Lizentiatsarbeit der Philosophischen Fakultät der Universität Zürich

German Terminology of Banking:

Linguistic Methods of Description

and

Implementation of a Program for Term Extraction

Referent:

Prof. Dr. Michael Hess

Author:

Angela Niederbäumer

November 2000

Contents

Introduction

The rapid progress in our society increases the need for communication among different language communities. This development requires quick translations of specialised texts. The growing demand for specialised translations brings about a large amount of special terminology. This represents a problem for a translator, who in most cases is not a subject specialist. To find the correct translation of special terms is a laborious task. A translator, though, has a major interest in reducing the amount of time spent on translations and cannot dedicate herself to time-consuming inquiries on term equivalents.

The importance of quick retrieval of terminology has not gone unnoticed in translation services, where a large amount of money is invested in order to give translators easy access to terminology. Today, modern translation services have their own terminology departments, where terms are entered in terminology data bases and kept up-to-date. The key issue in terminology departments is to gather a wide range of multilingual documents and to check them for usable terms. Because of the growing demand terminology must be retrieved and handled quickly. Manual term identification, though, is not an effective method. Instead, the employment of computer programs is required which help the terminologist in the retrieval of terms.

The recognition of terms is a task which requires complex cognitive abilities. A terminologist identifies terms because she understands the lexical and grammatical structure of a language, but most of all because she understands its semantic organisation and thus is able to map concepts into words. In cases of uncertainty about the nature of terms and their difference to non-terms a terminologist relies on answers provided by theoreticians in the field of terminology. The discipline of theoretical terminology is concerned with the conceptual character of terms, the semantic relations among candidate terms, and the determination of whether concepts are denominated by several terms or whether terms are denominating several concepts. It is the privilege of human beings to deal with these questions.

A computer program is not very effective in the semantic field, it cannot cope with the complex and interconnected nature of abstract concepts. In order to overcome this obstacle computational linguistics has provided alternative ways of looking at terms. Specialised terms are not analysed and extracted in a conceptual context, as done by humans, but according to their morphological and syntactic characteristics. A good theoretical background for a structural analysis of terms as opposed to non-terms originates from contributions in the field of language for special purpose (LSP). Studies on LSP have shown that noun phrases are dominant in LSP texts, while words of closed word classes are rare. Further, studies on term formation reveal that terms are mostly noun compounds.

The establishment of term characteristics in LSP is the foundation of methods for automatic terminology retrieval (ATR). Compared with a conceptual identification of terms as done by humans, though, a linguistic-based categorisation has many shortcomings. A reason for the disappointing results is the syntactic similarity of terms and non-terms. The automatic retrieval of terminology by means of syntactic characterisations only is connected with considerable editing work for a terminologist. The combination of a syntactic and a morphological component for term extraction is more promising, since this allows to extract terms also according to domain-specific morphemes.

There have been and still are valuable attempts of understanding and extracting terms on a structural and/or statistical basis. Most of these works focus on technical terminology, some on medical terminology. Some works follow domain-independent strategies. To my knowledge, there have been no ATR projects focusing on the language of banking.

A possible solution to the automatic extraction of banking terms in texts is presented in this work. A computer program to assist terminologists or translators in the retrieval of German banking terms from pertinent documents was implemented. A first component in this ATR process consists of a set of programs for natural language processing like a tokeniser, a morphological analyser, and a parser. These perform the task of recognising noun phrases in a text. A second component uses regular expressions to find occurrences of domain-specific morphemes among these noun phrases.

This thesis is structured in two main parts. Part 1 includes sections which examine theoretical questions of terminology: In section 1 the working methods of translation-oriented terminology are outlined and the new needs concerning terminology work, namely automatic term retrieval, are brought up. In section 2 the focus is on the theoretical criteria for distinguishing terms from non-terms. Since theoretical methods are not reliable for ATR, section 3 examines the terminological unit in relation to LSP; studies on lexicon and syntax of special language are discussed according to their relevance for ATR. The assumption that terminological units in LSP obey specific rules of syntactic formation will be examined empirically with the help of CS-term, a data base with banking terminology compiled manually. Section 4 traces the historical development of banking language, and identifies some particularities of banking terms.

Part 2 of this thesis is concerned with methods of ATR and with the implementation of a program for retrieval of banking terminology: In section 5 works with achievements relevant in the field of ATR are discussed. Section 6 refers to the characteristics and structure of CS-term, presents the method of analysis for CS-term, provides a description of the computer programs employed for term extraction, describes the evaluation, and presents the results achieved by this ATR program.

Part 1: Linguistic Methods for Describing Terms

In this part, first the difficulties of transferring conceptual knowledge on computers will be discussed, then the theoretical procedures humans rely on when identifying terms will be outlined, and finally characteristics of special language in general and of banking language in particular will be presented.

1 Translation-Oriented Terminology

1.1 Terminology in Specialised Translation

Translation of specialised language is quite distinct from the translation of general texts. In general translation problems frequently encountered are whether it is better to translate literally of freely, how to preserve rhyme and rhythm of the source language in the target language, or how to deal with phenomena like word plays and proverbial sayings. These questions are not at the centre of specialised translation. Specialised translation is meant to improve the communication among experts of a domain and to contribute to the distribution of knowledge across language boundaries. The specialised translator is responsible for transferring information among experts of various speech communities. The need for international communication in economy, politics, science, technical fields and more will continue to increase the necessity of specialised translations.

As we will see later, a typical characteristic of specialised texts is their complex syntactic structure, a difficulty, a trained translator can cope with. The major problem encountered in specialised texts is the terminology, which requires specific concepts from foreign languages to be transferred into one's mother tongue. Before starting to translate, the translator must become familiar with the terminology of the domain. To find the aptest equivalents to the terms in question is often a time-consuming task even for skilled translators. Translators are not free to define unknown terms which may be too specialised to occur in a dictionary. They have to search in previously translated documents, on the internet, or contact domain specialists. Studies show that a translator needs 40 % of the working time for inquiries on unknown terminology (Stellbrink 1988:203).

No matter how well-educated a translator is, if working in a special field there will always be gaps in her knowledge of term equivalents. The specialist translator will constantly be faced with new terms because specialised language tends to be productive in this regard. At the same time, translation services work on a time-saving basis and in order to remain competitive little time should be spent on inquiries of unknown terms. Despite these exigencies, the quality of the translations must be guaranteed.

One of the crucial requirements for delivering high-quality translations of texts is the systematic approach to domain-specific terminology. For this task, modern language services have their own terminology centres and it is the responsibility of the terminologists to supply the professional translator with relevant terms. A terminology centre makes inquiries on special terms in several languages and provides terminology data bases. These efforts ensure the consistency and correctness of terminology and speed up the translation process.

The growing importance of terminology is mirrored in the fact that ever more money, time, and effort are directed at the retrieval of terminology. Today, the installation of efficient terminology departments is a prerequisite in every translation service. This development has led to a growth of software firms which offer programs for terminology management. Further, private and political institutions support the development of methods and tools which enable efficient terminology work. Translation schools start to teach terminology as a discipline of its own, and an increasing number of courses in terminology is offered within the training programmes of translation centres. Translation-oriented terminology is highly developed in bilingual or multilingual institutions and communities like Switzerland, Canada, the European Union, or the United Nations. The linguistic requirements of these multilingual administrations, the economic systems in these countries, their policy, education systems, etc. create a great demand for terminology work.

The increasing need for terminology explains the strong interest in new working methods like the automatic retrieval of terms. But before we go into this subject, the conventional working procedure of multilingual terminology will be traced.

1.2 Working Methods in Terminology

The main task in terminology consists in building a terminological data base containing relevant terms. For this purpose the terminologist must identify current and future needs of translators. In cases in which terminological equivalents are unknown, a terminologist has to interrogate domain experts. Otherwise new terms are gathered from multilingual documentation. The manual identification of terms in documents is a laborious task. This process involves reading through the selected texts, listing all potential terms, and then entering those terms which are not yet included in the data base. In order to adjust the existing entries to possible changes in usage, the terminologist also needs to check newly identified terms against the ones existing in the data base.

A terminologist identifies terminological equivalents of different languages according to the conceptual relations between terms. The main linguistic utility a terminologist relies on in this process is the definition of the terms. Since in most cases terms are not provided with a definition but only with a contextual environment, the definition has to be inferred from the context. For each new entry, the terminologist provides additional information like synonymy, subject fields, sources, and grammatical information next to the term equivalents, definitions, and/or contexts.

As we will see soon, the most important question in theoretical terminology and also the key question in automatic term retrieval is the distinction between terms and non-terms within one language. Interestingly, this question is hardly ever discussed in practical terminology. An experienced terminologist seems to decide intuitively about the termness of a lexical unit and whether it needs to be entered in the data base. An explanation is that terminologists often have translation practice and know the needs of translators. Generally, terminologists consider a lexical unit which is not part of their general knowledge and is not yet in the terminology data base to be a valid term. For a terminologist, the distinction between term and non-term is a matter of intuition, almost of feeling. She does not verify whether a lexical unit fulfils the theoretical criteria concerning termness, partly because not even theoretical criteria can give an absolute answer to the question.

1.3 New Methods of Terminology Work

Every field of our modern society is in constant development, which means that in each one of these fields new terminology is continually coined. This process is so fast that it is impossible for terminologists to record all newly formed words, not even those created within their own institution. Neither the amount of time nor the number of terminologists needed to analyse only a small part of the texts containing potential terminology can be afforded. With automatic methods, the examination of large quantities of text could be completed within seconds. The terminologist just needs to select the correct terms among those provided by the computer tool. In this way many tedious tasks in terminology would be eliminated.

The employment of a program for term extraction does not mean that a computer is able to perform the task in exactly the same way in which an experienced terminologist does it manually. A terminologist retrieves terms thanks to her own understanding. A computer program, however, extracts candidate terms after a linguistic or statistical analysis of the lexical units. Today, automatically retrieved lexical units are far from being valid terms, although there are many computer programs which attempt at extracting terminology. There are still numerous limitations which arise from the use of term extraction programs in terminology, as we will see in the following section.

1.4 Problems of Automatic Terminology Retrieval

The manual identification of terms is solely based on concepts. One needs to describe concepts, to distinguish between concepts, to classify concepts within a conceptual system. This process succeeds because humans understand the texts they read, the context of the words, the semantic details. This understanding allows them to make decisions about potential terms.

As a matter of fact a computer program cannot (yet) be provided with such competence. A computer can only have the knowledge if it is programmed into it. And as is well-known, a full specification of all the factors used by the human mind to make decisions would be extremely difficult, if not impossible, to program. This is the central problem not only in ATR but in all automatic language analyses. Since a computer cannot recognise terms on a conceptual basis, most attempts at implementing ATR programs focus on the formal aspects of language; some methods employ statistical filters. An overview of these ATR methods will follow in the second part of this thesis.

The performance of available ATR tools is rather disappointing. Usually, ATR generated lists contain too many general words. This explains why ATR programs have not had a breakthrough in terminology practice. ATR tools can be of assistance to the terminologist, provided that the manual validation of the computer generated list does not become a laborious task. At the moment, the post-editing required is still time-consuming. In the course of this thesis we will see that the performance of available ATR programs is too poor for them to be a valid alternative to manual term retrieval. Yet, because of the increasing need for terminology the process of term identification cannot continue to be carried out on an entirely manual basis. Improvements of the performance of ATR tools are possible and will be widely discussed in this thesis. However, it is also important that terminologists lessen their expectations regarding ATR tools. Terminologists tend to expect an automatically generated term list which equals the results of a manual retrieval. They do not see any usefulness in a program which provides invalid candidate terms, even if a fair number of terms is correct. By becoming more familiar with the methods employed in ATR terminologists would have a better understanding of the factors which lead to errors. It is important for them to realise that the goal of ATR programs is not to replace them. The idea of ATR is to have a computer program which reads a text and produces a list of words among which the terminologist can choose those of relevance. Even if the output of an ATR program has to be post-edited, it makes terminology work much more efficient and accurate.

In this introduction to the world of practical terminology the most important concepts of this area have been touched upon. These elements shall be examined in greater detail in the next section.

2 Terminology Theory

2.1 Terms and Words

It is not easy to determine the criteria distinguishing terms from non-terms or words. A main difficulty is presented by the subjective and social factors underlying this question. An individual assessment of whether a lexical unit belongs to a particular domain is determined by the education a person has received. The lower a person's level of education, the narrower her view of what constitutes general knowledge will be. In fact, a person with a low level of education may attribute many more lexical items to special language than someone with a high level of education. On the other hand, a well educated person will have a wider conception of what is general knowledge and, at the same time, be able to distinguish between a greater number of special terms. In short, the same lexical item can belong both to special language and to general language depending on a person's education.

In practical terminology one is not concerned with these issues. In fact, as already stated in the previous section, the theoretical question of the termness of a lexical unit is hardly ever a point of discussion in terminology work. Terminologists rely on their own experience and intuition when differentiating terms from non-terms. For terminology theoreticians, however, it is not acceptable to limit the termness of a lexical item to social factors like the level of education. It is not a satisfying solution to rely on subjective matters like one's own intuition. Instead, they provide theoretically grounded answers to the question of how to define and delimit terms.

In the next sections, the terminological criteria used to distinguish between terms and words will be outlined. First there will be a brief account of the development of terminology as a discipline. It will be shown that contributions of the theoretician Eugen Wüster built the fundament for establishing terminology as a discipline. Then characteristics differentiating terms from words will be discussed.

2.2 Terminology Schools

Studies on the vocabulary of subject fields such as chemistry, zoology and botany, medicine, or mathematics started in ancient times. But the first attempts at establishing terminology as an independent linguistic discipline were made in the first half of the 20th century, as the early era of globalisation put more emphasis on the necessity of communication between different countries. In Western Europe the major exponent was the Austrian Eugen Wüster, in the former Soviet Union the influence came from D.S. Lotte (Laurén/Picht 1993:493). During these early stages, the efforts in the field of terminology were motivated by the necessity of standardization, particularly in the technical and scientific fields. In the first half of the 20th century three major orientations in terminology can be established: the schools of Prague, of Russia, and of Vienna (Laurén/Picht 1993:493-499). A more recent influence in terminology work comes from the Canadian community (Sager 1990:212).

2.2.1 The Prague School

According to Laurén/Picht (1993) the Prague terminology school originated from the Prague school of linguistics (495). Its major proponent was Drodz. This school is almost exclusively concerned with the structural and functional description of special languages. A further interest is the standardization of languages and terminologies.

2.2.2 The Russian School

The Russian school is based on works of D. S. Lotte and S. A. Caplygin (Laurén/Picht 1993:495-496). At the outset of the Russian interest in terminology lay the scientific and technological progress taking place in the country. From this development, and given the multilingualism of the former Soviet Union, the need for standardization of concepts and terms arose.

2.2.3 The Vienna School

Laurén/Picht (1993) maintain that the Vienna school (also called 'Western school') is based on the work "Allgemeine Terminologielehre" by Eugen Wüster, one of the founders of modern terminology theory (498). In his work Wüster treated themes like concept formation, conceptual systems, term formation, and term definition. All of these methods are adopted in today's terminology work (498). The Vienna school stemmed from the need of mathematicians, physicians, and technicians to standardize the terminology of these fields (497). The standardization of terms is thus an important feature of this school. Most countries in central and northern Europe work within the framework of the Vienna school (498). Wüster's contribution gave shape to the theoretical basis of modern terminology and established methodological principles in terminology work (499). Wüster's four-field model of semantic representation (section 2.3) contains many principles of theoretical terminology and is an attempt at resolving the problem of the distinction between terms and general words.

2.2.4 The Canadian Approach

The Canadian interest in terminology is motivated by the decision to make French a parallel official language to English (Sager 1990:212). Next to relying on western European experiences, Canada also developed working methods on its own, which regarded particularly the introduction of computers for terminology processing (Sager 1990:212). The focus of the Canadian approach is also on methods of term formation like the creation of neologisms (more in section 3.3.1.3).

2.3 Wüster's Model of Semantic Representation

In terminology theory the way in which the meaning is structured follows a different interpretation from the one presented in the traditional semantic triangle of C.K. Ogden and I.A. Richards (figure 1). The semantic triangle claims that meaning is essentially a threefold relationship between linguistic forms (symbol), concepts (thought of reference), and referents (objects identified by means of word or expression) (Crystal 1997:345).

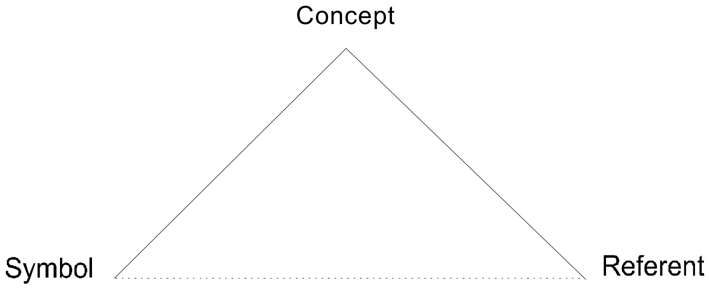
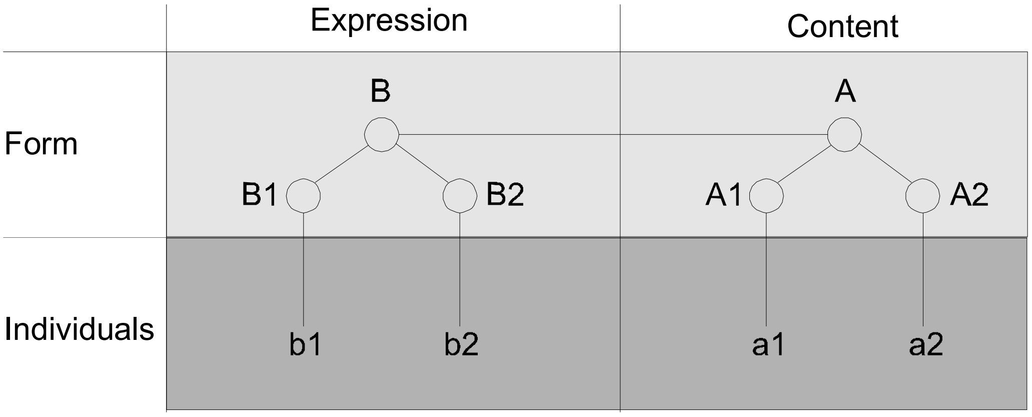


Figure 1: Ogden and Richards' semantic triangle. The dashed base line indicates that symbol and referent are not related.

