Automatic indexing: an approach using an index term corpus and combining linguistic and statistical methods

Timo Lahtinen

Academic dissertation to be publicly discussed, by due permission of the Faculty of Arts at the University of Helsinki in lecture room Unioninkatu 35, on the 11th of December, 2000, at 11 o'clock.

University of Helsinki
Department of General Linguistics
P.O. Box 4
FIN-00014 University of Helsinki
Finland

PUBLICATIONS
NO. 34
2000

ISBN 951-45-9639-0
ISBN 951-45-9640-4 (PDF)
ISSN 0355-7170
Helsinki 2000
Yliopistopaino

Abstract

This thesis discusses the problems and the methods of finding relevant information in large collections of documents. The contribution of this thesis to this problem is to develop better content analysis methods which can be used to describe document content with index terms. Index terms can be used as meta-information that describes documents, and that is used for seeking information. The main point of this thesis is to illustrate the process of developing an automatic indexer which analyses the content of documents by combining evidence from word frequencies and evidence from linguistic analysis provided by a syntactic parser. The indexer weights the expressions of a text according to their estimated importance for describing the content of a given document on the basis of the content analysis. The typical linguistic features of index terms were explored using a linguistically analysed text collection where the index terms are manually marked up. This text collection is referred to as **an index term corpus**. Specific features of the index terms provided the basis for a linguistic term-weighting scheme, which was then combined with a frequency-based term-weighting scheme. The use of an index term corpus like this as training material is a new method of developing an automatic indexer. The results of the experiments were promising.

Acknowledgements

Thank you

- Kimmo Koskenniemi, Fred Karlsson, and Lauri Carlson for guidance,
- Timo Järvinen, Pasi Tapanainen, Atro Voutilainen, Jussi Piitulainen, and Andrea Huseth for co-operation,
- friends, colleagues, and Rasti-Aspekti for back-up,
- my father Ville, my mother Sirkka, and my brother Vesa as well as other relatives for support,
- and my wife Tuuli, and our children Ilona, Henrikki, and Mikael for patience.

Contents

Part I

Introduction

This part will

- present the main contents and the structure of the thesis (*Chapter 1*)
- define some basic concepts of the thesis in short (*Chapter 1* and *Chapter 2*):
  - index term, index term corpus, automatic indexing, combining linguistic and statistical methods, and information retrieval (IR) (*Chapter 1*)
  - communication and information (*Section 2.2*)
  - relationship of information and index terms (*Section 2.3*)
- discuss briefly the contribution of language engineering to the challenge of the information age (*Section 2.1*)

Chapter 1

Overview

This overview will briefly describe the contents and the structure of the thesis, as well as some essential concepts.

The title of the thesis is **Automatic indexing: an approach using an index term corpus and combining linguistic and statistical methods**. Here is a short commentary on the title:

- **Index term** is an expression that describes the contents of a text and guides a user to the information.
- **Index term corpus** is a linguistically analysed text collection where the index terms are manually marked up. It is the training and test material of the new automatic indexing method of this thesis.
- **Automatic indexing** is the process of producing the descriptors (index terms) of a text automatically.
- **"Combining linguistic and statistical methods"** means that the automatic indexing method of this thesis combines the use of a syntactic parser with the detection of word frequencies.

One more important concept (not included in the title) is **information retrieval (IR)** which may be defined as the selective, systematic recall of logically stored information (Cleveland and Cleveland, 1983, p.33). Another important concept that is not included in the title is a new concept introduced in this thesis: **index-term-structure**, which is identified with 'weighted index terms in their context'. It can be seen as a new content analysis framework for information retrieval (cf. *Section 7.3*).

1.1 Research questions

The following research questions summarize the main points of the thesis.

1. Is there any point in using linguistic methods in automatic indexing?

   Automatic indexing relies typically on word frequencies. If the word occurs frequently in a given document, but does not occur in many other documents, it is possibly an appropriate document descriptor, and it should be weighted high by the indexer.

   Some linguistic methods, however, have been used as well. The weighting scheme developed in this thesis combines evidence from word frequencies and evidence from linguistic analysis provided by a syntactic parser. The results suggest that linguistic methods could be useful in automatic indexing.

2. Could linguistic methods offer any advantage over purely statistical indexing methods?

   The performance of the linguistic methods developed here is compared with the performance of purely statistical indexing methods. The indexing procedures are usually evaluated by the recall and precision rates[1] of retrieved documents, whereas in this thesis the automatic indexer is evaluated by the recall and precision rates of retrieved index terms using the test corpus where the index terms are manually marked up, as a benchmark. The results suggest that linguistic methods could offer some advantage over purely statistical indexing methods. The methods introduced in this thesis may help to improve precision without reducing recall.

3. How can we use linguistic methods in automatic indexing?

   One essential assumption of this thesis is that the parser provides useful hints for weighting index terms. Appropriate index terms are typically nouns or noun phrases, and the part-of-speech tagging distinguishes nouns from verbs and other parts of speech. The parser is also capable of recognizing proper nouns, which are typically appropriate index terms as well.

   The results of this thesis suggest that index terms have certain typical morphological, syntactical, and lexical features that provide useful information for weighting index terms.

   Another important advantage of using a parser in automatic indexing is that the parser can recognize noun phrases, which is the basis for recognizing appropriate multi-word index terms.

4. How can we combine linguistic and statistical methods in automatic indexing?

   Chapter 19 will introduce a new weighting scheme that combines linguistic and statistical methods in automatic indexing. *Section 1.3* will describe the weighting scheme briefly. The new weighting scheme can be seen as a kind of data fusion technique (cf. *Chapter 11*).

5. Is it possible to recognize **subtopics** by recognizing words that appear in the discourse at a certain point of the document, occur frequently for a while, and then disappear (that is, **bursty words**)?

[1] *Section 9.2* will define notions of **recall** and **precision**.

Section 18.1 will introduce a new method for recognizing bursty words. The results suggest that this method does not distinguish between terms and non-terms particularly well, but it does distinguish between subtopics and main topics with some accuracy. **Terms** are words marked up as terms in the index term corpus and **non-terms** are words not marked up as terms. **Main topics** are the central themes of the text and **subtopics** are the less central themes. Hearst's framework (cf. *Chapter 12*) characterizes text structure as a sequence of subtopical discussions that occur in the context of one or more main topic discussions (Hearst, 1997).

1.2 Materials and methods

Figure 1.1 presents a general picture of the materials and methods of the thesis. The issue will be discussed in *Part IV* in more detail.

The first steps of *Figure 1.1* describe how the material of the thesis was produced. The empirical study of this thesis is based on an index term corpus, which is a collection of texts where some information concerning index terms was encoded, both manually and automatically. The core of the index term corpus in this study consisted of five texts that were concerned with sociology and philosophy. All texts had manually generated back-of-the-book indexes.

The research aide identified and marked up the index terms for each text page using previously manually generated book indexes, that is, she marked up the closest equivalents of index terms found in the book indexes. After that, the linguistic analysis of the index term corpus was automatically provided by a dependency parser (FDG), and the textual location of words was analysed and marked up automatically, too. The textual location was encoded by tags that indicate if the word is in a title or subtitle, or in the first paragraph after or before a title or a subtitle, or in the first or last sentence of the paragraph. This encoding was done because of the assumption that some locations can have a special role in index term weighting (cf. *Section 7.2*).

The corpus was then divided into two parts: **a training corpus** and **a test corpus**. The features of index terms were explored using the training corpus, which is then the basis for the automatic indexer. The test corpus was used to test whether the results could be generalized beyond the context of the training corpus. The explored features of index terms included lexical, morphological and syntactical features, encoded by tags, as well as information about the location and the distribution of words (frequencies).

When we have all this information in the same corpus, it is possible to determine the set of single-word and multi-word index term patterns, and to calculate estimated index-term-likeness probabilities of a kind to these patterns. When we have calculated these probabilities, we can use them with any new text to estimate the index-term-likeness of the words and phrases of the text, that is, we can index texts automatically. The next section will describe the patterns and their estimated index-term-likeness probabilities in more detail.

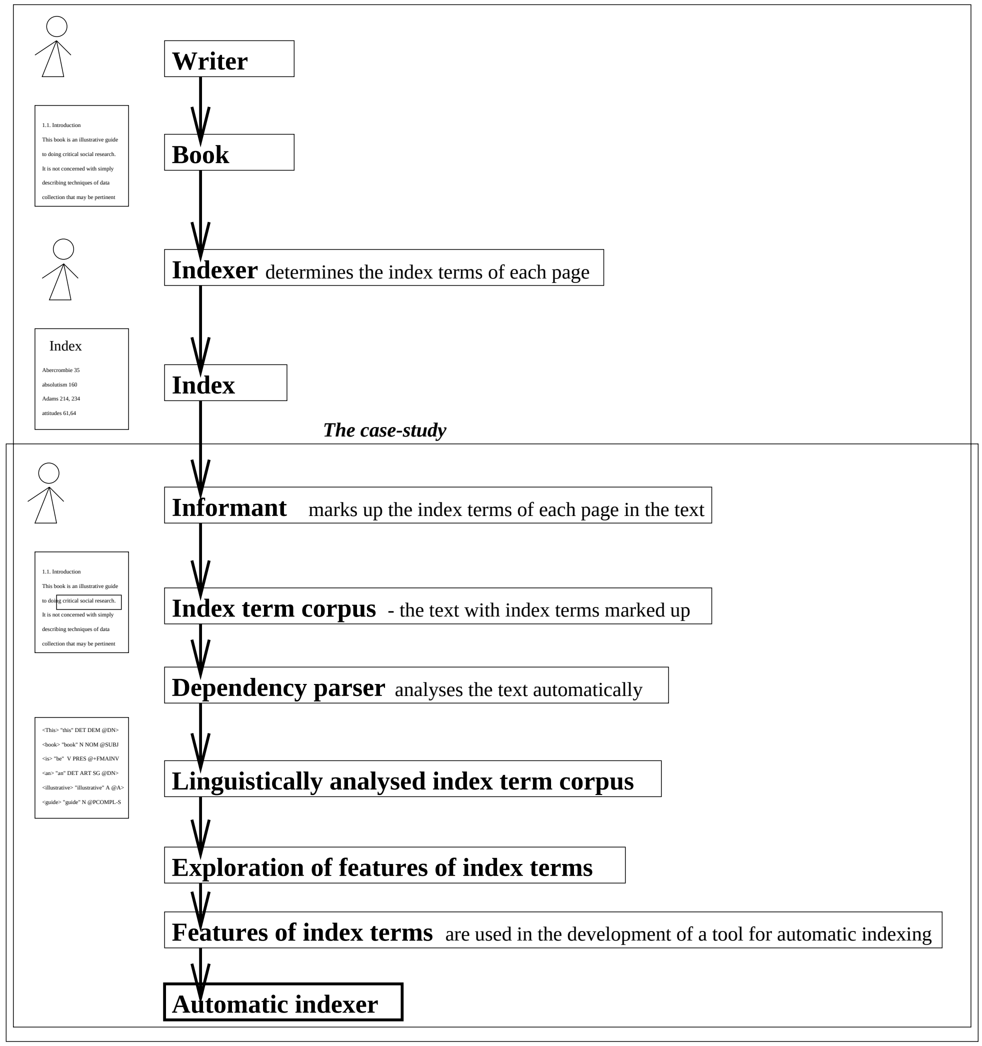

Figure 1.1: The course of the case-study

1.3 Weighting schemes used in the thesis

The thesis introduces three new weighting schemes:

- *TW* (**tag weights**), a weighting scheme based on linguistic analysis,
- *STW*IDF*, a weighting scheme that combines *TW* and the widely used *TF*IDF* weighting scheme, and
- a weighting scheme based on the **within-document burstiness**.

These weighting schemes are attempts to develop better content analysis methods for automatic indexing. *Section 7.2* will discuss some relevant issues concerning content analysis: lexical cohesion, anaphora resolution, and discourse analysis frameworks, among others.

Automatic indexing typically relies on shallow detection of **lexical cohesion**. If certain words occur in certain documents more frequently than in others, it may indicate that these words are topic words in those documents. This kind of lexical cohesion is related to **burstiness** discussed below. Different techniques have been developed in order to recognize also other cohesive ties than those of plain repetition, but the weighting schemes of this thesis rely on plain repetition.

Several **frameworks for discourse analysis** have been proposed, but in this thesis no such framework is applied. A robust discourse analyser that could reliably and automatically **resolve anaphora** and define the thematic structure of a text could contribute a great deal to automatic indexing, but unfortunately, no such analysis method is available. The weighting schemes of this thesis do not attempt to resolve anaphora in order to weight the index terms.

The weighting schemes described below are based on linguistic analysis provided by a parser and detection of distribution and location of words.

TW, a weighting scheme based on linguistic analysis

Tag weights (*TW*, cf. *Chapter 17*) combine all evidence provided by tag lists, that is, *TW* combines the linguistic evidence (tags provided by the parser and the location tags). *TW* weighting scheme was trained by using the **index term corpus** (*Chapter 6* and *Chapter 15*). In the index term corpus manually generated index terms were marked up by tags and their linguistic features were explored. On this basis, the set of single-word and multi-word index term patterns (*TW* patterns) was determined. Moreover, for each pattern an estimated index term probability was calculated by using the index term corpus as a training corpus. These index term probabilities are the weights of the *TW* weighting scheme.

The index term probabilities were obtained automatically by the following steps:

1. Count the number of all occurrences of a given pattern in running text (n_p). For instance, if a simple pattern "a noun with -ism ending" (tag combination `N and <DER:ism>`)[2] occurs 792 times in the training corpus, then $n_p = 792$.

2. Count the number of occurrences of this pattern that are marked up as index terms (n_i). If the pattern `N and <DER:ism>` occurs 453 times in the training corpus as an index term, then $n_i = 453$.

3. Divide the number of index term occurrences by the number of all occurrences (n_i/n_p). The index term probability of the pattern `N and <DER:ism>` is then $n_i/n_p = 453/792 = 0.572$. Thus, for example, the word `Marxism` has an index term probability of 0.572.

To sum up, *TW* weighting scheme weights index terms by using the index term probabilities of the patterns calculated from the training corpus.

[2]This is a simplified example. See real examples of patterns and their index term probabilities (i.e., weights) in *Section 21.1*.

STW*IDF, a weighting scheme based on linguistic analysis and word frequencies

Tag weights (*TW*) combine evidence provided by the parser and the location tags, but *TW* does not use evidence from **burstiness**. The notion of burstiness (cf. *Chapter 12*) characterizes two related phenomena (Katz, 1996):

- **Document-level burstiness** refers to multiple occurrence of a content word or phrase in a single document, which is contrasted with the fact that most other documents contain no instances of this word or phrase at all.

- **Within-document burstiness** (or burstiness proper) refers to close proximity of all or some individual instances of a content word or phrase within a document exhibiting multiple occurrence.

The phenomenon of burstiness is the underlying basis for most frequency-based indexing techniques. *STW*IDF* weighting scheme, as well as the widely used *TF*IDF* weighting scheme (cf. *Section 13.3*), uses evidence from document-level burstiness, and the third new weighting scheme (described below) uses evidence from within-document burstiness. Inverse document frequency (*IDF*) is based on the observation that words that are found in a fewer number of documents are often appropriate index terms. In the *TF*IDF* weighting scheme, *IDF* is multiplied by a number of occurrences of a given word or phrase in a document (*TF*). Thus, if a word occurs frequently in a given document (*TF*), but does not occur in many documents (*IDF*), it is weighted high by *TF*IDF*; such word is a typical bursty word.

*STW*IDF* (cf. *Chapter 19*) is a modified version of the standard *TF*IDF* weighting scheme and it is based on a well-known variant of the standard *TF*IDF* weighting scheme. Robertson and Sparck Jones refer to this variant as Combined Weight *CW* (Robertson and Sparck Jones, 1997). The main difference to the basic *TF*IDF*-formula is that *CW* takes into account the document length as well. *CW* also uses the so-called tuning constants which modify the extent of the influence of term frequency and the effect of document length. The values of tuning constants used in this thesis are those used by Robertson and Sparck Jones (Robertson and Sparck Jones, 1997). In this thesis, *CW* is referred to as *TF*IDF* and it is used to weight multi-word index terms as well as single-word index terms.

In *STW*IDF* weighting scheme *TF* is replaced by *STW*, which is the sum of the *TW* values of all occurrences of the term candidate in the test corpus (**summed tag weights** *STW*). If, for example, the frequency of the proper noun `Marx` and the frequency of the verb `suggest` is the same in the document, they have the same *TF* values. However, if the *TW* value of `Marx` is higher than the *TW* value of `suggest`, then the *STW* value of `Marx` is higher than the *STW* value of `suggest` as well. Thus, in *STW*IDF* weighting scheme *STW* gives extra weight to `Marx` compared with `suggest`, whereas in *TF*IDF* weighting scheme *TF* treats the words equally. In this way *STW*IDF* combines evidence based on linguistic annotation with evidence based on burstiness. Multi-word terms are weighted in the same way than single-word terms.

A weighting scheme based on the within-document burstiness

As mentioned above, **within-document burstiness** refers to close proximity of individual instances of a content word or phrase within a document. The purpose of the new weighting scheme based on the within-document burstiness (cf. *Section 18.1*) is to find words that appear in the discourse at a certain point of the document, occur frequently for a while, and then disappear. In other words, the purpose is to recognize **subtopics** by recognizing within-document bursty words, since subtopics could be assumed to be words that appear in the discourse at a certain point of the document, occur frequently for a while, and then disappear.

The new algorithm distinguishes between bursty words and words used throughout the text by counting the distances of the occurrences of individual words using paragraphs as units for measuring the distance. In this implementation, paragraphs were used as units, since paragraphs can be considered as topical units of discourse (cf. *Section 7.2*).

The within-document burstiness of different words is detected by determining the curves of the distribution functions of the words, and by computing areas above the curves of the words. This makes it possible to compare the within-document burstiness of words by using single values computed to each word. In the experiment of this thesis the values were computed only to single words, not to phrases, since at the moment the method does not include any mechanism for recognizing phrases. The results suggest that this method does not distinguish between terms and non-terms particularly well, but it does distinguish between subtopics and main topics with some accuracy.

1.4 Some main points of the thesis

The thesis is about

- **communicating information**. *Chapter 2* will briefly discuss some basic concepts related to communication of information.

- **communicating information by index terms**. *Part II* will describe the indexing task and *Part III* will discuss the use of index terms in information seeking process.

- **communicating information by index terms more effectively**. The purpose of the thesis is to improve the information seeking process by more precise content analysis of documents. The empirical part of the thesis (*Parts IV and V*) will introduce a new automatic indexing method that combines linguistic and statistical methods.

The topic of this thesis is the problem of finding the relevant information in large collections of documents. The main points of the thesis can be summarized as follows:

1. **The main problem: How to find the information that is needed?**
    - By discovering and describing ("understanding") the content of documents automatically.
2. **How to discover and describe the content?**
    - By an automatic and exhaustive content analysis that produces appropriate document descriptors (index terms) which are weighted according to their estimated importance for describing the content of a given document.
3. **How to determine effective document descriptors and their weights?**
    - By an automatic linguistic analysis of documents, including part-of-speech tagging, lexical and syntactic analysis, and analysis of location and distribution of words (frequencies).
4. **The main result of the thesis:**
    - An automatic indexer that extracts single-word and multi-word index terms and weights them according to their importance for describing the content of documents.

The following section will discuss the above presented points in more detail. Furthermore, it will reveal how the different sections of the thesis are connected to these points.

The main problem: How to find the information that is needed?

What is information and what is communication? *Chapter 2* (Language and information) will briefly discuss different definitions of these and other related concepts. *Section 2.1* (Language engineering and the information age) will briefly discuss what is the contribution of language engineering to the challenge of the information age. The various document collections contain a lot of information; how is the relevant information found? A more specific answer is sketched in the *Part III* (Index terms and information seeking), which discusses some theoretical and practical points related to information seeking - especially to information retrieval (IR):

- What are information retrieval systems? (*Chapter 9*)
- What are information seeking strategies? (*Chapter 10*)
- What are the techniques of information retrieval? (*Chapter 11 and 13*)

The empirical part of the thesis (*Part IV* and *Part V*) will focus on one specific, albeit important, subfield of information seeking: one way to improve the access to relevant information is to develop automatic techniques that are capable of discovering and describing the content of documents appropriately.

How to discover and describe the content?

Chapter 5 will present different information description languages. This thesis will focus on **index terms** as a description language of documents. Index terms are meta-information that describe documents and that are used for seeking information. The index terms of book indexes indicate to users ‘what is being written about and on what page’ and index terms of information retrieval systems are words/phrases that are weighted according to their importance for describing the content of a given document (*Part II*). *Section 4.2* will briefly discuss some principles of manual indexing as well although the main focus of this thesis is on automatic indexing. Automatic indexing produces lists of weighted index terms (*Chapter 13*).

The empirical part of the thesis (*Part IV* and *Part V*) will describe a technique of an automatic and exhaustive content analysis that produces weighted index terms that represent the content of documents.

How to determine effective document descriptors and their weights?

Automatic indexing has typically relied on word frequencies, but some natural language processing techniques have been used as well (*Chapter 11*). The weighting schemes used in information retrieval will be discussed in *Chapter 13*. Distribution of words provides useful evidence for weighting index terms; the **burstiness** of a given word often indicates a topical use of the word, that is, if the word occurs frequently in a given document, but does not occur in many other documents, it is possibly an appropriate document descriptor and it should be weighted high (*Chapter 12*).

The weighting scheme developed in this thesis (*STW*IDF*) combines evidence from burstiness and evidence from linguistic analysis provided by a syntactic parser (*Part IV*). The results suggest that appropriate document descriptors and their weights can be determined by an automatic content analysis of documents, including part-of-speech tagging, lexical and syntactic analysis, and analysis of location and burstiness of words (*Part V*).

The main result of the thesis

The main result of the thesis is an automatic indexer that extracts single-word and multi-word index terms and weights them according to their importance for describing the content of documents. The developed weighting scheme of the indexer (*STW*IDF*) combines evidence from burstiness and evidence from linguistic analysis and in the experiments of this thesis it outperformed weighting schemes based either on burstiness only or on linguistic analysis only.

The main point of this thesis is to illustrate the process of developing an automatic indexer (*Part IV* and *Part V*), but some theoretical background is given as well (*Parts I-III*). As Blair writes (Blair, 1990, p.122): The process of representing documents for retrieval is fundamentally a linguistic process, and the problem of describing documents for retrieval is, first and foremost, a problem of how language is used. Thus any theory of indexing or document representation presupposes a theory of language and meaning. Thus the focus of the theoretical discussion of this thesis is on the linguistic aspects considered as relevant to information retrieval.

So far the impact of linguistic tools within the field of information retrieval has been relatively modest. In recent years, however, more advanced linguistic techniques have been developed and several attempts have been made in order to improve retrieval performance of information retrieval systems by using these techniques. The successful application of linguistic techniques requires that linguistic tools are used for tasks in which they are best suited. In this thesis, the usefulness of a syntactic parser for the indexing task is considered.

1.5 Structure of the thesis

Parts I-III will present the theoretical basis of the research and give a brief overview of some techniques used in information retrieval. The essential concepts of this thesis will be discussed and defined. A number of different theoretical frameworks will be presented, as well as some new theoretical considerations of my own. The main purpose is to determine an appropriate theoretical framework to the empirical part of the thesis, but a kind of overall picture will be given as well.

Part IV Materials and methods will describe the process of creating the index term corpus and the methods that were used in order to explore the features of index terms.

Part V Results will present the explored features of index terms and evaluation of different indexing methods.

Part VI Discussion will interpret the results and consider their significance. It will also list the implications of this research and identify areas for further research.

Chapter 2

Language and information

Language is, among other things, a means of communicating information and index terms are units of language used as tools for communicating information. This is the approach of this study to language, information, and index terms. Information retrieval is a sub-discipline of information science which in a broad sense is concerned with information, knowledge, and understanding, i.e. essentially with *meaning* as perceived by a receiving mind and embedded in written records (Kochen, 1983). Ingwersen mentions the following four important sub-disciplines of information science (Ingwersen, 1992, p.12):

- *Informetrics*, i.e. the quantitative study of communication of information, such as co-citation.
- *Information management*, including evaluation and quality of textual and other media-based IR systems.
- *Information (retrieval) systems design*
- *Information retrieval interaction*

Figure 2.1 presents other disciplines providing valuable contributions to information science, such as computer science, psychology, sociology, and linguistics (Ingwersen, 1992, p.8). As the picture indicates, Ingwersen emphasizes the cognitive nature of information science and information retrieval. This thesis, however, does not focus on the cognitive aspects of information seeking process, but on the linguistic aspects. This chapter will briefly discuss some basic concepts related to communication of information.

2.1 Language engineering and the information age

The current age is often referred to as the information age. The vast amount of available information creates new opportunities, as well as new challenges. As more and more information becomes

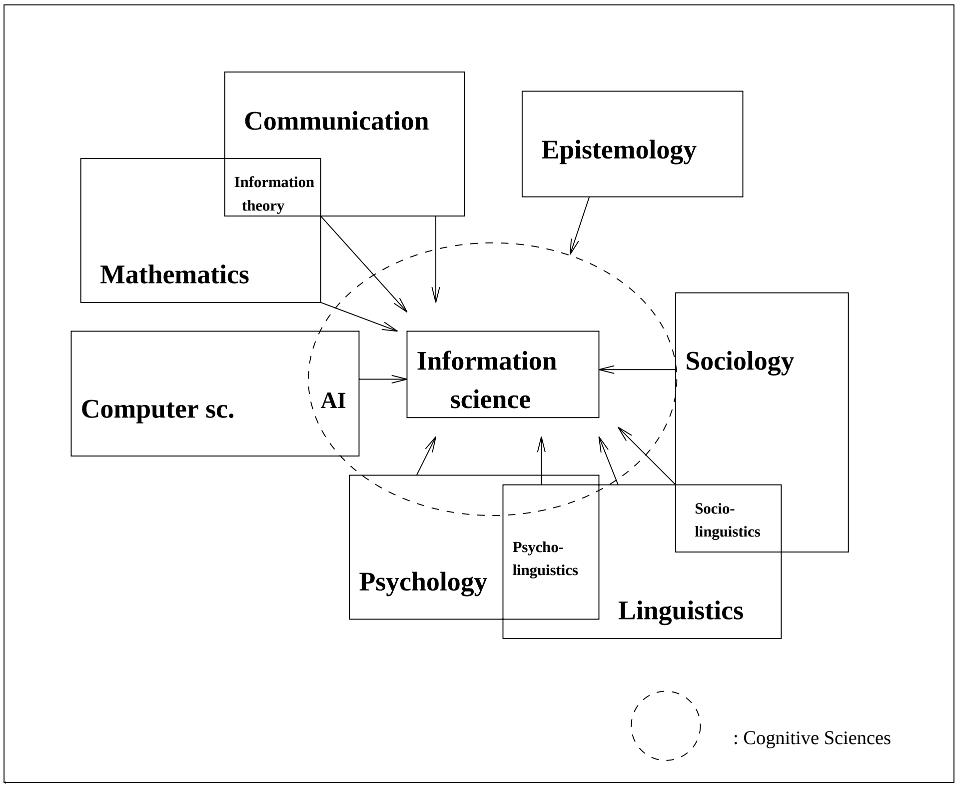

Figure 2.1: Scientific disciplines influencing information science (Ingwersen, 1992, p.8).

available from a wide range of sources the human recipients may find it increasingly difficult to select and assimilate what is useful: Language engineering software, embedded in information servers and in the search engines and 'intelligent agents' which are used to search them, provides the facilities to overcome these problems. The techniques developed within language engineering allow the analysis of the content of information sources, either in a quick 'shallow' sense, looking for information of potential interest on which to focus, or, within a specific subject area, to perform a complete analysis identifying specific information. In addition, the selected information can then be summarised for presentation to the user who can later decide to request the full information. This is clearly a very effective method of overcoming the problem of information overload. (Language engineering. Progress and prospects, 1997, p.32)

Figure 2.2 presents a general picture of activities which are involved in language engineering, from research to the delivery of products to end-users (Harnessing the power of language, pp.11-12).

As the picture shows, research leads to the development of techniques, the production of resources, and the development of standards. In practice, language engineering is applied at two levels, of which the first level includes a number of generic classes of application, such as:

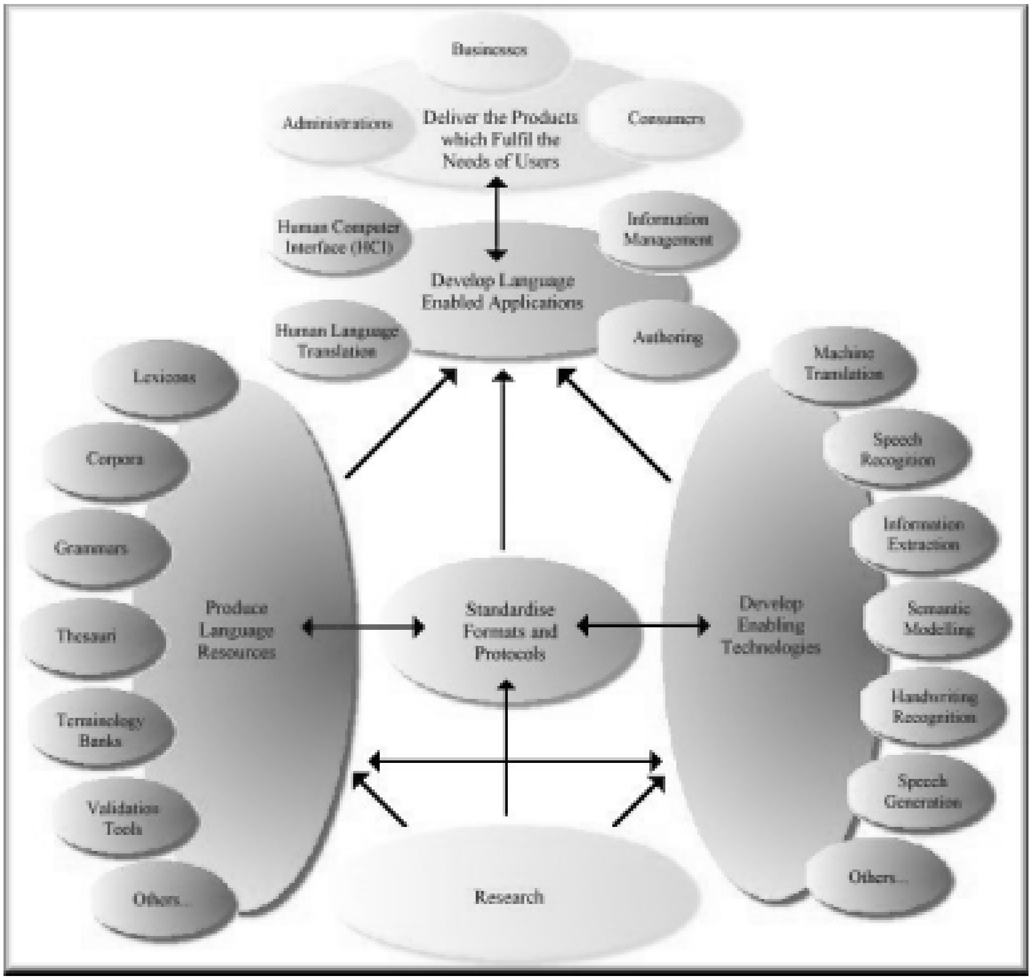

Figure 2.2: Model of language engineering activities (Harnessing the power of language, pp.11-12).

- language translation,
- information management (multi-lingual),
- authoring (multi-lingual), and
- human/machine interface (multi-lingual voice and text)

At the second level, these applications are applied to real world problems, for example:

- information management can be used in an information service, as the basis for analysing requests for information and matching the request, against a database of text or images, to select the information accurately
- authoring tools are typically used in word processing systems but can also be used to gener-

ate text, such as business letters in foreign languages, as well as in conjunction with information management, to provide document management facilities

- human language translation is currently used to provide translator workbenches and automatic translation in limited domains
- most applications can usefully be provided with natural language user interfaces, including speech, to improve their usability.

The purpose of this thesis is to contribute especially to the development of information management applications. Indexing from the point of view of information management applications will be discussed in more detail in *Part III*.

2.2 Communication of information

2.2.1 Concepts of the communication process

This thesis approaches language as a means of communicating factual information. The following section will briefly present some concepts related to the communication process. Foskett has found the following definitions in the Concise Oxford dictionary (1976) and the Macquarie Dictionary (1981) (Foskett, 1996, p.3):

- *knowledge*, is what *I* know
- *information* is what *we* know, i.e. *shared* knowledge
- *communication* is the imparting or interchange of ... information by speech, writing or signs, i.e. the *transfer* of information
- *data* [literally things given] any fact(s) assumed to be matter of direct observation.
- Additionally, a *document* is any physical form of recorded information

Collins COBUILD English Language Dictionary (1987), on the other hand, gives the following definitions:

- the *content* of a piece of writing, speech, television programme, etc is its subject matter and the ideas that are in it, in contrast to things such as its form and style
- the *meaning* of a word, expression, or gesture is the thing or idea that it refers to or represents and which can be explained by other words ... the *meaning* of what someone says or of a book, film, etc is the thoughts or ideas that are intended to be expressed by it.

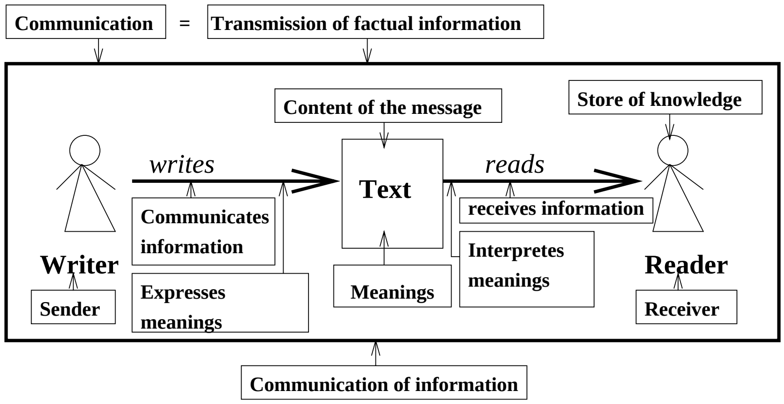

Figure 2.3: Communication process - the different concepts.

In linguistics, **meaning** is studied above all in semantics, but meaning is an important concept for text linguistics as well. According to Brown and Yule (Brown and Yule, 1983, p.26), the discourse analyst treats his data as the record (text) of a dynamic process in which language was used as an instrument of communication in a context by a speaker/writer to express meanings and achieve intentions (discourse). According to Lyons, the term **communication** can be defined, in a somewhat restricted way, as an intentional transmission of factual information: **communicative** means meaningful for sender, and **informative** means meaningful for receiver; receiver's store of factual **knowledge** is augmented in the communication process (Lyons, 1977, pp.32-39). Dretske emphasizes that a genuine theory of information would be a theory about the **content** of our messages, about the information we communicate (Dretske, 1981, p.40). *Figure 2.3* illustrates the overlap between the above mentioned concepts.

As mentioned above, the thesis approaches language as a means of communicating factual information. The thesis focuses on the linguistic features, which can be observed automatically, such as distribution of words, morpho-syntactic features, and endings of words. Thus, the theory of meaning and the cognitive aspects of communication will not be discussed here.

2.2.2 Different approaches to information

Thagard has found at least three different notions of information in the literatures of computer science, cognitive psychology, and philosophy (Thagard, 1990, pp.168-169):

- Information-processing approach,
- Ecological approach, and
- Mathematical approach

According to Thagard (Thagard, 1990, p.169), **the information-processing approach** to the notion of information is a typical approach of cognitive psychology, in which the notion of information is sometimes simply identified with the notion of knowledge. Information-processing psychology treats information primarily as a matter of mental representation, as computational structures in the minds of thinkers.

The ecological approach to the notion of information, on the other hand, emphasizes the presence of information in the world; information is seen as a property of facts or situations (Thagard, 1990, p.169).

The mathematical (or communication-theoretic or information-theoretic) notion of information was developed by Shannon (Shannon, 1949), and there the word 'information' is used in a special sense which differs from its ordinary, non-technical, everyday use. Weaver emphasizes that in particular, information in this sense must not be confused with meaning (Weaver, 1949, p.99). Shannon remarks that meaning and the semantic aspects of communication are irrelevant to the engineering problem (Shannon, 1949, p.3). The engineering problem is to maximize the efficiency of signal transmission, and information is a property of signal, in particular. The approach to information is statistical: the less probable signal, the more informative, as indicated by the formula:

$$I(s) = log_2 \frac{1}{p(s)}$$

The information (*I*) carried by a signal (*s*) is the logarithm of the reciprocal of the probability (*p*) of signal. Information is measured by using binary digits, bits, as units. The theory is based on the notion of **entropy**, borrowed from thermodynamics: if a given situation is highly organized, it is not characterized by a large degree of randomness or of a choice - that is to say, the information (or the entropy) is low (Weaver, 1949, p.103).

Lyons draws a terminological distinction between **signal information** and **semantic information**, even though they interact in a complex manner. There is, for instance, a link between these two senses of information with respect to the notion of surprise value, i.e., the principle of the proportion of signal-information: the greater a signal's probability of occurrence, the less signal-information it contains. 'Man bites dog' is in some sense a more significant item of news than 'Dog bites man'. When a signal has a probability of 1 and is thus totally predictable, it carries no signal-information. If somebody says something totally predictable, the utterance, in some sense, contains no semantic information. According to Lyons, the interaction of signal-information and semantic information must be taken into account in any theoretical model of the production and reception of speech. (Lyons 1977, pp.41-46)

However, as far as the distribution of index terms is concerned, it would be misleading to say that index terms are more informative if their entropy is high. On the contrary, an index term that occurs frequently in a limited passage of a text and then disappears from the discourse (i.e., the index term has low entropy) is a potential topic of that passage. Thus, it carries a lot of information about the content of the text, which makes it an informative index term.

Shannon introduced the classic model of communication, presented in *Figure 2.4* (Shannon, 1949, p.5). Shannon was an engineer at Bell Telephone, and the following interpretation of the model uses a telephone conversation as an example, even though the purpose of the model is to be

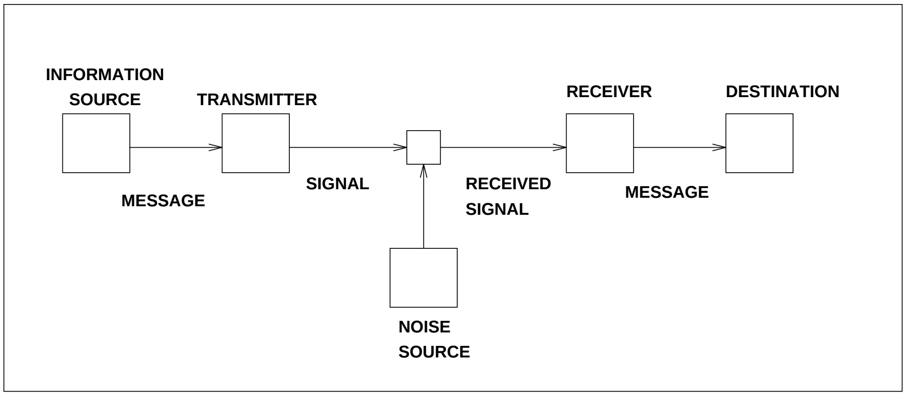

Figure 2.4: Schematic diagram of a general communication system by Shannon and Weaver (Shannon, 1949, p.5).

a general description of the communication process. The information source is a person speaking to a telephone, which is the transmitter that converts the speech (message) into an electric current (signal). The channel (the unlabelled box in the middle) is the medium (for instance a cable) that transmits the signal. Another telephone is the receiver and another speaker is the destination. The noise source is any additional stimuli that disrupts the conversation, for instance, a heavy traffic beside a telephone box. Lyons remarks that a certain degree of redundancy is essential in language in order to counteract the disturbing noise: by the means of redundancy, the receiver is able to recover the information lost caused by noise (Lyons 1977, pp.44-45). In this respect, Shannon's notion of noise has some linguistic importance as well.

	Intangible	*Tangible*
Entity	Information-as-knowledge: knowledge	Information-as-thing: data, document, recorded knowledge
Process	Information-as-process: becoming informed	Information processing: data processing, document processing, knowledge engineering

Figure 2.5: Buckland's matrix of different kinds of information (Buckland, 1991, p.6).

Figure 2.5 presents Buckland's matrix of different kinds of information (Buckland, 1991, p.6). This picture distinguishes between

1. Information as intangible entity: personal knowledge (private, mental, Popper's World 2 (Popper, 1972)). Brier calls this phenomenological knowledge (Brier, 1996, p.303).

2. Information as intangible process of knowing or becoming informed. Brier calls this cognition.

3. Information as tangible entity: objective/intersubjective materially registered knowledge (documents, part of Popper's World 3).

4. Information as tangible process: information/data processing, the mechanical manipulation of signals and symbols.

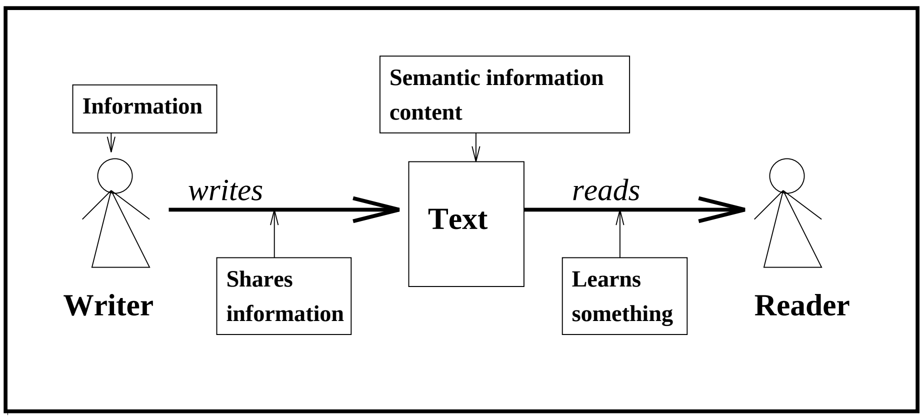

Figure 2.6: Everyday use of the word 'information'.

In this thesis, the focus is on the tangible aspects of information, and on describing the content of a document by means of index terms, in particular. The above described mathematical and cognitive aspects of communication are outside the scope of the study. The approach to information is based mainly on the **everyday use** of the word information: a writer has some information that is shared by a text. This information may originate from the world or from the writer's cognitive processes. A reader reads the text that has a certain semantic information content and learns something (*Figure 2.6*). Dretske refers to this everyday sense of the term 'information' as the *nuclear* sense (Dretske, 1981, p.45): A state of affairs contains information about X to just that extent to which is suitably placed observer could learn something about X by consulting it. This, I suggest, is the very same sense in which we speak of books, newspapers, and authorities as containing, or having, information about a particular topic, and I shall refer to it as the *nuclear* sense of the term "information". In this sense of the term, *false* information and *mis*-information are not kinds of information - any more than decoy ducks and rubber ducks are kind of ducks.

2.3 Information and index terms

Ingwersen gives the following definition to information retrieval (Ingwersen, 1992, p.228): The process involved in representation, storage, searching, finding, and presentation of *poten-*

tial information desired by a human user. Only when a user perceives potential information it becomes information to her. Potential information that is not perceived remains data (Ingwersen, 1992, pp.31-32)[1].

In this thesis, the distinction between 'data' and 'information' is not an essential question. Anyhow, the term potential information refers in this thesis to the semantic information content of documents.

From the point of view of the indexing task, the information of documents is always potential information: in principle, indexing takes into account all potential users with all potential information needs. Moreover, index terms do not contain the actual information of documents, but they are only pointers that guide a user to the information. Therefore, information of index terms can be considered as a kind of **meta-information**. van Dijk writes (van Dijk, 1977, p.122): First of all, it might be assumed that all (formal) INFORMATION IS PROPOSITIONAL, whatever the precise cognitive implications of this assumption. That is, we reconstruct knowledge as a set of propositions. A simple argument and predicate like 'the book' or 'is open' are not, as such, elements of information, only a proposition like 'the book is open'. In the same way, a simple index term, as such, is not capable of giving information. If, for instance, 'book' is an index term, then a user of the index is informed that the document contains information about a book or books. She must, however, read the document in order to find out that 'the book is open' (or whatever is said about books). On the other hand, multi-word terms may contain some potential information as well. For instance, the index term 'feelings as source of knowledge' (a real example from Griffiths and Whitford, 1988) contains more potential information than the index term 'feelings'. Typical index terms, however, are not propositional. The main function of index terms is not to present potential information, but to indicate 'what is being written about'. Thus it may be concluded that information of index terms is meta-information pointing to potential information of documents.

[1] Meadow distinguishes between data and information as follows (Meadow, 1992, p.22): An operational definition is that *information is data that changes the state of a system that perceives it*, whether a computer or a brain; hence, a stream of data that does not change the state of its receiver is not information.

Chapter 3

Summary

The following remarks summarize some main points of *Part I*:

- Language is a means of communicating information.
- Language engineering may provide methods of overcoming the problem of information overload.
- ‘The information content of the text’ is identified here with ‘the potential information content of the text’.
- Potential information becomes information when it is perceived.
- Index term is an expression that describes the contents of a text and guides a user to the information.
- Information of index terms is meta-information pointing to potential information of documents

The main focus of the study is on the potential information content of the text and on the exploring the linguistic features of index terms that guide to that information. The communication process as a whole is not under examination. Likewise, the cognitive and mathematical approaches to information and communication are outside the scope of the study.

Part II

Index terms

This part will discuss

- different approaches to indexing (*Chapter 4*):
  - What are index terms?
  - What kind of indexes can be found?
  - What is manual indexing about? Although this thesis will focus on automatic indexing, manual indexing is a relevant issue as well, since the index term corpus of this thesis is based on manually created indexes (*Section 4.2*).
  - What is the difference between index terms (objects used in the process of seeking information), topics (i.e., topic as a linguistic concept), and terminological terms (*Section 4.3*)?
- the information description languages in general (*Chapter 5*).
- the **method** of this thesis to improve indexing and information retrieval: the development of the automatic indexer by using the **index term corpus** (*Chapter 6*). This issue will be discussed in *Chapter 15* in more detail, but *Chapter 6* will give an overview.
- the **theoretical** contribution of this thesis: a new concept **index-term-structure** will be introduced in *Chapter 7*. This chapter will furthermore briefly discuss the empirical study of this thesis from the point of view of the index-term-structure.

Chapter 4

What are index terms?

As concluded in the previous part, information of index terms is meta-information pointing to potential information of documents. This chapter will discuss the index terms and indexing task in more detail.

4.1 Indexing task

According to ANSI 1968 Standard (American National Standards Institutes, 1968), an index is

a systematic guide to items contained in, or concepts derived from, a collection. These items or derived concepts are represented by entries in a known or stated searchable order, such as alphabetical, chronological, or numerical.

Indexing is

the process of analyzing the informational content of records of knowledge and expressing the informational content in the language of indexing system. It involves:

1. selecting indexable concepts in a document; and
2. expressing these concepts in the language of the indexing system (as index entries); and an ordered list.

An indexing system is

the set of prescribed procedures (manual and/or machine) for organizing the contents of records of knowledge for purposes of retrieval and dissemination.

An index term is an expression which contains a considerable amount of information (or meta-information) about the content of a text; for example, an index in a book consists of terms that refer to key content included in the book, such as concepts, persons, events. In information retrieval systems, an indexing language is the language that describes the documents and queries, and index terms (or descriptors or keywords) are the elements of the indexing language. Indexing can be done automatically or by human indexers, and index terms can be expressions derived from the text or expressions defined independently. So, index terms reflect the content of the text and even make a kind of shallow summary of the content. The main purpose of index terms, however, is to indicate to users ‘what is being written about’, not ‘what is written about certain issue’. Thus, the shallow summary provided by the index terms is a summary of ‘what is being written about’.

All indexing has the same underlying task of guiding a user to the relevant sources of information, but there are several different types and levels of indexes. Indexes of different kind could be categorized by using the following levels (Cleveland and Cleveland, 1983, pp.29-34):

1. word and name indexes,
2. book indexes,
3. periodical indexes, and
4. information retrieval system indexes

An example of a **word and name index** is a Bible concordance. This kind of index consists of the actual words of the text with no vocabulary control. In **book indexes** terms are manually generated and often in different form than in the text. **Periodical indexes** are in many ways similar to book indexes, only with broader scope. Periodical indexes are open-ended projects that involve a number of different authors with different styles and topics. The purpose of **information retrieval indexes** is to code the content indicators for effective retrieval of relevant documents. Often the index terms of information retrieval systems are word stems automatically derived from a document and weighted according to their distribution in a document collection.

Within the levels described above there are, for example, the following types (Cleveland and Cleveland, 1983, pp.35-44):

1. author indexes,
2. alphabetic subject indexes,
3. classified indexes, and
4. permuted title indexes

Author indexes guide the users to the titles of documents by way of authors. In **alphabetic subject indexes**, all index terms are in alphabetical order. **Classified indexes** are arranged in a hierarchy of related topics. Generic topics are on the top of the hierarchy and specific topics on the bottom. **Permuted title indexes** use the title words of documents as content indicators. In this thesis, book indexes with alphabetical order are the source of data, and the main objective of the study, on the other hand, is to develop a tool that automatically generates information retrieval system indexes.

4.2 Manual indexing

When a document is added to a collection, an indexer must ask several questions about the item (Lancaster, 1991, p.8):

1. What is it about?
2. Why has it been added to our collection?
3. What aspects will our users be interested in?

The characteristics and quality of indexes vary widely. For manual indexing there are procedures and instructions that guide the indexer's work. Indexing includes several activities (Cleveland and Cleveland, 1983, pp.62-74):

1. content analysis,
2. assigning of content indicators,
3. adding location indicators,
4. assembling the resulting entries, and
5. choosing the physical form in which the final index will be displayed

Careful content analysis is necessary in order to generate appropriate content indicators. Titles, subtitles and the abstract of text are good indicators of subject content, and likewise first and last sentences of paragraphs are considered to carry the message of the paragraph. Once the document has been analysed and subjects of the document have been determined, the next step is to convert the list of derived concepts into the controlled vocabulary of the indexing language. The derived concepts are checked in the thesaurus of standard index terminology and the final index terms are taken from there. They may be exact equivalents, synonyms, narrower terms, broader terms, or related terms. Many indexing rules have been designed in order to control the consistency and quality of indexes. Rules are not universal, and in different guides they may even be in contradictory. Following examples give a general idea of what rules look like (Cleveland and Cleveland, 1983, pp.62-64):

1. Refer singular to plural terms:

   Cat, *see* Cats

2. When writing modifications of terms, introduce the phrase with a word that stands out and catches the attention of the user:

   Sex, the use of TV in the teaching of

3. Use initials of authors:

   Jones, A. F.

4. Index to the maximum specificity signified by the author. (Don't "post up" to a more generic term if the author's specific word has an acceptable term at that level.) For example, if the author is talking about B-52 bombers and that is an acceptable term, don't substitute "airplanes".

An indexer must also define an appropriate depth of indexing, that is, the optimal number of topics that will be covered in the index. If there are too few topics, users may miss something. If there are too many topics, users may have to read irrelevant material. It is a difficult task to determine the optimal level of exhaustivity. (Cleveland and Cleveland, 1983, pp.70-71)

4.3 Index terms, topics, and terminological terms

Many books include a separate name index and subject index. A name index consists of index terms that refer to the proper names of the text and a subject index consists of index terms that refer to subjects (or subject matters) of the text. Borko and Bernier, on the other hand, distinguish between (Borko and Bernier, 1978, p.142):

- Subject index: *Subjects* are the foci of the work, the central themes toward which the attention and efforts of the author have been directed. They are those aspects of a work that contain novel ideas, explanations, or interpretations. And they should all be indexed.

