	be difficult to
123	45	att man inte	that one not
124	44	så är det	so is it
125	44	besked om att	information about that
126	44	att träffa andra	to meet others
127	43	delta om du	participate if you
128	43	att delta om	to participate if
129	43	att delta om du	to participate if you
130	43	så att det	so that it
131	42	tycker att det	think(s) that it
132	42	risken för att	the risk for that
133	42	att de inte	that they not
134	42	att det blir	that it becomes
135	41	minskar risken för	reduce(s) the risk of
136	41	så att de	so that they
137	41	som finns i	that is/are found in
138	40	när den ska	when it shall
139	40	risk att drabbas av	risk to be afflicted by
140	40	och om hur	and about how
141	39	även om man	even if one
142	39	det är viktigt	it is important
143	39	som gör att	that allows ‘which does that’
144	39	med din läkare	with your doctor
145	39	högt blodtryck och	high blood pressure and
146	38	till och med	even/up to ‘to and with’
147	38	på att det	that there ‘on that it’
148	38	är att det	is that it
149	38	för att ta	[for] to take
150	38	tyder på att	suggest(s) [on] that
151	38	för att kunna	[for] to be able
152	38	och på så	and [on] so
153	38	jag är en	i am a
154	38	det visar en	it shows a
155	37	kan vara svårt	can be difficult

Continued on next page...

Table 14.6- continued from previous page.

Rank	Freq	Swedish multiword unit	English equivalent
156	37	är viktigt att	is important that
157	37	det inte finns	it not is/are found
158	37	kan vara svårt att	can be difficult to
159	37	spridit sig till	spread [itself] to
160	36	det kan vara svårt att	it can be difficult to
161	36	är den vanligaste	is the most common
162	36	att jag har	that i have
163	36	en frisk cell	a healthy cell
164	36	det kan vara svårt	it can be difficult
165	36	för att det	so that it
166	36	en ökad risk	an increased risk
167	36	det är viktigt att	it is important to

14.3.1 Trigger phrases

Some phrases are typically used to signal a particular type of information. These trigger phrases may not contain much information in themselves, but indicate the existence of certain information. Examples of trigger phrases are: *The survey shows* and *In table*.

Many of the phrases in the tables of frequent multiword units in doctor and patient documents are not specific for any topic. However, some of them may be seen as trigger phrases indicating the target group of the documents. Trigger phrases could be used to give different priors to the search engines depending on the chosen user scenario. Potential trigger phrases for doctors and patients can be seen in tables 14.7 and 14.9.

It is not enough to pick out the most frequent multiword units in the documents of a certain target group in order to decide what could be used as trigger phrases. Among the most frequent phrases in the patient documents were phrases that were even more frequent in the doctor phrases, although they occurred further down in the frequency list. Two of the most frequent patient phrases were: *när det gäller* 'when it comes to' and *att drabbas av* 'to be afflicted by'. These occurred 102 and 95 times respectively in the patient documents but 575 and 255 times respectively in the doctor documents. This means that a larger number of doctor documents than patient documents would be retrieved if these multiword units were used as search terms (see table 14.8). This leaves few alternatives for effective patient trigger phrases. A possibility is instead if priors in the patients user scenario should be based on the absence

Table 14.7: Potential doctor trigger phrases.

Frequency	Swedish multiword unit	English equivalent
844	hos patienter med	in patients with
500	ökad risk för	increased risk of
459	av patienter med	of patients with
440	för behandling av	for treatment of
410	kan leda till	can lead to
401	har visat att	has shown that
352	vid behandling av	in treatment of
346	för patienter med	for patients with
344	de patienter som	those patients that
337	till följd av	as result of
314	talar för att	suggest(s) that
308	procent av patienterna	percent of the patients

Table 14.8: Phrases common for patients, but even more common for doctors.

Patient frequency	Doctor frequency	Swedish multiword unit	English equivalent
102	575	när det gäller	when it comes to
95	255	att drabbas av	to be afflicted by

Table 14.9: Potential patient trigger phrases.

Frequency	Swedish multiword unit	English equivalent
316	läs mer om	read more about
218	ställa din fråga	put your question
94	om du har	if you have
70	när behandlingen är	when the treatment is

of typical doctor phrases and give higher probability to documents that do not contain doctor trigger phrases.

14.4 Stylistic differences

The search for multiword units in the doctor and patient documents respectively, not only suggested trigger phrases for the two target groups, but also made stylistic differences apparent. There are stylistic differences, not only on phrase level, but also on term level. A closer look at the frequencies of *förmak∗* in the professional and lay person texts (presented in tables 14.2 and 14.3) reveals that not all frequencies are higher for professionals. The frequencies of

nouns in the definite form in the lay person texts are close to, equal to or higher compared to the same forms in the professional texts. The only term that for the definite forms had higher frequencies in the doctor documents than in the patient documents, is *förmaksmyocyt* 'atrium myocyte'. This, on the other hand is a specialized compound term, not well known by lay persons.

Table 14.10: Frequencies of terms beginning with *förmak* 'atrium' which are in the definite form in the patient documents. A comparison with frequencies in the doctor documents.

Word form	Frequency in doctor documents	Frequency in patient documents
förmaken	21	21
förmakens	1	2
förmaket	11	14
förmaksflimret	16	28
förmaksmyocyterna	2	1

Looking at all instances of strings beginning with *förmak* in the two sets of documents 66 tokens of 372, or 17.7% are nouns in the definite form, while the corresponding numbers for the doctor documents, are 89 of 932 tokens, or 9.6%.

Not only Swedish nouns, but also adjectives are inflected for definiteness and number. When comparing the word forms of adjectives in the doctor and patient documents, it is evident that the non-neuter/singular form is more common in the patient documents. The definite and the plural adjective inflections are identical to each other, but differ from the non-neuter singular and the neuter singular form for most adjectives, as can be seen in table 4.1 on page 44. Table 14.11 shows how the frequencies differ for a few adjectives.

If these tendencies hold for nouns and adjectives in general, morphological information could be used to give more weight to nouns and adjectives in the definite form when searching for lay person documents and less when searching for professional documents.

A difference that can be seen in the frequency tables 14.5 and 14.6 is that professionals seem to write in a generic sense. This gives high frequencies for phrases with meanings such as: *in patients with* or *of patients with*. Frequencies are also high for indefinite noun phrases such as: *in treatment of* or *for treatment of*. The patient documents often discuss specific patients or specific cases. This gives high frequencies for phrases that contain the pronoun *you* and noun phrases in definite form: *when the treatment is completed*.

Table 14.11: Frequencies of adjectives.

Term	Equivalent	Doctor documents		Patient documents	
		Non-neuter singular indefinite	Plural and/or definite	Non-neuter singular indefinite	Plural and/or definite
sjuk	sick	165	462	333	371
smittad	infected	115	501	332	320
fet	fat	67	137	219	193
tjock	thick/fat	59	15	152	28
smal	thin	22	21	41	25
gravid	pregnant	78	471	651	402
allergisk	allergic	364	210	432	282
överkänslig	hypersensitive	15	10	72	15
deprimerad	depressed	20	89	79	42

Overall the documents written for the doctor target group tend to be written in a more disassociated way compared with the patient documents which are more interactive in their approach, addressing the reader directly. While the professional documents tend to discuss research results or cases in general, the lay person documents often discuss specific cases. This difference in approach manifests itself, for example in the features described above with the patient documents containing more nouns in the definite form, and more pronouns in the first or second person, while doctor documents predominately have nouns in the indefinite form, and pronouns in the third person. The professional documents also tend to be written in a more formal way with many multiword phrases reoccurring with high frequencies.

As there is an apparent difference between the documents written for the professional and lay person target groups, these differences could be used for a precategorization of documents according to genre. Such a categorization could be stored in a separate field in the document representations.

14.5 Reflections

This thesis gives several examples of how lay person language and professional language differ, and how the differences can be used to find search keys that would be effective in order to find documents written for one of the target groups. It is easier to find features specific for the expert documents, than to find features specific for the non-expert documents. To find the expert docu-

ments it tends to be an effective strategy to use many synonyms, to use compounds and longer words, to use the names of specific drugs, and to use trigger phrases typical of the professional written language. On the other hand, there does not seem to be as many specific strategies to find non-expert documents. The only two found were to use typical lay person terms, terms of the kind that professionals would not use, and use these terms in combination with basic expert terms and to search for documents employing the definite form of nouns and adjectives.

Lemmatization in the indexing and search processes can improve recall as there will be more matches in the matching process. But as recall improves, precision deteriorates. The reason that words are inflected in different forms in the first place is that the inflectional endings carry information. If this information is removed, the possibility to use this information is also removed. As we could see earlier in this chapter medical professionals and lay persons use different ways of expressing themselves, alternatively, speak of different matters. Doctors have an interest in describing generic cases, while patients, or their close ones, are more interested in specific cases. The difference in expression between generic and specific often lies in inflectional endings.

In order to be able to use this information, it must be kept in some way. An interesting research question for future projects could be to study the benefit of lemmatizing inflected words, but keeping the inflectional information in tags, or recording the tendency of a text in terms of generic vs. specific. This could be a way to keep the higher recall gained by lemmatization, but still use inflectional information for discrimination.

Part V

Conclusions

15 THE END OF THE ROAD

I am therefore excused from saving Universes.

Life... chapter 6

Now that we have come to the end of the road, there is one more question: Did we reach our goal, the goal of finding answers to the four research questions we started with? The questions are given here once again, together with the answers we found along the way.

- What features do terms that are good search keys have? What features do terms that are bad search keys have? Can this knowledge be used to select compound constituents to use as search keys?

  - A good search key is a term that expresses one of the main subjects of the topic in question, and on the same level of specificity. It is a term with a frequency in the middle range of all term frequencies. It is clustered in few documents. A specific term with low frequency can be good if it expresses the topic well. At least it is not bad.
  - A bad search key is a term that brings noise to the top of the ranked list of retrieved documents. This is common with high frequency terms, terms with broad, non-specific meaning and terms spread evenly over the documents. A bad search key is a term that is not about the contents of the topic, or on another level of specificity.
  - A search key that is very infrequent is neither good nor bad. It lacks goodness in the sense that it does not add relevant documents to the answer set in any significant way. It lacks badness in the sense that it does not add non-relevant documents to the answer set.
  - Compound constituents which have collection frequencies that are further away than the whole compound from the middle range of the collection frequencies of all terms do not make the queries more effective and should not be used. Further, constituents that

have nothing to do with the meaning of the topic should not be used, for example constituents of non-compositional compounds.