The semantic triangle was proposed by Ogden and Richards in order to illustrate that there is no relevant relation between the symbol and the referent, that they are not directly connected. They did not mean to illustrate linguistic phenomena like polysemy (i.e. one expression form can refer to more than one concept form) and synonymy (i.e. one content form can be represented by two expression forms). In terminology, however, it is very important to identify these two types of relationships between words in order to achieve context independency of words.

In order to illustrate polysemy and synonymy, Wüster took the semantic triangle as a starting point for his four-field-model (figure 2). In accordance with Ogden and Richards' triangle he distinguished between the level of linguistic expression (*symbol* in figure 1; *b1* and *b2* in figure 2) and the level of semantic content (*referent* in figure 1; *a1* and *a2* in figure 2). He calls this the level of concrete individuals (dark grey in figure 2). Wüster's model takes into account that phonemes and graphemes are also concepts and that allophones and allographs are the individual realisations of these concepts. Thus, at the level of concepts he creates two separate fields to distinguish between the concept representation of the content (*A* in figure 2) and the concept representation of the expression (*B* in figure 2). He calls this the level of abstract concepts (light grey in figure 2). Further he splits the concept levels of both content and expression into two branches (leading to *A1, A2* and *B1, B2* in figure 2) in order to identify polysemy and synonymy.



a1 and a2	represent the individual objects in reality
A1 and A2	are the individual concepts representing a1 and a2
A	is the abstract concept representing A1 and A2
B	is the abstract concept of the symbolic representation of A
B1 and B2	are the individual abstract representations of a phonetic or graphic form
b1 and b2	are the individual phonetic representations of B1 and B2

Figure 2: Wüster's four-field model (adopted from Cabré 1999:41). The lower right field contains the individual objects, the upper right field depicts the concepts. The upper left field contains the signs as concepts (phonemes and graphemes) and the lower left field the respective realisation of these signs, i.e. the different allophones and allographs.

In the next section, characteristics provided by terminology theory concerning the differences between terms and words will be discussed.

2.4 Characteristics of Terms

Usually words are assigned to the vocabulary of general language while terms form the vocabulary of special languages, as the following definitions reflect:

> "The items which are characterised by special reference within a discipline are the 'terms' of that discipline, and collectively they form its 'terminology'; those which function in general reference over a variety of sublanguages are simply called 'words', and their totality the 'vocabulary'". (Sager et al. 1990:19)
>
> "Terms, like words in the general language lexicon, are distinctive and meaningful signs which occur in special language discourse." (Cabré 1999:80)

However, such straightforward characterisations are of little help. They are of a general nature and do not clarify the attributes which distinguish terms from words. In order to define terms one needs to contrast their special characteristics to the ones of words.

2.4.1 Context Independency

Words are investigated under specific view points like semantics, grammar, phonology, or orthography. Researchers concerned with terminology usually study the semantic aspect of words, since terms are defined by the content of lexical units. From a semantic point of view a word is "one of the smallest, completely satisfying bits of isolated 'meaning'". (Sapir 1921:34 [quoted from Kageura 1995:241]) In reality, many words have an imprecise meaning which cannot be isolated and described in a satisfying manner based on themselves alone. In many cases the meaning of a word depends on the context it is embedded in. The word 'table' for example, refers to a systematic display of numbers or other items in the context of this paper, but in the more general context of every-day life it refers to a raised board at which people may sit and take meals. This example displays the first major difference between words and terms:

> "Terms designate concepts which, within a specialised field, are independent of any context." (Desmet/Boutayeb 1994:310)

In a given discipline terms are used for one concept only, while in general language most words have multiple meanings. The lack of polysemy within one specialised domain accounts for the context independency of terms. In terminology, the polysemic nature of language can be avoided because terms are the result of convention, because they are formed based on an agreement between specialists of a given field, and because they are motivated by the intention to facilitate communication in a specialised domain (Sager 1990:56-57). Words, on the other hand, are the result of spontaneous formations, influenced by a number of social and cultural factors (Reinhardt et al. 1992:24). In word formation there is no authority which makes sure that new words are not sources of ambiguities and thus of misunderstandings.

As already stressed, the notion of context independency is valid only within one domain. Within texts referring to economic domains terms like *bank, liquid,* or *deposit* loose their lexical ambiguity and do not need an in-depth examination of their contexts. Within a hydrologic environment they are also understood without considering any context: the bank of a river, liquid as the aggregate form of water, or the deposit of sediments. Although these terms are polysemic with other specialised terms, an expert of a given domain can easily identify them in a specific concept. Within a general text though, one must rely on the context in order to solve the semantic ambiguity in the sentence *I saw the bank*. Only the context allows to determine whether *bank* refers to a financial institute or the board of a river.

2.4.2 Conceptual Representation

Another fundamental element of terms was repeatedly mentioned throughout the previous sections of this thesis and is also revealed by the above terminological definition of Desmet and Boutayeb's: their conceptual representation. This characteristic is also encountered in other definitions:

> "[A term is the] designation of a defined concept in a special language by a linguistic expression. [A concept is the] unit of thought constituted by those characteristics which are attributed to an object or to a class of objects." (ISO/DIS 1087 1988:7 [quoted from Arntz 1995:40 and 43])
>
> "The primary objects of terminology, the terms, are perceived as symbols which represent concepts. Concepts must therefore be created and come to exist before terms can be formed to represent them." (Sager 1990:22)
>
> "Le terme se définit comme unité signifiante constituée d'un mot (terme isolé) ou de plusieurs mots (termes complexes) qui désigne un concept, de façon univoque à l'intérieur d'un domaine." (De Bessé 1990:253)

The essential aspect in the characterisation of terms is the concept. Sager (1990: 22) sees concepts as "constructs of human cognition processes which assist in the classification of objects by way of systematic or arbitrary characterisation." A concept is thus the mental representation of an entity which enables people to recognise and understand the world. Terms as the denotation of concepts result from a convention, from an agreement between experts of a given domain. Their description of concepts "refers to the field of knowledge to which that concept belongs." (Desmet/Boutayeb 1994:309) Words on the other hand are not defined in terms of concepts but of semantic meaning. Conceptual representation versus semantic meaning is a parameter for distinguishing terms from general words.

It may be debated whether there is any reason to treat the conceptual content of terms differently from the semantic content of words. Several theoreticians have engaged themselves in this discussion, partly with different points of view. For Desmet/Boutayeb (1994) this apparent contradiction can be clarified. In fact, a term can be identified as such only if it refers to a concept which can be placed within a conceptual system of a particular domain. Words, however, "refer" less than terms (316). They appear in utterances, and it is the function of these utterances and not of the words themselves "to refer, that is, to describe a certain universe and to make assertions about that universe." (316)

For Kageura (1995), however, this distinction between terms and words is unsatisfactory. He remarks that "the relative status of terms, concepts, and their relationships in conceptual descriptions of terminological phenomena is exactly the same as that of words, meanings, and their relationships in semantic descriptions of words." (255) The independent pursuit of concepts and terms does not constitute a theoretical explanation of terms, as "it cannot theorise on the specific nature of terms without going into straightforward tautology." (251)

Generally, terminology theoreticians like Sager, Arntz, Picht, or Cabré take the first view. Their contributions can be broadly summed up as follows: a lexical unit is a term of a particular domain if it belongs to a structured system of concepts of that domain; a lexical unit is a word of general language if it cannot be placed within a systematic conceptual system of a given domain. Consequently, terms are understood as conceptual descriptions, whereas non-terms are attributed a semantic description.

If one adopts this last standpoint, the placement of a term within a given conceptual system is a criterion for determining its termness. This process is based on features, which can be seen as the atoms of concepts (Arntz/Picht 1995:54). Features themselves are also concepts. The concept 'shareholder' for example, is defined in *The New Shorter Oxford English Dictionary* (1993) as a "person who owns a *share* or shares" (2813). The concept 'stockbroker' is defined as a "*member of stock exchange* dealing in *stock* and *shares*" (3067). Thus, the definitions of these concepts are composed of other concepts like 'share', 'member of stock exchange', and 'stock'. We also identify a hyponymic relationship between the concept being defined and the concepts functioning as features in these definitions. Features of concepts have a central function in terminology. They fix the content of a concept, they establish systems of concepts by means of their hierarchical relationship, they are useful for the determination of synonymous terms (Arntz/Picht 1995:53). Features are also essential in the formation of new terms (section 3.4.1), as is evident from the examples above, which is of high significance for the automatic retrieval of terminology.

2.4.3 Terminological Definitions

The features of a concept are found in the definition of a specific concept. Desmet/Boutayeb (1994) understand the definition of a term as the representation of the concept (311). Terminological definitions differ from definitions of general words in the way that a "traditional definition begins with the generic characteristic. But, while in a [general] dictionary this characteristic is often limited to an approximation close to common experience, a terminological dictionary must provide a generic definiens as close as possible to the concept defined." (311) A closer investigation of the aspects which distinguish terminological definitions from other types of definitions can be found in De Bessé (1997:63-73).

2.4.4 New Approach: Formal Characterisation of Terms

After having characterised the specificity of terms as opposed to words, the question of the nature of a term might have become clearer for human beings. However, for ATR purposes this discussion is of no avail. The ways of identifying terms discussed so far require a complete understanding of conceptual meaning. A term is a lexical unit which designates a concept within a particular domain. When spotting terms a human deduces a variety of semantic and pragmatic details. This task is difficult to achieve with a computer program which would need to have a complete theory of human thought and world knowledge. Term identification on a conceptual level must be performed by an intelligent human being. If terms are to be retrieved by a computer, we need to explore them not at their conceptual level but at their lexical and syntactic level. In these areas computational linguistics has developed reliable methods for the automatic processing of language.

From a lexical point of view, the interesting question is how terms are formed and whether there are any recurring methods of terms formation. From a syntactic point of view the most evident difference between terms and words is that terms are mostly noun phrases. The question at this point is whether term formation and syntactic particularities of terms are typical of special language or whether they occur with the same frequency in general language. If there are any such characteristics in special language, they can be exploited for ATR purposes. This investigation will be carried out in the next section, in which the characteristics of special language relevant for ATR will be identified.

3 Language for Special Purpose (LSP)

3.1 Defining LSP

Special languages are subsets of the set of language as a whole. Among these subsets is also general language. Every single LSP can be said to "intersect with general language, with which it not only shares features but also maintains a constant exchange of units and conventions." (Cabré 1999:64) Further, special languages also intersect with each other. Figure 3 visualises these interchanges within the language system as a whole.

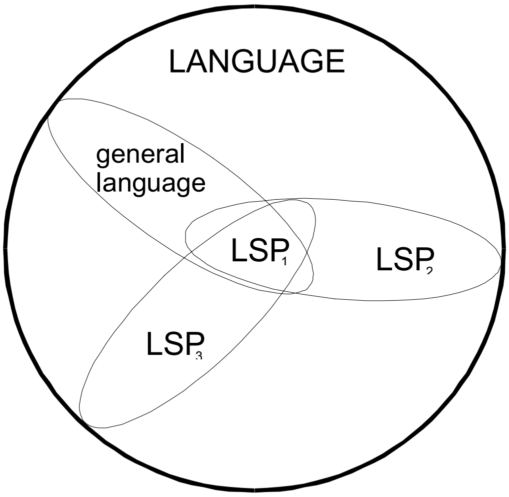


Figure 3: The subsets of special languages and general language in the set of language as a whole (adopted from Cabré 1999:66)

Special languages differ from general language particularly in their specialised vocabulary, or terminology; further in the structural characteristics of sentences (Fluck 1996:12). LSP is understood as a language which "is produced by a specific society and used by the group of people sharing the same profession or subject." (Nomoto 1980:562 [quoted from Kageura 1995:245]) In other words, LSP is special with respect to the content of the discourse and because it is used by people specialised in some profession. In everyday life speakers show little interest in special languages and their terminologies, even if these are a fundamental factor of our life. In fact, terminology is the foundation of communication not only in sciences like medicine, physics, chemistry, or engineering but also in less scientific areas of our culture. One just needs to think of the game of baseball, for which the knowledge of the meaning of terms like *fastball, slide ball, forkball, inning, home run, base stealer, balk* etc. are compulsory in order to understand the game.

When talking about special languages one has to be conscious of the fact that among these subsets of language there are fields which have very little in common with each other, not only concerning the knowledge required but also as regards linguistic features. Specialised articles about sports, theatre, or religion for example, are syntactically much closer to general language than to certain specialised languages like legalese. For this reason, it is important to choose the approach which best serves the aspects of investigation before describing a language of a special domain.

3.2 Approaches to LSP

Sager et al. (1980) identify three elements for a general definition of language: "the fact that a language is used by different groups of people, the fact that language makes reference to our knowledge of the world, and finally the fact that language is a system and therefore has particular structures and methods."(6) Accordingly they distinguish three fields of investigation of language: the field of pragmatics, of semantics, and of syntax. These fields are the basis for the description of LSP and for a classification of special language according to models. The following sections contain an overview of these approaches according to the arrangement in Sager et al. (1980:6-10).

3.2.1 The Pragmatic Approach

The pragmatic or user-oriented approach is the most frequent approach to LSP. It considers usage in connection with particular speech situations which is then contrasted with the occurrences of general language. Speech situations are the circumstances under which individuals use language. Examples of speech situations are the subject field, the type of user, the user's social status, the geographical location, or the type of environment in which a communication takes place. Sager et al. (1980:6) observe that special languages are recognised as "pragmatic or extra-linguistic subdivisions of a language" and that an attempt to explain LSP in terms of grammar is difficult. Still, pragmatic methods of describing LSP do not satisfy the needs of ATR. In order to identify terms automatically one must draw from grammatical resources.

3.2.2 The Semantic Approach

Semantic approaches do not consider the user of language but only the expression and its referent. Semantics is "the most important area of investigating whether there are any linguistic means of signalling special subject language items" (Sager et al. 1980:8). In agreement with Wüster's model of linguistic reference (section 2.3), this approach investigates relationships of lexical semantics like synonymy, polysemy, hyponymy.

As it was repeatedly pointed out, semantics holds a central role in the process of manual term identification but is of little use for ATR.

3.2.3 The Syntactic Approach

The syntax of language is not determined by semantics. A syntactic approach abstracts from the referents, or in terminological words from the concepts, and analyses only the structural relations between the expression. Since LSP cannot be explained in syntactic terms only, this approach has apparently received little attention. The problem of a syntactic approach is that "it is concerned with units no larger than the sentence, a level at which it is more difficult to single out special language features." (Sager et al. 1980:9) For the computational analysis of LSP, though, this approach is very important. The syntax of LSP will be discussed in depth in section 3.4.

3.2.4 A Model of Linguistic Features for Describing LSP

For a linguistic classification of LSP Sager and his fellow researchers developed a model which shows how the characterisation of special language is guided by pragmatic, syntactic, and semantic criteria (figure 4). Morphological criteria have been neglected in their description because these have little effect on LSP (9). However, a discussion of lexical features - a measure of analysis underlined in this model - also includes morphological considerations, as the extended discussion of this linguistic field in Sager et al. (1980:230-295) confirms.

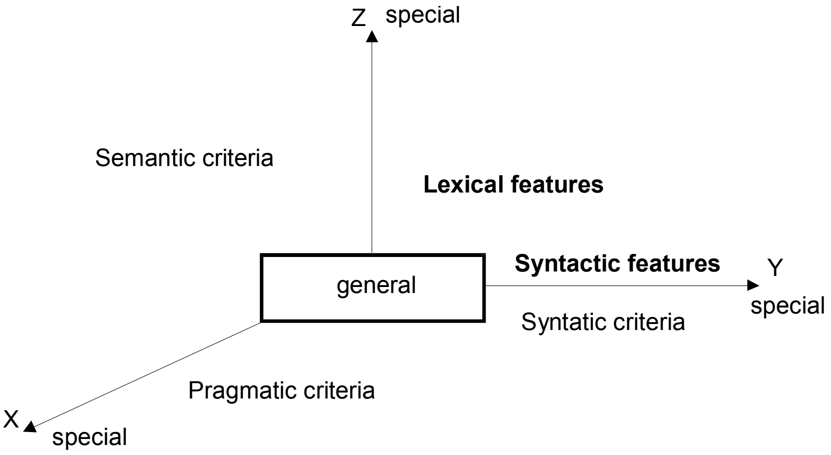


Figure 4: A model of linguistic features (adopted from Sager et al. 1980:9)

The authors' description of this model reads as follows:

> "The X axis would represent such major categories as report, handbook, contract, deposition, according to their distance from general language forms. The Y axis would give sentence and phrase structures as they occur in special language. The Z axis would indicate the increasing specialisation of lexical items." (Sager et al. 1980:9)

The advantage of this model is that it allows a gradual classification of LSP according to linguistic features. The approach taken in this thesis is situated along the YZ axis, thus investigating the different lexical and syntactic features of LSP and general language. The goal of analysing these aspects is to clarify the distinction between terms and non-terms for ATR purposes.

3.3 Lexical Level

As already stated in the previous section, LSP differs from general language particularly in its vocabulary. It is thus no surprise that the lexical aspect of LSP has been paid great attention. This area comprises knowledge of the phonological structure of lexicon entries, of their syntactic and morphological categorisations, and of their semantic representation. The lexicon is thus seen as an intersection where all the different information of language knowledge meet. For terminological purposes the phonological side of the lexicon is not of relevance. The semantic side is crucial in terminology work. However, this side is hardly considered in this thesis because meaning is difficult to fix automatically. The syntactic side will be treated extensively later. In the next section, the morphological side investigated is term formation. Rules of term formation disclose how general words are extended into terms and how the creation of new terminology takes place.

3.3.1 Term Formation

Sager (1997) defines term formation as the process of naming the concept required by a particular domain (25). He adds that term formation differs from general word formation "by its greater awareness of pre-existing patterns and models and its social responsibility for facilitating communication and the transmission of knowledge." (25) By means of the repeated act of creating new terms and of regulating existing terminology, a certain consistency of designation is achieved (25). Terms are on the whole less arbitrary and more consciously motivated and transparent than general words (27).

These reflections make evident that the wide-spread opinion that special vocabulary is mostly constituted by Latin, Greek, and nowadays English based terms is misleading. New terminology is not only created by importing foreign vocabulary, but new terms are also formed by employing other methods.