- Concept index: ... subjects may require introduction through other concepts, passing thoughts may be expressed, and examples may be used for illustration only. Such items are *concepts*; they aid in understanding the report, but they are not subjects and they need not to be subject indexed.

- Topic index: Many writings are divided into *topics* - often with subtitles. Indexing these topics (or their subtitles) creates an index to topics. Sometimes these topics are subjects, in which case they should be subject indexed. Usually, they are too broad for subject indexing; often they are concepts that serve to introduce, justify, prove, and amplify the subject studied and reported.

- Word index: An index to all words in a book is a concordance, or word index, not a subject index.

Word indexes are the most bulky; concept indexes are the next most bulky; topic indexes the next most; and the subject indexes the least bulky (Borko and Bernier, 1978, p.143). In this thesis, the central themes ("subjects") are referred to as **main topics** and the less central themes are referred to as **subtopics**. So, three kinds of index terms will be distinguished in the empirical case-study of this thesis:

1. Main topics,

2. Subtopics, and

3. Passing concepts and proper names

Topic is a frequently used term in linguistics as well. According to Brown and Yule (Brown and Yule, 1983, p.70) the notion of 'topic' is clearly an intuitively satisfactory way of describing the unifying principle which makes one stretch of discourse 'about' something and the next stretch 'about' something else, for it is appealed very frequently in the discourse analysis literature. In *Section 7.2* the notion of topic will be discussed in more detail, but at this point, topic (or discourse topic) is simply defined as 'what is being written about in the course of discourse'. The notion of topic has both similarities and dissimilarities with the notion of index term. Both describe the content of the text, but the point of view is different. For instance, a proper name mentioned only in parentheses is probably included in an index of a book. It is, however, unlikely interpreted as a topic of the text. When index terms are chosen, the criterion is to choose those items that someone might be interested in.

Terminology as a discipline has a notion of term which differs from the notion of index term. Terminology is concerned with collection, definition, standardization and presentation of terms which are well-defined lexical items belonging to special subject languages, and consisting of symbol, concept, referent and definition. Terms are often appropriate index terms as well, but they are not defined specially for information retrieval, as index terms are. Term definition ought to be as exact and universal as possible, whereas index terms in the first place describe a particular document. From the linguistic point of view, the theory of terms is, in principle, part of a theory of lexicology. Topic structure analysis, on the other hand, belongs to the study of text linguistics or discourse analysis. Terminology and text linguistics clearly have a different approach to language, but information retrieval is concerned with both of them. The weighting schemes applied in information retrieval systems aim at weighting the more essential topics of the discourse more highly. Indexing languages, however, include usually not only topics and terminological terms, but passing proper names and concepts as well. The overlap of terminological terms, topics, and index terms is illustrated by *Figure 4.1*.

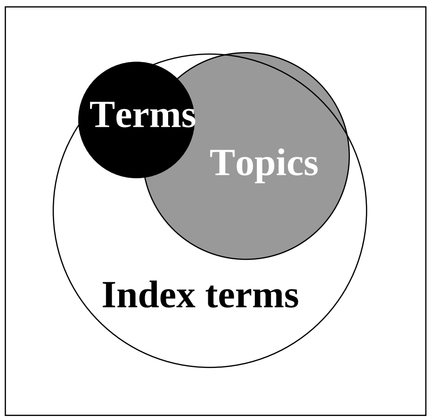

Figure 4.1: Terminological terms, topics and index terms.

Chapter 5

Information description languages

Harter arranged some major classes of information description languages along a continuum, by the degree of their departure from natural language prose (*Figure 5.1*). The left half of the continuum presents the natural language approaches to information representation and the right half of the continuum presents the controlled vocabulary approaches to information representation. The natural language approaches include full texts of documents, abstracts, titles, and identifiers extracted from the original text by indexers. The controlled vocabulary approaches include descriptors, subject headings, and hierarchical classification. The difference between identifiers and descriptors is that whereas identifiers are derived from the original text, descriptors are listed in thesauruses, which helps to deal with synonyms, homographs and such. The difference between descriptors and subject headings, on the other hand, is that whereas thesauruses are usually derived from existing document collections, subject heading lists are often *a priori* attempts to represent the whole structure of the universe instead of representing the vocabulary of specific document collection. Hierarchical classification scheme is an *a priori* representation of all human knowledge in a hierarchy, for example, Dewey Decimal Classification (DDC) used in United States primarily to classify books. (Harter, 1986, pp.42-51)

The index term corpus of this thesis is based on manually produced book indexes, in which many index terms are not directly derived from the text. For example, the expression `critical-dialectical perspective` of the text, is referred to as `dialectical analysis` in the index. The expressions in the index are often more general or more standardized than the expressions of the text. Book indexes are thus on the borderline between the natural language approach and the controlled vocabulary approach. In the process of constructing the index term corpus, the research aide marked up the index terms of book indexes into the texts and made thus an estimation of their textual origin and context. Constructing the index term corpus is then an attempt to transfer the description of the potential information content to the natural language side of the continuum.

All classes of information description languages along the continuum from identifiers to hierarchical classification represent, more or less, meta-information, whereas full texts, abstracts, and

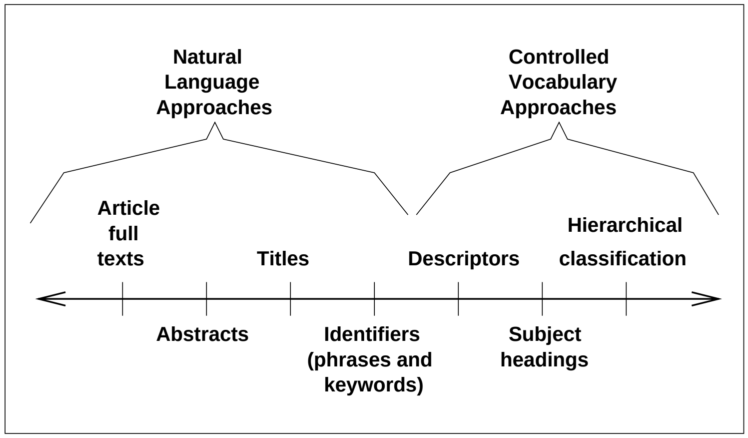

Figure 5.1: Information description languages, arranged by degree of departure from natural language (Harter, 1986, p.42).

perhaps also titles represent more or less potential information. It is assumed here then that this continuum might describe the degree of meta-information and potential information among the classes of information description languages as well. At the left edge of continuum (‘full text’) the degree of potential information is highest and toward the right edge of the continuum this degree decreases. Respectively, at the right edge of continuum (‘hierarchical classification’) the degree of meta-information is highest and toward the left edge of the continuum this degree decreases. It is obvious that a full text contains the highest degree of potential information, because it contains it all. Naturally, full texts include all the identifiers that represent meta-information as well. The degree of meta-information, however, is understood here as the degree of how fully potential information is *replaced* by meta-information, and in that sense in full texts the degree of meta-information is lowest. Hierarchical classification, on the other hand, may be considered to represent the highest degree of meta-information and the lowest degree of potential information, since its descriptions are most general and standardized; a more general description gives a more general impression of ‘what is being written about in the document, and what is written about it’. Consider the following example derived from the index term corpus. The book index includes the index term `women as inferior`, and the text includes the noun phrase `philosophical conceptions which exclude women` on the referred page. This noun phrase was marked up as an index term in the index term corpus. `Philosophical conceptions which exclude women` may be considered to contain more potential information than `women as inferior`, because it tells that women as inferior is related to philosophical conceptions in this discussion. A given a priori description could possibly be even more general than `women as inferior`. If a user is able to view the descriptions, then an index term that contains more potential information may give a better impression of the content of a document than a more general and standardized description. On the other hand, if a user is *not* able to view the descriptions, then a more general and standardized description may provide a more appropriate access to information than complex ad hoc index terms extracted from texts. A user is probably not able to formulate a search term `philosophical conceptions which exclude women`, whereas she may be able to use some more general and standardized descriptions, especially if she is familiar with the conventions of using the controlled vocabulary in information retrieval systems. Descriptions that provide more potential information are then more appropriate in interactive interfaces that let a user view the descriptions.

In a basic information retrieval system index terms are single words, and a query is constructed of single words combined using the boolean operators. A user interested in ‘women as inferior in philosophy’ might construct the query ‘women AND inferior AND philosophy’. Obviously, the user may retrieve a number of non-relevant documents that include those words but that do not discuss ‘women as inferior in philosophy’ (problem of **precision**). On the other hand, such relevant documents that do not include the word ‘inferior’ are not retrieved (problem of **recall**). Different techniques have been developed in order to overcome these problems. Proximity Searching refers to widely used technique which enables a user to search for terms situated within a specified distance to each other. For example, many systems include a NEAR operator which renders possible to determine the maximum distance between the terms. In the STAIRS retrieval system, the following operators are available in addition to AND, OR, and NOT (Salton and McGill, 1983, pp.34-41):

- ADJ specifying that two terms must be adjacent to one another,
- WITH indicating that the two terms must appear in the same sentence,
- SAME specifying that the two terms must appear in the same paragraph,
- SYN specifying that the two terms are to be considered as synonyms, and
- XOR (exclusive OR) indicating that a document is selected whenever it contains either of the specified terms but not both

These operators certainly improve the retrieval performance, but they are still incapable of expressing syntagmatic relations between the words as the index term `philosophical conceptions which exclude women` does. Potential information of this index term is based on these syntagmatic relations to a large extent. Using boolean combinations of keywords is in any case a more feasible way to construct a query than guessing what might be the complex index terms extracted from a document. Multi-word terms are used for indexing, in addition to single-word terms, in order to improve the precision. If a multi-word term occurring in a query is found in a given document, then the degree of potential information may be higher in the match of query term and document term, than in the case of single-word term match. The degree of potential information in term matching is strictly related to the precision of retrieval performance.

To sum up, information of index terms is meta-information, since index terms tell users ‘what is being written about’. Thus, from the point of view of the primary *function* of index terms, information of index terms is first and foremost meta-information. On the other hand, index terms may contain more or less potential information as well, since multi-word terms may tell users something about ‘what is written about a given topic’. So, in an interactive process where a user can view the index terms, multi-word terms that contain potential information have a secondary function of providing potential information as well. Potential information of multi-word terms, however, is probably in most cases insufficient to meet the information need of a user, but it can help to choose the relevant documents.

Chapter 6

Index term corpus

This chapter will give a brief overview of the index term corpus and the empirical study of this thesis. The issue will be discussed in *Part IV* in more detail.

The index term corpus of this thesis is a linguistically analysed text collection where the index terms are manually marked up by a research aide. The linguistic analysis of the index term corpus is done by a robust rule-based dependency parser, The Conexor Functional Dependency Grammar (FDG)[1] (Tapanainen and Järvinen, 1997) which is a relative of the Constraint Grammar framework (Karlsson *et al.*, 1995). The research aide identified and marked up the index terms for each document page using previously manually generated book indexes, that is, she marked up the closest equivalents of index terms found in the book indexes. Defining the content-bearing units of the text demands more or less subjective decisions, and a user of an index does not necessarily share the indexer's view. In any case, an index of a book represents an interpretation of the content of the text. An essential question of this case-study is whether these content-bearing units have such linguistic properties that make it possible to automatically distinguish between more appropriate and less appropriate index terms. The results of this study suggest that explored linguistic features of index terms provide a feasible basis to develop an automatic indexer. The corpus was divided into two parts: **a training corpus** and **a test corpus**. The features of index terms were explored using the training corpus, which is then the basis for the automatic indexer. The test corpus was used to test whether the results could be generalized beyond the context of the training corpus.

In information retrieval systems, index terms are usually weighted according to their importance for describing documents, and typically the weighting schemes are based on detection of word frequencies across the document collection. In the experiments of using natural language processing techniques to improve retrieval performance, the role of linguistic analysis is often restricted to discovery of multi-word phrases for indexing. These terms are then weighted by some frequency-based weighting technique. The weighting scheme of this thesis, however, combines evidence derived from word distributions with evidence derived from linguistic analysis.

[1]Originally developed at the Research Unit for Multilingual Language Technology at the University of Helsinki.

Chapter 7

Information structure, topic structure, and index-term-structure

This chapter will introduce a new concept **index-term-structure**. The term **information structure** has been used for many different purposes and so it will not be used in the framework of this thesis. Some examples of the different usages of this term are presented, however. All the different usages are related to the framework of this thesis in one way or another. The notion of **topic structure**, on the other hand, can be seen as a linguistic approach to the content analysis, and it is concerned with a number of relevant issues that are included in the framework of this thesis. **Index-term-structure** is identified with 'weighted index terms in their context' and it can be seen as a content analysis framework for information retrieval.

7.1 Information structure

Meadow (Meadow, 1992, p.1) characterizes **information** by the following definition (which he admits to be oversimplified): information is something that

1. is represented by a set of *symbols*,
2. has some *structure*, and
3. can be read and to some extent understood by *users* of information.

Also in Belkin's and Ingwersen's model of the cognitive communication system for information science (Belkin, 1978, p.81; Ingwersen, 1992, p.33) information is seen as structure. According to Harris, information has a language-like structure in certain aspects. In language, there are certain accretional steps in constructing a sentence and discourse, and correspondingly, information is structured by such accretional steps; when the form and meaning of sentences and discourses are decomposed into the accretional departures from equiprobability, the same structure is exhibited for the information therein (Harris, 1991, pp.322-356).

According to van Dijk (van Dijk, 1977, p.95), different aspects of **information distribution** in discourse, introduction, continuity, expansion, topicalization, focusing, etc., are **grammatically** interesting phenomena of the semantic structure of discourse: they are systematically associated with specific syntactic and morpho-phonological structures. If information has a language-like structure, it implies that analysis of linguistic structure could provide evidence for identifying informative expressions of text - index terms, for instance. The potential information of the text is represented by meanings which are constructed of meanings of words, meanings of clauses, and meanings of sentences. Natural languages provide countless ways in which to express meanings and only a human recipient is able to interpret them truly. However, weighting index terms involves only recognizing meta-information pointing to potential information, not interpreting the full meaning of a text. The question is then: how is this kind of meta-information encoded in texts? The results of this thesis suggest that index terms have certain typical morphological, syntactical, and lexical features, and that word frequencies provide likewise useful information for weighting index terms. The explored features of index terms provide the basis for defining automatically index-term-structure which can be considered as a kind of meta-information structure. It would be misleading then to refer to index-term-structure as information structure. Thus, in this thesis, the focus is on meta-information structure instead of information structure.

'Information structure' is, in fact, a highly ambiguous term. In linguistics, this term is used to refer to the organization of sentences in terms of the functions 'given' and 'new' which describe the status of information introduced into a discourse: **given information** is known to the addressee, and **new information** is unknown (Halliday, 1967, 1970a). van Rijsbergen, on the other hand, uses the term 'information structure' in a sense, which covers specifically a logical organisation of information, such as document representatives, for the purpose of information retrieval (van Rijsbergen, 1979, p.9). So, as may be seen, the term 'information structure' is used in many different ways, which is another reason for using the term 'index-term-structure' instead.

7.2 Topic structure

This section will discuss some relevant issues related to topic structure, such as lexical cohesion and location of topics in paragraphs. Lexical cohesion is related to **burstiness** and **TF*IDF**[1], which are important issues for the index term weighting scheme of this thesis. The empirical study of this thesis will furthermore detect whether the first and last sentences of a paragraph, and the first and last paragraphs of a section have a special role in index term weighting.

Topic structure is here thought to be a linguistic approach to describe the content of a discourse. There is a specific connection between 'discourse topic' and 'discourse content', the former consisting of the 'important' elements of the latter (Brown and Yule, 1983, p.107). The content of a discourse could be described by defining the topic structure that consists of the topics

[1] Burstiness and TF*IDF will be discussed in *Chapter 12* and *Section 13.3*.

and subtopics of the discourse.

Brown and Yule distinguish between **sentential topic**[2] and **discourse topic** (Brown and Yule, 1983, pp.70-71). Sentential topic describes sentence structure; Hockett makes a distinction between **topic** and **comment**[3] in a sentence as follows (Hockett, 1958, p.201): the speaker announces a topic and then says something about it [...] in English and the familiar languages of Europe, topics are usually also subjects and comments are predicates. The assumption that sentential topics are usually subjects is an interesting one from the point of view of automatic indexing. One essential assumption of this thesis is that syntactical analysis provides hints for weighting index terms, that is, it is assumed that some syntactical positions are more typical for appropriate index terms than others. There is an evident connection between discourse topics and index terms, since both of them are concerned with 'what is being written about in the course of discourse'. What is, however, the connection between discourse topics and sentential topics, and between index terms and sentential topics? In a way, sentential topics define 'what is being written about' on the sentence level, whereas discourse topics define 'what is being written about' on the discourse level. As Daneš points out, the themes of individual sentences (utterances) appear to be component parts of the global thematic structure of a text (Daneš, 1995, p.32). So, it might be assumed that all, or at least the great majority of discourse topics are sentential topics as well. On the other hand, it is clear that only a subset of sentential topics are discourse topics. Index terms, however, often include, among others, passing proper names and concepts that are not central discourse topics, but still something 'that is being written about'. Accordingly, the connection between index terms and sentential topics may be even closer than the connection between discourse topics and sentential topics. Thus, if sentential topics are usually subjects, it seems relevant to detect whether subject is a typical syntactical position for those index terms that should be highly weighted.

The discourse topic, on the other hand, expresses 'what is being talked/written about in the course of discourse'. Venneman considers the expression 'topic' or 'topic of discourse' as referring to a discourse subject on which attention of the participants of the discourse is concentrated (Venneman, 1975, p.317). Brown and Yule emphasize that there is always a set of possible expressions of the topic, and what is required is a characterisation of 'topic' which would allow each of the possible expressions, including titles, to be considered (partially) correct, thus incorporating all reasonable judgements of 'what is being talked about' (Brown and Yule, 1983, p.75). Such a characterisation can be developed in terms of a **topic framework** which is essentially a means of characterising the area of overlap in contributions to a discourse: identifying the elements in the topic framework enables us to produce a version of 'what is being talked about', i.e. the topic of conversation, which is more comprehensive than the single word-or-phrase-type title (Brown and Yule, 1983, pp.75-87).

[2]The sentential topic is related to the given/new framework (Halliday, 1970a, p.162)

[3]Or theme and rheme or topic and focus. For an automatic procedure for topic-focus identification, see Hajičová *et al.*, 1995.

Several frameworks for discourse analysis have been proposed[4], but in this thesis no such framework is applied. However, the ideas of these frameworks are highly relevant to content analysis and information retrieval. The following section will discuss some aspects of topic structure in more detail:

- coherence,
- cohesion,
- topic-shift, and
- paragraphs

The Cambridge encyclopedia of language (Crystal, 1997, p.423) gives the following definitions:

Coherence. The underlying logical connectedness of a use of language.

Cohesion. The formal linkage between the elements of a discourse or text.

Both notions are important for defining what a text is: A text plainly has to be *coherent* as well as cohesive, in that the concepts and relationships expressed should be relevant to each other, thus enabling us to make plausible inferences about the underlying meaning (Crystal, 1997, p.119).

Morris and Hirst developed a computational method for determining the structure of a text by means of lexical cohesive ties (Morris and Hirst, 1991)[5]. In this method, Roget's International Thesaurus was used as the major knowledge base for computing lexical chains which are sequences of successive nearby related words spanning a topical unit of the text. According to Morris and Hirst lexical chains tend to delineate portions of texts that have a strong unity of meaning, and thus they provide a valuable indicator of text structure. Determining the discourse structure related to cohesion is an essential step in determining coherence and the deep meaning of the text: Cohesion is a useful indicator of coherence regardless of whether it is used intentionally by writers to create coherence, or is a result of the coherence of text (Morris and Hirst, 1991, p.26).

Halliday and Hasan have recognized several types of cohesive relations (Halliday and Hasan, 1976, pp.2-27):

[4]such as those of van Dijk (macrostuctures; van Dijk, 1977, 1980, 1985), and Grosz and Sidner (Attention, Intentions, and the Structure of Discourse; Grosz and Sidner, 1986), Mann and Thompson (Rhetorical Structure Theory (RST); Mann and Thompson, 1988), and Suri and McCoy (Revised Algorithms for Focus Tracking (RAFT); Suri and McCoy, 1994)

[5]Hahn developed a text parsing system that determines underlying coherence by detecting lexical cohesive ties (Hahn, 1992). An example of an information retrieval system that attempts to determine coherence structures of texts explicitly is the Information System RUSSIA that constructs the thematic representation of a text using thesaurus knowledge about terms and property of text cohesion (Dobrov *et al.*, 1998).

- Conjunction,
- Reference,
- Lexical cohesion, and
- Substitution

In the framework of this thesis, lexical cohesion is the most relevant type, and it will be discussed in more detail below.

Halliday and Hasan distinguish between two types of lexical cohesion, **reiteration** and **collocation** (Halliday and Hasan, 1976, pp.274-292): collocation refers to cohesion that is achieved through the association of lexical items that regularly co-occur; reiteration is a form of lexical cohesion which involves the repetition of a lexical item or occurrence of a related item, for example:

- repetition (same word),
- synonym (or near synonym),
- superordinate, or
- generate word (such as people, person, man, thing, matter)

Automatic indexing typically relies on shallow detection of lexical cohesion. If certain words occur in certain documents more frequently than in others, it may indicate that these words are topic words in those documents. The repetition of certain words may indicate that the document also has a meaningful and relevant underlying semantic information content related to those words. The cohesive ties of the surface text may imply the underlying coherence, as seen above. A user is interested in the semantic information content of the document, but the search is done by matching words.

Different techniques have been developed in order to recognize other cohesive ties besides those of plain repetition. **Stemming** is a widely used method for collapsing together different words with a common stem. For instance, if a text includes words `Marx`, `Marxist`, and `Marxism`, it is reasonable to observe the distribution of the common stem `Marx` instead of three separate distributions of these words. These three words are associated with each other, and their co-occurrence in a document may imply an instance of lexical cohesion indicating a potential discourse topic. Accordingly, all of these three words might be considered as appropriate index terms. Synonymy, hyponymy, hypernymy, and other lexical relatedness of words, on the other hand, are detected by using for instance, **thesauruses** or techniques that define **semantic networks** of words. To sum up, in information retrieval systems, different lexical cohesive ties usually provide the basis for identifying ‘what is being written about in the course of discourse’.

Paragraphs may be regarded as highly cohesive entities. Hinds identifies paragraphs with units of writing that maintain a uniform orientation (Hinds, 1979, p.136). Brown and Yule use the term **topic-shift** in referring to the point at which the shift from one topic to the next is marked; the marking of topic-shift provides a structural basis for dividing up stretches of discourse into series of smaller units, each on a separate topic (Brown and Yule, 1983, pp.94-95). In written discourse paragraphs clearly are such smaller units. In practice, however, one paragraph may include more than one topics, and on the other hand, one topic may be discussed in more than one successive orthographic paragraphs. As Brown and Yule point out (Brown and Yule, 1983, p.95): Thus, it may be that the beginning of an orthographic paragraph indicates a point of topic-shift, but it need not do so. According to Longacre, the paragraph indentations of a given writer are often partially dictated by eye appeal; it may be deemed inelegant or heavy to go along too far on a page or a series of pages without an indentation or section break. A writer may, therefore, indent at the beginning of a subparagraph to provide such a break. Conversely, a writer may put together several paragraphs as an indentation unit in order to show the unity of a comparatively short embedded discourse (Longacre, 1979, pp.115-116).

Anyhow, paragraphs tend to be units indicating topic-shift, and thus, paragraph marking can be used as evidence, when it is determined 'what is being written about in the course of discourse'. If certain words occur in certain paragraphs more frequently than in others, it may be due to cohesion; and thus, it may indicate that these words are topic words in those paragraphs. According to Longacre in certain respects, a paragraph resembles a long sentence on the one hand and a short discourse on the other hand (Longacre, 1979, p.116). Accordingly, in information retrieval systems, many methods based on lexical cohesion could be applied to **paragraphs** as well as to documents. Zadrozny and Jensen argue that paragraph is the smallest domain in which topic and coherence can be defined; it is 'a unit of thought', and it makes more sense to talk about the meaning of a paragraph than about the meaning of a sentence (Zadrozny and Jensen, 1991, pp.172,207). Daneš considers paragraphs as central units of the thematic build-up of texts and describes the thematic structure of paragraph as follows (Daneš, 1995, pp.32-33): Thematic coherence is manifested by the fact that each paragraph has, in principle, a theme of its own, which appears as hypertheme in respect to the individual utterance themes that are subordinated to it.

A robust discourse analyser that could reliably and automatically resolve anaphora and define the thematic structure of a paragraph and of a text could contribute a great deal to automatic indexing, but unfortunately, no such analysis method is available. So, if, for instance, `Marxism` is referred to with a pronoun, it is not taken into account in counting the frequency of `Marxism`. If the pronouns referring to `Marxism` could be replaced by the word `Marxism`, it would increase the frequency of this word, and so the word would be weighted higher by an automatic indexer. The importance of anaphora resolution to information retrieval was studied by Pirkola and Järvelin who analysed the effects of ellipsis and anaphora resolution on proximity searching[6]

[6]Proximity searching is technique used in information retrieval, cf. *Chapter 5*.

in a text database and found that resolution was most relevant for person names (both anaphora and ellipses) and other proper name phrases (ellipses) and only marginal in other keyword categories (Pirkola and Järvelin, 1996).[7] The lack of world knowledge is a serious problem for resolving anaphora automatically, and thesauruses and such are probably not capable of providing all the world knowledge required for automatic anaphora resolution. Automatic indexing is typically based on word frequencies or word stem frequencies: the best terms are those of high frequency in some documents and low overall document collection frequency.

This thesis does not attempt to resolve anaphora in order to weight the index terms. Automatic indexing involves discovery of the appropriate expressions guiding a user to potential information of a given document, and topics are such expressions. If indexing is based on repetition of words or word stems only, it is evident that indexing gives only a rough approximation of what the topics of a document are. This technique is, however, fast and robust and easy to implement. In information retrieval systems, these rough approximations are compared with words or word stems of queries, and documents are retrieved and ranked according to the similarities. Repetition of words or word stems is an easily recognizable type of lexical cohesion which is detected by the automatic indexer of this thesis as well. The idea behind is that if a certain word occurs uncommonly frequently in a certain document, it indicates that the word may be a topic word of the document. Accordingly, if a certain word occurs uncommonly frequently in a certain paragraph (or in a group of neighbouring paragraphs) it may indicate that the word is a topic word of the paragraph, which means that the word is a topic word of the document. In this thesis, frequency-based techniques are combined with a linguistic technique that is based on exploration of the typical lexical, morphological and syntactical features of index terms.

One of the objects of this study is to detect whether the first and last sentences of a paragraph, on the one hand, and the first and last paragraphs of a section, on the other hand, have a special role in index term weighting. According to Gerdel and Slocum, paragraph topic occurs initially in the paragraph and often finally in the paragraph, thus indicating the beginning and the end of a topic and incidentally indicating the bounds of a paragraph (Gerdel and Slocum, 1976, p.275). Thus, it could be assumed that words in the first sentence and last sentence tend to include appropriate index terms.

Stark reports on an experiment in which 63 students were asked to judge what sentences of three essays (Russel, 1935; Didion, 1979; and Orwell, 1945) they considered to be important (Stark, 1988). Position in paragraph proved to have an effect: sentences at the beginnings of paragraphs were rated important more often than sentences in the middle or at the end. The last sentences of texts, however, were rated the most important sentences: they were rated important 62 % of the time, whereas the first sentences of texts were rated important 24 % of the time. The first sentences of paragraphs were rated important 46 % of the time, whereas the other positions were rated important only 20 % of the time. However, when the same texts were judged with-

[7]For an extensive study of anaphora in information retrieval, cf. Liddy, 1990.

out paragraph marking, the rated importance of these sentences shrinked from 46 % to 27 %. It suggests that judgements were not based only on the content of the sentence, but the paragraph markings were important cues to readers as well. On the other hand, when the same texts were judged with arbitrary misleading paragraph markings, only 21 % of the readers considered these non-natural paragraph-initial sentences to be important. Thus, putting a paragraph marking before a sentence did not make sentence more important. The effect of a paragraph cue could be considered as an interaction between the cue and the content. The results of Stark's experiment do not falsify the assumption that paragraph topic occurs initially in the paragraph, but they do suggest that the essence of potential information may be located in the middle of a paragraph as well. Given that topics make only a subset of index terms, it may be concluded that the first and last sentences of a paragraph may have a special role in weighting the index terms, but their importance is perhaps not to be overestimated.

7.3 Index-term-structure

The training corpus of this thesis is used to explore typical single-word and multi-word index term patterns. The automatic indexer weights representations of these patterns in running texts. Most of the words are included in the indexing language, but some obvious non-terms, such as `and`, `the`, and `she`, are excluded by the indexer[8]. When the automatic indexer marks up the weights into a text, it produces an analysis of the **index-term-structure** of the text. The index-term-structure could thus be identified with *weighted index terms in their context*. An index can then be made by extracting the index terms and by ranking them according to their weights. Information of index terms is here identified with meta-information pointing to potential information of documents, and the index-term-structure of a given document can thus be considered as a kind of meta-information structure.

Figure 7.1 presents the index-term-structure of the following example sentence (Harvey, 1990, p.48): `Marx saw social structures as oppressive`. The automatic indexer weights all words except `as` which is not considered as a potential index term here. The example sentence includes also one multi-word index term (`social structures`) that represents a typical index term pattern of the index term corpus. The pattern consists of two words: the first word is an adjective and a premodifier, and the other word is a noun and the head of a noun phrase. The sentence includes thus six index terms that are ranked as follows:

[8] A robust indexer should probably attach some weights to all words, because it is possible that `'and'`, for example, would be an appropriate index term of a text that discusses the function of `'and'` in English language.

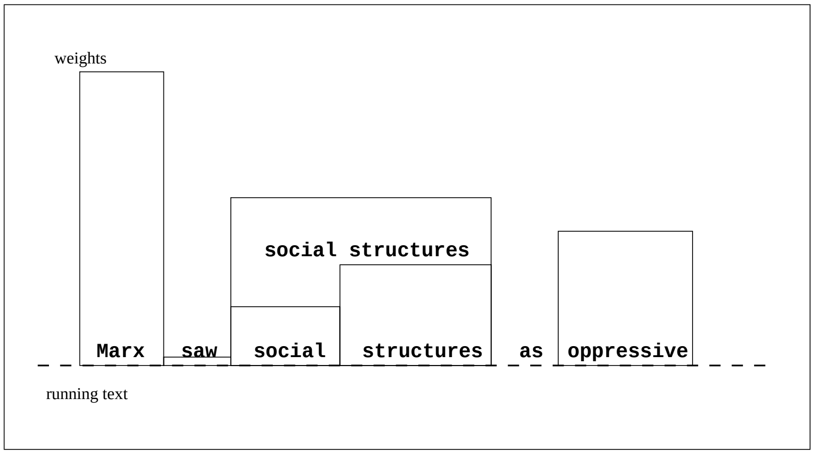

Figure 7.1: Index-term-structure of an example sentence.

```
Marx
social structures
oppressive
structure
social
see
```

The automatically produced index of the whole text would consist of all unique index terms of the text ranked by their weights[9]. Analysis of index-term-structure can also be used to measure the meta-information density[10] of a sentence, a paragraph, or a document. The meta-information density of a sentence is quantified by dividing the summed weights of the sentence by the number of running words in the sentence. The meta-information density of a paragraph or a document is quantified likewise. It might be assumed that a high meta-information density indicates a high information density as well. Anyway, this simple method could be applied to different tasks that are concerned with the quantification of information. For instance, it is possible to generate abstracts automatically by extracting sentences of the highest meta-information density. It is also possible to compare the informativeness of documents by ranking the documents according to their meta-information density.

The index-term-structure represents a kind of approximation of 'what is being written about in the course of discourse and where'. The **elements** of index-term-structure are the index terms, and the **relation** between these elements is, first and foremost, that of comparison. On the other hand, syntactic, cohesive and other linguistic relations of index terms are used to define the index-term-structure, that is, to calculate the weights, and thus, the index-term-structure has two levels:

[9]Low-weighted terms may be excluded, if, for instance, it is necessary to save space.

[10]Halliday's notion of **lexical density** provides a framework for measuring how closely information is packed. Lexical density is defined as the number of lexical items as a proportion of the number of running words. (Halliday, 1989, pp.61-67)

1. the underlying **linguistic level** and
2. the **meta-information level** described by weights

Moreover, the index-term-structure could express the thematic relations between index terms if a hierarchical topic structure in which the main topics are in the top of the hierarchy and subtopics in the bottom, was determined. Adding this kind of topic structure level into the index-term-structure would, however, require some method for detecting the semantic relations between index terms. For example, a thesaurus could determine standard terms and semantic relations between them, such as synonymy and hyponymy, and the main topics could be identified then using, for instance, the following criteria:

- The most frequent index terms are the most potential main topics.
- Index terms of the greatest number of synonymous or in some other way related terms are the most potential main topics.
- Index terms mentioned in titles and/or in the first or the last sentences of a document or a paragraph are the most potential main topics.
- Index terms that are neither too narrow nor too broad terms are appropriate main topic candidates.

In this thesis, the index-term-structure does not include the topic structure level, and the analysis of linguistic relations is used only to weight the terms, that is, to determine the meta-information level. The relation in focus involves then comparing the weights of index terms.

As discussed earlier, the index term corpus of this study is a linguistically analysed text collection where the index terms are manually marked up. A research aide identified and marked up the index terms for each document page using previously manually generated book indexes. In the book indexes, the index terms were usually simple noun phrases, but in texts the content of the noun phrases was sometimes expressed by using verbs, adjectives, or even clauses. For instance, the verb `oppress` and the adjective `oppressive` used in the text could be referred to with the noun `oppression` in the index. In such cases, the verb `oppress` or the adjective `oppressive` were marked up as index terms. So, the index terms of the index term corpus include more, for example, verbs, adjectives, adverbs, and complex noun phrases than the book indexes. Exploring the typical properties of index terms of the index term corpus provides the basis for developing the automatic indexer. The book indexes included roughly three types of index terms:

1. main topics,
2. subtopics, and
3. passing concepts and proper names

Sometimes it is difficult to define the borderline between main topics and subtopics, or between subtopics and passing concepts and proper names. Anyhow, the research aide marked up these categories of index terms into Grolier corpus which is a part of the index term corpus (cf. *Chapter 15*). It is important to be aware of this division, because different types of index terms are recognized to some extent by different criteria. Discourse topics, for example, can be recognized by their repetition in a text, whereas passing proper names and concepts have to be recognized, for example, by their lexical, morphological, and syntactical properties.

The results of this thesis suggest that index terms have certain typical morphological, syntactic, and lexical features. Word frequencies related to lexical cohesion provide likewise useful information for determining the index-term-structure. So, the automatic indexer determines the index-term-structure using, for example, the following linguistic evidence derived from training corpus:

- **parts of speech**, index terms of training corpus are typically nouns (or noun phrases)
- **syntax**, index terms of training corpus are more often subjects than objects
- **lexical features of words**, certain endings are typical for index terms of training corpus
- **lexical cohesion**, repeated expressions are often appropriate index terms

The typical properties of index terms will be discussed in *Chapter 21* in more detail, but some remarks are presented here as well. Why are index terms typically nouns or noun phrases and more often subjects than objects? Why are certain endings typical for index terms; and why does lexical cohesion provide a basis for weighting index terms? Nouns tend to carry such meta-information that characterize potential information of a text in a way that makes them the most appropriate index terms. Nouns refer, for instance, to such concrete objects of the world as persons and places that are often included in an index. Smeaton writes (Smeaton, 1992, p.272): It has always been assumed by researchers that in language it is the noun phrases that are the content-bearing units of information. This is not true for a full representation of meaning but noun phrases are good indicators of text content and for traditional information retrieval, that is what is wanted. As mentioned above, it is usual that verbs of a text are nominalized if they are included in an index (the index term is `oppression` instead of `oppress`), and consequently the index terms of the index term corpus include more verbs than the book indexes. The index term corpus of this study represents an abstract style that commonly uses nominalizations, and thus nouns are often used in a context in which it would be possible to use verbs as well. This increases the proportion of nouns and noun phrases in the index terms of the index term corpus. If, on the other hand, index terms were marked up into texts representing a less abstract style, the proportion of verbs could be higher. To sum up, the analysis of word classes (or parts of speech) provides important information for the automatic indexer, which, in principle, ranks nouns higher than adjectives, adjectives higher than adverbs, and adverbs higher than verbs[11]. Representatives of all other word classes are not weighted at all.

In the index term corpus some **syntactical positions** are more typical for index terms than others. As discussed above, subject is a typical syntactic position of sentential topics, that is, the subject of a sentence tends to describe 'what is being written about in the sentence', whereas other syntactical elements describe more or less 'what is written about the topic of the sentence'. Thus, it is not surprising that subject is a typical syntactical position for the index terms which are more concerned with 'what is being written about' than 'what is written about the topic'.

Concepts typical to the abstract style of the index term corpus have some **lexical features** that help in identifying the concepts automatically. Certain endings, such as `-ism`, `-ity`, and `-ogy`, are characteristic of these concepts which often are appropriate index terms as well. A number of these endings are in fact derivational endings, and accordingly these "lexical" features are then actually morphological features to a large extent. In any case, derivational endings are here included in lexical features. Another example of a useful lexical feature is related to the fact that with many concepts, such as "epistemology", "ontology", "ethics", "logic", and "objectivity", the indefinite article is not used. The dependency parser marks up such nouns by a `<-Indef>` tag. Although not all nouns with the `<-Indef>` tag are index terms in the index term corpus, this tag provides useful evidence for the automatic indexer. Another useful tag is the `<Proper>` tag which marks up the proper nouns; the dependency parser is capable of recognizing most of the proper nouns. To sum up, lexical features of words provide important information for the automatic indexer especially for recognizing passing concepts and proper names. These types of index terms may often be ranked low by an automatic indexer based on plain word frequencies or word stem frequencies.

The importance of **lexical cohesion** for automatic indexing was discussed earlier in this section. Repetition of words is an easily recognizable type of lexical cohesion which automatic indexing is typically based on. If certain words occur in certain documents more frequently than in others, it may indicate that these words are topic words in those documents; and accordingly, if a certain word occurs uncommonly frequently in a certain paragraph (or in a group of neighbouring paragraphs) it may indicate that the word is a topic word of the paragraph, which means that the word is a topic or a subtopic of the document. Thus, lexical cohesion provides important information for the automatic indexer especially for identifying discourse topics, whereas the lexical features of words are important for identifying passing concepts and proper names. Morphosyntactic and syntactic information may be useful both with regard to discourse topics and perhaps particularly with regard to passing concepts.

These examples presented above give an impression of how linguistic analysis can be used in defining the index-information-structure. In the process of defining the index-information-structure, the potential information content of a text is reduced to a meta-information-structure

[11] In practice, however, ranking is a more complicated matter, because of other linguistic information.

which is then a kind of description of the potential information content of the text. Blair writes (Blair, 1990, pp.137-138): First of all, there can be no necessary and sufficient (i.e. complete) representation of a text (other than the text itself and even *this* may not be sufficient for retrieval purposes [...]). Secondly, the standard to be used to judge the usefulness of a particular textual description is *not* that of "correctness", but one of "appropriateness". In other words, a textual description is neither correct or incorrect, but rather, more or less appropriate for a given task and situation. It must be asked then whether the index-information-structure is an appropriate description of the potential information content of the text. Answering this question properly would demand empirical testing. An interface should be developed that uses the method of this thesis in an information retrieval system, and then this method should be compared with other methods. This kind of testing is, however, beyond the scope of this study. Evaluation of the automatic indexer of this thesis is mainly done by using the test corpus of this study as a benchmark; we shall evaluate how well the automatic indexer ranks the index terms of the test corpus. What makes it difficult to compare the method of this thesis with other methods used in information retrieval systems, is the fact that there is, unfortunately, no other such experiment so similar to the experiment of this thesis that would make the comparison possible. Naturally, there are many automatic indexing procedures that can be compared with the automatic indexer of this thesis, and some comparison will be made in this thesis as well. The standard indexing procedures, however, rank single words or words stems according to their frequencies, and they are meant to be used in large-scale information retrieval systems that include thousands of documents. These methods tend to rank high topic words in particular, and they usually ignore multi-word terms totally, whereas the automatic indexer of this thesis pays attention to multi-word terms and passing concepts and proper names as well. In addition, frequency-based methods may attain better results if the document collection is larger than the one of this thesis. The standard indexing procedures are usually evaluated by the recall and precision rates of retrieved documents, whereas in this thesis the automatic indexer is evaluated by the recall and precision rates of retrieved index terms using the test corpus where the index terms are manually marked up, as a benchmark. To sum up, further research is needed in order to compare the appropriateness of the method of this thesis with other methods.

So, if the practical arguments for the use of index-term-structure are still under consideration, it should perhaps be asked whether the notion of index-term-structure is useful from the theoretical point of view. The ontology of index-term-structure relies clearly on the need for meta-information in information retrieval tasks, which gives a practical character to the notion of index-term-structure. The index-term-structure of a given text is not "natural", linguistically motivated structure as discourse structure or syntactical structure. Therefore, the notion of index-term-structure does not belong to the field of linguistic theory although the analysis of index-term-structure is based on linguistic analysis. The aims in the field of information retrieval are practical, which means that it is more important to develop effective methods than to determine why they are effective, what possible linguistic phenomena are behind them, or what are the appropriate theoretical concepts for describing the methods. This does not, naturally, mean that theoretical considerations would be neglected in the field of information retrieval. Anyhow, the aim of this thesis is to explore by an empirical experiment how linguistic analysis can be used in turning the potential information of a given text into meta-information that can be used in information retrieval tasks. Behind this aim there is a practical purpose of developing an effective automatic indexer. At the same time, however, this thesis tries, on the one hand, to understand the results of the exploration in the light of linguistic theory, and on the other hand, to define and use appropriate concepts for describing the discussed issue.

Index-term-structure is one of the theoretically important concepts of this thesis. It is identified with 'weighted index terms in their context', and it is related to the notions of information structure and topic structure. Index-term-structure, however, is not information structure but meta-information structure and discourse topics are only part of it. It is an important concept in this thesis, because, for example, the quantification of meta-information density of a sentence, a paragraph, or a document, is based on the 'weighted index terms in their context'. Index-term-structure produces a kind of profile for a given text that describes the amount of meta-information in different positions of the text. This profile could be of theoretical interest as well as practical interest. It may be assumed that the index-term-structure of a given text is not arbitrary, but there are certain (genre-specific) regularities that are based on the conventions of producing text. It is perhaps quite obvious that the producer of a text does not consciously produce meta-information, but if meta-information reflects the potential information, then the conventions of expressing potential information are reflected as the regularities of index-term-structures. These regularities can be identified by analysing manually index-term-structures of a number of texts, and the explored regularities can then be used for developing an automatic indexer that produces index-term-structures automatically. This method is applied in this thesis. Marking up index terms into the index term corpus produces a kind of index-term-structure in which the "weights" have values zero (non-term) or one (term)[12]. The automatic indexer based on these index-term-structures, however, weights index terms using probabilities varying from zero (i.e., non-weighted words) to a certain maximum value. Each index term has a number of different linguistic features, which are more or less typical for index terms, and the combination of these features determines the weight of a term.

According to Blair the nature of the informational task (the job) will determine what index terms (tools) will be used to search the collection (Blair, 1990, p.142)[13]. Words, sentences, and documents are tools, as well as index terms, for certain tasks. The task of index terms and index-term-structures is to help a user to find the information he needs. Blair states that information retrieval is fundamentally a process of communication. Inquirers are trying to

[12] Into one part of the index term corpus (Grolier) the research aide marked up the index terms using a scale from one to three (cf. *Chapter 15*). Grolier, however, was used only as a test corpus.

[13] Blair's idea of index terms as tools is based on "words-as-tools" metaphor discussed in the works of Zipf and Wittgenstein (Blair, 1990, pp.139-150).

describe the information they need in a way that indexers would understand, and indexers (or automatic indexing procedures) are trying to describe the content and context of documents in the collection in ways that would be understandable to the inquirers (Blair, 1990, pp.188-189). Different kinds of interfaces based on meta-information are needed between inquirers and documents in order to make information retrieval as effective as possible. An important means to improve the performance of an information retrieval system is to increase interaction between an inquirer and the system. This matter will be discussed below in more detail, but one example is presented here concerning the index-term-structure. Suppose an interactive interface in which a user is able to retrieve automatically produced document descriptions before retrieving full texts, in order to get a quick first look before choosing the most relevant full texts. A ranked index could be such an automatically produced document description. Viewing the highest ranked index terms certainly gives an impression of the potential information content of the document, but index terms represent only meta-information implying the potential information, because the contexts of index terms are not shown. The index-term-structure, however, makes it possible to view the ranked index terms in their context. Ranking sentences by their meta-information density produces a kind of abstract of a document. If a user views the highest ranked sentences, she will retrieve not only meta-information but fragments of potential information as well, because these sentences reveal at least something about 'what is written about the index terms of a document'.

Chapter 8

Summary

The following remarks summarize some main points of *Part II*:

- An indexing system is the set of prescribed procedures (manual and/or machine) for organizing the contents of records of knowledge for purposes of retrieval and dissemination (American National Standards Institutes, 1968).
- Index terms tell users ‘what is being written about’.
- Multi-word index terms may furthermore tell users something about ‘what is written about the matter that is being written about’.
- All indexing has the same underlying task of guiding a user to the relevant sources of information, but there are different types of indexes.
- In general, all topics are appropriate index terms, but not necessarily vice versa.
- **Index term corpus** is a linguistically analysed text collection where the index terms are manually marked up. It is the training and test material of the new automatic indexing method of this thesis.
- **Index-term-structure** is identified with ‘weighted index terms in their context’.
- The index-term-structure represents a kind of approximation of ‘what is being written about in the course of discourse and where’.
- The notion of index-term-structure is a kind of content analysis framework for information retrieval.
- Linguistic analysis provides evidence for determining the index-term-structure.

Part III

Index terms and information seeking

This part will discuss, among other things

- information seeking strategies (*Chapter 10*),
- the basic techniques of information retrieval,
- the state of art in natural language information retrieval, and
- some theoretical assumptions and models behind the information retrieval techniques (*Chapter 12*)

This section will try to answer the following two broad questions:

- What is information retrieval?
  - *Chapter 9* will briefly describe the basic information retrieval system.
  - *Section 9.1* will draw distinction between information retrieval, data retrieval, passage retrieval, and information extraction, although they are all connected to each other.
- What are the concepts, the theoretical assumptions and the techniques of information retrieval?
  - *Section 9.2* will define notions of **recall** and **precision**.
  - *Chapter 11* will present some points concerning the impact of natural language processing techniques on information retrieval.
  - *Chapter 12* will discuss some theoretical issues that are important to the frequency-based weighting schemes.
  - *Chapter 13* is one of the core sections of this part. It will discuss many essential issues concerning automatic indexing; a number of descriptions of algorithms, formulas, techniques, and experiments will be presented. The chapter will give fairly detailed examples of what have been done and how, since the issue is highly relevant to the empirical part of the thesis.

Chapter 9

Information retrieval systems

Salton and McGill give the following definition for the information retrieval system (Salton and McGill, 1983, p.xi): An information retrieval system is an information system, that is, a system used to store items of information that need to be processed, searched, retrieved, and disseminated to various user populations. 'Items of information' are usually documents in a document collection, and an information retrieval system is used to retrieve the relevant documents relating to the user's request (query). The present 'Information age' has challenged the traditional information retrieval systems which mostly have been used to handle bibliographic records and textual data in libraries and newspapers, and in legal and medical domains, among others. The explosive growth of the quantity of available information, information in Internet and also the new inventions like hypermedia, for instance, widen the scope of information retrieval and render necessary to improve the old methods and develop new ones.

Figure 9.1 presents a typical information retrieval system. A document collection consists of full texts which are represented by sets of index terms. A user formulates a query which in many information retrieval systems is constructed by using the boolean operators (AND, OR, NOT). The weighted index terms of the query are compared with the weighted index terms of the documents. Similarity is measured, for instance, by treating the index terms of the query and the index terms of the document as vectors in a n-dimensional space, and the similarity of the vectors is determined by calculating the cosine of the angle between the vectors. The output is usually a set of citations or document numbers ranked by the similarity which approximates the relevance of documents.

9.1 Information retrieval, data retrieval, passage retrieval, and information extraction

It is necessary to draw a distinction between information retrieval (IR, or document retrieval) and data retrieval (DR). *Table 9.1* presents some of the differences listed by van Rijsbergen (van Rijsbergen, 1979, p.2). Data retrieval systems store the data in data bases in explicit form, as

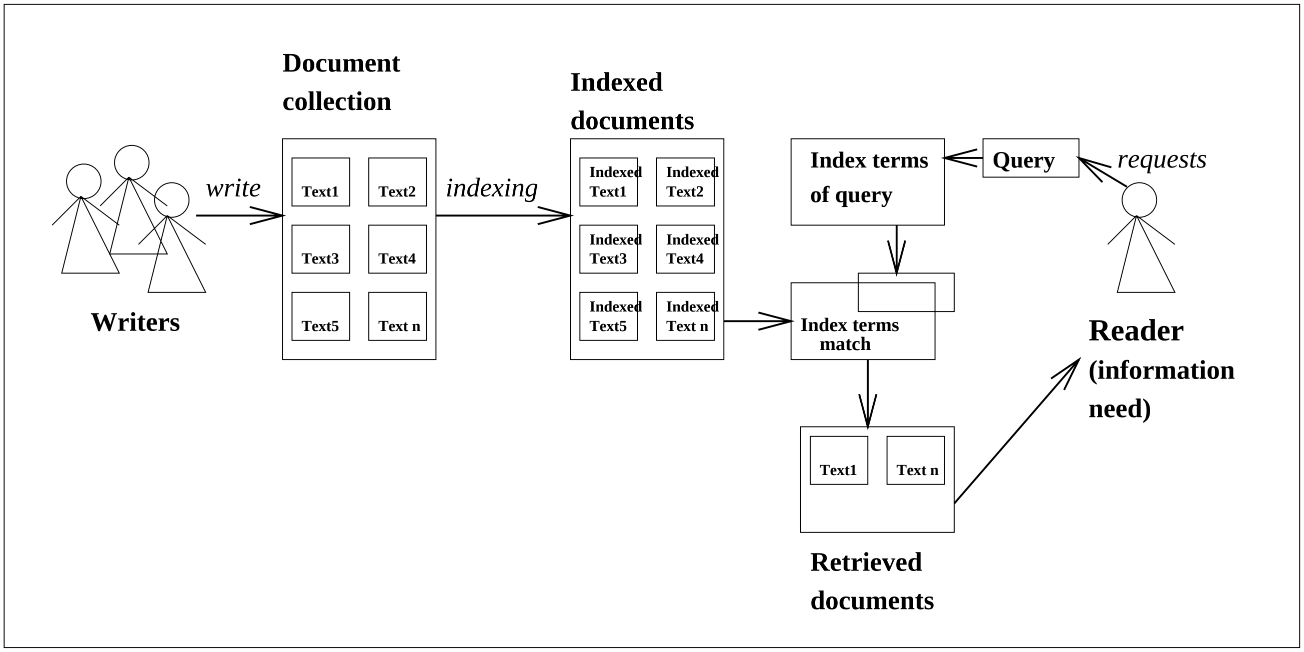

Figure 9.1: A typical IR system

tables, records and fields, which means that information does not appear as natural language text, whereas documents in information retrieval systems are usually items of natural language text. Data retrieval systems retrieve the actual information desired, an exact answer, if the information is available (exact match, deterministic model), whereas information retrieval systems retrieve documents, which are more or less relevant to the user (partial match, probabilistic model).