- Can specific features of professional language and of lay person language, respectively, be utilized when searching for medical documents for the two target groups?
  - This thesis reports on several measurable features that differ between the two registers. It should therefore be possible. However, it seems to be easier to point out specific features for professional language than for lay person language.
- Can the questions above be answered using a medical test collection with two indexes containing different representations of compounds, split and unsplit, and providing user group scenarios, professionals and lay persons? What other research questions can be answered with such a collection?
  - Even the few examples of the experiments with compounds presented in this thesis have made it clear that compounds should not all be treated the same way, and compound constituents should not be treated the same way. Some compounds, especially occasional and compositional, reveal their information first after they are split. For other compounds splitting means the information is destroyed.
  - The assessment of target groups and choice of user groups in the MedEval collection makes it possible to get a quick visualization of the difference in success of retrieval for the expert and non-expert groups.
  - Seeing that the indexes of MedEval, in the present form, contain the lemmatized forms of the collection terms, it is not possible to study the effects of using different word forms to retrieve documents for the two target groups. However, as the original documents are available, it would be possible to create full forms indexes, and use them to study the impact of different word forms on the two target groups.
  - As MedEval contains a collection of documents assessed for target groups, doctors and patients, this collection can be used to study differences in register: experts vs. non-experts. This can be done independently of the relevance assessments.
  - As information on how the documents written for the two target groups can be extracted from the assessed documents of MedEval,

this information could be used for a preclassification of documents not yet assessed.

The experiments described in this thesis should be seen as pilot studies. The author would like to see the results most of all as suggestions for further studies.

REFERENCES

Ahlgren, Per 2004. The effects of indexing strategy-query term combination on retrieval effectiveness in a Swedish full text database. Ph.D. diss., University College of Borås/Göteborg University. Publications from Valfrid, nr 28.

Allén, Sture 1970. *Frequency dictionary of present-day Swedish: Graphic words, homograph components*. Volume 1. Stockholm: Almqvist & Wiksell.

Baeza-Yates, Ricardo and Berthier Ribeiro-Neto 1999. *Modern information retrieval*. ACM-press, New York, New York.

Bauer, Laurie 1983. *English word-formation*. Cambridge: Cambridge University Press.

Berbyuk Lindström, Nataliya 2008. Intercultural communicaiton in health care – Non-Swedish physicians in sweden. Ph.D. diss., University of Gothenburg, Sweden.

Blåberg, Olli 1988. *A study of Swedish compounds*. Department of General Lingusistics, University of Umeå. Report 29.

Brin, Sergey and Lawrence Page 1998. The anatomy of a large-scale hypertexual web search engine. *Seventh International World-Wide Web Conference (WWW 1998)*.

Brodda, Benny 1979. Något om de svenska ordens fonotax och morfotax: Iakttagelser med utgångspunkt från experiment med automatisk morfologisk analys. *Pilus*. Department of Linguistics, Stockholm University. Also in: "I huvet på Benny Brodda" Festskrift till densammes 65-årsdag.

Buckley, Chris and Ellen M. Voorhees 2004. Retrieval evaluation with incomplete information. *Proceedings of the 27th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 25–32. Sheffield, UK: ACM Press.

Buckley, Chris and Ellen M. Voorhees 2005. Retrieval system evaluation. Ellen M. Voorhees and Donna K. Harman (eds), *TREC – Experiment and evaluation in information retrieval*. MIT Press.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form*. John Benjamins Publishing Company.

Carlberger, Johan, Hercules Dalianis, Martin Hassel and Ola Knutsson 2001. Improving precision in information retrieval for Swedish using stemming. Technical Report IPLab-194, TRITA-NA-P0116, NADA-KTH, Royal Institute of Technology, Stockholm, Sweden.

Carmel, David, Elad Yom-Tov, Adam Darlow and Dan Pelleg 2006. What makes a query difficult? *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Seattle, Washington, USA: ACM Press.

Cleverdon, C.W. 1967. The Cranfield tests on index language devices. *Aslib proceedings*, Volume 19, 173–192.

Cormack, Gordon V., Christopher R. Palmer and Charles L. A. Clarke 1998. Efficient construction of large test collections. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 282–289.

Cöster, Rickard, Magnus Sahlgren and Jussi Karlgren 2004. Selective compound splitting of Swedish queries for Boolean combinations of truncated terms. *Proceedings of Fourth Workshop of the Cross-Language Evaluation Forum, (CLEF)*.

Croft, W. Bruce, Donald Metzler and Trevor Strohman 2010. *Search engines – Information retrieval in practice*. Addison Wesley.

Croft, William and D. Alan Cruse 2004. *Cognitive lingustics*. Cambridge University Press, Cambridge.

Dalianis, Hercules, Martin Hassel and Sumithra Velupillai 2009. The Stockholm EPR corpus – Characteristics and some initial findings. *Proceedings of ISHIMIR 2009, Evaluation and implementation of e-health and health information initiatives: international perspectives. 14th International Symposium for Health Information Management Research*. Kalmar, Sweden.

Dioşan, Laura, Alexandrina Rogozan and Jean-Pierre Pècuchet 2009. Automatic alignment of medical terminologies with general dictionaries for an efficient information retrieval. Violaine Prince and Mathieu Roche (eds), *Information retrieval in biomedicine: Natural language processing for knowledge integration*, 78–105. Medical Information Science Reference.

Dura, Elżbieta 1998. Parsing words. Ph.D. diss., University of Gothenburg.

Elhadad, Noemie and Komal Sutaria 2007. Mining a lexicon of technical terms and lay equivalents. *BioNLP 2007: Biological, translational and clinical language processing*, 49–56. Association of Computational Linguistics, Prague.

Friberg [Heppin], Karin 2006. Sjuka uppslag. Åsa Abelin and Roger Källström (eds), *Från urindoeuropeiska till ndengerenko*, Volume 56 of *MISS, Meddelanden från Institutionen för svenska språket*, 1–21. Göteborgs universitet [University of Gothenburg]. ISSN 1102-4518.

Friberg [Heppin], Karin 2007. Bättre klimat med rätt information – Vi bygger MedEval, en medicinsk testkollektion för svensk forskning inom informationssökning. *Humanistdagboken nr 20, Klimat, Humanistdagarna 2007*, Volume 20, 45–52. Göteborgs universitet [University of Gothenburg]. ISBN 978-91-7360-355-3.

Friberg Heppin, Karin 2008. MedEval – The construction of a Swedish medical test collection. *Proceedings of the 2nd Swedish Language Technology Conference, SLTC 2008*. Stockholm.

Friberg Heppin, Karin 2009. MedEval – Six test collections in one. *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA*, 223–226. Odense, Denmark: NEALT Proceedings Series.

Friberg Heppin, Karin 2010. MedEval – A Swedish medical test collection with doctors and patients user groups. *Proceedings of the North American Conference of the Association for Computational Linguistics — Human Language Technologies, NAACL-HTL 2010*. Los Angeles, Ca.

Gillå, Urban 2005. *Medicinsk grundkurs*. Stockholm: Bonnier Utbildning.

Hahn, Udo, Martin Honeck, Michael Piotrowski and Stefan Schulz 2001. Subword segmentation – Leveling out morphological variations for medical document retrieval. *Proceedings AMIA Annual Symposium*, 229–233.

Hahn, Udo, Martin Honeck and Stefan Schulz 2002. Subword-based text retrieval. *Proceedings of the 36th Hawaii International Conference on System Sciences*.

Hahn, Udo, Kornél Markó and Stefan Schulz 2005. Subword clusters as lightweight interlingua for multilingual document retrieval. *Conference proceedings: the 10th Machine Translation Summit*, 17–24.

Harman, Donna 1991. How effective is suffixing? *Journal of the American Society for Information Science* 42 (1): 7–15.

Harman, Donna K. 2005. The TREC test collections. Ellen M. Voorhees and Donna K. Harman (eds), *TREC – Experiment and evaluation in information retrieval*. Cambridge, Massachusetts: MIT Press.

Harter, Stephen P. 1996. Variations in relevance assessments and the measurement of retrieval effectiveness. *Journal of the American Society for Information Science* 47 (1): 37–49.

He, Ying and Mehmet Kayaalp 2006. A comparison of 13 tokenizers on MEDLINE. Technical Report, ⟨http://www.lhncbc.nlm.nih.gov/lhc/docs/reports/2006/tr2006003.pdf⟩.

Hedlund, Turid 2002. Compounds in dictionary-based cross-language information retrieval. *Information Research* 7, no. 2.

Hersh, William, Chris Buckley, TJ Leone and David Hickam 1994. OHSUMED: An interactive retrieval evaluation and new large test collection for research. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 192–201.

Hersh, William, Henning Müller and Jayashree Kalpathy-Cramer 2008. The ImageCLEFmed medical image retrieval task test collection. *Journal of Digital Imaging*.

Hersh, William and Ellen Voorhees 2009. TREC genomics special issue overview. *Information Retrieval* 12, no. 1.

Hersh, William R. 2003. *Information retieval – A health and biomedical perspective*. 2. Springer.

Hersh, William R., Henning Müller, Jeffery R. Jensen, Jianji Yang, Paul N. Gorman and Patrick Ruch 2006. Advancing biomedical image retrieval: Development and analysis of a test collection. *Journal of the American Medical Informatics Association* 13 (5): 488–496.

Holmes, Philip and Ian Hinchliffe 2003. *Swedish: A comprehensive grammar*. 2. Routledge.

Hull, David A. 1996. Stemming algorithms – A case study for detailed evaluation. *Journal of the American Society for Information Science*.

INEX n.d. *INEX relevance assessment guide*. ⟨http://qmir.dcs.qmul.ac.uk/inex/Papers/INEX02_Relevance_Assessment_Guide.pdf⟩. The reference created May 16, 2007.

Ingwersen, Peter and Kalervo Järvelin 2005. *The Turn – Integration of information seeking and retrieval in context*. The Kluwer International Series on Information Retrieval. Springer.

Jensen, John T. 1990. *Morphology – Word structure in generative grammar*. Volume 70 of *Current Issues in Linguistic Theory*. John Benjamins Publishing Company.

Justeson, John S. and Slava M. Katz 1995. Technical terminology: Some linguistic properties and an algorithm for indentification in text. *Natural Language Engineering* 1 (1): 9–27.

Järborg, Jerker 2003. Semantisk uppmärkning – Metoder, problem och resultat. Research Reports from the Department of Swedish Language.

Järvelin, Kalervo and Jaana Kekäläinen 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20 (4): 422–446.

Järvelin, Kalervo, Ilkka Vähämöttönen, Heikki Keskustalo and Jaana Kekäläinen 2008. VisualVectora: An interactive visualization tool for cumulated gain-based retreival experiments. M. Sanderson & al. (ed.), *Workshop on novel methodologies for evaluation in information retrieval*. European Conference for Information Retrieval — ECIR '08, Glasgow, UK.

Karlgren, Jussi 2000. Stylistic experiments for information retrieval. Ph.D. diss., Department of Linguistics, Stockholm University.

Kokkinakis, Dimitrios 2001. A framework for the aquisition of lexical knowledge; Description and applications. Ph.D. diss., Department of Swedish, University of Gothenburg.

Kokkinakis, Dimitrios 2004. MEDLEX: Technical report. Technical Report, Department of Swedish, University of Gothenburg, ⟨http://demo.spraakdata.gu.se/svedk/pbl/MEDLEX_work2004.pdf⟩.

Kokkinakis, Dimitrios 2009. Lexical granularity for automatic indexing and means to achieve it: The case of Swedish MeSH. Violaine Prince and Mathieu Roche (eds), *Information retrieval in biomedicine: Natural language processing for knowledge integration*, 11–37. Medical Information Science Reference.

