A CORPUS LINGUISTICS STUDY OF TRANSLATION CORRESPONDENCES IN ENGLISH AND GERMAN

by

ALEKSANDAR TRKLJA

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of English
The University of Birmingham
November 2013

ABSTRACT

This thesis aims at developing an analytical model for differentiation of translation correspondences and for grouping lexical items according to their semantic similarities. The model combines the language in use theory of meaning with the distributional corpus linguistics method. The identification of translation correspondences derives from the exploration of the occurrence of lexical items in the parallel corpus. The classification of translation correspondences into groups is based on the substitution principle, whereas the distinguishing features used to differentiate between lexical items emerge as a result of the study of local contexts in which these lexical items occur. The distinguishing features are analysed with the help of various statistical measurements. The results obtained indicate that the proposed model has advantages over the traditional approaches that rely on the referential theory of meaning. In addition to contributing to lexicology the model also has its applications in practical lexicography and in language teaching.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Professor Wolfgang Teubert and Dr. Paul Thompson for their thoughtful and constructive supervision. I am indebted to Biman Chakraborty, Gabriela Saldanha, Geoff Barnbrook, Mary Snell-Hornby, Nick Groom, Oliver Mason, Pernilla Danielsson, Susan Hunston and my fellow research students for many discussions and useful advices. I also want to thank my family and friends and in particular my wife Barbara N. Wiesinger for her continued support during this long project.

Contents

List of Tables

List of Figures

List of Tables in Appendix

List of Figures in Appendix

Chapter 1 Introduction

1.1 This thesis

Imagine a car mechanic being captured by aliens who had crashed on Earth due to engine troubles. The aliens promise to release her only after she has repaired their rocket so that they can leave the Earth. She was chosen because of her skills. She has her car mechanic's tool kit with her so the situation is maybe not completely desperate. However, when she tries to use her wrench and put it on the socket to loosen the clutch bolts, the car mechanic realises that her tools are not suitable for repairing the spacecraft. The aliens provide her a tool box with suitable tools. However, she has no experience in working with such tools and she has never before seen such an engine. All the tools and the engine parts look the same to her, equally foreign and strange. Given that there is no manual available she needs to acquire the knowledge about using the tools from scratch. But how can she achieve this goal?

The situation is similar to that in which one tries to communicate in a language she has no or very limited knowledge of. No matter how skilful she is in her mother tongue, as long as she is not familiar with words and grammar from the foreign language they will all look the same to her. The comparison of language with tools was first proposed by Wittgenstein (1953: 11):

> "Think of the tools in a tool-box: there is a hammer, pliers, a saw, a screw-driver, a ruler, a glue-pot, glue, nails and screw. The functions of words are as diverse as the functions of these objects. (And in both cases there are similarities.) Of course, what confuses us is the uniform appearance of words when we hear them spoken or meet them in script and print. For their *application* is not presented to us so clearly." (italics in original)

Words and other linguistic units are like tools in that they have numerous functions and until the functions are known to us we cannot do much with these linguistic units. It is only after

their proper applications are presented clearly to us that they become intelligible. One of the ways to present these functions is by relying on our prior skills. Thus, in the above anecdote the car mechanic has a chance to survive if she recognises the functions and applications of the tools she needs to use and of the rocket engine. The car mechanic's knowledge is important because it can help her to create some conception of tools and of engine. Similarly, one of the ways of learning words and grammar constructions in a foreign language is by using knowledge of the mother tongue (Butzkamm and Caldwell, 2009). As Snell-Hornby (1987: 164) remarks:

> "experience in advanced language teaching and in translation teaching shows that the learner can understand a foreign language text better if unknown words are explained in terms of their own language system and against their sociocultural background without being rendered as foreign language equivalents which are often inadequate and contrived."

This task is usually supported by using bilingual dictionaries. As various studies indicate (Abu-Samak, 1996; Atkins and Knowles, 1990; Baxter, 1980; Nord, 2002; Tomaszczyk, 1979; Yong and Peng, 2007) learners prefer using bilingual to monolingual dictionaries. However, according to Atkins and Knowles (1990) monolingual dictionaries in practice prove to be more useful in helping users to understand the use of a word from a foreign language than bilingual dictionaries. This is because bilingual dictionaries usually lack the information regarding the context in which words are used and because they are based on single words. Thus what Durrell (2000: x) says in the context of German-English dictionaries is true for bilingual dictionaries in general:

> "Conventional bilingual dictionaries are often little help here, as they frequently give a fairly undifferentiated list of possible German equivalents for a particular English word without providing much detail on how those German

> equivalents are actually used or the types of context where one might be preferred to another."

So, it can be say that dictionary producers "shift the burden of choice to the user of the dictionary" (Martin, 1967: 56). This is not a minor burden given the fact that full synonymy is extremely rare (Cruse, 1986: 290). Similarly, the equivalence assumption presupposes the existence of some static and absolute *tertium comparationis* in relation to which "universal concepts are simply given different labels in various languages" (Snell-Hornby, 1987: 160). But, although usually taken for granted, the assumption about the existence of the universal language has never been proved (Teubert, 2010). For this reason, the basic issue of equivalency needs to be approached from a new angle by replacing "the principle of elementary approximation… by the principle of differentiation" (Snell-Hornby, 1987: 170).

The present thesis shares this view that the mother tongue can be a helpful resource for acquiring the foreign language and that bilingual dictionaries can contribute to this task significantly. In this thesis it will also be assumed that the principle of differentiation brings more advantages than the conventional approximation principle. The studies conducted in the thesis rely on the language in use theory of meaning which is discussed in detail in Chapter 3. The main objective of the research is to provide a corpus-informed, statistically-founded approach to the development of the differentiation principle that can be applied to practical bilingual lexicography. In addition to this practical purpose, the thesis also has a theoretical aspect: it aims at broadening our understanding of the structure of vocabulary from a cross-linguistic perspective. The thesis is situated at the intersection between contrastive lexicology, corpus-linguistics and translation study. In order to understand how the differentiation principle can be applied we need to briefly consider types of bilingual dictionaries

1.2 Bilingual lexicography and onomasiological dictionaries

In both monolingual and bilingual lexicography we can distinguish between two types of dictionaries: semasiological and onomasiological. The mainstream approach to bilingual lexicography is a semasiological one rather than onomasiological (Hüllen, 1999). The difference between the two is that "[whereas] a semasiological perspective investigates which concepts are associated with a given word, onomasiological research takes its starting-point in a concept, and investigates which words may be associated with that concept" (Geeraerts, 2003: 84). Words in the latter type of dictionaries are grouped according to thematic domains. Both approaches have developed side-by-side through history. For example, out of 1858 dictionaries published between 1467 and 1600 for German and other languages 475 were onomasiological dictionaries (Claes, 1977). One of the most important works in the history of English lexicography, *Ælfric's Glossary,* was also a topical or onomasiological dictionary for the language-pair Latin - Old English. It was compiled in the 10th century and organised around groups such as *birds, fish, animals, plants, trees* (Sauer, 2008: 34). Some words were ordered hierarchically "starting with the higher and proceeding to the lower, thus: 'God'–'angel'–'man'; or 'lord'–'servant, slave'" (Sauer, 2008: 34). Jacob Schöpper's *Synonyma* which was published in 1550 and which exerted a large influence on German lexicography was also a topical dictionary. It contained 34 classes labelled in Latin such as: *Memoria*, *Intellectus*, *Voluntas*, *Deus*, etc. Each group was further divided into subgroups. *Deus*, for example, consisted of *Diuinitas*, *Trinitas*, *Creare*, *Seruare*, etc. (Hahn, 2002: 61-67). These dictionaries were produced mainly for pedagogical purposes and served to help learners understand Latin texts and write in Latin. Some of these dictionaries were surprisingly innovative and continued to be used for a century or more.

To give one example of a historical bilingual onomasiological dictionary I will briefly discuss Comenius' *Orbis Sensualium pictus* which was initially prepared for Czech and Latin but was subsequently translated into various European languages. It became a standard learners' dictionary in its time and between 1631 and 1674 there were twenty-five different editions for English alone (Hüllen, 2009: 35). In this dictionary the terms were defined in the

form of a full-sentence which is akin to the full-sentence dictionary definitions developed for the Cobuild dictionary four hundred years later (Barnbrook, 2002). Such definitions are provided both in a mother tongue and a learning language. Illustrations were used for the terms that can be depicted. Figure 1.1 displays a class of words related to the Latin term *Piscatio* and its English cognate *Fishing* that belong to the topic *Senses* (*external and internal*). This is one of twenty topics into which the vocabulary is divided. So, we can see that the term *fisherman* is defined in terms of what a person does (*catcheth fish*) and this activity is further specified by means of other words. This type of description makes it possible eventually to understand all words considered to be relevant to *Fishing*. As one can see in the picture, the terms being defined are given in italic type. Some words are additionally defined through ostensive definitions by being represented in the form of objects on the image. A very important feature of this dictionary in the context of the present thesis is that the terms were explained by means of translation of complete sentences. This is very similar to the aligned sentences that form parallel corpora. Being given in sentences the terms are explained in the context in which they occur.

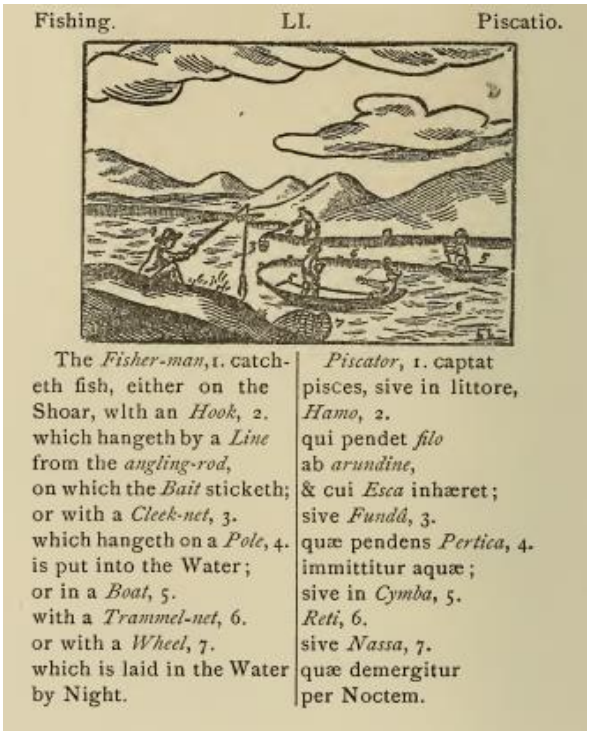


Figure 1.1: A dictionary entry from Comenius' *Orbis Sensualium pictus*

Nowadays, the discussion about advantages and disadvantages of semasiological and onomasiological dictionaries initiated by Leibniz has almost disappeared from contemporary lexicography (Hüllen, 1999). Onomasiological topical dictionaries are marginal in contemporary bilingual lexicography and alphabetically ordered dictionaries dominate both the market and the field (Goddard and Thieberger, 1997). However, there does not seem to be any purely lexicographically justifiable reason for this dominance.

First, onomasiological dictionaries have advantages for the task of language production. Hartmann (1983) observed that 75% of dictionary users need a dictionary for writing purposes. In this context learners have an idea what they want to say but do not know suitable words. For this purpose they usually do not need a dictionary that would define the meaning of a word that they know but instead they need a word that has a meaning they want to express. Onomasiological dictionaries have proved to be more useful in such a situation (Sierra and McNaught, 2000).

Second, conventional bilingual dictionaries have not followed the tremendous development that took place in the context of learners' dictionaries since the advent of corpus lexicography. The beginnings of corpus lexicography are linked to the Cobuild project in which John Sinclair (1987) acted as a principal researcher. The first edition of the Cobuild dictionary was published in 1987 and thereafter the use of corpora in compiling dictionaries has become a norm. "After Cobuild, the entry had additionally to be consistent with corpus data; if an entry did not fit the data satisfactorily, it needed reformulation, just as if theories do not fit data, they have to be replaced" (Moon, 2009: 457). The first bilingual corpus-based dictionary was Oxford-Hachette's English-French, French-English Dictionary (OXFA) published in 1994 (Roberts and Montgomery, 1996). Nevertheless, bilingual dictionaries have not experienced such profound changes as learners' dictionaries.

> "[T]he more innovative [dictionaries] may introduce a few new types of information (corpus frequency are the flavour of the month), but when it comes to setting out the meanings of words giving them definitions or equivalents in another language, including examples, idioms, pronunciations,

> usage notes, cross-references and the score or so of other kinds of information, tradition rules supreme. Most dictionaries are sublimely unaffected by the highly relevant work currently being done by linguists, especially in lexical semantics. The dictionary of the present is at heart little different from the dictionary of the past." (Atkins, 1996: 1)

However, it is questionable whether significant advancements are possible at all so long as bilingual dictionaries are based on the approximation principle. On the other hand, for a realisation of the differentiation principle, according to Snell-Hornby (1990), we need a completely new approach. The type of onomasiological dictionaries that she proposes would be based on semantic fields. In general, such dictionaries:

> "should rather aim at pinpointing the focal components of the lexeme concerned and at situating it both paradigmatically (or intralingually) and contrastively (or interlingually), i.e. both against other items in the semantic fields concerned and in contrast to similar items in the target language." (Snell-Hornby, 1990: 222)

This means that in such a dictionary a word would be explained both in terms of its synonyms and its translation equivalents. This remark will be used as a starting guiding principle in the present thesis for the development of a model for differentiation between lexical items. In addition, the model will also rely on the recent development in the corpus lexical semantics mentioned above by Atkins. Parallel corpora will be crucial for the identification of corresponding items in two languages. In addition, the model will also follow recent developments in the area of contrastive lexical semantics.

1.3 Contrastive corpus studies

Contrastive language studies were traditionally concerned only with the exploration of grammatical systems and sub-systems across languages. However, since the 1980's the focus has shifted from *langue* to *parole*; from language as a system to language in use. Simultaneously, in addition to the investigation of grammatical issues more and more studies concerned with lexical issues started to appear. As a result, contrastive studies became more related to lexicological and translation issues and more relevant for lexicographical issues (Hartmann, 2007: 43-44). The use of corpora gave this trend a further boost. Through corpora, in particular parallel corpora, researchers became aware of phenomena that had previously gone unnoticed. With the help of corpora it also became easier to explore "how languages differ, what they share and – perhaps eventually – what characterises language in general" (Johansson, 2007: 1).

The use of corpora in contrastive studies makes it possible to deal with a range of different issues. Aijmer and Altenberg (1996: 12) summarise the main applications of corpora to contrastive linguistic studies in the following way:

> • "they give new insights into the languages compared - insights that are likely to be unnoticed in studies of monolingual corpora;
> • they can be used for a range of comparative purposes and increase our understanding of language-specific, typological and cultural differences, as well as of universal features;
> • they illuminate differences between source texts and translations, and between native and non-native texts;
> • they can be used for a number of practical applications, e.g. in lexicography, language teaching, and translation."

For the purposes of the present thesis the first and fourth points are most relevant. In relation to the former, the thesis will demonstrate that the correspondence relations

between lexical items in two languages can be established purely by exploring their distribution in parallel corpora. The study of the context in which corresponding units occur will help to broaden our knowledge about lexical relations between sets of words in two languages.

As stated above, the thesis aims at providing a new descriptive model for dealing with lexical items across languages. In this sense, it differs from the majority of contemporary contrastive studies (e.g. Altenberg, 1999; 2002; Butler, 2008; Granger, 1996; Hasselgard, 2004; Johansson, 2003; 2007; Viberg, 2002; 2004) which are usually atomistic and deal with singular issues.

1.4 Language in use theory of meaning

The methodology used in this thesis is based on the language in use theory of meaning (Geeraerts, 2010: 164-178). This theory, which was chiefly defined by Wittgenstein and later adopted in linguistics by Firth and Sinclair (see Chapter 3), is an alternative to referential theories of meaning. Unlike the latter theories in which the meaning of a word is defined in relation to what it signifies, in the former the meaning is studied by focusing on a term's use in a specific context. This context may be either textual or situational. In the present thesis only the former will be explored. It must be stressed that I consider the latter equally relevant and that only the study of both contexts can provide a comprehensive description of the meaning of a term. Due to space restrictions the situational context cannot be dealt with here. The textual context will be studied by means of corpora. Corpora are suitable for this task because they "consist of traces of linguistic behaviour. What a corpus gives us is the opportunity to study traces and patterns of linguistic behaviour" (Hanks, 2008: 130). They therefore provide direct access to the real occurrence of lexical items.

One of the main insights of corpus research is that lexis like grammar is full of regularities and patterns. However, unlike in grammar these regularities are treated more as tendencies than law-like rules. At the same time, corpus research indicates that language is

full of variations but "in principled ways, which are at present imperfectly understood" (Hanks, 2008: 128). The task of studies that aim at providing descriptions of these tendencies is to improve our understanding of variations and patterns. This also means that these tendencies can be quantitatively explored and measured. In the current thesis the differentiation principle will be combined with various statistical measurements of the patterns which characterise the use of words in context.

In corpus linguistics one usually distinguishes between a corpus-based and a corpus-driven approach (Tognini-Bonelli, 2001). The former uses corpora to study phenomena by applying already existing categories and to test previous studies. As was mentioned in the previous section, contrastive linguistic studies usually use corpora in this manner. The corpus-driven approach, on the other hand, is concerned with the study of rough data and categories and findings here, therefore, emerge from the direct observation of data. The current thesis combines both approaches. Alongside part-of-speech categories that will be used in analyses some new categories will be created to describe specific occurrence of lexical items in raw data. The above distinction can also be understood in terms of the deductive vs. inductive method (Groom, 2007). If we accept this distinction the current thesis would draw from both methods but would also go beyond them and be closer to the abductive method which is a form of "thinking from evidence to explanation, a type of reasoning characteristic of many different situations with incomplete information" (Aliseda, 2006: 28). A purely inductive approach is impossible because there are no observations that are completely theory-free (Mahlberg, 2005: 2). For this reason, the abductive method seems more realistic because following this method we acknowledge that "an abductive explanation is always an explanation with respect to somebody of beliefs" (Aliseda, 2006: 30). In addition, "abduction is connected to both hypothesis construction and hypothesis selection" (Aliseda, 2006: 33). However, it is possible to control how heavily we rely on the previous beliefs and how much we assume prior to observations. Sinclair suggested a minimal assumption approach according to which at the beginning: "[w]e should only apply loose and flexible frameworks until we see what the preliminary results are in order to accommodate the new information that will come from the text" (Sinclair, 1994: 25). This view will be accepted in the present

study. Apart from relying on the language in use theory of meaning the study will not follow any specific linguistic theory such as cognitive linguistics or systemic-functional linguistics.

1.5 Research questions

In addition to the general objective stated in 1.1, the thesis also addresses the following two specific research questions:

- Is it possible to generate thematic sections to be used in onomasiological bilingual dictionaries through an investigation of textual contexts in which corresponding lexical items from two languages occur in a parallel corpus?
- Can we arrive at purely distributional distinguishing features for synonymous lexical items from L1 that share the same translation equivalents in L2 by exploring their occurrence in corpora from both an intralingual and interlingual perspective?

1.6 Outline of the thesis

Chapter 2 reviews previous cross-linguistic studies that deal with the classification of lexical items into semantic groups. First, traditional semantic field studies will be discussed, and then some more recent approaches. The assumptions, methodology and results relevant for the current thesis will be critically examined and discussed.

Chapter 3 introduces the theoretical and methodological framework on which the subsequent analyses will be based. Here, I begin with a general discussion of the language in use theory of meaning and then connect it to two specific linguistic approaches.

Chapter 4 is concerned with the first research question and sets out the methodology of identifying units of analysis and translation lexical domains by means of a parallel corpus.

Chapter 5 is designed to answer the second research question. Here, the distribution of lexical items belonging to the same lexical domain will be considered both from an intralinguistical and interlinguistical perspective.

Chapter 6 and Chapter 7 also deal with these two specific research questions and serve to test the findings obtained in the previous two chapters with a new set of data.

Chapter 8 summarises findings and discusses their significance with regard to theoretical and practical contributions of the thesis. In this chapter the limitations of the present study and suggestions for further research are discussed.

Chapter 2 Previous approaches

2.1 Introduction

This chapter provides an overview of how previous cross-linguistic studies approached relationships between words in two or more languages, discrimination between semantically related words, equivalence relations and the structure of lexicon. The chapter is divided into four sections.

In the first section two studies from the traditional lexical field theory based on componential analysis are discussed. These studies are included here for two reasons. First, the principle on which the classification of lexical items in the current thesis relies was partly developed in opposition to the model proposed by lexical field theorists. Second, componential analysis was the dominant approach to both monolingual and bilingual studies of lexicon and its achievements and limitations need to be acknowledged if for no other reason than to avoid repetitions of mistakes and failings. The second section deals with Fillmore's frame semantics which naturally builds on the lexical field theory; it emerged partly as a reaction to the isolationist view of words in the former theory. Unlike componential studies the approach of frame semantics is still very much alive as the ongoing FrameNet (e.g. Baker et al., 2003) project illustrates. Its potential for remedying the shortcomings of conventional bilingual dictionaries and helping to improve learners' monolingual dictionaries has been stressed on many occasions (e.g. Atkins, 1996; Atkins et al., 2003; Fillmore and Atkins, 1998).

The second section discusses two methods developed by two Scandinavian contrastive linguists. Both Viberg and Dyvik, whose works are reviewed here, suggest an innovative way of studying lexis in contrast by using translation corpora. In spite of this general similarity the two approaches depart from each other with regard to the proposed methodologies.

The third section reviews two approaches concerned with the study of translation equivalents beyond the single word: an international project based on the method proposed by John Sinclair and Dirk Siepmann's Bilexicon which aims to create bilingual dictionaries on the basis of native-like expressions.

In reviewing these studies special attention will be paid to the principles that underlie lexical fields, the nature of equivalence relations, the structure of the lexicon and the method used for differentiating between translation equivalents.

2.2 Componential approaches to semantic fields

Although various authors at the beginning of the 20th century, including Ferdinand de Saussure, discussed the notion of grouping lexemes into a semantic system, it was the German linguist Jost Trier (1931) who played the crucial role for the development of the notion of lexical fields (*Wortfeld*). His aim was to develop a theory of lexical fields that would deal with semantic changes and semantic relations. Subsequent studies showed that many of his assumptions were false or too general to be useful (Lyons, 1981). At the same time, various new methodologies and research questions appeared and the theory developed in different directions. What is common to different approaches is the idea that lexicon, language in general and relations between words can systematically be described by creating classes of semantically similar words. In its golden years (from 1960's to 1980's) the theory of lexical fields was dominated by componential analysis which was first introduced by Hjelmslev (1961). There are two branches of this approach. The first branch is related to the American componential analysis which follows the methodological principles suggested by Jerrold J. Katz and Jerry A. Fodor in their influential paper *The structure of a semantic theory* (1963). The second is the continental componential analysis based on Coseriu (1964; 1968). Common to both branches is that:

> "A minimal definition of the meaning of an item will be a statement of the semantic components necessary and sufficient to distinguish the meaning paradigmatically from the meanings of all other items in the language." (Bendix, 1971: 393)

Two studies which will be reviewed below are good illustrations of these two branches.

2.2.1 Adrienne Lehrer's analysis of *cooking* words

One of the most extensive studies in which the componential approach to lexical fields was applied is Adrienne Lehrer's book *Semantic fields and lexical structure* (1974). She adopts Trier's assumption that "the vocabulary of a language is structured, just as the grammar and phonology are structured" (Lehrer, 1974: 15) and combines it with Katz's componential analysis. The lexicon of a language, according to her, consists of word sets "which are related to conceptual fields and [that] divide up *semantic space* or the semantic domain in certain ways" (Lehrer, 1974: 15, italics in original). This is a very vague description of the structure of the lexicon given that the definition of word sets is based on the notion of conceptual fields of which we have only limited knowledge. Lehrer herself admits that "at this stage, the relationship between language and thought is still limited and must be considered as open one" (Lehrer, 1974: 17). In addition, it is unclear what she means by *semantic space*. From the above quotation it seems to be a synonym for the term *semantic domain* which itself remains undefined. Instead of a definition Lehrer only provides a list of some semantic domains borrowed from Lyons (1963; 1968): *a single text*, *the works of a single author*, *a single genre* and *texts which deal with the common subject matter*. It is difficult to see what all these examples have in common.

Like many other similar studies this work is also concerned with the description of relations between words that from a semantic field. To explore this issue she studied semantic properties of the words from the lexical field called *cooking*. Although all the words from this particular field are verbs she does not presuppose that all fields consist of only

words from the same word classes. Her opinion is that "it is useful to ignore parts of speech at times and contrast the meaning of items belonging to different word classes" (Lehrer, 1974: 197). The role of componential analysis comes to the fore in the meaning differentiation of words. The components are considered to be kinds of semantic primitives that seem to be part of human conceptualisation of the world. It remains unexplained how the author arrives at the components which describe the meaning of words from the field *cooking*. The most likely explanation is intuition. Ontologically, the components are not linguistic phenomena but physical properties of non-linguistic 'objects'. Thus, all components are related to different manners of cooking.

Lehrer identifies several semantic components that characterise the meaning of *cooking* words. Some of these components are *the use of water, the use of oil, cooking time, the use of cooking utensils*, etc. These components serve to differentiate between meanings of the words that belong to the semantic field under examination. We can illustrate this with the following example. The verb *boil,* according to this study, consist of the component cooking *with the use of water*, whereas the verb *fry* contains the component *cooking with the use of oil*. Similarly, *fry* is different from *sauté* because it does not have the component *the use of cooking liquid*.

Componential analysis plays a crucial role also when it comes to equivalence relations. For example, the German verb *kochen* corresponds both to *cook* and *boil* in English because it denotes both the general process of cooking and the process of cooking with water. Similarly, *braten* is an equivalent both to *fry* and *broil*, because its semantic components are both *the use of oil* and *no use of oil*. The comparison of semantic features, Lehrer claims, should help to establish semantic relations between languages and to specify semantic relations across languages. The case of *kochen* and *braten*, for example, shows that a word from one language may have a more general meaning than its correspondents from another language.

Another important issue that the author addresses is the structure of semantic fields. The structure was studied in terms of the generality of meanings of the words examined. This description also relies on componential analysis but the results were subsequently also tested

on informants. The structure of the semantic field *cooking* is schematically represented in Table 2.1

<table>
<tr><td colspan="8">cook</td></tr>
<tr><td rowspan="3">steam</td><td colspan="3">boil</td><td>roast</td><td colspan="2">fry</td><td colspan="2">broil</td><td rowspan="3">bake</td></tr>
<tr><td colspan="3">simmer</td><td></td><td>sauté</td><td>deep-fry
French-fry</td><td>barbecue
charcoal</td><td>grill</td></tr>
<tr><td>poach</td><td>stew</td><td>braise</td><td colspan="5"></td></tr>
</table>

Table 2.1: The semantic field *cooking* (adapted from Lehrer, 1974: 31)

According to the study, the meaning of *boil*, *steam*, *fry*, *broil*, *bake* and *roast* is subordinated to the meaning of *cook* which has the most general meaning. Similarly, *simmer* is subordinated to *boil* and *poach* to *simmer*. As we can see from the table above, the semantic field in question is not symmetrically structured. Thus, there are no hyponyms for the words *steam, roast* and *bake*. On the other hand, there are words with more specific meanings than *boil, fry* and *broil*. As the lexical items *deep-fry* and *French-fry* indicate, the borders between the senses of words are not always strict. These two words can mean both *fry* and *broil*. This example is also interesting because it shows that occasionally more than one word can consist of the same semantic components. Such words are considered to be synonyms.