Salton and McGill demonstrate the overlap among different types of information systems as presented in *Figure 9.2* (Salton and McGill, 1983, p.10). In question-answering systems, the data base consists of facts relating to special areas of discourse, together with general world knowledge. Questions and responses can be presented in natural language form. The query is analysed and compared with the stored knowledge, and the answer is assembled from the relevant facts. Data base management is concerned with storage, retrieval, updating, deletion and protection of data in explicit form, and management information adds analyzing and synthesizing procedures to data base management. So, these two last mentioned information systems are closely connected. Question-answering systems and information retrieval systems, on the other hand, both are concerned with natural language data, but the former retrieves facts and the latter retrieves documents.

Although this thesis is concerned with the area of document retrieval, two terms more or less related to question-answering systems have to be mentioned here. In **information extraction** (IE, also known as **message understanding**), unrestricted texts are analyzed and a limited range of key pieces of task specific information are extracted from them[1]. A typical information extraction

[1]SCISOR (The System for Conceptual Information Summarization, Organization, and Retrieval) is a well-known example of information extraction system (Jacobs and Rau, 1990).

	Data retrieval (DR)	*Information retrieval* (IR)
Matching	Exact match	Partial match, best match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Classification	Monothetic	Polythetic
Query language	Artificial	Natural
Query specification	Complete	Incomplete
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive

Table 9.1: Data retrieval and information retrieval (van Rijsbergen, 1979, p.2)

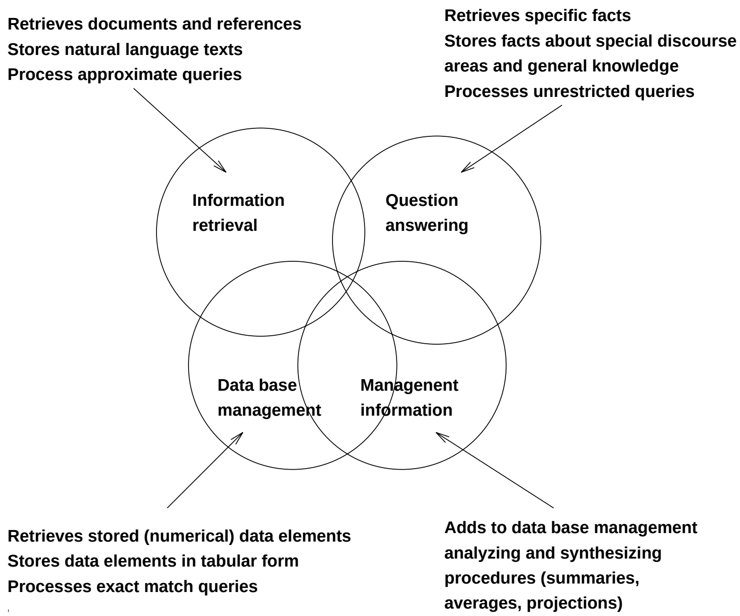

Figure 9.2: Overlap among types of information systems (Salton and McGill, 1983, p.10).

task is to extract information about management succession events from news articles. Outputs are in structured forms (e.g., case frames). In recent years efforts in this field have been focused through the US-government sponsored Message Understanding Conference (MUC) initiative[2]. Another term to be mentioned here is **passage retrieval**. According to Salton and McGill, syntactic analysis methods render possible directed retrieval activities that would take into account individual document portions such as sentences and paragraphs, and this has led to the so-called passage retrieval, where attempts are made to retrieve individual passages or sentences of documents rather than complete documents only [...]. Passage retrieval is based on

[2]cf. e.g., http://www.tipster.org/muc.htm, or http://www.muc.saic.com/

the analysis of the full text of documents, the aim being to retrieve either *answer reporting* passages, that is, passages from which an answer to a question can effectively be inferred, or alternatively *answer indicative* passages which indicate that the same document also contains an answer reporting passage. Passage retrieval may be advantageous because answers to questions could be immediately available instead of merely references to answers (Salton and McGill, 1983, p.284).

The distinction between passage retrieval and information retrieval, on the one hand, and between information extraction and information retrieval, on the other hand, is not always so clear; in some techniques these approaches are combined. Passage retrieval has been applied to information retrieval in order to improve retrieval effectiveness[3]. The MultiText system, for example, identifies small passages relevant to query and presents them to a user for the selection of additional query terms[4] (Clarke and Cormack, 1997). In some systems terms are extracted automatically from the relevant passages and added then to the query without asking from the user (Xu and Croft, 1996). In different passage retrieval approaches, documents may be divided into passages using ortographical paragraph division, even-sized windows, or analysed subtopic structure (cf. e.g., TextTiling in *Chapter 12*).

Information extraction techniques based on frames may be considered as a semantic or conceptual approach to information seeking. Usually these techniques are used for fairly narrow information extraction tasks, but some attempts have been made to apply them to information retrieval systems as well. Bear *et al.* from Stanford Research Institute (SRI) have made some experiments of adapting FASTUS[5] Information Extraction System to information retrieval tasks (Bear *et al.*, 1998)[6]. A typical FASTUS application extracts domain specific information by employing the following sequence of phases:

1. *Tokenizer*. This phase accepts a stream of characters as input, and transforms it into a sequence of tokens.

2. *Multiword Analyzer*. This phase is generated automatically by the lexicon to recognize token sequences (like "because of") that are combined to form single lexical items.

3. *Name Recognizer*. This phase recognizes word sequences that can be unambiguously identified as names from their internal structure (like "ABC Corp." and "John Smith").

4. *Parser*. This phase constructs basic syntactic constituents of the language, consisting only

[3]cf. e.g., Callan, 1994; Wilkinson, 1994; Wilkinson and Zobel, 1995; Hearst and Plaunt, 1993.

[4]Query expansion techniques are discussed in *Section 13.5*.

[5]FASTUS is an acronym for Finite State Automaton Text Understanding System.

[6]FERRET ("Flexible Expert Retrieval of Relevant English Text", cf. Mauldin (1991)) and DR-LINK (Liddy and Myaeng, 1993, 1994) are other examples of a conceptual approach to information retrieval. DR-LINK is a modular system that processes and represents text at the lexical, syntactic, semantic, and discourse levels of language. DR-LINK uses case frames in generating concept-relation-concept triples; documents and queries are represented as concept graphs which are matched during the retrieval process.

of those that can be nearly unambiguously constructed from the input using finite-state rules (i.e., noun groups, verb group, and particles).

5. *Combiner*. This phase produces larger constituents from the output of the parser when it can be done fairly reliably on the basis of local information. Examples are possessives, appositives, "of" prepositional phrases ("John Smith, 56, president of IBM's subsidiary"), coordination of same-type entities, and locative and temporal prepositional phrases.

6. *Domain or Clause-Level Phase*. The final phase recognizes the particular combinations of subjects, verbs, objects, prepositional phrases, and adjuncts that are necessary for correctly filling the templates for a given IE task.

The design of FASTUS was motivated by the design of MUC style scenario template tasks in which a narrowly defined prespecified information requirement was posed and fairly extensive effort was devoted to writing application grammars to answer that requirement. Adapting FASTUS to information retrieval tasks involved thus writing general, application-independent patterns for which it is possible to write application-specific instances which typically are tied to the argument structure of the topic-relevant verbs. In the experiment FASTUS was used then as a post-filter to the output of an information retrieval system. According to Bear *et al.*, the results while certainly not impressive, just as certainly do not foreclose the possibility of using IE technology in this way in IR applications. Rather they suggest that some care must be exercised in determining the proper range of application of this mixed-technology approach to IR, for there is little reason to think it is appropriate everywhere (Bear *et al.*, 1998, p.375).

9.2 Efficient information retrieval systems

An important issue for information retrieval systems is the notion of relevance; the purpose of an information retrieval system is to retrieve all the relevant documents (**recall**) and no non-relevant documents (**precision**). Recall and precision are defined as

$$Recall = \frac{Number\,of\,retrieved\,relevant\,documents}{Total\,number\,of\,relevant\,documents\,in\,collection}$$

$$Precision = \frac{Number\,of\,retrieved\,relevant\,documents}{Total\,number\,of\,retrieved\,documents}$$

Figure 9.3 presents a partition of document collection with respect to the relevance of the retrieved documents (cf. Salton and McGill, 1983, p.164). 'Silence' refers to the missing relevant documents and 'noise' to the retrieved non-relevant documents. Efficient information retrieval - minimizing the silence and noise - presupposes a good practice and theory of document representation.

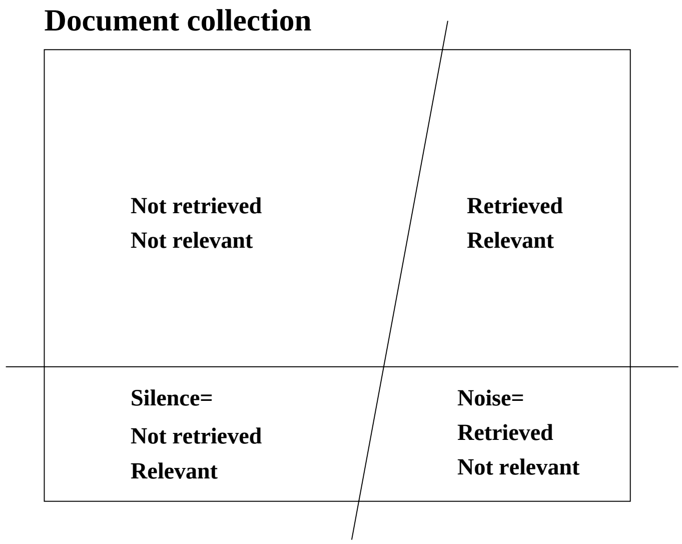

Figure 9.3: Partition of document collection.

Ingwersen states that in information retrieval, text representation has developed through four stages [7] (Ingwersen, 1992, p.22):

1. one book = one assigned class or index term, or single term extraction from the text
2. keyword phrases, morpho-syntactic term extraction, clustering
3. semantic values combined with request modelling
4. really adaptive, knowledge-based systems, pragmatic systems

According to Ingwersen, the first and the second stages represent level of traditional, system-driven information retrieval research which, combined with more user-oriented information retrieval, is reaching the third level. So far, the third stage has not been achieved completely, and the fourth stage cannot be achieved in computerized systems, except by direct support from humans, for whom all four stages are always available. (Ingwersen, 1992, pp.22-23)

The approach of this thesis is not user-oriented; the focus is not on developing interactive interfaces, but rather on developing tools that may be useful in developing interfaces. An index based on the index-term-structure represents meta-information and thus belongs to the second level. An abstract based on the index-term-structure, on the other hand, contains some potential information as well, and it could be of some use in the third level applications, if it was combined with an interactive interface. Swanson has characterized the limits of automatic indexing and retrieval in his “Postulates of Impotence” as follows (Swanson, 1985):

[7] Ingwersen applies here De Mey’s four stages through which thinking on information processing has developed (De Mey, 1977, p.xvii; De Mey, 1980, p.49).

- P5: Machines cannot recognize meaning and so cannot duplicate what human judgement can in principle bring to the process of indexing and classifying documents.

- P6: Word-occurrence statistics can neither represent meaning nor substitute for it.

- P9: [Therefore,] consistently effective fully automatic indexing and retrieval is not possible.

So, automatic semantic analysis must face two serious problems: machines cannot recognize meaning, and meaning, on the other hand, depends more or less on the individual human interpreters, that is, a given text, for instance, has not only one single meaning but as many as there are interpreters. Moreover, for different recipients different things represent meaningful information depending on recipient's state of knowledge and information need. Vickery and Vickery consider that it is unlikely that the semantic structures of retrieval systems can be tailored to the needs of the varied enquirers who will wish to use them. The emphasis must therefore be on iterative dialogue between system and user, to achieve a more effective match between the information want and the information available in the system (Vickery and Vickery, 1992, p.179).

Chapter 10

Information seeking strategies

This chapter will consider different information seeking strategies in order to demonstrate that querying is not the only way to seek information. This issue is relevant to the development of automatic indexer, since, for example, the requirements of hypertext index terms are quite different from the requirements of index terms used in query-based systems. If index terms are used for navigation and browsing, they are visible to users. An automatic indexer should then recognize the most relevant single-word and multi-word key terms and no noise (i.e, obscure terms) should be produced. On the other hand, if index terms are used in a traditional query-based system, they will not be visible to users. In that case the noise is not such a serious problem. Often the index terms of a query-based system are single stems with weights, and they are not necessarily the most revealing document descriptors for users to look at. One of the goals of this thesis is to develop an automatic indexer that produces index terms that are relevant to view by users. This is one reason why multi-word terms with more potential information are extracted as well as single-word terms.

Interaction with texts is a central process of information retrieval. Belkin makes the following assumptions related to information-seeking behaviour as human interaction with texts (Belkin, 1993):

1. Information-seeking is inherently an interactive process, and that process is characterized by the general features of people's interactions with texts;

2. The goal of IR systems is to support the range of information-seeking behaviors.

Belkin, Marchetti, and Cool characterized information-seeking strategies according to four dimensions which define the space of information-seeking strategies (ISS) (Belkin, Marchetti, and Cool, 1993). *Figure 10.1* summarizes the set of 16 prototypical information-seeking strategies based on these four dimensions which are:

1. the **method** of the interaction (scanning vs, searching)

2. the **goal** of the interaction (learning vs. selecting)

3. the **mode** of retrieval (recognizing vs. specifying)

4. the **type of resource** interacted with (information vs. meta-information)

ISS	METHOD		GOAL		MODE		RESOURSE	
	Sc	S	L	S	R	S	I	M
1	x		x		x		x	
2	x		x		x			x
3	x		x			x	x	
4	x		x			x		x
5	x			x	x		x	
6	x			x	x			x
7	x			x		x	x	
8	x			x		x		x
9		x	x		x		x	
10		x	x		x			x
11		x	x			x	x	
12		x	x			x		x
13		x		x	x		x	
14		x		x	x			x
15		x		x		x	x	
16		x		x		x		x

Method: Sc = Scan; S = Search
Goal: L = Learn; S = Select
Mode: R = Recognize; S = Specify
Resource: I = Information; M = Meta-Information

Figure 10.1: Information Seeking Strategies (ISS) (Belkin, Marchetti, and Cool, 1993).

Information-seeking strategies can be characterized according to their values along these dimensions, and a user of an information retrieval system may engage in several information-seeking strategies in the course of an information seeking episode. Belkin, Marchetti and Cool describe an interface, BRAQUE, which allows a user to move from one information-seeking strategy to another. This kind of approach presupposes establishing relations between user characteristics and specific information-seeking behaviours, but, at the moment, this issue is still a matter for further research. (Belkin, Marchetti, and Cool, 1993; Belkin, 1993)

Waterworth and Chignell, on the other hand, identified the following three dimensions of information exploration (Waterworth and Chignell, 1991):

1. **Structural responsibility**: navigational vs. mediated search,

2. **Target orientation**: querying vs. browsing, and

3. **Interaction method**: referential vs. descriptive interfaces

Structural responsibility is the dimension concerning who is responsible for searching and consequently who is responsible for the structure. From the system perspective, navigation is un-

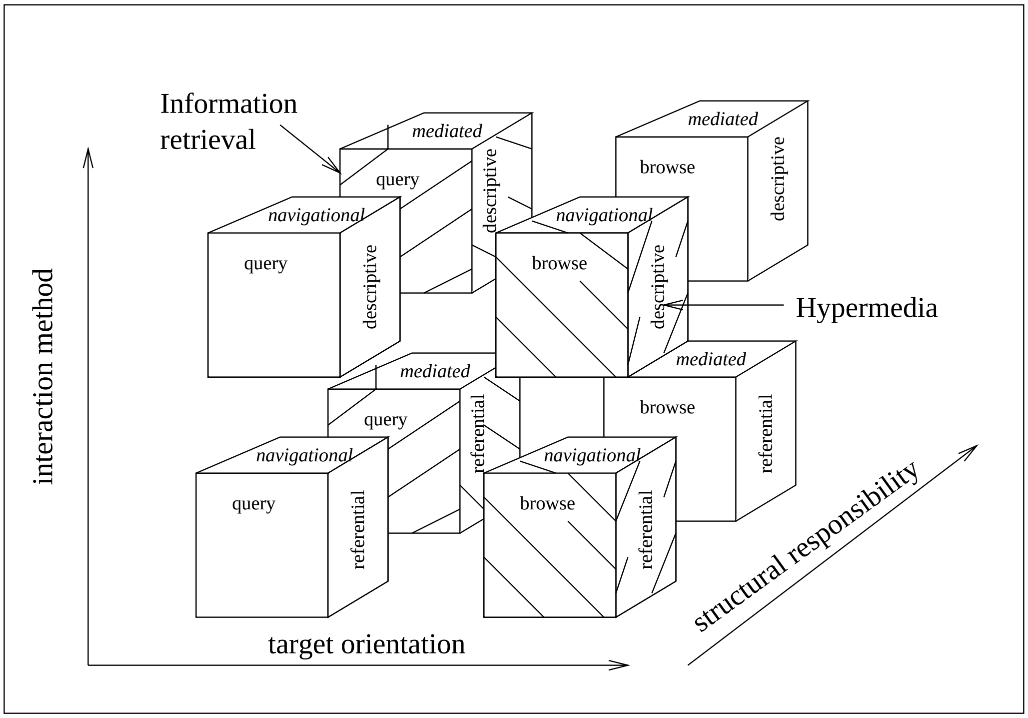

Figure 10.2: Three dimensions of information exploration: highlighting the components emphasized in information retrieval and hypermedia (Waterworth and Chignell, 1991, p.47).

structured but from the user's perspective, it is structured. In navigation the user is responsible for controlling the search process, whereas in traditional information retrieval, the system is responsible for searching and thus the system must be concerned with structure as well. **Target orientation** is the dimension which contrasts browsing and querying. Browsing can be distinguished from querying by the absence of a definite target in the mind of the user, and thus the distinction between browsing and querying is determined by the cognitive state of the user. User behaviours varying between querying and browsing may be thought of as a continuum that is characterised by the level of specificity of the user's informational goals. The **Interaction method** is the dimension which distinguishes between descriptive interfaces, typical to querying behaviour in traditional information retrieval, and referential interfaces, typical to hypertext browsing. *Figure 10.2* presents three dimensions of information exploration highlighting the components that are emphasized in information retrieval and hypermedia. Whereas traditional information retrieval systems have used mediated querying, hypermedia developers have emphasized navigational browsing. According to Waterworth and Chignell, One of the dilemmas for hypermedia enthusiasts is how to promote hypermedia as a general information seeking environment [...] How can our conceptualization of hypermedia and information retrieval be extended to provide more general information exploration mechanisms? In our view, the hypermedia model can be extended to provide mediated search and querying capabilities, if we extend our view of what a hypermedia link is and how it functions (Waterworth and Chignell, 1991, pp.46-47). Index linking model is an example of making hypermedia more usable for information retrieval. Nodes in the hypermedia system are described by a set of index terms which is organized as navigable hypermedia and which provides topic access points for a user. When reading the texts, the user may add relevant index nodes to her ‘hypermedia query’, which increases the weights of the nodes that are indexed by that term. The user may thus trigger mediated search by browsing through the index and marking the relevant index nodes. This model combines navigation and mediated search into an integrated paradigm: With index linking the navigational style of movement between nodes is preserved, but there is also the possibility of mediated search and of an interesting form of constrained navigation where the nature of the current query limits navigational choices while the act of navigation itself modifies the currently defined query. Ideally, information exploration strategies should be implemented in an environment where smooth transitions between browsing and querying, and between navigation and mediated search are possible. Index linking appears to be a useful strategy for making this possible (Waterworth and Chignell, 1991, p.51).

To sum up, both the four dimension model and the three dimension model described above illustrate the various aspects of information seeking, and at the same time these models make it evident that the current information retrieval systems are still incapable of offering a search environment which captures all different dimensions of information exploration.

Chapter 11

Natural language processing techniques and quantitative retrieval techniques

This section will present some estimations and comments concerning the impact of natural language processing (NLP) techniques on information retrieval. In addition, it will briefly describe the **data fusion** approach in which natural language processing techniques and quantitative retrieval techniques can be combined. The data fusion techniques described in this section are not applied in this thesis as such, but the weighting scheme of this thesis can be seen as a kind of data fusion technique.

van Rijsbergen remarks that linguistic analysis is expensive to implement and it is not clear how to use it to enhance information retrieval (van Rijsbergen, 1979, p.15). According to Smeaton, until recently, information retrieval was like other potential application areas for NLP in that it could not use NLP techniques as they were neither robust, efficient nor reliable enough. Now that has changed and information retrieval research, which has expended so much effort over the last 30 years developing statistical and keyword based approaches which have always had obvious limitations, is now starting to use NLP approaches to the processing of text in a constructive fashion (Smeaton, 1992, p.269).

A typical natural language processing technique used in information retrieval extracts multi-word terms for indexing by means of syntactic analysis. A basic quantitative retrieval method uses word stems as index terms and weights the terms basing on the word stem frequencies across the document collection. The retrieval performance can be executed using, for instance, the vector space model (VMS), or some probabilistic model. The vector space model measures the similarity between the query and documents by weighted inner product of overlapping terms. Queries and documents are converted into vectors which are weighted to give emphasis to the important terms. Query vectors are compared with document vectors and documents are ranked according to the similarities between the vectors. SMART[1] is a typical example of a system using the vector space

[1]For over 30 years, the SMART project at Cornell University has been one of the most prominent platforms in

model (cf. e.g., Salton and McGill, 1983). The probabilistic models, on the other hand, consider the probability that a term or concept appears in a document, or that a document satisfies the information need. INQUERY[2] is an example of a probabilistic model that uses a Bayesian network architecture (Croft, Callan, and Broglio, 1994). Retrieval is viewed as a probabilistic inference process which compares text presentations with queries. Documents are ranked according to their probability to meet user's information need.

Natural language processing modules can be incorporated into systems using the vector space model or some probabilistic model. SMART, for instance, has been frequently used as the statistical core engine in natural language information retrieval experiments. In INQUERY, on the other hand, text representations are based on different forms of linguistic and statistical evidence. Croft *et al.* describe INQUERY as follows (Croft, Callan, and Broglio, 1994): This approach (generally known as the interference net model and implemented in the INQUERY system) emphasizes retrieval based on combination of evidence. Different text representations (such as words, phrases, paragraphs, or manually assigned keywords) and different versions of the query (such as natural language and Boolean) can be combined in a consistent probabilistic framework. This type of "data fusion" has been known to be effective in the information retrieval context for a number of years, and was one of the primary motivations for developing the interference net approach.

So, INQUERY is an example of an approach in which multiple types of evidence are generated and combined within a single retrieval system. The **data fusion** approach, however, has been applied to combine the ranking results of distinct systems (often referred to as retrieval experts in this context), say, SMART and INQUERY, as well. Some promising results have been reported in this field (e.g., Bartell *et al.*, 1994), as well as some less promising results (e.g., Vogt *et al.*, 1997). Two important questions concerning mixture models are, what experts to use, and what fusion methods and parameters to use in order to maximize the performance of the system. According to Vogt, the best time to linearly combine IR systems is when both have reasonable performance of similar magnitude, but do not rank relevant documents in a similar fashion (Vogt, 1997). This notion makes sense intuitively, since retrieval techniques emphasizing different document and query features retrieve different relevant as well as non-relevant documents, and by optimal evidence combination it may be possible to combine the strengths of different techniques. Experts are typically combined linearly, but it has been proposed, for instance, to use non-linear neural network models as well (Bartell *et al.*, 1994; Bartell, 1994).

In their experiment[3] Bartell *et al.* combined two experts, the term expert and the phrase expert (Bartell *et al.*, 1994). The term expert was a standard vector-space retrieval system using white-

information retrieval research. SMART (System for Manipulating And Retrieving Text) was developed by the late Gerry Salton.

[2]INQUERY is the product of the Center for Intelligent Information Retrieval at the University of Massachusetts at Amherst.

[3]For other experiments, cf. e.g., Belkin, Cool, Croft, and Callan, 1993; Lee, 1997.

space delimited terms, and the phrase expert retrieved documents that contain phrases appearing in the query, but only if these documents were not retrieved by the term expert. Thus the phrase expert was meant just to improve the term expert's document ranking. Even though the phrase expert had very low performance compared with the term expert, the optimized combined system performed 12 % better than the term expert. Experts were combined linearly; the overall estimate $R_{\Theta,q}(d)$ for document *d* and query *q* was:

$$R_{\Theta,q}(d) = \Theta_1 E_1(q,d) + \Theta_2 E_2(q,d)$$

where $E_i(q,d)$ is the relevance estimate of expert *i*, and Θ_i is the scale of expert *i*. So, the set of Θ_i are free parameters for which values were determined automatically using a variation of Guttman's Point Alienation (Guttman, 1978) which is a statistical measure for rank correlation. Despite the low performance of phrase expert, the optimized phrase weight was found to be 0.738, and the term weight was 0.675. The reason for the higher phrase weight is that although the phrase expert often performs poorly, it can be very accurate when it performs well. Thus, Bartell *et al.* concluded that experts which perform poorly in isolation can still contribute positively to a combined solution (Bartell *et al.*, 1994).

The observations presented above support the hypothesis that data fusion is particularly effective if the experts are based on conceptually independent approaches which by their nature retrieve different sets of documents. Smeaton investigated whether the perceived conceptual independence of two retrieval strategies could be predictive of the improvement of the combined retrieval strategies by experiment of fusions of nine experts for Spanish. Smeaton found that fusing together independent retrieval strategies can yield improved retrieval effectiveness with pair-wise and triple-wise data fusion, even at high precision levels as has been observed elsewhere. However the results we have observed are not consistent and expected improvements in some fusion pairs did not materialise (Smeaton, 1998). Smeaton's results suggest that fusion of conceptually independent experts does not guarantee improvement in retrieval performance. Among other things, the quality of experts and the choice of optimal fusion methods and parameters are important factors as well. Data fusion is anyhow an approach in which linguistically motivated models could contribute to information retrieval, since natural language processing techniques may be considered as conceptually independent of quantitative techniques.

In recent years, different information retrieval techniques have been evaluated and compared quite exhaustively in Text REtrieval Conferences (TREC[4]). For example, fifty-one groups from 12 countries and 21 companies participated in TREC-6 which was held in 1997 (Voorhees and Harman, 1998b). All teams use the same training and test material, which renders possible the comparison of the results. Voorhees and Harman list the following goals for the TREC workshop series (Voorhees and Harman, 1998b):

[4]cf. Harman, 1993, 1994, 1995, 1996; Voorhees and Harman, 1997, 1998a.

- to encourage research in text retrieval based on large text collections;
- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.

Sparck Jones reports the results of TREC experiments by the following remarks (Sparck Jones, 1995):

Model questions.

- Are linguistically motivated models superior to statistically motivated ones? CMU's performance (CLARIT) shows that a linguistic approach performs perfectly well, but no better than statistical ones.
- Are linguistically grounded compound terms, as opposed to term conjunction in matching, valuable? There is no gain from using linguistically motivated, as opposed to adjacency defined compound terms.

Vocabulary questions.

- Does a holistic approach to the indexing vocabulary pay its rent? Only one of the TREC teams applied any strongly holistic vocabulary design methods, namely Bellcore, but without obvious benefit. However, some (e.g., Cornell and Dortmund) employed a pruned and normalised extracted phrase list (along with ordinary single terms), though this too, alone, does not appear to be significantly superior to simple single-term extraction and individual-item weighting.
- Is linguistic sophistication important? Manual thesauri were used primarily as adjuncts, but without noticeable effect (e.g., Siemens).

Description questions.

- Is linguistically motivated indexing superior to statistically motivated indexing, especially where this is derivative for individual documents? There is no clear gain from linguistic processing to select document terms (e.g., compounds, as in CMU).
- Are compound terms (whether linguistic or statistical) superior to single terms? There may, nevertheless, be a slight advantage from compounds (e.g., Amherst, Dortmund) as opposed to single terms.

- For the full texts used in TREC, what is the value of document-specific weighting schemes? There does appear, moreover, to be advantage in document-based weighting, practised by many participants, including, for example, Cornell, Dortmund.

Strzalkowski *et al.* summarize the experiences of TREC-6 in natural language information retrieval as follows (Strzalkowski *et al.*, 1998, p.365): The main observation to make is that thus far natural language processing has not proven as effective as we would have hoped in to obtain better indexing and better term representations of queries. Using linguistic terms, such as phrases, head-modifier pairs, names, does help to improve retrieval precision, but the gains remain quite modest.

The quotations above give a picture of the state of the art in natural language information retrieval. In the context of TREC-style retrieval tasks, it may seem uncertain whether natural language processing techniques will ever be mature enough to have an impact on information retrieval, and whether they can offer an advantage over purely quantitative retrieval methods. This, however, does not necessarily mean that it would be needless to continue the work in the area of natural language information retrieval. Even the best systems in TREC have not reached optimal results yet, and the development of retrieval techniques is still as urgent as ever. In addition, although the TREC tasks presented above include a wide range of different approaches to information retrieval, they do not represent the whole set of search strategies presented by Belkin, and by Chignell and Waterworth above. In querying users have a specific information need, but users may want to retrieve information by browsing as well. Users may want to learn something, or then they possibly want to select information. Different search strategies need different tools, and more sophisticated automatic content analysis is needed, if it is attempted to reach the semantic and pragmatic levels (cf. *Section 9.2*) in information retrieval systems. Possibly natural language processing techniques will have some impact on this development.

Chapter 12

Distribution of words in natural language

Although the approach of this thesis is clearly linguistic, not statistical, the problem of estimating probability distributions of words is discussed here briefly. The automatic indexer of this thesis relies on cohesion on one side, and cohesion is obviously related to the issue of word distribution. Various statistical methods based on different probability distributions, such as binominal distribution and poisson distribution, have been developed in order to find appropriate ways to approach natural language. But what is actually known about the distribution of words in natural language?

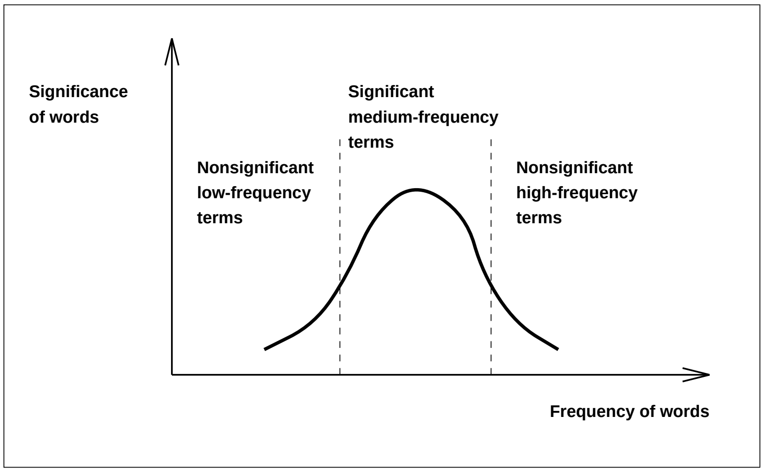

Figure 12.1: Medium-frequency terms are most content-bearing.

It is a general assumption in information retrieval that medium-frequency terms are most appropriate for indexing. This is, to a large extent, based on the work of Zipf (Zipf, 1949) and Luhn (Luhn, 1958). The famous constant rank-frequency law of Zipf

$$Frequency * rank \cong constant$$

states that if the word frequencies are multiplied by their rank order (i.e. the order of their frequency of occurrence), the product is approximately constant. Luhn remarks that the frequency of word occurrence provides a basis for a measurement of word significance, and states that medium-frequency words are most significant, as presented in *Figure 12.1* (Luhn, 1958; van Rijsbergen, 1979, p.16; Salton and McGill, 1983, p.62). The most frequent words (`the`, `of`, `and`, etc.), are least content-bearing, and the least frequent words are usually not essential for the content of a document either.

The observations presented above concerning the distribution of words are, however, of a very general nature. Dunning writes (Dunning, 1993, p.61): There has been a recent trend back towards the statistical analysis of text. This trend has resulted in a number of researchers doing good work in information retrieval and natural language processing in general. Unfortunately much of their work has been characterized by a cavalier approach to the statistical issues raised by the results. The approaches taken by such researchers can be divided into three rough categories.

1. Collect enormous volumes of text in order to make straightforward, statistically based measures work well.
2. Do simple-minded statistical analysis on relatively small volumes of text and either 'correct empirically' for the error or ignore the issue.
3. Perform no statistical analysis whatsoever.

Dunning argues that the last mentioned approach is typical for information-retrieval literature and mentions such central indexing techniques as inverse document frequency weighting (*IDF*), and Latent Semantic Indexing (LSI)[1] as examples of frameworks lacking sufficient statistical consideration. The basic problem of many statistical approaches is related to the fact that normal distribution is not suitable for statistical text analysis unless either enormous corpora are used, or the analysis is restricted to very common words. Using methods based on normal distribution tend to overestimate the significance of relatively rare events, which limits the ability of such methods to analyse rare events. Rare events, however, make up a large fraction of text. Dunning suggests the use of alternative statistical methods to those based on normal distribution, such as parametric statistical analysis based on binominal or multinominal distribution, methods based on poisson distribution, and distribution free methods. (Dunning, 1993)

It has been observed that function words tend to be closely modelled by a Poisson distribution, whereas content-bearing words do not tend to be, which means that words randomly distributed according to Poisson distribution are not informative in describing the content of a document. Thus, it is possible to identify content-bearing words by measuring the extent to which their distributions deviate from that expected under a Poisson process. Distribution of content-bearing words,

[1] *IDF* and LSI are discussed later in more detail.

on the other hand, has been described, for example, by a mixture of two Poisson distributions. (van Rijsbergen, 1979, pp.27-29)

Church and Gale made some experiments of comparing standard Poisson to Poisson mixture. The difference between these methods is described as follows: Under the standard Poisson, text is modelled as a homogeneous bag of words with a constant θ across documents, whereas under the proposed mixtures, the heterogeneity of text is modelled by allowing θ to vary over documents, subject to a density function ϕ, designed to capture dependencies on hidden variables such as genre, topic, author, etc. (Church and Gale, 1995, pp.188-189). The results of the experiment suggest that Poisson mixture fits the data better than standard Poissons, producing more accurate estimates, for example, of entropy (H) and inverse document frequency (*IDF*). (Church and Gale, 1995)

Katz, on the other hand, considers the mixture of Poisson distributions as an inappropriate model for distribution of content words, and presents patterns of word occurrences as an evidence against Poisson mixtures: the observed occurrences are not independent of each other, and multiple instances of content words are observed close to each other more often than it would be the case if they were an outcome of a Poisson process (Katz, 1996). Poisson mixtures as a two-stage stochastic mechanism for generating content words is thus incompatible with empirical data. In addition, discrete Poisson mixtures, such as the two-Poisson and the three-Poisson models, are limited in their capability to fit the data because of their faulty functional form. Katz approaches stochastic modelling of language as follows (Katz, 1996, p.16): it is desirable not to base probabilistic language modelling on *a priori* defined specific stochastic mechanisms (e.g., Poisson processes), but to represent dependencies on some *observable* language characteristics. We attempt such a constructive approach to statistical language modelling, capitalizing on basic discourse properties pertinent to text formation, and we use linguistically meaningful and observable text characteristics as model parameters. Katz's 'constructive approach to statistical language modelling' refers to his G-model which uses parameters rooted in the notion of topicality and in the distinction between single and multiple occurrence, as directly related to non-topical and topical use of a word or phrase.

According to Katz variations in document content correspond to variations in content words used, whereas variation in document length is mainly due to the varying number of different content words and to a much lesser degree, due to the varying number of instances of the individual content words used. Treatment of the concepts discussed in the documents introduces into discourse new content words and concepts related to these new words. Content words are accompanied by function words (determiners, prepositions, auxiliary verbs, etc.) making text formation possible. Individual content words and phrases are relatively rare and their treatment is usually narrowly localized in a document, whereas individual function words are used throughout the text. Thus, the numbers of instances of a given function word are approximately proportional to the document length, whereas the number of a given content word does nor explicitly depend on the document length; it is rather a function of how much a document is about the concept expressed by that word. A word or a phrase that is related to one of the main concepts of a document may be used throughout the entire document, which means that its occurrence does depend on the document size, but Katz argues that in such cases the document length can be viewed, not as a variable affecting the number of occurrences, but as a side effect of continuous and intensive treatment of the concept related to this word or phrase. Likewise, a tendency of longer documents to contain, on average, more instances of a given content word than the short documents is due to the fact that longer documents tend to discuss a given topic more exhaustively than short documents. The number of occurrences, per document, of the *individual* content words grows relatively slowly with the document length, but the number of *different* content words, per document, grows fast with the document length, which explains the linear growth of the *total* number of instances of content words with the document length. This weak dependence of the numbers of occurrences of content words on the lengths of documents is ignored by many statistical methods; content words are *not* scattered throughout the data collection, occurring in documents mostly as single instances, as a Poisson distribution with the rate $\lambda = f \cdot L \simeq 1$ implies. λ refers to the expected number of word instances in a document of length L. Relative frequency f, i.e. a proportion of instances of a particular word among all word instances determined from a large training corpus, refers to the expected number of word instances per word position. The problem with using relative frequency f to the description of distributional properties of content words is due to its incapability to take into account topicality. (Katz, 1996)

Katz's G-distribution, on the other hand relies on the notion of topicality. Documents of a document collection are classified into three groups according to their topicality with regard to a particular content word or phrase:

1. **Unrelated documents** contain no instances of this particular content word or phrase
2. **Non-topical documents** contain single instance of this particular content word or phrase
3. **Topical documents** contain more than one instance of this particular content word or phrase

Katz admits that it is possible that additional instances of a non-topically used word can be found, particularly in a long document, but for conceptual clarity this issue is ignored, as well as the possibility that a topically used word or phrase occurs in a document only once. In the mathematical treatment of the problem, single occurrence is simply equated with non-topical occurrence, and multiple occurrence with topical occurrence[2]. Katz presents results of an experiment made by Justeson and Katz as a support to the view that the number of instances of a word in a document indicates topicality (Justeson and Katz, 1995). In that experiment multiword terms occurring in a single paper were identified and classified as non-topical, topical, more topical, and highly topical, according to the degree of their topicality. The average frequency of terms was 3.00 for topical

[2]This somewhat crude generalization is understandable, but it raises the question whether this kind of generalization could be avoided somehow.

terms, 3.29 for more topical terms, and 4.75 for highly topical terms. Other experiments of this kind were made as well, and the results suggested generally high topicality of the most frequent terms (Justeson and Katz, 1995). (Katz, 1996)

Katz introduces the notion of **burstiness** as a fundamental concept related to topicality. Burstiness characterizes two closely related but distinct phenomena:

- **Document-level burstiness** refers to multiple occurrence of a content word or phrase in a single document, which is contrasted with the fact that most other documents contain no instances of this word or phrase at all.

- **Within-document burstiness** (or burstiness proper) refers to close proximity of all or some individual instances of a content word or phrase within a document exhibiting multiple occurrence.

According to Katz, a within-document burst of a given word always indicates that this multiple occurrence of the word is an instance of document-level burstiness as well, but not necessarily *vice versa*. A given content word or phrase may occur frequently in a document but at long intervals, without any within-document burst. A burstier word will be found in fewer number of documents than a less bursty word, and its multiple occurrence is usually narrowly localized in a document, often being confined to a single burst occupying just a paragraph or two. (Katz, 1996)

Katz's G-distribution models the distribution of content words and phrases along an extent of documents within a data collection:

- how likely the word occurs in a document

- how likely it is used topically when it occurs

- how intensively, on average, the word is used when it is used topically

(Katz, 1996)

The adequacy of G-model for the description of actually observed distributions of content phrases was tested empirically, and the results were satisfactory though the model passed statistical tests only when the range of document length was relatively small. G-distribution outperformed negative binomial distribution (Mosteller and Wallace, 1964) when compared how well they fit empirical data. Katz considers the adequacy of G-model as follows (Katz, 1996, p.50): We conclude that the assumption of equality of conditional probabilities in a document-level burst, which led to the derivation of the G-model, is consistent with the empirical data, and that, for a given document length or within a narrow range of the document lengths, the G-model provides an adequate description of empirical frequency distributions.

The phenomenon of burstiness is, in fact, the underlying basis for most frequency-based indexing techniques. Inverse document frequency (*IDF*), for instance, is based on the observation that bursty words that are found in fewer number of documents are often appropriate index terms.

In the widely used *TF*IDF* (often denoted as tf.idf) weighting scheme *IDF* is multiplied by a number of occurrences of a given word or phrase in a document (*TF*). Thus, if a word occurs frequently in a given document (*TF*), but does not occur in many documents (*IDF*), it is weighted high by *TF*IDF*; such word is a typical bursty word. Different weighting schemes are discussed in more detail in the following chapter.

Also in this thesis, both document-level burstiness and within-document burstiness are considered as central phenomena providing evidence for weighting index terms. From the linguistic point of view, burstiness may be viewed as a phenomenon closely related to lexical cohesion. Coherence, on the other hand, is related to topicality; a text is coherent as the writer of the text discusses a certain topic or topics. A text is lexically cohesive as the writer uses words or phrases related to the current topic; the writer either repeats the same words or uses synonyms or other related words. Repetition of the words related to the current topic is a prototypical instance of burstiness. It is evident, however, that observing the repetition of words or phrases can give only a shallow picture of the topic discussed. The treatment of a given topic does not necessarily mean that the same individual topical words or phrases are constantly repeated. Topics may be discussed using pronouns, synonyms, or other words or phrases more or less related to the topic; natural language offers a myriad of ways to express the things and ideas to be expressed, which makes it difficult to identify the topics of a discourse using only repetition of words or phrases as criterion. Thus, not only distribution of individual content words or phrases should be modelled, but distribution of related words or phrases as well. As seen above, it is not a trivial task to model distribution of content words in a general way, and certainly it is even less trivial to model distribution of related words.

Anyhow, topicality obviously determines distribution of related words as well as distribution of content words, which provides a basis for a number of techniques used in information retrieval, such as Latent Semantic Indexing (LSI), which organizes words into a semantic structure in order to overcome the problems of variability in word usage: the similarity of terms and documents is determined by the overall pattern of word usage in the entire collection, so that documents can be similar to each other, regardless of the precise words they contain. A description of terms, objects, and user queries based on the underlying semantic structure, rather than surface-level words, is used for representing and retrieving information. What this means from a user's perspective is that documents can be similar to a query even if they share no terms in common (Dumais, 1991, p.230).

Obviously, using Latent Semantic Indexing has some advantages in handling synonymy, polysemy, and term dependence, compared with the standard vector space model (VSM) which ignores the order and association between terms. Latent Semantic Indexing can be viewed as a kind of variant of the vector space model, but instead of representing documents and queries directly as sets of independent words, Latent Semantic Indexing represents them as continuous values on each of the *k* orthogonal indexing dimensions. Latent Semantic Indexing models the associative relationships using singular-value decomposition (SVD) which is a technique closely related to factor analysis and eigenvector decomposition. Singular-value decomposition provides a vector that represents the location of each term and document in the k-dimensional LSI representation, and the location of term vectors reflects the correlations in their usage across documents. Similarity between vectors is estimated based on the cosine or the dot product between vectors. Query terms are used to identify a point in the space, and documents are ranked according to their similarity to the query. Since both term vectors and document vectors are represented in the same space, similarities between any term combinations and documents can be obtained easily. Dumais mentions a number of systems that have the same idea of discovering and exploiting collection-specific inter-item associations, such as MatchPlus (Gallant *et al.*, 1993) and PhraseFinder (Strzalkowski *et al.*, 1995). (Dumais 1995; Hull, 1994)

In information retrieval systems, index terms are used to describe the documents of a collection, and thus the modelling the distribution of content words as well as detecting the distribution of related words is collection-dependent. For instance, parameters of Katz's G-model above were based on frequencies of content phrases across the collection, and in Latent Semantic Indexing inter-item associations are collection-specific. In other fields of language engineering, such as speech recognition, there is perhaps more need to model distribution of content words independently of a particular document collection. On the other hand, since within-document burstiness indicates document-level burstiness, it could be useful to detect within-document burstiness as well in order to identify topical content words and phrases more accurately; in particular if passage retrieval is used as a means to improve retrieval. Detecting within-document burstiness makes it possible to find relevant passages of documents, and these passages can be used in various ways, interactively or non-interactively, in order to improve the retrieval performance.

A number of techniques have been developed for analysing subtopic structures of texts. These text segmentation techniques typically detect changes in word distribution in order to identify the locations of topic shifts. Hearst's TextTiling is described here as an example of techniques which divide text into multi-paragraph subtopic passages (Hearst, 1997). Hearst's framework characterizes text structure as a sequence of subtopical discussions that occur in the context of one or more main topic discussions. The multi-paragraph approach is based on the observation that although in school one is taught that paragraphs should be written as coherent, self-contained units, complete with topic sentence and summary sentence, in real-world texts, however, these expectations are often not met. TextTiling uses three methods for detecting the locations of topic shifts:

1. Block comparison,

2. Vocabulary introductions, and

3. Lexical chains

The **block comparison** algorithm compares adjacent pairs of text blocks for overall lexical similarity. A lexical score is computed for the gap between every pair of sentences. Sentences are grouped into blocks, and the more words the blocks have in common, the higher is the lexical score in the gap between them. A low lexical score preceded by and followed by high lexical scores indicates a shift in vocabulary corresponding to a potential topic shift.

The **vocabulary introduction** method is based on Vocabulary-Management Profile technique described by Youmans (Youmans, 1991). It is kept track of how many first-time uses of words occur at the midpoint of every 40-word (35-word for Youmans) window in the text. Introduction of new vocabulary indicates a potential topic shift.

Lexical chains are detected in order to use analysis of lexical cohesion for identifying topic shifts. Hearst's technique is based on work of Morris and Hirst[3], but their method is not applied directly. Hearst does not use any thesauruses, only term repetition of morphological variants[4] of the same word is detected. Multiple chains are allowed to span an intention, and chains at all levels of intentions are analyzed simultaneously. A low number of active chains indicates a potential topic shift.

Hearst remarks that the ultimate goal of passage-level structuring is not just to divide texts into subtopic passages, but to identify and label the subject-matter of these segments as well. Hearst mentions hypertext display, information retrieval, text summarization, and text generation as examples of potential areas in which multi-paragraph segmentation could be useful (Hearst, 1997). The Xerox team used TextTiling in TREC-4 to partition documents into sets of multi-paragraph subtopical segments (Hearst *et al.*, 1996). This division provided a basis for TileBar visualization tool[5] which displays to a user the distribution of the query terms in the documents. The user may then choose to view the most appropriate documents and passages.

Salton and Allan report work on the automatic detection of hypertext links and theme generation using *TF*IDF* weighting and vector space model in order to find similarities among the paragraphs of large documents (Salton and Allan, 1993). All paragraphs, sections, and articles were given pairwise similarity scores, and those with the highest scores were linked together.

Lahtinen made some experiments of using within-document burstiness analysis as a means to identify appropriate index terms (Lahtinen, 1998). The frequency of a given word stem in a document (*SF*) was multiplied by inverse paragraph frequency (*IPF*):

$$SF * IPF = SF * log \frac{Number\, of\, paragraphs\, in\, a\, document}{Number\, of\, paragraphs\, containing\, the\, stem}$$

Thus, a stem occurring in many paragraphs has a low *IPF* and a stem occurring frequently in a small number of paragraphs has a high *SF*IPF*. This formula does not try to find index terms for

[3]cf. *Section 7.2*.

[4]Tokens are reduced to their roots by a morphological analysis function based on that of Karttunen, Koskenniemi, and Kaplan (Karttunen, Koskenniemi, and Kaplan, 1987).

[5]A number of tools have been developed for visualizing the search results in order to provide an easy interface for interactive retrieval process. cf. e.g., Nowell *et al.*, 1996; Veerasamy and Heikes, 1997; Hearst and Karadi, 1997; Smeaton *et al.*, 1998; Hemmje *et al.*, 1994.

each separate paragraph but for a whole document, just like *TF*IDF*. For this reason *SF* is defined as 'the number of times a stem occurs in a *document*' instead of 'the number of times a stem occurs in a *paragraph*'. Combined with the standard *TF*IDF* weighting scheme *SF*IPF* was found to give better results than either of the weighting schemes alone. *Section 18.1* will introduce another method for measuring within-document burstiness of words. This method detects the distances of the occurrences of individual words using paragraphs as units for measuring the distance.

Chapter 13

Automatic indexing

13.1 Representation and discrimination

A useful index term must fulfill a dual function (Salton and McGill, 1983, p.62, van Rijsbergen, 1979, p.29):

1. **Representation**: an index term must describe the potential information content of the document (recall function).

2. **Discrimination**: an index term must distinguish the document from the other documents of the collection (precision function).

According to van Rijsbergen, the emphasis on representation leads to a document-orientation which is typical for Artificial Intelligence (AI), for instance, and the emphasis on discrimination leads to a query-orientation which is typical for information retrieval (van Rijsbergen, 1979, p.30). In practice automatic indexing attempts to seek optimal trade-off between representation and discrimination, and several term weighting functions have been developed for this task. The following three term weighting functions are examples presented by Salton and McGill (Salton and McGill, 1983, pp.63-71):

1. The Inverse Document Frequency Weight,

2. The Signal-Noise Ratio, and

3. The Term Discrimination Value

The Inverse Document Frequency (*IDF*) weight is based on the assumption that term importance is inversely proportional to the total number of documents to which the term is assigned. That is, the smaller the number of documents containing the term, the greater the importance of the term for discriminating between the documents. A measure of the inverse document frequency for a term k can be written as (Sparck Jones, 1972):

$$IDF_k = log_2 \frac{n}{DOCFREQ_k} + 1 = log_2(n) - log_2(DOCFREQ_k) + 1$$

where *n* is the number of documents in the collection, and $DOCFREQ_k$ is the number of documents in which the term *k* occurs. *IDF* weighting provides a means to distinguish one document from another, but it does not necessarily describe the content of a given document very well. For instance, if the words `Marxism` and `throw` both occur in 100 documents, they have the same *IDF* weight even in a document in which `Marxism` occurs 20 times and `throw` only once. For this reason in widely used *TF*IDF* weighting, *IDF* is multiplied by the number of occurrences of a term *k* in a document *i* ($FREQ_{ik}$):

$$WEIGHT_{ik} = FREQ_{ik} \cdot [log_2(n) - log_2(DOCFREQ_k) + 1] = TF * IDF$$

Thus, *TF*IDF* weighting assigns a high degree of importance to terms occurring frequently only in few documents of a collection.