Kokkinakis, Dimitrios and Maria Toporowska Gronostaj 2006. Comparing lay and professional language in cardiovascular disorders corpora. *WSEAS Transactions on biology snf biomedicine*, Volume 3, 429–437.

Kraaij, Wessel 2004. Variations on language modeling for information retrieval. Ph.D. diss., Centre for Telematics and Information Technology, Enschede, The Netherlands.

Källström, Roger 1999. *Morfologi*. Volume 9 of *Guling*. Department of Linguistics, University of Gothenburg.

Lancaster, F. W. 1969. MEDLARS: Report on the evaluation of its operating efficiency. *American Documentation* 20: 119–142.

Larsen, Birger and Andrew Trotman 2006. INEX 2006 guidelines for topic development. ⟨http://inex.is.informatic.uni-duisburg.de/2006/inex06/pdf/TD06.pdf⟩.

Lemur n.d. *The Lemur Toolkit for language modeling and information retrieval*. Carnegie Mellon University and the University of Massachusetts. ⟨www.lemurproject.org/⟩.

Lieber, Rochelle and Pavol Štekauer (eds) 2009. *The Oxford handbook of compounding*. Oxford Handbooks in Linguistics. New York: Oxford University Press.

Lindskog, Bengt 2004. *Medicinsk miniordbok*. 6. Stockholm: Nordiska Bokhandelns Förlag.

Lovins, J. 1968. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics* 11: 22–31.

Luhn, H.P. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM journal of Research and Development* 1 (4): 309–317.

Luhn, H.P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2 (2): 159–165.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze 2008. *Introduction to information retrieval*. Cambridge University Press.

Markó, Kornél, Stefan Schulz and 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: 567–545.

Martin, Melanie J. 2010. Reliability and type of consumer health documents on the world wide web: an annotation study. *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents*, 39–45. Association for Computational Linguistics, Los Angeles, California.

MeSH . *Svenska MeSH: MeSH-resurser vid KIB*. Karolinska Institutet Universitetsbiblioteket, Stockholm.

Metzler, Don 2005. Indri retrieval model overview. ⟨http://ciir.cs.umass.edu/~metzler/indriretmodel.html⟩.

Metzler, Donald and Bruce Croft 2004. Combining the language model and inference network approaches to retrieval. *Information processing & management*.

Metzler Jr., Donald A. 2007. Beyond bags of words: Effectively modeling dependence and features in information retrieval. Ph.D. diss., Graduate School of the University of Massachusetts Amherst.

Meystre, S.M., G.K. Savova, K.C. Kipper-Schuler and J.F. Hurdle 2008. Extracting information from textual documents in the electronic health record: A review of recent research. *IMIA Yearbook of Medical Informatics*, pp. 128–144.

Page, Lawrence, Sergey Brin, Rajeev Motwani and Terry Winograd 1999. The PageRank citation ranking: Bringing order to the web. Technical report 1999-66. Stanford InfoLab.

Pirkola, Ari and Kalervo Järvelin 2001. Employing the resolution power of search keys. *Journal of the American Society for Information Science and Technology* 52 (7): 575–583.

Porter, M.F. 1980. An algorithm for suffix stripping. *Program* 14 (3): 130–137.

van Rijsbergen, C.J. 1979. *Information retrieval*. 2. London: Butterworths.

Robertson, S.E. and Karen Spärck Jones 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science* 27 (3): 129–146.

Robertsson, Stephen 2008. On the history of evaluation in IR. *Journal of Information Science* 34 (4): 439–456.

Rosell, Magnus and Sumithra Velupillai 2005. The impact of phrases in document clustering for Swedish. *Proceedings of the 15th Nordic Conference on Computational Linguistics, NODALIDA '05*. Joensuu, Finland.

Salton, Gerard 1981. A blueprint for automatic indexing. *SIGIR Forum* 16 (2): 22–38.

Salton, Gerard and Michael J. McGill 1983. *Introduction to modern information retrieval*. McGraw-Hill Book Company.

Sanderson, Mark and Hideo Joho 2004. Forming test collections with no system pooling. *SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 33–40. ACM.

Saracevic, Tefko 1975. Relevance: A review and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science* 39 (3): 321–343.

Saracevic, Tefko 1999. Information science. *Journal of the American Society for Information Science* 50 (12): 1051–1063.

Schulz, Stefan, Martin Honek and Udo Hahn 2002. Biomedical text retrieval in languages with a complex morphology. *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*, Volume 3, 61–68.

Sormunen, Eero 2002. Liberal relevance criteria of TREC – Counting on negligible documents? *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Sormunen, Eero, Kai Halttunen and Heikki Keskustalo 2002. Query Performance Analyser – a tool for bridging information retrieval research and

instruction. Research Notes RN 1. Department of Information Studies, University of Tampere.

Sormunen, Eero, Jaana Kekäläinen, Jussi Koivisto and Kalervo Järvelin 2001. Document text characteristics affect the ranking of the most relevant documents by expanded structured queries. *Journal of Documentation* 57 (3): 358–376.

Spärck Jones, Karen 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28 (1): 11–21.

Spärck Jones, Karen and C.J. van Rijsbergen 1975. Report on the need for and provision of an "ideal" information retrieval test collection. *British Library Research and Development Report 5266.* Computer Library, University of Cambridge.

Strohman, Trevor 2007. Efficient processing of complex features for information retrieval. Ph.D. diss., University of Massachusetts Amherst.

Strzalkowski, Tomek 1994. Robust text processing in automated information retrieval. *Proceedings of the fourth conference on Applied natural language processing, stuttgart*, 168–173. Association for Computational Linguistics.

Svensén, Bo 2009. *A handbook of lexicography – The theory and practice of dictionary-making*. Cambridge University Press.

Teleman, Ulf 1972. *Om svenska ord*. Gleerups.

Turtle, Howard and W. Bruce Croft 1990. Inference networks for document retrieval. *Proceedings of the thirteenth international conference on research and development in information retrieval*, 1–24. Association for Computing Machinery.

Velupillai, Sumithra 2009. Swedish health data – Information access and representation. Licentiate thesis, Stockholm University.

Velupillai, Sumithra, Hercules Dalianis, Martin Hassel and Gunnar H. Nilsson 2009. Developing a standard for de-identifying electronic patient records written in Swedish: Precision, recall and F-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics*.

Volk, Martin and Paul Buitelaar 2002. A systematic evaluation of concept-based cross-lingual information retrieval in the medical domain. *Proceedings of 3rd Dutch-Belgian Information Retrieval Workshop*. Leuven.

Voorhees, Ellen 2001. Evaluation by highly relevant documents. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 74–82.

Voorhees, Ellen M. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. *SIGIR'98*. Melbourne, Australia.

Voorhees, Ellen M. and Donna K. Harman (eds) 2005. *TREC – Experiment and evaluation in information retrieval*. Cambridge, Massachusetts: MIT Press.

Zipf, G.K. 1949. *Human behavior and the principle of least effort*. Addison Wesley Publishing.

Zobel, Justin 1998. How reliable are the results of large-scale information retrieval experiments? *Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 307–314.

A TOPICS

<TOP>
<TOPNO> 1 </TOPNO>
<TITLE> Fettsnål kosts inverkan på LDL och HDL. <TITLE>
<DESC> Hur påverkar en fettsnål, energifattig kost koncentrationerna av HDL och LDL? </DESC>
<NARR> Relevanta dokument ska innehålla information om lipoproteinerna HDL och LDL/lätta lipoproteiner och deras funktion, samt beskriva hur en kostförändring påverkar koncentrationen av dessa i blodet. Beskrivning av fettsnål kost är relevant.
</NARR>
</TOP>

<TOP>
<TOPNO>2</TOPNO>
<TITLE> Försiktighet vid behandling med erytromycin under graviditet
</TITLE>
<DESC> Vilka komplikationer kan vara gällande vid behandling av infektion hos gravida med erytromycin? </DESC>
<NARR> Relevanta dokument ska innehålla information om erytromycin/ Ery-Max/Erymax och dess verkan samt indikationer. Information om biverkningar med fokus på graviditet/havandeskap är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>4</TOPNO>
<TITLE> Blodpropp vid användning av östrogen och gestagen</TITLE>
<DESC> Vilka risker för utvecklandet av blodpropp finns vid användande av kombinationspreparat som innehåller östrogen och gestagen? </DESC>
<NARR> Relevanta dokument ska innehålla information om kombinationspreparat som innehåller östrogen och gestagen. Preparatens verkan ska beskri-

vas. Även risker, framförallt för blodpropp/trombos är relevant.
</NARR>
</TOP>

<TOP>
<TOPNO>5</TOPNO>
<TITLE> Reductil vid behandling av fetma </TITLE>
<DESC> När är det indicerat att börja behandla fetma med preparatet reductil? </DESC>
<NARR> Relevanta dokument ska innehålla information om reductil och dess verkan samt biverkningar och när det är indicerat att sätta in behandling för fetma/övervikt. Beskrivning av behandlingsstrategin är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>7</TOPNO>
<TITLE> Biverkningar vid cellgiftsbehandling av cancer </TITLE>
<DESC> Vilka biverkningar kan man räkna med vid behandling av cancer med cellgifter? </DESC>
<NARR> Relevanta dokument innehåller information om cellgifter/cytostatika, deras biverkningar och vilka typer av cancer som bör behandlas med dessa. Beskrivning av strategin vid cellgiftsbehandling är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>9</TOPNO>
<TITLE> Testosteron, fetma och blodtryck hos män </TITLE>
<DESC> Hur påverkar testosteronnivån mäns fetma och blodtryck?
</DESC>
<NARR> Relevanta dokument ska innehålla information om testosteronets roll i utvecklandet av fetma/övervikt hos män. Dokument som innehåller information om huruvida fetma påverkar testosteronproduktionen hos män eller som beskriver testosteronets verkan på blodtrycket hos män är relevanta.
</NARR>
</TOP>

<TOP>
<TOPNO>10</TOPNO>
<TITLE> Att behandla med xenical vid samtidig diabetes och/eller högt blodtryck </TITLE>
<DESC> Hur man går till väga när man behandlar fetma med xenical vid

samtidigt högt blodtryck och/eller diabetes typ II. </DESC>
<NARR> Relevanta dokument ska innehålla information om viktminskningspreparatet xenical samt dess indikation och eventuella biverkningar. Information om försiktighet vid fetmabehandling/behandling av övervikt av patienter som lider av diabetes/ sockersjuka typ II och/eller högt blodtryck är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>11</TOPNO>
<TITLE> Biverkningar hos reductil och xenical </TITLE>
<DESC> Vilka biverkningar har de två viktminskningspreparaten reductil och xenical? </DESC>
<NARR> Relevanta dokument ska innehålla information om xenical och reductil, samt de biverkningar som kan uppstå vid behandling av fetma/övervikt med dessa viktreduceringspreparat. </NARR>
</TOP>