Following the approach suggested by Berlin and Kay (1969) in their study of colour terms Lehrer further distinguishes between basic and peripheral words. In the above table the basic terms are displayed in the first and second rows. All other words are treated as peripheral. The basic words are supposedly more semantically general. Lehrer also assumes that there is a direct relation between the notion of subordination and the range of contexts in which a word occurs; "the more specific the meaning of the word, the fewer collocational possibilities there are" (Lehrer, 1974: 33). Unfortunately, no proof is offered for this assertion. I briefly examined this assumption by looking at the occurrence of the words from her semantic field in the BNC and ukWaC and it seems that the assumption is correct. For example, *cook* which is the most general term in the field is more frequent and occurs with a larger number of collocates than the words with more specific meaning such as *boil, fry* or *broil.* Similarly, *boil* which according to Lehrer's survey is the second most frequent term is more frequent and co-occurs with more collocates than the less general *simmer* and so on. So

there does seem to be a direct relation between frequency, generality of meaning and the number of collocates and this issue merits further investigation.

At first sight componential analysis seems attractive because it can handle the issue of meaning differences both from the perspective of one language and from a cross-linguistic perspective. Nevertheless, Lehrer's study suffers from shortcomings that are intrinsic to the componential approach in general and that have been repeatedly reported in various publications (e.g. Dixon, 1971; Geeraerts, 2010; Lyons, 1995; Van Roey, 1990). The first problem is related to the issue touched upon at the beginning of this section. Because of the ambiguously defined relationship between words and concepts the nature of semantic components and of their identification remains a mystery. Gordon's (2003: 2219) objection that "[t]here is nothing to suggest the existence of any objective or universally applicable means of establishing parameters for a componential analysis" applies also to Lehrer's study. It is through intuition, observation of the objects in the world and contemplation about mental objects that a researcher identifies these components. Like any other intuition-based study this one is also subjective and therefore prone to errors.

The next problem is that it is very doubtful if the approach can deal with words belonging to different semantic and grammatical categories. As Snell-Hornby (1990: 211) points out, componential studies are usually limited to words that refer to concrete objects, basic activities and stative adjectives. This is because it is relatively easy to identify semantic borders for such words "[b]ut many other vocabulary terms refer to 'things' which have features that are not neatly distinguishable, so that their meanings have 'fuzzy edges', i.e. contrast only vaguely and cannot be adequately described in terms of components" (Van Roey, 1990: 30). To give an example, consider the words *gleam*, *glisten*, *glitter* or *glow* or the lexical items *pose problems, cause problems, give rise to problems, create problems* that will be investigated in Chapter 5. Can we say that *gleam* is more central than *glisten* to the field of the light words? Or, would it make sense to say that *pose problems* has more general meaning than *cause problems*? Lehrer seems to be aware of this problem as she admits that "[the] lexical sets of words discussed were selected because they seemed amenable to the

field approach" (Lehrer, 1974: 201). This is a serious limitation because it restricts the analysis to specific types of words from the start.

The binary character of the sense components is another problematic issue. Such a restrictive representation of meaning does not assume that words from a field can have different shades of meaning. Similarly, it is not clear how many components are required to describe the meanings of words from the same field. Unlike in phonetics from where the componential analysis derives, in semantics we do not have a limited number of features that can be specified in advance and universally applied. As we study words from a new field new components emerge and the formerly used ones become irrelevant. Often a very large number of components must be listed which makes the approach uneconomic (Dixon, 1971: 441). Lehrer (1974: 201) admits that "the problem of determining the inventory of the lexical items in a field remains".

A final critique of Lehrer's and other similar component analysis approaches is related to the insistence on describing the meaning of words outside their context. Here, Lehrer follows Trier who himself did not consider syntagmatic relations to be important. At the time, this stance was heavily criticized by Porzig (1934) who suggested an alternative view that would incorporate syntagmatic relations into analysis. This view has been acknowledged but has never really been widely accepted by semantic field scholars. Lehrer seems to be aware of the restrictive character of collocates that occur with particular words as she devotes a whole chapter in her book to the relationship between grammar and lexicon but she fails to provide a plausible account of this relationship. All that is provided is a statement that "there is a tendency for words belonging to a field to share the same semantic restrictions" (Lehrer, 1974: 202) but without any attempt to explore this problem in detail.

The problem of focusing on single words is related to the above issues as well. In some cases this can paint a wrong picture such as when the theory deals with the issue of lexical gaps. But, as Durrell (1981) rightly points out, the ostensible lexical incongruence and lexical gaps that Lehrer spots in her study are in reality products of the componential method itself and not of the differences between languages. Thus, he shows that although German does not have a single word equivalent for the verb *simmer* one can express the same meaning by

means of a phrase *langsam kochen*. Therefore, the same meaning is lexicalised in different ways in two languages. And this is not an unknown phenomenon. According to Lyons (1977: 262) "[i]n many cases, one language will use a syntagm where another language employs a single lexeme with roughly the same meaning". Leherer admits that the approach might not be well equipped to deal with complex words, phrases and idioms. She concludes that "[t]his may turn out to be a fundamental mistake" (Lehrer, 1974: 201).

2.2.2 Karcher's study of *water* words in English and German

Karcher's (1979) contrastive study of English and German words from two corresponding semantic fields in English (*Water*) and German (*Gewässer*) follows the continental tradition of componential analysis. Components or *noems,* as he calls them, represent the conceptual and semantic core of words or sememes. He does not explain what these conceptual and semantic aspects are and, what's more, they seem to refer to the same phenomenon. The focus in this study is on the core elements and the author leaves the issue of non-core elements unexamined. The advantage of the method that Karcher applies over Lehrer's approach is that the members of his semantic fields are established not only by relying on intuition but also on reference books (monolingual dictionaries, thesauri and synonymy dictionaries). Similarly, equivalence relations are based on various bilingual German-English dictionaries and not only on the author's bilingual competence.

Karcher starts from the assumption that equivalence relations in two languages are more often many-to-many than one-to-one. In other words, for a word from L1 there will usually be more than one corresponding word in L2. According to him, componential analysis can help to explicate these relations and to find perfect matches or similarities and differences between corresponding terms. On the basis of 19 sense components, e.g. 19 binary ordered, extralinguistic properties which describe different qualities of *water* (e.g. *natural or artificial, flowing or stagnant, very large or very small, large or small, marshy or clear*) he reveals the structure of the semantic fields studied and the relationship between

equivalents in two languages. Each word that belongs to a semantic field in L1 is characterised by a specific configuration of components that cannot be found with other words. The same or very similar set of components is found with words that belong to the corresponding field in L2. Thus, it is only with the word *sea* from the English semantic field *water* that we find the following components: *natural, stagnant, maritime, steady* and *of large size*. The very same set applies to the German word *Meer* from the corresponding field *Gewässer.* No other German term is characterised by this particular constellation of components. Since there is a complete overlap between the components we can conclude that the two words perfectly match each other. Some other examples for the corresponding terms from two languages that rely on the collection of noems are: *Fluß* and *brook*, *Teich* and *pool*, *Kanal* and *canal*, *Strom* and *river*. According to this study, there is usually only one perfect match in another language. By establishing the configuration of semantic components words become monosemous and are reduced to equivalents that stand in one-to-one relationship to each other.

The issue of the structure of the vocabulary is not addressed in this study and the structure of semantic fields is touched upon only fleetingly. However, it seems that Karcher assumes that semantic fields follow a hierarchical organisation and that they can be divided into subfields. He describes the subfield *Wasserläufe*/*Watercourse* which is part of the more general semantic field *Gewässer*/*water* and is characterised by a constellation of the following noems: *flow, constant* and *inland waters*. This particular combination of components supposedly cannot be found with words belonging to other subfields.

Karcher does not stop here because his aim is to show that noems can help to explore the connotative meaning of words that belong to the same semantic field. According to him, the description of connotative meanings is central for a proper cross-linguistic analysis of differences between terms from the same field or subfield. According to Karcher, words mainly "differ in the connotative meaning of the meaning core" (Karcher, 1979: 149, my translation). Component analysis is here combined with the method of factor analysis and the study comprises several stages. For the purpose of the present thesis a somewhat simplified account will suffice.

At the beginning a set of connotational components or *attributes*, as they are called now, are established for all words from the English semantic field *Water* and the German field *Gewässer.* After that each attribute is loaded with specific values and native speakers of both languages are asked to fill in a survey by evaluating each word on the scale. Finally, the average values of factors are compared and the results show how similar or dissimilar the connotative meaning of words both from a monolingual and cross-linguistic perspective is.

To illustrate the approach three words (*brook, rivulet* and *river*) from the subfield *Watercourse* will be briefly considered. From the point of view of the denotative meaning these words are synonyms. However, differences between their senses emerge when we compare their connotative components. The highest values of the macro factors are associated with *river* which means that its meaning is associated more strongly than of other words with the following components: *being beautiful*, *positive*, *good*, *active*, *excitable*, *fast*, *strong*, *hard* and *masculine*. Similarly, *rivulet* is associated with *being active*, *being excitable* and *fast*. By comparing the assigned values of the components we also find differences between a word from L1 and its equivalent from L2. For example, the word *Pfütze* unlike its English correspondent *puddle* is associated with the following components *weakness*, *femininity*, *passivity* and *being bad*. The two words, therefore, have different connotative meanings.

Most criticism that was made in the context of Lehrer's study applies to Karcher's approach as well. Highly problematic in this study is the nature of the components (both denotative and connotative), the number of semantic components, their non-linguistic character that necessarily restricts the analysis to a specific set of lexemes, the binary character of semantic components, the focus on isolated words and ignorance of multi-word expressions, the lack of information regarding the syntactic framework in which words occur and the lack of real language data. Besides, the analysis of connotative meaning is highly abstract and clichéd. It is not clear for whom the word *rivulet* means activity and excitement. It seems that the author here assumes that this is a generally accepted interpretation by all users of one language. However, no attempt has been made to check this assumption on a representative sample. Additionally, the components *femininity*, *weakness* and *being passive*

follow the stereotypical gender representation and by no means can be treated as objective and purely analytic categories. Even more importantly, many connotative meanings are left unnoticed because the author does not analyse the use of words in context. For example, a brief exploration of the context in which the word *river* occurs reveals that the part of its emotional or connotative meaning is its co-occurrence with the prepositional phrase *of blood*. This seems to be specific only to this word as, according to the BNC, other words from the same field such as *brook* and *rivulet* do not collocate with this phrase. Finally, the author's aim to reduce all relations between words in two languages to one-to-one relations is too restrictive as it oversimplifies the issue of polysemy. For example, the German word *Strom* corresponds to the words that belong to the semantic field *Water* but it can also be used a translation equivalent of *power, electricity, flow, current* or *stream*.

2.3 Frame Semantics

The relationship between frame semantics and the theory of semantic fields is discussed in Post (1988). Fillmore describes this relationship in the following way:

> "The concept of semantic field can be captured by appealing to the notion of scheme, and the allied concept of vocabulary field can be identified with the notion of frame and with various linkages among frames." (Fillmore, 1977: 130-1)

Differences between two theories are more than just terminological. From the point of view of frame semantics the major flaw of the theory of semantic fields is that it fails to account for relations between components that characterise the meaning of words. It is precisely these relations which form the basis of frame semantics. Our knowledge about world around us and about mental concepts, according to this theory, is relational and semantic features or components represents values of the same attribute (Barsalou, 1992: 21-75). These relations are syntagmatic. For example, two features which characterise the frame for the noun *car* are

ENGINE and DRIVER and the relation between them is that DRIVER controls the ENGINE as in the following invented example.

She drove my old car.

Here, *she* is DRIVER and *my old car* is ENGINE. The names for attributes represent semantic roles. In frame semantics relations between semantic roles reflect our conceptualisation of the world which is why the frames are considered to have a cognitive foundation. This is the second important difference in relation to the theory of semantic fields, because frame semantics provides also a theoretical explanation for its lexical classification. The basic assumption in frame semantics is that the meaning of a word "can be understood only with reference to a structured background of experience, belief, or practices constituting a kind of conceptual prerequisite for understanding the meaning" (Fillmore and Atkins, 1992: 76-77). In other words, our understanding of words depends on our encyclopaedic knowledge which belongs to the realm of our cognitive functions. Semantic frames precede our understanding of meaning. We can grasp a word meaning only after we have already understood "the background frames that motivate the concept that the word encodes" (Fillmore and Atkins, 1992: 76-77). Or to speak in terms of frame semantics words evoke frames.

The theory provides an apparatus for the description of both grammatical and semantic environments in which words occur. The former environment is named *syntactic valence* which is described "in terms of the phrase types (e.g. noun phrase, prepositional phrase, etc.) of the possible complements, and in terms of the grammatical functions (e.g. subject, object, etc.) that the complements bear with respect to the word" (Fillmore et al., 2003: 236-237). The latter environment is described in terms of the concept called *semantic valence* which provides a description of frame-specific semantic roles (in the above example DRIVER and ENGINE). The sequence of frame elements reflects the semantic structure of a frame, or semantic relations from the syntagmatic point of view. Accordingly, the *Commercial_transaction* frame, for example, consists of the frame elements BUYER, SELLER, GOODS and MONEY (Fillmore and Atkins, 1992: 79). The following sentence annotated with

the frame elements is an example of this frame. The word that evokes a frame is called a *target word* and is labelled as *tg*.

[$_{<buyer>}$ Harry] *spent*^{tg} [$_{<money>}$ twenty dollars] [$_{<goods>}$ on a new tie].

For an accurate description of word meaning the information regarding both syntactic and semantic valence is needed. This is because frame elements can be differently distributed within one frame and because frame elements can have different grammatical realisations. The following table illustrates how a simple sequence of frame elements can acquire different grammatical forms.

	interlocutors	TARGET	topic
1	NP.Ext	argue.v	INI
2	NP.Ext	argue.v	PP_over.Comp
3	NP.Ext	argue.v	PP_about.Comp
4	NP.Ext	argue.v	PPing_about.Comp
5	NP.Ext	argue.v	Swhether.Comp

Table 2.2: A semantic and grammatical description of the verb *argue* (from Boas, 2002: 1367)

The theory of frame semantics was initially developed only for English but was subsequently also applied to contrastive language studies. The international DELIS project (e.g. Braasch, 1994; Heid, 1994; 1996) or Spanish, German and Japanese FrameNet data bases are a case in point. The aim of these projects was to provide a contrastive description of frames in several languages that would help to improve bilingual dictionaries. My discussion here will focus on a more recent study (Boas, 2002; 2005) of English and German frames.

The description of German data is based on the frames that have been identified in various studies for English and are stored in an electronic data base called *FrameNet* (Baker, Fillmore and Cronin, 2003). The English frames are in principle "understood as an independently existing conceptual system that is not tied to any particular language" (Boas, 2005: 466). At the beginning, English semantic frames are cleared from the information specific to English and are re-populated with lexical descriptions of German terms. After that,

German lexical units that evoke a particular semantic frame are identified in bilingual dictionaries. Their meaning is defined in terms of the frame elements associated with the given frame. The next step is to find "sentences that illustrate the use of each of the LUs [lexical units] in the frame" (Boas, 2005: 459) in a monolingual corpus. Corpora are used here for merely illustrative purposes. In the final stage, parallel frames are matched. Apart from bearing the same label the identified frames also consist of the identical sequence of frame elements which have similar grammatical forms. As a result, the frames that match both at the formal and semantic level are created. The following figure exemplifies how this was done with the English verb *argue* and its German equivalent *streiten*.

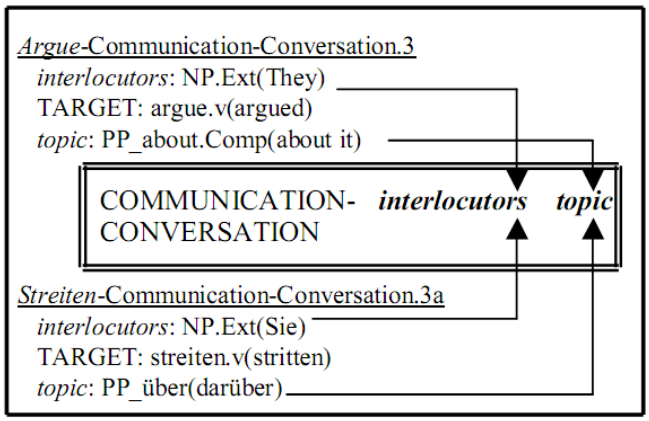


Figure 2.1: A frame semantics description of the English verb *argue* and its German equivalent *streiten* (from Boas, 2002: 1369)

According to this analysis, both words belong to the semantic frame COMMINICATION-CONVERSATION. The three shared frame elements are *interlocutors,* a *target word* and *topic.* The *interlocutor* role is realised in both languages as a noun phrase, the *target word* is a verb and the *topic* element is a prepositional phrase. The central part of the figure displays the common semantic elements.

The main advantage of frame semantics compared to the traditional semantic fields approach is that it provides a syntagmatic description of words that belong to the same frame. This is described both in grammatical and semantic terms. A serious limitation of frame semantics lies in the fact that semantic labels are based on *a priori* established categories. The problem is that the criteria that underlie these categories are not stable. For example, in the aforementioned analysis the verb *argue* belonged to the COMMUNICATION-

CONVERSATION frame whereas in the current version of FrameNet this meaning is described in terms of the QUARELLING frame. In cognitive psychology it is held that frames "are continually updated and modified due to ongoing human experience" (Evans and Green, 2006: 223). The reliance on pre-established categories is, therefore, associated with the risk of neglecting certain aspects of word meanings. As Hanks (2004: 6) notes, frame semantics "requires the researchers to think up all possible members of a Frame *a priori*, [which] means that important senses of words that have been partly analysed are missing and may continue to be missing for years to come." These omissions are discussed in Hanks (2004) and in Hanks and Pustejovsky (2005). One example discussed is the verb *toast* which is described only in terms of the *Apply_Heat* frame and it follows that frame semantics recognises only its cooking sense and neglects the celebrating sense. This is obviously an incomplete description of the term. It is questionable if it is possible at all to fix these problems without changing the fundamental principles of the theory. Hanks (2004: 6) is not very optimistic: "What is needed is a principled fix – a decision to proceed from evidence not frames. This is ruled out by FrameNet for principled reasons: the unit of analysis for FrameNet is the frame, not the word".

The application of frame semantics to contrastive analysis is also not without its problems. To begin with, it is very doubtful that English semantic frames applicable to different languages because of their universal character. As we have just seen, the interpretation of semantic frames changes over time even in the context of monolingual studies. There is no guarantee that the existent categories will not change in future studies. Additionally, the universal character of English categories overlooks the possibility that lexical units in other languages evoke frames that do not exist in English.

The insistence that lexical units should correspond both on grammatical and semantic level imposes serious limitations on the approach. The approach neglects the fact that the same or similar meaning in two languages can be realised through formally different grammatical constructions. This is actually a widely known fact in translation studies and contrastive linguistics since at least Catford's (1965) influential publication *A Linguistic Theory of Translation*. The advocates of contrastive frame semantics do not seem to be aware of this

phenomenon. It might have to do with the fact that they do not devote much attention to equivalence relations in general. They are much more concerned with parallel frames and their formal properties rather than with the occurrence of words in real language data. No wonder, therefore, that contrastive frame semantics does not deal with differences between translation equivalents.

2.4 Corpus approaches to semantic fields

The general interest of linguists in lexical or semantic fields has diminished since the 1980s. This may be partly due to the increased importance of frame semantics and the daughter theory construction grammar. However, this does not mean that the term has completely disappeared from linguistics. Although the theory is scarcely present in contemporary studies in the form as discussed above the term itself continues to be used in new approaches and methods. Two approaches that are based on the analysis of real language data are reviewed below.

2.4.1 Core words in semantic fields

The first approach to be discussed here has been advocated by Viberg in his various studies of Swedish motion (Viberg, 2008) and mental verbs (Viberg, 2005) and of Swedish and English verbs that denote motion and physical contact (Viberg, 2004; 2010). The author makes no reference to the traditional theory of semantic fields. Both semantics and grammar serve as criteria for a definition of semantic fields. He distinguishes between fields, macro fields and subfields. Semantic field is defined “as a set of words which belong to the same word class and which are closely related in meaning” (Viberg, 1993: 341). The words that belong to different word classes but are semantically similar create macro fields. Finally, semantic fields

can be divided into subfields but clear criteria have not been specified. It follows that the lexicon is hierarchically structured.

As for the structure of semantic fields, words are grouped according the centripetal principle whereby: "the semantic field is organised around a core concept" (Viberg, 1993: 341). The word that serves as the core or nuclear concept is usually the most frequent word (as far as verbs and adjectives are concerned) and the one with the most general meaning. The meaning of all other words from a semantic field is in some way related to these nuclear words. It is not explained in which way exactly. It is claimed also that not only the vocabulary from one language but the lexicon in general, cross-linguistically, can be reduced to nuclear words. Moreover, nuclear words are assumed to be identical or very similar in every language and therefore universal. "It turns out that 6 basic meanings are realised by one of the 20 most frequent verbs in all 11 languages (BE, CAN, GIVE, TAKE, SAY, SEE) and that 2 meanings are realised within this frequency range in all but one language (GO, MAKE)" (Viberg, 1993: 348). Capital letters here denote the lemma form of the core terms. Apart from being very frequent the core words are also typologically unmarked. Just as for Greenberg (1966) unmarkedness for Viberg means that the core words are phonologically and morphologically less complex, that they are not language specific and that they occur in many languages. Besides, they can be used in a large number of syntactic frameworks, other words can be derived out of them, they have a wide collocational range, they are stylistically neutral and are usually polysemic. Some of these characteristics overlap with the test of the coreness of vocabulary performed by Carter (1998: 34-49) and Stubbs (1986). Nuclear words are considered concepts and so they are treated as semantic primitives. Here Viberg relies on Miller and Johnson-Laird's (1976) definition of semantic primitives which are established on 'perceptual judgments' of the world around us and on the idea that perception underlies the meaning of all words.

One of the most significant achievements in these studies is the method of discrimination of senses of polysemous words. This is done through the analysis of grammatical and lexical environments in which words occur and by identification of translation equivalents in a parallel corpus. Each of the senses discovered in this way is assigned to particular semantic fields. One example is the Swedish verb *få* that serves as a

nuclear verb in the following four semantic fields: *Possession*, *Modal:Permission/Obligation*, *Inchoative* and *Causative*. These senses are associated with different grammatical patterns and collocates. Thus, when *få* denotes *Possession* it is followed by a noun phrase, whereas when it has a modal meaning an infinitive construction will follow. For each of the senses there is a corresponding set of translations in English. When used in the sense of *Possession* the verb corresponds to *get, have, give, receive, acquire* or *obtain.* On the other hand, when it is used in the modal sense it corresponds to the following English words: *can, be allowed to, must* (negated), *should* (negated) and *may*.

The first problem with this approach is the character of nuclear words. The very idea that perception is the basis of meaning for all words is questionable and unfortunately not sufficiently discussed. It does not seem to be a coincidence that Viberg in his studies only deals with different types of perception verbs. In addition, the fact that the basis of semantic primitives is judgment means that the primitives have a subjective character. It is, therefore, doubtful whether the phenomena defined by relying on subjective judgements are truly universal. The technique of derivation of meaning for particular words from nuclear words is not mentioned at all and it is not clear what sort of relation governs the structure of semantic fields.

The second problem is that his explanation of differences between equivalents is neither systematic nor precise. It seems that these differences are of probabilistic nature because the author mentions their raw frequency but this issue is not explored in detail. Similarly, no attempt has been made to summarise the tendencies with regards to the grammatical and lexical items which occur with words that belong either to the same field or to corresponding fields in Swedish and English. Thus, we do not know if all English items corresponding to the *Possession* sense of *få* occur in the same textual environment.

Finally, it remains unclear why semantic fields should consist only of words that belong to the same word classes. This is an unnecessary constraint that eliminates the candidates which according to a parallel corpus may serve as equally good equivalents as those that are grammatically akin to the target word. The fact that such candidates are included in macro fields seems to suggest either that they have a more general meaning or

that they are at a greater semantic distance from the nuclear words. However, Viberg does not explicitly discuss any of the two possibilities.

2.4.2 Semantic mirrors

Dyvik's (1998; 2004; 2005) corpus approach to contrastive semantic fields concerns semantic relations between Norwegian and English. He starts from the assumption that the meaning of words becomes visible in translation and that the translation mirrors the meaning of a word. The fact that perfect translations are impossible and that the target language is always "like a Procrustean bed for the source language" (Dyvik, 2005: 7) should not be seen as an obstacle. On the contrary, it is the difference between languages that can provide interesting insights about the semantics of words. "The anatomy of meaning emerges in the translational tension between languages" (Dyvik, 2005: 7). This suggests that knowledge obtained from a contrastive analysis depends on the language pair examined and that different conclusions may follow from different language pairs. We cannot know truly to what degree this assumption holds because the author only explores Norwegian and English lexemes.

The following two assumptions underlie the character of semantic fields in this approach: "semantically closely related words ought to have strongly overlapping sets of translations, and words with wide meanings ought to have a higher number of translations than words with narrow meanings" (Dyvik, 2004: 1). The analysis itself unfolds in the following stages.

First, a set of translations of a word from the source language is identified in a parallel corpus. This set is called a *t-image*. Afterwards the words from that t-image are translated back into the original language. This second set is called *an inverted t-image*. Finally, by observing distribution of words and overlapping translations of two images the author identifies their different senses. This process is illustrated in Figure 2.2 with the Norwegian word *tak* and other semantically related words in Norwegian and English. The first t-image, represented as the diagram labelled *E* (for English), displays results from the translation of the

Norwegian word into English. The inverted t-image (the diagram *N* that stands for Norwegian) is produced by a back-translation of the English words into Norwegian. The arrows indicate a direction of translation.