The major distinction made in term formation is primary versus secondary term formation (Sager 1997:27-28). "Primary term formation is the process of terminology creation that accompanies concept formation" (27) as a result of any kind of innovation. Thus, primary term formation takes place in a scientific environment, and there is no direct linguistic precedent of the new term created. "Secondary term formation is the process of creating a new term for an existing concept" (27). This occurs when a designation is changed as the result of terminology revision. In German banking terminology for example, the term *Mitarbeiteraktie* 'employee share' has substituted the older form *Belegschaftsaktie*. Secondary term formation is also involved when concepts are transferred among linguistic communities. When speaking of derivatives for example, the term *swap* is frequently used in German compounds (*Währungsswap* 'currency swap', *Zinssatzswap* 'interest rate swap', *Devisenswap* 'foreign exchange swap', *Schuldenswap* 'debt-for-equity swap', etc.).

The process of primary and secondary term formation is carried out by three different methods (arranged in this way by Sager 1990:71-80 and 1997:28-40):

- the use of existing resources (secondary term formation)
- the modification of existing resources (secondary term formation)
- the creation of new linguistic entities (primary term formation)

These three methods are employed in general word formation as well. However, while in general language the creation of new words is a spontaneous process, in terminology the formation of new terms is always motivated (Reinhardt et al. 1992:24). New terms are formed because particular circumstances require them. As will be shown later in this thesis, in German banking terminology existing lexical units are usually combined into new terms. This explains the fact that terms are in most cases compounds consisting of two or more stem words.

The following sections summarise the most common methods of term formation. It is a synthesis of contributions from Sager (1990:72-80 and 1997:25-51), Arntz/Picht (1995:119-127), and Reinhardt et al. (1992:19-23). According to Arntz/Picht (1997:119) these methods originate from a research done by Drodz and Seibicke in 1973. Sager and Reinhardt et al. do not mention any sources.

3.3.1.1 The Use of Existing Resources

This method refers to the extension of the meaning of an existing term so as to include a new concept. The most common extension of meaning is to exploit the polysemic nature of general words, as in the case of terminologisation. Another way of assigning new designations by means of existing linguistic resources is to use similes.

3.3.1.1.1 Terminologisation

Terminologisation is the metaphorical use of general words in LSP. Terms are created by terminologisation according to similarities in form, function, or position (Sager 1997:29). A term thus formed has features in common with its original word, which allows people to infer the meaning of a term. In the expression *Geldwäscherei* 'money laundering' for example, the activity of washing implies that one has to do with 'dirty' or illegal money.

Terms created by terminologisation are composed of general stem words. For a human, this makes it is easier to infer their meaning. For ATR, however, this type of term formation is less interesting. In fact, metaphorical use of general words provides no distinguishable linguistic features but needs a human's cognitive abilities to understand the meaning.

3.3.1.1.2 Simile

Terms formed by simile exhibit an analogy with existing designations (Sager 1997:28). Expressions like -*type* (*transaction-type code*), -*style* (*American-style option*), or -*like* (*bank-like finance company*) are used to determine new concepts. Term formation by means of simile can be exploited for automatic term retrieval. As will be shown later in section 5.5, in so-called contextual approaches such means of comparison are used for the extraction of terms.

3.3.1.2 The Modification of Existing Resources

In German, the most important method for designating new concepts is to modify existing terms by compounding. In English or Romance languages phrasal terms are created instead of compounding. Derivation or affixation is also a productive means of forming new terms from existing resources. Further methods are compression, i.e. term formation by means of shortening of expressions, and the conversion of terms from one part of speech (PoS) category into another without change of form.

3.3.1.2.1 Compounding

The meaningful elements of a word, the morphemes, are classified into free and bound morphemes (Finegan 1994: 83). Free morphemes can stand alone as words, bound morphemes can only be part of words. Compounding is the combination of two or more free morphemes into a "new syntagmatic unit with a new meaning independent of the constituent parts" (Sager 1997:34). Noun compounding is the most important method for the formation of new terms, as will be shown in section 6.4.3 by the analysis of the terminology data base of CREDIT SUISSE. Combinations of nouns with verbs, adjectives, or prepositions are also possible. In English, compounding leads to the formation of multi-word terms. In German free morphemes, also called lemmas, are mostly connected to form single-word terms. The examples given above indicate that German banking terminology makes frequent use of compounding.

In ordinary cases, the number of elements constituting a compound term is limited. Blank (1998:109) quotes a research by Ischreyt, who analysed 11'500 terms and found that two-element compounds were identified in 50 % of the terms and three-element compounds in 30%. In reality, there is no limit to the number of lemmas connected to form a term. However, compounds with more than three lemmas tend to have a grotesque character, as is the case with this example taken from the domain of engineering (Hoffmann 1985:122):
Ultrakurzwellenüberreichweitenfernsehrichtfunkverbindung.

Compounding is the most important method of term formation for automatic term retrieval. As will be shown in section 6.2.3.2, in German banking terminology most compound terms are formed by a limited number of free morphemes. Once recurring free morphemes are identified, they can be employed for the identification of unknown terms.

3.3.1.2.2 Phrasal Terms

German LSP theoreticians unanimously define a phrasal term as a term which is composed of two or more words separated by a space. For Reinhardt et al. (1992:18-19) phrasal terms (in their book called *Wortgruppenlexeme*) are a special case of compounding. While compounding strives for economical expressions, phrasal terms have the advantage of avoiding complicated compound words (so-called *Bandwurmwörter*). In Arntz/Picht (1995:121-122) phrasal terms are called multi-word terms (*Mehrwortbenennung*). They also consider multi-word terms as a special case of compounding, particularly used in Romance languages or in English.

In the previous section we have seen that in English compounding leads to the formation of multi-word terms. The way in which English multi-word terms formed by compounding differ from phrasal terms would need further discussion. From Sager's (1997:36-37) explanations it seems that phrasal terms include prepositions (*bridges with pin-joined members*) and adverbials (*artificially aged then cold worked*). However, among Sager's examples there are also syntactic patterns like (adjective-noun)-noun[1] which are once classified as compounds (*bending moment diagram*) and once as phrasal terms (*general purpose computer*).

For purposes of automatic term retrieval it is not necessary to disclose the exact difference between compounds and phrasal terms. The fact, however, that there are repeated syntactic structures in term formation provides an important means for an automatic identification of terms. The syntactic structure of multi-word terms is a reliable ATR methods for languages with a poor inflectional system and thus a static word order like English. In German this method is less effective.

3.3.1.2.3 Derivation

Derivation or affixation is "the addition of suffixes and/or prefixes for determining new concepts" (Sager 1997:30). By means of a suffix a word is transformed into another word with a meaning related to the original one but belonging to a different lexical class (*verbinden* 'to join' vs. *Verbindung* 'connection', *sichern* 'to secure' vs. *Sicherheit* 'security', *persönlich* 'personal' vs. *Persönlichkeit* 'personality', etc.). In some cases the suffixes substitute free morphemes in order to avoid awkward noun compounds. The suffix 'er' in *Rechner* 'computer' and *Bohrer* 'drill' is a simplification of the compound constructions *Rechenmaschine*, *Rechenanlage* and *Bohrwerkzeug* (Reinhardt et al. 1992:22).

By attaching prefixes the meaning of a word is changed but not its lexical class (*Überschuss* 'surplus', *Zwischenraum* 'blank space', *Vorsorgebereich* 'pension funds'). In certain German constructions it is hardly possible to define the boundary between prefixes and a free morpheme. This is the case when the meaning of a lemma with an affix differs from its meaning as an independent word. Reinhardt et al. (1992:22) illustrate this case with the example *aufbringen* ('to find' or 'to raise'). In combination with the prefix *auf* the meaning of

[1] the structure is indicated by brackets

the verb *bringen* 'to bring' is utterly changed. Prefixes of this kind are referred to as half-prefixes (in German *Halbpräfixe*).

The analysis of terms according to derivation proved to be a useful method for automatic term extraction. Studies on German terminology from the domain of automobile engineering showed that many terms contain typical prefixes (*mega, multi, mikro, radial, semi, trans, ultra, zentral,* etc.) and suffixes (*dicht, gramm, graph, werk, wesen, zeug,* etc.) which can be used as a filter for the retrieval of domain-specific terminology from text corpora (Heid et al. 1996).

3.3.1.2.4 Conversion

Conversion is the change of a word class without morphological alteration of the word (Sager 1997: 37). In terminology, conversion is used to create the noun concept associated with verbs (*schreiben* 'to write' vs. *das Schreiben* 'letter'), with adjectives (*blau* 'blue' vs. *das Blau*), or with adjectival participles (*vorsitzend* 'presiding' vs. *der Vorsitzende* 'chairman') (Arntz/Picht 1995:123). In German, conversion from verbs into nouns, so-called nominalization (section 3.4.1.3), is a frequent feature of LSP.

Terms created by conversion are useful for ATR if they are compounds containing domain-specific lemmas which can be exploited by a computer program.

3.3.1.2.5 Compression

Compression of the sort of regular abbreviations (*km*, *i.e., etc.*) is common in general language but of minor importance in terminology. A further type of compression consists in shortenings by omitting parts of words, as in the German words *Labor* (from *Laboratorium* 'laboratory') or *Band* (from *Tonband* 'tape recorder'). The most common type of compression and the most relevant for terminology is the formation of acronyms. These are created by joining the initials of a phrase. Examples from the banking domain are *SMI* (Swiss Market Index), *SOFFEX* (Swiss Options and Financial Futures Exchange), or *LIMEAN* (London Interbank Mean Rate). Names of institutions are frequently rendered as acronyms, as in *CSG* (Credit Suisse Group), *SBVg* (Schweizerische Bankiervereinigung), or *BIZ* (Bank für Internationalen Zahlungsausgleich). In most cases, acronyms form chains of capitalised letters, a feature which can be utilised in ATR.

3.3.1.3 The Creation of New Linguistic Entities (Neologisms)

Neologisms are either borrowed from other language communities or they are loan translations, i.e. word by word translations from a target language into a source language. In rare cases, neologisms are totally new creations.

3.3.1.3.1 Borrowing and Loan Translation

Today, all languages are influenced by English and widen their vocabulary by means of borrowing English words. In German speaking countries borrowing English words is a favoured way of enlarging the vocabulary. The fact that English is today's trendy language reflects the scientific, economic, and cultural dominance of Anglo-Saxon countries. In working environments, professionals within one subject field automatically adopt the English terminology, and prevent the creation of equivalent terms in their mother tongues. In German banking terminology there is a large number of English expressions which do not have a German equivalent (for example *Holding, Leasing, Check, Manager, Marketing*).

Loan translation is a word for word substitution of the lexical components of a source language term in the target language (Arntz/Picht 1995:124). Depending on the lexical rules of the target language, loan translation might necessitate some syntactic re-ordering of the compound elements. Common examples of loan translations are found in computational linguistics, where *maschinelle Übersetzung* originates from *machine translation maschinenunterstützte Übersetzung* from *computer aided translation*, or *automatische Termextraktion* from *automatic term extraction*.

The adoption of foreign vocabulary is a natural development of language. This evolution, though, has not always been favourably viewed, as Wüster points out:

> "Fremde Wörter werden von der nationalen Sprachnormung bewusst abgelehnt. Gegen die fremden Wörter spricht nicht nur ihre mangelnde Sprachreinheit (ein Gesichtspunkt der Sprachschönheit), sondern auch ihre geringe Merkhilfe." (quoted from Reinhard 1992:31)

Today, the interchange between different languages is widely accepted. Even authoritative institutions like the ISO consider the "internationalisation" of a language as a means of easing communication:

> "Der Wert der internationalen äusseren Formen besteht darin, dass sie unmittelbar die internationale Verständigung erleichtern. Darum kann bei der Wahl zwischen zwei gleichwertigen Synonymen in einer Sprache für denselben Begriff dasjenige vorgezogen

> werden, das in derselben Form auch in andere Sprachen übernommen worden ist." (quoted from Reinhard 1992:31)

In translation-oriented terminology borrowing represents a welcome way of avoiding time-consuming inquiries about the correct equivalent of a foreign term. For a computer program designed for one language only, borrowed terms are difficult to spot. As will be shown in section 6.2.2.3, most foreign terms are not stored in the lexicons used during the parsing process. Loaned translations are easier to retrieve. They can be detected automatically if they are identified by the morphological analyser and are composed of domain-specific lemmas (section 6.3.2).

3.3.1.3.2 New Creations

New creations are a relatively rare kind of word formation (Sager 1997: 38). In Britain and the US, the governments of Thatcher and Reagan have contributed to many English words to the field of economy (Gläser 1988:172-178). Words like *Reagonomics* and *Thatcherism* reflect the economic policies advocated by the two politicians. Financial measure, wage and working policy of this time were described by new words like *deregulation, rate-capping,* or *flextime*. New creations in connection with unemployment were *mobile worker, migrant work-seeker, make-work schemes, labour-market rigidity, anti-welfarism.* Word combinations like *stagflation* and *boomflation* refer to the simultaneous occurrence of *stagnation* and *booming* on the one hand, and *inflation* on the other hand[2].

In German, new inventions are even rarer since it tends to borrow terms, particularly from English. The problem of automatically identifying new creations are the ones mentioned in the preceding section: the retrieval of candidate terms is dependent on the size of the lexicon used in the parsing process, in which new creations might not yet be included. An extended discussion of the obstacles encountered during parsing will follow in section 6.3.

3.3.2 Some Statistics

3.3.2.1 Frequency of Word Classes as a Means of LSP Identification

Studies on German and English LSP in the fields of medicine, economy, and technology showed that there is a major difference between the number of word class occurrences in specialised and in general texts (Schefe 1975, Sager et al. 1980, Ahmad et al. 1994). These

[2] All examples are quoted from Gläser 1988: 173

works demonstrated that in LSP the use of nouns and adjectives is more than twice as frequent than in general texts. As a result of strong nominalization tendencies (section 3.4.1.3), verbs are much less frequent in LSP. Finite verb forms are more than twice as numerous in general texts than in special texts. Closed word classes like adverbs, pronouns and particles are almost inexistent in LSP. These studies are a confirmation of the strongly nominal style of LSP.

3.3.2.2 Frequency of Words as a Means of Subject Identification

Sager et al. (1980:236) refer to studies on the most frequent words in specialised dictionaries done by Hoffman in the seventies. The results of these researches show that the most frequent words found in the dictionaries allow a clear identification of the subject area. The five most frequent words in a medical dictionary for example, are *patient, blood, disease, cell,* and *fibre*. The most common words in a dictionary on electronics are *current, temperature, voltage, cathode,* and *energy*. The frequent use of a limited number of words is thus a characteristic of domain-specific terminology.

Hoffmann's study provides the foundation for ATR methods based on morphological analysis like the one implemented in this work. They are based on the assumption that, if whole words help identify a specific subject field, then parts of words, i.e. lemmas, might serve for the same purpose. By reducing words to their morphological parts, morphemes can be identified which are typical of a subject field and thus are crucial in forming its terminology. This brief excursion anticipates the method implemented in this thesis.

The next section explores the syntactic organisation of words in LSP.

3.4 Syntactic Level

At the sentence level, LSP has many characteristics differentiating it from general language. In terminology, though, the focus is not on the whole sentence but on its parts. Today researchers agree that the constituents of LSP sentences consist mainly of noun phrases (NPs) (Desmet/Boutayeb 1994, Roelcke 1999, Sager 1990, Arntz/Picht 1995, Heid et. al 1996, Voutilainen 1993, Justeson/Katz 1995). It will later be discussed that this LSP attribute alone is not sufficient for a successful term extraction. In fact, the nominal structure of LSP constituents can only be exploited for ATR purposes if they differ from general language structures. This is not always the case, and for this reason ATR programs restricted to NP extraction show a very poor performance.

Although not all syntactic features of LSP are helpful for ATR, they will be shortly discussed for the sake of completeness.

3.4.1 Syntactic constructions

In German, the syntactic characteristics of LSP concern the preference for certain sentence types, the tendency to use attributes and nominalization. These constructions increase the complexity of language while at the same time they ensure the transparency of statements (Roelcke 1999:80-81).

3.4.1.1 Sentence Types

Roelcke (1999) states that in German LSP, the most important sentence types are declarative sentences (80). Their function is to increase the clarity of a statement, which is what a strongly informative language like LSP intends to do. This sentence type also reflects the strife for anonymity and achieves a reduction of emotive elements. Declarative sentences often contain embedded relative clauses, which is a way of increasing the transparency of a statement (81). Imperative sentences are frequent in LSP, commonly encountered in texts like operating instructions (80) *(Zum Starten des Motors drehen Sie den Zündschlüssel nach rechts!)*[3].

Further constructions are formed by joining various kinds of clauses to declarative and imperative sentences. Conditional clauses are typical of LSP, either with conjunction (*Wenn Sie an dem Lastschrifteneinzugsverfaren teilnehmen möchten, füllen Sie bitte beiliegendes Formular aus*) or without conjunction (*Ist das Anti-Blockiersystem in Betrieb, erscheint die Leuchtanzeige Nr 7*). Final clauses with conjunctions (*Damit die Gleichung gelöst werden kann, muss erst der gemeinsame Nenner ermittelt werden*) or without conjunctions (*Zur Lösung der Gleichung ist zunächst der gemeinsame Nenner zu ermitteln*) are also common.

3.4.1.2 Noun Phrase Attributions

Characteristic attribute constructions in German NPs are adjective attributes (*staatliche Vorsorge* 'state provision')[4], participial attributes (*abnehmendes Kapital* 'decreasing capital'), prepositional attributes (*Abschreibung mit Zinseszinsen* 'effective interest method'), or genitive attributes (*Spaltung des Nennwertes* 'stock split').

[3] All examples in this section are quoted from Roelcke (1999: 80-84)

[4] All examples in this section are based on CS-term

The possibilities of embedding NP attributes and thus condensing a text (section 3.4.2) are basically unlimited. The only restriction is the comprehension ability of the language user. For terminology purposes, chains of NPs are relevant only for contextual retrieval. A terminologist is mainly interested in the individual concepts of a linguistic expression and will thus focus on its smaller constituent parts.