<TOP>
<TOPNO>12</TOPNO>
<TITLE> Xenical och mängden fett man tar upp från kost </TITLE>
<DESC> Verkningsmekanism för xenicals inverkan på fettupptag från kosten. </DESC>
<NARR> Relevanta dokument ska innehålla information om läkemedlet xenical, om dess indikation, terapeutisk strategi, samt biverkningar. Information om hur fettupptag från kost påverkas är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>13</TOPNO>
<TITLE> Concerta vid behandling av ADHD </TITLE>
<DESC> På vilka indikationer används läkemedlet concerta vid ADHD-behandling? </DESC>
<NARR> Relevanta dokument ska innehålla information om concerta och dess roll i behandlingen av ADHD/uppmärksamhetsstörning med hyperaktivitet. Information om regler för förskrivning är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>16</TOPNO>
<TITLE> Hormonbehandling under klimakteriet och risken för bröstcancer

</TITLE>
<DESC> Ökar eller minskar riskerna för bröstcancer då man undergår hormonbehandling under klimakteriet? </DESC>
<NARR> Relevanta dokument ska innehålla information om hormonbehandling under klimakteriet/menopaus och behandlingens risker. De ska beskriva relationen mellan behandling med hormoner och bröstcancer. </NARR>
</TOP>

<TOP>
<TOPNO>18</TOPNO>
<TITLE> Hormonbehandling under klimakteriet och risken för cervixcancer </TITLE>
<DESC> Ökar eller minskar riskerna för cervixcancer då man undergår hormonbehandling under klimakteriet? </DESC>
<NARR> Relevanta dokument ska innehålla information om hormonbehandling under klimakteriet/menopaus och behandlingens risker. De ska beskriva relationen mellan behandling med hormoner och cervixcancer/livmoderhalscancer. </NARR>
</TOP>

<TOP>
<TOPNO>19</TOPNO>
<TITLE> Johannesört och läkemedelsanvändning/läkemedelskonsumtion </TITLE>
<DESC> Hur påverkas läkemedel av samtidigt intag av johannesört? </DESC>
<NARR> Dokumenten ska innehålla information om johannesört/hypericum och dess eventuella interaktion/korsbiverkan med läkemedel, godkända av läkemedelsverket. </NARR>
</TOP>

<TOP>
<TOPNO>20</TOPNO>
<TITLE> Receptbelagda preparat vid hudsjukdom </TITLE>
<DESC> Vilket eller vilka receptbelagda preparat finns att tillgå vid hudsjukdom, exempelvis acne </DESC>
<NARR> Relevanta dokument ska innehålla information om receptbelagda preparat inom specialiteten hudsjukdomar/dermatos. Information om sjukdomar som kräver behandling med dylika preparat, exempelvis acne/akne/finnar, är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>21</TOPNO>
<TITLE> Antibiotika som bör undvikas före och under graviditet </TITLE>
<DESC> Vilka typer av antibiotika kan ge möjlig fosterskada eller skada på den gravida kvinnan och bör undvikas före och under graviditet/ havandeskap?
</DESC>
<NARR> Relevanta dokument ska innehålla information om olika sorters antibiotika som kan orsaka fosterskador. Beskrivning av vilka typer av fosterskador eller skadeverkningar på den gravida kvinnan orsakade av antibiotika som kan förekomma är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>23</TOPNO>
<TITLE> Risker vid användning av neuroleptika </TITLE>
<DESC>Vilka risker är förknippade med användandet av neuroleptika?
</DESC>
<NARR> Relevanta dokument skall innehålla generell information gällande neuroleptika/antipsykosmedel, dess indikationer, biverkningar och behandlingsalternativ. Information om de olika sjukdomstillstånd där neuroleptika används för behandling är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>24</TOPNO>
</TITLE> Behandling vid akut njursvikt </TITLE>
<DESC> Hur behandlar man en patient med tecken på akut njursvikt?
</DESC>
<NARR> Relevanta dokument ska innehålla beskrivning av njursvikt och dess symtom samt mekanismen bakom uppkomsten, renala, post- och prerenala. Beskrivning av behandlingsmetoder, såväl medicinska som kirurgiska, är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>25</TOPNO>
<TITLE> Symtom och behandling efter TIA </TITLE>
<DESC> Vilka symtom och vilken efterföljande behandling får man vid TIA?
</DESC>

<NARR> Relevanta artiklar ska innehålla information gällande TIA/hjärnischemi/cerebral ischemi, transitorisk ischemisk attack, och dess symtombild. Behandlingsalternativ och strategier är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>26</TOPNO>
<TITLE> Ställa diagnosen akut koronart syndrom </TITLE>
<DESC> Vilka kriterier och tekniker används för diagnosticering av akuta koronara syndrom? </DESC>
<NARR> Relevanta dokument ska innehålla beskrivning av akut koronart syndrom/koronarsjukdom/ kranskärlssjukdom samt de metoder som används för ställande av diagnos. Metoder som beskrivs kan vara såväl tekniska som laboratoriemässiga. </NARR>
</TOP>

<TOP>
<TOPNO>27</TOPNO>
<TITLE> Oxascand som ångestdämpande vid alkoholavgiftning </TITLE>
<DESC> Hur används oxascand som ångestdämpande preparat vid alkoholavgiftning? </DESC>
<NARR> Relevanta dokument ska innehålla information om preparatet oxascand som ångestdämpande preparat/anxiolytika, dess verkan, dosering och eventuell risk för tillvänjning vid alkoholavgiftning. </NARR>
</TOP>

<TOP>
<TOPNO>28</TOPNO>
<TITLE> Behandling med bensodiazepiner </TITLE>
<DESC> Vilka indikationer föreligger vid behandling med bensodiazepiner? När ska preparatet användas? </DESC>
<NARR> Relevanta dokument ska innehålla information om bensodiazepiner, när de ska sättas in som behandling, dosering och försiktigheter. Information om de tillstånd som behandlas med bensodiazepiner är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>31</TOPNO>
<TITLE> Behandling med SSRI-preparat vid depression hos gravida
</TITLE>
<DESC> Vilka preparat av typen SSRI kan användas vid behandling av de-

pression hos gravida? </DESC>
<NARR> Relevanta dokument ska innehålla information om SSRI-preparat/ serotoninupptagshämmare, som är lämpliga för behandling av depression/nedstämdhet hos gravida/havande. Skillnader mellan doser och terapiintervall samt försiktigheter är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>32</TOPNO>
<TITLE> Behandling med SSRI-preparat vid depression hos barn </TITLE>
<DESC> Indikationer för behandling av depression hos barn med SSRI.
</DESC>
<NARR> Relevanta dokument ska innehålla information om huruvida preparat av typen SSRI, serotoninupptagshämmare, kan användas för att behandla depression/nedstämdhet hos barn. Bekrivning av godkända preparat och deras effekt är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>36</TOPNO>
<TITLE> Effekter och interaktioner vid användning av waran </TITLE>
<DESC> Vilka effekter och interaktioner med andra läkemedel kan man förvänta sig vid användning av waran? </DESC>
<NARR> Relevanta dokument ska innehålla information om waran, indikation, effekter, biverkningar samt interaktion med andra läkemedel. Dokument som innehåller information om hur man tacklar uppkomna tillstånd är också relevanta. </NARR>
</TOP>

<TOP>
<TOPNO>37</TOPNO>
<TITLE> Allvarliga biverkningar vid nyttjande av naturläkemedel </TITLE>
<DESC> Hur yttrar sig biverkningar av allvarlig karaktär vid nyttjande av naturläkemedel? </DESC>
<NARR> Relevanta dokument innehåller information om vanliga typer av naturläkemedel, deras indikationer samt eventuella biverkningar av allvarlig karaktär. Dokument som beskriver hur man behandlar dessa biverkningar är också relevanta. </NARR>
</TOP>

<TOP>
<TOPNO>38</TOPNO>
<TITLE> Viagra och högt blodtryck </TITLE>
<DESC>Är det lämpligt att behandla en patient med viagra om patienten står på en reguljär behandling för högt blodtryck </DESC>
<NARR> Relevanta dokument ska innehålla information om viagra/sidenafil, erektil dysfunktion samt om högt blodtryck/hypertoni och behandling för detta. Information om huruvida man kan kombinera viagra med behandling för högt blodtryck är relevant samt vilka former av interaktion som kan vara aktuella. </NARR>
</TOP>

<TOP>
<TOPNO>39</TOPNO>
<TITLE> Medicinsk behandling vid Alzheimers sjukdom </TITLE>
<DESC> Vilken medicinsk behandling rekommenderas för en patient som drabbats av Alzheimers sjukdom. </DESC>
<NARR> Relevanta dokument ska innehålla information rörande Alzheimers sjukdom/åldersdemens, såsom klinisk bild, sjukdomsförlopp och prognos. Information om indikationer för och varianter av medicinsk behandling och deras effekter är relevanta. </NARR>
</TOP>

<TOP>
<TOPNO>41</TOPNO>
<TITLE> Erytropoietinbehandling vid anemi i samband med cancer
</TITLE>
<DESC> Hur används erytropoietin vid cancerrelaterad anemi? </DESC>
<NARR> Relevanta dokument ska innehålla information om erytropoietin/ erytropoetin/EPO och dess roll i behandling av anemi. Information om olika sorters cancer som kan ge anemi/blodbrist är relevant. Även information om cancerterapier som kan leda till anemi är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>42</TOPNO>
<TITLE> Antiviral behandling vid herpes simplex samt herpes zoster
</TITLE>
<DESC> Vad finns för behandling i samband med herpes zoster och herpes simplex i form av antiviral terapi/virushämmare. </DESC>
<NARR> Releventa dokument ska innehålla information om antivirala/virus-

hämmande preparat, deras administrationssätt och preparattyper. Tid för behandling, som när och duration är relevant, liksom information om infektionerna herpes simplex och herpes zoster. </NARR>
</TOP>

<TOP>
<TOPNO>43</TOPNO>
<TITLE> Kvinnors behandling vid uvi </TITLE>
<DESC> Vilken behandling erhåller kvinnor i samband med uvi? </DESC>
<NARR> Relevanta dokument ska beskriva behandling vid uvi/infektion i urinvägarna/urinvägsinfektion som drabbar kvinnor. Behandlingstid, preparattyper samt information om olika agens är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>44</TOPNO>
<TITLE> Kvinnors symtom vid uvi </TITLE>
<DESC> Vilka symtom uppvisar kvinnor vid uvi? </DESC>
<NARR> Relevanta dokument beskriver infektion i de övre urinvägarna/uvi/urinvägsinfektion hos kvinnor. Olika riskfaktorer och orsaker till uppkomst är relevant liksom symtom och diagnostiska kriterier. </NARR>
</TOP>

<TOP>
<TOPNO>46</TOPNO>
<TITLE> Vaccinationsprogrammet och allergier hos barn </TITLE>
<DESC> Vad innebär vaccinationsprogrammet och finns det risk att barn utvecklar allergi i samband med detta? </DESC>
<NARR> Relevanta dokument ska innehålla information om vaccinationsprogrammet med uppgifter om de vaccin som används och de sjukdomar som det vaccineras mot. Risker såsom utvecklandet av allergier samt andra komplikationer som kan drabba ett barn kopplat till vaccination är relevant.
</NARR>
</TOP>

<TOP>
<TOPNO>48</TOPNO>
<TITLE> Prognos vid cancer i matstrupen </TITLE>
<DESC> Vad är prognosen vid olika typer av cancer i matstrupen? </DESC>
<NARR> Relevanta dokument ska innehålla information om cancertyper i matstrupen/esofagus, såsom symtom, behandlingsalternativ och riskfaktorer.