The Signal-Noise Ratio is based on Shannon's communication theory (cf, *Section 2.2.2*) which states that the higher the probability of occurrence of a word, the less information it contains. By analogy to Shannon's information measure, it is possible to define the $NOISE_k$ of an index term *k* for a collection of *n* documents:

$$NOISE_k = \sum_{i=1}^{n} \frac{FREQ_{ik}}{TOTFREQ_k} log_2 \frac{TOTFREQ_k}{FREQ_{ik}}$$

where $TOTFREQ_k$ is the total collection frequency of a term *k*. $NOISE_k$ varies inversely with the concentration of a term in the collection. Since nonspecific terms tend to have more even distributions across the collection (high noise), an inverse function of $NOISE_k$ might be used as a function of term value:

$$SIGNAL_k = log_2 TOTFREQ_k - NOISE_k$$

If index terms of a document are ranked in decreasing order of the $SIGNAL_k$ value, it favours terms that distinguish only a few specific documents in which these high-signal terms exclusively occur. The $SIGNAL_k$ value can be multiplied by $FREQ_{ik}$:

$$WEIGHT_{ik} = FREQ_{ik} \cdot SIGNAL_k$$

This weighting scheme combines again the representation function and the discrimination function, but according to Salton and McGill, the $SIGNAL_k$ value does not give optimal performance in a retrieval environment, because it emphasizes term concentration in only a few documents of a collection (Salton and McGill, 1983, pp.66,73).

The Term Discrimination Value measures the degree to which the use of the term will help to distinguish the documents from each other. The discrimination value $DISCVALUE_k$ of a term *k* can be computed by comparing *AVGSIM*, the average document-pair similarity calculated by comparing the words of documents, and $(AVGSIM)_k$, the average document-pair similarity if the term *k* is removed from all the documents:

$$DISCVALUE_k = (AVGSIM)_k - AVGSIM$$

Index terms may be placed then into three rough categories according to their discrimination values:

1. The good discriminators with positive $DISCVALUE_k$ whose introduction for indexing purposes decreases the space density
2. The indifferent discriminators with a $DISCVALUE_k$ close to zero whose removal or addition leaves the similarity among documents unchanged
3. The poor discriminators whose utilization renders the documents more similar, producing a negative $DISCVALUE_k$

Once again a weight can be computed to each term in each document by combining the term frequency factor with the discrimination measure:

$$WEIGHT_{ik} = FREQ_{ik} \cdot DISCVALUE_k$$

13.2 Indexing exhaustivity and term specificity

Traditionally it has been thought that the effectiveness of an indexing system is controlled by two main parameters: **indexing exhaustivity** and **term specificity**. An exhaustive index includes a large number of terms in index, whereas a nonexhaustive index includes only the most important index terms. More exhaustive indexing means that more documents are retrieved, and recall is improved. At the same time, however, the proportion of non-relevant documents increases. Recall is thus favored at the expense of precision. Term specificity, on the other hand, refers to the ability of the index terms to describe topics precisely. If index terms are highly specific, it may improve precision, since fewer documents are retrieved, but most of them are likely to be relevant. Broad terms may not be able to distinguish relevant documents from non-relevant documents as accurately as narrow terms. Using narrow specific terms affects recall, however, since many relevant documents are then rejected as well as non-relevant. In this case, precision is favoured at the expense of recall. In practice, some kind of compromise must be reached between recall and precision.

It is possible to make low-frequency narrow terms with a $DISCVALUE_k$ close to zero broader, more general, and more usable using thesaurus in which related terms are suitably grouped. These related terms can be added to the initial terms. Respectively, it is possible to make high-frequency broad terms with a negative $DISCVALUE_k$ more specific using combinations of these words, that is, phrases instead of individual words. Consider, for example, the words `subject` and `matter`. Separately they may be too broad terms for indexing purposes, but the phrase `subject matter` could be an appropriate term. Salton and McGill describe thesaurus transformation and phrase transformation as presented in *Figure 13.1* (Salton and McGill, 1983, pp.86-87). Terms are arranged into three classes along the document frequency continuum. The good, medium-frequency terms with positive discrimination value are placed in the middle. The broad terms are transformed into phrases and the narrow terms are transformed into broader terms by means of a thesaurus.

This scheme brings forth an argument for indexing by phrases in addition to single words. Phrases tend to be more specific than single word terms; they tend to contain more potential information than single word terms. Using multi-word terms is thus a means to improve precision.

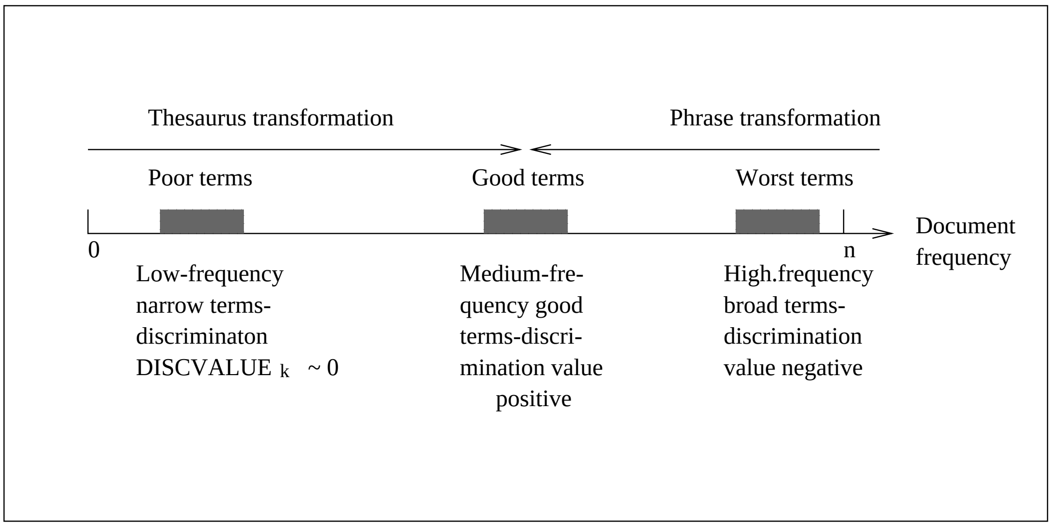

Figure 13.1: Term characterization in frequency spectrum (Salton and McGill, 1983, p.87).

13.3 Automatic indexing process

Salton presents the following blueprint for automatic indexing (Salton, 1989, p.304):

1. Identify the individual words occurring in the documents of collection.
2. Use a *stop list* of common function words (and, of, or, but, the, etc.) to delete from the text the high-frequency function words that are insufficiently specific to represent content.
3. Use an automatic *suffix-stripping* routine to reduce each remaining word to *word-stem* form; this reduces to a common form all words exhibiting the same stem (for example, analysis, analyzer, and analyzing are all reduced to stem ANALY).

4. For each remaining word stem T_j in document D_i, compute a weighting factor w_{ij} composed in part of the term frequency and in part of the inverse document-frequency factor for the term, for example

$$w_{ij} = tf_{ij} \cdot log(N/df_j).$$

5. Represent each document D_i by the set of word stems together with the corresponding weighting factors, that is,

$$D_i = (T_1, w_{i1}; T_2, w_{i2}; ....; T_t, w_{it}).$$

As proposed above, automatic indexing usually involves deletion of common high-frequency words, typically function words, by means of a stop list. InTEXT system, for instance, uses a large stoplist of some 2,000 words (Burnett *et al.*, 1996). In this system grammatical variants are combined using a modified Lovins algorithm (Lovins, 1968) which is a widely used stemming algorithm based on the longest match of a suffix. The Porter algorithm which uses multistep transformations is another popular stemming algorithm (Porter, 1980). Simple stemming algorithms based on the removal of the longest word endings matching any suffix on the suffix list are usually quite sufficient for defining stems in languages like English, which has a relatively simple inflectional system. Hull, for instance, found in his experiment no significant difference between using suffix stripping algorithms and using linguistically motivated morphological analysis (Hull, 1996). However, for the languages of more complicated morphology, such as Finnish (Koskenniemi, 1996) and French (Klavans *et al.*, 1997), morphological analysis may be necessary. Anyhow, stemming improves indexing performance since it renders possible to relate word variants.

The next step in Salton's blueprint was to weight the word stems using *TF*IDF* weighting. Many variants of this standard weighting scheme have been derived and used[1]. Robertson and Sparck Jones proposed one which takes into account the document length (Robertson and Sparck Jones, 1997; Robertson *et al.*, 1995). The following three sources of weighting data is used in this weighting scheme:

1. Collection frequency (IDF),
2. Term frequency (TF), and
3. Document length

The idea behind using **collection frequency** is that terms that occur in only a few documents tend to be more valuable than terms that occur in many documents. Collection frequency weight is defined for a term *t(i)* as

[1]cf. e.g., Salton and Buckley, 1988.

$$CFW(i) = logN - log(n(i))$$

where *n(i)* is the number of documents term *t(i)* occurs in, and *N* is the number of documents in the collection.

The idea behind using **term frequency** is that the more often a term occurs in a document, the more likely it is to be an important term for that document. The term frequency for a term *t(i)* in a document *d(j)* is

TF(i,j) = the number of occurrences of term t(i) in document d(j)

The idea behind using **document length** is that a term that occurs the same number of times in a short document and in a long document is likely to be more valuable index term for the short document. Length of a document *d(j)* is

DL(j) = the total of term occurrences in document d(j)

In the formula below, however, a normalized document length is used in which document length is divided by the length of an average document

NDL(j) = (DL(j))/(Average DL for all documents)

The Combined Weight for a term *t(i)* in a document *d(j)* is then

$$CW(i,j) = \frac{CFW(i) \cdot TF(i,j) \cdot (K1+1)}{K1 \cdot (1 - b + b \cdot (NDL(j)) + TF(i,j)}$$

where *K1* and *b* are tuning constants. *K1* modifies the extent of the influence of term frequency. The optimal value of this constant is determined through trials on the particular document collection. In TREC tests, the value *K1=2* was found to be effective. The tuning constant *b* modifies the effect of document length. It ranges between 0 and 1, and if *b=0*, it is assumed that documents are long because they are multitopic. If *b=1*, it is assumed that documents are long because they are repetitive. Thus, if *b* is set towards 1, it reduces the effect of term frequency on the ground that it is primarily attributable to verbosity. In TREC tests, the value *b=0.75* was found to be appropriate. This formula ensures that the effect of term frequency is not too strong, since doubling *TF* does not double the weight. For a term occurring once in a document of average length, the weight is just *CFW*. The overall score for a document *d(j)* is calculated by summing up the weights (CW(i,j)) of the query terms found in the document *d(j)*. Documents are ranked then in descending order according to their scores for presentation to the user. (Robertson and Sparck Jones, 1997)

So, usually weighting schemes are based solely on frequencies of words across the document collection, but in this thesis an attempt is made to combine evidence derived from word frequencies with evidence derived from linguistic analysis. An example of this kind of approach is provided by Smeaton *et al.* who used standard SMART-like term weighting in TREC-4, but also increased the weight of a term depending on whether the term occurred as a headnoun, modifying noun, verb, adjective, adverb or stopword (Smeaton *et al.*, 1996). InTEXT system, on the other hand, used the position of a phrase within the document (in heading, the first sentence of paragraph, etc.) as a weighting criterion in TREC-4 experiments (Burnett *et al.*, 1996). As mentioned earlier, one of the objects of this study is to detect whether the first and last sentences of a paragraph, on the one hand, and the first and last paragraphs of a section, on the other hand, have a special role in index term weighting.

13.4 Indexing by phrases

As discussed earlier, phrases may be used in indexing to improve precision, since phrases are often more specific than single word terms. According to Zhai *et al.* single words, as indexing units, may have two different kinds of problems (Zhai *et al.*, 1997, pp.347-348):

1. They may be misleading.

   In the context of lexical atoms[2], such as "hot dog", the contained single words do not carry their regular meanings and are thus very misleading if used as separate indexing terms;

2. They may be too general.

   For example, the individual words "junior" and "college" are not specific enough to distinguish "college junior" from "junior college".

[...] Based on these observations, we are inclined to propose the following two hypotheses:

1. The use of lexical atoms, such as "hot dog", to replace single words for indexing would increase both precision and recall;

2. The use of syntactic phrases, such as "junior college" to supplement single words would increase precision without hurting recall and using more such phrases results in greater improvement in precision.

These hypotheses were tested in TREC-5 NLP track by Zhai *et al.* (CLARIT^{TM} team[3]). Results suggested that exploiting lexical atoms to replace single words that form the lexical atoms increases the average precision consistently albeit slightly. Supplementing single words by various combination of syntactic phrases, on the other hand, resulted in a consistent and significant

[2]Evans and Zhai describe **lexical atoms** as follows: A lexical atom is a semantically coherent phrase unit. Lexical atoms may be found among proper names, idioms, and many noun-noun compounds. Usually they are two-word phrases, but sometimes they can consist of three or even more words, as in the case of proper names and technical terms. Examples of lexical atoms (in general English) are "hot dog", "tear gas", "part of speech", and "von Neumann" (Evans and Zhai, 1996, p.20). (Footnote by Lahtinen)

[3]CLARIT is a registered trademark of CLARITECH Corporation, and it is an acronym for *Computational-Linguistic Approaches to Indexing and Retrieval Text*. This system has been developed in Carnegie Mellon University (Laboratory for Computational Linguistics, CMU Pittsburgh).

improvement in retrieval performance. It was observed, however, that adding phrases was to advantage to some queries, but to disadvantage for others. Phrases were most useful when they occurred in the queries as well as in the documents. (Zhai *et al.*, 1997)

Most information retrieval systems use single words for indexing, but often indexing language is supplemented with phrases obtained using simple statistical methods. It might be assumed that phrases obtained using linguistic analysis would describe the content of a document more accurately than phrases discovered using simple statistical approaches. In the context of TREC, however, linguistic methods have not proven to be superior to statistical methods. As mentioned earlier, the results of TREC experiments suggested some advantage from using compound terms, but on the other hand there was no gain reported from using linguistically motivated, as opposed to adjacency defined compound terms (Sparck Jones, 1995).

A simple and classic phrase constructing method used, for instance, in SMART is to create index terms from all adjacent pairs of stemmed non-stopwords[4]. The final set of phrases used in indexing is composed of frequently occurring pairs of words, and these terms are weighted with the same scheme as single terms (Buckley *et al.*, 1995). The Xerox team compared this simple SMART method with light parsing method in TREC-5, and found that light parsing method was slightly better than the simple method, but at a cost of longer time of preprocessing. Hull *et al.* concluded (Hull *et al.*, 1997, p.177): Nonetheless, we are optimistic that this approach will prove useful in the long run for a number of reasons:

1. For non-English languages more work is being done on linguistic analysis than on information retrieval. This implies that morphological analysers for these languages may largely outperform simple stemming routines that have undergone the same maturation time as English stemming routines.

2. As machine become more powerful, the preprocessing times will continue to fall, making more complicated text analysis more economically feasible.

3. Robust parsing is progressing in the variety and accuracy of structures recognized.

Various more or less linguistically motivated techniques have been developed for discovering phrases as well as for indexing documents by them. Justeson and Katz proposed a simple algorithm for identifying technical terminology[5] in running text (Justeson and Katz, 1995). Typical grammatical properties of technical terms were detected using four dictionaries of technical terminology, and it was found that the majority of technical terms consist of more than one word. Typically these words are nouns, adjectives, and occasionally prepositions. The algorithm consists of two constraints applied to word strings in text: candidate strings must have the frequency

[4]**Mutual information** is another classic example of statistical method used to discover compound terms (Church and Hanks, 1990).

[5]Technical terms are well-defined lexical items belonging to special subject languages, and they are often appropriate index terms as well.

of two or more in the text, and their grammatical structure is specified by the regular expression $((A|N)^+|((A|N)^*(NP)^?)(A|N)^*)N$, where *A* in an adjective, *N* is a lexical noun, and *P* is a preposition.

The stream-based information retrieval model presented by Strzalkowski *et al.* is a good example of natural language information retrieval system evaluated in TREC experiments (Strzalkowski *et al.*, 1998). In this system, four different indexing methods (streams) are fused together:

1. Stem stream,
2. Phrase stream,
3. Normalized phrase stream, and
4. Proper name stream

Streams which are built using a combination of various indexing approaches, term extracting, weighting strategies, and even different search engines[6] perform in parallel. Term extraction includes the following steps:

1. Elimination of stop words.
2. Morphological stemming using a lexicon-based stemmer.
3. Phrase extraction using various shallow text processing techniques, such as part-of-speech tagging, phrase boundary detection, and word co-occurrence metrics.
4. Phrase normalization in which "Head+Modifier" pairs are identified in order to reduce syntactic variants to a common "concept". For example, "information retrieval", "retrieval of information", "retrieve more information", and "information that is retrieved" are all reduced to "retrieve+information" pair. Phrase normalization is intended to capture semantic uniformity across the variety of surface forms.
5. Proper name extraction in which people names and titles, location names, organization names, and such are identified.

Two types of phrases are then used for indexing. Unnormalized phrases are collected from part-of-speech tagged text using the following three patterns:

1. a sequence of modifiers (e.g., adjective and participles) followed by at least one noun, such as "air traffic control system",
2. proper noun sequences modifying a noun, such as "china trade", and
3. proper noun sequences, such as "Warren Commission"

[6] SMART, INQUERY, and NIST's Prise (Harman and Candela, 1989).

Phrase length is limited to seven words maximum. Normalized phrases, on the other hand, are derived through the following steps:

1. Part-of-speech tagging using Brill's rule based tagger (Brill, 1992).

2. Lexicon-based word normalization (extended "stemming").

3. Syntactic analysis with TTP (Tagged Text Parser) which is a full grammar parser based on Linguistic String Grammar (Sager, 1981). Each sentence is represented as a regularized parse tree which reflects the sentence's logical predicate-argument structure. Logical subjects and objects are identified in both passive and active sentences, and noun phrases are organized around their head elements.

4. Extraction of head+modifier pairs. The head is a central element of a phrase (e.g., main verb or main noun), whereas the modifier is one of the adjunct arguments of the head. The following types of pairs are extracted:

   - a head noun and and its left adjective or noun adjunct,
   - a head noun and the head of its right adjunct,
   - the main verb of a clause and the head of its object phrase, and
   - the head of the subject phrase and the main verb

5. Corpus-based disambiguation of long noun phrases. Nominal strings consisting of three or more words of which at least two are nouns are decomposed using distributional statistics of phrases. If a word pair occurs frequently in the corpus, it may be included in the index.

The streams were then weighted using standard SMART weighting schemes. In the comparison of the streams, stem stream was proven to outperform the other streams, as shown in *Figure 13.2*. According to Strzalkowski *et al.*, one possible explanation for the weak performance of normalized phrases was the quality of parse structures generated by the Tagged Text Parser. Another explanation could be the pair-based representation of phrases. For instance, the phrase "former Soviet president" is broken into two pairs "former president" and "Soviet president", even though the original phrase could be more appropriate for indexing.

Each stream ranks documents in order of relevance, and the final retrieval result is obtained by merging these rankings. The final ranking is derived by calculating the combined scores using the following formula:

$$finalscore(d) = \sum_{i=1...N} A(i) \cdot score(i)(d) \cdot prec\{ranks(i) | rank(i,d) \in ranks(i)\}$$

where *N* is the number of streams; *A(i)* is the stream coefficient; and *score(i)(d)* is the normalized score of the document against the query within the stream *i*; *prec(ranks(i))* is the precision estimate from the precision distribution table for stream *i*; and *rank(i,d)* is the rank of document *d*

RUNS	*short queries*	*long queries*
Stems	0.1070	0.2684
Phrases	0.0846	0.2541
H+M Pairs	0.0405	0.1787
Names	0.0648	0.0753

Figure 13.2: How different streams perform relative to one another (11-pt average precision) (Strzalkowski *et al.*, 1998).

in stream *i*. The coefficients of different streams (*A(i)*) are obtained empirically to maximize the performance of different combinations of streams. A better performing stream has a higher coefficient. *Figure 13.3* summarizes the precision improvements over the results obtained using only the stem stream. In this experiment, the combination of the stem stream and the unnormalized phrase stream outperformed the other fusions. (Strzalkowski *et al.*, 1998)

Streams merged	*short queries % change*	*long queries % change*
All streams	+5.4	+20.94
Stems+Phrases+Pairs	+6.6	+22.85
Stems+Phrases	+7.0	+24.94
Stems+Pairs	+2.2	+15.27
Stems+Names	+0.6	+2.59

Figure 13.3: Precision improvements over stem-only retrieval based on TREC-5 data (Strzalkowski *et al.*, 1998).

Different phrase normalization techniques have been developed in order to capture semantic uniformity across the variety of surface forms[7]. The CLARIT system, for instance, has used an automatically generated first order thesaurus for this task. The CLARIT system applies various

[7]Phrase normalization techniques usually normalize or ignore syntactic ambiguity. In some alternative approaches syntactic ambiguities have been incorporated into indexing by using different kinds of structured representations for matching between a query and documents. Smeaton *et al.*, however, reported somewhat discouraging results from an experiment of approach of this kind evaluated in TREC-3 (Smeaton *et al.*, 1995). The TINA project at Siemens, Münich (Schwarz, 1990) and the COP project at Pittsburgh (Metzler and Haas, 1989) are other examples of approach of this kind.

NLP techniques in order to identify candidate phrases which are then compared with the thesaurus. For indexing the phrases of a thesaurus are used, that is, the different surface forms are replaced by the forms found in the thesaurus (Evans *et al.*, 1991). Another kind of method applied to French, is described by Klavans *et al.* (Klavans *et al.*, 1997). Morphosyntactically related terms are conflated by generating morphological variants of words, and by identifying the different syntactic variants of phrases by means of rules describing different variants of multi-word terms, and by means of meta-rules that recognize the related variants. In index these variants are linked together.

Phrases are often weighted with the same scheme as single terms. Strzalkowski and Carballo, however, argue that a standard tf.idf weighting scheme (and we suspect any other uniform scheme based on frequencies) is inappropriate for mixed term sets (ordinary concepts, proper names, phrases) because:

1. It favors terms that occur fairly frequently in a document, which supports only general-type queries (e.g., "all you know about 'star wars'"). Such queries are not typical in TREC.

2. It attaches low weights to infrequent, highly specific terms, such as names and phrases, whose only occurrences in a document often decide of relevance. Note that such terms cannot be reliably distinguished using their distribution in the database as the sole factor, and therefore syntactic and lexical information is required.

3. It does not address the problem of inter-term dependencies arising when phrasal terms and their component single-word terms are all included in a document representation, i.e., launch+satellite and satellite are not independent, and it is unclear whether they should be counted as two terms (Strzalkowski and Carballo, 1996, p.253).

So, Strzalkowski and Carballo weighted the phrases more heavily by using the following formula:

$$weight(T_i) = (C_1 \cdot log(tf) + C_2 \cdot \alpha(N, i)) \cdot idf$$

where C_1 and C_2 are sufficiently large constants, and $\alpha(N, i)$ is 1 for $i < N$ and 0 otherwise. By the $\alpha(N, i)$ factor it is possible to weight the top *N* highest-idf matching terms more heavily. In TREC-3 experiments *N* was set to 15 or 20.

In this thesis, multi-word terms are weighted by the same weighting schemes than single-word terms. The problem of the low frequencies of multi-word terms is solved by using evidence from linguistic analysis. This issue will be discussed in *Parts V and VI*.

13.5 Query expansion and relevance feedback

Although the empirical part of the thesis is not concerned with query expansion and relevance feedback, this issue will be briefly discussed in this section, because of its great importance to efficient information retrieval. Query expansion techniques and relevance feedback techniques are often concerned with viewing some document descriptors to users. One of the goals of this thesis is to develop an automatic indexer that produces index terms that are relevant to view by users. This kind of indexer could then be a useful tool for query expansion techniques and relevance feedback techniques.

In **query expansion technique**, a query is expanded manually or automatically. Query expansion may be based on user's feedback about retrieved documents (**relevance feedback**), or it is also possible that synonyms and other related words or phrases are added to the query automatically by using thesauruses[8]. The use of query expansion and relevance feedback techniques has proven to improve retrieval performance greatly.

Query expansion is often limited to adding, deleting or re-weighting of query terms, but some approaches use full text expansion in which entire sentences, paragraphs, or other passages from documents are added to query. The relevant sentences can be found, for instance, by using information extraction techniques, and the relevant passages can be found by using automatic text summarizers or passage retrieval techniques. Strzalkowski *et al.* used FASTUS information extraction system (cf. *Section 9.1*) for extraction-based query expansion and GE Summarizer-Tool for summarization-based query expansion. In their framework the user decided whether a given summary was added to the query. (Strzalkowski *et al.*, 1998).

New terms for automatic query expansion can be discovered, for example, by taking frequent or highest weighted terms from the highest ranked documents or highest ranked passages. Different global and local techniques have been proposed, for example, by SMART and INQUERY teams (Buckley *et al.*, 1996; Buckley *et al.*, 1998; Xu and Croft, 1996). In automatic query expansion, the system provides pseudo-relevance-feedback by ranking the documents, but in interactive query expansion users may view documents, passages, term lists, or graphical representations produced by visualization tools. In interactive retrieval process users are thus able to contribute to the retrieval process, for example, by identifying relevant texts, by choosing new terms and deleting old ones, or by re-weighting the terms.

13.6 Automatic construction of hypertexts

The hypermedia approach provides new challenges to the traditional information retrieval systems. The World Wide Web, for instance, has recently become a very important source of information, which makes necessary to develop quite new search environments that support the wide variety of information-seeking behaviors. The smooth transitions between browsing and querying, and between navigation and mediated search, require the use of some appropriate technique, such as the index linking model discussed in *Chapter 10*.

[8]WordNet, for example, has been used for this kind of automatic query expansion in some experiments, but the results have not been particularly encouraging (Voorhees, 1994; Smeaton *et al.*, 1996).

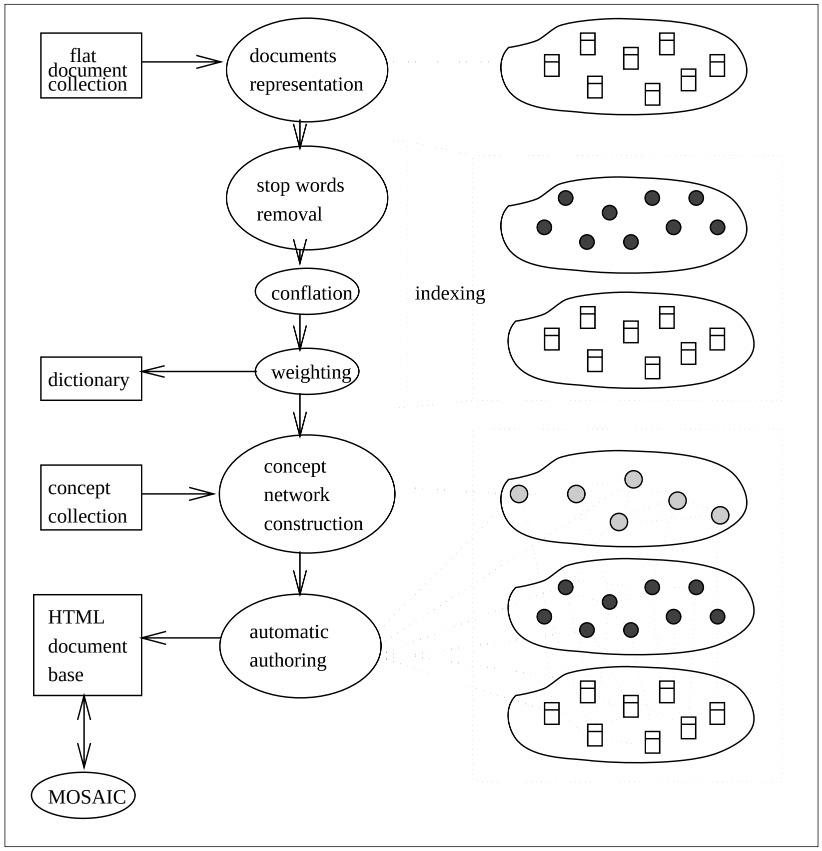

Figure 13.4: The automatic authoring process (Agosti *et al.*, 1995).

This section will not discuss, however, the development of hypermedia environments, but it will present an example of a tool which automatically builds up a hypertext from a document collection, proposed by Agosti *et al.* (Agosti *et al.*, 1995). TACHIR[9] is a tool for the automatic construction of hypertexts for information retrieval. In this framework, the structure of the hypertext reflects a three level conceptual model which includes the concept level, the index term level, and the document level. A network of links is constructed:

- within each collection: documents (D-D links), terms (T-T links), and concepts (C-C links).
- between a pair of collections: documents-terms (D-T links), terms-concepts (T-C links).

Automatic indexing is based on the traditional techniques: term extraction, stop terms removal, conflation, and weighting. Relationships between index terms (T-T links) and between documents (D-D links) are determined by using statistical methods. The concept network is constructed by means of a thesaurus (a concept collection of the domain). The thesaurus is also used for connecting index terms with the concept network (Agosti and Marchetti, 1992). The automatic authoring process is shown in *Figure 13.4*. The input is a large collection of multimedia documents, and

[9]Tool for the Automatic Construction of Hypermedia for Information Retrieval

the output is a collection of hypermedia documents, where the links between the terms and concepts construct a semantic network. Querying and browsing among documents, index terms, and concepts is then possible in the context of some World Wide Web interface. (Agosti *et al.*, 1995)

As discussed earlier, one of the goals of this thesis is to develop an automatic indexer that produces index terms that are relevant to view by users. The automatic authoring process described above, could be a task that profits from this kind of indexer.

Chapter 14

Summary

Part III discussed some elementary aspects of information retrieval, focusing on natural language processing in particular. Many issues have been discussed just briefly, and a number of important issues of information retrieval have been totally ignored, document classification methods, and machine learning techniques (neural networks, genetic algorithms), among other things. Anyhow, some main points of natural language information retrieval have been brought up, and this part will be concluded by reviewing the summary of Strzalkowski and Sparck Jones concerning natural language processing track at the fifth Text REtrieval Conference (Strzalkowski and Sparck Jones, 1997).

Figure 14.1 presents their "rather subjective view" of what might be the potential of natural language processing techniques for improving retrieval precision (Strzalkowski and Sparck Jones, 1997). This estimation was discussed at the NLP track workshop, and it summarizes the results of the natural language processing track in the fifth Text REtrieval Conference. Five teams participated in this track: GE/Rutgers/NYU/Lockheed Martin, Xerox, Mitre, Claritech, and ISS Singapore. Results were submitted by the first four teams which all outperformed the SMART statistical baseline system. Most of the techniques of *Figure 14.1* have been discussed above, but a brief recapitulation may be necessary:

- In their **full text query expansion** experiments, Strzalkowski *et al.* (1998) attached sentences and abstracts to queries.

- **Term-based query expansion** is the more usual query expansion technique.

- Mitre team used a preprocessor for **deleting extraneous text from queries**. A pattern-matching program removed phrases such as "a relevant document will contain" (Burger *et al.*, 1997).

- Strzalkowski *et al.* used occurrences of **hyphenated phrases** in text as a guide to extracting other multi-word terms for indexing. The hyphenated words, such as `alien-smuggle` and `per-capita`, were collected and their non-hyphenated occurrences were replaced by

NL technique	class	% change precision
Full-text query expansion	query build	40 to ???
Term-based query expansion	query build	15 to 25
deleting extraneous text from queries	query build	0 to 5
hyphenated phrases	phrases	-15
word bi-grams	phrases	5 to 10
extended bi-grams (windows)	phrases	-5
FSA phrases (noun groups)	phrases	7 to 25
Head+Modifier Pairs (full parsing)	phrases	2 to 15
proper names	concepts	1 to 3
concept tagging for indexing	concepts	0 to ???
concept tagging for re-ranking	concepts	0 to 3
stylistics	discourse	0 to ???
lexical normalization	stemming	5 to 8

Figure 14.1: Natural language processing results analysis (Strzalkowski and Sparck Jones, 1997).

the hyphenated forms. This technique, however, affected precision loss. (Strzalkowski *et al.*, 1997)

- As discussed earlier, Xerox team compared the simple phrase constructing method of **word bi-grams** with the light parsing method and found that the light parsing method was slightly better than the simple method.

- Xerox team used also the **extended bi-gram** method in which instead of just indexing by contiguous word pairs, any word pair within a window of ten non-stop words is used for indexing. The results, however, were somewhat discouraging.

- **FSA phrases** are phrases discovered by full syntactic analysis. The phrase indexing method of CLARIT team is an example of this approach.

- The normalized phrase stream of the stream-based information retrieval model is an example of the **head+modifier pair** technique.

- **Proper name recognition** was part of the stream-based information retrieval model as well.
- Strzalkowski *et al.* made some experiments in **concept tagging**. Words and phrases related to the concept 'foreign', such as `foreign`, `other countries`, `international`, as well as the names of foreign countries and cities, were tagged by a special 'foreign' token. This technique was meant to improve the retrieval performance in the cases that queries involved references to foreign countries. In the experiments, ten queries out of 45 were affected, and nine out of these ten showed improvement and only one showed a modest performance loss. (Strzalkowski *et al.*, 1997)
- Another experiment reported by Strzalkowski *et al.* was concerned with the connection between **stylistic variation** and relevance of retrieved documents (Strzalkowski *et al.*, 1997). It has been found, for instance, that highly ranked documents are longer than other documents - regardless of their actual relevance; and moreover, relevant documents tend to be textually, syntactically, and lexically more complex than non-relevant documents (Karlgren, 1996). Complexity can be determined by measuring different types of simple stylistic items, such as average word length, type/token ratios, pronoun counts, digit counts, and average sentence length (Karlgren and Cutting, 1994). Since complexity is related to relevance, these metrics were used to improve precision in TREC-5 experiments. It was found, however, that it is quite difficult to determine appropriate rules for this task. If rules to distinguish relevant documents from non-relevant were found for the entire corpus, these rules might degrade performance in the case of some individual queries. The conclusion was that to make use of stylistic variation for relevance grading, a query typology is needed; each query must be identified for style preferences.[1]
- All systems use some kind of **lexical normalization**.

To sum up, the results of the natural language processing track at the fifth Text REtrieval Conference suggest that query expansion techniques and phrase indexing techniques are the most promising approaches. On the other hand, as the question marks of *Figure 14.1* indicate, there are still questions to be answered in the field of natural language information retrieval as well as in the field of information retrieval in general. For instance, hypermedia is still a relatively new area in information retrieval tradition. As discussed earlier, in the context of TREC, linguistic methods have not proven to be superior to statistical methods. Sparck Jones summarizes the experiences of TREC-6 as follows (Sparck Jones, 1998, p.B-7): Thus term weighting, query expansion and so forth are valuable, and in automatic searching quite simple strategies can be as effective as more elaborated ones, so e.g., sophisticated natural language processing is not especially helpful. This has led to some convergence on what may be called the *generic tf*idf* paradigm with relevance feedback refinement. Time will show what the impact of natural language processing on information retrieval will be.

[1]For an exhaustive presentation of stylistic experiments for information retrieval, see Karlgren, 2000.

Automatic indexing techniques described above are designed for information retrieval systems, but many techniques could be adapted as well to generating semi-automatically back-of-the-book indexes. An example of a project in which linguistic analysis and phrase normalization is used for this kind of task is presented by Karetnyk, Karlsson, and Smart (Karetnyk, Karlsson, and Smart, 1991). An important difference between book indexes and information retrieval system indexes is related to the issue of the dual function of index terms: representation and discrimination (cf. *Section 13.1*). In the case of information retrieval systems, both functions are important, but in the case of book indexes, the focus is on the representation. The index term corpus of this thesis is based on book indexes, but the developed automatic indexer can be used to generate information retrieval system indexes as well as back-of-the-book indexes.

Part IV

Materials and methods

This part will discuss the materials and methods of the empirical part of the thesis:

- *Chapter 15* will describe the process of constructing the **index term corpus** which is a linguistically analysed text collection where the index terms were manually marked up by a research aide.
- *Chapter 16* will describe the process of linguistic annotation of the index term corpus.
- *Chapter 17* will describe the process of exploring the linguistic features of index terms by using the index term corpus as training material. A new weighting scheme based on the tag sets of words will be introduced.
- *Chapter 18* will describe the methods that measure the two kinds of burstiness of term candidates: **within-document burstiness** (*Section 18.1*) and **document-level burstiness** (*Section 18.2*).
- *Chapter 19* will introduce a new method that combines evidence based on linguistic analysis and evidence based on document-level burstiness.

Chapter 15

Index term corpus

The empirical study of this thesis is based on an index term corpus. It is a collection of texts where some information concerning index terms was encoded, both manually and automatically. The index terms were marked up manually into texts. Linguistic analysis was provided automatically by a parser, and the textual location of words was analysed and marked up automatically, too. The textual location was encoded by tags that indicate if the word is in a title or subtitle, or in the first paragraph after or before a title or a subtitle, or in the first or last sentence of the paragraph.

When we have all this information in the same corpus, it is possible to calculate estimated index-term-likeness probabilities of a kind to words and phrases. When we have calculated these probabilities, we can use them with any new text to estimate the index-term-likeness of the words and phrases of the text, that is, we can index texts automatically.

The core of the index term corpus in this study consisted of five texts that were concerned with sociology and philosophy. Four were essays[1] and the fifth was a longer document[2]. All texts had manually generated indexes. A research aide identified and marked up the index terms for each document page using the document index, that is, she marked up the closest equivalents of index terms found in the book indexes. The quality of the book indexes, and the consistency of the human indexers of the book indexes was not evaluated in this study. Manual indexing is a challenging task, and defining the index terms of the text demands many more or less subjective decisions. By my subjective account, however, the quality of the book indexes was good. Naturally, the results of this study might have been different if the book indexes were different. Automatic indexing is consistent, in a way, and it would have been possible to question some decisions of the human indexers of the book indexes. This kind of subjective evaluation was outside the scope of the study, though.

After marking up the index terms manually, the corpus was analysed linguistically by a parser and divided into two parts: **a training corpus**, consisting of two essays and 57 pages from the long text, and **a test corpus**, consisting of the remaining two essays and 16 pages. The idea was

[1] Together, these essays by different writers form sample ECV from the British National Corpus.

[2] A 73-page document taken from the Bank of English.

to include into both corpora texts that were concerned with somewhat different issues, and that are from different writers. The features of index terms were explored using the training corpus and the test corpus was used to test whether the results could be generalized beyond the context of the training corpus. In addition to this core corpus, another smaller test corpus was constructed of ten articles of **The Grolier Encyclopedia**. In this case, the research aide identified and marked up the index terms without any previously manually generated indexes. She also ranked the index terms of Grolier on a scale from one to three:

1. **Passing concepts and proper names**. The least important index terms of the article. The subjects concerning these index terms are not actually discussed in the article, but someone might still be interested in reading what is said about these index terms.

2. **Subtopics**. The index terms of some importance. The subjects concerning these index terms are not the main topics of the article, but something essential is said about them. Typically these index terms are subtopics and important names of the article.

3. **Main topics**. The most important index terms of the article. These index terms are the main topics and the most important names of the article.

Altogether, the index term corpus included 64,996 words in three parts that in this thesis will be referred to as:

- **training corpus** (three texts, 38,138 words),

- **test corpus** (three texts, 17,392 words)[3], and

- **Grolier** (ten articles, 9,466 words)

In the book indexes, terms are usually simple noun phrases, but in the texts the content of the noun phrases may be expressed by using verbs, adjectives, or even clauses. For instance, the index term of the book index `oppression of woman` may be expressed in the text by the clause `women are oppressed`[4]. In such cases, the research aide was instructed to mark up the closest equivalents. Sometimes index terms are nested, e.g., the expression `critical ethnography` found in the text includes two index terms of the book index, `critical ethnography` and `ethnography`. In these cases, both index terms were marked up.

In the training corpus of 38,138 words, a single word was marked up as an index term 2,145 times. A bigram was considered as an index term 1,183 times, and terms with more than two words had a frequency of 309. The most typical index term patterns found were simple noun

[3]These three texts are referred to as **test corpus**, because they form the primary test corpus. In the case of the smaller test corpus (**Grolier**) the research aide did not use any previously generated index.

[4]The text of the training corpus included two index terms of this pattern: `women are oppressed` and `women are controlled`.

phrases, for instance, `capitalism`, `biological determinism` and `methodology of philosophy`. Not surprisingly, most of the proper nouns in the text were included in the index.

In addition, there were other texts with no index term mark-up to supplement the index term corpus when applying frequency-based term weighting schemes:

- 767 articles from The Grolier Encyclopedia (1,576,908 words),
- 11 articles from The Times (8,152 words),
- a technical manual by Xerox (ScanWorX User's Guide, 38,852 words),
- a fragment of a novel by Edgar Rice Burroughs (At the Earth's Core, 3,879 words), and
- fifteen other texts that were concerned with history, linguistics, sociology, and politics, among other things.[5] (574,433 words)

The total corpus consisted of 2,267,220 words in 810 documents[6]:

- 64,996 words in texts with index term mark-up and
- 2,202,224 words in texts with no index term mark-up

[5] The texts were again taken from the British National Corpus (A6S, APD, CGF, CM6, CMN, CMR, CRF, EDH, F9K, F9V, FAC, GV5, GVA, H9F, and J2K).

[6] Grolier articles were considered as documents when the *IDF* values were calculated.

Chapter 16

Linguistic annotation

The linguistic annotation of all the 810 documents was done with a robust rule-based dependency parser, The Conexor Functional Dependency Grammar[1] (FDG, Tapanainen and Järvinen, 1997), originally developed at the Research Unit for Multilingual Language Technology at the University of Helsinki. FDG is related to the Constraint Grammar framework (Karlsson *et al.*, 1995), but it creates links between the elements of the sentence in addition to the shallow representation similar to the English Constraint Grammar (ENGCG) (Voutilainen, 1994; Järvinen, 1994). The parser also applies an ENGTWOL-style lexicon (Heikkilä, 1995; Koskenniemi, 1983), and a morphological disambiguator designed by Voutilainen (Voutilainen, 1995).

The following example has been taken from the training corpus:

```
"<Marx>"         "Marx" <Proper> N NOM SG @SUBJ subj:>2 </+INDEX-TERM>
"<suggested>"    "suggest" V PAST VFIN @+FMAINV #2 main:>0
```

This is the analysis of the clause `Marx suggested`. In the first place, each word is annotated with a base form, which is a useful feature for counting word frequencies (`"Marx"` and `"suggest"`). *TF*IDF* values were calculated using these base forms. Then, the tag list[2] carries information about the linguistic features of the individual words, e.g., the word `Marx` is

- a proper noun (<Proper> tag added by the parser),
- a noun (`N` tag added by the parser),
- in nominative case (`NOM` tag added by the parser),
- in singular number (`SG` tag added by the parser), and
- a subject (`@SUBJ` tag added by the parser)

The word `suggested` is

[1] For a demo, see: http://www.conexor.fi/

[2] For ENGTWOL and ENGCG tag descriptions, cf. Voutilainen *et al.*, 1992.

- a verb (V tag added by the parser),
- in past tense (PAST tag added by the parser),
- a finite verb (VFIN tag added by the parser), and
- a finite main predicator (@+FMAINV tag added by the parser)

The tag list carries information about the dependency links between the words as well, e.g., Marx is the subject of the sentence (subj:>2 added by the parser), and it has a link to the main verb suggest (#2 and main:>0 added by the parser). In addition to the tags annotated by the parser the following tags were added automatically by another program:

- <HEADLINE> tag if a word occurred in a title or a subtitle.
- <PAR1> tag if a word occurred in the first paragraph after a title or a subtitle.
- <PARn> tag if a word occurred in the first paragraph before a title or a subtitle.
- <SEN1> tag if a word occurred in the first sentence of the paragraph.
- <SENn> tag if a word occurred in the last sentence of the paragraph.

Furthermore, a research aide manually marked up all index terms by <+INDEX-TERM> tags, which were added to the automatically generated tag lists. The slash character (/) in the </+INDEX-TERM> tag indicates that the current word is the last word of the term. The <+INDEX-TERM> tag without the slash character indicates that the current word is not the last word of the term. So, if the tag list of the preceding word does not include the <+INDEX-TERM> tag and the tag list of the current word includes the </+INDEX-TERM> tag, it means that the current word is a single-word term. On the other hand, if the tag list of the preceding word does not include the <+INDEX-TERM> tag and the tag list of the current word does, it means that the current word is the first word of a multi-word term. If the tag list of the preceding word, however, does include the <+INDEX-TERM> tag as well as the tag list of the current word, it means that the current word is neither the first nor the last word of a multi-word term. If more than one index terms are nested, the different terms are distinguished by appending numbers into term tags, e.g.,

```
"<of>"           "of" PREP @<NOM-OF #5 mod:>4 <HEADLINE>
"<critical>"     "critical" A ABS @A> attr:>7 <HEADLINE> <+INDEX-TERM>
"<ethnography>" "ethnography" <-Indef> N NOM SG @<P #7 pcomp:>5 <HEADLINE>
                                               </+INDEX-TERM> </+INDEX-TERM-1>
```

In this example, critical (<+INDEX-TERM>) is the first word of the index term critical ethnography, and ethnography (</+INDEX-TERM>) is its last word. Furthermore, ethnography (</+INDEX-TERM-1>) is the first and the last word of the index term ethnography. This example comes from one of the subtitles of the training corpus, and for that reason the tag lists include the <HEADLINE> tag. *Appendix 1* presents a longer excerpt from the tagged training corpus.

Chapter 17

Term weights based on linguistic tags

The combination of linguistic annotation and index term mark-up made it possible to examine the linguistic structure of index terms. We occasionally find index terms consisting of verbs (e.g., understand), adverbs (e.g., historically), adjectives (e.g., empirical) and even clauses (e.g., women are oppressed), but the great majority of index terms were noun phrases. So, with the help of the linguistic analysis of the index term corpus, it becomes possible to try to identify index terms on the basis of their linguistic properties. A simple way to explore the typical features of index terms is to see how often each tag is included in the tag list of index terms. For example, the frequency of the N tag was 10,111 in the training corpus, and it appeared 1,987 times in the tag list of a single-word term. The training corpus gave an estimated index term probability of 0.197 (1,987/10,111) to the N tag, and similarly a probability of 0.755 to the <Proper> tag, a probability of 0.145 to the subj:> tag (subject of the sentence), a probability of 0.082 to the obj:> tag (object of the sentence), and so on. The probabilities for the different tags are obviously not independent, so they were calculated for all of the relevant **tag combinations**, i.e. for the combinations that in the training corpus distinguish index terms from non-terms in the most appropriate way. For instance, the tag combinations of the words Marx and suggested have the following index term probabilities:

```
"<Marx> "Marx" <Proper> N NOM SG @SUBJ subj:>2 </+INDEX-TERM> PROBABILITY:0.977
"<suggested>"  "suggest" V PAST VFIN @+FMAINV #2 main:>0 PROBABILITY:0.005
```

So, in the training corpus the tag combination

```
<Proper> N NOM SG @SUBJ subj:>
```

was a tag combination of a single-word term 42 times out of 43 (42/43=0.977). Exploration of relevant tag combinations and their estimated index term probabilities can be seen as a kind of decision tree, as *Figure 17.1* illustrates. On the first level of the decision tree, the words of the training corpus were divided into four groups: nouns, adjectives, adverbs, and verbs. On the second level, the nouns were divided into two groups: proper nouns and common nouns.

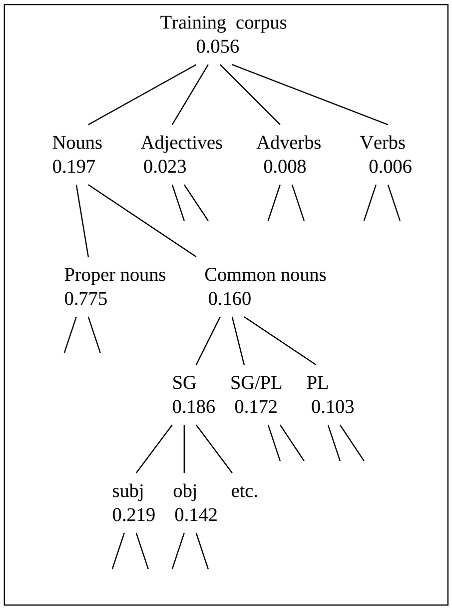

Figure 17.1: Exploration of relevant tag combinations and their estimated index term probabilities (single-word terms).

On the next level the common nouns, for example, were divided into three groups: singular, singular/plural[1] and plural nouns. The singular common nouns were divided according to their syntactic function. The singular common nouns as subjects were further divided according to their lexical features (e.g., endings), and so on. The leaf level of a tag combination was reached, when no more division was applied. A same kind of process of dividing words into smaller groups was applied to all nouns, adjectives, adverbs, and verbs. On the each level of the decision tree, the estimated index term probability values were calculated from the training corpus. For example, common nouns in singular have an index term probability of 0.219 if their syntactic function is subject (tag combination `N SG subj:>`). This tag combination, however, is not yet a leaf of the tree, but it is processed further by taking into account the lexical features of words (e.g., endings: `<DER:ism>` tag, `<DER:al>` tag, etc.) and the location of words (`<HEADLINE>` tag, `<PAR1>` tag, `<SEN1>` tag, etc.). The tag combinations of the leaves are the index term patterns that are then identified by the automatic indexer. Each pattern has an estimated index term probability that is calculated from the training corpus.

The multi-word terms were studied in the same way as the single-word terms. The patterns of index terms were explored by using the training corpus, and the index term probabilities of the different tag combinations were calculated. The tag combination of the following example (`autonomous person`), for instance, occurred 233 times in the training corpus, and of these, it was a tag combination of an index term 40 times. So, the noun phrase `autonomous person` is assigned the index term probability of 0.172 (40/233) in this context:

[1] The parser does not disambiguate some nouns in this respect, e.g., nouns like people and data.

```
"<An>"          "an" <*> <Indef> DET CENTRAL ART SG @DN> det:>3
"<autonomous>" "autonomous" A ABS @A> attr:>3 <+INDEX-TERM> 1/2-PROBABILITY:0.172
"<person>" "person" N NOM SG @SUBJ #3 subj:>4 </+INDEX-TERM> 2/2-PROBABILITY:0.172
"<is>"          "be" V PRES SG3 VFIN @+FMAINV #4 main:>0
```

The word `autonomous` is an adjective (`A`) and a premodifier (`attr:>`), and the head (`person`) is a noun (`N`) and a subject (`subj:>`).

Altogether, 89 different tag combinations were considered as relevant term patterns, and index term probabilities were calculated for these combinations based on the training corpus. The process of selecting the relevant term patterns was based on the grammatical logic, on one hand, and on the empirical evidence, on the other hand. For instance, as *Figure 17.1* illustrates, the first level of the decision tree includes four groups of parts of speech (nouns, adjectives, adverbs, and verbs). This division was based on grammar, in the first place, not on the empirical evidence. The empirical evidence, however, guided the process of selecting the most relevant, i.e., the most distinguishing grammatical features on the each level of the decision tree. For example, in the case of common nouns the number (SG, SG/PL, and PL) distinguished index terms from non-terms quite well, but in the case of proper nouns it did not. Consequently, the number was included in the decision tree of common nouns, but it was not included in the decision tree of proper nouns.

Towards the leaves of the decision tree, the number of occurrences of tag combinations became smaller and smaller, and the empirical evidence became more and more unreliable. If a certain tag combination had too few occurrences in the training corpus, it was not considered as a valid term pattern. So, to keep the term patterns robust, i.e., to avoid overtraining them, it was necessary to use relatively common grammatical features and to restrict the depth of the tree. To sum up, empirical evidence from the training corpus, together with grammatical logic and common sense provided the basis to build the decision tree and to select the 89 term patterns. *Section 21.1* will describe the selected term patterns in more detail.