3.4.1.3 Nominalization

Nominalization is the process of forming a noun from other word classes and it affects particularly verbs (*Bevollmächtigter* 'authorised person' from *bevollmächtigen* 'to authorise', *Absicherung* 'hedge' from *absichern* 'to hedge', *Anlage* 'investment' from *anlegen* 'invest', etc.). At a lexical level, nominalization is a method of term formation related to conversion and derivation. Nominalization is a way of "condensing" language (section 3.4.2). Its function in LSP is to abstract from the persons and/or things referred to in texts (Roelcke 1999:81), as shown in the following examples: *Die Universität bescheinigt den Studierenden, dass <u>sie immatrikuliert sind</u>* vs. *Die Universität stellt den Studierenden eine Bescheinigung aus, nach der <u>ihre Immatrikulation besteht</u>* (Roelcke 1999:81).

3.4.2 Condensation of Language

The syntactic constructions discussed above lead to an increase in sentence complexity. The predominance of NPs, the use of attributes and embedded clauses, the tendency of nominalization, together with the scarcity of function words make LSP a concise language. Entire sections of LSP texts often consist almost exclusively of nominal groups expanded by modifying NPs or by relative clauses, which increases sentence length. All these phenomena aim at clarifying the concepts while at the same time they add to the complexity of the sentence. Hahn (1983:117) describes this feature of LSP as a "condensation" of language. The general public will find such constructions rather impenetrable and tiring to read. Trained professionals, however, will not bother about their syntactic complexity but will appreciate the conceptual clarity emanating from them.

3.4.3 Informational Purpose of NPs

We have seen that NPs, either single-word or multi-word ones, are the most important constituents in LSP. The reason indicated by Sager et al. (1980) is that LSP is concerned less with actions and events than with the presentation of facts and theories, with the description of experiments and processes, or with the evaluation and explanation of results (204). The most appropriate grammatical structure for performing these functions is the nominal group, which therefore plays a much more important part in LSP than other syntactic phrases. Verbs for example are considered to be too vague for the exact definitions required in LSP, whereas nouns can be qualified by precise measurements and values (205). In the section on term formation we have seen that nouns can also be more easily modified than verbs, which permits noun groups to carry a greater information content than verbal groups. This characteristic results in their much higher frequency in special texts than in general texts.

3.4.4 Significance of Syntactic Characteristics for ATR

There are no limitations to the possible combinations of constituents within sentence structures, which is however no problem from an ATR perspective. In fact, since Chomsky's introduction of generative grammar it is well-known that it is possible to formalise an infinite set of sentences with a finite set of rules. Grammar theories like Phrase Structure (PS), Generalised Phrase Structure Grammar (GPSG), or Head-Driven Phrase Structure Grammar (HPSG) are good formalisms for a generalisation of language, and they are frequently employed in computational linguistics for parsing purposes.

In German, the problems of identifying terms according to syntactic criteria are not created by the syntactic complexity of candidate terms but rather by their lack of a distinctive structure. In approaches to German terminology a term is characterised almost exclusively as a single noun (compound). Contrary to English or the Romance languages, multi-word terms play a minor role in German terminology. Since in German LSP single noun terms are predominant, they do not provide any syntactic features distinguishing them from non-terms. As a matter of fact, identifying candidate terms only according to structural criteria does not yield satisfactory results. Later in this thesis we will see that the extraction of German candidate terms can not be very successful if performed with syntactic methods only.

This concludes the discussion of the lexical and syntactic characteristics of LSP. It is needless to say that LSP features many more differences to general language than the ones presented here. Word order, elliptical constructions, passive constructions, infinitive constructions, enumeration, and more linguistic characteristics are important in LSP. However, they seem to be of minor benefit to ATR, which is why they have not been touched upon in this overview.

Before the actual term extraction is described, there will be an introduction to the language of banking, the LSP field in which term extraction was performed in this work. First banking language is situated within a typological model of the language of economy. Then, the historical development of the language of economy and of its subset, the language of banking, will be traced.

4 Language of Economy

4.1 A Typological Model

Given the high priority of money in our way of life, the language of economy strongly intersects with general language. This makes it extremely difficult to delimit the concept of this special language. If we took into consideration all human activities which are connected with economy, we would have a never-ending diagram which would need to include occupations like selling, shopping, book-keeping and many more. In our daily life economic matters have often the highest priority. Since in everyday life we are used to talk in terms of 'shareholder value', 'market value', 'debt capital', 'investment funds', or 'fixed assets', specialised financial terms go unnoticed. Because of our society's strong interest in economy, the actual degree of specialisation of economic language is weakened.

Despite the infiltration of financial vocabulary into general language, the language of economy is a specialised language, and its features have been studied by many scholars (section 4.2). Attempts at establishing a typological model of the language of economy are based on Nomoto's definition of LSP (quoted in section 3.1). Accordingly, the special nature of the language of economy is distinguished by the subject field and by the type of user. Cabré (1999:64) refers to a study by Picht and Draskau who propose a typological characterisation by dividing LSP into different levels of specialisation. The highest level corresponds to the communication between highly qualified experts and the lowest one to general purpose information meant for the layman.

This thesis deals with the language of economy used in the banking sector, and it can be assumed that most users in this sector are subject specialists. Thus, models focusing on the subject field like the ones suggested by Hundt (1995) seem to be more appropriate. Hundt proposes two typological models for classifying the language of economy. The first model is based on the so called 'three-sector-hypothesis' (*Drei-Sektoren-Hypothese*), which categorises the entire economy in the primary, secondary, and tertiary sector. The second model is a classification of sciences related to economy like business data processing, economic geography, economic law, etc. Hundt positions banking language in the first model within the tertiary sector. He refers to the languages belonging to this model as *Institutionelle Wirtschaftssprachen* 'institutional languages of economy'. Figure 5 is a simplified version of Hundt's model.

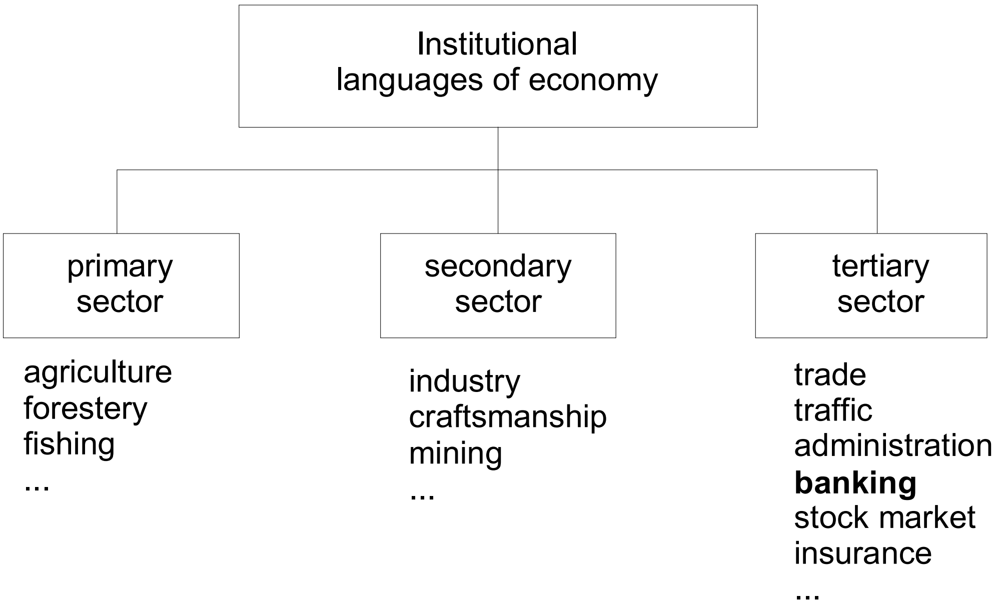


Figure 5: Typology of language of economy according to institutions (Hundt 1995:67)

In figure 5, the language of banking is a subset of the languages of economy of the tertiary sector. Among the languages listed in this sector are also the languages of stock market and insurance. CS-term includes many terms belonging to these subject fields. For the analysis of this data base, the terms having insurance as subject field were not examined. Terms belonging to the field of stock market, though, were considered as part of the banking vocabulary. As also Fluck (1996:62) recognises, stock market is closely related with banking business and provides many lexical units belonging to the banking vocabulary.

4.2 A Historical Account

Hundt (1995) remarks that LSP of the economic sector is still a "terra incognita" (11). Early studies on this special language concentrated on diachronic analyses. Hundt refers to Stieda's researches in 1894 on the Hanseatic language, on Fehr's dissertation in 1909 of the history of English trading language, or on the etymological development of words originating from Greek, Latin, and German done by Schrader in 1886 (17). An analysis at the conceptual level of economic language was done by Siebenschein in 1936, who looked closely at concepts like *kaufen* 'to buy', *verkaufen* 'to sell', *Handel* 'trade', *Geld* 'money', *Kapital* 'capital', *Zins* 'interest rate', *Vermögen* 'fortune', or *Schuld* 'debt'. However, he still focused on the etymological and historic context (21). Around 1936, Vancura studied the style of the language of economy by taking into account specific documentation like letters, contracts, economic articles and news, or advertisements (22).

In the 20th century the language of economy was frequently viewed in respect to political ideologies. Marxist groups saw economic texts as a means of manipulating people and abusing power, which is why they aimed at exposing such texts. Particularly the use of metaphors in economic texts was interpreted as an intended manipulation, for example by Schmitt, who pointed out:

> "Der Metapherngebrauch entspricht den kommunikativen Intentionen von Autoren, die ihren Lesern eine bestimmte Sicht der Wirtschaftsentwicklung vermitteln möchten." (Schmitt 1988:124 [quoted from Hundt 1995:40])

Next to metaphors, passive constructions were also interpreted as a strategy of the ruling class to disguise their deeds:

> "Die Urheber der Kursbewegungen, die Besitzenden, die Spekulanten werden ebensowenig genannt wie die Hintergründe von Aktientransaktionen und die gesellschaftlichen Verhältnisse, die sich darin manifestieren." (Arnold 1973:107 [quoted from Hundt 1995:33])

On a linguistic ground, the important development for the language of economy in this period is that first efforts are made to systematise its terminology. In the early forties works on standardization of economic terminology were done by Kutzelnigg and Koppelmann in "Terminologie der Warenkategorien" (24). They adopted Wüster's methodological guidelines regarding the conceptual organisation of terms by grouping them into a hyponymic relationship, by defining semantic features, and by structuring conceptual systems in different economic fields (24-25). Up to the seventies the lexicon and syntax of financial texts were studied with the help of text corpora by Schefe, De Cort and Hessmann, Duhme, Kvam, Piirainen, and Airismäki (26-28). The works focused on the frequency of certain word forms and on the sentence structure of financial texts (28).

4.3 Banking Language

An etymological reflection on today's banking expressions in German speaking countries throws an interesting light upon linguistic trends and customs. The terminology of banking reflects past and present influences of foreign countries on the German language community. In early centuries German financial terms were borrowed from Italian, where the roots of modern economic forms are. Words like *Giro* 'giro', *Konto* 'account', *Skonto* 'discount', *Valuta* 'foreign currency', *Saldo* 'balance', *Devisen* 'foreign exchange' are today approved German words and not considered as foreign vocabulary. Since the later 20th century there is the tendency to borrowing from English (see also section 3.4.1.3).

The internationalisation of language is evident in the terminology of CREDIT SUISSE. English expressions are sometimes preferred even if there is a German equivalent, as in the case of *Discount* instead of *Preisnachlass* or *Perpetual Bond* instead of *ewige Anleihe*. A further striking observation is the frequent occurrence of mixed forms like in the multi-word term *einen Margin Call erlassen* 'to generate a margin call', *Financial Futures-Limite* 'financial futures limit', *Währungs-Future* 'foreign exchange future', *Stockdividende* 'stock dividend', or *Value at risk des Vortages* 'previous day's value at risk'. In Switzerland, borrowing from French is also frequent, as in *Affaire liée/Ordre lié* 'contingent order', *Verkauf à découvert* 'short sale', *System à la criée* 'open outcry trading', *Décharge* 'discharge'. The French influence on the terminology of a Swiss bank is understandable since French is an official language of Switzerland.

4.4 ATR: New Perspectives in Banking Terminology Work

The increasing international co-operation in the banking sector is connected with an increasing need for terminology. An effective ATR program would be a necessary tool for terminologists and make terminology work much more efficient. Today, programs are available to assist the terminologist in the task of term identification, which, however, perform poorly. Providers of ATR programs promote their software as able to automatically build up multilingual term bases with a minimal intervention by the terminologist. In reality, the number of extracted words and phrases which are found to be actual terms is still very low. The terminologist is provided with a list of candidate terms among which many are useless and which needs a large amount of time for correction.

The reasons for the poor performance of ATR programs are manifold and will be discussed in the second part of this thesis. One point of failure can be anticipated: the strife for domain independency. A frequent argument of ATR providers is that one product can be used in any domain. Given the syntactic and morphological variation of different special languages, these diversities can by no means be covered by one single tool. For humans the notion of domain is crucial when identifying terms, and a computer program can also exploit domain-specific features of LSP to improve its performance.

A characterisation of LSP as done in this thesis gives a better understanding of the linguistic form of terms. The main objective of this investigation was to apply this theoretical foundation to the description of banking terminology. With the help of a term base, the next step will be to formalise and quantify all the syntactic and morphological characteristics of terms. In a second step these characteristics will help to program a tool for the automatic extraction of candidate terms from the domain of banking.

Part 2: Automatic Term Retrieval

In the first part of this paper the difficulties of transferring conceptual knowledge on computers have been widely discussed. In this second part of the work the focus is on procedures of automatic term retrieval (ATR). First, an overview of the ATR methods is given. Then the implementation of a self-developed method for the retrieval of German terms from the domain of banking is presented.

5 Related Work on ATR

In computational linguistics the dominant trends in the field of ATR regard the application of statistics and of linguistic methods. Statistically oriented approaches employ the frequency of words in a text to generate term lists. Linguistic approaches exploit morphological and syntactic characteristics of terms as opposed to non-terms. In rare cases, a linguistic approach is based on conceptual operations and on contextual information. Some approaches consider a combination of statistical and linguistic models for the extraction of terms. Basically, the linguistic ATR researches are similar to a large extent with each other. Their main difference lies in the computational infrastructure used.

5.1 Statistical Approach: Ahmad, Davies, Fulford, and Rogers

A statistical approach exploits the numerical information about words in texts. The assumption is based on a principle commonly applied in information retrieval (L'Homme et al. 1996:298), which maintains that if certain words occur more frequently than expected they carry significant information. A word which is mentioned often in a domain specific text describes a concept that is important in the domain and is thus likely to be a candidate term. Statistical analyses are performed on large corpora in which the dominance and suppression of certain word forms is easier to identify. The general rule is that a high number of open word classes is an indication of a domain-specific text; a high number of closed word classes is a typical feature of general language texts.

Studies on the relative frequency of open vs. closed word classes in corpora of LSP and general language have been performed by Ahmad et al. (1994). Looking at the frequencies of all words, they established that the ranking of the most common words is approximately the same. No matter which corpus is examined, the most frequent PoS are always prepositions, determiners, and pronouns. In LSP corpora though, in these closed classes some specialised words appear which, in a manual evaluation, are most often chosen as terms. Ahmad et al. conclude that open class words rank high in the frequency list if they describe important concepts. This method exploiting the frequency of word classes is used to match potential single word terms.

A method for the retrieval of multi-word terms has also been implemented by means of this approach. The output of the high ranking open class words is used to generate a list of collocations. First, the program searches the corpus for collocations in which the high-ranking open class word is the first word in the pair; the second word is not specified. Then the program searches for collocations in which the high-ranking word is the second in the pair; the first word is not specified. Although no linguistic information is used, the retrieval of correct collocations is frequent. The authors report very few cases in which the collocations are either not grammatical or not terminological. However, they do not mention the number of terminological collocations which have been missed by the program.

Ahmad et al. have integrated this statistical technique in *System Quirk*, a program for examining texts, extracting terms, and managing terminological work. The advantage of this method is its language independence. Before applying a statistical algorithm to other languages only the frequency threshold (i.e. the number representing the frequency) must be adapted. It is generally agreed that the disadvantage of a statistical approach is the inability of finding terms which occur with a low frequency in a text (L'Homme et al. 1996:299). Further, a statistical approach is acceptable for large corpora only. In tests with small corpora, a frequency threshold of two retrieves hardly any terms. This was also the case in the evaluation done for this thesis.

5.2 Linguistic approaches

Linguistic term extraction analyses morphological features and syntactic patterns of terms as opposed to non-terms. As shown in section 3.5, studies on LSP agree that terms are mainly NPs. Many ATR researchers promote NP extraction for the identification of terms. In the works presented in the next sections, the usefulness of syntactic criteria for term identification is underlined. However, morphological features of terms are rarely relied on.

5.2.1 Voutilainen and Arppe

In this research a program for the recognition of English NPs, called *NPtool*, is used for term retrieval. NPtool is based on components similar to the ones employed for the NP extraction in this work, these being two-level morphology, a NP grammar, and a parser (Voutilainen 1993; these modules will be explained in section 6.3.1 of this paper). From an input sentence NPtool extracts all NPs at their maximum length. Further, the program identifies all acceptable subsets of each maximum NP. The extraction of all possible permutations of NPs multiplies the number of candidate terms. For example, starting from the maximum length NP

exact form of the correct theory of quantum gravity

the following constituent NPs are retrieved (Arppe 1995):

exact form of the correct theory
exact form
form
form of the correct theory of quantum gravity
form of the correct theory
correct theory of quantum gravity
correct theory
theory of quantum gravity
theory
quantum gravity
gravity

This strategy leads to a total of twelve candidate terms only one of which is a term, *quantum gravity*. The disadvantage of this method is undoubtedly the additional work a terminologist is faced with when having to discard the incorrect candidate terms. As practical experience tells and the analysis on (German) terms performed for this work confirms (see section 6.2), long NPs are rare in terminology. The syntax of terminological NPs is rather simple, and there is little sense in presenting a terminologist with NP chains and all their possible sub-structures. On the one hand Arppe considers it easier to delete wrong candidates than to search for correct terms which were not extracted. On the other hand he admits that identifying all possible NPs is more useful for translation than for terminology, since the translation of a complex NP can be different from the one of its constituent NPs.

A further utility of NPtool is the integration of a database containing generic modifiers like *old, young, new, beautiful, soft, hard, long, short*, etc., also called *stop-list*. NPs with generic modifiers are removed from the list of extracted NPs in order to decrease the number of candidate terms. NPtool provides the extracted NPs with their frequency and arranges them into groups according to their grammatical heads.