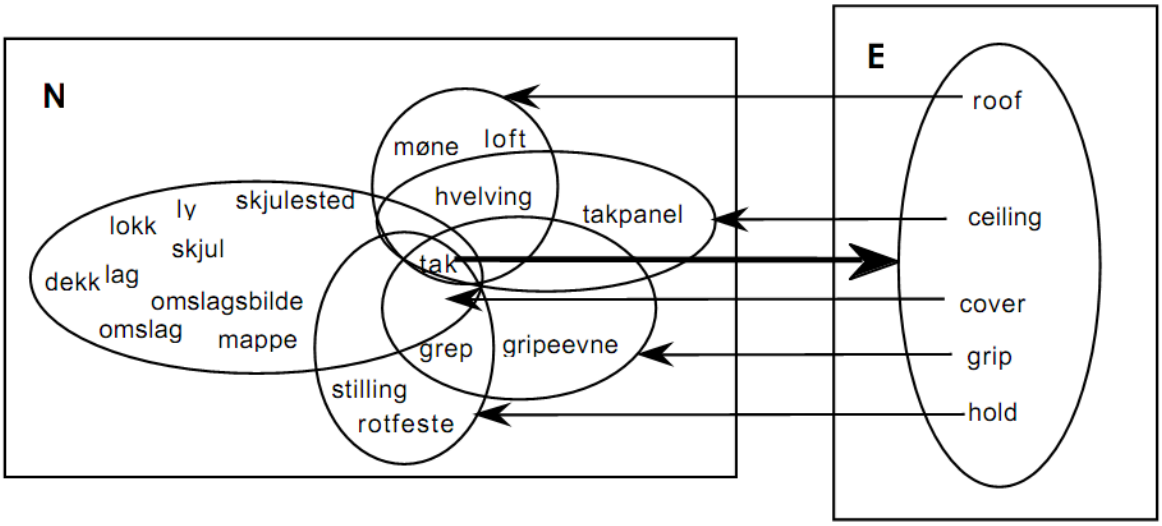


Figure 2.2: T-image and t-inverted image related to the word *tak* (from Dyvik, 2005: 38)

The following three overlapping areas can be observed above: *ceiling* and *roof* overlap when they correspond to *tak* and *hvelving*, the words *grip* and *hold* overlap when they correspond to *grep* and *tak* and *cover* does not overlap with other words. Two conclusions follow from here. First, *tak* is semantically most similar to *hvelving* and *grep* in Norwegian. Second, it has three senses that are displayed in Figure 2.3.

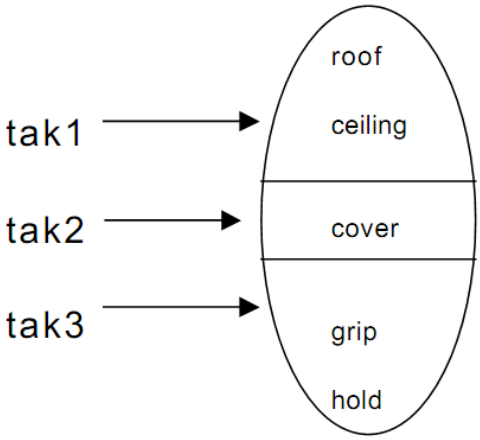


Figure 2.3: Three senses of the noun *tak* as they are reflected in translation (from Dyvik, 2005: 38)

It is interesting that differences between senses are interpreted in terms of semantic features or components like in the traditional theory of semantic fields even though they are established on completely different principles. Unlike in the traditional componential analysis these features do not refer to extralinguistic phenomena, nor do they stand for any type of semantic primitives. They are translation equivalents established in a translation corpus. The advantage is that description does not have to be limited to a specific set of words since this method can deal with any type of words. These features are used to establish two senses of the Norwegian adjective *lekker* and they are represented as: [*lekker*|*pretty*] and [*lekker*|*delicious*].

According to Dyvik, it is on the basis of these semantic features or senses that we can identify semantic fields. The following definition of semantic fields is offered: "two senses belong to the same semantic field if at least one sense in the other language corresponds translationally with both of them" (Dyvik, 2005: 11). It means that semantic fields do not consist of only monosemic words. No requirements regarding the grammatical nature of words are imposed here. Thus, in the above example the English words *roof, ceiling, cover, grip* and *hold* create one semantic field because all of them correspond to *tak*. It is also said that a semantic field contains "a set of senses that are directly or indirectly related to each other by a relation of semantic closeness." (Dyvik, 2004:317). The semantic closeness is discussed in terms of semantic relations such as synonymy, hyponymy and hypernymy. These terms suggest that semantic fields can be hierarchically structured. A more general sense is

always included as a feature of a more specific sub-sense. Or to put it the other way around, the features of more general senses are inherited by specific sub-senses. This general sense is also called a *peak* of the semantic field. For example, in the semantic field of the *meal words* the peak sense is [*mat1*|*supper2*]. This sense is inherited by the senses [*kveldsmat1*|*meail1*] and [*lunsj1*|*meail1*] which belong to the same field. All these senses are in turn inherited by [*aftensmat1*]. This is useful because it enables to find distinctions between words from the same field.

What is attractive about this approach is that equivalence relations are established in parallel corpora by observing the use of words and not by relying on intuition or dictionaries. It follows that the semantic similarity in two languages can be explored without having to define the semantic content of words or cognitive frames. Its main disadvantage is that it suffers from the same fundamental problem as the traditional semantic field theory: it focuses on isolated words. Unlike in frame semantics or in Viberg's approach, Dyvik does not explore the textual context of words. One possible reason is that it aims at providing a method for a compilation of a bilingual dictionary akin to Wordnet which itself is single-word-oriented. This is why it does not provide a method to distinguish between synonymous terms that belong to the same field. For example, the words *beautiful, attractive, charming, cute* are synonyms and correspond to the second sense of *sweet*. They are only listed as alike and we do not know when exactly they express this particular meaning of *sweet*. From what has been provided it seems that the author assumes that they have only this sense but this can hardly be acceptable. The Macmillan Dictionary lists two senses for *attractive* and *charming* and three senses for *beautiful* and *cute*. It follows that a specific meaning of *sweet* is described by means of polysemic words. The same problem is encountered in the cross-linguistic descriptions. In his list of cross-linguistic definitions Dyvik (2004), for example, describes one sense of *sweet* in terms of the Norwegian word *frisk.* But *frisk* will have different meanings in different contexts and we do not know in what specific textual context it occurs when it serves as the translation equivalent of *sweet*. It is certainly not when the Norwegian adjective is used in the collocation *frisk luft* because according to the multilingual OpenSubtitles corpus its English equivalent here is *fresh air* and not *sweet air*. In addition, the concordance lines

from the BNC and ukWaC show that it is *fresh air* which is an idiomatic expression in English and not *sweet air*.

2.5 Studies beyond single words

What is common to all the studies reviewed above is that the units of analysis are single words. Although frame semantics and the model proposed by Viberg introduce a syntagmatic element to the analysis and consider the textual context in which words are used all these approaches focus on single words. The following two models go beyond single words in the cross-linguistic study of lexical items.

2.5.1 Contextually defined translation equivalents

I will start with a review of an approach to cross-linguistic description of words based on the principles developed by John Sinclair and his colleagues in the Cobuild project. I will mainly discuss the method and results of the Multidict international project that was carried out in the early nineties and was coordinated by Sinclair himself. Illustrative of this approach are various papers by Sinclair (1996a; 1996b), Teubert (1996; 2001; 2002; 2004) and Tognini-Bonelli (1996; 2001; 2002).

The purpose of the initial project was to provide a description of "the shared meanings of languages in terms of the actual verbal contexts in which each instance is found" (Sinclair, 1996b: 174) that would serve as a basis for a construction of "a sample of a multilingual dictionary on the basis of evidence drawn from corpora in seven European languages" (Sinclair, 1996b: 179). Unfortunately, no such a database has ever been created. Like in Dyvik's approach, translation is seen as "a kind of disambiguation, with the differentiation of meaning shown by the way a word is translated" (Sinclair, 1996a: 176).

The basic assumption here is that "[t]here are likely to be parallels between the textual environment of a word in one language and a word that is used to translate it in another" (Sinclair, 1996b: 179). Thus, it is the textual context in which words occur that plays an essential role in the identification of equivalents. This assumption underlies Teubert's (2001) analysis of the translation of *sorrow* and *grief* into German. A translation of the two words into German indicates that they have three different senses: for the first sense their equivalent is *Trauer,* for the second it is *Kummer* and for the third *Gram.* Each of the three senses in both languages is associated with a specific context profile. At the time no large parallel corpora were available and the comparison of the context profiles in this study is conducted by means of English and German reference corpora. Teubert concludes that the collocation profile of *sorrow* and *grief* "will not differ much from the context profile for *Kummer* extracted from the German reference corpus, apart from it being in English instead of German" (Teubert, 2001: 148). One example is the noun *Stress* and *stress* that occurs both with *Kummer* and *sorrow*, respectively. In addition to the lexical context the grammatical context can also help to disambiguate the meaning of words. Thus, *know* will be translated into German as *wissen* when it is followed by a reported clause and as *kennen* when a noun phrase follows it (Sinclair, 1996b: 180). The examples illustrate that the study of textual contexts in which words occur is a more suitable way to identify equivalents than the matching of single words. A further example is a study of how *in the case of* and its Italian equivalent *in caso di* in Tognini-Bonelli (2001). The two expressions correspond to each other because they occur in a similar textual context. Tognini-Bonelli also shows that the notion of semantic preference can help to classify collocates. Thus, the collocates of the adverb *largely* can be grouped into three general semantic sets: *CAUSE/REASON*, *BASIS*/*RELATION* and *NEGATIVES*. In the first set we find collocations such as *largely because*, *largely thanks to* and *largely as a result.* This sense of *largely* is always translated into Italian by *soprattutto*. These semantic sets are reminiscent of lexical fields although the author makes no direct connection to it. Similarly, she does not address in depth the issue of the structure of the lexicon. It is only mentioned that the description of all parallel units of meaning should eventually lead to the creation of the webs "of equivalences based on contextual patterning" (Tognini-Bonelli,

2001: 149). One such web would consist of all parallel environments in which a lexical item from L1 and its equivalent from L2 occur. Following her analysis one can conclude that in such a web *largely* would belong to three different sets to which three parallel Italian sets would correspond.

The most important achievement of this approach is the disambiguation of word senses through observation of occurrence of words and multi-word expressions in translation. It shows that there is no need for external references to study meaning and describe equivalence relations. Unfortunately, the approach does not discuss the issue of discriminating between the equivalents that correspond to the same sense of a word. Because there were no adequate parallel corpora at the time, the equivalents in these studies were identified with the help of bilingual dictionaries or intuition. This may be the reason why usually only one translation equivalent is provided. For instance, in the last mentioned study *largely because* corresponds only to *soprattutto perché*. However, the English-Italian section of the EUparl corpus indicate some other options as well: *soprattutto* (*perché, a causa de*/*di*), *in gran parte* (*a causa de*/*di, perché, grazie a*), (*dovuto*) *in gran parte*, *principalmente* (*a causa della*), *ampiamente* (*a causa di*/*del, perché*), *in larga misura* (*a causa di*/*del, grazie a*), (*dovuto*) *in larga misura.* Therefore, the representation of relations between equivalents is too simplistic.

As noted above, the notion of semantic groups is discussed only in passing. The approach seems to be uninterested in general lexicology issues.

2.5.2 The Bilexicon project

The final section of this chapter is devoted to the multilingual Bilexicon project developed by the German linguist Dirk Siepmann (2005). Siepmann like Sinclair proposes that the study of correspondence relations across languages should go beyond single words. However, there is an important difference between two approaches. As we have seen in the previous section the issue of semantic grouping of words was only indirectly addressed. On the other hand, in

the approach proposed by Siepmann it plays a crucial role. This is not surprising given the fact that the aim of the Bilexicon project is to compile a multilingual onomasiological dictionary in which words would be ordered thematically

Siepmann deals with the following three languages: English, French and German. The thematically ordered entries are to consist of near-native lexical units. The following question is central for Siepmann: "what are the meaning units that native speakers use, and which of these have to be mastered to be able to perform at a near-native (or lower) proficiency level?" (Siepmann, 2005: 4). The principal units of analysis are topics that can be divided into sub-areas. One such topic is for example *motoring* and one of the sub-areas that it encompasses is *parking*. The notion of topic is ambiguously explained in terms of the concept of situation-type which itself is only vaguely defined. It seems that situation-types refer to the social context in which a text is produced. In this sense, they appear to be similar to Halliday's (1978) context of situation. The relation between a topic and sentence-type is explained in the following way:

> "One situation-type, such as a court hearing, can involve widely varying topics… conversely, the same topic, such as an account of an accident, can occur in several different situation-types or text-types, such as general conversation, court hearings, newspaper reports or insurance claims letters." (Siepmann, 2005: 7)

From this we can understand that there is no one-to-one relation between the topics and situation-types. However, it is still not clear on what basis the topics are defined.

From the outset, the vocabulary in the languages studied is divided into topic-specific sub-corpora. After that, the word list and a list of collocates specific to the subject areas are produced with the help of WordSmith tools for each language. Finally, the translation equivalents of collocations are established by relying on introspection. The results are the entries of the topic-specific collocations in one language and their translation in another language. Figure 2.4 displays such an entry for the English collocation *stream of traffic* and its German equivalents.

The entry shows how the choice of equivalents depends on the context in which a lexical unit from the source language is used. We can also observe that even larger lexical units can be ambiguous if not presented within a suitable context. According to the above entry, the *stream of traffic* will be a synonym of *flow of traffic* and *traffic flow* when it occurs with *steady* and in this case its German equivalents will be *der Verkehrsstrom* and *die Verkehrsflut*. On the other hand, when *stream of traffic* collocates with *endless* it has the same meaning as *solid line of cars* and *heavy traffic* and corresponds to *die Blechlawine*.

English	German
stream of traffic / flow of traffic / traffic flow	der Verkehrsstrom / die Verkehrsflut
the steady stream of traffic heading to St Sampsons	*die kontinuierliche Verkehrsflut in Richtung St. Sampsons (die sich nach St Sampsons ergießende Blechlawine)*
look behind early and move into the stream of traffic when safe	*schauen Sie sich frühzeitig um und ordnen Sie sich bei einer günstigen Gelegenheit in den fließenden Verkehr ein*
endless stream of traffic / solid line of cars / heavy traffic	die Blechlawine*
there is an endless stream of traffic from the Straße des 17. Juni going past the Brandenburg Gate	*von der Straße des 17. Juni rollt eine Blechlawine am Brandenburger Tor vorbei*
we go around a bend and there ahead of usis a solid line of cars as far as you can see	*wir fahren um eine Kurve und vor uns ergießt sich eine Blechlawine soweit das Auge reicht*

Figure 2.4: The expression *stream of traffic* and its German equivalents (from Siepmann, 2005: 19)

In addition to the insufficiently defined nature of the topics the approach provides an unsystematic representation of lexical units and lexicon in general. This is mainly due to the author's insistence on not using frequency as a criterion for the extraction of collocates. It is the entry compiler who, relying on his feeling for language, decides what the near-native expressions specific to a given topic are. As a result we have a highly subjective selection of collocations and of what is considered to be a topic and sub-topic.

An even more serious drawback is that the relationship between translation equivalents is not clearly defined. Thus, from the above entry one can conclude that any of the two German words *der Verkehrsstrom* and *die Verkehrsflut* can be used as equivalents for any of the following three English words: *stream of traffic, flow of traffic* and *traffic flow*. It seems that the author assumes that it makes little difference which of the offered equivalents will be used. Unfortunately, no large corpora are available where this assumption can be tested but as we know from the literature absolute synonyms, those that can be used interchangeable in every context, are extremely rare (Cruse, 1986). In addition, as it will be seen in the analysis conducted in the subsequent chapters translation equivalents do not occur always equally likely.

Nevertheless, the approach should be credited for recognising that the study of words in contexts is essential if dictionaries are to facilitate learners in acquiring idiomatic expressions. It also shows how the correspondence relationship between lexical units in two languages changes when we observe them in different contexts.

2.6 Conclusion

In this chapter different approaches to the contrastive study of semantic fields and relationships between translation equivalents have been critically reviewed. Special attention has been given to the following issues:

- if the approach demonstrates how equivalence relations have been established;
- if the approach deals with the issue of differentiation between translation equivalents that have the same meaning;
- if the approach deals with the structure of semantic fields and the structure of the lexicon in general.

The vague principles on which components and frames are established and reliance on introspection both in this matter and in establishing equivalents are problems which characterise both the traditional semantic field theory and frame semantics. More attractive for the current thesis are studies concerned with real language data. Identification of equivalents in parallel corpora by means of the textual context analysis appears a sound basis from which to begin. On the other hand, more can be learned about semantic groupings of words from the theory of semantic fields, frame semantics and the two Scandinavian approaches than from Sinclair's approach. All these issues will be further discussed in the following chapter which lays the theoretical and methodological foundation for the model proposed in this thesis.

Chapter 3 Theoretical background

3.1 Introduction

In this chapter I will discuss the theoretical and methodological framework on which the analysis of distribution of translation correspondences in English and German is based. I will also describe the data that will be used in the study as well as the analysis procedure. At the end of the chapter the definitions of the terms and conventions used in the present thesis will be provided.

The theoretical and methodological framework relies upon the language in use theory of meaning. This framework will be discussed in detail in Section 3.2. In 3.2.1 I will begin by providing an account of relevant features of this theory as they were formulated in Wittgenstein's (1953) ordinary language philosophy. In 3.2.2 a distributional approach to translation correspondences will be proposed as a model for operationalisation of this theory. An approach to the issue of translatability and translation correspondences from a point of view of hermeneutics will be introduced in 3.2.3. The subsection 3.2.4 explains how restricted distribution of lexical items which are used as translation correspondences will be dealt with in terms of local grammars. 3.2.5 and 3.2.6 are concerned with the corpus categories and

corpus tools which will be used to identify the correspondence relations between two languages and to describe distribution of lexical items.

Section 3.3 will deal with the parallel and monolingual corpora from which the data for the present analysis derive.

In 3.4 the procedure for identifying translation correspondences in a parallel corpus and for examining local contexts will be described.

3.5 will summarise the key terms which will be used in the analyses conducted in following chapters.

3.2 Theory and methodology

3.2.1 Language in use theory of meaning

A contextual view of meaning is summarised in Firth's often quoted sentence: "You will know a word by the company it keeps" (Firth, 1968: 179). According to this statement, in order to study the meaning of a word it is necessary to give an account of its occurrence. The occurrence of a word encompasses all environments in which it occurs. This idea is borrowed from Wittgenstein (1953; 1958) who is cited by Firth on the very same page where the above quote occurs. In linguists one usually does not go beyond acknowledging this relation. Here, I would like to claim that a more complete account of Wittgenstein's notion can provide better foundations for a contextualist corpus approach to meaning in general and to the study of translation correspondences in particular.

Wittgenstein's early work (1922) was concerned with developing a formal logic-based account of meaning. In the later works his interest shifted from ideal to ordinary language. A novel approach to meaning that he proposed in this second phase is summed up in the claim that "the meaning of a word is its use in the language" (Wittgenstein, 1953: 43). The idea relies on Frege's Context Principle (Reck, 1997) expressed as: "it is only in the context of a proposition that words have any meaning" (Frege, 1960: 73). However, the two philosophers further develop this idea in two different directions. Two key components of Wittgenstein's

notion of meaning are: language in use and the role of contexts. For Wittgenstein, there is no point in asking the question "What is a meaning of *X?* because "[i]f you want to know what a word means, look and see how it is used" (Waismann, 1965: 157). Language is considered an act or activity (*Handlung*) the purpose of which is to communicate meaning. This activity is not "something which just one man might do just once in his life" (Baker and Hacker, 2009: 28). It is, rather, something regular. The use of language as a regular activity is called a practice (*Praxis*). Therefore, "[i]n order to describe the phenomenon of language, one must describe a practice, not something that happens once, no matter of what kind" (Wittgenstein, 1956: 336). What we identify when we observe this practice are the patterns of action (*Handlungsmuster*) which are also called grammatical rules of language use (Busse, 1991: 10; Keller 1974: 10). These grammatical rules are not to be confused with Chomsky's (1957) syntactic rules that he considers to be internal property of the human mind and as such responsible for producing grammatically correct sentences. Neither should they be equated with the instructions, codifications and language standardisations (Heringer, 1977: 60-76). The grammatical rules are occurrences of words in typical contexts. The task of linguists is to discover these typical occurrences and to explicate them (Caillieux, 1974: 37). By doing this they describe the meanings of words.

There are different ways of identifying grammatical rules and I will discuss two of them which are relevant for the purpose of the present thesis. First, we can read rules off a language by looking at examples (Waismann, 1965: 148; Wittgenstein, 1953: 28). The examples display previous uses or occurrences of a word in specific contexts (Busse, 1991: 58) and they show what different occurrences of a word have in common (Baker and Hacker, 2009: 53). The examples must be collected systematically (Caillieux, 1974: 39), which means that the focus should be on the examples that illustrate frequent uses because the established rules should be representative of a given word. Second, we can define the meaning of a word by checking if it has more than one meaning. This may be done by testing if a word can be substituted by a new word (Waismann, 1965: 154). For Busse (1991: 51) this means to provide explanations through paraphrases (Busse, 1991: 51). Here we can conclude that "a word *a* means different things in different contexts, if in the one case it can be

replaced by *b* and in the other by *c* but not by *b*" (Waismann, 1965: 154). In both cases the meaning of a word is explained in terms of the typical environment in which it occurs. As we will see below, the two models can be successfully combined.

Wittgenstein also provides a term for a set of rules that describe occurrence of words. "[T]ypical local contexts in which an expression is given a proper role or use or meaning" (Penco, 2004: 286) is called a *language game*. The notion of language games stresses that typical contexts and uses of words are restricted in numbers. A word can occur only in a limited number of environments and will therefore be associated with a limited number of language games.

One may object that Wittgenstein's conception of grammatical rules is a philosophical one and different from the one used in linguistics. Wittgenstein himself refuted this objection. He insisted that his usage of the term *rules of grammar* was in the same sense as it was in ordinary grammar (Wittgenstein, 1980: 98). One can support this claim by showing that he, like linguists, talks about the grammar of words, expressions and sentences (Baker and Hacker, 2009: 59). The difference concerns the general aims and objectives. A philosopher studies grammars of language games to explore certain philosophical issues whereas a linguist is interested in purely linguistic phenomena. But, as the example of Firth illustrates, the model can successfully be used for the purposes of linguistics studies.

As we can see, Wittgenstein's theory of meaning is not only interesting because it served as a starting point for the contextualist approach to meaning in linguistics. It also provides some practical guidelines to the study of meaning through the notion of grammatical rules or patterns of occurrence, the role of context, explanation of meaning through examples and substitutions and the idea of language games.

3.2.2 Distributional approach

"There are likely to be parallels between the textual environment of a word in one language and a word that is used to translate it in another" (Sinclair, 1996b: 179). This is one of the

principles that guided the study of the equivalence relations in the international Multidict project reviewed in Section 2.5.1 in the previous chapter. According to this assumption, a close investigation of shared contexts in two languages will lead to an identification of translation equivalents. The proposition is an extension of the approach introduced by Zellig Harris (1952; 1954; 1970) and today known under the terms as *distributional hypothesis* (Sahlgren, 2008), *distributional semantics* (Schütze, 1998) or *distributional corpus analyses* (Geeraerts, 2010: 166-178). Henceforth, I will refer to the approach as the distributional hypothesis. The distributional hypothesis claims that "if we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words: difference of meaning correlates with difference of distribution" (Harris, 1970: 785). Distribution of an element is defined as all textual environments in which it occurs (Harris, 1952; 1954). The thesis can be understood as a re-formulation of the substitution principles introduced above even though it needs to be stressed that in Harris's work no direct references are made to Wittgenstein. Nevertheless, in both cases it was claimed that the lexical items that have same distribution will be semantically similar. According to Harris (1952), the lexical items that occur in the same context create an equivalence class. What qualifies items to belong to an equivalence class is that they are substitutable for each other. Historically, substitutability as an indicator of the equivalence relation is reminiscent of Leibniz's definition of identity relations: "Two things are the same if one can be substituted for the other without affecting the truth" (Lyons, 1977: 160). Relying on the substitution principle we arrived at a truly distributional and language in use definition of equivalence relations and semantic sets.

First, a comparative examination of the textual contexts in which lexical items from two languages occur will help to identify corresponding words in two languages. The correspondences identified will create an equivalence class. These correspondences will, therefore, be defined as sets of lexical items with the same distribution. This is similar to Zgusta's (1971: 314-315) proposal that lexicographers should establish the relationship between lexical items from a source and target language by comparing the contexts in which

they occur. Accordingly, the task of lexicographer is to find "real lexical units of the target language that, when inserted into the context, produce a smooth translation" (Zgusta, 2006: 236). For this reason, for Zgusta the lexical items from a target language are insertable equivalents and therefore their defining feature is insertability.

Second, using the distributional principle we can identify lexical items which create a class of semantically similar items. Thus, we can assume that if there are two or more lexical items from language A that create an equivalence set with one or more corresponding lexical items from language B then these items will form two substitution sets, each in its own language. According to the distributional principle the members of these sets will occur in similar contexts and denote similar meanings.

It is important to stress that the equivalence sets should be established from both an intralingual and interlingual perspective. This is because the words that occur in the same textual context are not automatically synonyms. Such words may also be antonyms or may belong to other types of semantic relations that have not been properly studied yet. For example, Church et al. (1994) show that words such as *pledge* and *contribute* may occur in the same context but are not synonyms. In order to make sure that such cases do not occur in our data, the substitution sets need to be established first in an intralinguistic analysis and then confirmed by an intralinguistic analysis. No two items from L2 will correspond to the same item from L1 and simultaneously occur in the same context unless they are synonymous**.**

3.2.3 Translation correspondences

So far, we have not considered the question where the equivalence sets come from. In the current model they will be seen as a product of a translation practice. A translation practice is an activity in which translators act as language users. They translate texts between languages in order to communicate meaning from a source into a target language. Translation activity as a type of communication can be explained with the help of the term *convergence* (Ervas,

2008). This term in ordinary language philosophy means “starting with two different sets of beliefs or theories, two speakers converge towards the same meanings, elaborating a common theory which is built up during the dialogue” (Penco and Vignolo, 2005: online). In translation instead of two speakers as communication participants we have two texts, an original text and its translation and translator. The sets of beliefs or theories in translation stand for the knowledge of a source and target language and previous experiences of translation. As such, they inform a translator’s decisions in the process of translation.

Convergence is an interpretative process in which “the interpreter will modify his initial theory in accordance with the entry information, building one or more *passing theories*” (Ervas, 2008: 22). Similarly, in order to translate a text one needs to understand and interpret its meaning (Gadamer, 2004: 386). By interpreting a foreign text a translator aims at making “what is alien our own” (Gadamer, 1977: 19). This process is taking place in language because interpretation and translation is not about engaging “in some kind of abstract thinking that is independent of both languages” (Yallop, 2004: 70), rather it is about “making the unknown comprehensible though known“ (Klöpfer, 1967: 69, my translation). The ‘known’ from the previous sentences are linguistic resources available in the target language.

It is also important to stress that “[o]ne's interpretation of a particular subject matter stands in a tradition of previous interpretations of the same subject” (Mendelson, 1979: 55). Thus, a translator’s decision to choose a specific term from a target language is always influenced by past decisions taken by members of a translation community of which she is a part. In other words, her theories and knowledge are inter-connected with the theories and knowledge of others. A translator’s decision to select a particular item depends on whether or not she is satisfied with the given option. Or in Wittgenstein’s words: “What happens is not that this symbol cannot be further interpreted, but: I do no interpreting. I do not interpret, because I feel at home in the present picture“ (Wittgenstein, 1967: 44).

Now, we can establish a link between this view and what was said in 3.2.1 and 3.2.2. Past translations can be considered examples of substituting the items from two languages for each other. Materially, these past translations can be collected and stored in a corpus. Depending on the size of a corpus and its diversity a corpus will more or less successfully

reflect current dominant interpretations of meaning of an item from a source language by means of the items from a target language. Distribution of translation correspondences in corpora will also mirror the current degree of agreement among translators regarding the substitutability of a term from L1 by one or more terms from L2.