The automatic indexer uses the index term probabilities of the patterns for weighting representations of the index term patterns. In this thesis, the weights based on tag combinations (i.e., the index term probabilities of the patterns) are referred to as **tag weights** (*TW*). The automatic indexer sums up all the tag weights of each word or phrase (**summed tag weights** *STW*) in a certain document and combines evidence from tag weights with other evidence, when it weights the words and phrases of a running text. The following matrix includes ten examples from the test corpus:

TERM CANDIDATE	TF	MAX-TW	STW
abandon	1	0.019	0.019
ability	2	0.394	0.589
ability of neo-capitalism	1	0.037	0.037
abortion-decision	1	0.407	0.407
Abraham	1	0.810	0.810
absent	1	0.093	0.093
abstract	5	0.067	0.180
abstract conception	2	0.268	0.536
abstract conception of justice	1	0.121	0.121
abstract labour	2	0.267	0.534

`TF` is the frequency of the term candidate in the test corpus, `MAX-TW` is the highest *TW* value among the occurrences of the term candidate in the test corpus. `STW` is the sum of the *TW* values of all occurrences of the term candidate in the test corpus. *TW* values of a certain term candidate may vary along the document, depending on, for instance, the syntactic function or the location of the occurrence. If, for example, the word `gender` occurs in a title with the following tag combination, it has a *TW* value of 0.541:

```
"<Gender>" "gender" <*> N NOM SG @<P <HEADLINE> </+INDEX-TERM> PROBABILITY:0.541
```

In another context its *TW* value is 0.375:

```
"<gender>"      "gender" N NOM SG @<P cc:>4 </+INDEX-TERM> PROBABILITY:0.375
```

Chapter 18

Term weights based on burstiness

As discussed earlier, two kinds of burstiness can be defined in order to weight index terms: **within-document burstiness** which refers to close proximity of individual instances of a term candidate within a document, and **document-level burstiness** which refers to multiple occurrence of a term candidate in a single document, which is contrasted with the fact that most other documents contain no instances of this candidate at all.

18.1 Within-document burstiness

This section will describe a new algorithm for measuring within-document burstiness of words by using distribution functions[1]. The new method counts the distances of the occurrences of individual words using paragraphs as units for measuring the distance[2]. Consider the following example: The document includes 50 paragraphs labelled as paragraph-1, paragraph-2,...,paragraph-50. The word `Marxism` occurs once in the paragraph-12, twice in the paragraph-13, and once in the paragraph-19.

The distances of the occurrences of `Marxism` are measured by using paragraphs as units:

```
Marxism paragraph-13 - paragraph-12 =  1
Marxism paragraph-13 - paragraph-13 =  0
Marxism paragraph-19 - paragraph-13 =  6
```

Distances are then ordered decreasingly and each occurrence adds one fourth of the total 100 % (since `Marxism` occurs four times in the document):

[1]The development of the algorithm was inspired by the ideas of Kimmo Koskenniemi, Seppo Mustonen, and Lauri Tarkkonen.

[2]Another possibility would be to measure distances between words. In this implementation, paragraphs were used as units, since paragraphs may be considered as topical units of discourse, as discussed earlier.

```
WORD      DISTANCE     PERCENTAGE
--------------------------------
Marxism (0)            25 %
Marxism  0             50 %
Marxism  1             75 %
Marxism  6             100 %
```

The first zero above (in parenthesis) stands for the first observation of `Marxism`. It adds one fourth of the total 100 % as well, even though there is no actual distance in that case; between four words, there are only three distances[3]. However, the figures actually used in the distribution function, instead of the absolute distances (e.g., 0,1,6), are calculated by the following formula:

$$distance = log(1 + \frac{freq \cdot D}{N})$$

where *freq* is the total frequency of the term candidate in the document, *D* is the absolute paragraph distance of the term candidate (e.g., 0, 1, or 6), and *N* is the number of the paragraphs in the document. The formula is not based on a mathematical or statistical theory, but on experiments and empirical observations. The formula tries to compensate for certain unwanted effects. The use of *log* scales the distances more compactly, for convenience. The use of constant 1 in the formula is necessary, because the absolute distance *D* can be zero, and we do not take the logarithm of zero. The use of *freq* reduces the advantage of the most frequent words. For example, the word `the` occurs frequently in all paragraphs, and so the distances are short. The purpose of this method, however, is not to find this kind of words but words that appear in the discourse at a certain point of the document, occur frequently for a while, and then disappear. In other words, the purpose is to recognize subtopics. The use of *N* reduces the effect of the document length; in longer documents word frequencies are higher, and so the formula uses a kind of relative frequencies (*freq/N*). The distribution function of `Marxism` is as follows:

```
WORD      DISTANCE (x)  PERCENTAGE (y)
-------------------------------------
Marxism   (0)             25 %
Marxism   0.000           50 %
Marxism   0.077           75 %
Marxism   0.392           100 %
```

Figure 18.1 presents the curves of the word `housework` and the word `see` (examples from the training corpus). The burstiness of different words can be compared by computing the areas above the curves of the words (the shaded areas of *Figure 18.2*)[4]:

$$Area = \sum_{i=0}^{max} (x_i \cdot (1 - y))$$

[3]The fourth distance could be measured from the last occurrence to the first occurrence, as a loop. For example, the last occurrence of `Marxism` is found in the 19th paragraph, 31 paragraphs from the end. The first occurrence of `Marxism` is found in the 12th paragraph. So, the distance between the last and the first paragraph is 43 (31+12).

[4]Naturally, there would be other possible ways to compare the curves as well.

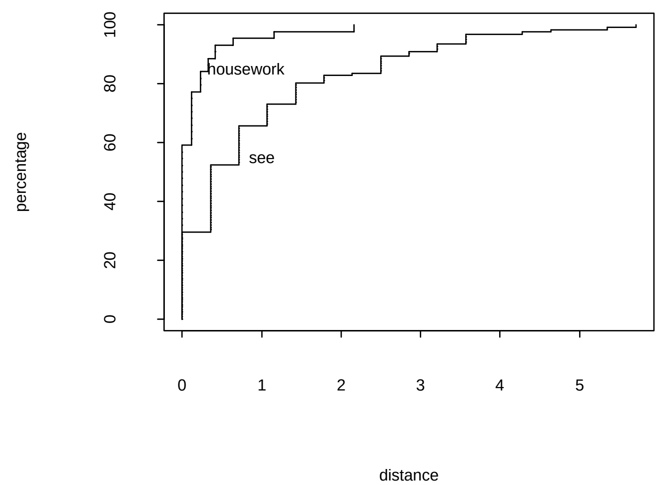

Figure 18.1: Distribution functions of `housework` and `see`.

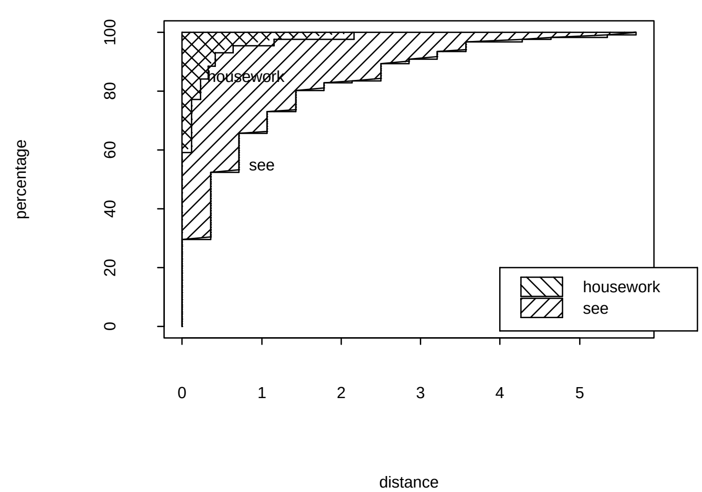

Figure 18.2: The shaded area of `housework` is smaller than the shaded area of `see`, that is, the word `housework` is burstier than the word `see`.

where x_i is the distance (from 0 to the maximum value, which is 0.392 in the example of `Marxism` above), and y is the percentage (e.g., when x is 0.077, y is 75 % in the example above).

The smaller the area, the burstier the word. As *Figure 18.2* indicates, the word `housework` is burstier than the word `see`. The following matrix includes ten examples from the test corpus, five bursty words and five non-bursty words. The `FREQ` column tells the frequency of the word in the test corpus.

```
WORD          FREQ   AREA
-------------------------------
Rawls         6      0.000
Marxist       3      0.000
allow         5      0.000
patriarchy    7      0.033
division      17     0.085
.
.
.
Willis        36     1.442
make          22     2.343
moral         57     2.301
case          32     3.564
woman         137    6.694
```

`Rawls`, `Marxist`, and `allow` are words that occur in only one paragraph (`AREA`=0). The first two of them are subtopics and index terms, but `allow` is not. It is just a verb that is used frequently in only one paragraph, for some reason. `Patriarchy` and `division` are bursty words as well, but only the former one is an index term. `Willis`, on the other hand, is a frequently used word all over the document. It is not bursty word, but it is an index term. Moral and women are important themes of the test corpus, and so these words occur frequently all over the test corpus. `Make` and `case`, on the other hand, are common words in many texts. Their high frequency does not imply that they are themes of the test corpus. As these examples indicate, main themes are often non-bursty words and subthemes are often bursty words. Combined with other weighting schemes the method described in this section might be useful for distinguishing subtopics from main topics, among other things.

18.2 Document-level burstiness

In this thesis, document-level burstiness is measured by the formula presented in *Section 13.3*, the Combined Weight for a term *t(i)* in a document *d(j)* (Robertson and Sparck Jones, 1997):

$$CW(i,j) = \frac{CFW(i) \cdot TF(i,j) \cdot (K1+1)}{K1 \cdot (1-b+b \cdot (NDL(j)) + TF(i,j)}$$

CFW stands for Collection Frequency Weight and it is defined for a term *t(i)* as

$$CFW(i) = logN - log(n(i))$$

where *n(i)* is the number of documents term *t(i)* occurs in, and *N* is the number of documents in the collection (810 documents in this case). Usually *CFW* is referred to as Inverted Document Frequency (*IDF*).

TF stands for Term Frequency and it is defined for a term *t(i)* in a document *d(j)* as

TF(i,j) = the number of occurrences of term t(i) in document d(j)

NDL stands for Normalized Document Length and it is defined for a document *d(j)* as

NDL(j) = (DL(j))/(Average DL for all documents)

where *DL(j)* is the total number of running words in document d(j)[5]. In the corpus of this thesis, the average Document Length (*DL*) is 2799 words (2,267,220/810).

K1 and *b* are tuning constants. *K1* modifies the extent of the influence of term frequency, and the value *K1=2* is used in this thesis, since it was found to be effective in some TREC tests (Robertson and Sparck Jones, 1997). The tuning constant *b* modifies the effect of document length, and the value *b=0.75* is used in this thesis, since it was found to be effective in some TREC tests (Robertson and Sparck Jones, 1997).

Robertson and Sparck Jones refer to this formula as Combined Weight *CW* (Robertson and Sparck Jones, 1997), but as discussed earlier, it is a variant of the standard *TF*IDF* weighting scheme. The main difference to the basic *TF*IDF*-formula is that *CW* takes into account the document length as well. In this thesis, *CW* is referred to as *TF*IDF*, and it was used to weight multi-word index terms as well as single-word index terms. So, *IDF* values of 18,654 single-word and multi-word term candidates were calculated by using all the 810 documents. These 18,654 term candidates include all term candidates of the texts with index term mark-up (64,996 running words). *TF*IDF* values were calculated by using the base forms provided by the parser, and they were not calculated for stop words or arbitrary word sequences, but only for those term candidates that represent the 89 defined patterns. In this way, the precision was improved, since a lot of totally impossible term candidates were excluded, such as `they`, `need of`, and `but also in`. Most of the poor term candidates were excluded on the basis of the tag lists, but a short stop list was applied as well. For example, certain adjectives such as `certain`, `whole`, and `same` were included in the stop list.

In addition, *TF*IDF* values were calculated for bigrams formed by using the simple phrase constructing method described in *Section 13.4* (Buckley *et al.*, 1995): all adjacent pairs of base forms of non-stopwords were considered as two-word term candidates. The stop list of 390 words included articles, pronouns, verbs (e.g., `be`, `became`, and `must`), and adverbs, among others. This stop list was much longer than the stop list of the pattern matching method, since in that

[5]Document length can be measured in different ways. In this thesis, it is measured by counting the number of running words.

method a number of stop words would have been redundant because they would have been excluded on the basis of their tag lists. For example, it was not necessary to list all different pronouns, because all words with the `PRON` tag (i.e., all pronouns) were excluded. So, two different sets of bigrams were created: one by using the pattern matching method and one by using the simple method described above, and for the candidates of the both sets *IDF* values were calculated by using all the 810 documents. *TF*IDF* values were then calculated for both sets, and the results were compared.

Chapter 19

Term weights based on linguistic tags and burstiness

An important assumption of this thesis is that combining evidence based on linguistic annotation with evidence based on burstiness offers a profitable approach to developing tools for information retrieval tasks. In this thesis, the combination of evidence is done by replacing the *TF* values (term frequencies) by the *STW* values (summed tag weights) in the *TF*IDF*-formula (or *CW*-formula, as Robertson and Sparck Jones call it). The new formula is referred to as *STW*IDF*, and it is the weighting scheme of the automatic indexer developed in this thesis:

$$STW * IDF(i,j) = \frac{IDF(i) \cdot STW(i,j) \cdot (K1+1)}{K1 \cdot (1-b+b \cdot (NDL(j)) + STW(i,j)}$$

So, instead of counting the plain occurrences, i.e. the frequencies, the individual term occurrences are weighted according to their tag patterns. In the example

```
"<Marx>"         "Marx" <Proper> N NOM SG @SUBJ subj:>2 </+INDEX-TERM> TW:0.977
"<suggested>"    "suggest" V PAST VFIN @+FMAINV #2 main:>0 TW:0.005
```

the *TW* value of `Marx` is higher than the *TW* value of `suggest`. If the frequency of these words is the same in the document, they have the same *TF* values, but since the word `Marx` has higher *TW* values, its *STW* value is higher than the *STW* value of word `suggest`. On the other hand, if two words have similar tag weights, but one of them occurs more frequently, the more frequent word has a higher *STW* value than the less frequent one. In this way evidence based on linguistic annotation is combined with evidence based on burstiness.

The following matrix includes ten examples from the test corpus:

TERM CANDIDATE	TF	TF*IDF	STW	STW*IDF
abandon	1	1.126	0.019	0.026
ability	2	1.942	0.589	0.780
ability of neo-capitalism	1	3.827	0.037	0.173
abortion-decision	1	4.526	0.407	2.126
Abraham	1	2.163	0.810	1.830
absent	1	2.143	0.093	0.250
abstract	5	3.971	0.180	0.299
abstract conception	2	7.388	0.536	2.709
abstract conception of justice	1	4.526	0.121	0.683
abstract labour	2	6.430	0.534	2.243

`TF` is the frequency of the term candidate in the test corpus and `TF*IDF` is its *TF*IDF* value. `STW` is the sum of the *TW* values (summed tag weights) of the term candidate in the test corpus and `STW*IDF` is its *STW*IDF* value.

Chapter 20

Summary

To sum up, the *STW*IDF* weighting scheme described above combines evidence from burstiness and evidence from linguistic analysis provided by a syntactic parser. The weighting scheme was trained by using an **index term corpus** which is a linguistically analysed text collection where the index terms were manually marked up by a research aide.

Another new weighting scheme was introduced as well. This method measures within-document burstiness.

Part V

Results

This part will present the results of the experiments of the thesis:

- *Chapter 21* will present a summary of findings in corpora with manual index term mark-up. These findings provide the basis for the weighting schemes that use evidence from linguistic analysis.
- *Chapter 22* will evaluate the weighting scheme that is based on tag sets of words only.
- *Chapter 23* will evaluate the weighting schemes that are based on evidence from burstiness only, and the new weighting scheme that combines evidence based on linguistic analysis (tag sets) and evidence based on document-level burstiness.

Chapter 21

Summary of findings in corpora with manual index term mark-up

21.1 Patterns of index terms

The combination of linguistic annotation and index term mark-up in the training corpus made it possible to examine the linguistic structure of index terms.

Figure 21.1 presents the seven main patterns of index terms. All these patterns included several subpatterns, so that altogether 89 different tag combinations were considered as relevant term patterns. The lengths of the most common index term patterns varied from one to three words. Only 95 index terms were longer (50 of which showed `of`-constructions[1]), and some of these had only a few representatives. For example, three patterns consisted of five words each, but the training corpus included only 15 such terms in total (e.g., `surface of oppressive structural relationships`). Nine index terms were even longer than five words. The reason why the index term corpus includes a few long and complicated index terms is that the index term mark-up was based on book indexes. Sometimes a simple index term of the book index was expressed in a more complicated way in the text. For example, the book index included the index term `women as inferior`, and the text included the noun phrase `philosophical conceptions which exclude women` on the referred page. This noun phrase was marked up as an index term in the index term corpus.

The most common term pattern (`A-N`) consisted of an adjective as a premodifier and a noun as a head. `Post-Freudian theory` is an example of such index terms[2]. `Same time` is an example of non-terms of the `A-N` pattern. Single common nouns (`N`) comprised the next largest group of index terms. `Capitalism` and `housework` are examples of terms of the `N` pattern, and `amount` and `example` are examples of non-terms of the `N` pattern. Index terms as well as index term patterns may be nested, e.g., the expression found in the text `critical ethnography`

[1]Altogether 92 index terms contained the preposition `of`, and 42 of them consisted of three words.

[2]In the book index this term was in the form `post-Freudism`.

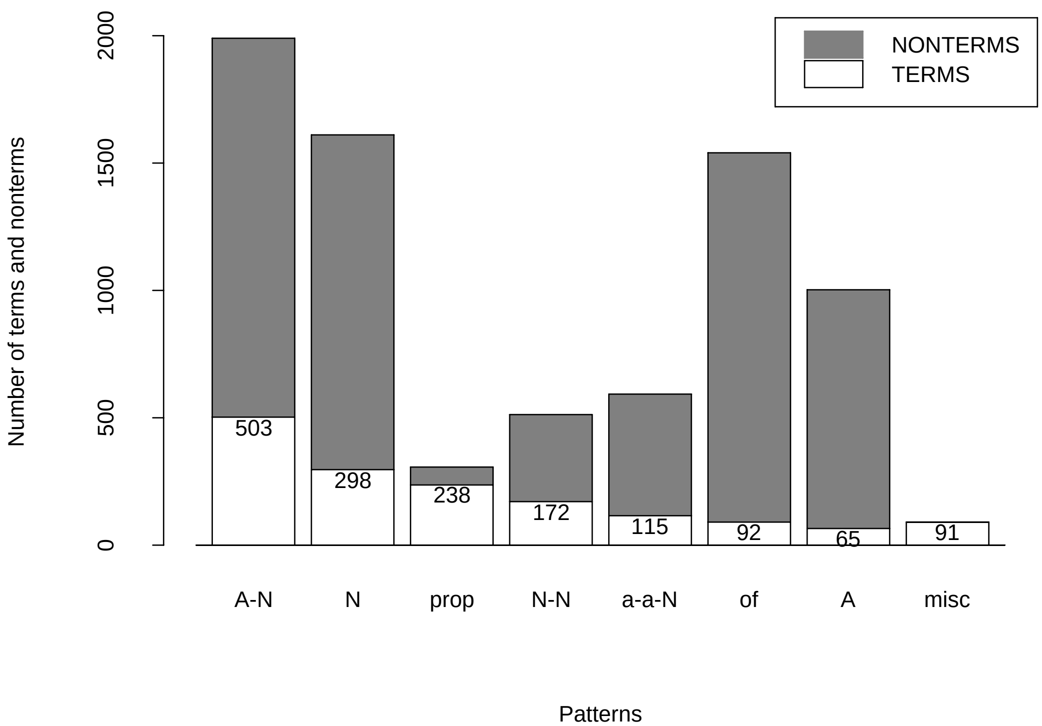

Figure 21.1: Different term patterns and their frequencies in the training corpus.

includes two index terms, `critical ethnography` (`A-N` pattern) and `ethnography` (`N` pattern). Proper nouns (`prop`) included all of the proper noun terms of various lengths[3]. The pattern of two successive nouns (`N-N`) contained a few genitive constructions, such as `women's oppression` (term) and a number of compounds, e.g., `mass media` (term). `People's words` and `core concept` are examples of non-terms of `N-N` pattern.

Two successive premodifiers (`attr:>`) in the `a-a-N` pattern may be either nouns or adjectives, for instance, `Marx's scientific socialism` (term) and `oppressive social structures` (term). `Rather different way` is an example of non-terms of the `a-a-N` pattern. Genitive constructions using `of`-preposition (`of`) included 92 different index terms of various lengths, e.g., `oppression of women` and `structural and historically specific nature of capitalism`. `Variety of ways` is an example of non-terms of the `of` pattern.

The word `and` was included in 29 different index terms of the training corpus. The most common term pattern that included `and` was `N-and-N` pattern. For instance, `deconstruction and reconstruction` occurred nine times in the training corpus, and it was in the book index in the same form as well. `Time and place` is an example of non-terms of the `N-and-N` pattern. The only major non-NP group was formed by 65 index terms consisting of single adjectives (`A`). The adjectives of the training corpus included some index terms, such as `male-biased` and `Freudian`, as well as a number of non-terms, such as `possible` and `long`. Only 91 index terms (`misc`) did not fall into any one of the seven above-mentioned main categories. This group included verbs (e.g., `reconstruct`[4]), adverbs (e.g., `dialectically`[5]), and some longer term patterns (e.g., `political affiliation for critical social research`[6]), among other things.

Part-of-speech tagging provides important information for distinguishing between the main term pattern types, but further information is needed in order to identify term candidates successfully. For example, not all adjective-noun-sequences or noun-noun-sequences are noun phrases; in the sentence

```
She argues that this perception of discontinuity and dominance has
consequences to the way experience finds expression in the work of
male philosophers.
```

the noun `way` and the noun `experience` do not form the noun phrase `way experience`. Syntactic analysis reveals that there is a clause boundary between these words.

Part-of-speech tagging is the first step of determining the tag weights of term candidates, but other information in the tag lists is useful as well: syntactic functions, lexical features, and location

[3]Proper names will be discussed later in more detail.

[4]In the book index this term was in the form `reconstruction`.

[5]In the book index this term was in the form `dialectical analysis`.

[6]In the book index this term was in the form `politics and research`.

of term candidates. The following 17 term patterns will give a picture of the 89 different tag combinations that were considered as relevant term patterns. Each term pattern has an estimated index term probability that was calculated by using the training corpus as training material. These probabilities are same as tag weights (*TW*), as discussed earlier. The automatic indexer observes tag combinations of words and weights term candidates using the probabilities that were calculated from the training corpus. *Chapter 17* described the process of creating the index term patterns in more detail. In the following patterns, tags are combined using the boolean operators (AND, OR, NOT).

Single verb pattern, example 1:

```
tag combination: ING AND <SVO>
weight: 0.019
```

That is, -ing forms (`ING`) of monotransitive verbs (`<SVO>`) will have a weight of 0.019, e.g., `deconstructing`.

Single verb pattern, example 2:

```
tag combination: (V OR EN) AND <SVO>
weight: 0.007
```

That is, monotransitive verbs with the `V` tag (verb) or the `EN` tag (-ed/en forms) will have a weight of 0.007, e.g., `oppressed`. So, -ing forms were more typical to index terms in the training corpus than -ed/en forms.

Single adverb pattern:

```
tag combination: ADV AND <DER:ly>
weight: 0.040
```

That is, adverbs (`ADV`) with the `<DER:ly>` tag (derived adverb in -ly) will have a weight of 0.040, e.g., `dialectically`.

Single adjective pattern, example 1:

```
tag combination: A AND attr: AND NOT (SUP OR CMP OR <DER:ian> OR <*> OR <->)
weight: 0.016
```

That is, adjectives (`A`) that are premodifiers (`attr:`) but are not in superlative or comparative form and do not have the `<DER:ian>` tag (derived adjective in -ian) or the `<*>` tag (the first letter in higher case) or the `<->` tag (the word includes an hyphen) will have a weight of 0.016, e.g., `empirical`.

Single adjective pattern, example 2:

```
tag combination: A AND (<END:ogy> OR <DER:ian> OR <END:ism> OR
 <DER:less> OR <DER:al> OR <DER:ive> OR <DER:ic> OR <?> OR <*>) AND
 NOT (attr: OR @<NOM OR SUP OR CMP OR <DER:ble>)
weight: 0.219
```

That is, adjectives that are not premodifiers or postmodifiers (`@<NOM`) and are not in superlative or comparative form and do not have the `<DER:ble>` tag (derived adjective in -ble) and do have an ending -ogy, -ian, -ism, -less, -al, -ive, or -ic, or the `<?>` tag (word is not found in the lexicon of the parser) or the `<*>` tag will have a weight of 0.219, e.g., `egalitarian`. Certain endings were typical to index terms in the training corpus. `<?>` and `<*>` tags provided useful information as well. These issues will be discussed below in more detail.

Single common noun pattern, example 1:

```
tag combination: N AND subj: AND
 (<-> OR <?> OR <*>) AND
 NOT (<Proper> OR PL OR <DER:ism>)
weight: 0.654
```

That is, common nouns (`N` AND NOT `<Proper>`) that are subjects (`subj:`), with the `<->`, `<?>`, or `<*>` tag, and that are not in plural (`PL`), and that do not have the ending -ism (`<DER:ism>`), will have a weight of 0.654, e.g., `philosophy`. Index terms were quite often subjects, as the next section will show.

Single common noun pattern, example 2:

```
tag combination: N AND <HEADLINE> AND
 (<-> OR <?> OR <*>) AND
 NOT (<Proper> OR PL OR <DER:ism> OR subj:)
weight: 0.541
```

That is, common nouns in titles and subtitles (`<HEADLINE>`) with the `<->`, `<?>`, or `<*>` tag, and that are not subjects (`subj:`), and not in plural (`PL`), and that do not have the ending -ism (`<DER:ism>`), will have a weight of 0.541, e.g., `Totality`. Titles and subtitles contain often index terms, as *Section 21.4* will show.

Single common noun pattern, example 3:

```
tag combination: N AND pcomp: AND <DER:ism> AND
 NOT (<Proper> OR PL)
weight: 0.619
```

That is, common nouns that are prepositional complements (`obj:`), with the `<DER:ism>` tag, and that are not in plural (`PL`) will have a weight of 0.619, e.g., `capitalism`. The words with the ending -ism were quite often index terms in the training corpus, as *Section 21.3.1* will show.

Single common noun pattern, example 4:

```
tag combination: N AND obj: AND
 NOT (<Proper> OR PL OR <DER:ism> OR <-> OR <?> OR <*>)
weight: 0.195
```

That is, common nouns that are objects (`obj:`), with no `<DER:ism>`, `<->`, `<?>`, or `<*>` tag, and that are not in plural (`PL`) will have a weight of 0.195, e.g., `ontology`.

Single common noun pattern, example 5:

```
tag combination: N AND attr: AND
 NOT (<Proper> OR PL OR <DER:ism> OR <-> OR <?> OR <*>)
weight: 0.020
```

That is, common nouns that are premodifiers (`attr:`), with no `<DER:ism>`, `<->`, `<?>`, or `<*>` tag, and that are not in plural (`PL`) will have a weight of 0.020, e.g., `public`.

Single proper noun pattern, example 1:

```
tag combination: N AND <Proper> AND subj:
weight: 0.977
```

That is, proper nouns that are subjects will have a weight of 0.977, e.g., `Lloyd`.

Single proper noun pattern, example 2:

```
tag combination: N AND <Proper> AND pcomp:
weight: 0.810
```

That is, proper nouns that are prepositional complements will have a weight of 0.810, e.g., `Hegel`.

Two-word noun phrase pattern, example 1:

```
tag combination of the first word: A AND attr:
tag combination of the second word: N AND (<DER:ism> OR <DER:bility>) AND
 NOT <Proper>
weight: 0.658
```

That is, a two-word noun phrase that has an adjective as a premodifier and a common noun with an ending -ism or -bility, as a head, will have a weight of 0.658, e.g., `moral scepticism`.

Two-word noun phrase pattern, example 2:

```
tag combination of the first word: A AND attr: AND <*> AND
 (<DER:ist> OR <DER:ian> OR <DER:al> OR <DER:ic>)
tag combination of the second word: N AND <*> AND
 NOT (<Proper> OR <DER:ism> OR <DER:bility>)
weight: 0.485
```

That is, a two-word noun phrase that is a subject and has an adjective as a premodifier and a common noun as a head, and the first letters of both words are in the higher case, and the first word has an ending -ist, -ian, -al, or -ic, and the second word does not have the ending -ism or -bility, will have a weight of 0.485, e.g., `Ethnographic Approach`.

Two-word noun phrase pattern, example 3:

```
tag combination of the first word: N AND attr: AND NOT (<*> OR GEN)
tag combination of the second word: N AND <Proper>
weight: 0.500
```

That is, a two-word noun phrase that has a common noun (NOT `<*>`) that is not in genitive case (NOT `GEN`) as a premodifier and a proper noun as a head, will have a weight of 0.500, e.g., `mid-twentieth-century America`.

Three-word noun phrase pattern, example 1:

```
tag combination of the first word: A AND attr:
tag combination of the second word: A AND attr: AND (<DER:ic> OR <DER:al>)
tag combination of the third word: N AND pcomp: AND NOT <Proper>
weight: 0.368
```

That is, a three-word noun phrase that is a prepositional complement and has two adjectives as premodifiers of which the second one has an ending -ic or -al, and that has a common noun (NOT `<Proper>`) as a head will have a weight of 0.368, e.g., `critical ethnographic technique` .

Three-word noun phrase pattern, example 2:

```
tag combination of the first word: N AND pcomp:
tag combination of the second word: @<NOM-OF
tag combination of the third word: N AND pcomp:
weight: 0.088
```

That is, a three-word noun phrase that is a prepositional complement, and where the first word is a noun, the second word is `of`, and the third word is a noun, will have a weight of 0.088, e.g., `division of labour`.

The following sections will describe different features of index terms in more detail.

21.2 Syntactic functions of terms

This section will examine what are the typical syntactic functions of index terms. We shall focus on the single nouns and noun phrases with one or two premodifiers; this covers the five most common term patterns described in the previous section. *Figure 21.2* (training corpus) and *Figure 1*[7] (test corpus) present the five most common syntactic functions of noun phrases (including single nouns; nouns as premodifiers are excluded):

```
@<P          Complement of preposition  (e.g., "divorced from ETHICS")
@SUBJ        Subject (e.g., "PSYCHOANALYSIS might offer")
@OBJ         Object (e.g., "She cites HEGEL")
@NH          Stray noun phrase head (e.g., "(GOLDTHORPE et al., 1969)")
@PCOMPL-S    Subject complement  (e.g., "It is not MARXISM")
```

The sixth group (`misc`) includes, for example, indirect objects, object complements, and appositions. Complements of prepositions (`@<P`) had the highest absolute number of index terms (1,546), but stray noun phrase heads (`@NH`) had the highest proportion of terms (the order is same for the test corpus):

```
Function     Proportion  Term/All
--------------------------------------------
@NH          0.478       (200/418)
@SUBJ        0.332       (718/2160)
@<P          0.310       (1546/4982)
@OBJ         0.247       (433/1750)
@PCOMPL-S    0.154       (87/566)
```

In the training corpus a premodifying noun was a part of a multi-word term 509 times. There were, however, cases like `Hughes's paper`, where the genitive premodifier `Hughes's` is an index term, but the whole phrase is not. In the training corpus a premodifying noun alone was an index term 134 times (the proportion of terms was 0.177 in the test corpus):

[7] *Figure 1* is in *Appendix 2*, as a number of other figures of **Result** part. If a figure of the training corpus and a figure of the test corpus are very much alike, the figure of the test corpus is in the appendix 2.

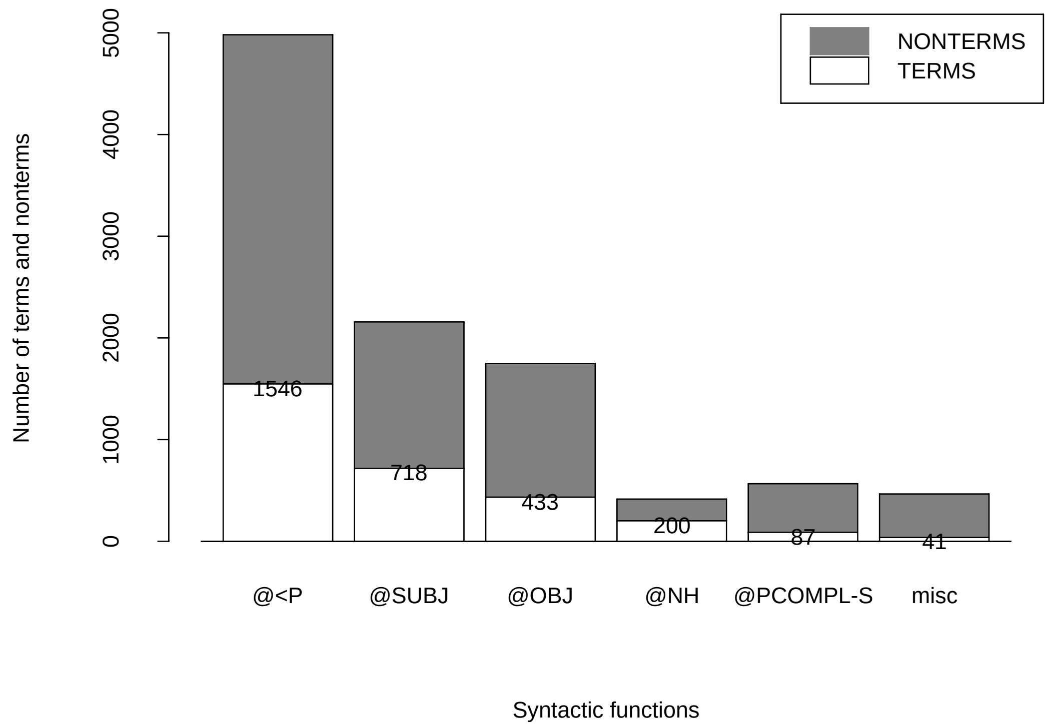

Figure 21.2: Syntactic functions of index terms (training corpus).

```
Function                      Proportion  Term/All
---------------------------------------------------
@A> (premodifier of a noun)  0.142       134/945
```

In the training corpus an adjective was a part of a multi-word term 1230 times. 118 times a single adjective was marked up as a term, of which 40 times the adjective was a premodifier. For example, the phrase `Aristotelian distinction` was not marked up as a term, but the adjective `Aristotelian` was. In the index, however, the term was not `Aristotelian`, but `Aristotle`. An adjective as a subject complement was marked up as an index term 34 times in the training corpus. For example, the adjective `positivist` in the clause `no method of data collection is inherently positivist` was marked up as a term. In the index this term was in the form `positivism`. Other functions of single adjective terms included, for example, apposition and complement of preposition of which the adjective `oppressive` in the sentence `Marx saw social structures as oppressive` is an example.

To sum up, in the index term corpus some syntactic functions were more typical for index terms than others. The differences between functions proved to be quite similar in the training corpus and in the test corpus. Accordingly, the automatic indexer of this thesis weights index terms also according to their syntactic functions. For example, subjects are weighted higher than objects, as the term proportion is higher with subjects than with objects.

21.3 Lexical features of terms

Index terms marked up into the index term corpus had certain **lexical features** that are useful for the automatic indexer. In this section seven endings typical of index terms will be studied: `-ist` (e.g., `capitalist`), `-ogy` (e.g., `epistemology`), `-ism` (e.g., `feminism`), `-ory` (e.g., `history`), `-ity` (e.g., `objectivity`), `-al` (e.g., `philosophical`), and `-ic` (e.g., `linguistic`). Three tags provided by the dependency parser will be considered as well:

- `<Proper>`,
- `<-Indef>`, and
- `<?>`

The `<Proper>` tag identifies proper nouns, the `<-Indef>` tag identifies nouns with which indefinite article can not be used, e.g., `ontology`, `ethics`, and `logic`. The `<?>` tag is attached to words that are not found in the lexicon of the parser, e.g., `deconstructive-reconstructive`.

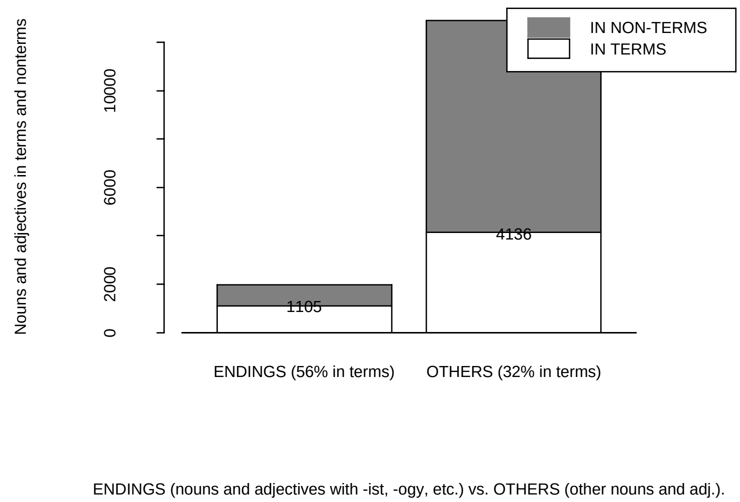

Figure 21.3: Nouns and adjectives as terms and non-terms, or as parts of terms and non-terms (training corpus). 56 % of the words with one of the seven endings (-ist, -ogy, -ism, -ory, -ity, -al, or -ic) has a term tag (`<+INDEX-TERM>`).

21.3.1 Endings

Figure 21.3 shows how often the seven above mentioned endings were found in the index terms of the training corpus[8]. 1,985[9] adjectives and nouns had some of these endings, and 1,105 of these words (56 %) were index terms or parts of index terms, as the word `linguistic` is a part of the term `linguistic philosophy`. 12,892 adjectives and nouns did not have any of these endings, and 4,136 of them (32 %) were index terms or parts of index terms. Altogether, 5,241 adjectives and nouns were marked up as index terms or as parts of index terms, and 21 % (1,105/5,241) of them had some of these endings. In other words, these endings were relatively typical for index terms of the training corpus. In the test corpus, however, this proportion was not as high (13 %, 265/2,074)[10]. But as in the training corpus, also in the test corpus adjectives

[8]The figures are based on the tags provided by the parser, not directly on the actual endings of words. For example, the word "philosophical" had the `<DER:al>` tag, but the word "real" did not, since in the case of "philosophical" `-al` is derivational ending, but in the case of "real" it is not. So, "real" is not taken into account here.

[9]1,985 adjectives and nouns included all occurrences of these words (**tokens**), that is, there were fewer distinct adjectives and nouns (**types**).

[10]In other words, the test corpus included less index terms with these endings, relatively. This observation demonstrates that texts are different in this respect.

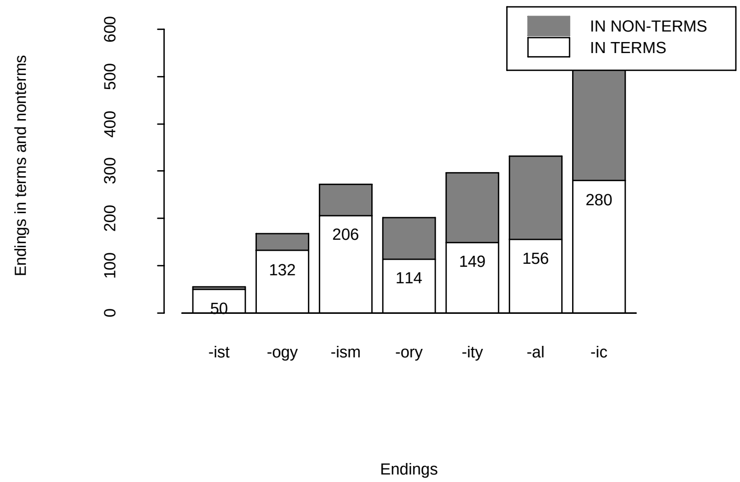

Figure 21.4: Seven endings in terms and non-terms (training corpus).

and nouns with these endings were terms or parts of terms clearly more often, relatively (52 %), than adjectives and nouns without these endings (31 %), as presented in *Figure 2*[11]. Consequently, the automatic indexer weights the words with these endings higher than words without these endings[12].

Figure 21.4 presents the differences between the endings in the training corpus, and *Figure 21.5* presents the differences between the endings in the test corpus. In the training corpus, `-ic` was the most frequent ending of the terms (280 occurrences). In the test corpus, however, the most frequent ending of the terms was `-ity` (74 occurrences).

On the other hand, when the *proportions* of term occurrences were compared between these two corpora, the differences were not so great. In the following matrix occurrences of endings in terms are divided by the total number of occurrences. As *Figures 21.4 and 21.5* illustrate, some endings are typical for some texts and some other endings for other texts. For example, in the training corpus `-ory` occurred altogether 201 times (114 times in terms), but in the test corpus `-ory` occurred only 20 times (8 times in terms). The big difference is not explained by the lengths of the corpora alone: the training corpus consisted of 38,138 words and the test corpus consisted of 17,392 words [13]. This kind of differences between these two corpora are not surprising, however,

[11] In *Appendix 2*.
[12] The final weight depends, naturally, on all evidence (word frequencies, syntactic functions, and so on).
[13] 17,392/38,138=0.456 and 20/201=0.010

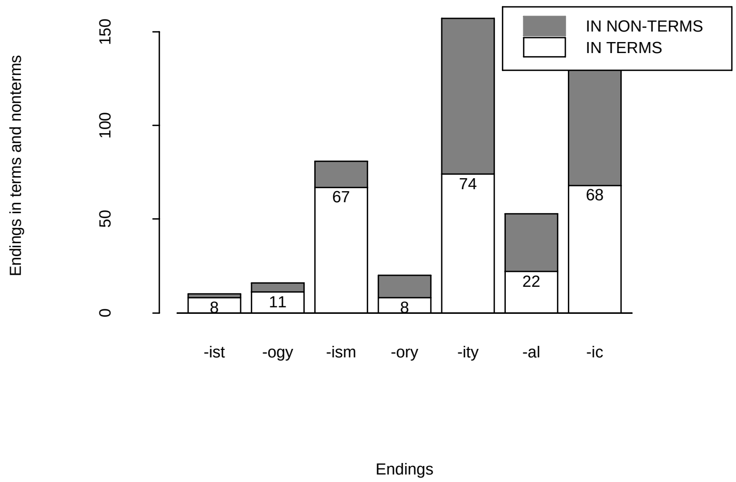

Figure 21.5: Seven endings in terms and non-terms (test corpus).

since the corpora are relatively small, and the discussed issues and the vocabulary are more or less different in these corpora. On the other hand, some endings happened to be quite common in both corpora, such as `-ism`, `-ity`, and `-ic`.

ENDING	TRAINING CORPUS: PROPORTION (TERM/ALL)	TEST CORPUS: PROPORTION (TERM/ALL)
-ist	91 % (50/55)	80 % (8/10)
-ogy	79 % (132/168)	69 % (11/16)
-ism	76 % (206/272)	83 % (67/81)
-ory	57 % (114/201)	40 % (8/20)
-ity	50 % (149/297)	47 % (74/157)
-al	47 % (156/332)	42 % (22/53)
-ic	45 % (280/623)	44 % (68/156)

With the low frequency endings (e.g., `-ist`) the differences of proportions between the corpora tend to be greater than with the high frequency endings (e.g., `-ic`). Anyhow, the proportions of the test corpus seem to have a lot in common with the proportions of the training corpus, and thus the automatic indexer uses the analysis of the endings for weighting the terms.

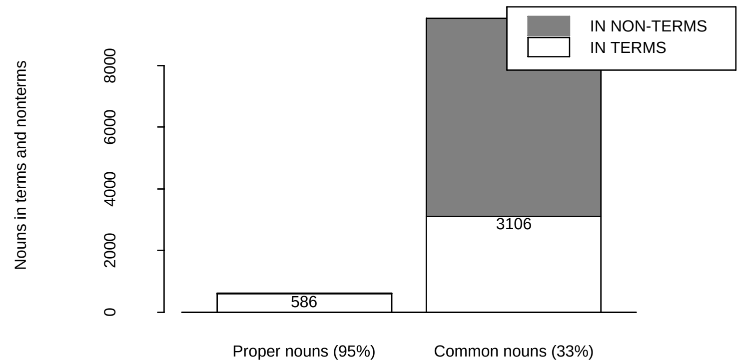

Figure 21.6: Proper nouns and common nouns as terms and non-terms, or as parts of terms and non-terms (training corpus).

21.3.2 Proper nouns

As discussed earlier, most of the proper nouns were included in the indexes of the corpora. So, both in the training corpus and in the test corpus 95 % of the words with the `<Proper>` tag had a `<+INDEX-TERM>` tag[14] as well, as illustrated in *Figures 21.6 and 3*[15]. The words with the `<Proper>` tag and with no `<+INDEX-TERM>` tag were mainly first names (such as `Margaret` in `Margaret Whitford`), which usually are easy to identify by their context. So, `<Proper>` tags provide highly useful evidence for the automatic indexer.

[14] In this context the `<+INDEX-TERM>` tag represents `</+INDEX-TERM>` tags, `<+INDEX-TERM-1>` tags etc. as well.

[15] In *Appendix 2*.

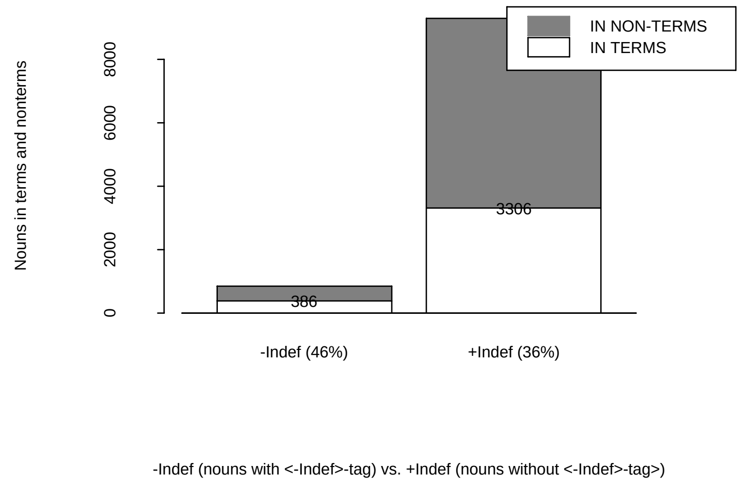

Figure 21.7: Nouns with and without the `<-Indef>` tag in terms and non-terms (training corpus).

21.3.3 Words without an indefinite article

Another useful tag is the `<-Indef>` tag which identifies nouns with which the indefinite article is never used. Such nouns are `feminism`, `epistemology`, and `knowledge`, among others. Often these words denote abstract concepts that tend to be appropriate index terms as well. In the training corpus nouns with the `<-Indef>` tag had the `<+INDEX-TERM>` tag in their tag list 386 times out of 840 (46 %; -Indef in *Figure 21.7*), whereas nouns without the `<-Indef>` tag had the `<+INDEX-TERM>` tag in their tag list 3306 times out of 9296 (36 %; +Indef in *Figure 21.7*).

As seen in *Figure 4*[16], also in the test corpus the words with the `<-Indef>` tag had a higher term proportion than the words without the `<-Indef>` tag.

21.3.4 Words not found in the lexicon

As mentioned above, the `<?>` tag is attached to words that are not found in the lexicon of the parser[17]. These words are often proper names or compound nouns, such as `gender-drama`.

[16]In *Appendix 2*.

[17]The fact that a given word is not found in the lexicon does not prevent the morphological analysis of the word and syntactic analysis of the sentence. The heuristic rules of the parser make the parser robust in this respect.

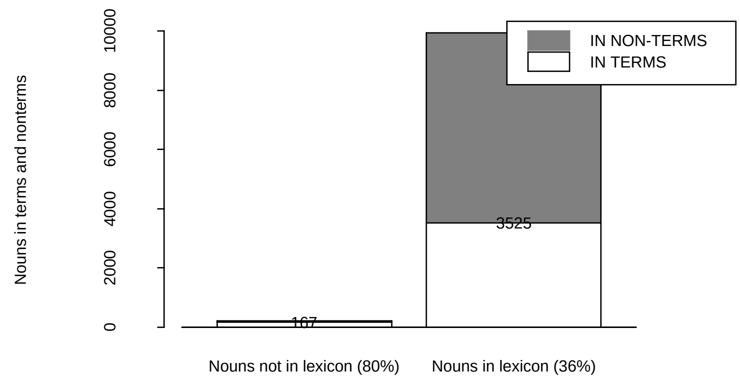

Figure 21.8: Nouns not in lexicon (<?> tag) and nouns in lexicon as terms and non-terms, or as parts of terms and non-terms (training corpus).

Quite often these words are appropriate index terms as well. In the training corpus nouns not found in the lexicon had a manually attached `<+INDEX-TERM>` tag in their tag list 167 times out of 209 (80 %), whereas nouns found in the lexicon had the `<+INDEX-TERM>` tag in their tag list 3525 times out of 9927 (36 %) as illustrated in *Figure 21.8*. In the test corpus (*Figure 5*[18]) 70 % of the nouns with the `<?>` tag had the `<+INDEX-TERM>` tag in their tag list.

To sum up, endings as well as `<Proper>` tags, `<-Indef>` tags, and `<?>` tags provide evidence that combined with other evidence is useful for the automatic indexer, since these endings and tags seem to be typical for many index terms, as shown above.

21.4 Location of terms

This section will evaluate the assumption that titles, subtitles, the first and last sentences of a paragraph, and the first and last paragraphs of a section, contain more topical expressions than the text does on average. We will investigate whether analyzing the location of words and phrases provides evidence for index term weighting. So, the words in the first and last sentences and in the first and last paragraphs were marked up by tags, as well as the words in titles and subtitles. Titles and subtitles were not originally marked up in the corpora, and tagging with specific location tags

[18]In *Appendix 2*.

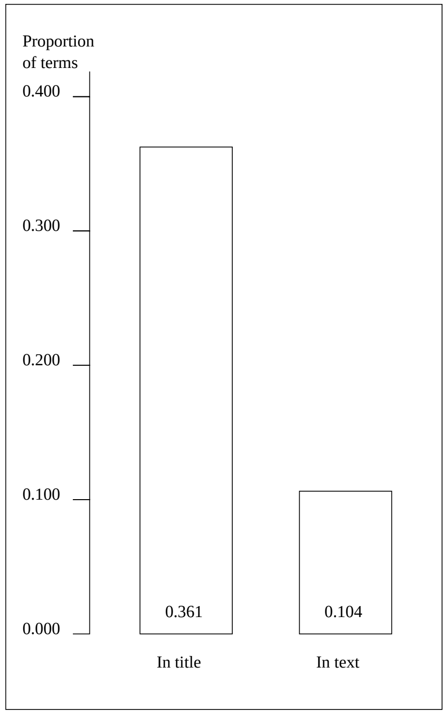

Figure 21.9: Proportion of terms in titles/subtitles and in text (training corpus).

was done by a simple automatic procedure which marked up all words of a short paragraph less than seven words[19] if the paragraph did not end with a full stop.

The titles and subtitles of the training corpus included 169 words and 61 single-word and multi-word terms, and the proportion of terms was calculated by dividing the number of terms by the number of words. Words with a `<HEADLINE>` tag had a term proportion of 0.361 (61/169), whereas words without a `<HEADLINE>` tag had a term proportion of 0.104. *Figure 21.9* presents the proportions in the training corpus, and *Figure 6*[20] presents the proportions in the test corpus (in titles/subtitles 0.398 and in elsewhere 0.099). The results suggest, as expected, that the automatic indexer should, in general, weight words and phrases in titles and subtitles higher than words and phrases elsewhere.