5.2.2 Bourigault

Bourigault's ATR method is also based on parsing techniques for the recognition of NPs. The extraction of NPs is carried out in four consecutive steps (adopted from Aussenac-Gilles et al. 1995):

Tagging PoS are assigned to words.

Splitting PoS like verbs, pronouns, conjunctions, and determiners are used to identify the boundaries between NPs; the result is a set of NPs of maximum length.

Parsing The retrieved maximum NPs are parsed for the identification of smaller NP constituents. Every NP subset is divided into two constituents: a constituent in head position representing a super-ordinate concept and a constituent in expansion position functioning as modifier.

Structuring Each of these constituents is linked with other terms which use the same constituent words.

The final result is a terminological hypertext network composed of nodes and links (see also section 5.5). The nodes are of two types: there are term nodes which are associated with other candidate terms and there are text nodes associated with textual units of the original text.

The advantage of this approach is that it allows a hierarchical ordering of terms, an important procedure in terminology work. The text nodes are useful for the retrieval of context. This ATR-method is at the core of *LEXTER* (Logiciel d'Extraction de Terminologie), a system for the retrieval of technical terms from French and English texts.

5.2.3 Heid, Jauβ, Krüger and Hohmann

Heid et al. (1996) focus on the extraction of single and multi-word terms. An additional investigation includes the morphological components of candidate terms, which makes this approach a progression towards the ones discussed till now. The extraction procedure is based on two steps:

- linguistic preanalysis and annotation of text material
- corpus query

The analyses of the raw corpus in step one includes the identification of word and sentence boundaries (tokenising), the identification of grammatical categories and morphological features (morphosyntactic analysis), the disambiguation of the morphosyntactic analysis (part-of-speech tagging), and the identification of lemma candidates (lemmatisation). After the corpus has been annotated with PoS and lemmatised, the corpus queries of step two aim at extracting characteristic morphosyntactic elements. The queries consist of three types of regular expressions: on characters (to identify abbreviations), on morphemes (to identify single-word terms and their appearance in multi-word terms), and on PoS (to identify multi-word terms). The languages worked on are German, French, and English. The corpus is taken from automobile engineering.

The queries on single-word terms (and on their appearance in multi-word terms) rely on the assumption that many terms in the domain of automobile engineering contain typical affixes. For the purpose of term extraction Heid et al. search for recurring morphemes across the lemmatised text. Next to extracting single nouns the system also retrieves adjectives and verbs. The best results are achieved in the identification of nouns. Queries for adjectives and verbs produce a considerable amount of noise, apparently because these word forms have few affixes typical of specialised language. The authors, however, give no examples of affixes which retrieve terminological adjectives and verbs but also lead to noise.

The queries on multi-word terms identify typical noun-verb collocations, noun groups, and prepositional groups. It is reported that in many cases the retrieved multi-word candidate terms are not well-formed. Ungrammatical constructs are particularly frequent among complex noun groups and prepositional groups. The authors point out that parsing techniques would alleviate this problem.

The paper also includes a report on a comparison of the extraction results with Ahmad's statistical approach. This comparison shows that the instances retrieved by Heid's system are all contained in the output of the statistical approach. However, the statistical approach produces much more noise.

The exploitation of morphological features is an important procedure particularly for the extraction of German terms. In the implementation of my own method for ATR I have relied a great deal on this approach as far as the search for domain-specific morphemes is concerned.

5.2.4 Church and Dagan

Like the preceding approaches, the contribution by Church and Dagan (1994) shows that (English) terms are NPs with a limited structural variance. Single-word candidates are retrieved by means of a list of all words that occur in the text and do not appear in a selfdefined stop-list with general words. Multi-word terms are extracted from the PoS-tagged text with the help of regular expressions defined by Justeson and Katz (1995) (section 5.3.1). Finally, the list of extracted terms is sorted according to the frequency of their head words. The authors specify that frequent head words are likely to generate a number of terms. However they do not apply any frequency constraint. The program also provides access to concordance lines useful for the identification of correct candidates or for spotting missing terms.

Church and Dagan employed this procedure for the development of the ATR program *Termight*. In addition to the monolingual task just presented, Termight has a bilingual component which makes use of word alignment to extract the translation of the candidate terms. Termight is used at AT&T Business Translation Services, apparently with great success. Terminologists at AT&T are reported to construct terminology lists of 150-200 terms within one hour. No further information is given by the authors about the performance of this system. One can assume, though, that Termight has a high recall but a low precision value, as it is common for tools relying mainly on syntactic constraints.

5.3 Combination of Linguistic and Statistical Methods

In order to enhance the chances of extracting good candidate terms there is the possibility of applying statistics combined with linguistic methods. Yet, the stress in these studies is on the linguistic element. Statistics are limited and employed sparingly.

5.3.1 Justeson and Katz

From observations of technical terms in English dictionaries Justeson and Katz (1995) conclude that these consist mostly of NPs containing adjectives, nouns, and occasionally prepositions. Further, the authors observe that terminological NPs are more susceptible to repetition than non-terminological ones. Based on these findings they present a domain-independent algorithm for the identification of English terminology. The algorithm requires the satisfaction of two constraints:

- Frequency: candidate NPs must have a frequency of 2 or more in the text.
- Grammatical structure (retrieved by means of regular expressions): a candidate term is a multi-word noun phrase; and it is either a string of nouns and/or adjectives, ending in a noun, or it consists of two such strings, separated by a single preposition.

For the implementation a lexical database of about 100'000 entries is used. This database is connected with a morphological analyser. In the retrieval process each word in a string being tested passes through the lexicon and morphological analyser. There, if necessary, it is segmented in morphemes and assigned a PoS. The annotated word strings are accepted as candidate terms if they meet the requirements defined in the regular expression and if the frequency constraint is fulfilled. Interestingly, while in the previous approach (section 5.2.3) parsing was considered a better method for the retrieval of multi-word terms, Justeson and Katz declare that they did not use any parsing in order to keep the error rate down.

The evaluation presented by the authors reveals a nearly perfect recall value, however, the precision value is rather poor. As already mentioned, the dominance of recall over precision is typical of term extraction methods which rely on syntactic patterns only. The authors stress that an increment of the frequency constraint would improve the precision rate. However, experience tells that an improvement of the quality of the candidate terms correlates with a poorer recall rate and thus a loss of correct terms.

Justeson and Katz' algorithm has been implemented for a program called *TERM* which is used in IBM translation centres. This method is acknowledged to be "[one] of the best defined procedures for the identification and extraction of technical terms" (Boguraev and Kennedy 1998:18).

5.3.2 Heid

The focus in Heid (1999) is on the acquisition of terminologically relevant collocations in German technical texts. The type of collocations searched for are noun-noun (i.e. noun+preposition+noun and noun+article+genitive noun), noun-verb, and adjective-noun. The computer infrastructure adopted is the same as in Heid et al. (1996), namely a tokeniser, a PoS tagger, and a lemmatiser.

At the outset of this work is the analysis of collocations made by Marie-Claude L'Homme[5], in which the following characteristics of terminological collocations are identified:

- In adjective-noun and noun-verb collocations the nominal component is a term. In noun-noun collocations the base is a term, the collocate is a nominalization of a verb or adjective.

[5] The study appeared in Eurolex 1998 under the title "Caractérisation des combinaisons lexicales spécialisée par rapport aux collocations de langue générale"

- Combinations of terms with a collocate are a matter of convention; for example, some companies use special collocations to distinguish their corporate language from that of competitors.
- Terminological collocations can lead to series in which several terms share the same collocate.

The extraction procedure consist of two subtasks: first single word terms are searched for which are to function as bases of the collocations; then regular expressions search for the desired collocations exploiting the grammatical annotation of the word sequences. The identification of single-word candidate terms, next to using typical morphological features (Heid et al. 1996), is based on the relative frequency method postulated in Ahmad et al. (1992). Heid uses a stop list of verbs and of adjectives taken from a collection of general language collocations in order to avoid collocations not relevant to his purpose.

5.4 Contextual Approach

The EAGLES Lexicon Interest Group (1998) argues for the inclusion of contextual information for term retrieval. The authors note that the linguistic and statistical information used for the extraction of terms is 'internal', i.e. stemming from the candidate term itself. A method of retrieving multi-word candidate terms is to exploit their 'external' information, namely information derived from the context of the candidate term. The idea is that terms of the same domain share some common context: They give the example that the form *shows* of the verb *to show* in medical domains is almost always followed by a term. Thus, a word preceded by this verb is a candidate term. A further form, *is called*, from the verb *to call*, is mostly used to define terms. A word following this verb is extracted as a candidate term. These and other kinds of words are thus incorporated in the retrieval of terms.

The same idea for identifying terms is found in Condamines (1995:225), although she does not mention it in relation to ATR but for detecting relations between concepts. She proposes to use lexical markers occurring in some precise context as indications of semantic relations. For example, a structure such as "... V N1 *specially* N2" can indicate a hyponymic relationship between N1 and N2, as in the sentence *Paul likes flowers and especially roses.* (Condamines 1995:225) A lexical marker like *specially* embedded in such syntactic rule is a possible way of retrieving candidate terms.

In principle, this method can be regarded as collocation extraction. The difference is that the EAGLES procedure drops the 'collocates' (*shows, is called*) once a term has been identified.

In true collocation retrieval, though, the collocate is a compulsory element of the terminological unit.

5.5 Conceptual Approach

The aim in the preceding ATR approaches has been to find out how terms are composed, in what kind of phrasal structure they occur, and whether their frequency correlates with termness. A conceptual or semantic approach looks for the termness of words by manually tracing semantic relations like synonymy, hyponymy, inclusion, part-whole, cause and effect, etc. between them. The aim is to combine linguistic data (provided by terminologists) and world-knowledge data (i.e. semantic relations, provided by knowledge engineers) for the development of terminology knowledge bases (Condamines 1995). Conceptual approaches originate from the fields of knowledge engineering and artificial intelligence, which is why their discussion has not been included in the linguistic approaches. An involvement of these sciences has been initiated by of Aussenac et al. (1995) [6]. Their objective has been to combine the ATR-tool *LEXTER* with a knowledge acquisition tool called *MACAO*. Basically, LEXTER provides a terminological hypertext output and MACAO provides a knowledge representation language to investigate concepts. The knowledge engineer models the LEXTER output according to related concepts, identifies terms and builds with them a lexicon which characterises the vocabulary of the domain.

The usefulness of this method for ATR is doubtful. Firstly, it requires extensive human intervention for the constitution of a conceptual network. An experienced terminologist must establish the terminology of a particular domain (by using the ATR generated output) before the knowledge engineer makes links between terms and concepts. Secondly, once all types of relations between concepts have been established, these relations are not used for the extraction of new terms. Currently, conceptual networks, also called knowledge bases, are used in knowledge acquisition. In terminology work they can be of help in the manual compilation of a terminological database to establish the conceptual relation between terms, but not for the retrieval of new terms.

[6] The working group calls itself "Terminology and AI".

5.6 Summary

The main stream in ATR concentrates on the recognition of NP structures. The approaches presented differ mainly in the computer infrastructures used for the ATR process. The NPs are identified either by means of regular expressions applied to PoS-tagged texts or with parsing techniques. Morphological features are currently exploited for German only. Some approaches stress the importance of term frequency in selecting good candidate terms from large corpora. One method incorporates contextual information for the identification of terms. In conceptual approaches, automatically retrieved terms are employed for labelling the relationships between concepts.

Most approaches aim to extract as many candidate terms as possible rather than miss a term. This leads to a large amount of noise in the extraction results and thus to time-consuming editing work for the terminologist. Working experience in language services teaches that the acceptance of language processing tools is rather low. A tool with a high error rate will be rejected, no matter how many well-formed results it has provided. An improvement of the quality of the extracted terms has been provided by Heid: besides a computer-based NP extraction, it is useful to check this output against a list of lemmas considered to be significant. Ideally, the user should be able to edit the content of this lemma list in order to influence the performance of an ATR tool. This method has been adopted for the implementation of a banking-specific ATR tool, which will be the subject of the following section.

6 Self-Developed Tool: ATR from German Banking Texts

Previous studies on ATR generally acknowledge that domain specific terminology consists mainly of NPs. Depending on the domain, different structures for terminological NPs and non-terminological NPs have been recognised. A further observation is that morphological features are typical of the vocabulary of a certain domain. Nearly all of the previous studies discussed deal with the linguistic characterisation and extraction of technical terminology. My study is concerned with linguistic elements characteristic of German terms from the domain of banking, a domain which has been largely neglected in the field of ATR[7]. Starting from an empirical study of the linguistic properties of a terminology database, a method for the retrieval of banking terminology has been implemented. The ATR-tool was developed in the following steps:

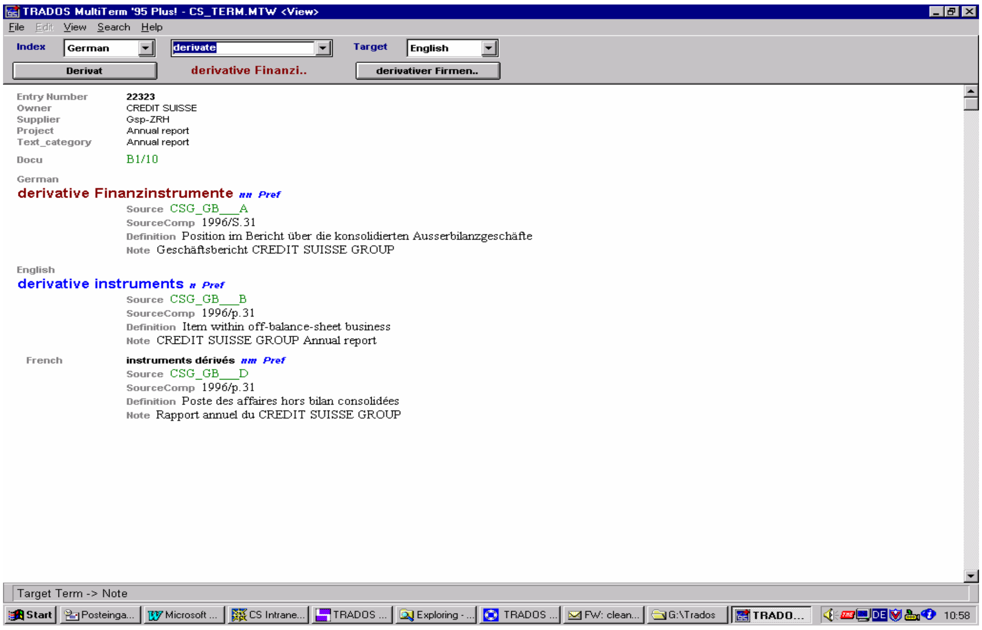
- Supply of a banking-specific terminology database which has been compiled manually
- Linguistic analysis of the terminology database with the aim of identifying banking-specific characteristics in the structure of the entries and in the composition of single words
- Adaptation of (available) tools for tokenisation, morphological analysis, and parsing purposes
- Definition of a grammar for the identification of domain specific NPs
- Implementation of an algorithm for the identification of domain specific morphological features
- Corpus tests
- Evaluation

In the following sections these steps are discussed in depth.

[7] To my knowledge, there has been only one study on financial language for computational purposes, namely Andrew Sanvay and Khurshid Ahmad's analysis of the role of metaphors in financial texts.

6.1 CS-term - a Terminology Database for the Banking Domain

The material for the linguistic analysis of banking terminology has been provided by CS-term, the terminological database of CREDIT SUISSE. CS-term has been compiled for years by terminologists and translators for translation purposes. The entries are stored in four different languages, German, English, French, and Italian. The terminology management system worked with is *Multiterm* of the TRADOS company. Figure 6 shows a sample CS-term entry:



Index: German — derivate — Target: English

Entry Number 22323
Owner CREDIT SUISSE
Supplier Gsp-ZRH
Project Annual report
Text_category Annual report

Docu B1/10

German
derivative Finanzinstrumente *nn Pref*
Source CSG_GB___A
SourceComp 1996/S.31
Definition Position im Bericht über die konsolidierten Ausserbilanzgeschäfte
Note Geschäftsbericht CREDIT SUISSE GROUP

English
derivative instruments *n Pref*
Source CSG_GB___B
SourceComp 1996/p.31
Definition Item within off-balance-sheet business
Note CREDIT SUISSE GROUP Annual report

French
instruments dérivés *nm Pref*
Source CSG_GB___D
SourceComp 1996/p.31
Definition Poste des affaires hors bilan consolidées
Note Rapport annuel du CREDIT SUISSE GROUP

Figure 6: Screen shot of CS-term entry in Multiterm

CS-term consists of about 23'000 entries. The entries are provided with definition, context, grammatical information, and with cross-reference to corresponding document sources. As a further aid to understanding the terms, each entry is assigned at least one subject field ranging from banking, economy, politics, insurance to natural science, culture, sports, etc. There is, however, no systematic conceptual organisation of the subject fields. For example *banking* and *electronic banking* are considered as two equivalent subject fields although within a typological model the second would be a sub-domain of the first. For this work all subject fields which can unequivocally be considered as characterising banking terminology were filtered out. The total number of German entries belonging to these subject fields is 8830. The purpose of the linguistic analysis on these terms is to illustrate their syntactic structure and morphological features. The procedure of this linguistic analysis and its results are the subject of the next section.

6.2 Linguistic Analysis of CS-term

The linguistic analysis of CS-term included the exploration of the syntactic structure and domain-specific morphological features of every single entry. Recurring linguistic elements found were used as criteria for the implementation of a grammar and for the construction of a lexicon with domain specific morphemes. This linguistic analysis was performed with GERTWOL. In the next sections this program is introduced, then the method of analysis is presented. Finally, the linguistic characteristics of CS-term which were identified are discussed.