An issue that needs to be addressed here is that of translation equivalence. It has presented a central question in translation studies and bilingual lexicography for decades (see for example Baker, 1998: 77-80; Duval, 1991: 2817-2824; Munday, 2001: 35-55; Zgusta, 2006: 230-261). Nevertheless, not all scholars even agree that the term *equivalence* is really appropriate here. Snell-Hornby (1988; 1990) dismisses it because it has been adopted from mathematics and as such it creates "an illusion of symmetry between languages which hardly exists beyond the level of vague approximations and which distorts the basic problems of translation" (Snell-Hornby, 1988: 22). Besides, the notion of equivalence also presupposes that "a word in one language must necessarily be lexicalized to fulfil the same function in another language" (Snell-Hornby, 1990: 209-210). This simplistic belief has been continually refuted in translation practice. A discussion of equivalence relations, according to her, is premature and should be put on hold until other important questions are dealt with.

Johansson (2007) takes a similar view. According to him, the issue of equivalence should not be addressed at the early stage of research. We should start rather by exploring correspondence relations between terms in two languages. By studying "the correspondences we may eventually arrive at a clearer notion of what counts as an equivalent across languages" (Johansson, 2007: 5). The notion of correspondence, unlike that of equivalence, does not presuppose that the items in two languages are identical. It must be stressed that the term *correspondence* here does not have the same meaning as Catford's (1965) term *textual correspondence* or Koller's (1978) term *correspondence*. For Johansson correspondence simply denotes the relationship between lexical items revealed in a corpus of parallel texts. He does not specify what linguistically these correspondences are but from his studies one can conclude that they range between single words and multi-word units. In the present thesis the size of correspondences remains flexible and not determined in advance as it is the case in some other approaches. It means that any two expressions that occur in the

same context regardless of their size can be considered translation correspondences. The lexical items from two languages will be treated as translation correspondences and the issue of translation equivalence left aside. This view will make possible the development of a model that will be applicable to a range of lexical items regardless of their size or grammatical class.

3.2.4 Sublanguages and local grammars

Above it was said that the notion of grammatical rules or typical patterns indicated that the meaning of words was locally defined. In addition, the notion of language game was accepted because it indicates that contexts in which words occur are always restricted in number. Finally, it was considered that the lexical items that had similar distribution could be grouped into substitution classes. The local character of the grammar rules and their restricted number underlie classification of lexical items to semantic domains in linguistics and need to be discussed in greater detail.

I will propose that the issue can be approached with the help of the notion of sublanguage introduced by Harris (1968). A sublanguage is seen here as a semantically restricted domain with particular distributional features. These domains are based on common subject-matter. Thus, we have sublanguages of immunology or surveys (Harris, 1988; Harris et al., 1989), technical manuals for aviation (Kittredge, 1982) or task-oriented dialogues (Grosz, 1982). Below I will argue that the approach can be extended to other semantic sets identified on the distribution principle.

The grammars of sublanguages are based on the distribution principle, that is, on "a purely word-combinatorial investigation" (Harris, 1988: 40) and not on *a priori* concepts that supposedly underlie the meaning of words. One argument against using the idea of sublanguage in the present thesis is that according to some authors (e.g. Kittredge, 1982; Moskovich, 1982) the grammars of sublanguages are substantially different from general language grammar. According to this view, the grammars of sublanguages unlike the grammar of general language are full of irregularities. However, it is not very difficult to find examples in the grammar of general language that disprove this view. According to general

grammatical rules, when a noun occurs with a verb it is normally preceded by an article. The following expressions that belong to the general language clearly violate this rule: *take place, take part*, *take advantage*, *take responsibility*, *take action*, *take photo* or *take advice*. One may question the validity of this example and argue that in this case we deal with idiomatic expressions that often do not obey grammar rules (Burger, 1998). But, then, what about *take a look*, *take the opportunity*, *take the form* or *take a step*? These are also idiomatic expressions yet in accord with grammar. So, the idiomaticity does not seem a strong argument here. Furthermore, some lexical items may occur in both 'grammatical' and 'ungrammatical' combinations such as: *take time* and *take a long time*. This behaviour is obviously not specific to the *take* constructions. Consider *make use*, *make sense* and *make trouble* as opposed to *make a difference*, *make a decision*, *make a claim* or *make a statement*. Finally, we say *make a note* but *take note*.

What all these examples indicate is that the issue of sublanguage has not only to do with specific text types. Rather it is related to the general problem of conventional grammatical categories which are both overgeneralising and too crude. This problem is addressed by Waismann (1965) in his paper on grammatical rules. He notes that general grammatical categories cannot explain why *north-east* does not occur in the contexts *. . . of the North Pole* and *. . . of the South Pole* such as in *north-east of the South Pole*. He notices that the only explanation that the general grammar can provide is that *north-east* can be followed either by a noun or pronoun. The reason why it cannot explain the type of occurrence is because "our division of words into separate types probably follows principles that are too rough" (Waismann, 1965: 136). In other words, conventional categories of parts-of-speech are too general to grasp this fine difference. It follows that

> "it would be arbitrary to accept that it is a rule of grammar that 'north-east of' must be followed by a noun or pronoun in the accusative, yet to deny that it is a rule of grammar that these must themselves be designators of a place, an object or person at a place, or of an event occurring at a place." (Baker and Hacker, 2009: 63)

Thus, the problem has more to do with our descriptive categories than with different text types. It is these categories that create an impression that sublanguages and language requires different types of grammatical categories. In fact, to provide a more detailed description of occurrence of words we must "dig deeper, pushing aside the outward division of words into noun, adjective, etc." (Waismann, 1965: 136) and create categories that describe local contexts in which words occur.

The categories used to describe local contexts of sublanguages belong to local grammars of these sublanguages. The term *local grammar* was proposed by Gross (1993). Currently various local grammars are available such as a local grammar of dictionary definitions (Barnbrook and Sinclair, 1995; 2001), evaluative language (Hunston and Sinclair, 2000) or cause and effect in biomedicine (Allen, 2006). The great merit of "local grammars [is that they] employ functional categories specific to the area under description" (Butler, 2004: 158). This is one example:

Evaluative category	*Hinge*	*Thing evaluated*
The important point	is	to involve them in the decision.

Table 3.1: Representation of a local grammar of evaluation (adapted from Hunston and Sinclair, 2000: 91)

As we can see, the model makes it possible to develop a precise description of the function of lexical items in a specific context. This is because the assigned category labels "are far more transparent than the highly general ones" (Hunston and Sinclair, 2000: 80). Additionally, local grammars may contain categories that are multi-word units which makes it possible to develop an analysis that goes beyond the traditional parts-of-speech categories. Therefore, the model of local grammars seems to be suitable for studying local contexts of the lexical items used as translation correspondences.

Using the notion of sublanguages and local grammar we can define the substitution sets and the relations between the translation correspondences more specifically. It will be said that belonging to the same substitution sets create two corresponding sublanguages characterised by similar locally-defined grammatical rules or local grammars. By analogy with

the aforementioned definition of sublanguages I will refer to the substitution sets of translation correspondences with the term *lexical domains*. The term seems to be more suitable than the traditional terms *lexical fields* or *semantic fields* which, as was seen in the previous chapter, nowadays have somehow ambiguous meanings and are used interchangeably with the terms *semantic sets or semantic groups* that are used as descriptive rather than technical categories.

What speaks in favour of using the notions of sublanguage and local grammar in the study of translation correspondence is its successful application in the field of machine translation (Kittredge, 1987; Lehrberger, 1982). Here, it was shown that restriction of translations to semantic domains has made it possible to automate translation. The analysis of various text types indicated that the idiosyncratic features of one language can resemble those from another. The studies showed that "automatic translation from L1 to L2 does not require complete grammars of L1 and L2, only context sensitive transfer rules to obtain the proper lexical items in L2 and some rules for restructuring the resulting string of lexical items in L2" (Lehrberger, 1982: 98-99). It is of little significance that in these studies sublanguages are defined in the traditional way because, as was argued above, the sublanguages do not have to be defined only in relation to specific subject-matters.

3.2.5 Corpus categories and corpus tools

In this section categories and tools used in the analysis of the distribution of lexical items will be described. These categories and tools will help to "separate from the mush of general goings-on those features of repeated events which appear to be part of a patterned process" (Firth, 1968: 187).

It was stressed above that a detailed study of word occurrence and the associated rules should be based on an observation of systematic examples, e.g. the examples that illustrate typical occurrence of words. To identify such examples a large number of word occurrence need to be investigated. This can be done by observing the word occurrence in a

corpus. The term *corpus* refers to "a collection of naturally-occurring language text, chosen to characterise a state or variety of a language" (Sinclair, 1991: 171). A corpus consists of natural and authentic language, not of artificially constructed sentences. As such "[i]t is a record of performance, usually of many different users, and designed to be studied, so that we can make inferences about typical language use" (Stubbs, 2001: 239-40).

Nowadays, all corpora are electronic which makes it possible to explore occurrence of lexical items by means of computational tools. One of the basic computational corpus tools is the KWIC concordance that displays contexts in which a word occurs. This tool lists the occurrences of words on the concordance lines. With the KWIC "[the] word-form under examination appears in the centre of each [concordance] line, with extra space on either side of it" (Sinclair, 1991: 33). The word in the centre is called a *node word* and the textual environment in which it occurs is a *span*. A span can be of varying length but previous experience teaches us that the optimal span for the study of typical contexts in which a node occurs is four words before and after the node word (Sinclair et al. in Krishnamurthy, 2004: xix). The words that repeatedly occur within a defined span with a node word create typical contexts for a node word. These repeated words are referred to as *collocates.* Accordingly, the frequent co-occurrence of a node with its collocates are called *collocations*. These two terms in this meaning will be used in the present thesis.

The study of collocations is the basic way of studying contexts of lexical items by means of electronic corpora. It is through the observation of collocations that we detect typical patterns and typical usages of lexical items (Sinclair, 1966). The study of collocations therefore reveals relationships "a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey, 1991: 6-7). The issue of typicality is here addressed in terms of probability. This manoeuvre makes it possible to approach typicality as a variable to be explored with the help of statistical measurements. This is explained in 3.1.6 below.

So far the terms *word* and *lexical item* were used interchangeably in the present thesis. But, note that in the above quotation collocations involve relationships between lexical items and not words. Sinclair (2004: 132) considers that the term *lexical item* is more

appropriate here because words are products of writing conventions. These conventions are not always straightforward and may change over time. Thus, we write *another*, *maybe* and *wherever* as one word but *in order to*, *as if*, and *of course* as two or three-word expressions. The term *lexical item*, on the other hand, is neutral in this sense because it covers both linguistic units smaller than a word and multi-word units. In the present thesis the term *lexical item* will be used instead of *word*. In addition, in the same sense I will also use the terms *lexical unit* and *linguistic unit*.

Further studies have shown that the study of collocates can be modelled at a more abstract level.

First, various collocation studies indicate that "[p]articular syntactic structures tend to co-occur with particular lexical items, and – the other side of the coin – lexical items seem to occur in a limited range of structures" (Francis, 1993: 143). In other words, the lexical items that share distribution will also tend to share general grammatical structures in which they occur. This phenomenon is called *colligation* and is defined as "the co-occurrence of grammatical phenomena" (Sinclair, 2004: 142). A colligation study consists of classifying contexts in which lexical items occur in terms of the parts-of-speech categories.

Second, the collocates of a node word tend also to be semantically related. This phenomenon is labelled *semantic preference*. For example, a set of collocates that occur with the adjective *large* create a semantic group of quantities and sizes (Stubbs, 2001: 65).

Both types of classification can contribute to the description of translation correspondences. Grammatical classification will be a first step in developing local grammars of the items from the same lexical domain. A similar application of conventional grammatical categories is found in other local grammar approaches (Allen, 2006; Barnbrook, 2002). Semantic classification will serve to classify the specific sets of collocates according to their meaning.

The field of corpus studies has always been closely connected with technological developments. Thus, it was only after a rapid development of computational tools that Sinclair was able to implement a method of collocation studies suggested some 20 years earlier. In the past 20 years the size of corpora has increased significantly. This created new

challenges for corpus linguists because an investigation of huge corpora based on concordance lines is time consuming. However, along with this development, new concepts and more powerful corpus tools were created. Three such tools will be used in the present thesis: the IMS Corpus Workbench (Evert and Hardie, 2011), Sketch Engine (Kilgarriff et al., 2004) and RCQP package (Desgraupes and Loiseau, 2012).

The IMS Corpus Workbench is a set of tools that uses the powerful CQP query language that makes possible more complex queries than the traditional corpus tools such WordSmith (Scott, 2008) or AntConc (Anthony, 2011). One of the options that it provides is to sort and group results in many different ways. For example, one can group collocates of a node word in the form of lemma words or in terms of parts-of-speech. This is especially helpful in dealing with a very long list of collocates. Classification according to the word classes simplifies the task of identification of colligation patterns considerably.

Two particular tools in Sketch Engine that will help in the investigation of the local grammars are Word Sketch and Sketchdiff. Word Sketch is a lexical profiling tool that provides "a kind of statistical summary which reveals the salient facts about the way a word most typically combines with other words" (Atkins and Rundell, 2008: 109). Sketch Engine uses the CQP query language and unlike previous corpus tools this statistical summary is not 'grammatically blind' because it produces "separate collocate lists for different grammatical patterns" (Kilgarriff and Rundell, 2002: 5). Like in the aforementioned IMS Corpus Workbench, collocates can be grouped here as also as lemmata or according to their word classes. The Sketchdiff tool compares "word sketches for the two words, showing the collocations that they have in common and those they do not" (Kilgarriff and Kosem, 2012: 46). Word sketches are based on the collocation strengths of the compared collocations. Unfortunately, Sketchdiff allows only comparison of single words and given the fact that many lexical items from the lexical fields studied in the present thesis are multi-word units it is not possible to use the tool.

The shortcomings of Sketchdiff will be overcome using the RCQP package. This is a package that can be installed to the programming language R. It combines various statistical

packages with the tools available at the IMS Corpus Workbench. As a result, it produces the same results as Sketchdiff without the need to restrict the comparison to the single words.

In addition to the above tools the parallel concordancer ParaConc (Barlow, 2008) will be used. With ParaConc one can search corresponding words in parallel corpora. The texts in parallel corpora are aligned at the sentence level and the software produces all lines in which a term from L2 corresponds to a term from L1 occur.

3.2.6 Probability and differences between lexical items

In 3.2.1 grammatical rules were defined as the typical distribution of lexical items. In 3.2.5 it was said that the study of typicality was about distinguishing between random and non-random distribution and that the notion of typicality can be studied in terms of probability. According to Halliday (1991: 42) "[a] linguistic system is inherently probabilistic in nature" which suggests that the distributions of lexical items are not equally probable. The notion of probability makes it possible to study typicality by means of statistical measurements. The language is full of variability "but it is vague and variable in principled ways, which are at present imperfectly understood" (Hanks, 2008: 128). A statistical approach seems to be suitable here given the fact that the purpose of "descriptive statistical methods is to increase our understanding of the nature of variability in a population" (Peck et al., 2008: 5).

The variability of distribution of lexical items from the translation lexical domains will be studied in relation to the following three issues. First, using statistical measurements we will be able to specify typical contexts in which a lexical item occurs. Second, statistical analysis will serve to find differences between lexical items that belong to the same lexical domains. Third, they will help to distinguish between more and less probable translation correspondences from a target language. Therefore, the differentiation principle will be realised by applying various statistical measurements to the analysis of distribution of lexical items. In addition, these measurements will also serve to study local grammar rules and other tendencies that occur in data.

The measure used to study the probability of co-occurrence or collocation strength is called Dice's coefficient. This measure was proposed by Lee Raymond Dice (1945) in his study of association strength of species in nature. Initially, the test was introduced to replace Forbe's coefficient association test. The latter is similar to the Mutual Information test which is often used in corpus linguistics. The problem with both these tests is that they return a very high score for low frequency word pairs (Evert, 2005). Rather than indicating the association strength of a collocation they show the "amount of the deviation of the number of their occurrence together from the number expected by chance" (Evert, 2005: 298). Dice's coefficient proved to be superior to other tests in the study of words' probabilities (Curran, 2004). This test relies on the measurement of a coincidence index and measures the joint occurrence of *a* and *b* divided by the total occurrence of *a* and *b* separately in two samples. It provides information similar to that gained from the measurement of relative frequency since it measures the proportion of joint frequency in relation to the total number of their separate occurrence. However, Dice's coefficient is more reliable and exact. "The values of the association indices and of the coincidence index range from 1.0, which indicates association of the two species in all the samples examined, to 0.0, which indicates complete failure of association under the conditions of observation" (Dice, 1945: 298-299). In reality, the occurrence with the value 0.100 or higher will mean that we deal with typical collocations. On the other hand, values lower than 0.001 will indicate that the observed co-occurrences are less typical. This test was used with the corpus tools Word Sketch and Sketchdiff to help lexicographers to distinguish between more and less typical collocations. It is stable on different corpus sizes and types of corpora. Since the results of Dice's coefficient range between 0.0 and 1.0 and in some cases it is difficult to estimate the differences in probabilities a new log version of the test was proposed (Rychlỳ, 2008) under the name *logDice test*.

The main features of the logDice test are:

- "Theoretical maximum is 14, in case when all occurrence of X co-occur with Y and all occurrence of Y co-occur with X. Usually the value is less then [sic] 10.
- Value 0 means there is less than 1 co-occurrence of XY per 16,000 X or 16,000 Y. We can say that negative values means there is no statistical significance of XY collocation.
- Comparing two scores, plus 1 point means twice as often collocation, plus 7 points means roughly 100 times frequent collocation.
- The score does not depend on the total size of a corpus. The score combine [sic.] relative frequency of XY in relation to X and Y" (Rychlỳ, 2008: 9).

Because of its good performance logDice will be used in the current thesis for two purposes. First, a detailed investigation of distribution of a lexical item will help to identify typical collocates of translation correspondences. Second, a comparative analysis of collocation strength will be used to distinguish between more typical expressions and less typical expressions that lexical items from the same lexical domain create. In other words, the logDice test will be used to compare the typicality of collocations.

To estimate if the differences in frequency and the number of collocates with which lexical items from the same lexical domain occur are significant or not the approach of testing the null hypothesis will be applied. Here we start by formulating a null hypothesis that the differences between studied values are random and then

> "compute the probability *p* that the event would occur if H_0 were true, and then reject H_0 if is too low (typically if beneath a significance level of *p* 0.05, 0.01, 0.005, or 0.001) and retain H_0 as possible otherwise." (Manning and Schütze, 1999: 163)

It means that one does not try to prove an alternative hypothesis but rather to show that certain differences are significant by rejecting the null hypothesis. To test if the observed differences are statistically significant I will use the chi- square and t-test. The former test is suitable for the purposes of our analyses because it does not assume normal distribution of data and linguistic data is not normally distributed. The chi-square test is based on the comparison of expected and observed values in data. If the difference between the values is large it will follow that the null-hypothesis which claims that this difference is by chance can be rejected. However, this test is not suitable in cases when data contain numbers which are smaller than ten. For such cases the t-test will be used.

To explore similarities and differences between translation correspondences the relations between the following variables will be examined: the number of the items from a source language to which an item from the target language corresponds, the raw frequency of translation correspondences, the number of collocates and the values of association strength. Relations between these variables will be calculated by means of a correlation coefficient. The correlation coefficient is concerned with the following two questions: “Do high values of variable X tend to go together with high values of variable Y? Or do high values of X go with low values of Y?” (Butler, 1985: 137). In the first case we talk about positive and in the second about negative correlations. “The correlation coefficient is bounded by −1 (a perfect negative correlation) and +1 (a perfect positive correlation)” (Baayen, 2008: 87). A negative correlation means that X has high and Y low values and a positive correlation occurs when both X and Y have positive values. The closer the value of the coefficient to zero, the weaker the correlation between the variables will be.

3.2.7 Conclusion

The method described above relies on the language in use theory of meaning introduced by Wittgenstein. The theory makes it possible to avoid dealing with the unresolved issues that are intrinsic to the approaches based on the referential theory of meaning. In these approaches the meaning of terms either stand for ideas placed in the mind or to things that

are somewhere outside language. For Harris (2005: 3-4) both views are naïve and part of the same language myth because none of the two assumptions can never be proved. As such they cannot properly deal with the issue of meaning. The former is the psychocentric version and the latter the reocentric version of the myth. On the other hand, according to the language in use theory of meaning the essential question is not what the content of a lexical unit is but how it is used. Seen from this perspective, it is quite irrelevant if the speakers of a language refer to the same entity or not when they produce messages. Wittgenstein illustrates this with the following thought experiment:

> "Suppose everyone had a box with something in it: we call it a 'beetle'. No one can look into anyone else's box, and everyone says he knows what a beetle is by looking at his beetle. Here it would be quite possible for everyone to have something different in his box. One might even imagine such a thing constantly changing. But suppose the word 'beetle' had a use in these people's language? If so, it would not be used as the name of a thing. The thing in the box has no place in the language-game at all; not even as a something: for the box might even be empty. No one can 'divide through' by the thing in the box; it cancels out, whatever it is." (Wittgenstein, 1953: 100)

No one can look into anyone else's head or find some essential substances to which words can be reduced. But, this is an irrelevant issue as long as we use signs in a similar way.

I would claim that the distributional principle and the method of corpus linguistics provide a sound basis for the development of an analytic model for language study that relies on the language in use theory of meaning. It must be stressed that this model does not presuppose objectivity. In the present thesis the units of analysis will be considered the product of interactions between text corpora and available corpus and statistical tools.

3.3 Corpora

This section will describe the corpora used to study distribution of translation correspondences in the present thesis. A corpus was defined above as a collection of texts which recorded performance of language use. Two types of corpora will be required in the present thesis: a translation or parallel corpus and monolingual reference corpora for English and German. The former type presents the record of the translators' interpretation of texts from a source language by means of resources from the target language and the latter of the language use of English and German speakers. The translation corpus will be used to identify translation correspondences and the monolingual corpora will be used to handle the issue of typical contexts in which these correspondences occur. Two important issues needs to be discussed here briefly.

Ideally, corpora should be representative of a studied language but hardly any contemporary corpus meets this aim. For example, although most of the language that we use is probably spoken language, most corpora have a bias towards the written language. This is true even of the British National Corpus which usually serves as a model for other corpora: "90 percent of the BNC is writing and only 10 percent speech" (Meyer, 2002: 32). In relation to the corpora of translation texts we depend on the existing translations and they are often specific to a genre or authors. For example, Bernardini and Zanettin (2004) showed that most of the original English texts from their bidirectional English-Italian corpus were popular literary texts, whereas the Italian original texts were part of 'serious literature'. In addition, four out of ten translations studied from Italian into English are translated by the same author. It means that an investigation of these texts would tell us more about the language of that particular translator than about language use in general. As for the monolingual corpora, they are usually compiled independently from each other for the purposes of monolingual studies. This is why they are not always based on the same composition design. Even if we want to use the same composition design the question that arises is what can be considered to be comparable in two languages at all. The same genres do not necessarily share the same features in two languages. To give one example, there is no evidence that the inverted

pyramid structure of hard-news English newspaper articles exists in other languages (Thomson et al., 2008).

One issue that brought about a heated discussion in the field of translation studies is that of the so-called translationese (Baker, 1993; Baroni and Bernardini, 2006; Johansson, 2007; Mauranen and Kujamäki, 2004). The term refers to an assumption that a translated text bears certain translation features that distinguish them from purely monolingual texts. Although this hypothesis has not been confirmed so far one must be aware of the not wholly 'natural' character of translated texts. One way to make sure that translation correspondences identified in a translation corpus reflects the language use of native speakers is to use reference corpora. As said above, this approach will be adopted in the present thesis.

The corpora used in the current thesis present a compromise solution in relation to the above issues.

The Europarl corpus will be used as a parallel corpus. It is a collection of discussions extracted from the proceedings of the European Parliament (Koehn 2005). It contains parallel tests for 21 European languages. In the present thesis the version 6 compiled between November 2009 and December 2010 will be used. The last version was created after the major part of the analyses in the thesis had been completed.

The most serious disadvantage of this corpus is that we do not know in what language texts were originally composed. In addition, the corpus is not representative of English and German in general because it is biased towards the specific EU jargon. The first issue means that in future it might be necessary to check the results of the back translations. To deal with the latter issue an attempt was made to select for analysis lexical items with general meaning. The major advantage of the Europarl is its size and free availability. Translation correspondences are often multi-word units and other available parallel corpora such as INTERSECT (Salkie, 1995) do not provide enough data. For example, one of the terms to be studied *give rise to problems* occurs only twice in the latter corpus.

The reference corpus for English used in the present thesis is the ukWaC corpus. ukWaC is compiled by crawling web pages in the .uk domain it contains 1.9 billion tokens.

Although one would at first doubt its representativeness, a comparison with BNC does not indicate significant differences (Ferraresi et al., 2008). It even seems that ukWaC is more representative of spoken language than BNC. This is due to the fact that it contains text types such as emails or online forum discussions which tend to resemble spoken language. For the German a cognate of ukWaC called deWaC will be used. This corpus that consists of 1.7 billion words is also constructed from texts available on Internet and it has similar features as ukWaC. Both English and German corpora have the merit of being very large. The size of the corpora is important because some of the examined lexical items are very long and smaller corpora do not provide enough data to study them. For example, the lexical item *a considerable number of* occurs 153 times in BNC and with only three nouns on the R1 position. On the other hand, in ukWaC it occurs 1531 times and collocates with more than 400 noun collocates on the R1 position.

3.4 Data analysis procedure

Data analysis consists of three phases: i) identification of translation correspondences and lexical domains, ii) study of differences between lexical items from the same lexical domain, and iii) analysis of distribution of translation correspondences. The phases are described below.

i) Identification of translation correspondences

First a lemma word from English is selected and its German correspondences will be identified in the Euparl corpus. After that one can either classify those German lexical items that correspond to particular English collocations into the same sub-domains or simply treat them as the members of one lexical domain. Practice shows that the first option is suitable if an English lexical item has a very large number of correspondences. Classifying items into small groups makes the analysis less time-consuming than if we were to process a large number of correspondences at once. The second option is appropriate if a lexical item has less than ten translation correspondences. At the next stage the identified German

correspondences will be translated back into English. After that new back-translations from English into German and vice versa will be carried out until no new item appears in the lexical domains or sub-domains in any of the two languages.

ii) Study of differences

After all translation correspondences have been identified a comparative analysis of the occurrence of lexical items from an English and German domain and sub-domain will be conducted with the help of reference corpora. First, typical contexts in which the lexical items occur will be examined in terms of the word classes. This will be done by using the Word Sketch tool and the commands available with the IMS Corpus Workbench. These contexts will then be described with the help of the local grammar categories. Second the frequency of lexical items, the number of collocates, grammatical structures and the values of the association strength of the lexical items that belong to the same domain or sub-domain will be compared. Association strength will be based, as already mentioned, on the logDice coefficient, whereas the significance test will indicate if the observed differences are statistically significant or not. Finally, the correlation test will show if there is a strong relationship between specific variables. All this will help to establish typical distribution.

iii) Distribution of translation correspondences

Relationships between lexical items from L1 and L2 will be described in terms of the following variables. First, the number of correspondences that the lexical items from L2 will be determined. After that it will be observed to what extent each lexical item from L2 is used as a translation correspondence to the items from L1. The first variable will be called *the number of correspondence relations* and the second *correspondence degree*. The values of the two variables will be added to each other and the result will be interpreted in terms of the variable called *correspondence potential*. Here, differences within each of the variables and relations between them will be measured using the significance test and correlation test.