Figure 21.10 (training corpus) and *Figure 21.11* (test corpus) present index term proportions in different locations of texts. The first sentences of the first paragraphs of texts in the training corpus contained 649 words and 101 single-word and multi-word terms, and so the proportion of terms was 0.156 (101/649). The index term density was highest in the first sentences of the first paragraphs of texts both in the training corpus and in the test corpus (0.134) and lowest in the last sentences of the last paragraphs (0.094 in training corpus, and 0.041 in test corpus). The second highest index term density was in the first sentences of the last paragraphs (0.136 in training

[19]Seven words proved to be an appropriate threshold for this corpus.
[20]In *Appendix 2*.

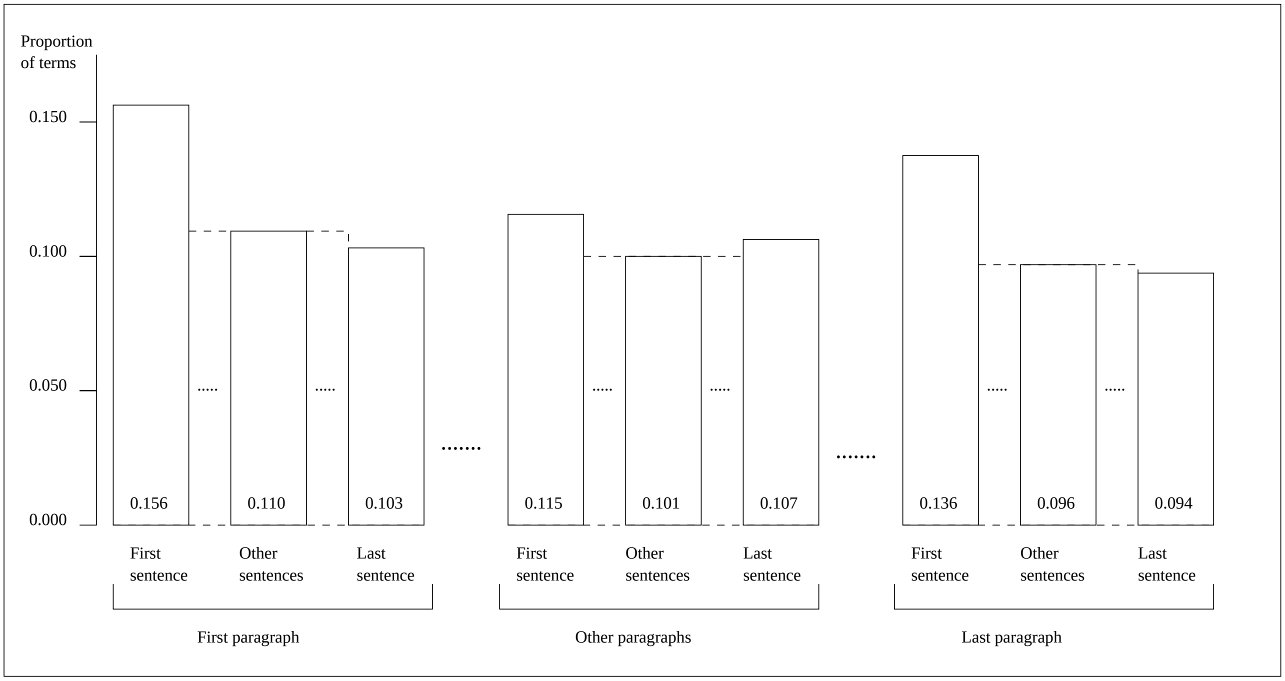

Figure 21.10: Index term density in training corpus.

corpus, and 0.130 in test corpus).

Even though the last sentences of the last paragraphs contained less index terms than other sentences, on average, those terms may be considered as more important than terms on average. Each term was attached with a **term frequency tag** which tells how many times the term was marked up as an index term by the research aide. The higher term frequency indicates the greater importance of the term as a content descriptor. In the training corpus the average term frequency of the terms of the first sentences of the first paragraphs was 34.455, the average term frequency of the terms of the last sentences of the last paragraphs was 24.837, and the average term frequency of the terms of the other sentences was 18.145. Both in the training corpus and in the test corpus, the first sentences of the first paragraphs and the last sentences of the last paragraphs contained more high-frequency terms than the other sentences, as presented in *Figure 21.12* (training corpus) and in *Figure 21.13* (test corpus)[21].

To sum up, the results of this experiment suggest that the location of words and phrases provides useful evidence for recognizing and weighting index terms. Naturally the small size of the corpus (six documents and 518 paragraphs) must be taken into account when these results are considered, and in a different genre the results would possibly be different. In any case, the indexer of this thesis uses this evidence for weighting index terms. Location tags (`<HEADLINE>`, `<PAR1>`,

[21]Note that the scales are different in *Figure 21.12* and *Figure 21.13*. The average term frequencies of the training corpus are higher than the average term frequencies of the test corpus due to some very high frequency main topics of the training corpus.

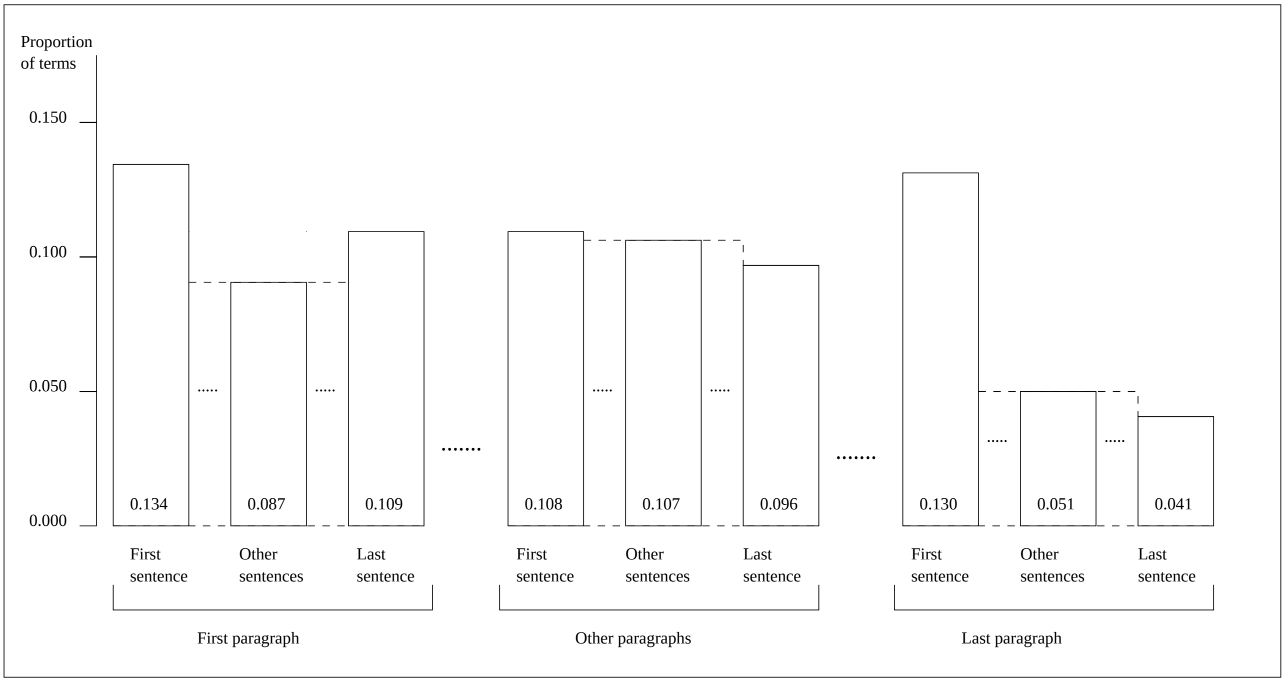

Figure 21.11: Index term density in the test corpus.

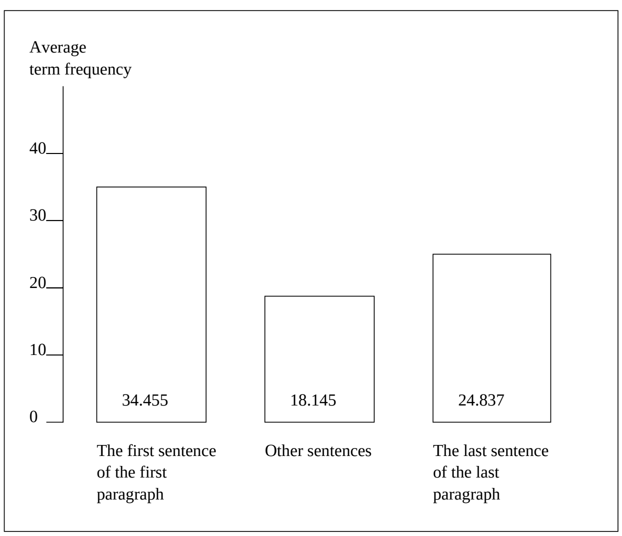

Figure 21.12: Average term frequencies in the training corpus.

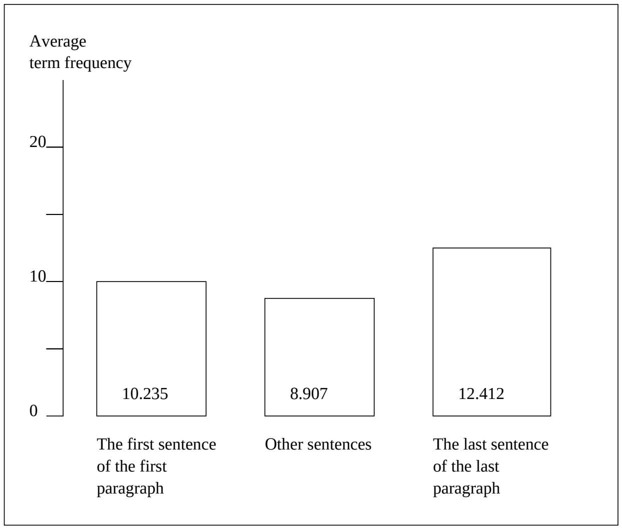

Figure 21.13: Average term frequencies in the test corpus.

etc) were treated in the same way as tags provided by the parser. One subject of future research will be to develop different techniques for combining the location evidence with other evidence.

Chapter 22

Tag weights distinguish between terms and non-terms

As discussed earlier (*Chapter 17*), tag weights (*TW*) combine all evidence provided by tag lists, that is, *TW* combines the linguistic evidence (tags provided by the parser and the location tags), but *TW* does not use evidence from burstiness. The automatic indexer observes tag combinations of words and weights term candidates accordingly. This chapter will evaluate how well tag weights based on training corpus distinguish between terms and non-terms in the test material. Tag weights were given to each single-word and multi-word term candidate. In texts the *TW* values among the occurrences of a given term candidate varied, but the highest *TW* value of the term candidate (*MAX-TW*) was chosen[1]. So, the output was a list of unique term candidates ranked by their *MAX-TW* values. Since index terms were marked up in the test corpus as well, it was possible to evaluate how reliably the index terms were ranked highest. The evaluation of the ranked term candidate lists is illustrated by recall-precision curves, with the points representing the level of precision (the number of found terms divided by the number of scanned words) at different recall percentages (the number of found terms divided by the total number of terms). In the optimal case the precision would be 100 % all the way.

As *Figure 22.1* indicates, *TW* values based on the training corpus predict the index-term-likeness of the term candidates rather well. Both single-word and multi-word terms are included and three *MAX-TW* curves are presented:

- training corpus,
- test corpus, and
- Grolier

[1]Another solution would be to calculate an average weight to an index term using all *TW* values of the term candidate. Both solutions were evaluated in this study and no great differences appeared between the results.

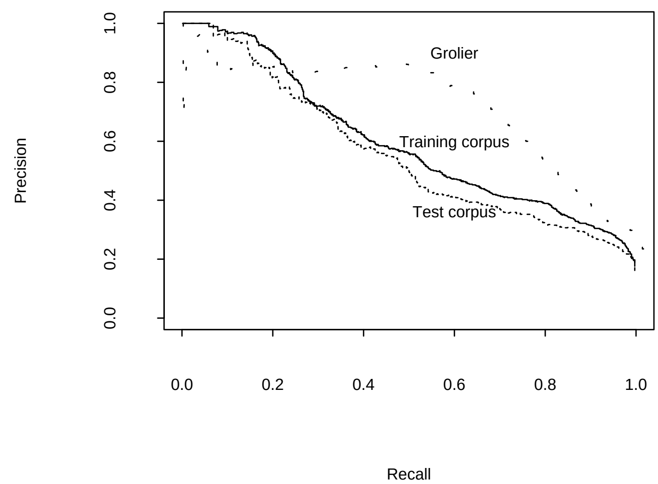

Figure 22.1: Evaluation of tag weights (*MAX-TW* of the test corpus - all terms).

There is no big difference between the curve of the training corpus and the curve of the test corpus, and the Grolier curve is even better than the other curves. At the point of 0.5-recall, the precision is still 0.560 with the training corpus, 0.504 with the test corpus, and 0.854 with Grolier. In other words, when half of the terms are found, more than half of the scanned term candidates are terms. If all the term candidates of these three corpora were ordered randomly, the precision would be 0.177, since the total number of terms in these corpora is 2788 and the total number of term candidates is 15766 (2788/15766=0.177).

Somewhat surprisingly the precision-values of Grolier are highest. One explanation is that Grolier includes more proper names than the training corpus and the test corpus. As discussed above, proper names are usually marked up as index terms and they are relatively easy to identify. Another explanation is that in Grolier the index term density is higher in general, since encyclopedic texts tend to contain much information in a compact form. The higher the index term density, the easier the identification of index terms. Anyhow, the success with Grolier suggests that the tag weights based on the training corpus can distinguish between terms non-terms rather well even in a case in which the index terms are not determined by using previously generated index, and in which the discussed topics and the style of writing are more or less different than in the training corpus.

The following ten examples are taken from the ranked list of the test corpus. The interval between the ranks is 72, so that the recall of the tenth example has reached the point of 0.5-recall.

A plus sign (`+`) refers to a term, and a minus sign (`-`) to a non-term; the `MAX-TW`-column contains the highest tag weight value of the candidate in the test corpus, the `RECALL`-column shows the increasing recall, the `PRECISION`-column shows the decreasing precision, and the `RANK`-column shows the rank of the candidate:

```
  TERM CANDIDATE                   MAX-TW  RECALL  PRECISION RANK
-------------------------------------------------------------------
+ Simone de Beauvoir               1.000   0.002   1.000     1
+ Laporte Chemical                 0.942   0.106   0.945     73
+ morality                         0.654   0.191   0.855     145
+ fantasy representation           0.508   0.249   0.747     217
- comparison group                 0.476   0.311   0.699     289
- decision-making procedure        0.443   0.352   0.634     361
+ ethnographic account             0.420   0.392   0.589     433
- thought-experiments              0.407   0.437   0.562     505
- work                             0.394   0.475   0.536     577
- workplace                        0.394   0.500   0.501     649
```

The index terms are ranked significantly higher than non-terms (Mann Whitney's U, $p > 0.95$). In this list the 45 first term candidates are all terms. *Appendix 3* presents a list of the top 100 term candidates of the test corpus, ranked by the *MAX-TW* values.

So, tag weights distinguish between terms and non-terms, but do they distinguish between important terms and less important terms? As discussed above, the location of terms (titles, first and last sentences and paragraphs) provides evidence that distinguishes between important terms and less important terms, but the location of terms is only one criterion to weight terms by their tag lists. In the Grolier corpus the importance of index terms was estimated and marked up on a scale from one to three. It is possible then to evaluate the capability of tag weights to distinguish between important terms and less important terms. The question is, are the tag weights of more important terms higher than the tag weights of less important terms.

Figure 22.2 suggests that tag weights do not distinguish between important terms and less important terms particularly well. The figure presents the average tag weights of non-terms (0), passing concepts and proper names (1), subtopics (2), and main topics (3). Both single-word and multi-word terms are included. The average tag weight of non-terms is lowest, as expected, but the average tag weight of main topics is not highest, as one would expect it to be. Again the proper names explain the results. Grolier includes a number of proper names that have been marked up as subtopics by the research aide, since even though they have some importance in the article, they are not the main topics. The tag weights of proper names are always high, and so the average tag weight of subtopics is high as well.

The following matrix shows the observed *MAX-TW* value distributions of the non-terms (0), the passing concepts and proper names (1), the subtopics (2), and the main topics (3) of Grolier. Both single-word and multi-word terms are included. For convenience, the *MAX-TW* values are rounded to the nearest tenth:

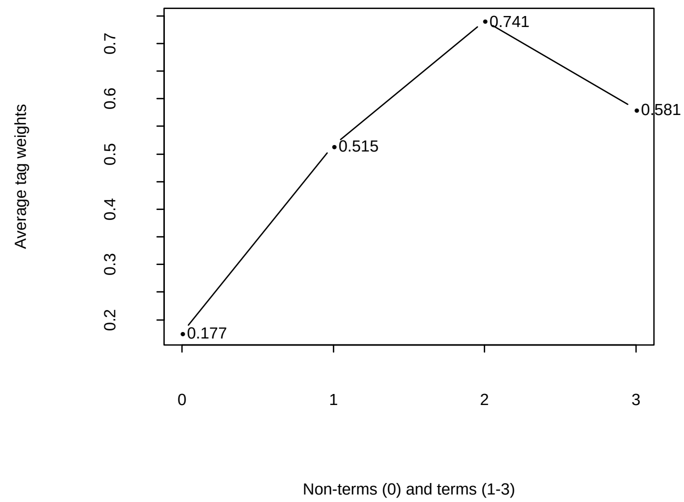

Figure 22.2: Average tag weights in Grolier (all terms). 0=non-terms 1=passing concepts and proper names 2=subtopics 3=main topics.

MAX-TW	0	1	2	3
	NUMBER OF NON-TERMS AND TERMS:			
0.0	1014	11	6	0
0.1	728	30	28	9
0.2	249	9	9	1
0.3	344	28	18	15
0.4	293	38	36	10
0.5	56	10	12	9
0.6	2	4	7	1
0.7	24	6	8	5
0.8	47	23	32	3
0.9	48	42	144	19
1.0	27	13	81	9
SUM:	2832	214	381	81

As the matrix illustrates, if, for example, term candidates with *MAX-TW* value less than 0.5 were excluded from the index, 2628 non-terms and 248 terms (116 passing concepts and proper names (1), 97 subtopics (2), and 35 main topics (3)) would be excluded, and 204 non-terms and 428 terms (98 passing concepts and proper names (1), 284 subtopics (2), and 46 main topics (3)) would be included. So, this threshold would filter a great deal of noise, but 35 main topics as well.

The recall of all terms would be 0.63 (428/676) and the proportion of terms, that is, the precision, would be 0.68 (428/632). If the threshold was 0.1, however, 1014 non-terms and *no* main topics would be excluded.

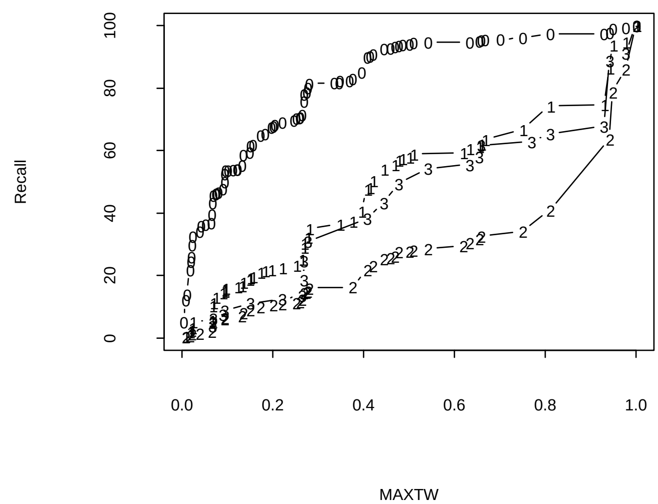

Figure 22.3: Recall of *MAX-TW* (single-word and multi-word term candidates of Grolier). 0-line=non-terms, 1-line=passing concepts and proper names, 2-line=subtopics, 3-line=main topics.

Figure 22.3 shows the recall values (*y*-axis) of non-terms (0) and terms (1-3) at different points of *MAX-TW* values (*x*-axis). 0-line shows the recall curve of single-word and multi-word non-terms of Grolier, 1-line shows the recall curve of passing concepts and proper names, 2-line shows the recall curve of subtopics, and 3-line shows the recall curve of main topics of Grolier. For example, 94 % of non-terms (0), 58 % of passing concepts and proper names (1), 28 % of subtopics (2), and 49 % of main topics (3) have *MAX-TW* value less than or equal to 0.5. The steep shape of the recall curve of non-terms (0) indicates that the great majority of non-terms (0) have low *MAX-TW* values, as they should have. However, the recall curve of subtopics (2) should be above the recall curve of main topics (3) - not below. In other words, the main topics do not have the highest *MAX-TW* values as they should have.

To sum up, choosing the highest tag weight of a term candidate (*MAX-TW*) in a text provides a good basis to distinguish between terms (1-3) and non-terms (0), but a poor basis to distinguish between the most important terms (3) and less important terms (1 and 2). An ideal weighting scheme should weight the words and phrases so that the weights get higher along the continuum (non-terms (0), passing concepts and proper names (1), subtopics (2), and main topics (3)). The conclusion is that more evidence is needed. In the following chapter, evidence from tag weights will be combined with evidence from burstiness.

Chapter 23

Burstiness distinguishes between important terms and less important terms

23.1 Within-document burstiness

In this thesis, the within-document burstiness of term candidates was measured by the method described in *Section 18.1*. The areas, however, were computed only to single-word term candidates.

Figure 23.1 indicates that measuring the within-document burstiness of term candidates does not provide sufficient evidence to distinguish between terms and non-terms. The evaluation of two ranked word lists is illustrated by recall-precision curves, with the points representing the level of precision at different recall percentages. In the `Most bursty`-list (dotted line in *Figure 23.1*) the single-word term candidates of the test corpus are ranked in increasing order of areas, that is, the most bursty words are ranked highest. In the case that many words have the same *BURST* value, the more frequent words are ranked higher. In the `Least bursty`-list (solid line in *Figure 23.1*) the same term candidates are ranked in decreasing order of areas, that is the least bursty words are ranked highest. The result is poor either way. All the 1058 words that occur only once in the test corpus were excluded, since the method measures burstiness, and there is no point to measure the burstiness of the words that occur only once.

The following list includes the ten least bursty (rank 1-10) and the ten most bursty (rank 968-977) words of the test corpus. Again the words that occur only once in the test corpus were excluded. In the case that many words have the same *BURST* value, the less frequent words are ranked higher. Altogether 385 words out of the total 977 words had the *BURST* value 0, and 59 of them were marked up as index terms (59/385=0.15). The words with the *BURST* value 0 are words that occur only in one paragraph. A plus sign (`+`) refers to a term, and a minus sign (`-`) to a non-term. The `FREQ`-column shows the frequency of the word, and the `BURST`-column shows

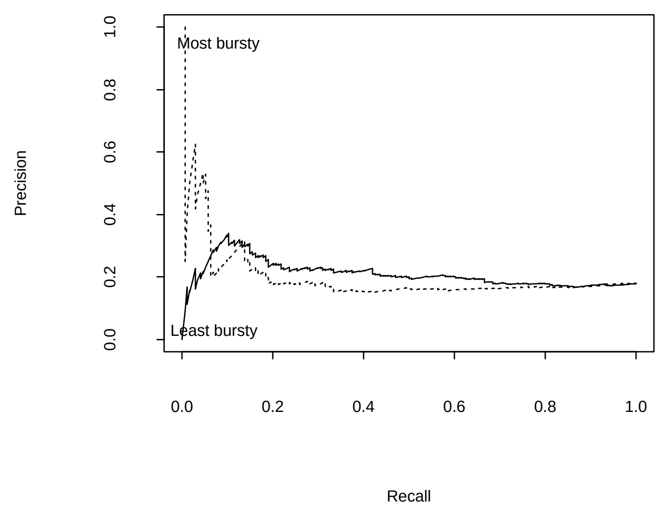

Figure 23.1: Recall and precision of within-document burstiness (single-word term candidates of the test corpus).

the burstiness of the word; the higher the *BURST* value, the less bursty word.

TERM CANDIDATE	FREQ	BURST	RECALL	PRECISION	RANK
the least bursty:					
- woman	137	6.694	0.000	0.000	1
- work	80	6.128	0.000	0.000	2
- man	77	5.981	0.000	0.000	3
- life	55	5.409	0.000	0.000	4
- see	38	5.340	0.000	0.000	5
- social	47	4.733	0.000	0.000	6
- argue	40	4.088	0.000	0.000	7
- case	32	3.564	0.000	0.000	8
- view	36	3.461	0.000	0.000	9
- give	29	2.932	0.000	0.000	10
.					
.					
.					
the most bursty:					
- impulse	4	0.000	0.971	0.175	968
- dilemma	4	0.000	0.971	0.174	969
+ Luton	4	0.000	0.977	0.175	970
+ Kafka	4	0.000	0.983	0.176	971
+ Nozick	5	0.000	0.989	0.177	972
+ irrational	5	0.000	0.994	0.178	973
- Hammertown	5	0.000	0.994	0.178	974
- allow	5	0.000	0.994	0.177	975
- population	6	0.000	0.994	0.177	976
+ Rawls	6	0.000	1.000	0.178	977

Appendix 4 presents a list of the 100 least bursty term candidates of the test corpus and *appendix 5* presents a list of the 100 most bursty term candidates of the test corpus. The lists include only single word term candidates.

A number of least bursty words, such as `and`, `the`, and `be` were automatically excluded by choosing the same term candidates that are used by the other weighting schemes (*MAX-TW*, *STW*IDF*, and *TF*IDF*). The poor precision values in the list above suggest that within-document burstiness alone provides insufficient evidence for distinguishing between terms and non-terms. In some cases, however, evidence from within-document burstiness seems to distinguish between terms and non-terms more accurately than evidence based on tag combinations (*MAX-TW*), evidence based on document-level burstiness (*TF*IDF*), or evidence based both on tag combinations and on document-level burstiness (*STW*IDF*). The following matrix presents ten such examples from the test corpus. The words are within-document bursty index terms which are weighted relatively low by *MAX-TW*, *STW*IDF*, and *TF*IDF*, that is, these weighting schemes do not identify these terms accurately[1].

[1]The *BURST* values vary from 0 to 6.694 in the test corpus, and the values of all these ten examples are zero or

INDEX TERM	FREQ	DISTR	BURST	MAX-TW	STW*IDF	TF*IDF
surface	2	199	0.037	0.067	0.129	1.348
age	2	309	0.042	0.375	0.298	1.063
penetrate	2	62	0.000	0.019	0.047	2.467
authority	2	184	0.000	0.111	0.218	1.635
body	4	282	0.005	0.171	0.544	1.736
conservative	3	93	0.000	0.092	0.396	2.687
contradiction	2	26	0.000	0.065	0.306	3.302
slavery	2	43	0.000	0.394	1.431	3.327
contradictory	2	23	0.000	0.093	0.446	3.419
media	2	51	0.000	0.171	0.618	2.655

The `FREQ`-column shows the frequency of the index term in the test corpus and the `DISTR`-column shows in how many documents out of 810 the index term occurs. All ten index terms have a low frequency in the test corpus and they all occur in a number of documents. Thus their *TF*IDF* values are low, that is, they are not document-level bursty words in the document collection, even though they are within-document bursty words in the test corpus (*BURST* values $\cong 0$). The *MAX-TW* values are relatively low as well, that is, the syntactic functions, the lexical features, and the locations of these words do not provide enough evidence for identifying them with index terms. Consequently, the *STW*IDF* values based on *TW* values and document-level burstiness are low as well. So, these examples indicate that in some cases evidence from within-document burstiness could be helpful to identification of index terms.

On the other hand, *Figure 23.2* suggests that measuring the within-document burstiness of term candidates could provide evidence to distinguish between subtopics (2) and main topics (3). The figure presents the average within-document burstiness (computed areas) of non-terms (0), passing concepts and proper names (1), subtopics (2), and main topics (3). Only the single-word terms of the longest Grolier article (67 paragraphs) were included[2], and all the 873 words that occur only once in the article were excluded. *The higher the average value, the less bursty group.* The average value of subtopics is lowest, that is, subtopics are the most bursty words, on average. The average value of main topics is highest, that is, main topics are the least bursty words, on average. However, the article included only 15 main topics and 44 subtopics, which must be taken into account when the significance of the results is considered.

almost zero, that is, they are all within-document bursty words. The *MAX-TW* values vary from 0.003 to 1.000 in the test corpus, the *STW*IDF* values vary from 0.002 to 14.190 in the test corpus, and the *TF*IDF* values vary from 0.302 to 16.710 in the test corpus.

[2]Short articles are inappropriate data for this method which measures within-document burstiness by observing paragraph distances. In short articles all distances are short.

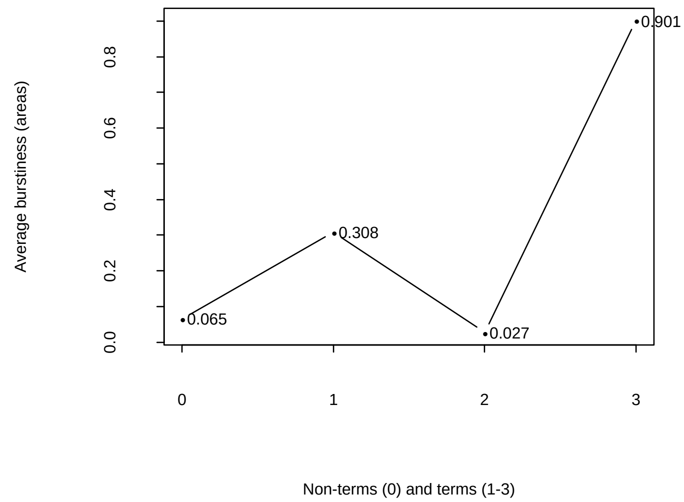

Figure 23.2: Average within-document burstiness (i.e. average areas) in Grolier. **NOTE:** The higher the average value, the less bursty group. 0=non-terms 1=passing concepts and proper names 2=subtopics 3=main topics.

The following matrix shows the observed *BURST* value (area) distributions of the non-terms (0), the passing concepts and proper names (1), the subtopics (2), and the main topics (3) of the longest Grolier article. The higher the *BURST* value, the less bursty word. Only single-word terms are included. For convenience, the *BURST* values are rounded to the nearest tenth:

NUMBER OF NON-TERMS AND TERMS:				
BURST (AREA)	0	1	2	3
0.0	201	18	32	6
0.1	42	2	7	1
0.2	40	1	1	0
0.3	12	2	1	2
0.4	7	4	0	0
0.5	9	1	2	0
0.6	4	0	0	0
0.7	3	1	0	1
0.8	2	0	0	0
0.9	4	1	0	0
1.0	1	2	0	0
1.1	2	1	1	0
1.2	1	0	0	0
1.3	0	1	0	0
1.4	3	0	0	0
1.5	2	0	0	0
2.0	2	0	0	1
2.1	1	0	0	0
2.2	1	0	0	0
2.6	1	0	0	0
2.7	1	0	0	0
2.8	0	0	0	1
2.9	0	0	0	1
3.0	1	0	0	0
3.1	1	0	0	0
3.4	0	0	0	1
5.0	1	0	0	0
9.2	0	0	0	1
SUM:	342	34	44	15

The title of the article was *American art and architecture*. The least bursty words (high *BURST* value) included main topics, such as `art` (frequency of 42) and `architecture` (frequency of 25), and non-terms, such as `new` (frequency of 41) and `work` (frequency of 25)[3]. On the other hand, the most bursty words (*BURST* value zero) included six main topics as well, such as `expressionism` (frequency of 3) and `impressionism` (frequency of 3). These words were considered as main topics even though they occurred only locally in the discourse. In a way, however, they are not perhaps such main topics as `art` and `architecture` which are used throughout the text. In addition, the most bursty words (*BURST* value zero) included 201 bursty non-terms, such as `national` (frequency of 4) and `meet` (frequency of 3).

So, measuring the within-document burstiness of term candidates could provide an appropriate approach to distinguish between subtopics and main topics, or between local and global topics, if this method was combined with some other method that could recognize non-terms more accurately. In some cases local and global topics can be distinguished by the frequencies of the

[3]A number of least bursty words, such as `and`, `the`, and `be` were automatically excluded by choosing the same term candidates that are used by the other weighting schemes (*TW*, *STW*IDF*, and *TF*IDF*).

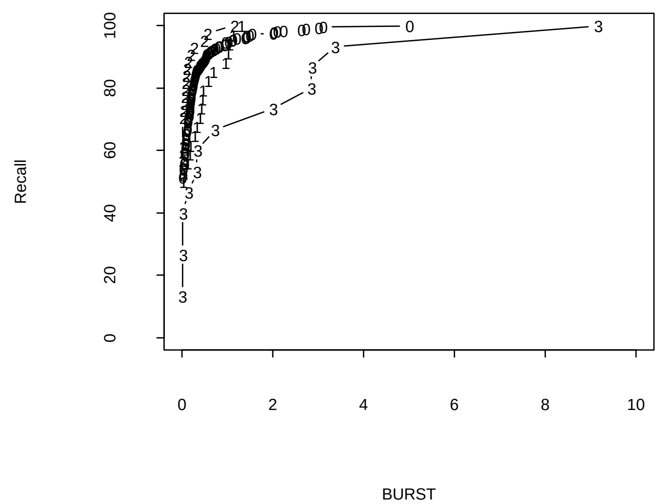

Figure 23.3: Recall of *BURST* values (areas of single-word term candidates of Grolier). 0-line=non-terms, 1-line=passing concepts and proper names, 2-line=subtopics, 3-line=main topics.

words; for example, the local topic `expressionism` has the frequency of 3 and the global topic `architecture` has the frequency of 25. The higher frequency does not, however, always indicate that the word is a main topic. For example, the main topic `painter` is used throughout the text, and it has the *BURST* value of 1.0 and its frequency is 9, whereas the local subtopic `Whistler` has the *BURST* value of 0.0 and its frequency is 11. The word frequency is an insufficient criterion for recognizing those main topics which are used throughout the text, but which are not mentioned very frequently, for some reason. This kind of main topics may have been referred to with a synonym or a pronoun, for example.

Figure 23.3 shows the recall values (*y*-axis) of non-terms (0) and terms (1-3) at different points of *BURST* values (*x*-axis). 0-line shows the recall curve of single-word non-terms of the longest Grolier article, 1-line shows the recall curve of passing concepts and proper names, 2-line shows the recall curve of subtopics, and 3-line shows the recall curve of main topics. For example, 92 % of non-terms (0), 82 % of passing concepts and proper names (1), 98 % of subtopics (2), and 60 % of main topics (3) have *BURST* value less than or equal to 0.6. The recall curve of subtopics (2) is the highest curve, which indicates that subtopics are the most bursty words. The recall curve of main topics (3) is the lowest curve, which indicates that main topics are the least bursty words. However, the recall curves of non-terms (0), passing concepts and proper names (1), and subtopics (2) are quite similar, which indicates that this method is not able to distinguish between these groups accurately.

To sum up, the method described above distinguishes between bursty words and words used throughout the text. The results suggest that this distinction could be helpful to identification of index terms as well as to classification of index terms, if the method was combined with some other methods based on, for instance, detection of document-level burstiness and tag combinations. More extensive experiments are needed, however, in order to fully evaluate the usefulness of this method to automatic indexing.

23.2 Document-level burstiness

In this section two weighting schemes, *STW*IDF* and *TF*IDF*, will use document-level burstiness as evidence for weighting term candidates. Both weighting schemes use the term candidates provided by the pattern matching method, with one exception: the set of two-word term candidates will be created by two ways. One is to use the pattern matching method and the other is to use the simple phrase constructing method described in *Section 13.4*. Furthermore, both weighting schemes use the base forms provided by the parser.

23.2.1 Terms and non-terms

Figure 23.4 presents the evaluation of three ranked term candidate lists of the test corpus by recall-precision curves, with the points representing the level of precision at different recall percentages: *TF*IDF* (term candidates provided by the pattern matching method), *STW*IDF*, and the highest tag weight values (*MAX-TW*). Both single-word and multi-word terms are included. The figure reveals that *MAX-TW* and *STW*IDF* distinguish between terms and non-terms better than *TF*IDF*. At the point of 0.5-recall, the precision is still 0.561 with *STW*IDF*, and 0.504 with *MAX-TW*, whereas it is only 0.239 with *TF*IDF*.

The following ten *TF*IDF*-examples are taken from the ranked list of the test corpus. The interval between the ranks is 149, so that the recall of the tenth example has reached the point of 0.5-recall. A plus sign (`+`) refers to a term, and a minus sign (`-`) to a non-term; the `FREQ`-column shows the frequency of the term candidate in the test corpus, the `DISTR`-column shows in how many documents out of 810 the term candidate occurs, the `TF*IDF`-column contains the *TF*IDF* value of the candidate in the test corpus, the `RECALL`-column shows the increasing recall, the `PRECISION`-column shows the decreasing precision, and the `RANK`-column shows the rank of the candidate:

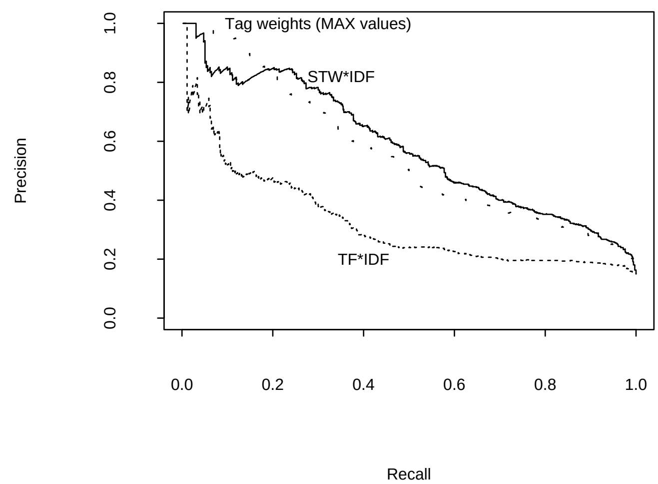

Figure 23.4: Recall and precision of *TF*IDF*, *STW*IDF* and *MAX-TW* (single-word and multi-word terms of the test corpus).

TERM CANDIDATE	FREQ	DISTR	TF*IDF	RECALL	PRECISION	RANK
+ counter-school	21	1	16.710	0.002	1.000	1
- life of woman	2	1	7.388	0.116	0.493	150
+ critical case study	2	2	5.764	0.217	0.465	299
- Kant argument	1	1	4.678	0.286	0.408	448
- feminist writer Marge Piercy	1	1	4.678	0.330	0.353	597
- personal anecdote	1	1	4.526	0.370	0.318	746
+ moral response	1	1	4.526	0.397	0.284	895
- perspective be replace	1	1	4.526	0.430	0.263	1044
+ moral presumption	1	1	4.526	0.461	0.247	1193
- inversion	2	10	4.219	0.502	0.239	1342

The index terms are ranked significantly higher than non-terms (Mann Whitney's U, $p > 0.95$). *Appendix 6* presents a list of the top 100 term candidates of the test corpus, ranked by the *TF*IDF* values. The term candidate `Kant argument` is in the text in the form `Kant's argument`, and the term candidate `perspective be replace` is in the text in the form `perspective is replaced`. In the ranked lists words are in their base forms.

The next ten examples from the test corpus are ranked by *STW*IDF* values. The interval between the ranks is 64, so that the recall of the tenth example has reached the point of 0.5-recall. A plus sign (`+`) refers to a term, and a minus sign (`-`) to a non-term; the `STW`-column contains the summed tag weights of the term candidate in the test corpus, and the `STW*IDF`-column contains the *STW*IDF* value of the candidate in the test corpus:

	TERM CANDIDATE	STW	STW*IDF	RECALL	PRECISION	RANK
+	Willis	34.618	14.190	0.002	1.000	1
-	the affluent worker	1.665	5.071	0.084	0.831	65
+	neo-capitalism	1.000	3.827	0.167	0.829	129
+	love	2.382	3.156	0.250	0.829	193
+	pornographic magazine	0.485	2.578	0.306	0.763	257
-	male researcher	0.443	2.293	0.355	0.707	321
+	male domination	0.508	2.145	0.394	0.655	385
-	reductionist view	0.467	1.989	0.431	0.615	449
+	contractual relation	0.416	1.883	0.473	0.591	513
-	grandmother	0.375	1.771	0.502	0.556	577

The index terms are ranked significantly higher than non-terms (Mann Whitney's U, $p > 0.95$). *Appendix 7* presents a list of the top 100 term candidates of the test corpus, ranked by the *STW*IDF* values.

So, it seems that use of the tag list information (*STW*IDF*) improves the capability of distinguishing between terms and non-terms. One reason for this is that documents included a lot of term candidates that appeared in the entire document collection only once. For example, in the test corpus 1675 term candidates out of the total 4281 term candidates appeared only once (`FREQ`=1 and `DISTR`=1). 225 of them were marked up as terms, so it is not feasible to exclude them from the set of term candidates. With burstiness-based methods it is impossible to distinguish between terms and non-terms if they occur in the document collection only once. This problem is demonstrated in *Figure 23.5* which presents the recall-precision curves of three ranked term candidate lists of Grolier. Both single-word and multi-word terms are included. *MAX-TW* values rank the term candidates of Grolier even better than the term candidates of the test corpus. However, the results of burstiness-based weighting schemes are worse with Grolier, the results of *TF*IDF* in particular. As discussed earlier, an explanation for good results with Grolier is that Grolier includes more proper names than the test corpus, and proper names are usually terms and easy to identify. On the other hand, an explanation for poor results with burstiness-based weighting schemes is that most documents in Grolier are short encyclopedia articles that do not necessarily contain a lot of repetition of term candidates, that is, burstiness.

The problem of single occurrence is a problem of multi-word terms in particular. Most of these single-occurrence candidates are multi-word terms. So, perhaps the results of *TF*IDF* would be better if the multi-word terms were excluded? *Figure 23.6* presents the recall-precision curves of the ranked term candidate lists of the test corpus. In this case only single-word terms are included. The results are slightly better with the burstiness-based weighting schemes (*TF*IDF* and *STW*IDF*) here than in the *Figure 23.4* which presents the results with all term candidates. The difference, however, is not great. To sum up, in these experiments the weighting schemes

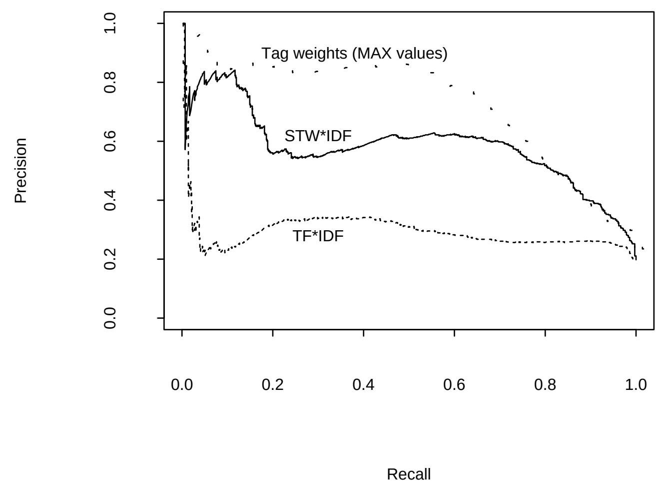

Figure 23.5: Recall and precision of *TF*IDF*, *STW*IDF* and *MAX-TW* (single-word and multi-word terms of Grolier).

based on tag lists distinguished between terms and non-terms better than the weighting scheme based on burstiness only.

23.2.2 Two-word terms

In this section two methods of creating the set of two-word term candidates will be compared:

- pattern matching method and
- the simple method of constructing phrases using a stoplist

The pattern matching method uses two-word term patterns based on the training corpus. This method produced 1170 term candidates from the test corpus out of which 660 appeared in the entire document collection (810 documents) only once. The simple method of constructing phrases considers all adjacent pairs of base forms of non-stopwords as two-word term candidates. This method produced 2075 term candidates from the test corpus out of which 1302 appeared in the entire document collection only once. So, the pattern matching method produced only half of the term candidates produced by the simple method. However, all the 1039 extra term candidates produced by the simple method were non-terms, that is, noise.

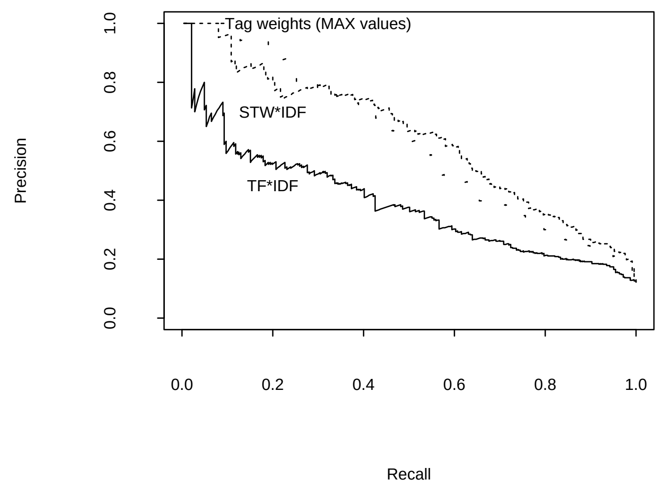

Figure 23.6: Recall and precision of *TF*IDF*, *STW*IDF* and *MAX-TW* (only single-word terms of the test corpus).

Figure 23.7 presents the recall-precision curves of the three ranked term candidate lists of the test corpus:

- *TF*IDF* (all bigrams): the simple method,
- *TF*IDF* (patterns): the pattern matching method, and
- *STW*IDF*

The pattern matching method outperforms the simple method in this experiment. Again the problem of single occurrences must be considered. Usually the simple method excludes the low-frequency term candidates automatically. However, if those term candidates that occur only once in the entire document collection were excluded, it would mean in this case that 126 terms out of the total 292 terms were excluded as well. For the sake of the recall, the single-occurrence term candidates were included in this experiment.

The following ten *TF*IDF*-examples are taken from the ranked list where the term candidates have been produced by the simple method from the test corpus. The interval between the ranks is 96, so that the recall of the tenth example has reached the point of 0.5-recall. A plus sign (`+`) refers to a term, and a minus sign (`-`) to a non-term; the `FREQ`-column shows the frequency of the term candidate in the test corpus, the `DISTR`-column shows in how many documents out of 810

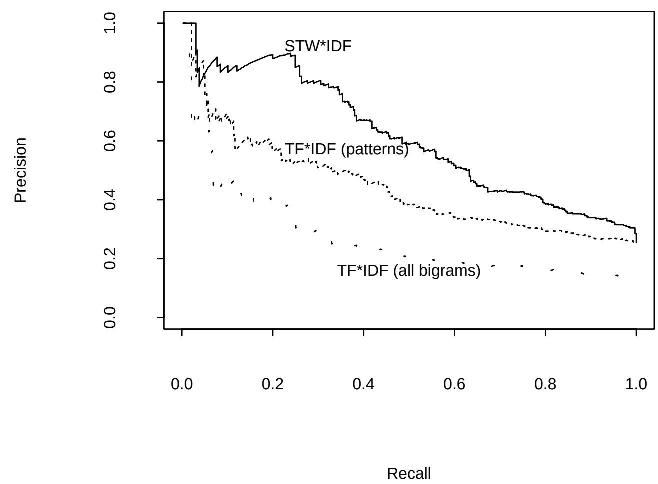

Figure 23.7: Recall and precision of *TF*IDF* (simple method), *TF*IDF* (pattern matching method), and *STW*IDF*. Two-word terms of the test corpus.

the term candidate occurs, the `TF*IDF`-column contains the *TF*IDF* value of the candidate in the test corpus, the `RECALL`-column shows the increasing recall, the `PRECISION`-column shows the decreasing precision, and the `RANK`-column shows the rank of the candidate:

```
  TERM CANDIDATE             FREQ   DISTR  TF*IDF   RECALL   PRECISION  RANK
------------------------------------------------------------------------------
+ counter-school culture      21     1      16.710   0.003    1.000        1
- Hegel view                   2     2      6.803    0.130    0.392       97
- woman passive                1     1      4.678    0.233    0.352      193
- Filmer patriarcha            1     1      4.678    0.277    0.280      289
- someone_else moral           1     1      4.678    0.322    0.244      385
- Kohlberg dilemma             1     1      4.526    0.373    0.227      481
- sexual apartheid             1     1      4.526    0.390    0.198      577
- interact personal            1     1      4.526    0.411    0.178      673
+ moral response               1     1      4.526    0.469    0.178      769
- Nozick point                 1     1      4.526    0.503    0.170      865
```

The index terms are ranked significantly higher than non-terms (Mann Whitney's U, $p > 0.95$).

To sum up, the results of this experiment suggest that pattern matching method could be a useful approach to identification of multi-word term candidates - whatever is the applied weighting scheme. The pattern matching method does not depend on the size of the document or the size of the document collection: term candidates can be extracted on the basis of their tag lists.

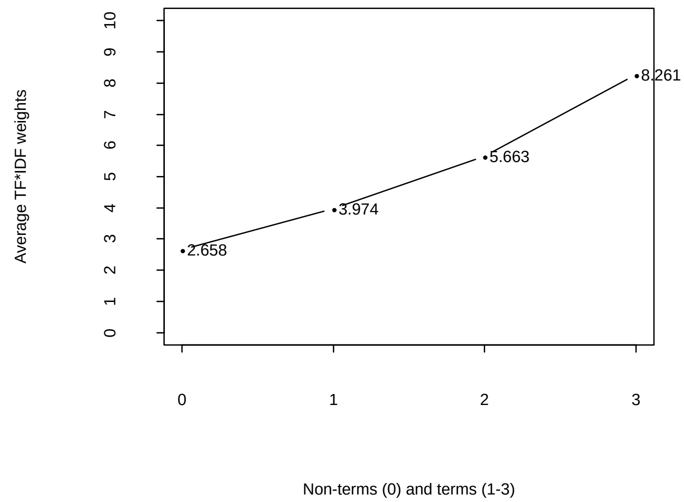

Figure 23.8: Average *TF*IDF*-weights in Grolier (single-word terms).

23.2.3 Important terms and less important terms

So far the focus has been on distinguishing between terms and non-terms. The purpose of a weighting scheme, however, is not only to distinguish between terms and non-terms, but also to weight the index terms according to their importance as descriptors of the document content. As seen above, tag weights did not distinguish accurately between subtopics and main topics, whereas within-document burstiness did.

23.2.4 Single-word terms

Figure 23.8 suggests that *TF*IDF* is useful in distinguishing between important single-word terms and less important single-word terms. The figure presents the average *TF*IDF* values of non-terms (0), passing concepts and proper names (1), subtopics (2), and main topics (3). The average *TF*IDF* value of non-terms is lowest, and the average *TF*IDF* value of main topics is highest. The average *TF*IDF* values grow as the importance of terms grows.