6.2.1 The GERTWOL System

GERTWOL is a system for the lemmatisation of German words and for the automatic recognition of morphosyntactic elements like PoS, case, gender, and number. The GERTWOL lexicon contains approximately 85,000 words taken from the *Collins German Dictionary*. GERTWOL is based on the so-called 'Two-Level Morphology' (TWOL), a theory which treats morphological phenomena in two ways, on the lexical and on the surface level (Covington 1994:276). A morphological analyser like GERTWOL "determines as early as possible what word it is looking at, and chooses morphological rules taking this knowledge into account." (Covington 1994:276)

For this work, a list of 8830 German terms filtered from CS-term was assigned to GERTWOL. Figure 7 displays the form of a sample after being analysed by this program.

```
"<das>"
                "das"  ART DEF SG NOM NEUTR
                "das"  ART DEF SG AKK NEUTR
                "das"  PRON DEM SG NOM NEUTR
                "das"  PRON DEM SG AKK NEUTR
                "das"  PRON RELAT SG NOM NEUTR
                "das"  PRON RELAT SG AKK NEUTR
```

```
"<Erlebensfallkapital>"
                         "Er|leb~en\s#fall#kapital"  S NEUTR SG NOM
                         "Er|leb~en\s#fall#kapital"  S NEUTR SG AKK
                         "Er|leb~en\s#fall#kapital"  S NEUTR SG DAT
                         "Er|leb~en\s#fall#kapital"  S(A) POS SELTEN
                         "er|leb~en\s#fall#kapital"  * A POS UNDEKL
```

Figure 7: Gertwol output of the NP *das Erlebensfallkapital*

GERTWOL verticalises the single words of the NP, performs a morphological segmentation and provides the PoS structure of the words. These characteristics will be explained in detail:

6.2.1.1 Segmentation of Words

The segmentation in GERTWOL allows to find free and bound morphemes of the word. GERTWOL has four segmentation characters:

#	strong segmentation (Erlebens # fall # kapital)
|	weak segmentation (Er | leb)
\	linking element (Erleben \ s)
~	suffix (Erleb ~ en)

A strong compositional boundary (#) separates elements which can also appear as independent words, i.e. free morphemes. This boundary allows thus to identify domain-specific lemmas. For example, in the case of *Er|leb~en\s#fall#kapital* only the lemma *kapital* would be chosen as characteristic of the domain of banking. In the analysis of CS-term the boundary '#' was taken as a lemma separator.

A weak compositional boundary (|) separates bound morphemes of the derivational type which occur in compounds, for example prepositions and prefixes. A weak composition boundary also occurs in combination with so-called half suffixes, i.e. if the meaning of a word in a compound is different from its meaning as an independent word (*Erleb* vs. *leb*). In the analysis of CS-term this boundary was not considered as lemma separator because potential domain-specific lemmas are lost (for example *Ver + mögen* vs. *Vermögen, An + lage* vs. *Anlage, Gut + haben* vs. *Guthaben, Hypo + thek* vs. *Hypothek,* etc.).

Linking elements (\) separate grammatical features like number and case from their stem word (*Börse \ n, Bank \ en*). They alter the form of a word without changing its central meaning. These elements were removed from the domain-specific lemmas retrieved in CS-term.

Suffixes (~) are recognised by GERTWOL as either occurring after stem words (*zahl~ung, rechn~ung*) or after other suffixes (*kredit|würd~ig~keit, Börse\n#kapital~is~ier~ung*). Suffixes do not alter the meaning of the stem word, however, their exclusion might lead to the retrieval of unwanted general words. In a banking text a compound containing the element *schuld~ner*, *zahl~ung,* or *rechn~ung* is likely to be a candidate term (*Garantieschuldner, Schuldnerverzug, Zahlungsterminal, Deckungsbeitragsrechnung, Einsatzgebietskosten-rechnung*). The elements *schuld*, *zahl* or *rechn,* though, occur mostly in words of general character (*Schuld, Schuldbekenntnis, Zahl, Zahltag, Zahlstelle, Rechen, Rechenschaft, Rechenaufgabe*, etc.). In a first step, the automatically retrieved morphemes included stem words retaining their suffixes. Later, in a manual editing, suffixes were removed from those stem words which were unmistakably banking-specific (for example *bank* vs. *bankier, debit* vs. *debitor, hypothek* vs. *hypothekar, invest* vs. *investor, zins* vs. *verzinslich*).

6.2.1.2 Grammatical Categories

In figure 7 (section 6.2.1) we see that GERTWOL assigns to the word *das* the PoS article (ART) and pronoun (PRON). For the word *Erlebensfallkapital* it identifies the possible PoS noun (S), nominal adjective (S[A]), and adjective (A). GERTWOL marks further grammatical features of the words like number, case, and gender. For the recognition of the syntactic structure of a term entry, only the information on PoS were extracted.

6.2.2 Analysis of the GERTWOL Output based on CS-term

The GERTWOL output based on CS-term contained several PoS and segmentation ambiguities which had to be resolved previous to a morphosyntactic characterisation of the entries.

6.2.2.1 PoS Ambiguities

As has just been observed, certain words are assigned more than one PoS by GERTWOL. In the NP *das Erlebensfallkapital* the article *das* is assigned the appropriate PoS *ART* and the wrong PoS *PRON*; *Erlebensfallkapital* is provided with the correct reading *S* and the wrong readings *S(A)* and *A*. For these cases a procedure which selects the most probable PoS had to be defined in the program.

The most frequent PoS ambiguities were found in combination with nouns. Since most terms are nouns, this PoS was preferred to all others. The ambiguity between defined articles and relative pronouns was also frequent. Furthermore, some adjectives may formally correspond to verbs or nominal adjectives (if capitalised). Adjectival participles are additionally also identified as verbs. In table 1 the PoS preferences for the identification of the syntactic structures are listed.

Table 1: Preferred PoS in case of ambiguity

Preferred PoS	Rejected PoS
S	all
ART	PRON
A	S(A), V, ADV
A(PART)	V

6.2.2.2 Segmentation Ambiguities

Ambiguous segmentation is encountered when a word has one PoS but is lemmatised in two or more ways. Nouns are not only the most frequent word forms both in general and specialised language; they are also a frequent source of ambiguous lemmas when analysed automatically. A study by Volk (1999) on the lemmatisation of nouns in a newspaper corpus has shown that 10% of all noun types are assigned more than one lemma by GERTWOL. In order to resolve ambiguous lemmatisation, Volk performed an analysis of the most probable segmentation. In this study he formulates the principle of the "least internal complexity", stating that in cases in which the same noun is segmented by GERTWOL in a weak and a strong way, the weak way should be preferred. In order to discard segmentation ambiguity for nouns, Volk's program was run on the GERTWOL output of CS-term.

Some examples from CS-term which are segmented by GERTWOL in a weak and a strong way are given in table 2.

Table 2: CS-term entries with ambiguous morphological segmentation

<table>
<tr><th>Word</th><th>Weak segmentation</th><th>Strong segmentation</th></tr>
<tr><td>Einzahlung</td><td>ein|zahl~ung</td><td>ein#zahl~ung</td></tr>
<tr><td>Annahmeerklärung</td><td>an|nahm~e#er|klär~ung</td><td>an|nah#meer#klär~ung</td></tr>
<tr><td>Antragsteller</td><td>an|trag|stell~er</td><td>an|trag\s#teller</td></tr>
<tr><td>Rückstellungen</td><td>rück|stell~ung</td><td>rück#stell#lunge</td></tr>
<tr><td>Bankeinzugsverfahren</td><td>bank#ein|zug\s#ver|fahren</td><td>bank#ein#zug\s#ver|fahr~en</td></tr>
<tr><td>Kreditüberwachung</td><td>kredit#über|wach~ung</td><td>kredit#üb#er|wach~ung</td></tr>
<tr><td>Leasingfinanzierung</td><td>leas~ing#finanz~ier~ung</td><td>lea#sing#finanz~ier~ung</td></tr>
</table>

6.2.2.3 Treatment of words unknown to GERTWOL

As already stated, GERTWOL disposes of a lexicon of 85'000 words. Some words in CS-Term were not recognised by GERTWOL. Most of these words are English or French nouns like in the case of *Briefing, Cash Dispenser, Cash-Flow-Statement, Bulletin*. Some unknown words are a mix of German and English lemmas like *Offshore-Gesellschaft, Offshoreprinzip, Doppel-Sharing, Contractinggeber*. Finally, there are German words like *Investitions-rentabilität, Invalidierungswahrscheinlichkeit,* or *Dekotierung* which are too specialised to be found in the GERTWOL dictionary. Since the great majority of the CS-term entries are nouns (see next section), it was decided that words unknown to GERTWOL should be assigned the category *S* (for noun) if they were capitalised.

6.2.3 Results of GERTWOL Analysis

6.2.3.1 Syntactic Characteristics

The syntactic structures and the overall percentage of the CS-term entries analysed by GERTWOL are arranged in table 3.

Table 3: PoS-structures and their frequency in percentages identified in CS-term; the last column gives examples of wrong PoS assignment

PoS Pattern	%	Examples Noise
S	63	AHVG, AHVV, AI, Aggregiert, Blankowürdig, Zinsreagibel
A S	7	-
S S	4	Antizipierter Abrechnunsposten, Affaire Liée, Wertzuberichtigendes Engagement, Antizipative Aktiven, Blended Cover, First Mover
ABK	3	-
A(PART) S	2	-
A	1	-
S V	1	Capital Asset, Cross Rate
S ART S	1	-
S CONJ S	0.5	-
S S S	0.5	Weighted Average Cost
S ABK	0.5	-
others	16.5	

Table 3 reveals that the overwhelming majority of terms in CS-term are single nouns. These make up 63% of all entries. Another fairly frequent construction is the collocation adjective-noun (7%). With decreasing frequency noun chains (4%), acronyms (3%), and adjectival participles followed by nouns (2%) are found. The remaining structures are also of nominal character. Special collocations, which were the subject of other ATR approaches, like noun-verb and noun-genitive noun constitute only 1% of the identified structures. No relevant occurrences of noun+preposition+noun group collocations were identified. A comparison of the length of terminological NPs identified in CS-term with NPs extracted in previous ATR approaches reveals some syntactic differences. Whereas in most approaches long NP chains were extracted as candidate terms, the majority of CS-term entries have neither a preposition, nor a pronoun, nor an adverb. CS-term entries are mostly composed of nouns and adjectives and have a very short structure. Words of closed classes seem to function as terminological boundaries within long NPs and will thus be eliminated in the implementation of the ATR process.

In the third column of table 3 we see that noise affects almost exclusively single nouns (S) and noun chains (S S). This outcome was to be expected, since nouns make up the largest number of terminological entries. A frequent cause for these errors is the default PoS *S* assigned to capitalised words unknown to GERTWOL. Of course, not all of these words are nouns. Another source of errors are capitalised adjectives, adjectival participles, or nominal adjectives which have the same form as nouns. As we have seen in section 6.2.2.1, in case of PoS ambiguities involving these word classes, the PoS *S* for noun is always given priority. Some of the noise encountered in noun-verb collocations is caused by English nouns which have the same form as German verbs, like in the example of *asset* (past tense of *essen* 'to eat') and *rate* (imperative and 1. person present of *raten* 'to guess'). Manual scrolling through the GERTWOL output revealed no noise in the other structures. A quantification of the error rate would have been too laborious and was thus not performed.

Overall, 66 different syntactic patterns were found in CS-term. The 11 structures listed in table 3 account for 83.5 % of all entries. The remaining 16.5 % of CS-term entries are composed of 55 different structures. The prevalence of the structures listed in table 3 is a motive for considering them in the grammar for banking terminology to be defined later.

The results of the CS-term analysis confirm the studies on German LSP, namely that terms have a strong nominal character. Nevertheless, as can be observed in table 3, and as the candidate terms extracted on the basis of these structures will show (section 6.3.6), terms are not dramatically different from non-terms as regards structure. An automatic term extraction relying on syntactic structures only does not lead to success.

6.2.3.2 Morphological Characteristics

The goal of the morphological analysis was to compile a list of lemmas found in banking terms. These will be used for term validation during the ATR process. Before choosing a candidate term, the ATR program verifies whether there are occurrences of the word or of its segments in this list. If this proves to be true, the word is chosen as a candidate term. This banking-specific list of lemmas will from now on be called *lemma-lexicon.*

As the discussion of the GERTWOL segmentation (section 6.2.1.1) revealed, the lemma-lexicon consists of stem words found in the CS-term entries. As mentioned before, stem words were sometimes retrieved with suffixes in order to enhance their domain specificity.

From the 8830 entries in CS-term GERTWOL extracted 2248 different stem words. In example 1 the banking-specific stem words which have a frequency higher than 20 are arranged (listed in descending frequency)[8].

(1) kredit, wert, bank, risiko, geschäft, markt, aktie, konto, schrift, rechnung, kapital, kunde, anlage, zins, wechsel, vermögen, system, zahlung, kurs, satz, kosten, geld, finanz, limit, börse, depot, check, grund, fond, privat, pfand, stelle, rück, buch, währung, gewinn, spar, leasing, bilanz, handel, deckung, schein, option, brief, anteil, prämie, folio, management, liquidität, invest, brutto, gold, effekten, konzern, inkasso, hypothek, transaktion, kassa, umsatz, giro, gebühr, dollar, diskont, bonus, valuta, rendit, dividend, bonität, valoren, transfer, tarif, portefeuille, makler, euro, erwerb, einzahlung, derivat, darlehen, cash, tresor, skonto, schuldner, saldo, rabatt, parität, säule, marge, fund, scheck, defizit, debit

Information on the frequency of these lemmas is deliberately omitted because they do not only compose terms but are also found in general words. For example, *schrift* is a frequent lemma in banking terms like *Wert#schrift, Zins#gut#schrift,* or *Last#schrift#ver|fahr~en*, but it also appears in CS-term entries such as *Ver|sicher~ung\s#zeit#schrift, Zeit#schrift\en#ver|lag,* or *Block#schrift*. It is legitimate to ask why the latter words occur in a banking-specific terminology database. This example discloses that CS-term contains not only terms but also entries which are not strictly terminological.

If standing alone most lemmas listed in example 1 are of general character. They unfold their terminological relevance only in combination with other morphemes or as part of multi-word terms, as table 4 depicts.

[8] The GERTWOL segmentation boundaries ~ and | were deleted to improve legibility.

Table 4: 'General' lemmas and their terminological character when occurring in compounds

Lemma	Examples
stelle	Unterzahlstelle, Vorkostenstelle, Zahlstelle, Zahlstellenklausel, Belehnungsstelle, Zulassungsstelle, Abrechnungsstelle, Ausgleichsstelle
system	Clearingsystem, Dispoverwaltungssystem, Dreisäulensystem, Einzelkontoführungssystem, Kostenverrechnungssystem, Währungssystem, Überweisungssystem
kunde	Firmenkunden, Firmenkundensegment, Kundensegment, Kundenakquisition, Kundenausleihungen, Kundenfestgelder, Individualkundengeschäft, Kleinkundengeschäft
geschäft	Kreditgeschäft, Kollektivgeschäft, Kommissionsgeschäft, Kompensationsgeschäft, Komptantgeschäft, Kostgeschäft
buch	Restbuchwert, Buchhaltung, Buchwert, Abwicklungsbuchhaltung, Aktienbuch, Buchführungsgrundsätze, Bankenbuch
währung	Anlagewährung, Währungsklausel, Basiswährung, Doppelwährungsanleihe, abwertungsverdächtige Währung, konvertierbare Währung, Währungs-Futures
wechsel	nationalbankfähiger Wechsel, nicht bankfähiger Wechsel, kommerzieller Wechsel

6.2.4 Summary

The linguistic analysis of the term bank CS-term confirmed that syntactically terms are of nominal character. The syntax of most terms is represented by eleven different structures, nine of them being NPs. Almost all these NPs lack articles, prepositions, adverbs and other words of closed classes. Morphologically, the majority of the CS-term entries are composed of few stem words. Frequently recurring stem words were stored in a lemma-lexicon and will help spot banking terms. In the ATR process, any word containing a lemma listed in this lexicon can be identified, no matter what the morphological variation of the word is or in what position the lemma is found in a compound word. This method is particularly useful for German, which makes frequent use of noun compounds.

The results of the CS-term analysis were used for the implementation of an ATR tool, which will be the focus of the next section.

6.3 Term Identification: Implementation

The implementation of the program for term extraction is based on two steps:

1. Identification of NP structures by means of parsing techniques; from now on this step will be called *NP-tool*.
2. Regular expressions processing the output of the NP-tool for the identification of domain-specific morphemes; from now on this step will be called *lexicon-lookup*.

The flow chart of this ATR process is shown in figure 8. The single steps involved in the extraction process will be explained in the next sections.

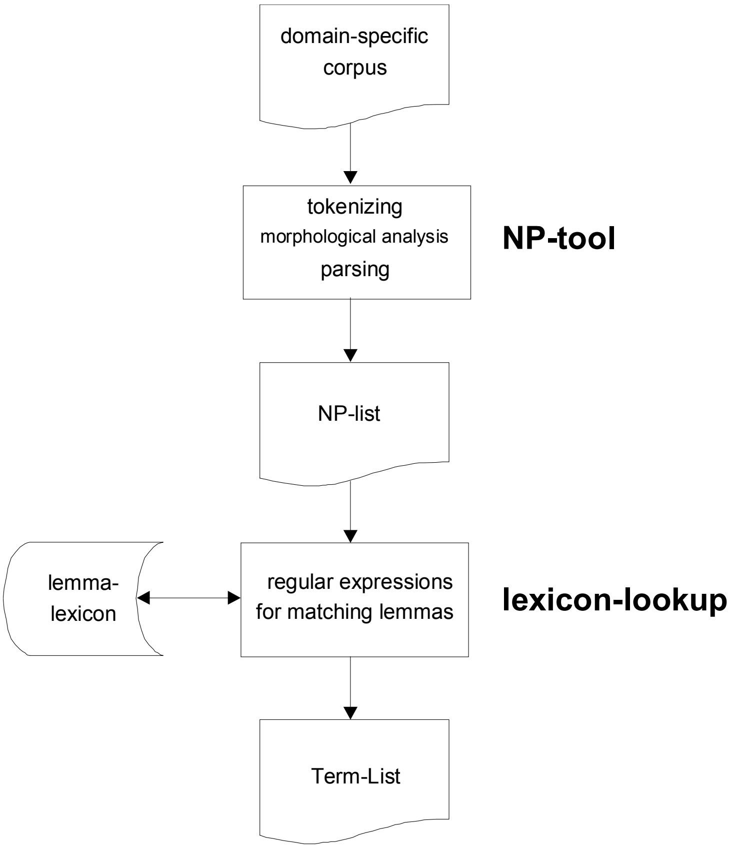


Figure 8: Flow chart of term extraction process

6.3.1 NP-tool

The NP-tool is a program for the extraction of NPs from German texts. It was developed at the University of Zürich within the scope of a project on answer extraction concerning student registration. In my work, the NP-tool was adapted for the extraction of terminological NPs.