Beskrivning av prognos beroende av stadie och grad, samt lokal är relevant. Information angående könsskillnader vid prognosen är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>49</TOPNO>
<TITLE> Hereditet och risken för hjärt- och kärlsjukdomar hos kvinnor
</TITLE>
<DESC> Vad innebär hereditet som riskfaktor för kvinnors utvecklande av hjärt- och kärlsjukdom. </DESC>
<NARR> Relevanta dokument ska innehålla information om risken för hjärt-kärlsjukdom hos kvinnor samt vilken roll hereditet/ärftlighet spelar i utvecklandet av ovan nämnda sjukdomar. Information om vilka typer av hjärt-kärlsjukdomar på vilka hereditet har störst inverkan är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>50</TOPNO>
<TITLE> NSAID och magens slemhinna vid ulcus </TITLE>
<DESC> Hur påverkar preparat av typen NSAID magens slemhinna vid ulcus? </DESC>
<NARR> Relevanta dokument ska innehålla information om uppkomsten av ulcerationer i magslemhinnan/magsår och vanliga symtom. Information om NSAID-preparat/icke-steroida antiinflammatoriska medel och dess verkan på magslemhinnan är relevant.
</NARR>
</TOP>

<TOP>
<TOPNO>51</TOPNO>
<TITLE> Anemi och cancer </TITLE>
<DESC> Varför kan en patient med cancer drabbas av anemi? </DESC>
<NARR> Relevanta dokument ska innehålla information om vad anemi/blodbrist är, symtom, behandling och orsaker. Information om cancerrelaterad anemi dels utlöst av cancern och dels utlöst av cancerbehandlingen är relevant.
</NARR>
</TOP>

<TOP>
<TOPNO>53</TOPNO>
<TITLE> Studier av protein vid Alzheimers sjukdom </TITLE>
<DESC> Beskriv metoder som används för att studera protein som specifikt uppträder vid Alzheimers sjukdom. </DESC>
<NARR> Relevanta dokument ska innehålla information om molekylärbiologiska metoder som används för att studera protein som uppträder specifikt vid Alzheimers sjukdom/åldersdemens. Information om hur man histopatologiskt ställer diagnosen är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>54</TOPNO>
<TITLE> Degeneration av gula fläcken </TITLE>
<DESC> Varför utsätts gula fläcken för degeneration, vad innebär det och vilken behandling är lämplig?</DESC>
<NARR> Relevanta dokument ska innehålla information om gula fläcken/ macula lutea och om hur och varför den degenereras/makuladegeneration. Information om riskfaktorer, symtom, komplikationer och behandling är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>55</TOPNO>
<TITLE> Hodgkins och non-hodgkins lymfom </TITLE>
<DESC> Vilka likheter och skillnader finns det mellan de två typerna av lymfom: hodgkins och non-hodgkins? </DESC>
<NARR> Relevanta dokument ska innehålla definitioner av hodgkins lymfom/malignt lymfogranulom och non-hodgkins lymfom, symtom samt rekommenderad behandling för de två tillstånden och prognos för båda. </NARR>
</TOP>

<TOP>
<TOPNO>56</TOPNO>
<TITLE> Gastric bypass för reducering av kroppsvikt </TITLE>
<DESC> Varför utförs gastric bypass för viktreducering och hur tillämpas det? </DESC>
<NARR> Relevanta dokument ska innehålla information om indikation för att gastric bypass/magsäcks-bypass/reduktion av magsäcken ska genomföras, tillvägagångssätt samt förväntat resultat i form av viktnedgång. </NARR>
</TOP>

<TOP>
<TOPNO>57</TOPNO>
<TITLE> Symtom vid förgiftning med paracetamol </TITLE>
<DESC> Vad innebär förgiftning med paracetamol och vilka symtom uppvisar patienten? </DESC>
<NARR> Relevanta dokument beskriver paracetamol och definierar överdosering med mängd aktiv substans. Paracetamols verkan ska beskrivas liksom symtom vid överdosering och/eller förgiftning samt behandling. </NARR>
</TOP>

<TOP>
<TOPNO>58</TOPNO>
<TITLE> Leverpåverkan av kombinationen paracetamol och alkohol
</TITLE>
<DESC> Hur påverkas levern av samtidigt intag av paracetamol och alkohol?
</DESC>
<NARR> Relevanta dokument ska innehålla information om paracetamol och dess verkan på levern samt alkohol och dess verkan på levern. De ska även innehålla information om hur samtidig konsumtion av ovan nämnda substanser kan påverka leverns funktion. Beskrivning av skadliga mängder ska ingå. </NARR>
</TOP>

<TOP>
<TOPNO>62</TOPNO>
<TITLE> Smärtbehandling med opioider </TITLE>
<DESC> Hur bedömer man att en smärta kräver behandling med opioider?
</DESC>
<NARR> Relevanta dokument ska informera om smärta, smärtlindring, opioidpreparat, administrationssätt samt indikationer för insättande av opioider.
</NARR>
</TOP>

<TOP>
<TOPNO>63</TOPNO>
<TITLE> Biopsi vid misstänkt magsäckscancer </TITLE>
<DESC> Vilka tekniker och redskap används vid biopsi av magsäck vid cancermisstanke. </DESC>
<NARR> Relevanta dokument ska behandla indikation för magsäckscancer/ventrikelcancer/magcancer/magsäckstumörer/ventrikeltumörer, när man ska ta

en biopsi/vävnadsprov samt risker med att punktera magsäckens vägg. Information om hur man praktiskt går till väga vad gäller redskap, premedicinering samt förberedelse både för patient och för vårdpersonal är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>65</TOPNO>
<TITLE> Antihistaminbehandling vid pollenallergi </TITLE>
<DESC> Hur används antihistaminer vid behandling av pollenallergi?
</DESC>
<NARR> Relevanta dokument ska innehålla information om pollenallergi och dess behandling med fokus på antihistaminer. Beskrivning av indikation, administrationssätt samt försiktigheter vid användandet av antihistamin/histaminblockerare/histaminantagonister är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>66</TOPNO>
<TITLE> Behandling av allergisk anafylaxi </TITLE>
<DESC> Hur behandlas anafylaxi till följd av allergi? </DESC>
<NARR> Relevanta dokument ska innehålla information om anafylaxi/allergisk chock, med fokus på allergi samt dess behandling. Information om vilka preparat och behandlingsstrategier som används är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>67</TOPNO>
<TITLE> Steroider vid akut astma </TITLE>
<DESC> Hur används steroider vid behandling av akut astma? </DESC>
<NARR> Relevanta dokument ska innehålla information om astma samt steroiders roll vid behandling, primärt vid akuta situationer. Beskrivning av indikation, administrationssätt samt försiktigheter vid användandet av steroider är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>68</TOPNO>
<TITLE> Symtom och behandling vid DVT </TITLE>
<DESC> Vilka symtom associeras med DVT, djup ventrombos, och hur ser behandlingen ut? </DESC>

<NARR> Relevanta dokument ska innehålla information om DVT/djup ventrombos/venblodpropp, vilka symtom som manifesteras samt hur behandlingen ser ut. Information om diagnostik och väsentliga preparat är relevant.
</NARR>
</TOP>

<TOP>
<TOPNO>69</TOPNO>
<TITLE> Differentialdiagnoser vid symtomet bröstsmärta </TITLE>
<DESC> Vad finns det för differentialdiagnoser hos patient med bröstsmärta som symtom? </DESC>
<NARR> Relevanta dokument ska innehålla information om symtomet bröstsmärta samt viktiga differentialdiagnoser. Information om väsentlig diagnostik är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>73</TOPNO>
<TITLE> Appendicit som diagnos vid buksmärta </TITLE>
<DESC> Hur ställs diagnosen appendicit vid buksmärta? </DESC>
<NARR> Relevanta dokument ska innehålla information om appendicit/blindtarmsinflammation och om hur diagnosen ställs samt kliniska manifestationer såsom buksmärta. Information om väsentlig provtagning samt undersökningar är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>75</TOPNO>
<TITLE> Diagnostik av gastroenterit vid obehag från buken </TITLE>
<DESC> Hur ställs diagnosen gastroenterit vid obehag från buken?
</DESC>
<NARR> Relevanta dokument ska innehålla information om gastroenterit/gastroenterit/mag-tarmkatarr/magsjuka/mag-tarminflammation och om hur diagnosen ställs samt kliniska manifestationer såsom obehag från buken. Orientering om vanliga agens samt väsentlig provtagning och undersökningar är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>76</TOPNO>
<TITLE> Behandling av KOL-recidiv </TITLE>

<DESC> Hur behandlas ett recidiv av KOL? </DESC>
<NARR> Relevanta dokument ska innehålla information om KOL/kroniskt obstruktiv lungsjukdom med fokus på recidiv/återfall och hur behandlingen ser ut. Information om preparat och dess användande är relevant samt orientering kring orsaker till uppkomsten av KOL. </NARR>
</TOP>

<TOP>
<TOPNO>77</TOPNO>
<TITLE> Symtom av sepsis såsom feber och påverkan på blodtrycket
</TITLE>
<DESC> Vilka symtom ger sepsissjukdomen med fokus på feber och blodtrycksförändringar? </DESC>
<NARR> Relevanta dokument ska innehålla information om sepsis/blodförgiftning/septikemi, dess påverkan på blodtryck och feber. Information om bakomliggande orsaker till utvecklandet av sepsis samt vanliga agens är relevant.
</NARR>
</TOP>

<TOP>
<TOPNO>82</TOPNO>
<TITLE> Orsakerna bakom kramp och dess behandling </TITLE>
<DESC> Vilka är orsakerna bakom uppkomsten av kramp och hur behandlas det? </DESC>
<NARR> Relevanta dokument ska innehålla information om kramp/spasm och bakomliggande orsaker. Information om behandling är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>83</TOPNO>
<TITLE> Orsak till synkope och dess behandling </TITLE>
<DESC> Vilka är de bakomliggande orsakerna till synkope och hur behandlar man det? </DESC>
<NARR> Relevanta dokument ska innehålla information om synkope/svimning, de bakomliggande orsakerna samt behandling. </NARR>
</TOP>

<TOP>
<TOPNO>85</TOPNO>
<TITLE> Blödning vid överdosering av waran </TITLE>
<DESC> Hur påverkas risken för blödning vid överdosering av waran?