The following matrix shows the observed *TF*IDF* value distributions of the non-terms (0), the passing concepts and proper names (1), the subtopics (2), and the main topics (3) of Grolier (all ten articles). Only single-word terms are included. For convenience, the *TF*IDF* values are rounded to the nearest integer:

	NUMBER OF NON-TERMS AND TERMS:			
TF*IDF	0	1	2	3
0	23	0	1	0
1	464	5	2	0
2	512	19	14	0
3	283	19	40	2
4	142	20	37	1
5	68	13	20	3
6	52	7	17	9
7	28	8	9	2
8	18	4	10	1
9	15	0	15	2
10	9	0	9	1
11	11	0	13	1
12	5	1	5	1
13	4	0	3	2
14	0	0	0	2
16	0	0	0	1
17	0	0	0	1
SUM	1634	96	195	29

As the matrix illustrates, if, for example, term candidates with *TF*IDF* value less than 5 were excluded from the index, 1424 non-terms and 160 terms (63 passing concepts and proper names (1), 94 subtopics (2), and 3 main topics (3)) would be excluded, and 209 non-terms and 160 terms (33 passing concepts and proper names (1), 101 subtopics (2), and 26 main topics (3)) would be included. So, this threshold would filter a great deal of noise, but half of the index terms as well, i.e. the recall would be 0.50 (160/320). More than half of the words of the index would be non-terms. The proportion of terms, that is, the precision, would be 0.43 (160/369).

Figure 23.9 shows the recall values (*y*-axis) of non-terms (0) and terms (1-3) at different points of *TF*IDF* values (*x*-axis). 0-line shows the recall curve of single-word non-terms of Grolier (all ten articles), 1-line shows the recall curve of passing concepts and proper names, 2-line shows the recall curve of subtopics, and 3-line shows the recall curve of main topics. For example, 89 % of non-terms (0), 76 % of passing concepts and proper names (1), 53 % of subtopics (2), and 14 % of main topics (3) have *TF*IDF* value less than or equal to 5. The order of the curves from the highest curve to the lowest curve is as it should be: 0,1,2, and 3. The great majority of non-terms (0) have *TF*IDF* values less than 5, and the great majority of main topics (3) have *TF*IDF* values higher than 5.

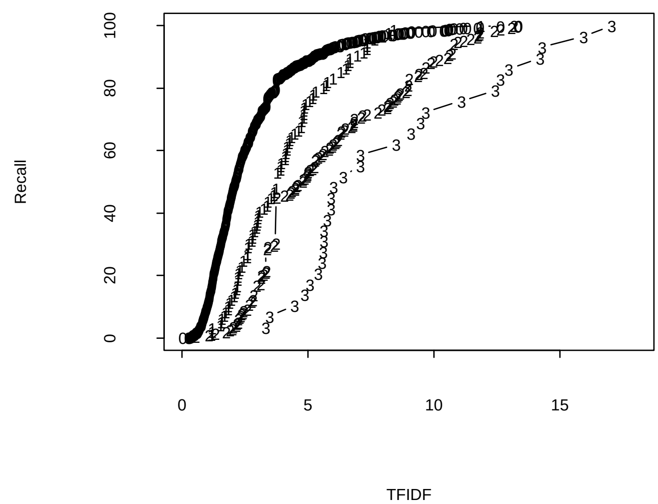

Figure 23.9: Recall of *TF*IDF* (single-word term candidates of Grolier). 0-line=non-terms, 1-line=passing concepts and proper names, 2-line=subtopics, 3-line=main topics.

Figure 23.10 indicates that also *STW*IDF* is useful in distinguishing between important single-word terms and less important single-word terms. The average *STW*IDF* value of non-terms is lowest, and it is even lower than the average *TF*IDF* value of non-terms. This observation confirms the previous result that *STW*IDF* distinguishes better between terms and non-terms than *TF*IDF*. The average *STW*IDF* values grow as the importance of terms grows, and the average *STW*IDF* value of main topics is highest, as with the average *TF*IDF* values.

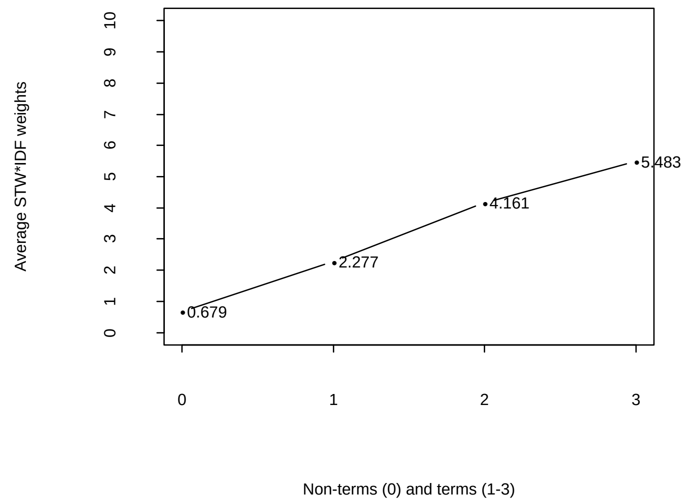

Figure 23.10: Average *STW*IDF*-weights in Grolier (single-word terms). 0=non-terms 1=passing concepts and proper names 2=subtopics 3=main topics.

The following matrix shows the observed *STW*IDF* value distributions of the non-terms (0), the passing concepts and proper names (1), the subtopics (2), and the main topics (3) of Grolier (all ten articles). Only single-word terms are included. For convenience, the *STW*IDF* values are rounded to the nearest integer:

NUMBER OF NON-TERMS AND TERMS:

STW*IDF	0	1	2	3
0	1105	11	9	0
1	306	30	21	4
2	108	16	31	3
3	54	17	42	5
4	23	13	27	2
5	14	6	14	4
6	8	1	6	2
7	5	0	8	0
8	6	1	11	3
9	1	1	13	1
10	2	0	9	0
11	2	0	4	2
12	0	0	0	1
14	0	0	0	2
SUM	1634	96	195	29

As the matrix illustrates, if, for example, term candidates with *STW*IDF* value less than 2 were excluded from the index, 1411 non-terms and 75 terms (41 passing concepts and proper names (1), 30 subtopics (2), and 4 main topics (3)) would be excluded, and 223 non-terms and 245 terms (55 passing concepts and proper names (1), 165 subtopics (2), and 25 main topics (3)) would be included. So, this threshold would filter a great deal of noise, but 75 index terms as well: the recall would be 0.77 (245/320). However, more than half of the words of the index would be index terms; the precision would be 0.52 (245/468). On the other hand, if the term candidates with *STW*IDF* value less than 1 were excluded from the index, 1105 non-terms and only 20 index terms would be excluded. No main topics would be excluded. In this case the recall would be rather high (300/320=0.94) and the precision would still be 0.36 (300/829).

Figure 23.11 shows the recall values (*y*-axis) of non-terms (0) and terms (1-3) at different points of *STW*IDF* values (*x*-axis). 0-line shows the recall curve of single-word non-terms of Grolier (all ten articles), 1-line shows the recall curve of passing concepts and proper names, 2-line shows the recall curve of subtopics, and 3-line shows the recall curve of main topics. For example, 91 % of non-terms (0), 50 % of passing concepts and proper names (1), 25 % of subtopics (2), and 21 % of main topics (3) have *STW*IDF* value less than or equal to 2. The order of the curves from the highest curve to the lowest curve is as it should be: 0,1,2, and 3. The great majority of non-terms (0) have *STW*IDF* values less than 2, and the great majority of main topics (3) have *STW*IDF* values higher than 2.

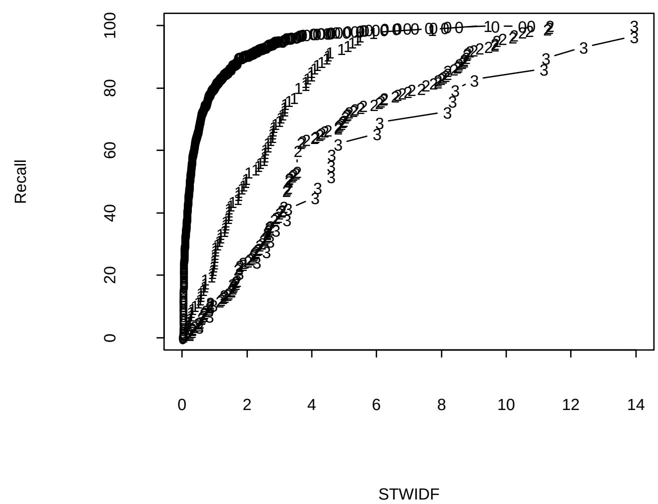

Figure 23.11: Recall of *STW*IDF* (single-word term candidates of Grolier). 0-line=non-terms, 1-line=passing concepts and proper names, 2-line=subtopics, 3-line=main topics.

23.2.5 Multi-word terms

So, as far as the single-word terms are concerned, both *TF*IDF* and *STW*IDF* distinguish between important terms and less important terms. However, when the same experiment is done with multi-word terms, *STW*IDF* outperforms *TF*IDF*. *Figure 23.12* indicates that *TF*IDF* can identify multi-word main topics, but other multi-word terms and non-terms have about the same average values. On the other hand, as with single-word terms, also with multi-word terms the average *STW*IDF* values grow as the importance of terms grows. The average *STW*IDF* value of non-terms is clearly lowest and the average *STW*IDF* value of main topics is highest, as presented in *Figure 23.13*.

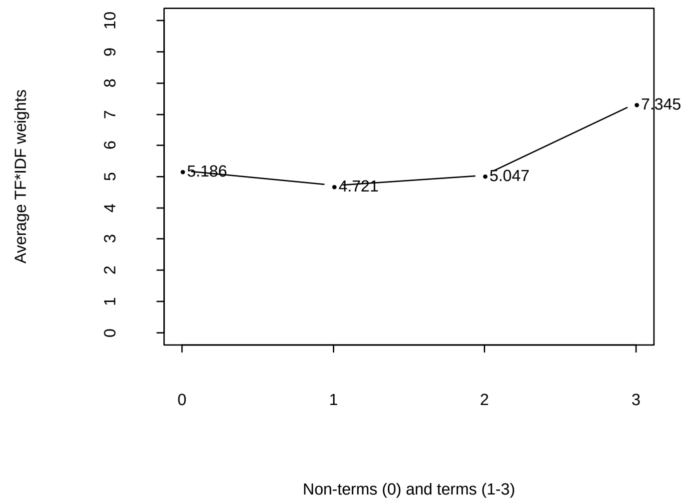

Figure 23.12: Average *TF*IDF*-weights in Grolier (multi-word terms). 0=non-terms 1=passing concepts and proper names 2=subtopics 3=main topics.

The following matrix shows the observed *TF*IDF* value distributions of the non-terms (0), the passing concepts and proper names (1), the subtopics (2), and the main topics (3) of Grolier (all ten articles). Only multi-word terms are included. For convenience, the *TF*IDF* values are rounded to the nearest integer:

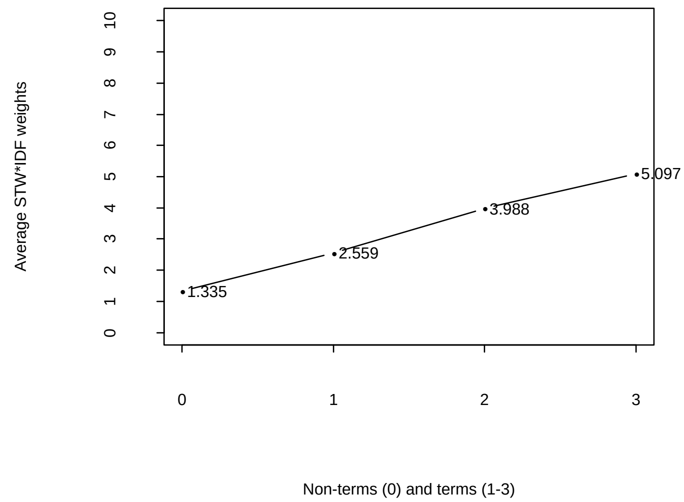

Figure 23.13: Average *STW*IDF*-weights in Grolier (multi-word terms). 0=non-terms 1=passing concepts and proper names 2=subtopics 3=main topics.

NUMBER OF NON-TERMS AND TERMS:

TF*IDF	0	1	2	3
0	1	0	0	0
1	7	0	0	0
2	54	4	2	1
3	166	35	51	4
4	665	47	88	5
5	17	5	3	7
6	26	8	1	7
7	8	0	0	2
8	18	4	1	7
9	12	2	14	4
10	18	2	5	6
11	94	6	9	6
12	80	4	8	1
13	30	0	4	1
14	1	0	0	1
15	1	0	0	0
16	0	1	0	0
SUM	1198	118	186	52

As the matrix illustrates, if, for example, term candidates with *TF*IDF* value less than 5 were excluded from the index, 893 non-terms and 237 terms (86 passing concepts and proper names (1), 141 subtopics (2), and 10 main topics (3)) would be excluded, and 305 non-terms and 119 terms (32 passing concepts and proper names (1), 45 subtopics (2), and 42 main topics (3)) would be included. So, this threshold would filter a great deal of noise, but a great deal of the index terms as well: the recall would be only 0.33 (119/356), and the precision would be only 0.28 (119/424).

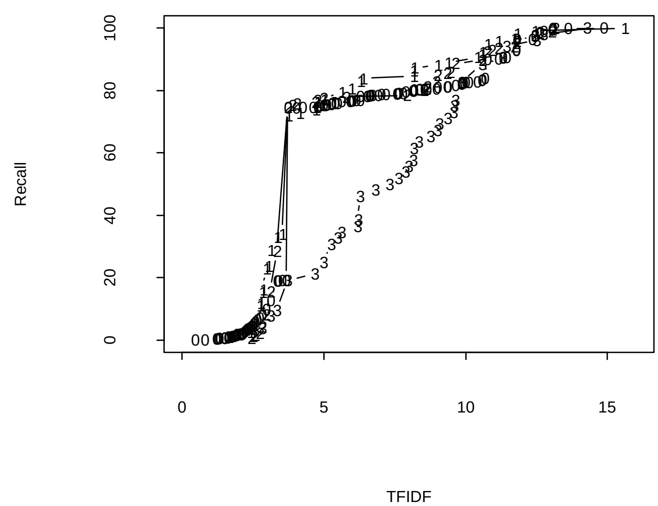

Figure 23.14: Recall of *TF*IDF* (multi-word term candidates of Grolier). 0-line=non-terms, 1-line=passing concepts and proper names, 2-line=subtopics, 3-line=main topics.

Figure 23.14 shows the recall values (y-axis) of non-terms (0) and terms (1-3) at different points of *TF*IDF* values (x-axis). 0-line shows the recall curve of multi-word non-terms of Grolier (all ten articles), 1-line shows the recall curve of passing concepts and proper names, 2-line shows the recall curve of subtopics, and 3-line shows the recall curve of main topics. For example, 75 % of non-terms (0), 77 % of passing concepts and proper names (1), 77 % of subtopics (2), and 25 % of main topics (3) have *TF*IDF* value less than or equal to 5. These curves illustrate that *TF*IDF* values do not distinguish between non-terms (0) passing concepts and proper names (1), and subtopics (2). Main topics (3), however, have a flatter curve than the other groups. The multi-word main topics are often candidates that are relatively frequent in the given document (*TF*), but relatively infrequent in the whole document collection (*IDF*). On the other hand, most of the multi-word term candidates are relatively infrequent in the whole document collection[4],

[4]Altogether, *IDF* values were calculated to 12,350 multi-word term candidates, and 10,004 (81 %) of them occurred

which means that *TF*-factor is far more important than *IDF*-factor as far as multi-word terms are concerned. Most of the multi-word non-terms, passing concepts and proper names, and subtopics are so infrequent that *TF*IDF* values do not distinguish between them.

*STW*IDF* values, however, do distinguish between different groups of multi-word term candidates. The following matrix shows the observed *STW*IDF* value distributions of the non-terms (0), the passing concepts and proper names (1), the subtopics (2), and the main topics (3) of Grolier (all ten articles). Only multi-word terms are included. For convenience, the *STW*IDF* values are rounded to the nearest integer:

```
        NUMBER OF NON-TERMS AND TERMS:
STW*IDF       0      1      2      3
---------------------------------------
0             457    11     0      0
1             438    34     15     6
---------------------------------------
2             123    14     18     5
3             55     27     41     10
4             28     16     77     5
5             31     8      5      6
6             36     2      4      5
7             16     2      3      1
8             5      2      3      5
9             4      2      10     1
10            2      0      6      6
11            1      0      4      1
12            2      0      0      1
----------------------------------------
       SUM    1198   118    186    52
```

As the matrix illustrates, if, for example, term candidates with *STW*IDF* value less than 2 were excluded from the index, 895 non-terms and 66 terms (45 passing concepts and proper names (1), 15 subtopics (2), and 6 main topics (3)) would be excluded, and 303 non-terms and 290 terms (73 passing concepts and proper names (1), 171 subtopics (2), and 46 main topics (3)) would be included. So, this threshold would filter a great deal of noise, but 66 index terms as well: the recall would be 0.81 (290/356). However, almost half of the candidates of the index would be index terms. The proportion of terms, that is, the precision, would be 0.48 (290/593).

Figure 23.15 shows the recall values (*y*-axis) of non-terms (0) and terms (1-3) at different points of *STW*IDF* values (*x*-axis). 0-line shows the recall curve of multi-word non-terms of Grolier (all ten articles), 1-line shows the recall curve of passing concepts and proper names, 2-line shows the recall curve of subtopics, and 3-line shows the recall curve of main topics. For example, 82 % of non-terms (0), 45 % of passing concepts and proper names (1), 14 % of subtopics (2), and 15 % of main topics (3) have *STW*IDF* value less than or equal to 2. The order of the curves from the highest curve to the lowest curve is as it should be: 0,1,2, and 3[5]. The great majority of

only in one document out of 810.

[5]Except that the curves of subtopics and main topics are quite similar with *STW*IDF* values less than 3.5.

non-terms (0) have *STW*IDF* values less than 2, and the great majority of main topics (3) have *STW*IDF* values higher than 2.

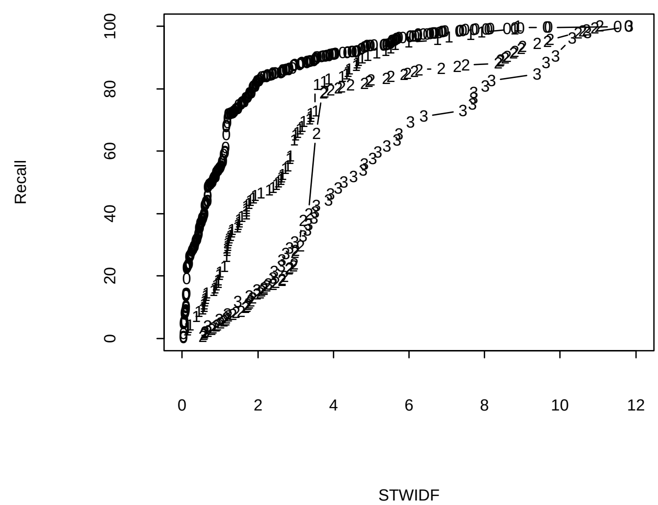

Figure 23.15: Recall of *STW*IDF* (multi-word term candidates of Grolier). 0-line=non-terms, 1-line=passing concepts and proper names, 2-line=subtopics, 3-line=main topics.

Chapter 24

Summary

To sum up, *STW*IDF* seems to be able to combine the strengths of tag weights and *TF*IDF*. Tag weights distinguish quite well between terms and non-terms, but poorly between important terms and less important terms. *TF*IDF*, on the other hand, distinguish poorly between terms and non-terms, but reasonably well between important terms and less important terms. The above described experiments suggest that *STW*IDF* distinguishes reasonably well both between terms and non-terms, and between important terms and less important terms. The combination of tag list information and analysis of burstiness seems to be then a feasible way to improve the performance of weighting schemes.

Part VI

Discussion

Chapter 25

Promising results

The results suggest that it is possible to define a number of typical features of index terms in order to develop an automatic indexer. In general, the index terms of the test corpus shared the features of the index terms of the training corpus. All texts of the corpus, however, represented more or less the same genre[1], which must partly explain the promising results. If a text of a different genre were used as test material, the results would possibly not be as good. A robust indexing tool will require a large corpus of different texts as training material.

Altogether, 89 different tag combinations were considered as relevant term patterns, of which the great majority were different simple noun phrase patterns. Index term probabilities were calculated for these combinations by using the training corpus. The automatic indexer of this thesis uses the index term probabilities of the patterns for weighting representations of the index term patterns by tag weights (*TW*). Part-of-speech tagging is the first step of determining the tag weights of term candidates, but other information in the tag lists is useful as well: syntactic functions, lexical features, and location of term candidates. The results suggested that tag weights distinguish reasonably well between terms and non-terms, but poorly between important terms and less important terms, and the conclusion was that more evidence is needed. The next step was to combine evidence from tag weights with evidence from burstiness.

Two kinds of burstiness were detected in order to weight index terms: **within-document burstiness** which refers to close proximity of individual instances of a term candidate within a document, and **document-level burstiness** which refers to multiple occurrence of a term candidate in a single document, which is contrasted with the fact that most other documents contain no instances of this candidate at all. *Section 18.1* introduced a new method of measuring within-document burstiness. This method did not distinguish between terms and non-terms particularly well, but it did distinguish between subtopics and main topics with some accuracy. Document-level burstiness was measured by a variant of the standard *TF*IDF*-weighting scheme.

In this experiment, the *TF*IDF*-weighting scheme could not distinguish between terms and non-terms as accurately as the weighting schemes that used the tag list information (*TW* and

[1]Except that the genre of Grolier may be considered as somewhat different from the genre of other corpora.

*STW*IDF*). This was the case with multi-word term candidates in particular. A larger corpus could possibly somewhat improve the performance of *TF*IDF* though. On the other hand, as mentioned above, tag weights alone did not distinguish accurately between important terms and less important terms. Thus, the results support the assumption that combining the linguistic and the frequency-based evidence would be a profitable approach to developing tools for information retrieval tasks. For instance, the index term `industrialism` occurred in the documents of the test corpus only once, and consequently, it was ranked low by the *TF*IDF*-weights. On the other hand, because of its tag list, it was ranked high by the *TW*-weights. Another index term, `biological`, is an adjective, and so it was ranked low by the *TW*-weights. However, because of its distribution, the word was ranked high by the *TF*IDF*-weights. In both cases, one weighting scheme overlooked an index term that was highly ranked by the other. If the weighting schemes are combined, the recall-precision curve can be improved, as the results of the previous part have indicated. The weighting scheme of the automatic indexer of this thesis, *STW*IDF*, does combine evidence from burstiness and linguistic analysis, and it outperformed both *TW* and *TF*IDF*.

As discussed earlier, in the context of TREC-style retrieval tasks, *TF*IDF* has established a certain status as a standard weighting scheme. Document collections of TREC are large, and the previously given queries determine the query terms, and thus the highly-weighted noise[2] does not necessarily disturb the retrieval process - particularly if the term lists are not viewed by the users. On average the query terms are weighted higher in relevant documents than in non-relevant documents. This does not mean, however, that the performance of *TF*IDF* should not and could not be improved - both with regard to precision and recall.

25.1 Precision

The results of the previous part indicated that weighting schemes based on only burstiness do not distinguish particularly well between terms and non-terms. If index terms are weighted in order to identify such document descriptors that are viewed by users, it is important to reduce the noise. This is the case, for example, in interactive interfaces where users view index terms in order to choose relevant documents, or in automatic hypertext generation where hot words (index terms) for hypertext links are identified automatically. In such tasks it is important that inappropriate term candidates are excluded as far as possible.

As discussed earlier, multi-word terms tend to contain more potential information than single-word terms, since phrases tend to be more specific than single word terms. Thus multi-word terms can be used to improve the precision of retrieval. For example, the word `subject` and the word `matter` may be too broad terms for indexing purposes, but the phrase `subject matter` could be an appropriate term for an automatic retrieval process. Furthermore, if a user views an

[2]For example, if a writer of a document should frequently use some less usual and informationally empty word, such as `thrice`, a “blind” burstiness-based weighting scheme probably weights this “term” high.

Precision	Recall	
Very high	Not so high	Viewed index terms
Not so high	Very high	Not viewed index terms

Figure 25.1: Trade-off between precision and recall.

index term list that includes the term `subject matter`, she can probably get a more precise impression about the content of the document than if the term list included two terms `subject` and `matter`.

The pattern matching method of *STW*IDF* produces single-word and multi-word terms that represent typical patterns of index terms. These terms are also weighted by *STW*IDF* according to their index-term-likeness (*TW*) and burstiness. In the ranked term lists of *STW*IDF* noise was weighted much lower than in the ranked term lists of *TF*IDF*. The weighted terms produced by *STW*IDF* should be quite appropriate document descriptors for tasks in which precision is demanded.

25.2 Recall

It is possible that a given index term is an important document descriptor even though it does not appear frequently in the document. This may be the case, for example, if the document is short or the index term is long, as seen above. The results of this thesis suggest that in this kind of situations the linguistic analysis may be able to increase the weight, so that the importance of the term is not ignored despite the lack of evidence based on burstiness.

If *TF*IDF* ignores or weights low a multi-word term which appears only once but which is still an important document descriptor, it is possible that a relevant document is missed by a user. *STW*IDF* uses evidence from tag lists, which may reveal the importance of the index term, and so the recall is improved.

Figure 25.1 summaries some points presented above. If index terms are viewed by users (e.g., hot words, interactive interfaces), it is important that the viewed index terms are informative and relevant document descriptors. It is useful to view multi-word terms that tend to contain more potential information than single-word terms, since phrases tend to be more specific than single word terms. It is also necessary to weight the index terms accurately, so that only the best document descriptors can be identified and viewed. Thus, in this task precision is more important criterion than recall. The methods introduced in this thesis may be highly useful in this kind of task.

On the other hand, index terms are not necessarily viewed by users, as in the case of traditional information retrieval systems. If the recall of retrieval is more important issue for users than the precision of retrieval, the index may include all words of a document without any weights or multi-word terms at all. However, if users want to avoid noise (i.e., non-relevant documents), it is necessary to weight terms accurately. Moreover, as discussed earlier, the use of multi-word terms is one way to improve precision as well. The methods introduced in this thesis may help to improve precision without reducing recall.

25.3 Weights without evidence based on burstiness

The *TW*-weighting scheme has one remarkable advantage over the burstiness-based weighting schemes. Once the probabilities of the tag combinations have been calculated, a sentence is a sufficient input for weighting the words, i.e., no document or document collection is needed. For instance, the words of the query `What role does Islam play in restricting women in Pakistan?` are weighted as follows:

```
Islam          0.985
Pakistan       0.766
woman          0.103
role           0.082
restrict       0.027
do             0.005
play           0.005
in             0.000
what           0.000
```

The *TW*-weighting scheme can be used for weighting query terms automatically in order to retrieve the most relevant documents by comparing query weights to document weights.

25.4 The use of pattern matching method to identify term candidates

In an experiment two methods of creating the set of two-word term candidates were compared:

- pattern matching method and
- the simple method of constructing phrases using a stoplist

The pattern matching method used two-word term patterns based on the training corpus, and the simple phrase constructing method considered all adjacent pairs of base forms of non-stopwords as two-word term candidates. *TF*IDF* values were then calculated for both sets, and the results were compared. In this experiment the pattern matching method outperformed the simple method when evaluated by the recall-precision curves.

The results of this experiment suggested that pattern matching method could be a useful approach to identification of multi-word term candidates - whatever is the applied weighting scheme.

The pattern matching method produced only half of the term candidates produced by the simple method. However, all the 1039 extra term candidates produced by the simple method were non-terms, that is, noise. The pattern matching method does not depend on the size of the document or the size of the document collection: term candidates can be extracted on the basis of their tag lists.

25.5 Textual variation

Textual variation appears in various ways, and it seems apparent that index-term-structure depends on textual variation as well. The index term corpus of this thesis includes, for example, a number of proper names as index terms, but an index of a software manual, on the other hand, would probably include very few if any proper names. So, the most easily recognizable index terms, proper names, are missing from that genre to a large extent. There would probably be other differences as well between the index-term-structure of a software manual and index-term-structures of the index term corpus. Accordingly, a robust indexing tool based on linguistic analysis must obviously take into account textual variation.

If the index-term-structures of different genres are different, it means that a robust indexing tool requires a large corpus of a wide range of genres as training material. Automatic indexing based solely on word stem frequencies is usually considered as a robust technique for all genres, but the automatic indexer of this thesis relies on linguistic features as well, which probably makes it more genre-specific. The differences between the index-term-structures of different genres might be an interesting subject of future research.

Chapter 26

Conclusion

Chapter 7 introduced a new concept: **index-term-structure**. Index-term-structure was identified with ‘weighted index terms in their context’, and it provides a content analysis framework for information retrieval. The analysis of the index-term-structure of a text can be seen as an analysis of a kind of meta-information structure of the text. The weights of index-term-structure can be calculated by any appropriate weighting scheme, but in this thesis a new weighting scheme, *STW*IDF*, was developed and applied. *STW*IDF* combines evidence from linguistic analysis and evidence from document-level burstiness.

Evidence from linguistic analysis was based on an index term corpus, which is a linguistically analysed text collection where the index terms were manually marked up. The typical linguistic features of index terms were explored using the index term corpus, and these findings provided the basis for a linguistic term-weighting scheme. The use of index term corpus of this kind as training material is a new approach to develop an automatic indexer.

The corpus was divided into two parts: a training corpus and a test corpus. The features of index terms were explored using the training corpus, which provided the basis for the automatic indexer. The test corpus was used to test whether the results could be generalized beyond the context of the training corpus. On the basis of the training corpus, the set of single-word and multi-word tag combination patterns of index terms was determined, and for each pattern an estimated index term probability was calculated by using the training corpus as training material. The automatic indexer used the index term probabilities of the tag combination patterns for weighting representations of the index term patterns, and these weights were referred to as **tag weights** (*TW*). The use of tag combination patterns of this kind is a new method of gathering evidence for an automatic indexer. The tags may provide information about the location of term candidates, or about the lexical, morphological, and syntactic properties of term candidates, and all this information can be expressed by a single tag weight. The dependency parser provides rich information on the linguistic features of index terms for the purpose of developing an automatic indexer, and it is possible to make this indexer more robust by constructing a larger training corpus of a wide range of genres.

*TF*IDF*-weighting scheme is a standard method of measuring document-level burstiness. *STW*IDF* combined evidence from tag weights with evidence from document-level burstiness by using summed tag weights (*STW*) instead of term frequencies (*TF*) in *TF*IDF*-formula. This is a new method to combine evidence from linguistic analysis and evidence from document-level burstiness. In the experiments of this thesis, *STW*IDF* was able to combine the strengths of tag weights and *TF*IDF*. Tag weights distinguished reasonably well between terms and non-terms, but poorly between important terms and less important terms. *TF*IDF*, on the other hand, distinguished poorly between terms and non-terms, but reasonably well between important terms and less important terms. The results suggested that *STW*IDF* distinguishes reasonably well both between terms and non-terms, and between important terms and less important terms. The combination of tag list information and analysis of burstiness seems to be then a feasible way to improve the performance of weighting schemes.

This thesis also introduced a new algorithm to measure within-document burstiness. The algorithm counts the distances of the occurrences of individual words using paragraphs as units for measuring the distance. A subject of future research will be to develop the algorithm and to lay a mathematical foundation for it. Another way to measure distances would be to use the word distances as units. In this implementation, paragraphs were used as units since paragraphs may be considered as topical units of discourse.

The within-document burstiness of different words was detected by determining the curves of the distribution functions of the words, and by computing areas above the curves of the words. This made it possible to compare the within-document burstiness of words by using single values computed to each word. Naturally, there would be other possible ways to compare the curves as well, and here is another subject of future research.

So, this algorithm distinguishes between bursty words and words used throughout the text. The results suggested that this distinction could be helpful to identification of index terms as well as to classification of index terms, if the method was combined with some other method that could recognize non-terms more accurately. In this thesis, however, no such combination was done, and thus more extensive experiments will be needed in order to fully evaluate the usefulness of this method to automatic indexing.

To sum up, this experiment in combining a linguistic weighting scheme with a variant of the standard *TF*IDF*-weighting scheme has offered promising results. Since the index terms were explicitly marked up in the corpus, it proved to be a relatively straightforward task to determine the basis for a simple linguistic weighting method, as well as to evaluate the performance of different weighting schemes. *STW*IDF* combined evidence from document-level burstiness and linguistic analysis by using summed tag weights (*STW*) instead of term frequencies (*TF*) in *TF*IDF*-formula. Naturally, there would be other ways to combine different kinds of evidence as well. Moreover, *STW*IDF* did not use evidence based on within-document burstiness. Subjects of future research will be, among the others, the evaluation of different techniques for combining the linguistic information and evidence based on the two kinds of burstiness, as well as the differences between the index-term-structures of different genres.

Bibliography

[1] Agosti, Maristella, Massimo Melucci, and Fabio Crestani. 1995. Automatic Authoring and Construction of Hypermedia for Information Retrieval. *Multimedia Systems*, vol.3(1), pp.15-24.

[2] Agosti, Maristella and Pier Giorgio Marchetti. 1992. User Navigation in the IRS Conceptual Structure through a Semantic Association Function. *The Computer Journal*, vol.35(3), pp. 194-199.

[3] American National Standards Institutes. 1968. *Basic Criteria for Indexes*. ANSI Z39.4, New York.

[4] Bartell, Brian T. 1994. *Optimizing Ranking Functions: A Connectionist Approach to Adaptive Information Retrieval*. PhD thesis, Department of Computer Science & Engineering, The University of California, San Diego.

[5] Bartell, Brian T., Garrison W. Cottrell, Richard K. Belew. 1994. Automatic Combination of Multiple Ranked Retrieval Systems. In Croft, W. Bruce and C.J. van Rijsbergen (editors). *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'94, 3-6 July 1994, Dublin, Ireland. Springer-Verlag, London, pp.173-181.

[6] Bear, John, David Israel, Jeff Petit, and David Martin. 1998. Using Information Extraction to Improve Document Retrieval. In Voorhees Ellen M. and Donna K. Harman (editors). *The Sixth Text REtrieval Conference (TREC-6)*. NIST Special Publication 500-240, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.367-377.

[7] Belkin, Nicholas J. 1978. Information concepts for information science. *Journal of Documentation*, vol.34(1), pp.55-85.

[8] Belkin, Nicholas J. 1993. Interaction with texts: Information retrieval as information-seeking behavior. In *Information retrieval'93. Von der Modellierung zur Anwendung*, Universitaetsverlag Konstanz, Konstanz, pp.55-66.

[9] Belkin, N. J., C. Cool, W.B. Croft, J.P. Callan. 1993. The Effect of Multiple Query Representations on Information Retrieval System Performance. In Korfhage, Robert, Edie Rasmussen and Peter Willett (editors). *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'93, June 27-July 1, 1993, Pittsburgh, PA USA, pp.339-346.

[10] Belkin, N. J., P.G. Marchetti, and C. Cool. 1993. BRAQUE: Design of an interface to support user interaction in information retrieval. *Information Processing & Management*, vol.29(3), pp.325-344.

[11] Belkin, Nicholas J., A. Desai Narasimhalu, and Peter Willett (editors). 1997. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'97, 27-31 July 1997, Philadelphia, Pennsylvania.

[12] Blair, David C. 1990. *Language and representation in information retrieval*. Elsevier Science Publishers B.V., Amsterdam, The Netherlands.

[13] Bookstein, Abraham, Yves Chiaramella, Gerard Salton, and Vijay V. Raghavan (editors). 1991. *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR-91, October 13-16, 1991, Chicago, Illinois USA.

[14] Borko, Harold and Charles L. Bernier. 1978. *Indexing concepts and methods*. Academic Press Inc., New York.

[15] Brier, S. 1996. Cybersemiotics: A new interdisciplinary development applied to the problems of knowledge organization and document retrieval in information science. *Journal of Documentation*, vol.52(3), pp.296-344.

[16] Brill, Eric. 1992. A Simple Rule-Based Part Of Speech Tagger. *Proceedings of the Third Conference on Applied NLP*. Trento, Italy, pp. 152-155.

[17] Brown, Gillian and George Yule. 1983. *Discourse analysis*. Cambridge University Press, Bath, Great Britain.

[18] Buckland, Michael. 1991. *Information and information systems*. Greenwood Press, Westport, United States of America.

[19] Buckley, Chris, Gerard Salton, James Allen, and Amit Singhal. 1995. Automatic query expansion using SMART: TREC-3. In Harman, Donna K. (editor). *Overview of the Third Text REtrieval Conference (TREC-3)*. NIST Special Publication 500-225, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.69-80.

[20] Buckley, Chris, Amit Singhal, Mandar Mitra, and (Gerard Salton). 1996. New Retrieval Approaches Using SMART: TREC 4. In Harman, Donna K. (editor). *The Fourth Text REtrieval*

Conference (TREC-4). NIST Special Publication 500-236, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.25-48.

[21] Buckley, Chris, Mandar Mitra, Janet Walz, and Claire Cardie. 1998. Using Clustering and SuperConcepts Within SMART: TREC 6. In Voorhees Ellen M. and Donna K. Harman (editors). *The Sixth Text REtrieval Conference (TREC-6)*. NIST Special Publication 500-240, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.107-124.

[22] Burger, John D., John S. Aberdeen, David D. Palmer. Information Retrieval and Trainable Natural Language Processing. 1997. In Voorhees Ellen M. and Donna K. Harman (editors). *The Fifth Text REtrieval Conference (TREC-5)*. NIST Special Publication 500-238, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.433-435.

[23] Burke, J. 1986. *The Day the Universe Changed*. Little, Brown & Co., Boston Massachusetts.

[24] Burnett Mark, Craig Fisher, and Richard Jones. 1996. InTEXT Precision Indexing in TREC4. In Harman, Donna K. (editor). *The Fourth Text REtrieval Conference (TREC-4)*, NIST Special Publication 500-236, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.287-294.

[25] Callan, James P. 1994. Passage-Level Evidence in Document Retrieval. In Croft, W. Bruce and C.J. van Rijsbergen (editors). *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'94, 3-6 July 1994, Dublin, Ireland. Springer-Verlag, London, pp.302-310.

[26] Church, K. and P. Hanks. 1990. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, vol.16(1), pp.22-29.

[27] Church, Kenneth W. and William A. Gale. 1995. Poisson mixtures. *Natural Language Engineering*, vol.1(2), pp.163-190.

[28] Clarke, Charles L.A. and Gordon V. Cormack. 1997. Interactive Substring Retrieval (MultiText Experiments for TREC-5) In Voorhees Ellen M. and Donna K. Harman (editors). *The Fifth Text REtrieval Conference (TREC-5)*. NIST Special Publication 500-238, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.267-278.

[29] Cleveland, Donald B. and Ana D. Cleveland. 1983. *Introduction to indexing and abstracting*. Libraries Unlimited, Inc., Littleton, Colorado.

[30] Cole, Peter and Jerry L. Morgan. 1975. *Syntax and Semantics. Volume 3. Speech Acts*. Academic Press, Inc., New York.

[31] *Collins COBUILD English Language Dictionary*. 1987 William Collins Sons & Co Ltd, London.

[32] *The concise Oxford dictionary*. 1976. 6th edition, Clarendon Press, Oxford.

[33] Coppen Peter-Arno, Hans van Halteren, and Lisanne Teunissen (editors). 1998. *Computational Linguistics in the Netherlands 1997. Selected Papers from the Eighth CLIN Meeting.* CLIN'97, 12 December 1997, Nijmegen, Netherlands.

[34] Croft, W. Bruce and C.J. van Rijsbergen (editors). 1994. *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'94, 3-6 July 1994, Dublin, Ireland. Springer-Verlag, London.

[35] Croft, Bruce, James Callan, and John Broglio. 1994. Routing and Ad-Hoc Retrieval Evaluation using the INQUERY System. In Harman, Donna K. (editor). *The Second Text REtrieval Conference (TREC-2).* NIST Special Publication 500-215, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.75-83.

[36] Crystal, David. 1997. *The Cambridge encyclopedia of language*. 2nd edition, Cambridge University Press.

[37] Daneš, František. 1974. Functional sentence perspective and the organization of the text. In Daneš, František (editor). *Papers on functional sentence perspective*, Academia, pp.106-128.

[38] Daneš, František (editor). 1974. *Papers on functional sentence perspective*, Academia.

[39] Daneš, František. 1995. The Paragraph - a Central Unit of the Thematic and Compositional Build-up of Texts. In Wårwik, Brita, Sanna-Kaisa Tanskanen, and Risto Hiltunen (editors). *Organization in Discourse. Proceedings from the Turku Conference*, Anglicana Turkuensia, No 14, Turku, Finland, pp.29-40.

[40] Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer, and R.A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, pp.391-407.

[41] Didion, J. 1979. The Getty. In *The white album*. Simon & Schuster, New York, pp.74-78.

[42] van Dijk, Teun A. 1977. *Text and Context. Explorations in the semantics and pragmatics of discourse*. Longman Group Ltd, London.

[43] van Dijk, Teun A. 1980. *Macrostructures: an interdisciplinary study of global structures in discourse, interaction, and cognition*. Lawrence Erlbaum Associates, Inc., New Jersey.

[44] van Dijk, Teun A. 1985. Semantic Discourse Analysis. In van Dijk, Teun A. (editor). *Handbook of discourse analysis. Volume 2. Dimensions of discourse*. Academic Press, Inc. Ltd, London, pp.103-136.

[45] van Dijk, Teun A. (editor). 1985. *Handbook of discourse analysis. Volume 2. Dimensions of discourse*. Academic Press, Inc. Ltd, London.

[46] Dobrov, Boris V., Natalia V. Loukachevitch, and Tatyana N. Yudina. 1998. Conceptual Indexing Using Thematic Representation of Texts In Voorhees Ellen M. and Donna K. Harman (editors). *The Sixth Text REtrieval Conference (TREC-6)*, NIST Special Publication 500-240, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp. 403-454.

[47] Dretske, Fred I. 1981. *Knowledge and the flow of information*. Basil Blackwell Publisher, Oxford.

[48] Dumais, S.T., G.W. Furnas, T.K. Landauer, and Deerwester, S. 1988. Using latent semantic analysis to improve information retrieval. In *CHI'88 Conference Proceedings: Human Factors in Computing Systems*, May 1988, ACM, New York, pp.281-285.

[49] Dumais, Susan T. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments, & Computers*, vol.23(2), pp.229-236.

[50] Dumais, Susan T. 1995. Latent Semantic Indexing (LSI): TREC-3 Report. In Harman, Donna K. (editor). *Overview of the Third Text REtrieval Conference (TREC-3)*. NIST Special Publication 500-225, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.219-230.

[51] Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, vol.19(1), pp.61-74.

[52] Evans, David A., Kimberly Ginther-Webster, Mary Hart, Robert G. Lefferts, and Ira A. Monarch. 1991. Automatic Indexing Using Selective NLP and First-Order Thesauri. In *Proceedings of RIAO '91*. April 2-5, 1991, Autonoma University of Barcelona, Spain, pp.624-644.

[53] Evans, David A. and Chengxiang Zhai. 1996. Noun-Phrase Analysis in Unrestricted Text for Information Retrieval. In *Proceedings of the 34th Annual Meeting of Association for Computational Linguistics*. June 24-28, 1996, Santa Cruz, California, pp.17-24.

[54] Foskett, A.C. 1996. *The Subject Approach to Information*. Fifth Edition, Library Association Publishing, London.

[55] Frei, Hans-Peter, Donna Harman, Peter Schäuble, and Ross Wilkinson (editors). 1996. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'96, August 18-22 1996, Zurich, Switzerland.

[56] Gallant, Stephen I., Robert Hecht-Nielsen, William R. Caid, Kent Pu Qing, Joel Carleton, David Sudbeck. 1993. HNC's MatchPlus System. In Harman, Donna K. (editor). *The First*

Text REtrieval Conference (TREC-1). NIST Special Publication 500-207, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.107-111.

[57] Gerdel, F. and M.C. Slocum. 1976. Paez Discourse, Paragraph, and Sentence Structure. In Longacre, R.E. (editor). *Discourse Grammar, Part I*. The Summer Institute of Linguistics, Dallas.

[58] Givón, Talmy (editor). 1979. *Syntax and Semantics. Volume 12. Discourse and Syntax*. Academic Press, Inc., New York.

[59] Grice, H. Paul. 1975. Logic and Conversation. In Cole, Peter and Jerry L. Morgan. *Syntax and Semantics. Volume 3. Speech Acts*. Academic Press, Inc., New York, pp.41-58.

[60] Griffiths, Morwenna and Margaret Whitford (editors). 1988. *Feminist perspectives in philosophy*. The Macmillan Press Ltd, London.

[61] Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, vol.12(3), pp.175-204.

[62] Guttman L. 1978. What is not what in statistic. *The Statistician*, 26, pp.81-107.

[63] Hahn, Udo. 1992. On text coherence parsing. In *Proceedings of the Fifteenth International Conference on Computational Linguistics*. COLING-92, Nantes, France, pp.25-31.

[64] Hajičová Eva, Hana Skoumalová, and Petr Sgall. 1995. An Automatic Procedure for Topic-Focus Identification. *Computational Linguistics*, vol.21(1), pp.81-94.

[65] Halliday, Michael A.K. 1967. Notes on transitivity and theme in English, Part 2. *Journal of Linguistics*, vol.3(2), pp.199-244.

[66] Halliday, Michael A.K. 1970a. Language structure and language function. In Lyons, John (editor). *New Horizons in Linguistics*. Penguin Books, Harmondsworth.

[67] Halliday, Michael A.K. and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman Group Ltd, London.

[68] Halliday, Michael A.K. 1989. *Spoken and written language*. Oxford University Press, Oxford.

[69] Hanson, Philip P. (editor). 1990. *Information, language, and cognition*. The University of British Columbia Press, Vancouver, Canada.

[70] Harbo, O. (editor). 1980. *Theory and Application of Information Research*. Mansell, London.

[71] Harman, D. and Candela, G. 1989. Retrieving records from a gigabyte of text on a minicomputer using statistical ranking. *Journal of the American Society for Information Science*, vol.41(8), pp.581-589.

[72] Harman, Donna K. (editor). 1993. *The First Text REtrieval Conference (TREC-1)*. NIST Special Publication 500-207, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html).

[73] Harman, Donna K. (editor). 1994. *The Second Text REtrieval Conference (TREC-2)*. NIST Special Publication 500-215, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html).

[74] Harman, Donna K. (editor). 1995. *Overview of the Third Text REtrieval Conference (TREC-3)*. NIST Special Publication 500-225, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html).

[75] Harman, Donna K. (editor). 1996. *The Fourth Text REtrieval Conference (TREC-4)*. NIST Special Publication 500-236, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html).

[76] *Harnessing the power of language*. A brochure published by the *LINGLINK* team at Anite Systems, on behalf of the participants in the Language Engineering Sector of the Telematics Applications Programme.

[77] Harris, Zellig. 1991. *A theory of language and information: a mathematical approach*. Oxford University Press, Oxford.

[78] Harter, Stephen P. 1986. *Online information retrieval*. Academic Press, Inc., Orlando, Florida.

[79] Harvey, Lee. 1990. *Critical social research*. Unwin Hyman Ltd, London.

[80] Hearst, Marti A. and Christian Plaunt. 1993. Subtopic Structuring for Full-Length Document Access. In Korfhage, Robert, Edie Rasmussen and Peter Willett (editors). *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'93, June 27-July 1, 1993, Pittsburgh, PA USA, pp. 59-68.

[81] Hearst, Marti, Jan Pedersen, Peter Pirolli, Hinrich Schütze, Gregory Grefenstette, and David Hull. 1996. Xerox Site Report: Four TREC-4 Tracks. In Harman, Donna K. (editor). *The Fourth Text REtrieval Conference (TREC-4)*. NIST Special Publication 500-236, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.97-119.

[82] Hearst, Marti A. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, vol.23(1), pp.33-64.

[83] Hearst, Marti A. and Chandu Karadi. 1997. Cat-a-Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results using a Large Category Hierarchy. In Belkin, Nicholas J., A. Desai Narasimhalu, and Peter Willett (editors). *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'97, 27-31 July 1997, Philadelphia, Pennsylvania, pp.246-255.

[84] Heikkilä, Juha. 1995. ENGTWOL English lexicon: solutions and problems. In Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, and Arto Anttila (editors). *Constraint Grammar: a language-independent system for parsing unrestricted text, volume 4 of Natural Language Processing*. Mouton de Gruyter, Berlin and New York.

[85] Hemmje, Matthias, Clemens Kunkel, and Alexander Willet. 1994. LyberWorld - A Visualization User Interface Supporting Fulltext Retrieval. In Croft, W. Bruce and C.J. van Rijsbergen (editors). *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'94, 3-6 July 1994, Dublin, Ireland. Springer-Verlag, London, pp.249-259.

[86] Hinds, John. 1979. Organizational patterns in discourse. In Givón, Talmy (editor). *Syntax and Semantics. Volume 12. Discourse and Syntax*, Academic Press, Inc., New York, pp.135-157.

[87] Hockett, C.F. 1958. *A Course in Modern Linguistics*. Macmillan, New York.

[88] Hull, David A. 1994. Improving Text Retrieval for the Routing Problem using Latent Semantic Indexing. In Croft, W. Bruce and C.J. van Rijsbergen (editors). *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'94, 3-6 July 1994, Dublin, Ireland. Springer-Verlag, London, pp.282-291.

[89] Hull, David A. . 1996. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, vol.47(1), pp.70-84.

[90] Hull, David A., Gregory Grefenstette, B. Maximilian Schulze, Eric Gaussier, Hinrich Schütze, and Jan O. Pedersen. 1997. Xerox TREC-5 Site Report: Routing, Filtering, NLP, and Spanish Tracks. In Voorhees Ellen M. and Donna K. Harman (editors). *The Fifth Text REtrieval Conference (TREC-5)*. NIST Special Publication 500-238, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp. 167-180.

[91] Ingwersen, Peter. 1992. *Information retrieval interaction*. Taylor Graham Publishing, London.

[92] Ingwersen, P. and N.O. Pors (editors). 1996. *Proceedings CoLIS, 2nd International Conference on Conceptions of Library and Information Science: Integration in Perspective*. Copenhagen, Oct. 13-16, 1996. The Royal School of Librarianship, Copenhagen.

[93] Jacobs, Paul S. and Lisa F. Rau. 1990. SCISOR: Extracting Information from On-line News. *Communications of the ACM*, vol.33(11), pp.88-97.

[94] Justeson, John S. and Slava M. Katz. 1995. Technical Terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, vol.1(1), pp.9-27.

[95] Järvinen, Timo. 1994. Annotating 200 Million Words: The Bank of English Project. In *COLING-94 . The 15th International Conference on Computational Linguistics Proceedings*, volume 1, Kyoto, Japan, pp. 565-568.

[96] Karetnyk, David, Fred Karlsson, and Godfrey Smart. 1991. Knowledge-based Indexing of Morpho-syntactically Analysed Language. *Expert Systems for Information Management*, vol.4(1), pp.1-29.

[97] Karlgren, Jussi and Douglas Cutting. 1994. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In *Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING'94)*. COLING-94, August 1994, Kyoto, Japan, (http://xxx.lanl.gov/abs/cmp-lg/9410008).

[98] Karlgren, Jussi. 1996. Stylistic Variation in an Information Retrieval Experiment. In *Proceedings of the NeMLaP-2 Conference*. September 1996, Bilkent University, Ankara, Turkey.

[99] Karlgren, Jussi. 2000. *Stylistic Experiments for Information Retrieval*. PhD Dissertation at the department of linguistics, Stockholm University. SICS Dissertation series 26, Edsbruk, Sweden.

[100] Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, and Arto Anttila (editors). 1995. *Constraint Grammar: a language-independent system for parsing unrestricted text, volume 4 of Natural Language Processing*. Mouton de Gruyter, Berlin and New York.