Each of the three procedures is linked to one of the research questions introduced in Chapter 1. The first procedure will serve to group lexical items into semantic groups or, as they will be called here, into translation lexical domains and sub-domains. This type of

grouping is an essential prerequisite for compilation of an onomasiological dictionary. The second procedure will help to answer the question regarding the cases when more than one option is available in a target language for a lexical item from a source language. It will also specify typical contexts in which lexical items correspond to each other and provide information regarding grammatical structures in which items occur. The third procedure will enable us to distinguish between more and less typical translation options.

3.5 Terms explained and conventions

In this section the terms and conventions that will be used in the present thesis will be briefly defined.

Lexical item, lexical unit or linguistic unit will be used to refer to single words and multi-word expressions.

Translation correspondence is a lexical item from a target language that occurs in the same context as a corresponding lexical unit from a source language. Translation correspondences are identified in parallel corpora.

Translation lexical domain is a set of mutually substitutable lexical items. Lexical domains may consist of more lexical domains. Translation correspondences create corresponding lexical domains in two languages.

Grammatical rules are sets of lexical patterns that summarise typical contexts in which lexical items occur.

Source and *target language* and *L1* and *L2*: Due to the nature of the Europarl corpus in the present thesis the term *source language* will not refer to the language in which a text was originally composed but to the language that serves as a point of departure for the process of establishing translation correspondences. Similarly, the term *target language* is a language in which translation correspondences are found. I will also refer to the source language as L1 and target language as L2.

All lemmata will be placed within angle brackets. A lemma can take the form both of a single word or a multi-word expression (e.g. <problem> and <give rise to problem>). The names of lexical domains will be written in capital letters and put between curly brackets. They will be named after the most frequent lexical items and represented with capital letters, fonts at the 12-point size and curly brackets (e.g. {CAUSE PROBLEM}). For the consistency reasons these labels will be used also for sub-domains. For example, in the sub-domain that will be studied in Chapter 7 the most frequent lexical item is <a number of>. However, since the more general domain is labelled <many> this label will be used also for the given sub-domain. The local grammar categories will be coded with capital letters, fonts at the 10-point size and square brackets. Alternative elements will be represented by a vertical bar (e.g. [PROCESS|DECISION] <cause> [INTENSIFIER|QUANTITY] <problem|difficulty>).

The following grammatical tags are used in the present thesis:

The English PENN tagset
CC Coordinating conjunction
DT Determiner
IN Preposition, subordinating conjunction
IN Preposition, subordinating conjunction
JJ Adjective
MD Modal verb
NN Common noun, singular or mass noun
NNS Common noun, plural
RB Adverb
TO Any use of 'to'
WDT Wh-determiner
WRB Wh-adverb

The German STTS tagset
APPR Preposition
ART Determiner

Chapter 4 Identification of translation correspondences

4.1 Introduction

This chapter is concerned with identification of translation correspondences and translation lexical domains through the observation of lexical items in the Europarl parallel corpus. The procedure consists of three steps. First a lemma from English is randomly selected and its common German corresponding items are searched for. Following the substitution assumption discussed in 3.2.2 the German lexical items that occur in same contexts are classified into same substitution sets. The items from the same set are thereafter translated back into English in order to find out if there are any other items which along with the initial English lexical item create a substitution set. Substitution sets containing the lexical items from two languages are then considered to form corresponding translation lexical domains.

The lexical item that serves as the starting point of the analysis is the lemma <rise>. It was chosen randomly because the proposed method should be applicable to any type of lexical items. What makes this particular lexical item interesting is that it is semantically ambiguous which is partly due to the fact that it can be used both as a noun and verb. One can, therefore, expect that the context in which it occurs will help to define it senses. In addition, it does not belong to the specialist jargon of the EU parliament and that it occurs frequently enough that its correspondences in German can be studied.

4.2 Identification of the TLD {CAUSE PROBLEM} and {PROBLEM BEREITEN}

<rise> occurs 5447 times in the English-German section of the Europarl corpus; 2614 times as a noun and 2833 times as a verb. The noun form was selected for further investigations and the verb was left aside. To find out if the noun occurs in an idiomatic expression Word Sketch that provides information regarding collocational strength was used. The results are displayed

in Table 4.1. The collocates of <rise> are classified into grammar classes. The image below displays four classes of collocates with which this noun most often occurs. We can see that when it occurs with verbs <rise> is used either in the subject or object position. It occurs also in the subject position when it colligates with adjectives and the verb <to be>. Finally, <rise> can be modified by other lexical items but it can also modify other nouns. Although the results obtained with Word Sketch can contain erroneous information such as listing among collocates the noun <rise> itself, the tool in general successfully summarises the context in which the given noun typically occurs. The second column below indicates the frequency of collocates and the third column displays the logDice values or the association strength of the collocations created between <rise> as a node word and other lexical items.

As we can see, the noun <rise> occurs more frequently and more typically with the verb <give> than with any other collocate.

rise *(noun)* **ukWaC freq = 76389** (48.8 per million)

object_of	30411	2.9
give	14778	8.26
chart	167	7.05
predict	234	6.59
witness	97	5.83
halt	66	5.72
see	1803	5.26
report	273	5.22
attribute	51	5.09
forecast	40	5.05
combat	48	5.05
watch	185	4.99
trace	66	4.94
stem	42	4.92
fuel	41	4.84
expect	216	4.83
chronicle	26	4.68
project	42	4.64
cause	302	4.61
blame	41	4.56

subject_of	3902	0.6
backdate	10	6.04
offset	27	5.54
parallel	7	4.89
evidence	8	4.74
accompany	60	4.71
average	10	4.43
trigger	15	4.18
coincide	7	4.17
mirror	9	3.89
fuel	9	3.86
characterise	14	3.86
squeeze	7	3.81
prompt	10	3.52
accelerate	8	3.44
slow	8	3.27
threaten	16	3.13
hit	31	3.08
predict	12	2.89
overlook	8	2.63

adj_subject_of	627	0.7
inevitable	16	4.17
due	103	3.57
equivalent	8	2.98
flat	8	2.0
likely	37	1.91
less	16	1.49
low	18	0.47
high	31	0.18
necessary	9	0.09

modifier	33115	1.4
pay	1870	8.85
sea-level	403	8.57
meteoric	359	8.44
cent	738	8.21
rapid	550	7.79
sharp	532	7.79
steady	271	7.33
dramatic	378	7.29
steep	276	7.19
sudden	183	6.73
inexorable	105	6.65
sea	535	6.62
slight	196	6.59
temperature	538	6.56
tax	733	6.53
%	1593	6.5
price	1358	6.45
sun	327	6.37
gradual	102	6.28

modifies	2285	0.1
flat	89	6.34
fall	51	4.72
block	106	4.62
apartment	27	3.78
building	157	3.14
lock	13	3.04
percent	13	3.01
tower	12	2.48
hotel	32	2.4
estate	21	2.4
festival	15	2.35
rise	15	2.31
decline	7	2.14
economy	24	1.95
living	11	1.9
working	10	1.78
housing	19	1.65
thank	8	1.14
increase	14	1.12

Table 4.1: Typical collocates of the noun <rise>

The co-occurrence of <rise> with <give> results in the idiomatic expression <give rise to>. The expression is regularly listed as an idiomatic expression in both the bilingual and the English

language dictionaries. In the electronic version of the Oxford English Dictionary it is defined as:

> "to give rise to: to be the origin of; to cause, bring about, result in." (viewed 6.4. 2011)

A definition from Macmillan English dictionary: for advanced learners reads as:

> "give rise to something to make something happen or begin, especially something or unexpected." (Rundell, 2007: 1287)

Therefore, we can select for the further analysis the whole item and investigate its translation correspondences in German.

The lexical item <give rise to> occurs 1368 times in the Europarl corpus. There are 24 corresponding German lexical items in target texts, most of which are verbs. There are three possible types of relationships between the lexical items from two languages and between translation correspondences. The first possibility is that the term <give rise to> has 24 distinct uses to which 24 German lexical items correspond. Second, <give rise to> has only one usage and 24 German lexical items are synonyms and are therefore mutually interchangeable. Third, <give rise to> has a range of uses and some of them have more than one German translation correspondence. This would mean that such translation correspondences are substitutable only in certain contexts. The further analysis will show which of the three possibilities is correct. German translation correspondences of <give rise to> are displayed in the first column of Table 4.2.

	<give rise to problem>	<give rise to concern>	<give rise to fear>	<give rise to debate>	<give rise to confusion>	<give rise to difficulty>	<give rise to doubt>	<give rise to question>	<give rise to cost>
<zu Problem|Schwierigkeit|Sorge|Verwirrung|Kosten führen>	√	√			√	√			√
<Problem|Schwierigkeit auftreten>	√					√			
<Problem|Sorge|Debatte|Verwirrung auslösen>	√	√		√	√				
<Problem|Sorge|Debatte|Zweifel|Frage hervorrufen>	√	√		√			√	√	
<zu Sorge|Angst|Debatte|Zweifel|Frage Anlass geben>		√	√	√			√	√	
<Problem|Verwirrung|Schwierigkeit|Frage|Kosten entstehen>	√				√	√		√	√
<Problem|Schwierigkeit (mit sich) bringen>	√					√			
<Problem es gibt>	√								
<zu Debatte|Verwirrung kommen>				√	√				
<mit Problem|Schwierigkeite verbunden sein>	√					√			
<Debatte stattfinden>				√					
<Problem schaffen>	√								
< für Debatte|Verwirrung sorgen>				√	√				
<Debatte provozieren>				√					
<Debatte entbrennen>				√					
<Problem sich ergeben>	√								
<Problem|Kosten verursachen>	√								√
<Ursache sein>	√								
<Problem|Schwierigkeit|Frage aufwerfen>	√					√		√	
<Angst|Zweifel aufkommen>			√				√		
<Problem|Frage sich stellen>	√							√	
<Angst wecken>			√						
<Problem|Schwierigkeit bereiten>	√					√			
<Verwirrung stiffen>					√				

Table 4.2: <give rise to> and its German translation correspondences according to the Europarl corpus

The first row shows the nouns that occur at least twice in the Europarl corpus. They create the most typical context in which <give rise to> occurs. The central part of the table indicates in which cases lexical items from two languages correspond to each other. For example, <give rise to> is translated into German as <auftreten> only when it occurs with the nouns <problem> and <difficulty>, that is, only when the German verb collocates with the corresponding nouns <Problem> and <Schwierigkeit>. Similarly, <stattfinden> corresponds to <give rise to> only when it collocates with <Debatte> and when the English lexical item co-occurs with <debate>.

The following conclusions can be derived from the above table. First of all, the correspondence relations clearly depend on the context in which <give rise to> occurs. The German lexical items do not simply correspond to the unit <give rise to>. Secondly, not all German lexical items correspond to all English collocations. For example, <führen zu> and <Anlass geben> correspond to five collocations created with <give rise to> whereas <entbrennen> and <wecken> occurs only in two translation correspondences. It follows that the former correspond to the English lexical item to the greater extent than the latter because they occur in a higher number of similar contexts as <give rise to>. It also follows that the German lexical items that do not correspond to the same collocation are not mutually interchangeble. One can, therefore, conclude that <give rise to> does not have only one usage and we cannot consider the German items to be synonyms. But, given the fact that an English collocation has more than one translation correspondence we can also conclude that <give rise to> does not have 24 distinct uses but rather that often more than one German lexical item correspond to one of its uses. It means that the relationship between <give rise to> and its German translation correspondences can be explained in terms of the third types of relations hypothesised above.

We can proceed further by selecting any of the collocations formed with <give rise to> for a further analysis. In order to determine if there are other English lexical units with a similar use we need to translate the corresponding German items into English. In this way we will identify the translation lexical domains in which the items from the two languages are used. Thus, in order to establish all domains in which <give rise to> occur it would be

necessary to focus on each of the above given collocations that it creates and its German correspondences.

Due to space restriction only one of the above collocations and the corresponding German items will be selected for the further investigation. I decided to choose the collocation <give rise to problem> because it has the largest number of translation correspondences.

According to the Europarl corpus <give rise to problem> can be translated into German by 14 different lexical items. Like the original English lexical unit the German correspondences also consist of two elements: one element that corresponds to <give rise to> and another that corresponds to <problem>. The German correspondences are displayed in Table 4.3.

German lexical items	Frequency in the Europarl corpus
<Problem bereiten>	3
<Schwierigkeit bereiten>	2
<Problem (mit sich) bringen>	4
<Schwierigkeit (mit sich) bringen>	2
<zu Problem führen>	16
<zu Schwierigkeit führen>	5
<Problem schaffen>	4
<Schwierigkeit schaffen>	1
<Problem verursachen>	8
<Problem aufwerfen>	6
<Schwierigkeit aufwerfen>	3
<Problem entstehen>	5
<Schwierigkeit entstehen>	2
<Problem ergeben sich>	4
<Ursache DET Problem sein>	4

Table 4.3: German translation correspondences for <give rise to problem>

This table also shows the frequency of the German lexical items in Europarl. Although there are in total 14 distinctive translation correspondences some elements occur in more than one context. There are only two items (e.g. the nouns <Problem> and <Schwierigkeit>) that correspond to <problem> and they are used in several constructions. Since most of the

elements that correspond to <give rise to> collocate with both nouns they also occur more than once. For example, <bereiten> corresponds to <give rise to problems> both when it collocates with <Problem> and <Schwierigkeit>. Therefore, in reality there are nine items that correspond to <give rise to problems> when they collocate with two particular nouns. One can see from Table 4.3 that the correspondences occur with different frequency. This issue will be left aside for the moment. We will rather focus on the question of other English items that occur in the same lexical domain as <give rise to problems>.

When we translate back the above German lexical items into English by using the ParaConc tolls we obtain the list of English correspondences displayed in Table 4.4.

English translation correspondences	
<cause of problem>	<pose problem>
<cause difficulty>	<present difficulty>
<cause problem>	<present problem>
<create difficulty>	<problem arise>
<create problem>	<raise difficulty>
<difficulty arise>	<raise problem>
<give rise to difficulty>	<result in difficulty>
<lead to difficulty>	<result in problem>
<lead to problem>	<there be difficulty>
<pose difficulty>	<there be problem>

Table 4.4: English lexical items corresponding to the German lexical units displayed in Table 4.3

Since there are 20 additional lexical items in English that correspond to the same items as <give rise to problems> we can conclude that we deal with many-to-many relationship. One can notice that the above table contains the collocation <give rise to difficulty> that was previously encountered in Table 4.2. It follows that this lexical unit and <give rise to problem> are also mutually substitutable. Several additional back translations were performed in both directions to ensure that the list includes all relevant cases. Some new items identified in German are: <Problem darstellen>, <Schwierigkeit darstellen>, <Problem auftreten>, <Schwierigkeit auftreten>, <es gibt Problem>, <es gibt Schwierigkeit> and <problematisch sein>. Similarly, new items found in English are <to be problematic> and <it be a problem>.

The lexical items from both languages that will be examined in the following chapter are displayed in Appendix A in Tables A1 and A2. The first table contains English expressions that correspond to German lexical items and the second table provides translations of English lexical units. The tables also indicate the frequency of both the translated items and their correspondences. Their correspondences are ordered by frequency. Both the translation items and their correspondences are displayed as lemmata. In reality, any element can take different word forms. For example, the nouns <problem> and <difficulty> can be used both in plural and singular. In German there are also complex noun compounds such as in <Gesundheitsproblem> or <Wirtschaftsproblem>. Similarly, verbs can be used in the word forms such as *cause, causes, caused*. Finally, these expressions are not fixed. Thus, the nouns <problem> and <difficulty> and the corresponding German nouns can be modified by an adjective, noun or other multi-word expressions. Therefore, the lexical item <create problem> contain also the cases such as <create considerable problem>, <create serious problem> or <create health problem>.

Since the purpose of the present study is to explore typical translation correspondences two criteria have been introduced. First, the items that occur only once will not be taken into consideration. Second, only the lexical items that correspond to at least two lexical units from another language will be considered. Finally, the items that correspond to less than two percent of the occurrence of an item from L1 will be ignored. This is an AND condition, which means that the study will deal only with the items which meet all three criteria.

4.3 Conclusion

In the previous section, a set of translation correspondences in English and German were identified by comparing local contexts in which lexical items occur. Following the terminology introduced in Chapter 3 we can say that these correspondences create substitution sets. These substitution sets have two levels. From the cross-linguistics point of view the lexical items from L1 are substitutable with the lexical items from L1. Seen from the monolingual perspective the lexical items from the same language are also mutually substitutable since they occur in the same context. As earlier said, the technical term for such sets is *translation lexical domains* (TLD). These domains are similar to the traditional lexical fields or to Tognini-Bonelli's (2001: 150) web of translation units, Viberg's (1983; 1993; 2004) and Dyvik's (1998 2004; 2005) semantic fields. However, the TLD are not based on researchers' intuition and the referential theory of meaning. Rather they derive from the distributional analysis of the occurrence of lexical units in a parallel corpus.

The above identified domains consist of the lexical units in English and German that correspond to each other. For this reason, we can say that these domains are corresponding lexical domains. The name of the English domain will be {CAUSE PROBLEM} and the name of the German domain will be {PROBLEM BEREITEN}. In choosing labels I follow Apresjan's (2000: 217) suggestion not to use artificial terms but words from the object language that are intuitively comprehensible. The domains in question are labelled according to the most frequent items which will be explained in the next chapter.

It should be added that other similar TLD can be established by using the same method. The only difference is that other collocations created with <give rise to> would serve as a starting point. By starting, for example, from <give rise to costs> we would first need to identify all its German correspondences. A translation of these correspondences into English would show if there are other English lexical items that together with the initial term create a substitution set. Further back-translations would finally reveal all members of this translation lexical domain. After all domains in which <give rise to> occurs have been established one could study the relations between them. At this stage it is too early to speculate on the

results. However, given the fact that some German lexical items correspond to more than one collocation of <give rise to> one can expect that at least some of the created domains would be part of a larger domain. Furthermore, it would be a technical decision whether to refer to them as macro domains or to treat the lower level domains as sub-domains. This issue will be dealt with in Chapter 7. Provisionally, the sets identified above will be referred to as translation lexical domains.

In Chapter 5 we will proceed by investigating whether the distributional approach can provide distinguishing features that would make it possible to differentiate between translation correspondences which belong to the same TLD.

Chapter 5 Lexical items from the TLD {CAUSE PROBLEM} and {PROBLEM BEREITEN}

5.1 Introduction

In Chapter 3 we defined both the correspondence relations between the lexical items from two languages and the relations between lexical items belonging to the same translation lexical domain in terms of substitutability. Thus, it was said that a translation correspondence is a lexical item from L2 that can substitute for an item from L1. The two occur in similar contexts with the difference that these contexts are in two different languages. Similarly, two lexical items belong to the same domain if they are mutually substitutable. It is again the context in which they occur that plays a crucial role.

In the present chapter the distribution of lexical items from the TLD {CAUSE PROBLEM} and {PROBLEM BEREITEN}, which have been identified in the previous chapter, will be examined both from an intranlinguistic and interlinguistic perspective. The results will show if the model of these intralingual and interlingual investigations successfully implements the differentiation principle.

The intralinguistic analyses will reveal typical contexts in which lexical items occur and their local grammar rules. Comparisons of these contexts will show if the model can provide a set of purely distributional distinguishing features for the given lexical items. Positive results will mean that we can replace previous approaches based on the referential theory of meaning with a new approach that relies on the observation of language in use. The distributional distinguishers will point at the differences between the items from the same domain. More generally, positive results should show to what degree an item is substitutable by another item from the same domain. As explained in Chapter 3, in the current thesis language is considered to be of a probabilistic nature. Therefore, the results which will follow from analyses will be interpreted in terms of patterns and tendencies.

As for the occurrence of translation correspondence, the analyses aim at finding how the occurrence of the lexical items from L2 differ in relation to their uses as correspondences for the items from L1. The results will indicate to what degree the items from L2 are

substitutable for those from L2. To begin with, the distribution of English and then of German lexical items will be studied. The data in the monolingual analysis derive from ukWaC and deWaC for English and German data, respectively. The analysis concerned with relations of lexical items in two languages will be based on the data from the Europarl parallel corpus. The results will be displayed using the conventions introduced in 3.5.

5.2 TLD {CAUSE PROBLEM}

The study of the distribution of lexical items that belong to the current TLD unfolds in the following way. First, the general grammar structures associated with the lexical items will be established. The corpora that were tagged by parts-of-speech for both languages simplify this above task enormously. The frequent structures, subsequently, will serve as the point of departure for establishing local grammar categories. Distributional distinguishers will follow from a comparative analysis of the occurrence of items belonging to these grammar categories. In particular, I will here compare the frequency of lexical items and the values of the collocation strength. The correlation test and chi-square test will be used to determine the strength of relationships between the studied variables and significance of identified differences.

5.2.1 Grammar structures

5.2.1.1 A local grammar of the lexical items from the TLD {CAUSE PROBLEM}

Table 5.1 contains ten English lexical items from the TLD {CAUSE PROBLEM}. These items meet the pre-defined conditions explained in the previous chapter, e.g. they correspond to at least two German lexical items and occur at least two times.

Lexical items	
<cause *problem*>	<present *problem*>
<create *problem*>	<*problem* arise>
<give rise to *problem*>	<raise *problem*>
<lead to *problem*>	<result in *problem*>
<pose *problem*>	<there be *problem*>

Table 5.1: Lexical items from the TLD {CAUSE PROBLEM}

We will start investigating differences and similarities between the current items by looking at their general grammatical properties. All lexical items consist of a verbal and a nominal element. Nominal elements contain two nouns that can take four word forms: *problem, problems, difficulty* and *difficulties*. As will be shown below, these four word forms are not equiprobable. The verbal elements are transitive verbs in all but two cases. The exceptions are the intransitive verb <arise> and the existential <there be>. The difference between the transitive and intransitive verbs can be explained in terms of syntactic roles. The clauses with intransitive constructions involve only one participant which syntactically occurs in the subject position. In our case this participant is either the noun <problem> or <difficulty>. Transitive constructions, on the other hand, involve two or three participants. Here, in addition to the subject, clauses also contain a direct object. It is in this position in which one of the two nouns is used. The verbs <cause> and <present> are sometimes also used ditransitively, that is, with an indirect object. The analysis below will show types of items that occur in this position. Another difference between current transitive verbs is that some of the (<cause>, <create> and <pose>) can also be used in passive. <there> which is used in the existential constructions is a so-called dummy subject. With this lexical item a speaker states that a problem or difficulty exists without specifying how it came into being.

Differences between verbal elements can also be explained in terms of what Halliday (1994:37) calls “thematic structure”. This structure concerns how we organise a textual message. Two main elements in the structure are called *Theme* and *Rheme*. “The Theme is the element which serves as the point of departure of the message; it is that with which the clause is concerned” (Halliday, 1994: 37). It is followed by a Rheme, the element in which the information from the Theme is developed. Thus, in the constructions formed with the

transitive verbs from the present TLD the point of departure is the piece of information in which the sources of problems or difficulties are stated. The message is further developed by verbal and nominal elements. On the other hand, with <arise> it is either <problem> or <difficulty> that serves as the point of departure. The Rheme is the verb itself and other items that might follow it. In the existential construction <there be>, <there> serves as a Theme. Here, "the point of departure is precisely the fact that a participant… is to be introduced" (Gómez-González, 2001: 124). In our case, the introduced participant is <problem> or <difficulty>.

The context in which lexical items from the present domain occur can be also accounted for in terms of general grammatical structures. These structures have been identified by sorting the context into parts-of-speech categories. This task was performed by using the function 'sort by pos' available with the IMS Corpus Workbench tools. At first the usual span of five lexical items to the left and right of the verbal and nominal elements was selected. However, subsequent analysis indicated that the relevant context consisted of one or two items preceding the verbal and following the nominal elements and from one to three items inserted between the two elements. The following table illustrates how the initial structures looks like for the context in which <cause problem> occurs.

Frequency	**Grammar structures**			
1212	MD	<cause>	<problem>	IN
280	RB	<cause>	<problem>	IN
228	TO	<cause>	<problem>	IN
118	MD	<cause>	<problem>	RB
105	MD	<cause>	<problem>	WRB
93	NN	<cause>	<problem>	IN
88	MD	<cause>	<problem>	DT
84	MD	<cause>	<problem>	CC
76	NN	<cause>	<problem>	IN
68	MD	<cause>	<problem>	JJ
60	MD	<cause>	<problem>	TO
60	WDT	<cause>	<problem>	IN
59	NNS	<cause>	<problem>	IN

Table 5.2: Grammatical structures for <cause problem>

The numbers on the left-hand side indicate the frequency of patterns. The grammatical categories are coded according to the English PENN tagset as already explained in 3.5. For example, the first pattern indicates that one context in which <cause problem> occurs is a modal verb on the left-hand side and prepositions on the left-hand side. Subsequently, these patterns have been closely investigated and compared for every single item and across the whole domain and as a result we have obtained a repertoire of typical colligations.

The contexts identified in this manner are divided into two general groups according to their functions. The first group consists of lexical units which modify verbal elements and the second of those which modify nominal elements. In the first group we find modal verbs and adverbs. Nominal elements can be preceded by a determiner or modified by an adjective or another noun phrase. In addition, it can be followed by a prepositional phrase which also has a modifying function. The lexical units containing <cause> and <present> also occur with noun phrases used as indirect objects. Schematically, the combination with colligations can be summarised in the following way:

[MD] [ADV] <CAUSE^{TR}> [NP] [DT|NP|ADJ] <problem> [for NP]
<CAUSE^{EX}> [MD] [ADV] [NP|ADJ] <problem>
[NP] [NP|ADJ] <problem> [MD] [ADV] <CAUSE^{ITR}>

These classes present typical contexts in which the current lexical items are found in my corpus. Or in Wittgenstein's terms, they present a language game associated with this set of lexical units.

As explained in Chapter 3, the square brackets indicate that an element is optional, whereas the horizontal bar means that either of the elements occurs in the given position. The superscripts specify the type of verbal elements. This representation is, however, too general and does not indicate the functions of the items belonging to the specific word class. Therefore, a further specification that would reveal local functions and the character of these elements is required.

First, a description of the lexical items that precede <problem> and <difficulty> will be provided. The purpose of these lexical items is to specify the meaning of the nouns <problem> and <difficulty>. Here we can distinguish between determiners on the one hand and modifiers on the other. The definite determiners that occur here are <the>, <these> and <this>. The indefinites are <a|an>, <some> or zero determiner that we find with the plural form of <problem> and <difficulty> and the singular form of <difficulty>. Modifiers are either adjective or noun phrases. We will see below that items from both groups can be combined in various manners.

Modifiers encompass a very large number of lexical items. They can be classified in terms of the functions that they perform into several local grammar classes. These classes are: INTENSIFIERS, QUANTIFIERS, SORTALS and COMPARATORS. The functions of each of them will be described below. Some of the most frequent members of the classes are displayed in Table 5.3.