</DESC>
<NARR> Relevanta dokument ska innehålla information om hur waran påverkar risken för blödning/hemorragi. Information om dosering av waran är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>90</TOPNO>
<TITLE> Behandling vid acidos hos patient med diabetes </TITLE>
<DESC> Hur behandlas diabetiker med acidos? </DESC>
<NARR> Relevanta dokument ska innehålla information diabetesinducerad acidos/ketoacidos. Dokumenten ska innehålla information om lämplig behandling av patient med acidos med fokus på diabetespatienter/sockersjuka.
</NARR>
</TOP>

<TOP>
<TOPNO>92</TOPNO>
<TITLE> Behandling av eksem med steroider </TITLE>
<DESC> Hud: Hur går man till väga vid behandling av eksem med steroider?
</DESC>
<NARR> Relevanta dokument ska innehålla information om hur en hudläkare går till väga vid behandling av eksem med steroider. Information om olika typer av steroider som används vid behandling av eksem, samt när, var och hur man använder dem är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>94</TOPNO>
<TITLE> Spridning av HIV </TITLE>
<DESC> Vilka smittvägar och vilka risker finns för HIV-infektion?
</DESC>
<NARR> Relevanta dokument ska innehålla information om HIV/human immunodeficiency virus/humant immunbristvirus, om hur viruset smittar samt vilka smittvägar som finns. Information om viruset och dess egenskaper är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>96</TOPNO>
<TITLE> Instabil angina pectoris </TITLE>

<DESC> Vad innebär instabil angina pectoris? </DESC>
<NARR> Relevanta dokumenten ska innehålla information om angina pectoris/kärlkramp och andra hjärtsjukdomar. Information om vad som särskiljer en instabil angina pectoris från andra hjärtrelaterade sjukdomar är relevant. </NARR>
</TOP>

<TOP>
<TOPNO>97</TOPNO>
<TITLE> Hypertension och andra symtom i samband med hjärtsvikt
</TITLE>
<DESC> Hur påverkar hypertension och andra symtom en patient med hjärtsvikt? </DESC>
<NARR> Relevanta dokumenten ska innehålla information om hjärtsvikt/hjärtinkompensation/hjärtinsufficiens. Information om hypertension/hypertoni/högt blodtryck och förhållandet mellan detta symtom och hjärtsvikt, samt andra symtom som kan vara aktuella i samband med hjärtsvikt är relevant.
</NARR>
</TOP>

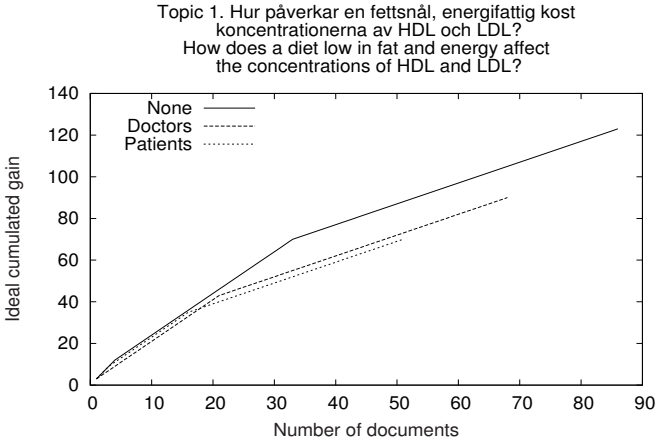
<TOP>
<TOPNO>100</TOPNO>
<TITLE> Rubbning av TSH-insöndring, påföljande hypertyreos och symtom
</TITLE>
<DESC> Vilka symtom uppvisar man vid hypertyreos orsakad av störd TSH-insöndring? </DESC>
<NARR> Relevanta dokument ska innehålla information om hypertyreos/sköldkörtelsjukdom. Hur TSH/tyreotropin påverkar thyreoidea/tyreoidea/sköldkörteln. Information om de symtom som en patient med sjukdomen uppvisar är relevant. </NARR>
</TOP>

B IDEAL CUMULATED GAIN

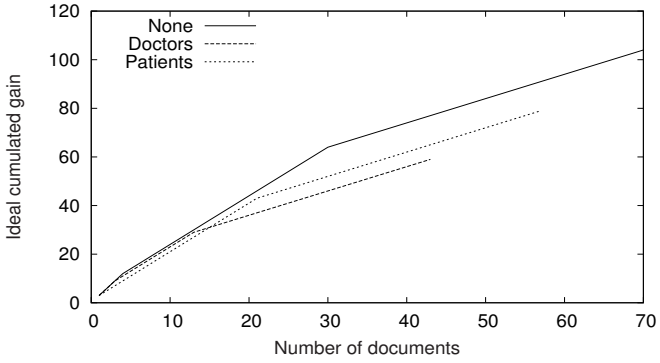
The graphs below show the ideal cumulated gain for the None, Doctors and Patients scenarios in the 30 topics with the largest sets of known relevant documents.

Note that the scales of the graphs are not identical. They are adjusted to the cumulated gain of each topic. First are topics with a fairly even number of doctor and patient documents. On the following pages are first topics with predominantly doctor documents, followed by pages with predominantly patient documents.

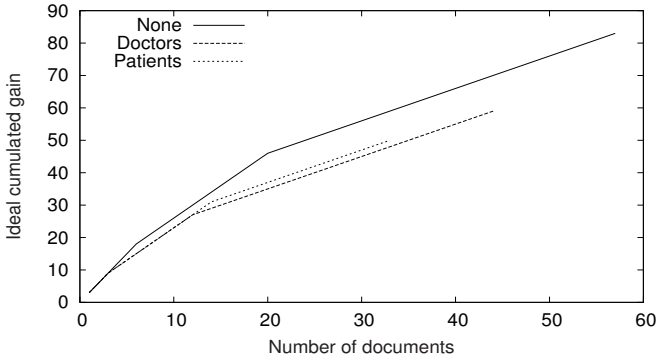
B.1 Topics with fairly even distribution of doctor and patient documents



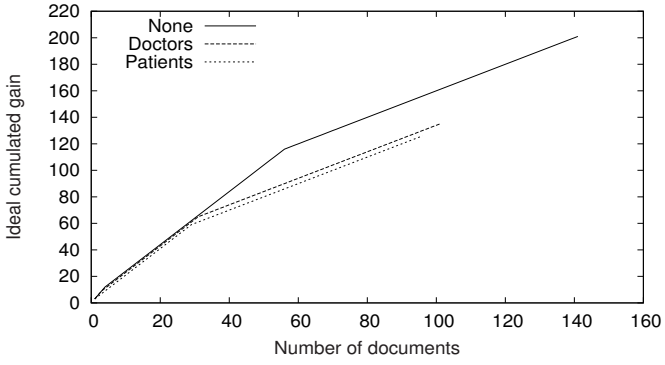
Topic 4. Vilka risker för utvecklandet av blodpropp finns vid användande av kombinationspreparat som innehåller östrogen och gestagen?
Which risks for developing a blood clot exist when using combination drugs containing estrogen and gestagen?



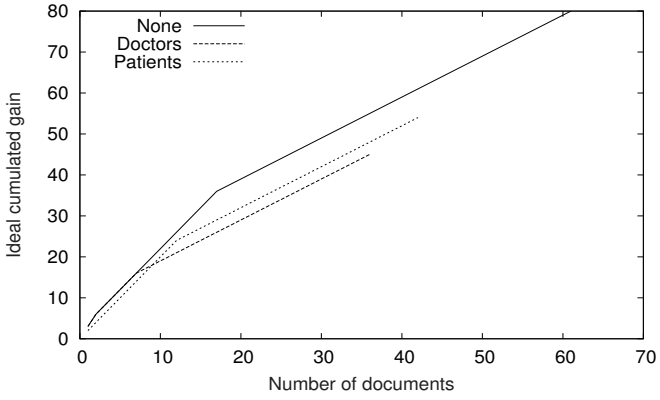
Topic 26. Vilka kriterier och tekniker används för diagnosticering av akuta koronara syndrom?
What criteria and which techniques are used for diagnosing acute coronary syndromes?



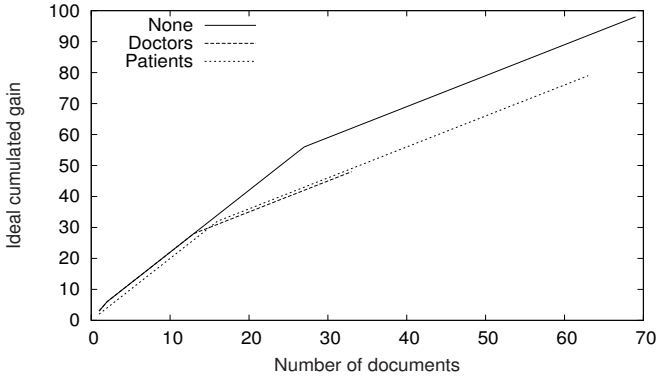
Topic 36. Vilka effekter och interaktioner med andra läkemedel kan man förvänta sig vid användning av waran?
Which effects and interactions with other medicines can be expected with the use of waran?



Topic 41. Hur används erytropoietin vid cancerrelaterad anemi?
How is erythropoietin used in cancer related anemia?

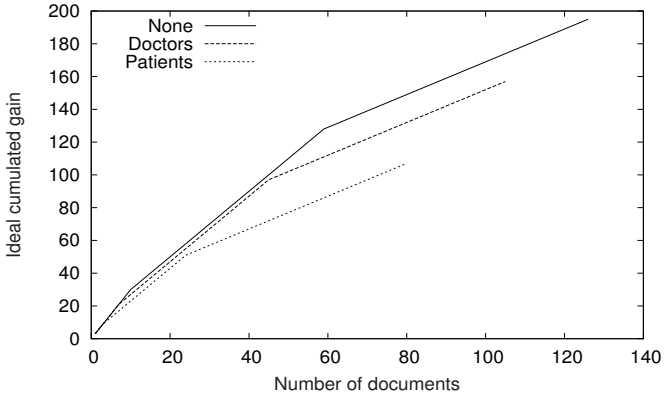


Topic 75. Hur ställs diagnosen gastroenterit vid obehag från buken?
How is the diagnosis gastroenteritis made
in discomfort from the abdomen?

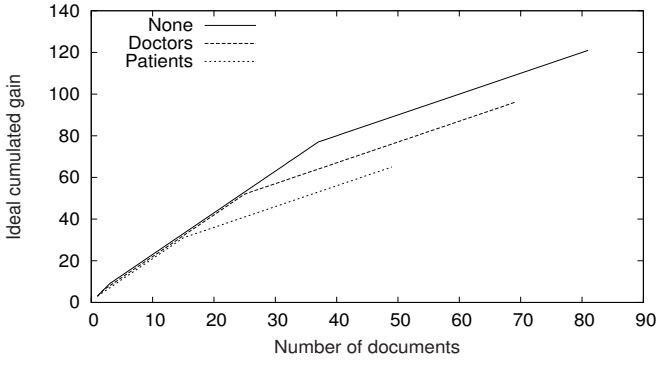


B.2 Topics with predominantly doctor documents

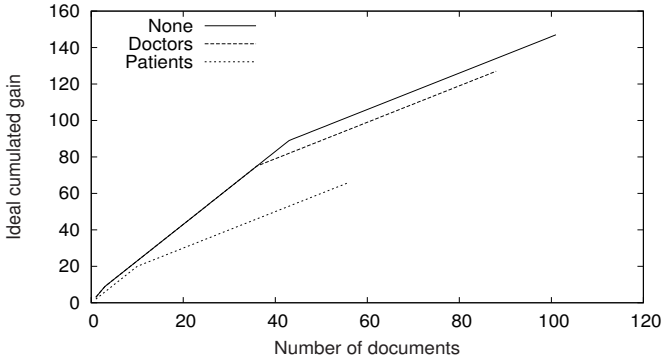
Topic 23. Vilka risker är förknippade med användandet av neuroleptika?
Which risks are associated with the use of neuroleptics?