The major adaptations affected the grammar, which will be discussed later. In the NP-tool, the processing of a sentence implies three phases: a tokeniser splits sentences into words, GERTWOL performs a morphological description, and a parser maps the NP-structure. With the sample sentence *Kotierte Aktien werden zum Marktwert bilanziert.* the various operations performed by these three modules in the intersections of NPs will be traced:

6.3.1.1 Tokenisation

In a tokeniser, a sentence is defined as a string of words terminated by a punctuation mark (which is also treated as a word). In the tokenising process the sentence and word boundaries are determined and the words are represented as a list of atoms. Capitalised words are put between quotes in order not to be mistaken as variables by Prolog, the programming language.

The sentence *Kotierte Aktien werden zum Marktwert bilanziert.* is tokenised in the following way:

['Kotierte', 'Aktien', werden, zum 'Marktwert', bilanziert, '.']

6.3.1.2 Morphological analysis

Once the sentence has been split up, the elements of the list are analysed by GERTWOL. The performance of this program and the form of its output have been the subject of section 6.2.1. Before submitting the GERTWOL output to the parser a perl script excludes unwanted PoS readings or morphological features. Words which need different PoS features than those assigned by GERTWOL are stored separately in a user defined lexicon. Words which are not recognised by GERTWOL (mostly titles, abbreviations, proper names, and non-German nouns) are also stored in this lexicon. In the end, each word in the sample sentence is assigned morphosyntactic features and submitted to the parser.

6.3.1.3 Parsing

Parsing is the process of assigning grammatical structure to a surface string. The parser used is a bottom-up chart parser. Bottom-up parsing combines words and their lexical categories into phrases and those into sentences. The advantage of chart parsing is that it saves all partially completed results in a chart thus avoiding reparsing well-formed syntactic constituents. The NP rules of the parsing process were defined using phrase structure (PS) grammar. The NP constituents are complex feature structures. A complex feature structure is

a set of features and values. For example the NP constituent *S* for noun can have the feature structure:
case=nominative
number=singular
gender=feminine
The operations are carried out by recursive list processing and unification of feature structures. The parsing output can take the form of a nested sublist, of a tree diagram, or of SGML annotations in running text. For this work nested sublists were chosen as output form (figure 9). From the list of parsed NPs only those with maximum length are chosen as term candidates.

Kotierte Aktien
(N2(AdjP(A(A(PART)(Kotiert))))(N1(S(Aktie))))
dem Marktwert
(N2(ART(der))(N1(S(Markt#wert))))

Figure 9: Parser output (shaded) of the sentence *Kotierte Aktien werden zum Marktwert bilanziert.*

This output visualises the hierarchical manner in which words are combined to form NPs. The PS grammar contains a rule which states that a NP (N2) can be broken down into an adjectival phrase (AdjP) and a noun group (N1) or into an article (ART) and a noun group. Noun groups can have several elements. In figure 9 they are composed of one noun (S). The adjectival phrase is composed of one adjective (A) which is further identified as a participial adjective (A[PART]). All morphosyntactic features attached by GERTWOL to the words in the tokenised list are compared to match these and other patterns. In this way the parser resolves the remaining morphosyntactic ambiguities.

In figure 9 we see that after the parsing process the words retain their GERTWOL segmentation (*Markt#wert*). This segmentation will be needed for the identification of banking-specific lemmas.

The next section discusses the modifications which had to be made in the grammar for the purpose of term extraction.

6.3.1.4 Adaptations in the NP-tool Grammar for ATR Purposes

The goal of the design of the original NP-tool grammar had been to have a complete coverage of the German NP syntax in real-world texts. The NP grammar took into account the preference of the German language for building long sentences with embedded phrases. However, complex NP structures necessarily lead to many ambiguous analyses, mostly caused by recursive NPs and by embedded prepositional phrases (PP). A NP with an embedded PP, which in turn has other embedded NPs, can easily provide up to 50 different readings. Although NP chains are common constructs in German, they are not very relevant from a terminological point of view. This observation was reinforced by the analysis of CS-term, which showed that the great majority of terminological entries are just single nouns and that very few entries are constituted of NP chains.

If rules retrieving short NPs are applied, the parsed output is undoubtedly of higher quality. The parser hardly ever fails to recognise the correct structure. For terminological purposes the only problem of retrieving short NP structures is the exclusion of recursive genitive NPs. In the case of institution names, for example *Eidgenössische Diplomprüfungskommission der Bankwirtschaft* or *Vereinigung der Schweizer Börsen,* the term is composed of the entire word sequence and an extraction of its parts is a misinterpretation. There are, however, cases like *Management der Fristentransformation* in which the terminological information is in the second noun and an extraction of the whole NP does not necessarily improve the content of the term.

Generally, NP chains are rarely valid terms and thus of little advantage to automatic term identification. In addition, they are a cause for parsing errors. In order to reduce the rate of wrong NP structures NP chains were not considered. The NP rules which were defined for the retrieval of candidate terms are given in figure 10.

$$\begin{array}{lll}
\text{N2} & \longrightarrow & \text{(DET) (AdjP) N1} \\
 & & \text{N2 CONJ N2} \\
\text{N1} & \longrightarrow & \text{S}^{+} \\
 & & \text{ERSTGLIED} \\
 & & \text{ABK} \\
\text{S} & \longrightarrow & \text{S(A)} \\
 & & \text{S(PART)} \\
\text{AdjP} & \longrightarrow & \text{A}^{+} \\
 & & \text{A(PART)}
\end{array}$$

Figure 10: NP-rules for the extraction of candidate terms

Figure 10 gives a simplified form (i.e. without feature structure) of the NP rules that were defined to retrieve candidate terms. Constituents enclosed in brackets have an optional character. The operator '+' indicates the possibility of more than one occurrence. The full names of the constituents are listed in the appendix.

6.3.2 Lexicon-lookup

The identification of lemmas in potential candidate terms is performed by regular expressions over the (segmented) words. The input strings in this module are the extracted NPs. For every input NP, lemmas are matched by regular expressions. Lemmas are identified if one of the constituents *S, ERSTGLIED, ABK, S(A), S(PART), A,* or *A(PART)* are followed either by a sequence of letters enclosed in brackets (word is not segmented) or by a sequence of letters preceding or following the sign '#' (word is segmented). If at least one lemma in the NP has a match in the lexicon, the whole NP is stored in the list of candidate terms. In this way single-word and multi-word terms are retrieved. The flow-chart in figure 11 visualises this process:

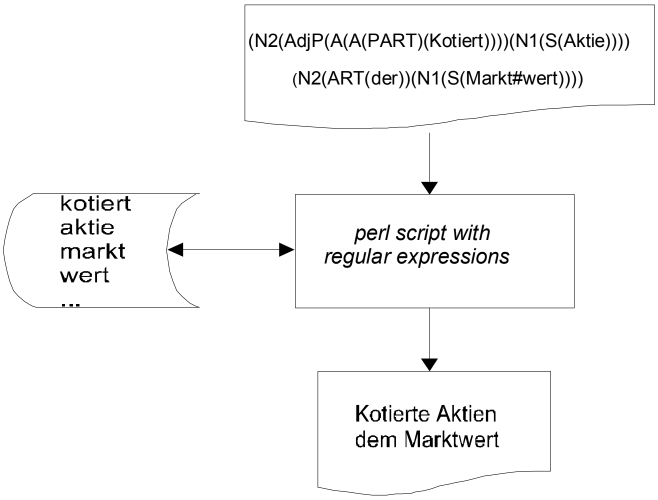


Figure 11: Flow-Chart of lexicon-lookup module

The most important component in this system is the lexicon, counting roughly 100 banking-specific lemmas frequently used in CS-term entries (listed in example 1). Based on this small lexicon almost all terms in the test corpus were retrieved (section 6.3.6). Infrequently used domain-specific lemmas which were not retrieved by the analysis of CS-term are a cause for the loss of relevant terms in the extraction procedure. To improve the ATR performance, the user can enter missing lemmas in the lexicon with little effort once they have been identified.

As described above, if the lexical lookup finds a lemma it puts the whole NP on the candidate list, no matter how many lexical elements a NP is composed of. Retrieving the whole NP may mean extracting general words. This is not necessarily a disadvantage, considering the lexical character of individual parts of terminological multi-word NPs (Justeson and Katz 1994:2). General words can be important elements for determining the meaning of a multi-word term. For example, the term *Aktie* 'share' refers to a sum used or available for investment; the term *junge/neue Aktie* 'new share' designates those shares which are newly issued and which are not entitled to dividend payment in the past business year. If the general modifiers *junge* or *neue* are omitted from this lexical NP a different concept is referred to. This example illustrates the danger of stop words. Eliminating generic modifiers might lead to missing important multi-word terms.

6.3.3 Software and Hardware Platform

NP-tool was implemented in Prolog, a language especially suited for the processing of natural language. Prolog provides a built-in unification operation which can be exploited in the search of lexical, morphological, and syntactic information. Further, Prolog is able to handle recursiveness, which plays a fundamental role in the processing of natural language. The program for the lexicon-lookup was implemented in Perl, a programming language which allows regular expressions to process string of words.

Both the NP-tool and the lexicon-lookup were developed on Sun workstations under the Unix operating system. The aim of the development of these tools was to provide a technique (which incorporates some ideas of other ATR researches) for the acquisition of banking terms. While these tools will remain in the area of research, this simple but - as will be shown in the next section - efficient technique might be integrated into a commercial system for term retrieval.

The next section deals with the performance of the ATR process. There will first be a quantification of the results on the basis of the NP extraction only. In a second step, the positive influence of the lemma-lookup on the quality of the NPs finally chosen as candidate terms will be demonstrated.

6.4 Evaluation

Strategies followed in previous ATR approaches are governed by the belief that for a terminologist it is easier to delete wrong candidate terms than to manually find those missed by the automatic extraction. This explains why comparisons of the number of terms identified manually and of those extracted automatically showed that test persons identify much fewer items as terms than an ATR system does (L'Homme et al. 1996:303). This evaluation will demonstrate that this observation is true only for the NP extraction, whereas the number of terms retrieved after the lexicon-lookup is close to the one of manually identified terms.

6.4.1 Test Corpus

The test corpus counts a total of 2712 words. It was taken from two annual reports of CREDIT SUISSE dated 1997 and 1998. Annual reports are known to be a good source for domain-specific terminology. To improve the results of the NP extraction non-textual data like formula and figures were manually removed from the corpus.

6.4.2 Method

The method applied for this evaluation was to compare the list of terms manually identified first with the one produced by the NP-tool and then with the one extracted by the lexicon-lookup module. The manual term identification was performed previous to the runs of the programs. This list is assumed to be perfect, although it is a fact that no absolute judgement exists for the distinction between terms and non-terms. The output of the NP-tool was first checked for its terminological content and then it was processed by the lexicon-lookup module. The lists of candidate terms provided by the NP-tool and lexicon-lookup module were organised in alphabetical order and the frequency rate of the candidate terms were computed. Items which differed in their inflectional form (for example *Bank/Banken*, *Eigenkapital/Eigenkapitals*, etc) were manually reduced to their basic form. Examples for noise and silence will depict the errors of the two modules. Rates of noise and silence will be used to investigate the effectiveness in terms of recall and precision of the NP-tool and of the lexicon-lookup module.

6.4.3 Results

In the manual analysis of the test corpus a total of 248 terms were identified. These terms will from now on be referred to as *relevant items.* Syntactically, all relevant items were NPs. For the evaluation, the list of relevant items was compared to the lists generated by the automatic extraction. The goal was to identify those elements that are

- common to all three lists (manual, NP extraction, and lexicon-lookup)
- unique to the manual list and to the ATR lists.

An overview of the results of the NP-tool and lexicon-lookup module is given is table 5.

Table 5: Automatically retrieved and validated candidate terms in the NP-tool and lexicon-lookup

Module	Retrieved items	Retrieved relevant items	Noise	Silence
NP-tool	571	241	330	7
lexicon-lookup	229	213	16	35

The total number of candidate terms identified by the NP-tool is 571, 241 of which coincide with the relevant ones. The total number of items identified by the lexicon-lookup is 229, 213 of which correspond to the manual list. The errors produced by the ATR modules are measured in terms of *noise* and *silence*. L'Homme (1996:32) gives the following definitions of these two categories:

- noise: combinations that are listed, but are not terminological
- silence: terminological units present in a text, but omitted by the system

The output of the NP-tool contains much noise but little silence. In fact, 330 of the 571 extracted NPs are not terminological, however, only seven of the relevant terms are missed. The noise for the lexicon-lookup is quite low, since it extracts only 16 items which were not intended. However, it fails to recognise 35 of the relevant terms. The silence rate of the lexicon-lookup module would have been lower if the seven terms not retrieved by the NP extraction had been available. In fact, six of the seven NPs not extracted by the NP-tool are composed of stem words stored in the lemma-lexicon.

Example 2 indicates a selection of items accounting for noise in the NP-tool. In example 3 all noisy words extracted by the lexicon-lookup are listed[9].

[9] Initial determiners are left out because they are not part of terms.

(2) Änderungen, Anfang, angekündigter Zusammenschluss, Anhang, August, ausserordentlicher Ertrag, Bank- und Versicherungsbedürfnisse, Banken oder Kunden, Beispiel Marktrisiken, bessere Ergebnisse, Bestimmungen, Einsatz, Erfolg, Ertragsrückgänge, Erwartungen, Finanzdienstleistungen, Finanzergebnis, Forderungen oder Verpflichtungen, Geschäft, Geschäftsbereich, Geschäftsjahr, hervorragende Performance, historische Simulationsmethode, Jahresgewinn, Jahresrechnungen, Kunden, Kundengruppe und Märkte, Kundensegment, Rechnung und Risiko, Risiken und Versicherungsrisiken, Risikoarten, Schweizer Vorschriften, USA, USD, Vergleich, Versicherungsgeschäft, Vorjahr, Wiederverkauf, Ziel, Zweiten Weltkriegs

(3) Bank- und Versicherungsbedürfnisse, Banken oder Kunden, Beispiel Marktrisiken, Finanzdienstleistungen, Finanzergebnis, Geschäft, Geschäftsbereich, Geschäftsjahr, Jahresgewinn, Jahresrechnung, Kunden, Kundengruppe und Märkte, Kundensegment, Rechnung und Risiko, Risiken und Versicherungsrisiken, Risikoarten, Versicherungs-geschäft

As the words listed in example 2 reveal, among the errors produced by the NP-tool single nouns are by far the most frequent. We are thus confronted with the major problem of an ATR method based simply on syntax: the repetition of single nouns is not an indication of terminological NPs. In a language like German, even the repetition of multi-word NPs does not provide a reliable distinction for term retrieval. The majority of multi-word terms are adjective-noun collocations, a construction also prevalent in general language. In German, an overgeneration of wrong candidate terms cannot be avoided with structural criteria only.

The lexicon-lookup produces little noise (example 3). Most of the incorrect items extracted by the NP-tool are filtered by the morphological component. Noise is mostly caused by the morphemes *ertrag, bank, finanz, geschäft, kunden, rechung,* and *risiko*. In principle, an exclusion of these errors would be achieved by deleting the corresponding morphemes in the lemma-lexicon. However, there is a large risk of increasing the silence rate, since morphemes can lead to noise in some cases but at the same time be a reliable source of correct term identification. In table 6 there is a selection of lemmas which account for the retrieval of valid terms but are also a cause for noise.

Table 6: Lemmas which in the test corpus are found both in terms and in non-terms

Lemma	Term	Non-term
ertrag	Kapitalertragssteuer*	Ertragsrückgänge
finanz	Allfinanz	Finanzergebnis
geschäft	Ausserbilanzgeschäft	Geschäft, Geschäftsjahr
kunde	Firmenkundengeschäft	Kunden
rechung	Fremdwährungsumrechnungskurse	Jahresrechnung
risiko	Risikokapitalbindung*	Risikoarten

* would have been retrieved because of the lexicon entry *kapital*

Sample terms accounting for silence in the NP-tool output are listed in example 4. The causes are either words unknown to GERTWOL (*rückforderbare Quellensteuern*) or NP structures which are not defined in the NP grammar. These missing NP structures are chains of (English) nouns (*Credit Suisse Group Risk Coordination Committee*), recursive genitive NPs (*Kotierungsreglementes der Schweizer Börse*), NPs with embedded PPs (*Verordnung über die Banken und Sparkassen*), and elliptical constructions (*kalkulatorisch ermittelten und effektiven Steueraufwand*). The articles and prepositions in the terms listed in example 4 are treated as NP boundaries by the parser, which not only leads to silence but also to a slight increase of noise. Considering the small number of these structures encountered in the tests, they can be neglected. I am aware, though, that with a larger corpus, the lack of these structures might lead to a distortion of the results.

(4) Credit Suisse Group Risk Coordination Committee, europäische Aktien- und Investmentbankgeschäft der BZW, Kotierungsreglementes der Schweizer Börse, Securities Lending und Borrowing, Verordnung über die Banken und Sparkassen, rückforderbare Quellensteuern, kalkulatorisch ermittelten und effektiven Steueraufwand

The candidate terms which were not extracted by the lexicon-lookup (listed in example 5) were either not on the input list (i.e. all terms listed in example 4) or were constituted of lemmas missing in the lemma-lexicon. The lemma-lexicon had not been enlarged previous to the extraction process. It contained only those lemmas identified in CS-term. If the lemma-lexicon had been completed with additional lemmas, more relevant terminology would have been retrieved. The addition of the two missing lemmas *anleihe* and *methode*, for example, would have lowered the silence rate by ten terms. This example reveals the advantages of a lemma-lexicon. The lemma-lexicon can easily be revised according to the characteristics of the domain or of the type of text, and with very few entries a large number of terms can be retrieved.

(5) Accrual-Methode, Aktiven, Amortized Cost-Methode, Anleihe, Annual Credit Provision, Asset Backed Securities, Beteiligungstitel, CS life, Economic Capital, Financial Products, GRCC, Group Chief Risk Officer, Hedging, Interessenzusammenführungs-Methode, Konsolidierungs- und Bewertungsgrundsätze, nicht kotierte Titel, operativer Reinverlust, Passiven, Pooling-of-Interests-Methode, Spartenergebnisse, Staatsanleihe, Swaps, Value at Risk-Methode, Varianz-Kovarianz-Methode, VaR-Methode, Volatilitäten, Vollkonsolidierung, Wandelanleihe, Winterthur Columna, Winterthur Leben (plus six terms from example 4)

6.4.3.1 Recall and Precision: Two Parameters of Retrieval Effectiveness

Noise and silence rates allow to measure the effectiveness of a system in terms of *recall* and *precision*. The recall and precision measures adopted for this evaluation are based on the following definitions (Salton 1989:248):

- Recall: the ratio *retrieved relevant items(a)/all relevant items (a+d)*
- Precision: the ratio *retrieved relevant items(a)/all retrieved items (a+b)*

These definitions are visualised in figure 12.