INTENSIFIERS	QUANTIFIERS	SORTALS	COMPARATORS
big	a few	access	additional
considerable	a great range of	behaviour	another
enormous	a lot of	communication	certain
great	a number of	engineering	different
huge	a series of	environmental	distinct
key	a small number of	ethical	further
large	all kind of	financial	new
major	all sort of	health	other
minor	fewer	legal	particular
serious	many	logistical	same
severe	more	management	similar
significant	numerous	noise	special
small	several	operational	typical
substantial	some	performance	unique
subtle		political	various
		pollution	
		practical	
		safety	
		security	
		technical	

Table 5.3: Local grammar classes of modifiers that occur with <problem> and <difficulty>

The function of INTENSIFIERS is to specify to what extent a problem or difficulty is serious. On the one hand, there are items denoting a high degree of intensification, and on the other those that denote a low degree of intensification. The INTENSIFIERS occur with both the singular and plural form of <problem> and <difficulty>. The most frequent positive INTENSIFIERS that occur with all or almost all lexical items from the present lexical domain are <serious>, <significant>, <severe>, <considerable>, <great> and <real>. Three relatively frequent negative INTENSIFIERS are <minor>, <small> and <subtle>. INTENSIFIERS usually co-occur directly with <problem> or <difficulty> but they can also combine with some SORTALS such as: <health>, <financial>, <environmental>, <welfare>, <technical>, <logistical>, <safety>, <practical>, and <social>. This is exemplified in the following two sentences.

1. Failure to guarantee this could **cause considerable financial difficulties** for tenants…
2. The injected substance **poses serious health problems**, even with limited use.

The lexical items from the category QUANTIFIERS describe if problems are large or small. They occur with the plural form of <problem> and <difficulty>. QUANTIFIERS are mostly one-word long but they occasionally also take the form of multi-word expressions. Some of the most frequent QUANTIFIERS from the positive end are <many>, <numerous>, <a number of>, <a great range of>, <a series of>. On the negative side there are <a few>, <fewer>, <a small number of>. In-between are <several> and <some>. QUANTIFIERS either precede the nouns <problem> and <difficulty> directly or collocate with the members of other three categories such as in the following two examples.

3. This has **resulted in a number of serious problems** to society.
4. Unfortunately **there were a few technical problems** with the sound during this act…

The category SORTAL has more members than any other local grammar class. The function of this type of lexical items is to specify what kind of problems or difficulties we deal with.

SORTALS classify problems and difficulties into kinds, which usually have to do with health issues (e.g. <health>, <breathing>, <hearth>, <skin>, <liver>, <eye>, <sleeping> or <dental>), communication (e.g. <language>, <communication>, <email>, <access> or <understanding>), security (e.g. <safety>, <security>, <health and safety>, <flooding>), technology (e.g. <technology>, <engineering>, <navigational>, <operational> or <technical>) and various socio-economic issues (e.g. <traffic>, <social>, <behavioural>, <unemployment>, <financial>, <environmental>, <economic>, <pollution> and <noise>). As one can see, the lexical items from this group are either nouns or adjectives. SORTALS are mostly one-word long but they can occasionally be modified by the items from other three groups such as in (4) above.

COMPARATORS are items that show if problems or difficulties discussed in a given message are of the same type as those which arose in other contexts. These other contexts can be explicitly mentioned or inferred from the message itself. Thus, in (5) one can infer that the underlined items from the first two sentences denote certain problematic situations and that the noun <problem> from the third sentence refers to other similar situations. On the other hand, in (6) in the first sentence the activity was explicitly named a problem and the second sentence indicates that further problems may also occur.

5. There were calls to deal with a roof blowing off a house in Middlemarch Road, Radford, and a tree falling on a car in Holyhead Road.

Station officer Danny Moynihan said crews from the station were also called to help out in Solihull, which was hit harder.

Warwickshire police reported more than 100 weather-related calls in the north of the county, as flooding in some areas **caused further problems.**

6. Clearance can be a problem. Pins can **create some other problems.**

The nouns <same>, <similar>, <different> and <distinct> can also be used to compare whether a problematic activity or event produces the same effects on different recipients. In (7) men and women face different problems.

7. These varying circumstances **pose different problems** for women and men.

One can here again draw a distinction between a positive and negative pole. There are fewer positive than negative COMPARATORS. Three most representative items of the former type are <same>, <similar> and <typical>. The difference between the three is that only the last term involves the sense of regularity. Among the negative COMPARATORS the most frequent ones are <different>, <new>, <other>, <another> and <particular>. What distinguishes them is that the items <particular>, <unique>, <specific> and <special> are more specific than <new>, <different>, <other> or <certain> which only denote general difference. The lexical items from this group mostly occur directly with <problem> or <difficulty>, but some items may be preceded by INTENSIFIERS or PLURAL QUANTIFIERS or followed by SORTALS.

The most frequent determiners that occur with <problem> and <difficulty> are the articles <the> and <a>, the demonstrative pronoun <this> and the plural zero determiner. More will be said about these determiners in the next section.

Finally, the singular form of the noun <problem> is in some cases followed by a post-modifier. Two types of prepositional phrases are found here: <of NP> and <with NP>. The former is usually preceded by the lexical item *the problem* and the latter by *a problem*. Both serve to provide more information about types of problems and have therefore the same function as SORTALS. Thus, *the problem of access* from (8) could also be rephrased as *an access problem* as one can see from (9) without changes in meaning. Similarly, *a problem with the camera* from (10) can be transformed into *a camera problem* as in (11). The prepositional phrases formed with <with> are more frequent. Both types of post-modifiers occur mostly with <there be>.

8. However, **there is** still **the problem of access**: how do people reach the waterfront?
9. One issue I would flag is that **there is an access problem**.
10. My initial thought was that **there was a problem with the camera**.
11. To determine if **there is a camera problem**, all the variables

We will now look at lexical modifiers of verbal elements which are realised as modal verbs, adverbs and adverbial expressions. There are three types of modality judgments we can find here: probability, usuality and duration.

The probability items express to what degree a speaker is certain that problems will arise. Here, we can distinguish between the items that express a higher level of certainty (<will>, <would>, and <should>) from those that express a lower level of certainty (<can>, <may>, <could> and <might>). It is interesting that the probability adjuncts such as <probably>, <certainly>, <maybe> do not occur here very often and that probability is mainly expressed through modal verbs. The item <should> differs from other modal verbs because it is mostly used in a negative form such as in (12). This type of modals will be coded as PROBABILITY_OPERATORS.

12. For all normal purposes this **should not pose a problem**…

The usuality items express judgment regarding how often a problem or difficulty occurs. On the one end, we find <always> and <often> which denote that they occur often, and on the other end there is the item <never> that expresses the opposite meaning. Between these two extremes are <usually>, <sometimes>, <from time to time>, and <occasionally>. The usuality items will be coded as USUALITY.

The final class of modality items express how persistent a problem or difficulty is. This meaning is realised by means of adverbial expressions. Again, on the one end there are the items indicating continuity (<still>, <most of the time>), and on the other those that denote the opposite sense (<no longer>). The items performing this function will be denoted as DURATION.

Regarding the distribution of verbal modifiers, PROBABILITY_OPERATORS mostly collocate directly with verbal elements but some also co-occur with the USUALITY and DURATION items. The items from the latter two groups are mutually exclusive.

It was mentioned above that the verbs <cause> and <present> are occasionally used with an indirect object. An indirect object occurs between the verbal element and the nominal element. The lexical items used as indirect objects are usually the object pronouns

us, you and *them* or the general nouns denoting a mass of individuals such as <people>, <population> or <students. The expressions created with <present> undergo minor modifications when they occur with the indirect object by being extended with the preposition <with> as in (13). The same meaning is also expressed with the prepositional phrase <for NP> which follows the nominal element. Nouns used in this prepositional phrase are usually those referring to government-related institutions (e.g. <government>, <authorities>, <planning body>, <community council>, <Civil Service>, <the Queen>), members of a community (e.g. <the elderly>, <farmers>) or to the general public (e.g. <people>, <majority of the population>, <visitors>). In contrast to the indirect object constructions the prepositional phrase colligates with all lexical items. The items that occur as indirect objects or in the prepositional phrase <with NP> will be marked as RECIPIENT.

13. As in Donne's satire, they **present us with problems** of conflicting authority…

An attempt has also been made to identify the classes of items that occur in the subject position with transitive verbs. However, no regularities have been observed apart from the fact that the noun phrases that occur in this position denote activities, events or processes and not denote animate subjects. For this reason, I will refer to such terms with the general term THING.

5.2.1.2 Conclusion

In the above section general properties of the lexical items belonging to the English TLD {CAUSE PROBLEM} were studied. After identifying two main elements that compose the lexical items, the type of clauses in which they occur were briefly discussed. Differences between them were explained in terms of transitivity and textual metafunction. In the second part, the local grammar classes specific to the lexical items from the present lexical domain were described. These classes are based on the functions that linguistic items perform in

relation to the nominal and verbal element. Using the categories and metalanguage introduced above and in Chapter 3 the distribution of the classes can be summarised in the following way:

THING [PROBABILITY_OPERATOR] [CONTINUITY|USUALITY] CAUSE^{TR} [RECIPIENT] [*INTENSIFIER*|QUANTIFIER|SORTAL|COMPARATOR] PROBLEM [RECIPIENT]
CAUSE^{EX} [PROBABILITY_OPERATOR] [CONTINUITY|USUALITY] [INTENSIFIER|QUANTIFIER|SORTAL|COMPARATOR] PROBLEM
INTENSIFIER|QUANTIFIER|SORTAL|COMPARATOR] PROBLEM[PROBABILITY_OPERATOR] CAUSE^{INTR}

Table 5.4: Local grammar structures for lexical items from the TLD {CAUSE PROBLEM}

The above structures are a schematic representation of local grammar rules and they yield a slightly oversimplified picture. The purpose of this representation, however, is to display relative positions of the classes in relation to each other rather than to cover all possible combinations. Below, we will see how the members of these classes are typically distributed.

5.2.2 An intralinguistic analysis of items from the TLD {CAUSE PROBLEM}

5.2.2.1 General distribution

Above we identified typical contexts in which the lexical items from the present TLD occur. Now, we can explore the assumption that differences between semantically closely related lexical items can be established on purely distributional grounds. This will be done in two ways. First, typical contexts common for two or more lexical items will be compared. The comparison will show if these contexts are significantly different. Second, it will be explored if there are lexical items that occur in unique contexts. In both cases, verbal and nominal modifiers will be examined.

We will begin by examining the general raw frequency of lexical items from the present TLD. The frequency data include the occurrence with <problem> and <difficulty> as

well as with the members of all local grammar classes. The results are represented in Table 5.5.

Lexical items	Raw frequency
<there be problem|difficulty>	17763
<cause problem|difficulty>	10338
<problem|difficulty arise>	5793
<present problem|difficulty>	3575
<create problem|difficulty>	3062
<pose problem|difficulty>	2673
<lead to problem|difficulty>	2378
<raise problem|difficulty>	662
<result in problem|difficulty>	386
<give rise to problem|difficulty>	284

Table 5.5: Raw frequency of English lexical items in ukWaC

There are some obvious differences between the figures in the above table. The lexical item <there be problem|difficulty> is 1.7 times more frequent than the next most common <cause problem|difficulty>. In addition, it occurs almost 50 times as frequently as the least frequent item from the list. To examine the significance of these differences the chi-square test was used. The following numbers can be observed: χ^2 = 57869.53, df = 9, p-value < 2.2e-16. The results indicate that differences are larger than it would be the case if the data were randomly distributed. The value of 2.2e-16 is the smallest value that can be calculated accurately with the programming language R and is far below the predefined threshold (0.05). Similarly, the value of the chi-square test is much higher than the critical value for the degree of freedom of 9 which is 16.7.

Table 5.6 displays the co-occurrence of verbal elements with the nouns <problem> and <difficulty> and in Table 5.7 one can observe the occurrence with four word forms (*problem, problems, difficulty* and *difficulties*). One can notice that the percentage values which describe the occurrence of verbal elements in both tables vary considerably.

Lexical items	Co-occurrence with <problem>	Co-occurrence with <difficulty>
<there be>	87%	13%
<cause>	91%	9%
<arise>	76%	24%
<present>	80%	20%
<create>	90%	10%
<lead to>	81%	19%
<pose>	90%	10%
<raise>	82%	18%
<result in>	68%	32%
<give rise to>	66%	34%

Table 5.6: Co-occurrence of verbal elements with the lemmata <problem> and <difficulty>

The lemma <problem> tends to co-occur with verbal elements with higher frequency than the lemma <difficulty>. The average value for the co-occurrence with the former noun is 81% and with the latter only 19%. But there are some individual differences between verbal elements here. Thus, <give rise to>, <result in> and <arise> are most common with <difficulty> and <cause>, <create> and <pose> are least typical verbal elements in this context. The figures above show that every third or fourth occurrence of the former verbal elements in the present corpus is with <difficulty> and with the latter it is every tenth or eleventh occurrence.

According to Table 5.7, verbal elements occur more frequently with the plural form of <problem> and <difficulty> than with the singular form. The median value for the co-occurrence with the plural form of the first noun is 75% and there are seven out of ten items that occur with this or higher value. The median for the plural form of the second noun is 77% and we find six out of ten items that have this or a higher percentage value. The occurrence of <cause> with the singular form of the two nouns is extremely seldom and <lead to> is also not very typical in this context. This context is, on the other hand, most typical of <there be>. The verbs <present> and <pose> occur almost equally probably with the two word-forms of <problem>.

Lexical items	Co-occurrence with *problem*	Co-occurrence with *problem*	Co-occurrence with *difficulty*	Co-occurrence with *difficulties*
<there be>	57%	43%	40%	60%
<cause>	0.7%	99.3%	0.2%	99.8%
<arise>	24%	76%	26%	74%
<present>	46%	54%	23%	77%
<create>	21%	79%	14%	86%
<pose>	45%	55%	23%	77%
<lead to>	8%	92%	17%	83%
<raise>	25%	75%	23%	77%
<result in>	12%	88%	31%	69%
<give rise to>	13%	87%	31%	69%

Table 5.7: Co-occurrence of verbal elements with four word forms of the lemmata <problem> and <difficulty>

5.2.2.2 Modifiers of verbal elements

After this general description of the co-occurrence with the two nouns, a detailed analysis of the distribution of lexical items from the local grammar classes will be carried out. First, the items which modify verbal elements will be studied.

The general distribution of the constructions with modal verbs is summarised in Figure 5.1. The modal verbs in question are <can>, <may>, <could>, <might>, <will>, <would> and <should>. The analysis shows that no strong relationship can be established between the frequency of lexical items and the distribution of modal verbs. The correlation coefficient for these two variables is r=-0.23. Thus, among four lexical items that most often occur with modals (between 32% and 38% of the time) we find two items that have low frequency (<result in problem|difficulty> and <give rise to problem|difficulty>), one medium frequency word (<lead to problem|difficulty>) and one word with high frequency (<cause problem|difficulty>). It seems, therefore, that the occurrence with modals is a property of individual lexical items.

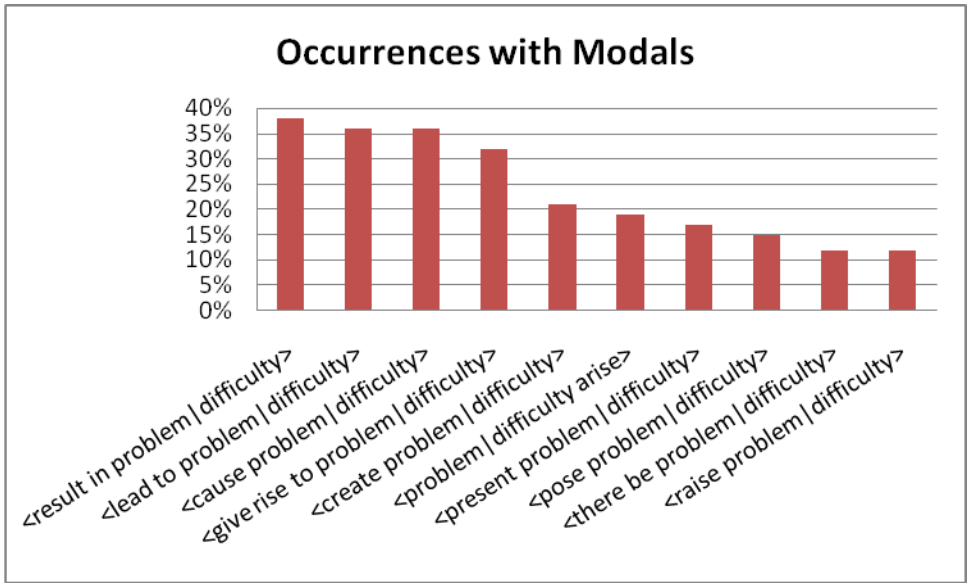


Figure 5.1: Co-occurrence of lexical items from the TLD {CAUSE PROBLEM} with modal verbs

With the exception of the expressions formed with <there be> and <raise> the item <can> is the most frequent modal in the present context. Collocations with <can> account for between 11% and 25% of all occurrences of the first four lexical items from the table with modal verbs. The modals <could>, <may>, <will> and <would> occur equally likely with all lexical items. One exception is <arise> which is most typical with <may>. The modal verbs <should> and <might> occur with low frequency in the present context.

The most frequent items from the class USUALITY are <often> and <always>. These items are used to stress that problems or difficulties occur regularly as a result of some other events or activities. The next in the series are <usually> and <frequently> and the items that denote rarity are not used very often. There are no significant differences with regard to the general co-occurrence of the USUALITY items with lexical units from the present field.

The items from the class DURATION are less common than the USUALITY items. The most typical item from this class is <still> and it most typically occur with <cause problem|difficulty>, <present problem|difficulty> and <pose problem|difficulty>

The above description can be summarised in the form of the following features that can be used to distinguish between lexical items from the present domain. The first feature observed is that <result in problem|difficulty>, <lead to problem|difficulty>, <cause problem|difficulty> and <give rise to problem|difficulty> occur more often with the set of

modals studied above. The second feature is that the DURATION item <still> most typically occurs with <cause problem|difficulty>, <present problem|difficulty> and <pose problem|difficulty>. Finally, the lexical item <arise> most commonly collocates with <may> and is the only item that occurs in the conditional expressions with <should> and <if>. On the other hand, no important differences are observed with regards to the USUALITY items. Similarly, the distribution of the particular modals is uniform across the field.

As we saw above the category RECIPEINT is realised in two ways in the current data. The lexical items from this class are used either as an indirect object in which case they occur between nominal and verbal elements, or as the prepositional phrase <for NP> when the items follow the nominal element. As for the distribution of the items used in the first position the most common are expressions with <cause problem|difficulty>. 6% of all occurrences of this lexical item are used with the indirect object. The indirect object is here mainly realised as <us> or <you>. Similarly, 3% of all the occurrences of <present problem|difficulty> are found with an indirect object. The co-occurrence with <pose problem|difficulty> is much less typical and combinations with other lexical items are not found in the ukWaC corpus. On the other hand, the expressions with prepositional phrases occur with every item but with very low frequency. Two far more typical combinations are <pose problem|difficulty for NP> and <create problem|difficulty for NP>. These two items colligate with the prepositional phrase <for NP> in more than one-quarter of all cases. Apart from two intransitive lexical items and <give rise to problem|difficulty>, which are very infrequent here, other expressions have similar distribution.

5.2.2.3 Modifiers of nominal elements

In this subsection the distribution of lexical items that modify nominal elements will be explored. Due to greater complexity of their distribution a more detailed analysis will be needed than in 5.2.2.2. The analysis proceeds in the following way. First, the occurrence without modifiers for both <problem> and <difficulty> will be examined. Here, both the plural

and singular forms will be studied separately. Thereafter, the distribution of modifiers will be compared. Here, both the number of modifiers and the values of collocation strength will be investigated. The analysis will also show if the differences between the distribution of particular classes of modifiers are significant.

The nouns <problem> and <difficulty> can be specified by a determiner, modifier or both. Determiners specify whether the nouns are of a specific or general kind and whether modifiers, as we have seen above, provide information about the two nouns. There are certain modifiers that in the present context can occur with and without determiners.

Figure 5.2 displays the percentage of the occurrence of the two word forms of the lemma <problem> without modifiers or without the combination of a determiner plus a modifier in the current corpus. As far as the singular form of <problem> is concerned the expressions without modifiers comprise between 55% and 88% of its total distribution. The expressions formed without modifiers are most typical of <give rise to>. The figures for plural constructions range between 20% and 72%. It means that modifiers in general play a more important role with the plural than with the singular form of <problem>. The graph indicates that modifiers are most common here with <create> *problems* and <raise> *problems* and that the expressions without modifiers are most common with *problems* <arise>.

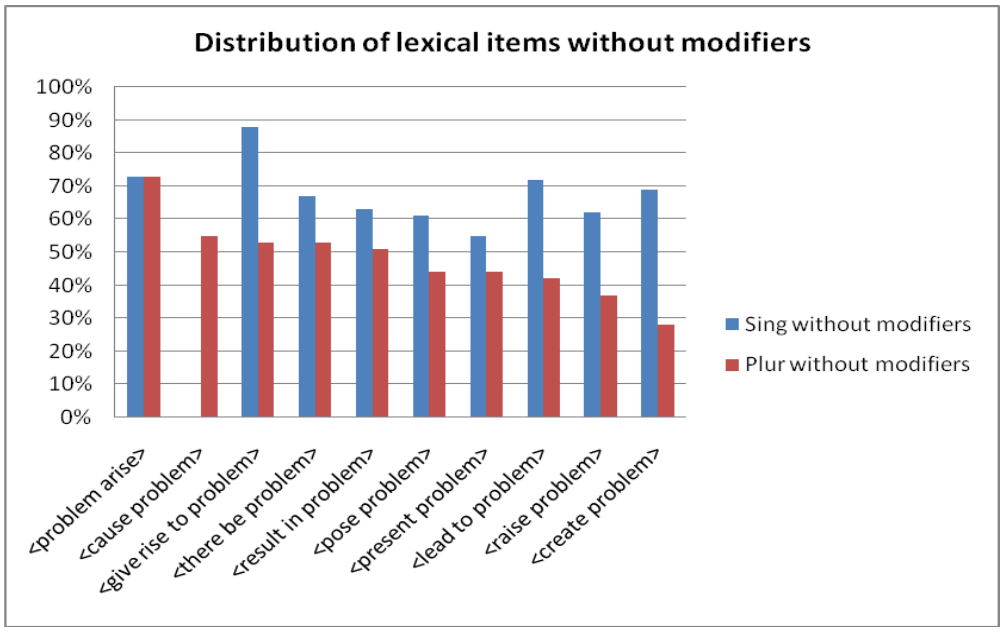


Figure 5.2: Distribution of lexical items that occur with the noun <problem> without modifiers

Further differences emerge when particular combinations are examined. The items that precede both word forms of <problem> are either definite/indefinite determiners or the negation item <no>. Definite determiners are not typical of <there be>, <present>, <pose> and that they are more common with <arise>, <raise> and <create> when these items collocate with *problem*. The opposite is true when it comes to the collocations formed with *no problem*. The indefinite *a problem* is most typical with <there be>, <create>, <present> and <pose>. The indefinite constructions make almost half of the occurrences with the word form *problem* for these lexical items.

Definite determiners are infrequent with the word form *problems*. Collocations with <no> are more specific of <there be>, <create>, <present> and <pose> than of other items. Two verbal elements that tend to occur less typically with the zero plural indefinite determiner are <create> and <raise>. No significant differences can be observed for other lexical items.

A similar analysis was also carried out for the expressions formed with the noun <difficulty>. Figure 5.3 displays results regarding its distribution without modifiers. The expressions formed with *difficulty* and without modifiers range from 32% to 95%. The highest number of modifiers occur with <there be>, <present> and <pose> and in other cases they play only a minor role. One feature that was not observed with *problem* is the occurrence of *difficulty* in the expressions that contain neither a determiner nor a modifier such as in (14) and (15). This word is used here as an uncountable noun.

14. ...returning to work from maternity leave Job share in supervisory or managerial roles may **lead to difficulty** for staff working with two managers.
15. However, **there could be difficulty** if urgent access to the airway is required.

With the plural form *difficulties* modifiers occur in between 32% and 50% of all cases. The figures that represent the occurrence without modifiers are very similar to the above results. One exception is <there be> which occurs less frequently in the expressions formed with

<difficulty> and without modifiers than in the expressions formed with <problem> and without modifiers. The verb <cause> in both cases avoids constructions with the singular form of the two nouns.

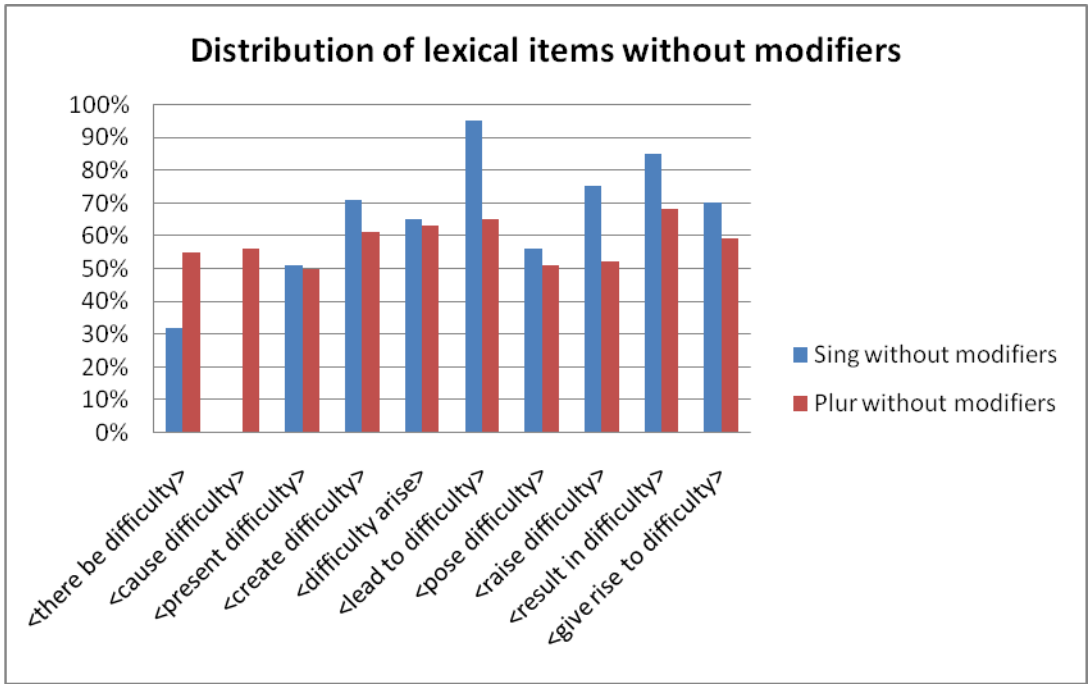


Figure 5.3: Distribution of lexical items that occur with the noun <difficulty> without modifiers

Most typical combinations with *the difficulty* are <arise> and <raise>. As we saw above, these two verbs and <create> were also the most frequent collocates of *the problem*. Similarly, the same three items as above (<there be>, <present> and <pose>) are the most frequent collocates of <no>. Indefinite constructions are most strongly associated with the verbs <create>, <result in> and <lead to>. Here, however, the indefiniteness is more often realised with the zero determiner rather than with <a>.