[101] Karttunen, Lauri, Kimmo Koskenniemi, and Ronald M. Kaplan. 1987. A compiler for two-level phonological rules. In Dalrymple, Mary (editor). *Tools for Morphological Analysis*. Center for the Study of Language and Information, Stanford, CA.

[102] Katz, Slava M. 1996. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, vol.2(1), pp.15-59.

[103] Keenan, E.L. (editor). 1975. *Formal Semantics of Natural Language*. Cambridge University Press.

[104] Klavans, Judith L., Evelyne Tzoukermann, and Christian Jacquemin. 1997. Effective Use of Natural Language Processing Techniques for Automatic Conflation of Multi-Word Terms: The Role of Derivational Morphology, Part of Speech Tagging, and Shallow Parsing. In Belkin, Nicholas J., A. Desai Narasimhalu, and Peter Willett (editors). *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'97, 27-31 July 1997, Philadelphia, Pennsylvania, pp.148-155.

[105] Kochen, Manfred. 1983. Library science and information science. Broad or Narrow? In Machlup, Fritz and Una Mansfield (editors). *The Study of Information*. John Wiley & Sons, Inc., New York, pp.371-377.

[106] Korfhage, Robert, Edie Rasmussen and Peter Willett (editors). 1993. *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'93, June 27-July 1, 1993, Pittsburgh, PA USA.

[107] Koskenniemi, Kimmo. 1983. Two-level morphology: a general computational model for word-form recognition and production. Publications No. 11. Department of General Linguistics, University of Helsinki.

[108] Koskenniemi, Kimmo. 1996. Finite-state morphology and information retrieval. In *Proceedings of ECAI-96 Workshop on Extended Finite State Models of Language.* Budapest, Hungary, pp.42-45.

[109] Lahtinen, Timo. 1998. The Use of an Index Term Corpus for the Development of an Automatic Indexer. In Coppen Peter-Arno, Hans van Halteren, and Lisanne Teunissen (editors). *Computational Linguistics in the Netherlands 1997. Selected Papers from the Eighth CLIN Meeting.* CLIN'97, 12 December 1997, Nijmegen, Netherlands, pp.59-75.

[110] Lancaster, Frederick Wilfrid. 1991. *Indexing and abstracting in theory and in practice*. Library Association Publishing Ltd., London.

[111] *Language engineering. Progress & prospects*. 1997. A brochure published by the *LINGLINK* team at Anite Systems, on behalf of the participants in the Language Engineering Sector of the Telematics Applications Programme.

[112] Lee, J. H. 1997. Analyses of Multiple Evidence Combination. In Belkin, Nicholas J., A. Desai Narasimhalu, and Peter Willett (editors). *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'97, 27-31 July 1997, Philadelphia, Pennsylvania, pp.267-275.

[113] Liddy, E. DuRoss. 1990. Anaphora in natural language processing and information retrieval. *Information Processing & Management*, vol.26(1), pp.39-52.

[114] Liddy, Elizabeth D. and Sung H. Myaeng. 1993. DR-LINK's Linguistic-Conceptual Approach to Document Detection. In Harman, Donna K. (editor). *The First Text REtrieval Conference (TREC-1)*. NIST Special Publication 500-207, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.113-129.

[115] Liddy, Elizabeth D. and Sung H. Myaeng. 1994. DR-LINK: A System Update for TREC-2. In Harman, Donna K. (editor). *The Second Text REtrieval Conference (TREC-2)*. NIST Special Publication 500-215, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.85-99.

[116] Longacre, R.E. (editor). 1976. *Discourse Grammar, Part I*. The Summer Institute of Linguistics, Dallas.

[117] Longacre, R. E. 1979. The paragraph as a grammatical unit. In Givón, Talmy (editor). *Syntax and Semantics. Volume 12. Discourse and Syntax*, Academic Press, Inc., New York, pp.115-134.

[118] Lovins, J. , 1968. Development of a Stemming Algorithm. *Mechanical Translations and Computational Linguistics*, 11, pp.11-32.

[119] Luhn, H.P. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2, pp.159-165.

[120] Lyons, John (editor). 1970. *New Horizons in Linguistics*. Penguin Books, Harmondsworth.

[121] Lyons, John. 1977. *Semantics 1*. Cambridge University Press, Cambridge, Great Britain.

[122] Machlup, Fritz and Una Mansfield (editors). 1983. *The Study of Information*. John Wiley & Sons, Inc., New York.

[123] *The Macquarie dictionary*. 1981. Macquarie Library Pty Ltd., St Leonards, NSW.

[124] Mann, William C. and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Towards a functional theory of text organization. *Text*, vol.8(3), pp.243-281.

[125] Mauldin, Michael L. 1991. Retrieval Performance in FERRET. A Conceptual Information Retrieval System. In Bookstein, Abraham, Yves Chiaramella, Gerard Salton, and Vijay V. Raghavan (editors). *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR-91, October 13-16, 1991, Chicago, Illinois USA, pp.347-355.

[126] Meadow, Charles T. 1992. *Text information retrieval systems*. Academic Press, Inc., San Diego, California.

[127] Metzler, D.P. and Haas S.W. 1989. The Constituent Object Parser: Syntactic Structure Matching for Information Retrieval. *ACM Transactions on Information Systems*, vol.7(3), pp.292-316.

[128] De Mey, M. 1977. The cognitive viewpoint: its development and its scope. In *CC-77: International Workshop on the Cognitive Viewpoint*, Ghent University, Ghent, pp.xvi-xxxii.

[129] De Mey, M. 1980. The relevance of the cognitive paradigm information science. In Harbo, O. (editor). *Theory and Application of Information Research*. Mansell, London, pp.48-61.

[130] Morris, Jane and Graeme Hirst. 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, vol.17(1), pp.21-48.

[131] Mosteller F. and Wallace D. 1964. *Inference and disputed authorship: The Federalist*. Addison-Wesley.

[132] Nowell, Lucy Terry, Robert K. France, Deborah Hix, Lenwood S. Heath, and Edward A. Fox. 1996. Visualizing Search Results: Some Alternatives To Query-Document Similarity. In Frei, Hans-Peter, Donna Harman, Peter Schäuble, and Ross Wilkinson (editors). *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'96, August 18-22 1996, Zurich, Switzerland, pp.67-84.

[133] Orwell, George. 1945. Some thoughts on the common toad. In Orwell, S. and I. Angus (editors). *The collected essays, journalism and letters of George Orwell: Vol.4. In front of your nose*. 1945-1950. Harcourt Brace & World, New York, pp.141-145.

[134] Orwell, S. and I. Angus (editors). 1945-1950. *The collected essays, journalism and letters of George Orwell: Vol.4. In front of your nose*. Harcourt Brace & World, New York.

[135] Pirkola, Ari and Kalervo Järvelin. 1996. Recall and precision effects of anaphor and ellipsis resolution in proximity searching in a text database. In Ingwersen, P. and N.O. Pors (editors). *Proceedings CoLIS, 2nd International Conference on Conceptions of Library and Information Science: Integration in Perspective*. Copenhagen, Oct. 13-16, 1996. The Royal School of Librarianship, Copenhagen, pp. 459 - 475.

[136] Popper, Karl R. 1972. *Objective Knowledge*. Oxford University Press, Oxford.

[137] Porter, M. F. 1980. An algorithm for suffix stripping. *Program*, 14, pp.130-137.

[138] van Rijsbergen, C.J. 1979. *Information Retrieval*. Second edition. Butterworth & Co Ltd, London.

[139] Robertson, S.E., S. Walker, S. Jones, M.M. Hancock-Beaulieu, M. Gatford. 1995. Okapi at TREC-3. In Harman, Donna K. (editor). *Overview of the Third Text REtrieval Conference*

(TREC-3). NIST Special Publication 500-225, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.109-126.

[140] Robertson, S.E. and Karen Sparck Jones. 1997. Simple, proven approaches to text-retrieval. Technical report 356, Computer Laboratory, University of Cambridge.

[141] Roeh, I. 1982. *The rhetoric of news*. Studienverlag, Bochum.

[142] Russel, B. 1935. On comets. In *In praise of idleness*. Allen & Unwin, London, pp.223-225.

[143] Sager, N. 1981. *Natural language information processing*. Addison-Wesley, Reading, MA.

[144] Salton, Gerard and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., Singapore.

[145] Salton, Gerard and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, vol.24(5), pp.513-523.

[146] Salton, Gerard. 1989. *Automatic Text Processing. The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Publishing Company, Inc., Reading, MA.

[147] Salton, Gerard and James Allan. 1993. Selective text utilization and text traversal. In *Proceedings of ACM Hypertext'93*.

[148] Schwarz, C. 1990. Content Based Text Handling. *Information Processing & Management*, vol.26(2), pp.219-226.

[149] Shannon, Claude E. 1949. The Mathematical Theory of Communication. In Shannon, Claude E. and Warren Weaver. *The Mathematical Theory of Communication*. The University of Illinois Press, Urbana, pp.3-91.

[150] Shannon, Claude E. and Warren Weaver. 1949. *The Mathematical Theory of Communication*. The University of Illinois Press, Urbana.

[151] Smeaton, Alan F. 1992. Progress in the Application of Natural Language Processing to Information Retrieval Tasks. *The Computer Journal*, vol.35(3), pp.268-278.

[152] Smeaton, Alan F., Ruairi O'Donnell, and Fergus Kelledy. 1995. Indexing Structures Derived from Syntax in TREC-3: System Description. In Harman, Donna K. (editor). *Overview of the Third Text REtrieval Conference (TREC-3)*. NIST Special Publication 500-225, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.55-67.

[153] Smeaton, Alan F., Fergus Kelledy, and Ruairi O'Donnell. 1996. TREC-4 Experiments at Dublin City University: Thresholding Posting Lists, Query Expansion with WordNet and POS

Tagging of Spanish. In Harman, Donna K. (editor). *The Fourth Text REtrieval Conference (TREC-4)*. NIST Special Publication 500-236, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.373-389.

[154] Smeaton, Alan F. 1998. Independence of Contributing Retrieval Strategies in Data Fusion for Effective Information Retrieval. To appear in *Proceedings of the 20th BCS-IRSG Colloquium*. Springer-Verlag Workshops in Computing, April 1998, Grenoble, France, (http://www.compapp.dcu.ie/ asmeaton/pubs-list.html).

[155] Smeaton, Alan F., Fergus Kelledy, and Gerard Quinn. 1998. Ad hoc Retrieval Using Thresholds, WSTs for French Mono-lingual Retrieval, Document-at-a-Glance for High Precision and Triphone Windows for Spoken Documents. In Voorhees Ellen M. and Donna K. Harman. *The Sixth Text REtrieval Conference (TREC-6)*. NIST Special Publication 500-240, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.461-476.

[156] Sparck Jones, Karen. 1972. A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, vol.28(1), pp.11-20.

[157] Sparck Jones, Karen. 1995. Reflections on TREC. *Information Processing & Management*, vol.31(3), pp.291-314.

[158] Sparck Jones, Karen. 1998. Summary Performance Comparisons TREC-2, TREC-3, TREC-4, TREC-5, TREC-6 (Appendix B). In Voorhees Ellen M. and Donna K. Harman (editors). *The Sixth Text REtrieval Conference (TREC-6)*. NIST Special Publication 500-240, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), Appendix B, pp. B1-B8.

[159] Stark, Heather A. 1988. What Do Paragraph Markings Do? *Discourse Processes*, vol.11(3), pp.275-303.

[160] Strzalkowski, Tomek, Jose Perez Carballo, and Mihnea Marinescu. 1995. Natural Language Information Retrieval: TREC-3 Report. In Harman, Donna K. (editor). *Overview of the Third Text REtrieval Conference (TREC-3)*. NIST Special Publication 500-225, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.39-53.

[161] Strzalkowski, Tomek and Jose Perez Carballo. 1996. Natural Language Information Retrieval: TREC-4 Report. In Harman, Donna K. (editor). *The Fourth Text REtrieval Conference (TREC-4)*. NIST Special Publication 500-236, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.245-258.

[162] Strzalkowski, Tomek and Karen Sparck Jones. 1997. NLP Track at TREC-5. In Voorhees Ellen M. and Donna K. Harman (editors). *The Fifth Text REtrieval Conference (TREC-5)*. NIST

Special Publication 500-238, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.97-100.

[163] Strzalkowski, Tomek, Louise Guthrie, Jussi Karlgren, Jim Leistensnider, Fang Lin, Jose Perez-Carballo, Troy Straszheim, Jin Wang, and Jon Wilding. 1997. Natural language information retrieval: TREC-5 report. In Voorhees Ellen M. and Donna K. Harman (editors). *The Fifth Text REtrieval Conference (TREC-5)*. NIST Special Publication 500-238, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.291-314.

[164] Strzalkowski, Tomek, Fang Lin, and Jose Perez-Carballo. 1998. Natural Language Information Retrieval TREC-6 Report. In Voorhees Ellen M. and Donna K. Harman (editors). *The Sixth Text REtrieval Conference (TREC-6)*. NIST Special Publication 500-240, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.347-366.

[165] Suri, Linda Z. and Kathleen F. McCoy. 1994. RAFT/RAPR and Centering: A Comparison and Discussion of Problems Related to Processing Complex Sentences. *Computational Linguistics*, vol.20(2), pp.301-317.

[166] Swanson, Don R. 1985. Historical Note: Information Retrieval and the Future of an Illusion. *Journal of the American Society for Information Science*, vol.36(3), pp.92-98.

[167] Tapanainen, Pasi and Timo Järvinen. 1997. A non-projective dependency parser. In *The Proceedings of the 5th Conference on Applied Natural Language Processing*. ACL, April 1997, Washington, D.C.

[168] Thagard, Paul. 1990. Comment. Information and Concepts. In Hanson, Philip P. (editor). *Information, language, and cognition*. The University of British Columbia Press, Vancouver, Canada, pp.168-174.

[169] Veerasamy, Aravindan and Russel Heikes. 1997. Effectiveness of a graphical display of retrieval results. In Belkin, Nicholas J., A. Desai Narasimhalu, and Peter Willett (editors). *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'97, 27-31 July 1997, Philadelphia, Pennsylvania, pp.236-245.

[170] Venneman, T. 1975. Topic, sentence accent, and ellipsis: a proposal for their formal treatment. In Keenan, E.L. (editor). *Formal Semantics of Natural Language*, Cambridge University Press.

[171] Vickery, Brian C. and Alina Vickery. 1992. *Information science in theory and practice*. Butterwoth & Co Ltd, Bowker-Saur, West Sussex, Great Britain.

[172] Vogt, Christopher C. 1997. When Does it Make Sense to Linearly Combine Relevance Scores? In Belkin, Nicholas J., A. Desai Narasimhalu, and Peter Willett (editors). *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Poster presented at SIGIR'97, 27-31 July 1997, Philadelphia, Pennsylvania.

[173] Vogt, Christopher C., Garrison W. Cottrell, Richard K. Belew, Brian T. Bartell. 1997. Using Relevance to Train a Linear Mixture of Experts. In Voorhees Ellen M. and Donna K. Harman (editors). *The Fifth Text REtrieval Conference (TREC-5)*. NIST Special Publication 500-238, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.503-516.

[174] Voorhees, Ellen M. 1994. Query Expansion using Lexical-Semantic Relations. In Croft, W. Bruce and C.J. van Rijsbergen (editors). *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'94, 3-6 July 1994, Dublin, Ireland. Springer-Verlag, London, pp. 61-69.

[175] Voorhees Ellen M. and Donna K. Harman (editors). 1997. *The Fifth Text REtrieval Conference (TREC-5)*. NIST Special Publication 500-238, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html).

[176] Voorhees Ellen M. and Donna K. Harman (editors). 1998a. *The Sixth Text REtrieval Conference (TREC-6)*. NIST Special Publication 500-240, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html).

[177] Voorhees Ellen M. and Donna K. Harman. 1998b. Overview of the Sixth Text REtrieval Conference (TREC-6). In Voorhees Ellen M. and Donna K. Harman (editors). *The Sixth Text REtrieval Conference (TREC-6)*. NIST Special Publication 500-240, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.1-24.

[178] Voutilainen, Atro, Juha Heikkilä and Arto Anttila. 1992. Constraint Grammar of English. A Performance-Oriented Introduction. Publications No. 21, Department of General Linguistics, University of Helsinki.

[179] Voutilainen, Atro. 1994. Designing a Parsing Grammar. Publications No. 22, Department of General Linguistics, University of Helsinki.

[180] Voutilainen, Atro. 1995. Morphological disambiguation. In Karlsson, Fred, Atro Voutilainen, Juha Heikkilä, and Arto Anttila (editors). *Constraint Grammar: a language-independent system for parsing unrestricted text, volume 4 of Natural Language Processing*. Mouton de Gruyter, Berlin and New York, pp.165-284.

[181] Waterworth, John A. and Mark H. Chignell. 1991. A model for information exploration. *Hypermedia*, vol.3(1), pp.35-58.

[182] Weaver, Warren. 1949. Recent Contributions to the Mathematical Theory of Communication. In Shannon, Claude E. and Warren Weaver. *The Mathematical Theory of Communication*. The University of Illinois Press, Urbana, pp.94-117.

[183] Wilkinson, Ross. 1994. Effective Retrieval of Structured Documents. In Croft, W. Bruce and C.J. van Rijsbergen (editors). *Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'94, 3-6 July 1994, Dublin, Ireland. Springer-Verlag, London, pp.311-317.

[184] Wilkinson, Ross and Justin Zobel. 1995. Comparison of Fragmentation Schemes for Document Retrieval. In Harman, Donna K. (editor). *Overview of the Third Text REtrieval Conference (TREC-3)*. NIST Special Publication 500-225, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp. 81-84.

[185] Wårwik, Brita, Sanna-Kaisa Tanskanen, and Risto Hiltunen (editors). 1995. *Organization in Discourse. Proceedings from the Turku Conference*. Anglicana Turkuensia, No 14, Turku, Finland.

[186] Xu Jinxi and W. Bruce Croft. 1996. Query Expansion Using Local and Global Document Analysis. In Frei, Hans-Peter, Donna Harman, Peter Schäuble, and Ross Wilkinson (editors). *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR'96, August 18-22 1996, Zurich, Switzerland, pp. 4-11.

[187] Youmans, Gilbert. 1991. A new tool for discourse analysis: The vocabulary-management profile. *Language*, vol.67(4), pp.763-789.

[188] Zadrozny, Wlodek and Karen Jensen. 1991. Semantics of Paragraphs. *Computational Linguistics*, vol.17(2), pp.171-209.

[189] Zhai, Chengxiang, Xiang Tong, Nataša Milić-Frayling, and David A. Evans. 1997. Evaluation of Syntactic Phrase Indexing - CLARIT NLP Track Report. In Voorhees Ellen M. and Donna K. Harman (editors). *The Fifth Text REtrieval Conference (TREC-5)*. NIST Special Publication 500-238, National Institute of Standards and Technology, Gaithersburg, MD, (http://trec.nist.gov/pubs.html), pp.347-358.

[190] Zipf, H.P. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, Massachusetts.

Appendix 1. An excerpt from the beginning of the training corpus

```
"<$<p>>"
"<$<s>>"
"<Introduction>"        "introduction" <*> N NOM SG @NH main:>0 <END:ion> <HEADLINE>
"<$<s>>"
"<$<p>>"
"<$<s>>"
"<Morwenna>"    "morwenna" <*> <?> N NOM SG @A> attr:>2 <+INDEX-TERM> <HEADLINE>
"<Griffiths>"   "Griffiths" <*> <Proper> N NOM SG @NH #2 main:>0 </+INDEX-TERM> </+INDEX-TERM-1> <
"<and>" "and" CC @CC cc:>2 <HEADLINE>
"<Margaret>"    "Margaret" <*> <Proper> N NOM SG @A> attr:>5 <+INDEX-TERM> <HEADLINE>
"<Whitford>"    "Whitford" <*> <Proper> N NOM SG @NH #5 cc:>2 </+INDEX-TERM> </+INDEX-TERM-1> <HEA
"<$<s>>"
"<$<p>>"
"<$<s>>"
"<Philosophy>"  "philosophy" <*> N NOM SG @SUBJ subj:>2 <PAR1> <SEN1>
"<is>"  "be" <Irreg> <SVC/A> <SVC/N> V PRES SG3 VFIN @+FMAINV #2 main:>0 <PAR1> <SEN1>
"<in>"  "in" PREP @ADVL #3 loc:>2 <PAR1> <SEN1>
"<urgent>"      "urgent" A ABS @A> attr:>5 <PAR1> <SEN1>
"<need>"        "need" <Count> N NOM SG @<P #5 pcomp:>3 <PAR1> <SEN1>
"<of>"  "of" PREP @<NOM-OF #6 mod:>5 <PAR1> <SEN1>
"<a>"   "a" <Indef> DET CENTRAL ART SG @DN> det:>9 <PAR1> <SEN1>
"<feminist>"    "feminist" <Count> N NOM SG @A> attr:>9 <+INDEX-TERM> <PAR1> <SEN1>
"<perspective>" "perspective" N NOM SG @<P #9 pcomp:>6 </+INDEX-TERM> <PAR1> <SEN1>
"<$.>"
"<$<s>>"
"<For>" "for" <*> PREP @ADVL #1 tmp:>8 <PAR1>
"<centuries>"   "century" <Count> <ADV-N> N NOM PL @<P pcomp:>1 <PAR1>
"<the>" "the" <Def> DET CENTRAL ART SG/PL @DN> det:>4 <PAR1>
"<practice>"    "practice" N NOM SG @SUBJ #4 subj:>7 <PAR1>
"<of>"  "of" PREP @<NOM-OF #5 mod:>4 <PAR1>
"<philosophy>"  "philosophy" N NOM SG @<P pcomp:>5 <PAR1>
"<has>" "have" <Irreg> <SVO> <SVOC/A> V PRES SG3 VFIN @+FAUXV #7 v-ch:>8 <PAR1>
"<been>"        "be" <Irreg> <SVC/A> <SVC/N> EN @-FMAINV #8 main:>0 <PAR1>
"<overwhelmingly>"      "overwhelming" <DER:ly> ADV @ADVL man:>8 <PAR1>
"<the>" "the" <Def> DET CENTRAL ART SG/PL @DN> det:>11 <PAR1>
"<prerogative>" "prerogative" <Count> N NOM SG @PCOMPL-S #11 comp:>8 <PAR1>
"<of>"  "of" PREP @<NOM-OF #12 mod:>11 <PAR1>
"<men>" "man" N NOM PL @<P pcomp:>12 <PAR1>
"<but>" "but" CC @CC cc:>8 <PAR1>
"<it>"  "it" <NonMod> PRON NOM SG3 SUBJ @SUBJ subj:>16 <PAR1>
"<is>"  "be" <Irreg> <SVC/A> <SVC/N> V PRES SG3 VFIN @+FMAINV #16 cc:>8 <PAR1>
"<only>"        "only" ADV @AD-A> ad:>18 <PAR1>
"<recently>"    "recent" <DER:ly> ADV @ADVL #18 man:>16 <PAR1>
"<that>"        "that" <**CLB> CS @CS pm:>23 <PAR1>
"<feminist>"    "feminist" <Count> N NOM SG @A> attr:>21 <PAR1>
"<analysis>"    "analysis" N NOM SG @SUBJ #21 subj:>22 <PAR1>
```

```
"<has>" "have" <Irreg> <SVO> <SVOC/A> V PRES SG3 VFIN @+FAUXV #22 v-ch:>23 <PAR1>
"<made>"        "make" <SVO> <SVOO> <SVOC/A> <SVOC/N> <into/SVOC/A> <P/for> <P/of> <InfComp> EN @-
"<it>"  "it" <NonMod> PRON ACC SG3 @OBJ obj:>23 <PAR1>
"<possible>"    "possible" <DER:ble> A ABS @PCOMPL-O #25 oc:>23 <PAR1>
"<to>"  "to" INFMARK> @INFMARK> pm:>27 <PAR1>
"<see>" "see" <Vcog> <SVO> <as/SVOC/A> <InfComp> V INF @-FMAINV #27 mod:>25 <PAR1>
"<the>" "the" <Def> DET CENTRAL ART SG/PL @DN> det:>30 <PAR1>
"<distorting>"  "distort" <ING> <Nominal> A ABS @A> attr:>30 <PAR1>
"<effect>"      "effect" N NOM SG @OBJ #30 obj:>27 <PAR1>
"<of>"  "of" PREP @<NOM-OF #31 mod:>30 <PAR1>
"<this>"        "this" DET CENTRAL DEM SG @DN> det:>34 <PAR1>
"<historical>"  "historical" <DER:ic> <DER:al> A ABS @A> attr:>34 <+INDEX-TERM> <PAR1>
"<fact>"        "fact" N NOM SG @<P #34 pcomp:>31 </+INDEX-TERM> <PAR1>
"<$.>"
"<$<s>>"
"<The>" "the" <*> <Def> DET CENTRAL ART SG/PL @DN> det:>2 <PAR1>
"<articles>"    "article" N NOM PL @SUBJ #2 subj:>6 <PAR1>
"<in>"  "in" PREP @<NOM #3 mod:>2 <PAR1>
"<this>"        "this" DET CENTRAL DEM SG @DN> det:>5 <PAR1>
"<book>"        "book" <Count> N NOM SG @<P #5 pcomp:>3 <PAR1>
"<demonstrate>" "demonstrate" <Vcog> <SVO> V PRES -SG3 VFIN @+FMAINV #6 <PAR1>
"<in>"  "in" PREP @ADVL #7 loc:>6 <PAR1>
"<a>"   "a" <Indef> DET CENTRAL ART SG @DN> det:>9 <PAR1>
"<variety>"     "variety" N NOM SG @<P #9 pcomp:>7 <PAR1>
"<of>"  "of" PREP @<NOM-OF #10 mod:>9 <PAR1>
"<ways>"        "way" <Count> <ADV-N> N NOM PL @<P #11 <PAR1>
"<where>"       "where" ADV WH @<P pcomp:>10 <PAR1>
"<the>" "the" <Def> DET CENTRAL ART SG/PL @DN> det:>14 <PAR1>
"<bias>"        "bias" N NOM SG @SUBJ #14 subj:>15 </+INDEX-TERM> <PAR1>
"<occurs>"      "occur" V PRES SG3 VFIN @+FMAINV #15 mod:>11 <PAR1>
"<and>" "and" CC @CC cc:>15 <PAR1>
"<how>" "how" ADV WH @ADVL man:>21 <PAR1>
"<it>"  "it" <NonMod> PRON NOM SG3 SUBJ @SUBJ subj:>19 <PAR1>
"<might>"       "might" V AUXMOD VFIN @+FAUXV #19 v-ch:>20 <PAR1>
"<be>"  "be" <Irreg> <SVC/A> <SVC/N> V INF @-FAUXV #20 v-ch:>21 <PAR1>
"<redressed>"   "redress" <SVO> EN @-FMAINV #21 cc:>15 <PAR1>
"<$.>"
"<$<s>>"
"<They>"        "they" <*> <NonMod> PRON PERS NOM PL3 SUBJ @SUBJ subj:>3 <PAR1> <SENn>
"<also>"        "also" ADV ADVL @ADVL meta:>3 <PAR1> <SENn>
"<show>"        "show" <Vcog> <SVO> <SVOO> V PRES -SG3 VFIN @+FMAINV #3 <PAR1> <SENn>
"<that>"        "that" <**CLB> CS @CS <PAR1> <SENn>
"<redressing>"  "redress" <SVO> ING @-FMAINV <PAR1> <SENn>
"<it>"  "it" <NonMod> PRON NOM SG3 SUBJ @SUBJ   "it" <NonMod> PRON ACC SG3 @OBJ <PAR1> <SENn>
"<is>"  "be" <Irreg> <SVC/A> <SVC/N> V PRES SG3 VFIN @+FMAINV #7 <PAR1> <SENn>
"<a>"   "a" <Indef> DET CENTRAL ART SG @DN> det:>9 <PAR1> <SENn>
"<matter>"      "matter" N NOM SG @PCOMPL-S #9 comp:>7 <PAR1> <SENn>
```

```
"<of>"  "of" PREP @<NOM-OF #10 mod:>9 <PAR1> <SENn>
"<importance>"  "importance" <-Indef> N NOM SG @<P pcomp:>10 <PAR1> <SENn>
"<to>"  "to" PREP @<NOM #12 @ADVL #12 <PAR1> <SENn>
"<feminists>"   "feminist" <Count> N NOM PL @<P pcomp:>12 <PAR1> <SENn>
"<as_well_as>"  "as_well_as" CC @CC cc:>12 <PAR1> <SENn>
"<to>"  "to" PREP @ADVL #15 cc:>12 <PAR1> <SENn>
"<philosophers>"        "philosopher" <Count> N NOM PL @<P pcomp:>15 <PAR1> <SENn>
"<$.>"
"<$<s>>"
"<$<p>>"
"<$<s>>"
"<Feminist>"    "feminist" <*> <Count> N NOM SG @A> attr:>2 <+INDEX-TERM> </+INDEX-TERM-1> <SEN1>
"<ideas>"       "idea" <Count> N NOM PL @SUBJ #2 subj:>3 </+INDEX-TERM> <SEN1>
"<are>" "be" <Irreg> <SVC/A> <SVC/N> V PRES -SG1,3 VFIN @+FAUXV #3 v-ch:>4 <SEN1>
"<interrelated>"        "interrelate" <SVO> EN @-FMAINV #4 main:>0 <SEN1>
"<with>"        "with" PREP @ADVL #5 ha:>4 <SEN1>
"<philosophical>"       "philosophical" <DER:ic> <DER:al> A ABS @A> attr:>7 <+INDEX-TERM> <SEN1>
"<ideas>"       "idea" <Count> N NOM PL @<P #7 pcomp:>5 </+INDEX-TERM> <SEN1>
"<$,>" cc:>4
"<but>" "but" CC @CC cc:>4 <SEN1>
"<most>"        "much" <Quant> DET POST SUP SG @QN> qn:>12 <SEN1>
"<feminist>"    "feminist" <Count> N NOM SG @A> attr:>12 <SEN1>
"<writing>"     "write" <ING> N NOM SG @SUBJ #12 subj:>13 <SEN1>
"<would>"       "would" V AUXMOD VFIN @+FAUXV #13 v-ch:>15 <SEN1>
"<not>" "not" NEG-PART @ADVL neg:>13 <SEN1>
"<be>"  "be" <Irreg> <SVC/A> <SVC/N> V INF @-FAUXV #15 v-ch:>16 <SEN1>
"<recognised>"  "recognise" <Vcog> <SVO> EN @-FMAINV #16 cc:>4 <SEN1>
"<as>"  "as" PREP @ADVL ha:>16 <SEN1>
"<$_’>"
"<philosophy>"  "philosophy" N NOM SG @SUBJ @APP @A> @<P <SEN1>
"<$_’>"
"<$.>"
"<$<s>>"
```

Appendix 2. Figures

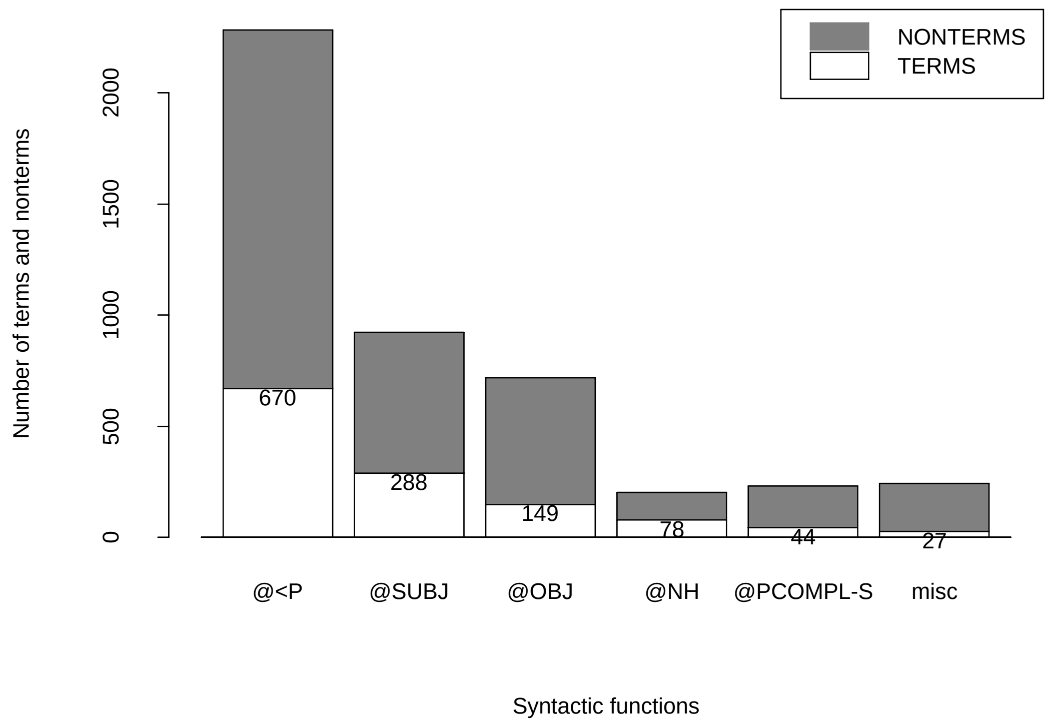

Figure 1: Syntactic functions of index terms (test corpus).

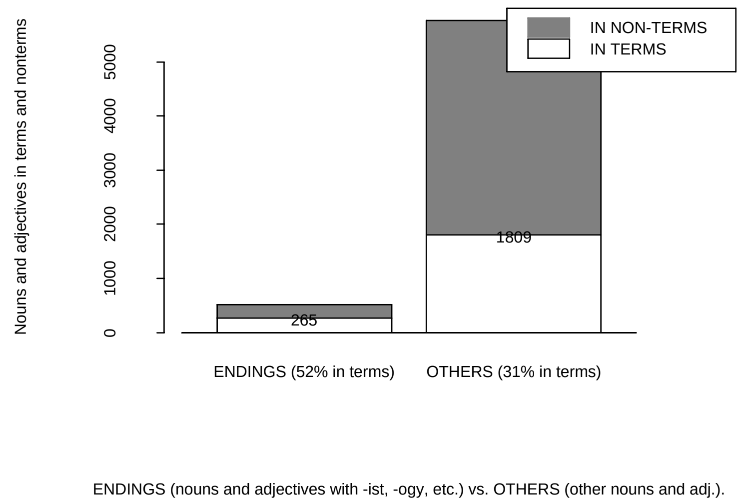

Figure 2: Nouns and adjectives as terms and non-terms, or as parts of terms and non-terms (test corpus). 52 % of the words with one of the seven endings (-ist, -ogy, -ism, -ory, -ity, -al, or -ic) has a term tag (`<+INDEX-TERM>`).

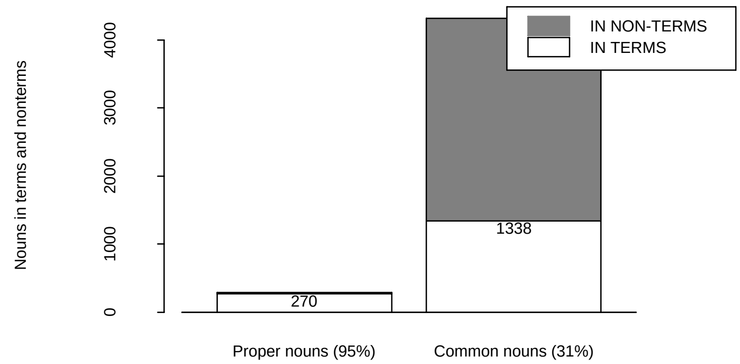

Figure 3: Proper nouns and common nouns as terms and non-terms, or as parts of terms and non-terms (test corpus).

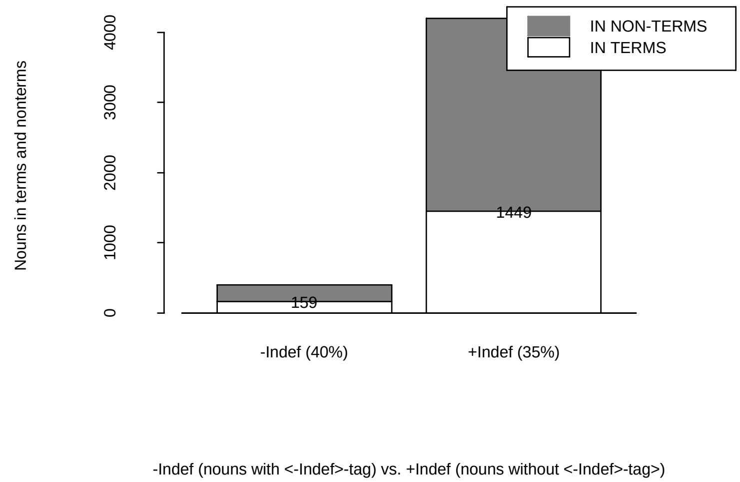

Figure 4: Nouns with and without the `<-Indef>` tag in terms and non-terms (test corpus).

Figure 5: Nouns not in lexicon (<?> tag) and nouns in lexicon as terms and non-terms, or as parts of terms and non-terms (test corpus).

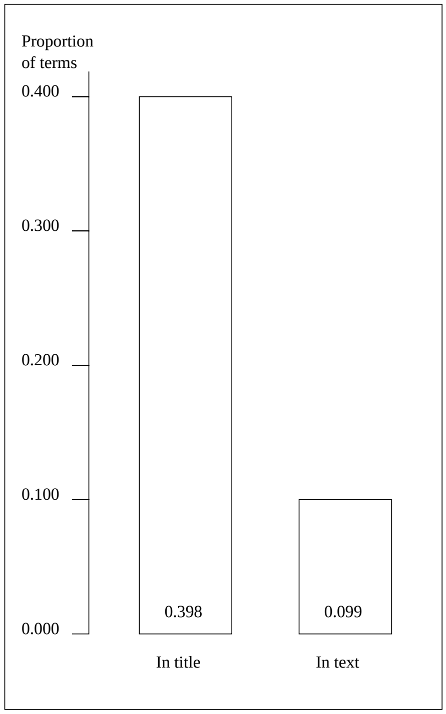

Figure 6: Proportion of terms in titles/subtitles and in text (test corpus).

Appendix 3. The top 100 term candidates ranked by *MAX-TW*

TERM CANDIDATE	MAX-TW
+ Simone de Beauvoir	1.000
+ Hegel	1.000
+ Hume	1.000
+ R. E. Vernede	1.000
+ Kafka	1.000
+ Kant	1.000
+ individualism	1.000
+ eroticism	1.000
+ Rawls	1.000
+ Nozick	1.000
+ Norman Mailer	1.000
+ neo-capitalism	1.000
+ Plato	1.000
+ Willis	1.000
+ Robert Nozick	1.000
+ inegalitarianism	1.000
+ Carol Gilligan	1.000
+ Plath	1.000
+ Keith Graham	1.000
+ Lockwood	1.000
+ Gilligan	1.000
+ Sade	1.000
+ Kohlberg	1.000
+ Sylvia Plath	1.000
+ suburbanism	1.000
+ Marx	1.000
+ J. H. Goldthorpe	1.000
+ determinism	1.000
+ capitalism	1.000
+ de Beauvoir	1.000
+ Richard Wollheim	1.000
+ feminism	1.000
+ de Sade	1.000
+ conformism	1.000
+ Lawrence Kohlberg	1.000
+ industrialism	1.000
+ woman become object	1.000
+ Graham	1.000
+ stoicism	1.000
+ D. Lockwood	1.000
+ sexism	1.000
+ Dworkin	0.977
+ Goldthorpe	0.977
+ Engels	0.977

+ Solomon	0.977
- Justine	0.977
- Samsa	0.977
+ Mead	0.977
+ McMillan	0.977
+ Platt	0.977
+ Freud	0.977
+ Sartre	0.977
+ Eisenstein	0.977
+ Luton	0.977
+ Beauvoir	0.977
+ Moye	0.948
+ Assiter	0.948
+ Almond	0.948
+ Firestone	0.948
+ Williams	0.948
+ Blauner	0.948
+ Rousseau	0.948
+ Selby-Bigge	0.948
- Rene	0.948
+ Kulfik	0.948
+ Wollheim	0.948
+ Bechhofer	0.948
+ Bernard	0.948
- Mayfair	0.948
+ Sigmund Freud	0.942
+ David Hume	0.942
+ John Rawls	0.942
+ Laporte Chemical	0.942
+ Marge Piercy	0.942
+ Margaret Mead	0.942
+ Hester Eisenstein	0.942
+ Arthur Kulfik	0.942
+ Edmund Blunden	0.942
+ Andrea Dworkin	0.942
+ Paul Willis	0.942
+ Brenda Almond	0.942
- Gregor Samsa	0.942
+ Alison Assiter	0.942
+ Jean Piaget	0.942
+ Carol McMillan	0.942
+ Greenham Common	0.942
+ Pauline Reage	0.942
+ Vauxhall Motor	0.942
+ Lewis Carroll	0.942
- lady Diana	0.942
+ Helen Weinreich-Haste	0.942

+ self consciousness	0.929
+ working-class consciousness	0.929
+ commodity consciousness	0.929
+ master-slave dialectic	0.929
+ capitalist ideology	0.929
+ working-class collectivism	0.929
+ Weinreich-Haste	0.810
+ Reage	0.810
- Cambridge	0.810

Appendix 4. The 100 least bursty term candidates of the test corpus

	TERM CANDIDATE	BURST
-	woman	6.694
-	work	6.128
-	man	5.981
-	life	5.409
-	see	5.340
-	social	4.733
-	argue	4.088
-	case	3.564
-	view	3.461
-	give	2.932
+	nature	2.846
+	male	2.597
-	suggest	2.479
-	people	2.404
-	make	2.343
-	take	2.342
-	process	2.311
-	relation	2.304
+	moral	2.301
-	change	2.182
+	fact	2.104
+	power	1.983
-	actual	1.937
-	set	1.932
-	form	1.901
-	involve	1.869
-	human	1.868
-	conception	1.833
-	live	1.812
-	use	1.755
-	real	1.711
+	class	1.709
+	labour	1.671
-	world	1.667
-	analysis	1.626
-	will	1.610
+	value	1.606
-	material	1.593
+	worker	1.562
+	individual	1.550
+	right	1.547
+	person	1.545
-	account	1.504
+	working-class	1.472

- concern	1.460
- think	1.460
+ Willis	1.442
+ attitude	1.437
- question	1.431
+ autonomy	1.419
+ morality	1.407
- notion	1.403
+ cultural	1.403
- end	1.391
- society	1.390
- direct	1.357
- manual	1.349
- describe	1.330
- action	1.326
- wide	1.323
+ autonomous	1.313
- sense	1.289
+ culture	1.286
- aspect	1.285
- simple	1.262
- ground	1.224
- reflect	1.181
+ Goldthorpe	1.179
+ structure	1.175
- means	1.166
- member	1.159
- study	1.149
+ limit	1.126
- relationship	1.122
- relate	1.105
- accept	1.103
- group	1.096
- main	1.094
+ experience	1.076
- perspective	1.070
- critical	1.069
+ difference	1.054
- lead	1.029
- situation	1.008
+ control	1.002
- say	0.994
- result	0.989
- essential	0.986
- contrast	0.985
- research	0.953
- ask	0.948

- thesis	0.946
- draw	0.938
- idea	0.934
- link	0.927
+ Lockwood	0.918
- turn	0.917
- mean	0.917
- represent	0.915
- answer	0.912

Appendix 5. The 100 most bursty term candidates of the test corpus

	TERM CANDIDATE	BURST
+	Rawls	0.000
-	population	0.000
-	allow	0.000
-	Hammertown	0.000
+	irrational	0.000
+	Nozick	0.000
+	Kafka	0.000
+	Luton	0.000
-	dilemma	0.000
-	impulse	0.000
-	masculinity	0.000
-	unfavourable	0.000
+	Eisenstein	0.000
+	Marxist	0.000
+	Sade	0.000
-	assembler	0.000
+	belief	0.000
-	carry	0.000
-	concrete	0.000
-	confirm	0.000
+	conservative	0.000
-	consumption	0.000
-	craftsman	0.000
-	creature	0.000
-	crisis	0.000
-	destroy	0.000
-	drive	0.000
-	economical	0.000
-	enquiry	0.000
+	fieldwork	0.000
-	frame	0.000
-	frequent	0.000
-	grasp	0.000
-	inferior	0.000
-	machinist	0.000
-	maintain	0.000
-	mass	0.000
-	measure	0.000
-	model	0.000
-	nearby	0.000
-	novel	0.000
-	permit	0.000
-	plant	0.000
-	procedure	0.000

- pupil	0.000
- quality	0.000
- questionnaire	0.000
- relativist	0.000
- representation	0.000
- run	0.000
- schedule	0.000
- sentiment	0.000
- sequence	0.000
- sit	0.000
- spirit	0.000
+ subjectivity	0.000
- taxman	0.000
- teacher	0.000
- theatre	0.000
- transform	0.000
+ unequal	0.000
+ virtue	0.000
- willing	0.000
- woman-centered	0.000
+ Dworkin	0.000
+ Engels	0.000
+ Freud	0.000
- Greek	0.000
+ Hume	0.000
- Justine	0.000
+ McMillan	0.000
+ Piaget	0.000
+ Plath	0.000
+ Plato	0.000
+ Platt	0.000
+ Sartre	0.000
- academic	0.000
- accurate	0.000
- achieve	0.000
- achievement	0.000
- acknowledge	0.000
- additional	0.000
- adequate	0.000
- adore	0.000
- advocate	0.000
- affect	0.000
- affirm	0.000
- afford	0.000
- alter	0.000
- analogy	0.000
- ancient	0.000

- appropriate	0.000
- appropriation	0.000
- articulation	0.000
- assert	0.000
- assume	0.000
- attribute	0.000
+ authority	0.000
- bad	0.000
- ball	0.000

Appendix 6. The top 100 term candidates ranked by *TF*IDF*

	TERM CANDIDATE	TF*IDF
+	counter-school	16.710
+	counter-school culture	16.710
+	porn	16.263
+	love-making	15.350
+	embourgeoisement	14.720
+	affluent worker	14.410
+	Willis	14.250
-	lad	14.176
-	pornographic	14.131
-	Goldthorpe and Lockwood	14.036
+	Kohlberg	14.000
+	Goldthorpe	13.955
-	Penthouse	13.661
+	Lockwood	13.588
+	working-class culture	13.588
+	embourgeoisement thesis	13.575
+	labour power	13.251
+	pornographic eroticism	12.969
+	Rawls	12.771
-	manual labour	12.640
+	Gilligan	12.138
+	patriarchy thesis	11.904
+	white-collar worker	11.762
+	master-slave	11.627
+	eroticism	11.420
+	master and slave	11.016
+	master-slave dialectic	11.016
-	man and wife	11.016
-	ethnographic	10.873
+	oppositional culture	10.861
-	Hammertown	10.861
-	standard quantitative	10.861
+	nozick	10.672
+	moral perspective	10.672
-	respondent	10.526
-	oppositional	10.451
+	Hegel	10.411
+	working-class	10.397
+	labour	10.275
-	para	10.124
+	white-collar	10.003
+	moral ideal	9.951
+	moral development	9.951
-	woman moral	9.951

+ lover	9.924
+ sexism	9.914
+ Wollheim	9.876
+ pornography	9.852
+ ethnographic material	9.832
+ working-class lad	9.742
+ Luton	9.742
- work situation	9.737
- the affluent worker	9.737
+ moral outlook	9.685
- Playboy	9.603
+ autonomous being	9.575
- page of Penthouse	9.575
- personhood	9.575
- reader of Penthouse	9.575
+ Kant	9.566
- patriarchy	9.440
- docility	9.361
- woman-centered	9.361
- relativist	9.361
+ experience of pregnancy	9.361
- taxman	9.361
+ case study	9.327
+ fantasy	9.288
- affluent	9.285
- satisfaction	9.094
+ work class	8.955
- thesis	8.797
+ counter-culture	8.734
- non-manual	8.734
+ coerce	8.591
- Kant and Hegel	8.584
+ de Sade	8.584
+ rational being	8.584
+ Sade	8.584
- woman and man	8.581
+ morality	8.479
+ moral dilemma	8.392
- life-experience	8.392
+ Eisenstein	8.392
- different moral	8.392
- systematic difference	8.392
- manual	8.372
- husband and wife	8.369
- realise	8.350
- comparison group	8.314
- world of work	8.314

- penetration and limitation	8.314
- embourgeoised	8.314
+ Skefko	8.314
- unfavourable status	8.314
- mental work	8.314
+ educational paradigm	8.314
- feminist	8.234
- manual work	8.144
- course of action	8.137

Appendix 7. The top 100 term candidates ranked by *STW*IDF*

	TERM CANDIDATE	STW*IDF
+	Willis	14.190
+	Kohlberg	13.923
+	Goldthorpe	13.865
+	counter-school culture	13.745
+	Lockwood	13.508
+	counter-school	13.419
+	love-making	13.026
+	Rawls	12.400
+	Gilligan	12.085
+	porn	11.840
+	embourgeoisement	11.695
+	labour power	10.987
+	pornographic eroticism	10.951
+	working-class culture	10.936
+	Hegel	10.368
+	embourgeoisement thesis	10.325
+	eroticism	10.209
+	master-slave dialectic	9.997
+	Kant	9.422
-	Penthouse	9.316
+	Luton	9.138
+	working-class	9.058
+	affluent worker	8.985
+	Nozick	8.976
+	de Sade	8.584
+	sexism	8.434
+	labour	8.329
+	Sade	8.254
+	Wollheim	8.051
+	Eisenstein	7.979
+	master-slave	7.903
-	Samsa	7.279
+	self consciousness	7.244
+	Hester Eisenstein	7.111
-	Gregor Samsa	7.111
-	Justine	7.065
-	manual labour	6.966
+	capitalism	6.954
+	patriarchy thesis	6.784
-	work situation	6.739
+	Plath	6.575
+	ethnographic material	6.548
-	Hammertown	6.506
+	autonomy	6.357

+ pornography	6.335
+ morality	6.291
+ Platt	6.264
- Playboy	6.083
+ working-class lad	6.025
- taxman	6.011
+ counter-culture	5.999
+ Dworkin	5.931
+ Moye	5.854
+ Kulfik	5.854
+ McMillan	5.718
+ case study	5.691
+ work class	5.642
+ Kafka	5.612
+ Laporte	5.545
+ white-collar worker	5.501
+ individualism	5.319
- docility	5.264
+ Weinreich-Haste	5.252
+ Hume	5.074
- the affluent worker	5.071
+ moral perspective	5.067
+ Rousseau	5.029
+ fantasy	4.966
+ commodity consciousness	4.904
+ Sartre	4.862
+ Bechhofer	4.858
+ oppositional culture	4.828
+ wife	4.800
+ master-slave relation	4.735
+ fantasy relation	4.735
- life-experience	4.719
+ racism	4.698
+ moral development	4.697
+ Keith Graham	4.678
- personhood	4.617
- systematic difference	4.555
+ R. E. Vernede	4.526
+ Lawrence Kohlberg	4.526
+ Carol Gilligan	4.526
- Mayfair	4.489
+ Skefko	4.485
- satisfaction	4.480
- lady Diana	4.467
+ Arthur Kulfik	4.467
+ Pauline Reage	4.467
+ Andrea Dworkin	4.467

+ Marge Piercy	4.467
+ feminism	4.458
- thesis	4.443
+ ethnographic work	4.440
- lad	4.421
- comparison group	4.416
+ working-class job	4.408
+ moral ideal	4.398
- affluence	4.367