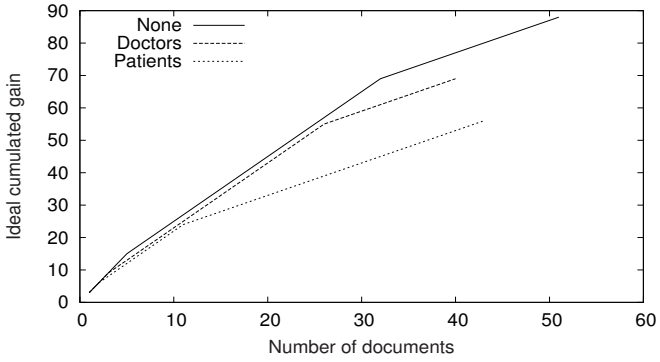
Topic 25. Vilka symtom och vilken efterföljande behandling
får man vid TIA?
Which symptoms and what subsequent treatment
does one get with TIA?



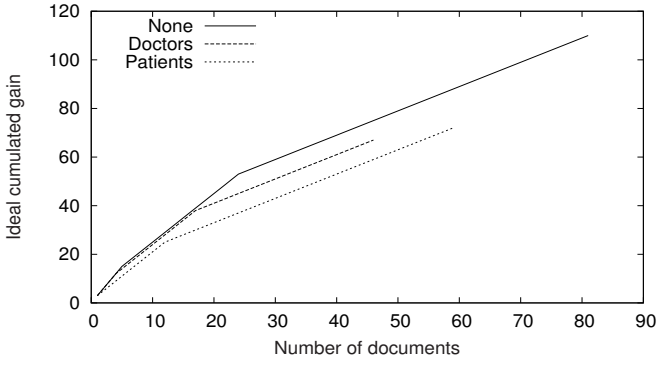
Topic 28. Vilka indikationer föreligger vid behandling
med benzodiazepiner? När ska preparatet användas?
Which indications exist for treatment with
benzodiazepines? When should the drug be used?



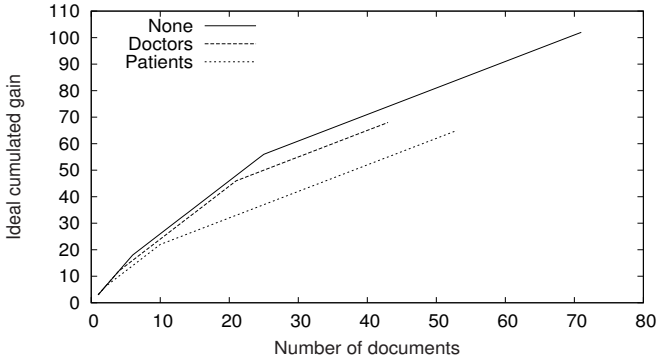
Topic 37. Hur yttrar sig biverkningar av allvarlig karaktär
vid nyttjande av naturläkemedel?
How are side effects of serious nature manifested
in the use of herbal remedies?



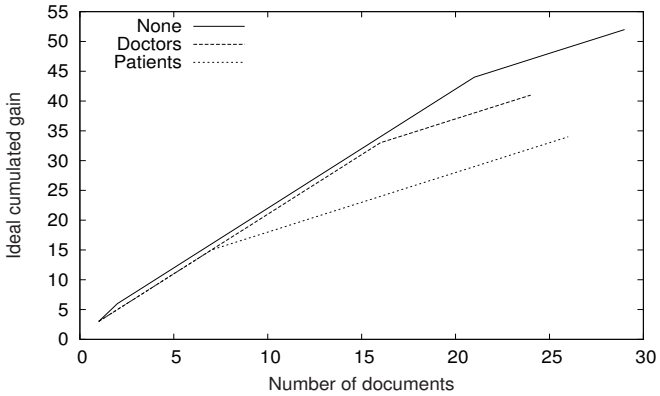
Topic 39. Vilken medicinsk behandling rekommenderas för en patient
som drabbats av Alzheimers sjukdom?
What medical treatment is recommended for a patient
who suffers from Alzheimer's disease?



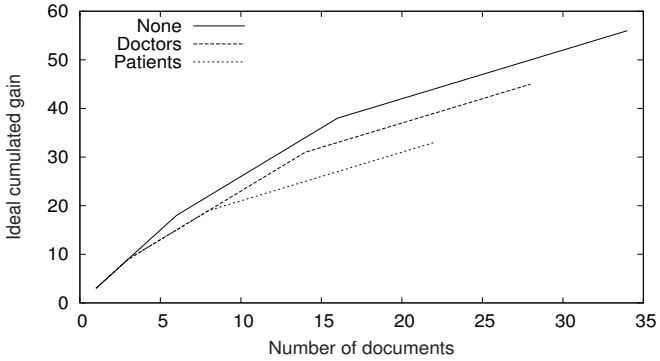
Topic 46. Vad innebär vaccinationsprogrammet och finns det risk att barn utvecklar allergi i samband med detta?
What is the vaccination program and is there a risk that children develop allergies in relation to this?



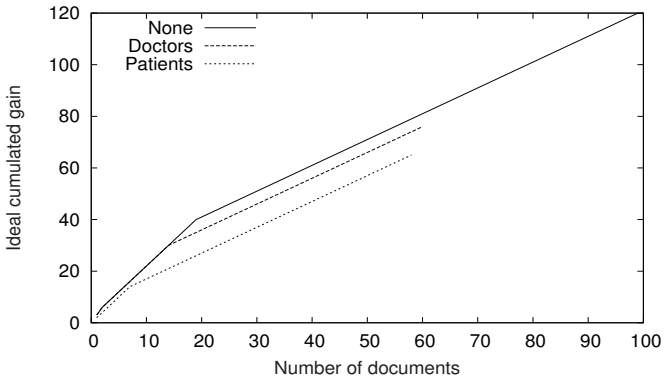
Topic 48. Vad är prognosen vid olika typer av cancer i matstrupen?
What is the prognosis of various types of cancer of the esophagus?



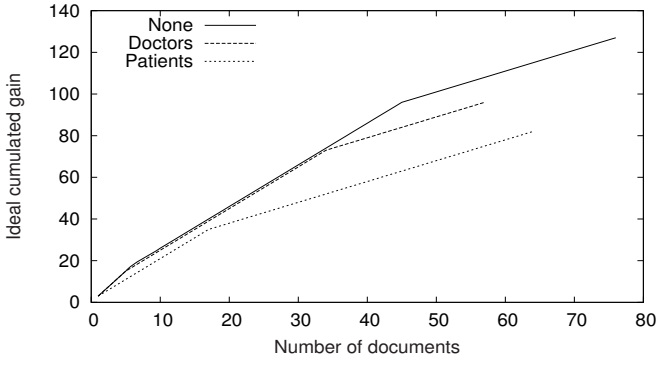
Topic 57. Vad innebär förgiftning med paracetamol och vilka symtom uppvisar patienten?
What is poisoning with paracetamol and what symptoms does the patient present?



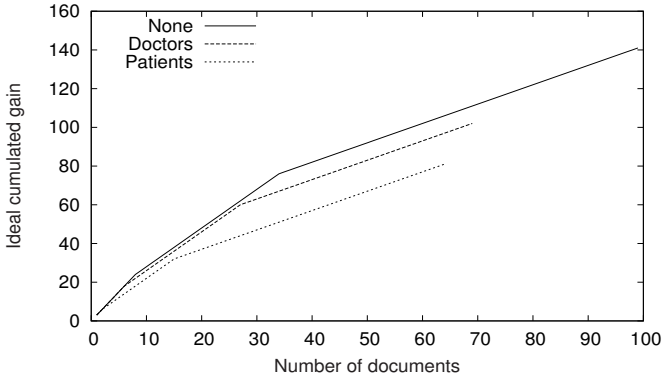
Topic 62. Hur bedömer man att en smärta kräver behandling med opioider?
How does one judge that a pain requires treatment with opioids?

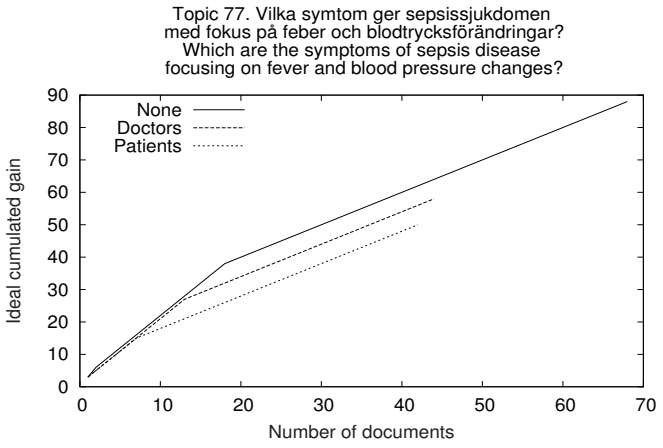


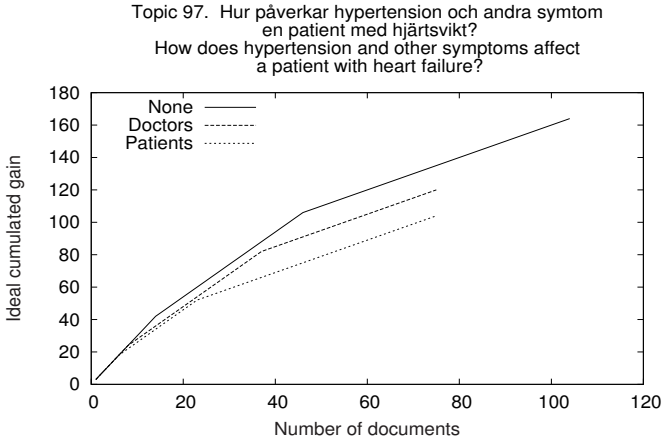
Topic 68. Vilka symtom associeras med DVT, djup ventrombos, och hur ser behandlingen ut?
Which are the symptoms associated with DVT, deep venous thrombosis, and what does the treatment look like?



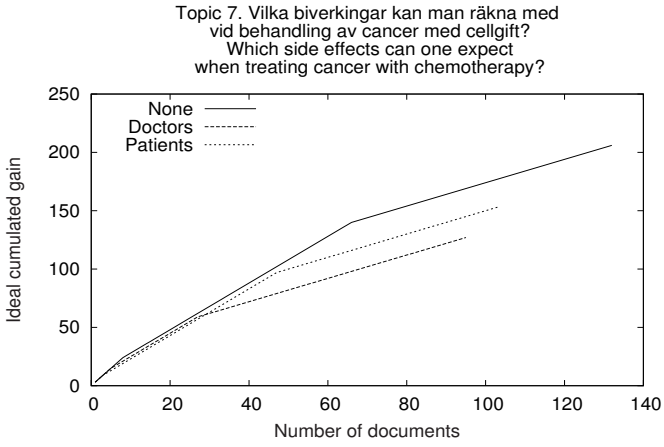
Topic 69. Vad finns det för differentialdiagnoser hos patient med bröstsmärta som symtom?
Which differential diagnoses exist in patients with chest pain symptoms?