One important difference between the distribution of the word form *difficulties* and *problems* in the current context is that the former item very seldom occurs with <no>. The only significant occurrence in our data is recorded with <present> and <pose>. Another difference is greater prominence of the definite article <the> with <arise> and <raise>. What is common for both word forms is that the most frequent expressions are formed with the

zero plural indefinite determiner. Thus, in about half of all cases the word form *difficulties* occurs in this form.

The following distinguishing features are identified in the above analysis:

- Modifiers occur typically in the constructions formed with <create>, <raise> and <lead to> and *problem* on the one hand, and with <create> and *difficulties*, on the other. They are very infrequent with <pose> *problems* and <result in> *difficulties*. Modifiers are in general less often used with the singular form of the noun elements. One exception is <there be> *difficulty*.
- The most typical definite constructions are <create> *the problem, the problem|difficulty|difficulties* <arise> and <raise> *the problem|difficulty|difficulties*.
- The most frequent indefinite expressions are <there be> *a problem*, <create> *a problem*, <present> *a problem*, <pose> *a problem, <create> difficulty,* <result in> *difficulty* and <lead to> *difficulty.* It is worth mentioning here that the second most frequent verbal element <cause> does not occur with the singular form of the nouns <problem> and <difficulty>.

We will now consider the distribution of modifiers in relation to different word forms of the nouns <problem> and <difficulty>. First, modifiers that occur with *problems* will be examined.

Table 5.8 provides a summary regarding the distribution of modifiers that occur with *problem.* The first column gives a list of lexical items from the present TLD that occur in this context. The curly bracket and numbers indicate the span within which modifiers occur. The second column shows the frequency of the expressions formed with the word form *problems* and modifiers, whereas the third column indicates the number of modifiers found with individual lexical items. The span defined is based on the observation of grammatical structures as displayed above in Table 5.1.

Lexical items	Frequency with modifiers	Number of modifiers
<cause> {1,3} *problems*	4396	405
<there be> {1,3} *problems*	3827	345
{0,3} *problems* <arise>	1085	166
<lead to> {1,3} *problems*	1017	195
<create> {1,3} *problems*	1002	137
<present> {1,3} *problems*	933	133
<pose> {1,3} *problems*	790	92
<raise> {1,3} *problems*	259	49
<result in> {1,3} *problems*	137	43
<give rise to> {1,3} *problems*	85	25

Table 5.8: The number of modifiers and frequency of lexical units that collocate with the word form *problems*

The first feature observed here is a strong correlation between the frequency of lexical items and the number of modifiers. The more frequent a lexical item is, the larger number of modifiers it will have. This is confirmed by the correlation test that yields a very high value of r=0.97.

The second important fact is that there are significant differences in data for both variables as the following results of the chi-square test indicates. The chi value is very high and the p-values are far below the predefined threshold of 0.05.

```
data: frequency with modifiers
χ² = 15124.63, df = 9, p-value < 2.2e-16
data: number of modifiers
χ² = 915.8365, df = 9, p-value < 2.2e-1
```

The results also indicate that more frequent items tend to occur with the majority of collocates of less frequent items. The degree of overlap is greater for the items that considerably differ in their frequencies. In other words, the overlap tends to be smaller for lexical items that occur with similar frequency and greater for those with extreme values. Finally, shared modifiers tend to be the most frequent collocates. The frequency of shared modifiers and the degree of overlap for the lexical items investigated is illustrated in Figure 5.4. The former feature is denoted by the colour blue and the latter by red. It can be seen that the degree of overlap increases as we move from left to right, that is, from the items with

higher to those with lower frequency. The figure displays only the graphs for the two most frequent and the two least frequent items. A complete list of graphs is provided in Appendix B in Figure B1.

We can also observe that although <cause> {1,3} *problems* occurs with only slightly more than 30% of the collocates found with <there be> {1,3} *problems* these items make up more than 60% of the total frequency of constructions formed with modifiers.

There are a few minor exceptions to these general patterns. For example, the overlap between modifiers that occur with <present> {1,3} *problems* and with <pose> {1,3} *problems* is larger than between the two items and <result in> {1,3} *problems* although the latter is less frequent. It indicates greater similarity between the former item and the greater difference between the two former items and the latter item. Similarly, the behaviour of <there be> {1,3} *problems* is more similar to that of {1,3} *problems* <arise> than to that of <lead to> {1,3} *problems* although the latter is less frequent than the expressions formed with <arise>.

The tendency that shared collocates have very high frequency can be illustrated through the following two cases. 87 out of 137 modifiers that occur with <create> {1,3} *problems* collocate also with the more frequent <cause> {1,3} *problem*s. Out of 50 modifiers that occur only with <create> {1,3} *problems* 42 items occur twice and eight occur three times. Other items occur between five and 150 times, or on average 12 times. Similarly, although only 39% of modifiers that occur with <raise> {1,3} *problems* also occur with <lead to> {1,3} *problems* those items that are found only with the former item occur only twice. On the other hand, shared items occur between three and 22 times or on average seven times. In general, 20% of the most frequent collocates account for 60% or more of all expressions formed with modifiers.

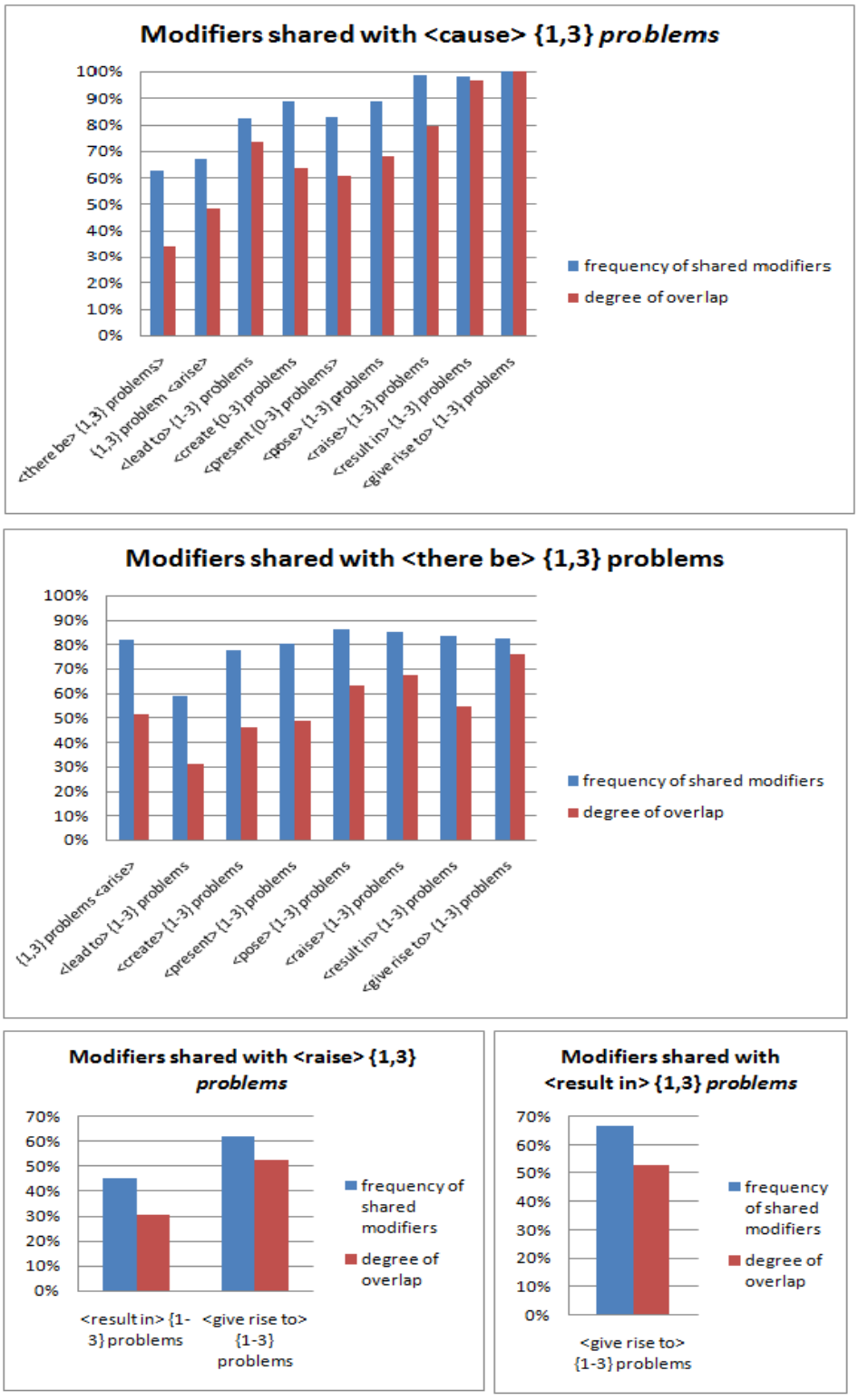


Figure 5.4: Frequency and degree of overlap of shared modifiers that occur with the word form *problems* in the TLD {CAUSE PROBLEM}

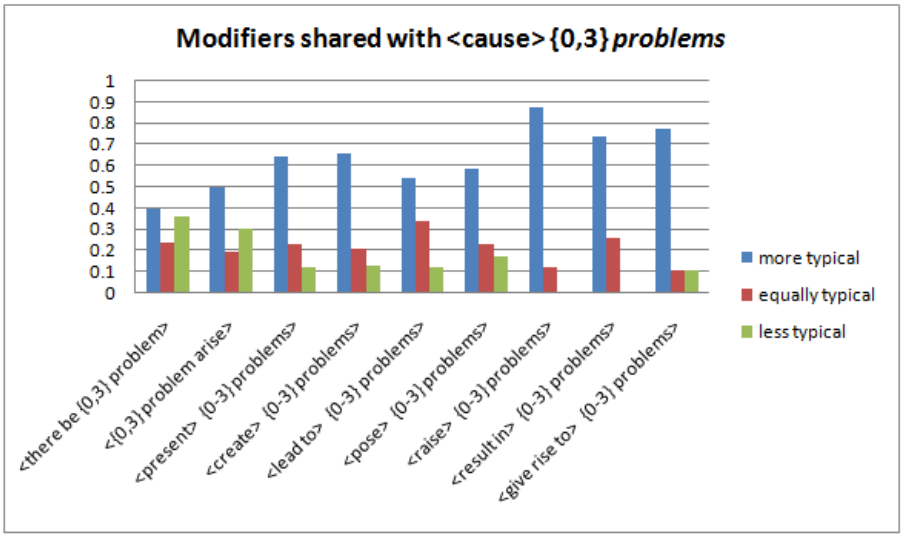
The great majority of combinations with modifiers are made up of only one lexical item (between 95% and 100% of the occurrence of all lexical items). The most frequent modifiers are the items from the local grammar class QUANTIFIERS: <many>, <a lot of>, <a number of>, <several>, <some>, and <a few>. Other frequent items include <new> <other>, <real>, <further>, <particular> <further> and the INTENSIFIERS <serious>, <significant> and <major>. Among the most frequent SORTALS we find <health>, <technical>, <financial> and <environmental>. The multi-word modifiers mostly consist of two modifiers where we typically find combinations of a QUANTIFIER or INTENSIFIER and a SORTAL such as in: <many social>, <a number of financial>, <serious health> and <significant environmental>.

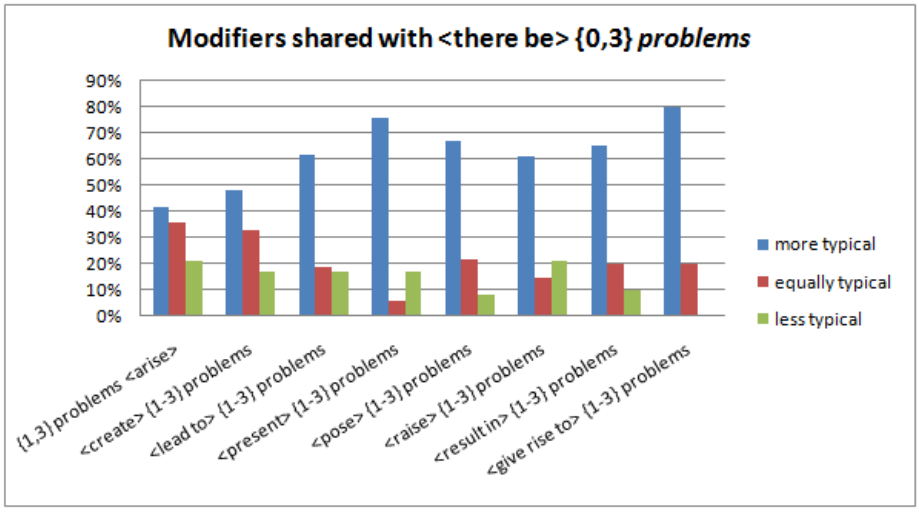
A closer look at the distribution of shared modifiers reveals the following tendencies. The number of stronger collocations is almost always larger for the items with higher frequency. This result follows from a comparative analysis of differences between the values of association strength for the shared items. In addition, the number of stronger collocations tends to increase as the frequency difference between two items becomes larger. Finally, the number of modifiers that have similar or lower values in these cases tends to be more similar. This is illustrated with Figure 5.5 that contains the data for the same lexical items as the previous figure. The graphs for other lexical items are provided in Figure B2 in Appendix B. The values of association strength are based on the logDice values.

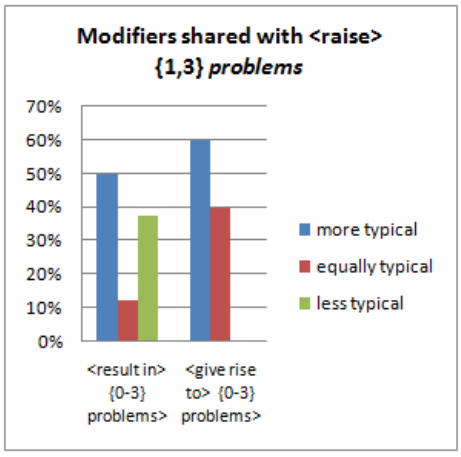
There are some exceptions to the tendencies identified above. Thus, <lead to> {1,3} *problems* is more frequent than <present> {1,3} *problem* although the number of stronger collocations is larger for the latter lexical item. Similarly, <cause> {1,3} *problem* has a larger number of stronger collocations compared to <create> {1,3} *problem* than when compared to <pose> {1,3} *problem* although the second item is less frequent than the first one. <pose> {1,3} *problem* in general tends to create stronger collocations than one would expect from its frequency.

The above results provide additional information which can help to distinguish between the uses of lexical items from the present lexical domain. The investigation of the distribution of lexical items without modifiers revealed some individual differences between these items. The present analysis, on the other hand, points to the differences of a more

general nature. The differences are explained in terms of tendencies. The results suggest that more frequent items can usually substitute for the less frequent ones. The opposite is not true because this will often result in non-idiomatic expressions.







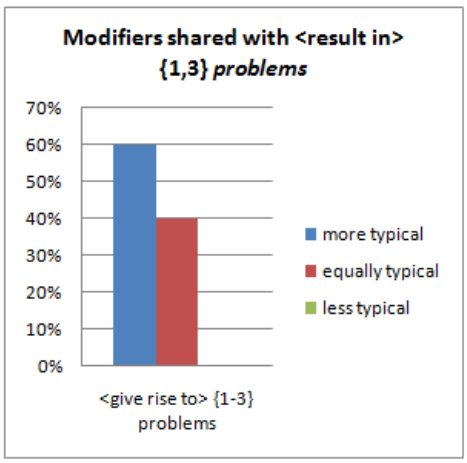


Figure 5.5: Association strength values for the expressions made up of verbal elements, shared modifiers and *problems*

The next analysis will be concerned with the co-occurrence of modifiers with the word form *problem*. First we will consider the variables frequency and the number of collocates. Similar to the results that were based on the data displayed in Table 5.8, the analysis of the lexical items from Table 5.9 indicates that in the current data there is a strong relationship between the two variables (r=0.99). Because of their very low frequency and a small number of modifiers we can exclude from the analysis the lexical items with <cause>, <result in> and <give rise to>.

Lexical items	**Frequency with modifiers**	**Number of modifiers**
<there be> {1,3} *problem*	2852	414
<present> {1,3} *problem*	568	89
<pose> {1,3} *problem*	417	75
{1,3} *problem* <arise>	334	50
<create> {1,3} *problem*	160	51
<raise> {1,3} *problem*	52	20
<lead to> {1,3} *problem*	40	22
<cause> {1,3} *problem*	28	5
<result in> {1,3} *problem*	12	7
<give rise to> {1,3} *problem*	3	4

Table 5.9: Frequency and the number of modifiers for lexical units that collocate with the word form *problem*

The chi-square test indicates that the differences within each of the two columns are significant:

```
data: frequency with modifiers
χ² = 9459.69, df = 6, p-value < 2.2e-16
data: number of modifiers
χ² = 1132.660, df = 6, p-value < 2.2e-16
```

The same general tendencies as above can be observed again. More frequent lexical items tend to occur with a significant portion of modifiers of the less frequent items. The coverage increases when the difference in frequency between two items rises. Thus, <there be> {1,3} *problem* collocates with about 50% of modifiers that occur with <present> *problem* and

<pose> *problem* and with more than 75% of those that co-occur with other lexical items. Like above, shared collocates are at the same time the most frequent modifiers and they usually cover between 70% and 80% of all occurrences of lexical items in this context.

The range of modifiers is here more limited than it was above. The most frequent are the following INTENSIFIERS: <major>, <serious>, <big>, <huge>, <significant> and <significant> and COMPARATORS <specific>, <particular>, <same>, <similar>, <different>, <new>, <another>. The most frequent SORTALS are <security>, <health>, <medical>, <access>, and <technical>. In the vast majority of cases modifiers are preceded by an indefinite article. This tendency is not observed with <same> which occur with <the> and <major> and <big> which may occur with both a definite and indefinite determiner. Modifiers are almost always one-word long. If the two-word units occur they are usually combinations of an INTENSIFIER or a COMPARATOR and a SORTAL such as *a serious health problem* or *another financial problem*.

A comparative analysis of the values of association strength yields results similar to those from above. Usually, more frequent lexical items create the most typical collocations. It means that the strongest collocations are formed with <there be> *problem*, <present> *problem* and <pose> *problem*. The proportion of more typical collocations increases with the difference in frequency. Thus, almost all modifiers that are shared by <there be> *problem* and <raise> *problem*, <lead to> *problem* are more typical with the former item. On the other hand, out of 42 modifiers that occur with both <present> {1,3} *problem* and <pose> {1,3} *problem* 13 are more typical with the first item, 11 with the second and 18 are equally typical with both.

Now we will look at the distribution of the modifiers that occur with the lemma <difficulty>. Again, the plural form of the noun will be considered first.

Table 5.10 displays the frequency of lexical items formed with modifiers and the number of associated modifiers. The correlation between the two variables is equally strong as in the previous two cases (r=0.98).

Lexical items	Frequency with modifiers	Number of modifiers
<there be> {1,3} *difficulties*	623	92
<cause> {1,3} *difficulties*	390	74
<present> {1,3} *difficulties*	372	60
<create> {1,3} *difficulties*	198	47
{1,3} *difficulties* <arise>	167	35
<lead to> {1,3} *difficulties*	126	25
<pose> {1,3} *difficulties*	112	22
<raise> {1,3} *difficulties*	32	12
<result in> {1,3} *difficulties*	29	8
<give rise to> {1,3} *difficulties*	29	10

Table 5.10: The number of modifiers and frequency of lexical units that collocate with the word form *difficulties*

The chi-square test indicates that the frequency differences are significant in this case, as well.

data: **frequency with modifiers**
χ^2 = 1660.364, df = 9, p-value < 2.2e-16

data: **number of modifiers**
χ^2 = 196.5844, df = 9, p-value < 2.2e-16

If we compare the data with those from Table 5.8 that contain collocations created with the plural form of the noun <problem> we notice certain differences in the upper part of the current table. The expressions created with <there be> are here one and a half times more frequent than that with <cause>. Almost the opposite was the case in the previous table. Similarly, the constructions formed with <present> and <create> are more frequent than those formed with <arise> and <lead to>. The opposite is true when these verbs collocate with *problems*. Also, the frequency values of the lexical items formed with <cause> and <present> on the one hand and with <lead to> and <pose> on the other are more similar than was the case above.

However, these exceptions do not alter the general tendencies observed in the previous two cases. Thus, <there be> {1,3} *difficulties* occurs with 30% of modifiers that collocate with <cause> {1,3} *difficulties* and with more than 60% of modifiers that one finds

with the four items from the bottom of Table 5.10. The relationship between degree of overlap and frequency is similar with other lexical items as well. Shared collocates are again most frequent modifiers. The linguistic unit <cause> {1,3} *difficulties* is used with not more than 70% of the collocates of other lexical items but these common collocates account for between 72% and 96% of total occurrence of modifiers.

The most frequent modifiers are the QUANTIFIERS <some>, <many>, <a number of>, and the INTENSIFIERS <serious>, <major>, <significant>, <great>, <considerable> and <enormous>. The last three items are more typical here than they were with the noun <problem>. Frequent COMPARATORS in the present context are <additional>, <further>, and <particular> and among the most prominent SORTALS are <breathing>, <communication>, <financial> and <learning>. It should be noted that the multi-word modifiers are extremely rare here.

We can again observe the trend that more frequent lexical items are associated with the larger number of stronger collocations. Also, the lexical items with similar frequency values and a similar number of shared collocates have similar distribution. A comparison of the distribution of <cause> {1,3} *difficulties* and <present> {1,3} *difficulties* with that of <present> {1,3} *difficulties* and <raise> {1,3} *difficulties* will illustrate this. There are seven stronger collocations formed with each of the two items from the first pair and seven with similar association strength. On the other hand, eight out of nine modifiers are more typical with <present> {1,3} *difficulties* than with <raise> {1,3} *difficulties* and one item occurs equally likely with both these items.

The word form *difficulty* in the current data is very infrequent compared to other three word forms. When we exclude extremely infrequent lexical units only five items remain for further analysis. The results for these five items are displayed in Table 5.11.

Lexical items	Frequency with modifiers	Number of modifiers
<there be> {1,3} *difficulty*	624	48
<present> {1,3} *difficulty*	96	20
{1,3} *difficulty* <arise>	53	13
<pose> {1,3} *difficulty*	27	11
<create> {1,3} *difficulty*	24	6

Table 5.11: The number of modifiers and frequency of lexical units that collocate with the word form *difficulty*

Due to a very small number of items it is not possible to perform a correlation test and the chi-square test but it seems that two variables comply with each other. We can observe that both the number of collocates and the frequency of items decreases. With the exception of <pose> {1,3} *difficulty* and <create> {1,3} *difficulty* other items occur at least twice as frequently as the next most common unit. <there be> {1,3} *difficulty* occurs with almost all modifiers that collocate with the last three items from the table and with 60% of collocates of <present> {1,3} *difficulty.* Similarly, <present> {1,3} *difficulty* share between 80% and 100% of modifiers with the less frequent items. The lexical items that they share occur with a very high frequency. Among the most frequent modifiers are the COMPARATORS <further>, <particular>, <new>, <another> and the INTENSIFIERS <great>, <serious>, <main> and <considerable>. All expressions are single-word units and apart from <great> and <main> they are preceded only by the indefinite article <a>. The adjective <great> occurs with both <the> and <a> and <main> only with <the>. Finally, the strongest collocations are found with more frequent lexical items while the collocations that contain low frequency items are less typical.

5.2.2.4 Unique collocates

Above we studied textual contexts which are common for all lexical items from the present domain. In the current subsection the focus will be on unique constructions. This analysis will show how numerous such constructions are and if they can be used as additional

distinguishing features. Only those constructions which are relatively frequent will be taken into account.

The prepositional phrase <with NP> in the present context occurs only with <there be> and we find it typically following *a problem*. The most numerous nouns that occur in this prepositional phrase refer to various features of technology as the following items illustrate: <brake>, <car>, <camera>, <email address>, <server>, <printer>, <motherboard>. Common are also the nouns related to delivering service such as: <order>, <ticket>, <delivery>, <account>, <booking>. The nouns from the latter set is usually preceded by *your*. The collocation <there be> *the problem* also occurs with the prepositional phrase <of NP|V-ing>. The noun <problem> here serves as the head of a noun phrase that consists of another noun or a gerund. These nouns typically denote some health or social issues (16): <alcoholism>, <apathy>, <racism>, <obesity>, <abuse>. The most frequent verbs are about completing some tasks or achieving something (17): <get>, <find>, <have>, <determine>, <obtain>, <define>. The meaning of these verbs is further specified through the expressions that refer mainly to successful completion of tasks. A similar meaning can be expressed with the construction <there be> *a problem* <of how to V>..

16. In America **there is the problem** of obesity due to the fast food generation.
17. Then **there's the problem** of getting it past EU competition authorities.

30% of the time the verb <arise> occurs in the word form *arising.* This form is used in a noun phrase in which <problem> or <difficulty> serve as its head and *arising* as a postmodifying element. The most frequent are expressions formed with the word form *problems*. The whole noun phrase is usually followed by <from NP>. This additional element provides information about the reasons that led to problems or difficulties. It is not possible to identify specific categories of the nouns that occur in this prepositional phrase. Semantically, this prepositional phrase has the same function as nouns that occur in the subject position with the transitive verbs from the present domain. Finally, the whole noun phrase is usually

preceded by a verb denoting mitigation of unpleasant situations (18-19). Some of these verbs are: <deal>, <cope>, <tackle>, <sort out>, <overcome>, <resolve>, or <prevent>.

18. The group struggles to overcome **problems arising** from the pantomime and…

19. This way the store owner will be able to deal with any **problems that may arise** from them being placed in a new environment…

The verb <arise> collocates also with <should> or <if> and they together form conditional clauses. There are three general forms of this type of utterance. In the first type <should> occurs at the beginning of the sentence as in (20). The second type consists of both <should> and <if> such as in (21). The most common is, however, the third type that contains only <if> as in (22).

20. **Should any problems arise,** please inform your supervisor and/or visiting tutor.

21. There is a local part-time caretaker who can be contacted in an emergency **if a problem should arise** with the building.

22. Examine feet regularly and seek medical attention **if any problems arise.**

These conditional expressions occur in the service-oriented or advice-giving sort of communication. These clauses are typically accompanied with an independent clause in which advice is provided on how to tackle a problematic situation.