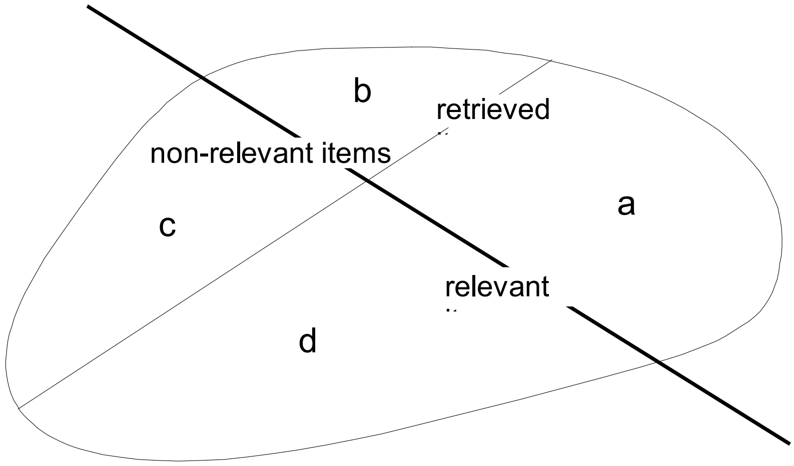


Figure 12: Recall-precision proportions

In figure 12, noise is located in *b* and silence in *d*. A system would perform faultlessly if it had maximum recall (d=0) and maximum precision (b=0). The results previously presented disclose that both the NP-tool and the lexicon lookup are partly far from this ideal performance. As shown in table 7, the NP-tool has a recall rate of 97%, which means that it identifies most items included in the manual list. The high recall ratio is connected with a poor precision rate. In fact, only 42% of the recovered NPs are terms. Thus, the NP-tool recovers a large number of well-formed NPs which are rarely valid terms and must be discarded in time-consuming postediting. The lexicon-lookup has a recall of 85% and misses a fair number of desired terms. However, the low recall rate of the lexicon-lookup is compensated by a high precision rate. Among the items extracted 93% are valid terms.

Table 7: Recall and precision of the NP-tool and lexicon-lookup

	NP-tool	lexicon-lookup
Recall	97%	85%
Precision	42%	93%

The NP-tool yielded high recall and low precision, whereas the lexicon-lookup produced the reverse result. In the previous section it was stated that the recall rate of the lexicon-lookup could easily be enhanced if the lexicon was enlarged with missing lemmas. At this point the risk of such a procedure becomes obvious: if the recall rate is enhanced there is not only an increase of the total number of relevant items retrieved; but the total number of non-relevant retrieved items also grows, and thus decreases the precision. One must keep in mind that any attempt to improve recall can have negative effects on the precision rate.

6.4.3.2 Frequencies of extracted candidate terms

Some researches use statistics to improve the precision rate of an ATR tool. As already mentioned, the disadvantage of a statistical method is that it retrieves only recurring words. Therefore, a large corpus is needed in order to apply statistical constraints. Due to the small test corpus statistics were not emphasised in this work. At the end of the extraction the number of occurrences of each term was computed, though. The outcome showed that if a frequency operation with a cut-off value of two had been used in this ATR process, only one out of nine terms would have been retrieved. Frequency constraints undoubtedly reduce the size of the output, but only by missing a large number of terms.

6.4.4 Summary

The results presented confirm that a limitation on syntactic patterns is not a reliable method for terminology retrieval. The NPs extracted meet all the criteria found in the CS-term analysis. However, mostly they do not contain relevant information. Terminological NPs are not different from general NPs particularly because single nouns constitute the largest subset of candidate terms. The list of retrieved NPs has many useful terms, but also contains a considerable number of words which would never have been extracted on terminological grounds. A retrieval of well-formed syntactic structures is useful; it is, however, inadequate to consider a list of NPs effective for terminology work.

A method of reducing the occurrences of terminologically irrelevant NPs is to exploit their morphological character. The morphological component in this work proved to be a reliable indicator for termness. Most of the items extracted by the lexicon-lookup match those in the manual list. Due to an uncompleted lexicon, a large number of terms were missed by the program. However, this erroneousness could be eliminated with little labour by expanding the lexicon with more domain-specific lemmas. The effectiveness of the lemma-lookup can be influenced according to the type of text and its recurring morphemes. With a more complete lexicon the lemma-lookup would be a solid system for ATR which would certainly make terminology work more efficient.

The tests on the NP-tool and lexicon-lookup have been made on a small corpus. The representativeness of the results discussed is thus uncertain. A major problem in evaluating automatically extracted terms is that no absolute judgement exists for the distinction between terms and non-terms, not even in practical terminology. Term identification depends not only on the terminologist's knowledge of the domain but also involves a fair degree of subjectivity. For this evaluation I relied on my experience in practical terminology work, however, I am aware that the subjective factor could not be eliminated completely.

7 Future Work

There is much interesting work that remains to be done to improve this ATR program and extend its functions. As a start, I propose the following enhancements:

- *Bilingual ATR tool*: In this work the focus was on candidate term search within monolingual texts. The idea to implement an ATR tool, however, was motivated by the needs of terminology work in translation. The implementation of a bilingual program involves two steps. First a linguistic examination of source and target language is needed, for example according to the method presented in this thesis. Then, term equivalents are retrieved by means of alignment techniques (for a discussion of alignment methods see Blank 1998:31-60).
- *Context retrieval*: Theoretically, the difference between terms and non-terms is that the former are independent of context. In practice, however, translators are not subject specialist and they do not know the exact concept of every term. In practical terminology work, in order to provide a better understanding of the terms, they are usually entered in the data base with their context. The most obvious limitation of automatically generated term lists is that they remove the terms from their original contexts. A valuable addition to the ATR program presented in this thesis would be the retrieval of contextual information for the candidate terms.
- *Stop list*: Despite the danger of using stop lists (section 6.3.2) the user should be given the option of defining words which the program has to ignore during the extraction process.
- *Domain exchange:* The method which was used for analysing banking language can be applied to other terminology databases and deliver the syntactic and morphological resources of any domain.
- *User interface*: The program has served for research purposes. No utilities are available to make the process of term extraction easier. The tool could be provided with a graphical user interface which would allow the user to extract, view, and edit terms by means of simple operations.

8 Conclusion

The goal of this thesis was to explore the methods and principles of terminology, to investigate the characteristics of terms in respect to non-terms, and to implement a method for automatic terminology retrieval. In section 1 of the first part of this thesis, it was recognised that our multi-lingual world depends increasingly upon the accurate translation of a wide range of specialised documents. The need for efficient and cost-effective specialised translation presents a challenge to terminologists, who have to provide translators with the appropriate terminology. The increasing need for terminology calls for automatic term extraction methods which provide a quick retrieval of terms.

In section 2 it was shown that, from a theoretical point of view, the most important criterion for identifying terms is the notion of concept, which is fixed in definitions. A further criterion for delimiting terms from non-terms is the context independency of the former. The conceptual character of terms and their context independency can easily be apprehended by a human being but are extremely difficult to render in computer applications.

Due to the difficulties of encoding conceptual knowledge in computers, the task of ATR consists in understanding how the structure of terms differs from that of non-terms. Insights in this respect are provided by studies on language for special purpose. In section 3 LSP was analysed at the lexical and syntactic level. First rules of term formation were dealt with. It was shown that terms formed by compounding, phrasal composition, and derivation are very frequent in LSP and that these provide a number of criteria for ATR. Other patterns of term formation which can be exploited for ATR are similes and compression. Often noun terms are formed by conversion from verbs, which for ATR is also an evidence of termness. The syntactic characteristics of special language are mostly accounted for at the text or sentence level. At the phrase level LSP studies single out the preference for noun phrases in specialised texts.

Section 4 concentrated on a sub-language in the field of economy, namely banking language. This special language has received little attention within the domain of LSP. Studies focus mostly on the etymological and socio-historical aspects of banking language. In this thesis the linguistic aspects of banking language were examined with the aid of the terminology data base CS-term.

In the second part of this thesis, section 5 discussed researches in the field of statistics and linguistics related to ATR. In a statistical approach termness is determined by the frequency of words and word-classes. In linguistic approaches termness is suggested by the nominal character of words, by special PoS collocations, by the identification of domain-specific morphemes, or by the presence of contextual information.

Section 6 presented the implementation of a self-developed ATR program and discussed its performance. Firstly, the syntactic and morphological characteristics of the entries in the terminology data base CS-term were investigated. Term extraction grounded on syntax was based on the assumption that the entries in CS-term have typical structures. The analysis of CS-term showed that this assumption does not hold true. In fact, the majority of terms are single word terms, which is not sufficient to isolate them from general words. From a morphological point of view, banking specific lemmas are recurring in CS-term entries, which is a useful criterion for the automatic identification of banking terms.

Secondly, section 6 explained the computational framework used in the process of ATR. The first component in the ATR process is the NP-tool. The computer modules included in the NP-tool are a tokeniser for the detection of sentence and word boundaries, a morphological analyser for the segmentation of words, and a parser for the retrieval of the syntactic structure. The second component is the lexicon-lookup. In this module, regular expressions processing the NP-tool output filter only those NPs which are formed by banking-specific lemmas. The domain-specific lemmas are stored in a lexicon which was generated from the analysis of CS-term. The final output of this ATR process is an alphabetically-sorted list of candidate terms.

Thirdly, section 6 discussed the results of the NP-tool and lexicon-lookup. The NP-tool has a high recall rate but a very poor precision. This is an indication that syntactic criteria are not sufficient to isolate terms. The lexicon-lookup shows a poorer recall rate than the NP-tool. However, while it misses some terminology typical of the domain it has an excellent precision rate. The recall rate of the lexicon-lookup could be improved by entering missing domain-specific lemmas. However, while every additional domain lemma adds to the retrieval of more terms, it represents also a risk of increasing the error rate.

In this thesis it was repeatedly stressed that in order to communicate across the language barriers the easy accessibility of terminology is of utmost importance. Computational linguistics has recognised this need and numerous contributions in this field are changing terminology work in many ways. Many programs for the automatic retrieval of terminology provide the terminologist with candidate terms. With the development of reliable tools the efficiency and consistency in terminology work will certainly increase compared to manual methods. The problems in automatic terminology retrieval can be overcome by implementing tools which rely more strongly on domain-specific linguistic resources.

The advance of terminology programs also requires that terminologists acquire a better knowledge of the computational methods behind these tools. Terminologists should no longer consider computer tools to carry out the process of term extraction independently. Introducing terminologists to the problems and principles of automatically processing language would increase their understanding of the sometimes poor performance of terminology tools. On the other hand, computer experts should get familiar with the needs of terminologists, should integrate these needs into the tools they program. By being aware of the problems of finding terms in documents a software engineer may perceive new ways for improving the performance of a tool. A constructive attitude of both parts would lead not only to mutual enrichment but also to progress in the respective work. This is important for an effective retrieval and management of large amount of terminology and for a successful solution of the present and future demands in terminology.

Appendix: List of GERTWOL-Tags

A	adjective
A(PART)	adjectival participle
ABK	abbreviation
AdjP	adjective phrase
AKK	accusative
ART	article
CONJ	conjunction
DAT	dative
DEF	definite
DEM	demonstrative
ERSTGLIED	first part
NEUTR	neuter
NOM	nominative
POS	positive
PP	prepositional phrase
PRON	pronoun
RELAT	relative
S	noun
S(A)	nominal adjective
S(PART)	nominal participle
SELTEN	rare form
SG	singular

Bibliography

Ahmad, K., Davies, A., Fulford, H., Rogers, M. (1994). What is a term? The semi-automatic extraction of terms from text. In: Snell, M., Pöchhacker, F., Kaindl, K. (eds.) *Translation Studies: An interdiscipline*, pages 267-278. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Arntz, R., Picht, H. (1995). *Einführung in die Terminologiearbeit*. Hildesheim, Zürich, New York: Georg Olms Verlag.

Arppe, A. (1995). Term extraction from unrestricted text. In: *The 10th Nordic Conference on Computational Linguistics*. Helsinki: Department of General Linguistics, University of Helsinki.

Aussenac-Gilles, N., Bourigault, D., Condamines, A., Gros, C. (1995). How can knowledge acquisition benefit from Terminology?. In: *9th Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff.*

Blank, I. (1998). *Computerlinguistische Analyse mehrsprachiger Fachtexte, CIS-Bericht-98-109*. Ludwig Maximilians-Universität, München.

Boguraev, B., Kennedy, Ch. (1998). Applications of term identification technology: domain description and content characterisation. In: *Natural Language Engineering*, 5 (1), pages 17-44. Cambridge University Press.

Cabré, M.T. (1999). *Terminology: Theory, Methods and Applications*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Church, K., Dagan, I. (1994). *Termight*: Identifying and translating technical terminology. In: *Proceedings of the 4th Conference on Applied Natural Language Processing*, *Stuttgart*, pages 34-40.

Condamines, A. (1995). Terminology: New needs, new perspectives. In: *Terminology*, 2 (2), pages 219-238. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Covington, M. (1994) *Natural Language Processing for Prolog Programmers*. New Jersey: Prentice-Hall.

Crystal, D. (1997). *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell Publishers Ltd.

De Bessé, B. (1997). Terminological Definitions. In: Wright, S.E., Budin, G. (eds.) *Handbook of Terminology Management. Volume 1*, pages 63-74. Amsterdam, Philadelphia: John Benjamins Publishing Company.

De Bessé, B. (1990). La définition terminologique. In: Chaurand, J., Mazire, F. (eds.): *La Définition. Actes du colloque la définition, Paris, 18-19 novembre 1988*, pages 252-262. Paris: Larousse.

Desmet, I., Boutayeb, S. (1994). Terms and words: Propositions for terminology. In: *Terminology*, 1 (2), pages 303-325. Amsterdam, Philadelphia: John Benjamins Publishing Company.

EAGLES Lexicon Interest Group. (1998). EAGLES: Preliminary Recommendations on Semantic Encoding, Interim Report. http://www.ilc.pi.cnr.it/EAGLES96/rep2/rep2.html

Finegan, E. (1994). *Language, its Structure and Use.* Orlando: Harcourt Brace & Company.

Fluck, H.-R. (1996). *Fachsprachen: Einführung und Bibliographie*. Tübingen, Basel: A. Franke Verlag.

Gläser, Rosemarie (1988). Neulogismen im englischen Wortschatz der Ökonomie und Probleme der Übersetzung. In: Bungarten, Theo (eds.): *Sprache und Information*, pages 172-182. Tostedt: Attikon Verlag.

Heid, U. (1999). Extracting terminologically relevant collocations from German technical texts. In: *Proceedings of the TKE '99 International Congress on Terminology and Knowledge Engineering, Innsbruck, August 1999*, pages 241-255. Frankfurt: Indeks.

Heid, U., Jauβ, S., Krüger, K., Hohmann, A. (1996). Term extraction with standard tools for corpus exploration - Experience from German. In: *Proceedings of TKE '96 International Conference on terminology and Knowledge Engineering*, pages 1-11. Frankfurt: Indeks.

Hahn von, W. (1983). *Fachkommunikation. Entwicklung, Linguistische Konzepte, Betriebliche Beispiele.* Berlin: Walter de Gruyter.

Hoffman, L. (1985). *Kommunikationsmittel Fachsprache: Eine Einführung*. Tübingen: Gunter Narr Verlag.

Hundt, Markus. (1995). *Modellbildung in der Wirtschaftssprache: Zur Geschichte der Institutionen- und Theoriefachsprachen der Wirtschaft*. Tübingen: Max Niemeyer Verlag.

Justeson, J.S. and Katz, S.M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. In: *Natural Language Engineering*, 1 (1), pages 9-27. Cambridge University Press.

Kageura, K. (1995). Towards the theoretical study of terms - A sketch from the linguistic viewpoint. In: *Terminology* 2 (2), pages 239-257. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Laurén, C., Picht, H. (1993). *Ausgewählte Texte zur Terminologie*. Wien: Termnet.

L'Homme, M.C., Benali, L., Bertrand, C., Lauduique, P. (1996). Definition of an evaluation grid for term-extraction software. In: *Terminology,* 3 (2), pages 291-312. Amsterdam, Philadelphia: John Benjamins Publishing Company.

The New Shorter Oxford English Dictionary. Volume 1 and 2. 1993. Oxford: Oxford University Press.

Reinhardt, W., Köhler, C., Neubert, G. (1992). *Deutsche Fachsprache der Technik.* Hildesheim, Zürich, New York: Georg Olms Verlag.

Roelcke, T. (1999). *Fachsprachen*. Berlin: Erich Schmidt Verlag GmbH.

Sager, J. (1997). Term Formation. In: Wright, S.E., Budin, G. (eds.): *Handbook of Terminology Management. Volume 1*, pages 25-41. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Sager, J. (1990). *A Practical Course in Terminology Processing*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Sager, J., Dungworth, D., McDonald, P.F. (1980). *Englisch Special Languages. Principles and Practice in Science and Technology*. Wiesbaden: Oscar Brandstetter Verlag.

Salton, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Reading: Addison-Wesley Publishing Company, Inc.

Schefe, P. (1975). *Statistische syntaktische Analyse von Fachsprachen mit Hilfe elektronischer Rechenanlagen am Beispiel der medizinischen, betriebswirtschaftlichen und literaturwissenschaftlichen Fachsprache im Deutschen*. Göppingen: Verlag Alfred Kümmerle.

Stellbrink, H.J. (1988). Effizienz und Produktivität in einem Fremdsprachendiest in der Industrie. In: Bungarten, T. (eds.): *Sprache und Information*, pages 196-205. Tostedt: Attikon Verlag.

Volk, M. (1999) Choosing the right lemma when analysing German nouns. In: *Multilinguale Corpora: Codierung, Strukturierung, Analyse. 11. Jahrestagung der GLDV,* pages 304-310. Frankfurt.

Voutilainen, A. 1993. Nptool. A detector of English noun phrases. In: *Proceedings of the Workshop on Very Large Corpora*. Columbus, Ohio: Ohio State University.