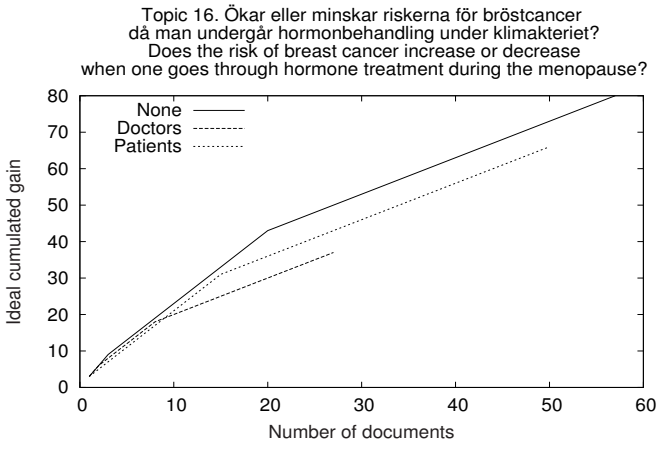



Topic 77. Vilka symtom ger sepsissjukdomen
med fokus på feber och blodtrycksförändringar?
Which are the symptoms of sepsis disease
focusing on fever and blood pressure changes?
None
Doctors
Patients
Ideal cumulated gain
Number of documents
0
10
20
30
40
50
60
70
80
90

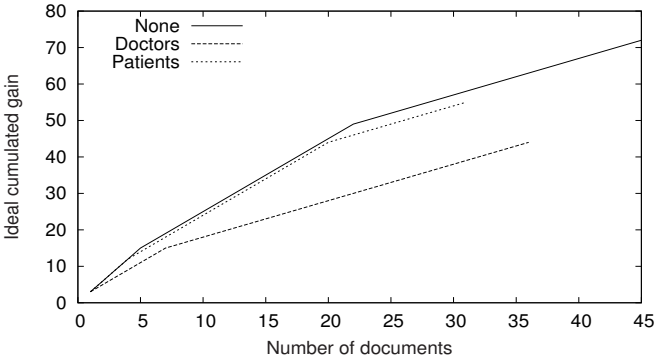

Topic 97. Hur påverkar hypertension och andra symtom
en patient med hjärtsvikt?
How does hypertension and other symptoms affect
a patient with heart failure?
None
Doctors
Patients
Ideal cumulated gain
Number of documents
0
20
40
60
80
100
120
140
160
180

B.3 Topics with predominantly patient documents

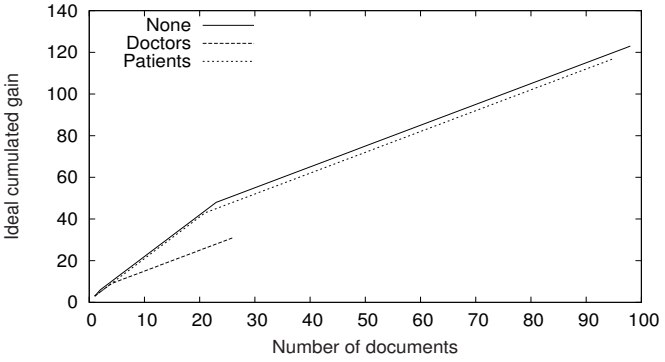




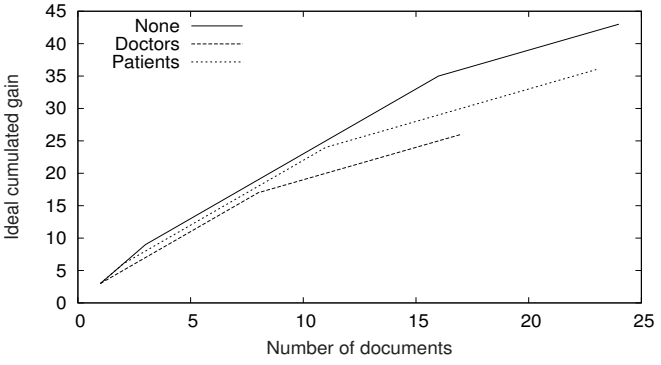
Topic 18. Ökar eller minskar riskerna för cervixcancer
då man undergår hormonbehandling under klimakteriet?
Does the risk of cervix cancer increase or decrease
when one goes through hormone treatment during the menopause?



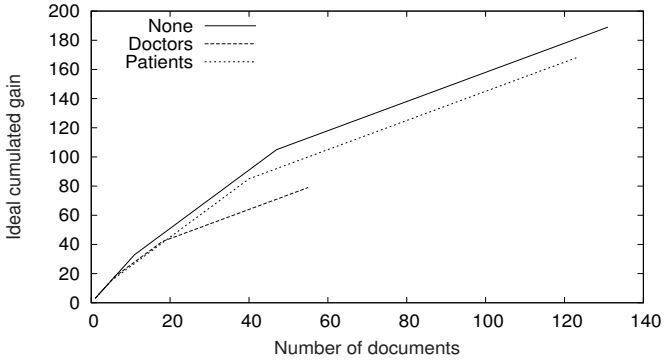
Topic 20. Vilka receptbelagda preparat finns att tillgå vid hudsjukdom,
exempelvis acne?
Which prescription drugs are available for skin disease,
for example acne?



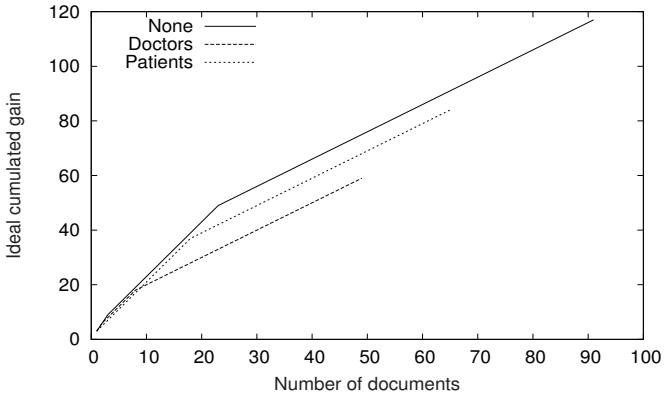
Topic 31. Vilka preparat av typen SSRI kan användas vid behandling
av depression hos gravida?
Which preparations of the type SSRI can be used in treatment
of depression in pregnant (women)?



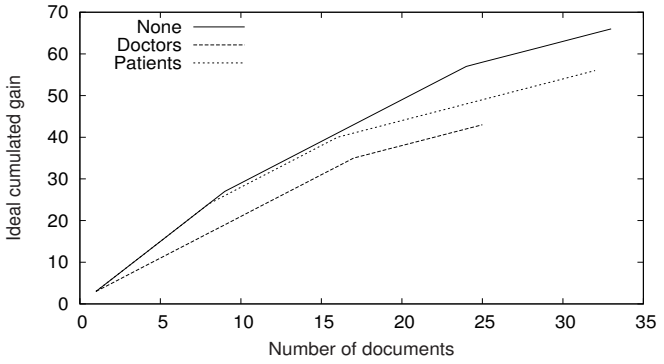
Topic 42. Vad finns för behandling i samband med herpes zoster och herpes simplex i form av antiviral terapi/virushämmare?
Which treatment exists for treatment of herpes zoster and herpes simplex in the form of antiviral therapy/virus inhibitors?



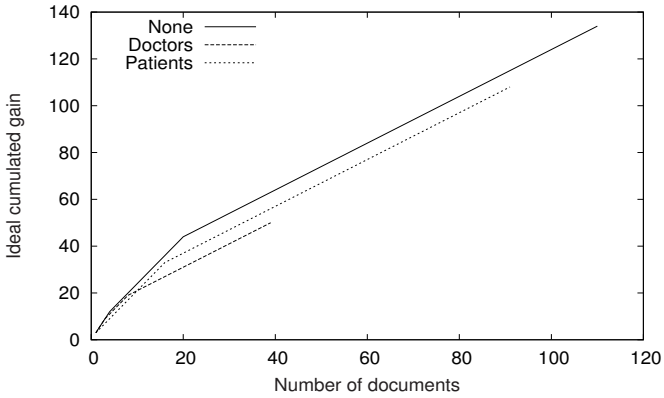
Topic 51. Varför kan en patient med cancer drabbas av anemi?
Why may a patient with cancer contract anemia?



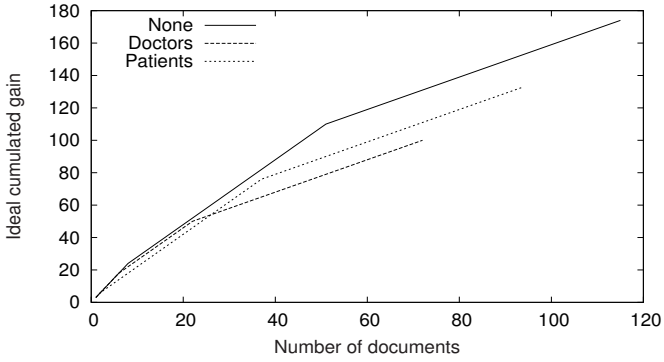
Topic 56. Varför utförs gastric bypass för viktreducering och hur tillämpas det?
Why is gastric bypass performed for weight reduction and how is it applied?



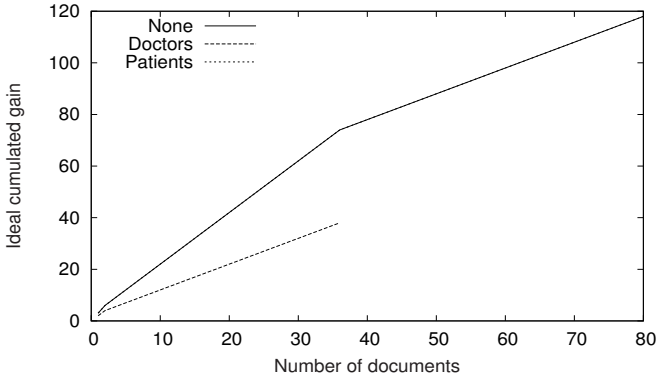
Topic 65. Hur används antihistaminer vid behandling av pollenallergi?
How are antihistamines used in treatment of pollen allergy?



Topic 82. Vilka är orsakerna bakom uppkomsten av kramp
och hur behandlas det?
Which are the reasons behind the occurrence of cramp
and how is it treated?



Topic 92. Hud: Hur går man tillväga vid behandling
av eksem med steroider?
Skin: How does one perform treatment of eczema with steroids?



So Long, and Thanks for All the Fish

The title of the fourth book of the Trilogy in Four Parts.