Finally, in the current corpus we also find one construction which is typical of the verb <create>: <more… than it solve>. The use of this expression is illustrated by the concordance lines below. Although the construction in question occurs also with <cause> and <present> such combinations are far less common. In about 70% of the cases when <create> collocates with <more> it is used in this construction. The noun <problem> occurs more frequently than the noun <difficulty>.

```
rains alone can sometimes <create more problems than> it solves . Because of t
ew such an approach would <create more problems than> it solved . In fact they
configuration control can <create more problems than> too few as visibility of
m of violence which would <create more problems than> it would solve . If seve
bout the second world war <creates more problems than> it solves . They
s . " However , this idea <creates more problems than> it solves , by far . I
u warned , the acts could <create more problems than> they solved .
thugs , on duty , who can <create more problems than> they prevent . This is p
, but I thought it might <create more problems than> it solved . My own perso
previously been suggested <create more problems than> they solve . Thus , even
more valued - or do they <create more problems than> they solve top What coun
e practical problems . It <created more problems than> it solved . The social
, " the reasoning goes , <creates more problems than> it solves : Buddhist l
arbours , said this could <create more problems than> it solved . "
increased choice does not <create more problems than> it solves . This may cre
resent economic model has <created more problems than> benefits for
s or could the technology <create more problems than> it solves . I
igh density housing could <create more problems than> it solves . In some coun
, Fuller and Stecker 1997 <creates more problems than> it solves . The archae
```

5.2.2.5 Conclusion

In previous sections we explored whether a detailed analysis of local contexts in which the lexical items from the TLD {CAUSE PROBLEM} occur can help to identify features that can be used to distinguish between these items. Unlike in the approaches that rely on the referential theory of meaning where such features are based on semantic components, the current approach proceeded from the assumption that the features can derive from the observation of the distribution of lexical items. The following distributional distinguishing features were identified at several levels.

- First, we distinguished between transitive, intransitive and existential expressions. These differences are explained in terms of what serves as a point of departure in constructing a message. We saw that some but not all transitive verbs can be used in passive constructions.
- After that, significant differences between the general frequency of lexical items were observed. The difference in frequency was also observed when the focus was on particular word forms of the lemmata <problem> and <difficulty>.
- An analysis of verbal elements revealed that only two lexical items occurred with an indirect object. A further investigation indicated that not all items were equally probable in these colligations. We also saw that lexical items differed in terms of their occurrence with the items expressing modal meaning. The greatest

differences, nevertheless, were discovered when we turned to the contexts in which the nouns <problem> and <difficulty> occurred. Here, we observed that the probability of lexical items occurring with and without modifiers differed.

- The analysis of the expressions with nominal modifiers indicated a strong correlation between the frequency of lexical items, the number of modifiers and the values of collocation strength.
- The results showed that when lexical items co-occurred with the same collocates that those with higher frequency had higher substitution potential than those with lower frequency. This tendency was recorded with all four word forms of the nouns <problem> and <difficulty>.
- Finally, it was observed that three lexical items were associated with several unique constructions that were atypical of other lexical items.

Table 5.12 summarises the above results in terms of the distinguishing features observed. In more particular, it shows the behaviour of lexical items with regard to transitivity, occurrence in the passive, occurrence with the lexical items from the local grammar category RECIPIENT and with modal verbs. The typicality of occurrence with the modals in questions is displayed in terms of ranking, with the value one indicating the highest degree of typicality. The term *typicality* here refers to the association strength of the combinations examined. The differences observed are explicitly represented in this table. For example, one can see that <create problem|difficulty> shares the feature transitivity with seven other verbs. A combination of the features transitivity and passive voice is found only with <create problem|difficulty> and <cause problem|difficulty>. The table also shows that RECIPIENTS occur with <create problem|difficulty> only in the form of the prepositional phrase <for NP> and that <create> is the fourth most typical verb that occurs with modal verbs.

Lexical items	Transitivity	Passive	RECIPIENT	Typicality of occurrences with modal verbs
<there be problem|difficulty>	EX			8
<cause problem|difficulty>	TR	√	INDIR/FOR+NP	2
<present problem|difficulty>	TR		INDIR/FOR+NP	6
<create problem|difficulty>	TR	√	FOR+NP	4
<problem|difficulty arise>	INTR			5
<lead to problem|difficulty>	TR			2
<pose problem|difficulty>	TR	√	FOR+NP	7
<raise *problem*|difficulty>	TR		FOR+NP	8
<result in problem|difficulty>	TR			1
<give rise to problem|difficulty>	TR			3

Table 5.12: Distinguishing features for lexical items from the TLD {CAUSE PROBLEM}

In a similar fashion, we can also summarise the results that describe the co-occurrence of four word forms of the nouns <problem> and <difficulty> and modifiers. Table 5.13 displays frequency and association strength for the lexical items made up of *problem* or *problems* and determiners or modifiers. For the sake of simplicity all differences are displayed in terms of ranking. The value one means that the collocation in question has the highest association strength, the value two indicates a slightly less typical collocation and so on.

As with the previous table, Table 5.14 displays differences in terms of ranking for combinations of verbal elements with the singular or plural form of the lemma <difficulty> and with determiners or modifiers.

In general, these results can be interpreted in terms of the substitution potential of lexical items. Substitution potential indicates to what degree one lexical item can replace other linguistic units from the same domain. Thus, it can be said that the top ranked lexical items or the items with higher frequencies and with a larger number of collocates have higher substitution potential than the lexical units associated with lower values. Substitution potential of lexical items may vary depending on how typically they occur in a given context. For example, <present> has higher substitution potential than <cause> when it occurs with *problem* and *difficulty* but the opposite is true when it collocates with *problems* and *difficulties*. Substitution potential, therefore, can help to find similarities and differences between the uses of lexical items that belong to the same domain.

Variable Lexical items	Frequency *problem*	Typicality <no> + *problem*	Typicality Definite + *problem*	Typicality Indefinite + *problem*	Typicality Modifiers + *problem*	Frequency *problems*	Typicality <no> + *problems*	Typicality Modifiers + *problems*
<there be>	1	1	6	1	2	2	1	2
<cause>	9				1	1	2	1
<present>	3	2	6	3	5	5	3	6
<create>	5	4	4	2	4	7	2	3
<arise>	2		2	5	3	3		4
<lead to>	6		3	4	4	4		5
<pose>	4	3	5	2	6	6	4	7
<raise>	7	4	3	6	7	8		8
<result in>	8		3	7	8	9		9
<give rise to>	10		1	4	9	10		10

Table 5.13: Distinguishing features for the expressions formed with *problem and problems*

Variable Lexical items	Frequency *difficulty*	Typicality <no> + *difficulty*	Typicality Definite + *difficulty*	Typicality Indefinite + *difficulty*	Typicality Modifiers + *difficulty*	Frequency *difficulties*	Typicality <no> + *difficulties*	Typicality Definite + *difficulties*	Typicality Modifiers + *difficulties*
<there be>	1	3	3	7	1	1			1
<cause>	10					2			2
<present>	2	1		6	2	3	1		3
<create>	4			3	4	4			4
<arise>	3		2	9	3	5		1	5
<lead to>	5			1		6			6
<pose>	6	2		5	4	7	2		7
<raise>	8		1	7		8		2	8
<result in>	7			2		9			9
<give rise to>	9			4		10			9

Table 5.14: Distinguishing features for the expressions formed with *difficulty* and *difficulties*

Substitution potential also indicates the position of a lexical item in a TLD which is related to the structure of lexicon. As we saw in Chapter 2, in previous studies this issue was explored by means of the notion of basic (Lehrer, 1974) or core words (Viberg, 2002) and by exploring how general the meaning of terms that belong to the same field is. High values of substitution potential, nevertheless, should not be confused with the generality of meaning. The generality of meaning is inseparable from the referential theory of meaning and cannot be dealt with from the language in use perspective. Substitution potential indicates the generality of use and not of meaning. This issue will be touched upon in more again in 5.2.3.2.

5.2.3 An interlinguistic analysis of the items from the TLD {CAUSE PROBLEM}

5.2.3.1 General principles

The previous analysis was concerned with the substitutability of lexical items belonging to the TLD {CAUSE PROBLEM}. The analysis was carried out from an intralingual perspective. The present section explores the distribution of lexical items from the domain in question from an interlingual perspective. In particular, the analysis will compare the use of English lexical items as translation correspondences for German items. The results that followed from the intralingual analysis were interpreted in terms of their substitution potential. Similarly, the results of the interlingual investigation will indicate the correspondence potential of lexical items.

The investigation will be based on two variables. The first variable is the number of lexical items from L1 to which an item from L2 can correspond. Here, I will start from the null-hypothesis that the items from the same TLD have an identical number of correspondence relations. The second variable is concerned with the percentage with which an item from L2 is used as a translation correspondence. Again, it will be assumed that all translation correspondences are used at an equal rate. Thus, if we suppose that there are five items from L2 that correspond to an item from L1, according to the null-hypothesis each will be used 20%

of the time. Correspondence potential will be calculated as the sum of values of two variables.

The number of correspondence relations that one lexical item establishes will show its general substitutability. If an item from a TLD in L2 corresponds to all items from a TLD in L1 we will conclude that it has very wide usage. On the other hand, if an item corresponds only to say two items from L1 it will follow that it has very restricted usage as a translation correspondence.

Without knowing how often one item can replace another term we have only a partial picture of the behaviour of lexical items. Thus, it is possible that an item with wider usage is never used in more than five percent of the time as a translation correspondence. On the other hand, if an item corresponds once to 60% and once to 70% of the occurrence of the item from L1 this will indicate very common usage of the given item in this context. This is why the percentage of usage needs to be taken into account. Clearly, not all percentage values are equally significant and 60% of correspondence is more consequential than five percent. In order to calculate these values in respect to their importance, different values will be assigned to different percentages. This is displayed in Table 5.15. The importance of assigned values increases with the percentage values. The highest assigned value will have a L2 lexical item that corresponds to the items from L1 between 90% and 100% of the time.

Percentages	Assigned values
1-9%	1
10-19%	2
20-29%	3
30-39%	4
40-49%	5
50-59%	6
60-69%	7
70-79%	8
80-89%	9
90-100%	10

Table 5.15: The values that describe the percentage of use of lexical items as translation correspondences

In the next step, the assigned values are calculated by adding the number of times an item is used with the given percentage and by finding its average value. For example, if an item is used twice with the percentage below 10% and once with 33% its final value will be 2 since: (2x1)+(1x4)=6/3=2. The former numbers in the parentheses indicate the number of items from L1 to which an item from L2 corresponds. The latter numbers indicate the percentage of its uses as a translation correspondence. Finally, correspondence potential in this case will be 5 because the item is used three times and the percentage of its use as a translation correspondence converted into assigned values is two.

I will here examine the distribution of only the most frequent lexical items. In more particular, the study will deal only with the translation correspondences that contain the plural form of the noun <problem>. As has been the case so far, only the items that correspond to at least one German lexical item in at least three percent of the cases will be included in the analysis.

5.2.3.2 Correspondence potential of English translation correspondences

The English lexical items from the present domain do not correspond to the same number of German items. The differences can be seen in the fourth column of Table 5.16 below. According to the chi-square test these differences are statistically significant: p-value=0.03. The average number of established correspondences is six. Three items have a higher and four a lower number of correspondence relations. A lexical item which corresponds to the largest number of translation correspondences is <cause> *problems*. It corresponds to all but one lexical item from the TLD {PROBLEM BEREITEN}. Three items that correspond only to two German lexical items are <to be problematic>, <lead to> *problems* and <result in> *problems.*

The second column shows the sum of the percentage with which the current lexical items are used converted into the assigned values. We can see that these values are not identical. The p-value for this set of data is again more below the critical level: 6.76e-09. There is even a correspondence between these values and the values that describe the

number of correspondences (r=0.88). However, a comparison with the next column indicates that the items correspond to a different extent to the German items. For example, the total value in the second column for <give rise to> *problems* is seven and for <raise> *problems* five. From the fourth column we can see that the former corresponds to seven and the latter to only three German items.

Lexical items	**Correspondence degree:total**	**Correspondence degree:average**	**Number of correspondence relations**	**CP**
<cause> *problems*	24	2.2	11	13.2
<there be> *problems*	24	2.4	10	12.4
<create> *problems*	18	1.8	10	11.8
problems <arise>	16	2.7	6	8.7
<give rise to> *problems*	7	1	7	8
<pose> *problems*	7	1.2	6	7.2
<present> *problems*	6	1	6	7
<raise> *problems*	5	1.7	3	4.7
<to be problematic>	5	2.5	2	4.5
<lead to> *problems*	4	2	2	4
<result in> *problems*	4	2	2	4

Table 5.16: Distribution of translation correspondences from the TLD {CAUSE PROBLEM} in relation to the items from the TLD {PROBLEM BEREITEN}

Relatively high numerical values for the lexical items from the bottom of the table (<to be problematic>, <lead to> *problems*, <result in> *problems*) are due to the fact that each of them corresponds at least once with a high percentage to the items from the TLD {PROBLEM BEREITEN}. But, only the first four items correspond to more than one German lexical item in more than 10% of the cases. Interestingly, apart for the verbs <give rise to>, <present> and <result in> all other items are used at least once in more than 10% of the cases as translation correspondences. In these cases the lexical items serve as the most frequent translation correspondences of German terms. For example, <raise> *problems* is the most preferred translation correspondence for *Probleme* <aufwerfen> and for <zu> *Problemen* <führen> it is <lead to> *problems*. Other differences can be read off from the table as well. For example, we can observe that <present> *problems* has higher correspondence potential than <result in> *problems.* This has to do with the number of correspondence relations that these two items

establish and with the frequency with which they occur as translation correspondences. The former expression corresponds only to two German items and occurs with very high frequency. The latter expression, on the other hand, does not have very high frequency as a translation correspondence but it corresponds to half of the items from the TLD {PROBLEM BEREITEN}.

The above results can be compared with those obtained in the intralingual analysis. Here it is interesting to compare whether correspondence potential is related to the general frequency of combinations with the plural form of the noun <problem> and to the values of association strength of these lexical items.

Lexical items	**Total frequency in ukWaC**	**CP**
<cause> *problems*	9372	13.2
<there be> *problems*	6672	12.4
problems <arise>	3965	8.7
<create> *problems*	1939	11.8
<lead to> *problems*	1765	4
<present> *problems*	1477	7
<pose> *problems*	1321	7.2
<raise> *problems*	404	4.7
<result in> *problems*	230	4
<give rise to> *problems*	163	8

Table 5.17: Relationship between the distribution of lexical items from the TLD {CAUSE PROBLEM} in the reference corpus and their correspondence potential

Table 5.17 contains the information regarding the frequency of the English lexical items in question according to their distribution in the reference corpus ukWaC as well as their correspondence potential. The data are ordered according to the frequency variable. The findings related to the item <to be problematic> are not included because it is beyond the division singular vs. plural nouns and <problem> vs. <difficulty>

In general, more frequent lexical items tend to have higher correspondence potential. This is what one can conclude from calculating the correlation coefficient (r=0.8). However, this correlation is not perfect s there are some serious deviations. For example, <lead to> *problems* has far lower correspondence potential in our data than would be expected from its

frequency in the reference corpus. It is the fifth most frequent item but next to the last item according to its correspondence potential. Exactly the opposite is true for <give rise to> *problems*. There are also minor differences between the pairs <present> *problems* and <pose> *problems* and *problems* <arise> and <create> *problems*. In both cases, the former items are less frequent in the reference corpus but have stronger correspondence potential. Therefore, one cannot assume that the fact that a lexical item occurs with high frequency in the reference corpus will mean that this item will automatically have higher correspondence potential. In other words, the lexical items with higher substitution potential do not necessarily have higher correspondence potential.

The results following from the analysis of correspondence potential can be interpreted also in terms of the structure of a TLD. Thus, it can be said that the centrality of an item to a TLD depends on its use as a translation correspondence. This is displayed in figure 5.6 below in which the structure of the TLD {CAUSE PROBLEM} is represented through concentric circles. The position of lexical items in the figure reflects their correspondence potential. The further a lexical item is from the centre, the lower correspondence potential it has.

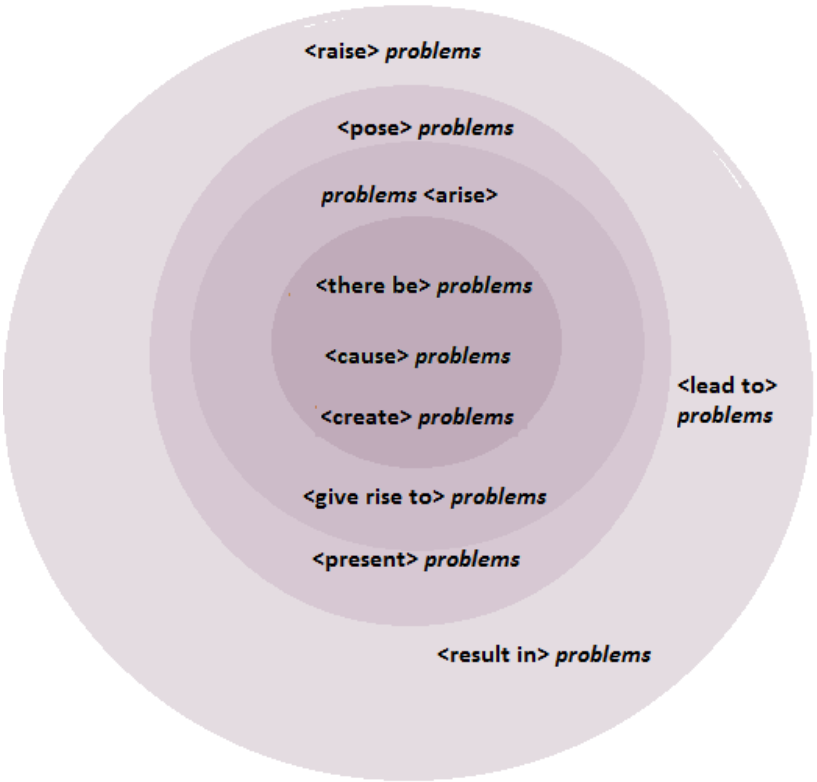


Figure 5.6: Structure of the TLD {CAUSE PROBLEM} according to correspondence potential values of lexical items

5.2.3.3 Conclusion

The above investigations were concerned with the application of the differentiation principle to the study of translation correspondences both from an intralingual and interlingual perspective by means of a distributional model.

The intralinguistic analysis indicates that lexical items that belong to the same TLD can be differentiated by focusing on local contexts in which they occur. Two types of distinguishing features were discovered here. A set of specific distinguishing features are related to the occurrence of items in particular collocations or colligations. A set of general features are related to broader tendencies. Both types of features showed how mutually substitutable lexical items are and this feature is called *substitution potential*.

The interlinguistic analysis showed that the lexical items differed significantly in terms of their use as translation correspondence. There are two factors which determine these differences. The first one is due to the different number of items from the German TLD to which the English items correspond. The second difference has to do with how often an English lexical item corresponds to German items. Both factors are equally important and in the above analysis they were merged into one variable called correspondence potential.

Although the substitution and correspondence potential tend to be related to a certain extent they are not completely correlated. For this reason, when talking about the structure of a TLD or the position of lexical units in a TLD it is necessary to distinguish between a view based on the intralingual and interlingual analysis. It can be concluded that the results obtained indicate that the differentiation principle can be successfully applied to the study of distribution of lexical items.

5.3 TLD {PROBLEM BEREITEN}

In this section the distribution of German lexical items from the TLD {PROBLEM BEREITEN} will be studied. The analysis goes through the same stages as the analysis of English lexical items.

Hence, first the analysis from an intralingual and afterwards from an interlingual perspective will be conducted.

The intralingual analysis will start by examining the grammar structures and typical contexts in which lexical items occur. After that, the distribution of linguistic units in these contexts will be compared. The comparative analysis will show whether the distinguishing features of German lexical units can be identified in the same way as was the case with the English items. Here, both individual differences and general tendencies will be examined. The final analysis will explore significant unique collocates that can also serve additional distinguishers. The results will give information about the substitution potential of the German lexical items examined. The data used in this section derive from the German reference corpus deWaC.

The interlingual study will explore how German linguistic units are used as translation correspondences of corresponding English items. Here, the focus will be on the following variables: correspondence degree, the number of correspondence relations and correspondence potential.

5.3.1 Grammar structures

5.3.1.1 A local grammar of the lexical items from the TLD {PROBLEM BEREITEN}

As we saw in Chapter 4 there are 13 lexical items that belong to the TLD {PROBLEM BEREITEN}. These items are displayed in Table 5.18.

Lexical items	
<es geben Problem|Schwierigkeit>	<zu Problem|Schwierigkeit führen>
<Problem|Schwierigkeit auftreten>	<Problem|Schwierigkeit schaffen>
<Problem|Schwierigkeit aufwerfen	<Problem|Schwierigkeit sich ergeben>
<Problem|Schwierigkeit bereiten>	<Problem|Schwierigkeit verursachen>
<Problem|Schwierigkeit bringen>	<problematisch sein>
<Problem|Schwierigkeit darstellen>	<Ursache GEN|für Problem|Schwierigkeit sein>
<Problem|Schwierigkeit entstehen>	

Table 5.18: Lexical items from the TLD {PROBLEM BEREITEN}

There are several differences between the grammatical structures of the lexical units from the present domain. There are 11 items that consist of a nominal and verbal element. There is also one noun phrase and one predicative adjective. As for the nominal element, its structure is similar to that of corresponding English cognates. The following two nouns are used interchangeably: <Problem> or <Schwierigkeit>. The verbal element occurs in seven expressions in the form of a transitive verb, three times as an intransitive verb and once as an existential expression. Transitive verbs are <aufwerfen>, <bereiten>, <bringen>, <führen zu>, <schaffen> and <verursachen>. The verb <bringen> occasionally occurs with <mit sich> but this is an optional element. Intransitive verbs are <auftreten>, <entstehen> and <sich ergeben>, and <es gibt> is an existential construction. Differences between these three types of constructions can be explained both in grammatical terms and in terms of textual metafunction. The nouns <Problem> and <Schwierigkeit> occur with transitive verbs in the direct object position, and with intransitive verbs they serve as the subject of a clause. In the existential clause the item <es> is a dummy subject and the two nouns have the same function as in transitive clauses. In terms of textual meaning, if <Problem> and <Schwierigkeit> occur with the transitive verbs they serve as Rheme and if they occur with intransitive verbs they are topicalised and perform the function of Theme. The verbs <bereiten> and <verursachen> differ from other transitive verbs because the two nouns with them may also occur as Theme such as in (23).

23. **Probleme bereitet** der Abschnitt Katzenelnbogen-Zollhaus.

This is a marked usage and it serves to topicalise <Problem> and <Schwierigkeit>. A similar effect can be achieved with the use of transitive verbs in passive. All verbs except <führen zu> can theoretically occur in this manner but such expressions are infrequent in my corpus. Finally, the transitive verbs <bereiten> and <bringen> stand out from other items because they colligate with substantives used as indirect object.

The noun phrase <Ursache DT|für Problem|Schwierigkeit> and the adjective <problematisch> collocate with the verb <sein>. The latter item may occasionally occur with

other linking verbs such as <scheinen> and <erscheinen>. These items denoting problems typically occur in Rheme but as (24) illustrates the adjectival expression may also be used in the topicalised position.

24. **Sehr problematisch ist** die Farbe bei Topasen, weil viele ihren Farbton durch Behandlung erhalten haben.

General grammar structures of current lexical items were investigated by means of the CQP tools. These structures in many respects resemble those that were identified for corresponding English lexical items. Regarding transitive, intransitive and existential expressions, verbal elements can be modified by a modal verb or an adverbial expression. Similarly, nominal elements can be modified by an adjective or a noun phrase. In the latter case, modifiers are typically merged with the noun <Problem> into a compound expression (25).

25. Sie sei der einzige Bereich, so Kowalik, der **Umweltprobleme verursachen** kann.

The verb <sein> that co-occurs with <problematisch> is also occasionally modified by a modal verb or adverb. The adjective itself may also be modified by an adverb. Similarly, the nouns <Problem|Schwierigkeit> that collocate with <Ursache GEN|für> are sometimes preceded by a modifier.

Word order in German is more flexible than in English. Thus, the verbal element that occurs in a transitive, intransitive or existential expression can either precede or follow nominal elements. The latter is typically encountered in relative clauses. In addition, when the verbs <bereiten> and <verursachen> occur in Theme they are found in the position after <Problem> and <Schwierigkeit>.

The study of local grammars will begin with the items that specify or modify the meaning of <Problem> and <Schwierigkeit>. The following types are found here: modifiers, determiners and the negative word <kein>. Modifiers either precede the two nouns in which

case they are usually adjectival expressions or together with the noun <Problem> create a compound noun. Regardless of this formal grammatical classification into adjectival expressions and compound nouns they can be grouped into several local grammar classes in relation to the functions they perform. These classes of items are summarised in Table 5.19. The table contains also typical members of these classes.

As we can see, these are the same four classes that were identified for English lexical items (see 5.2.1 above) and they denote the same senses. What was said before also applies here. The items are mainly adjectives but INTENSIFIERS and SORTALS also contain nominal expressions. The modifiers are mostly one-word long but QUANTIFIERS can sometimes be realised as multi-word items as well.

INTENSIFIERS	QUANTIFIERS	SORTALS	COMPARATORS
<arg>	<eine Reihe>	<beruflich>	<alt>
<echt>	<eine Vielzahl>	<finanziell>	<änlich>
<enorm>	<einig>	<gesellschaftlich>	<besonder>
<enorm>	<einzig>	<gesundheitlich>	<gewiß>
<erheblich>	<ein paar>	<Gesundheitsprobleme>	<gleich>
<ernst>	<kaum>	<intern>	<irgendwelch>
<ernsthaft>	<mehr>	<Kommunikationsproblem>	<neu>
<gering>	<wenig>	<Kreislaufproblem>	<neuerlich>
<gewaltig>	<zahlreich>	<logistisch>	<speziell>
<gravierend>		<mental>	<unterschiedlich>
<groß>		<organisatorisch>	<weiter>
<Hauptproblem>		<viel>	<zusätzlich
<Kernproblem>		<politisch>	
<klein>		<praktisch>	
<massiv>		<psychisch>	
		<rechtlich>	
		<sozial>	
		<Sprachproblem>	
		<strukturell>	
		<technisch>	
		<Verständnisproblem>	
		<weiter>	
		<wirtschaftlich>	

Table 5.19: Modifiers that frequently occur with the German lexical items from the TLD {PROBLEM BEREITEN}

The most frequent INTENSIFIERS express importance, such as <enorm>, <ernsthaft> or <Kernproblem>. The items that play down the importance of a problem are less numerous and less frequent. We find here only two items with relatively high frequency <gering> or <klein>. But these are also less frequent than those that have amplifying meaning. Although INTENSIFIERS usually directly occur with <Problem> or <Schwierigkeit> some of them occasionally combine with SORTALS. The adjective <problematisch> denotes a similar meaning like INTENSIFIERS hen it collocates with the adverbs that either downgrade (e.g. <wenig>, <ein bisschen> and <teilweise>) or strengthen (e.g. <besonders>, <sehr>, <äußerst>, <durchahus> and <ziemlich>) its meaning.

QUANTIFIERS refer to a large or small number of problems. The former type is more frequent. They occur only with the plural form of the two nouns. Most common are direct collocations with *Probleme*|*Problemen* or *Schwierigkeiten* but some QUANTIFIERS also co-occur with SORTALS and COMPARATORS.

SORTALS can be classified into three major semantic subgroups: the items that refer to health issues (e.g. <gesundheitlich>, <psychisch>, <Gesundheitsproblem>, <mental> Kreislaufproblem), the items that denote socio-political issues (e.g. <gesellschaftlich>, <innenpolitisch>, <finanziell>, <juristisch>) and those concerned with communication (e.g. <Kommunikationsproblem>, <Sprachproblem>, <Verständnisprobleme>). Finally, there are items which while do not belong to any of these groups and are too few in number to constitute a separate group (e.g. <Personalproblem>, <Abgrenzungsproblem> or <Kapazitätsproblem>). They will be set aside in the current study. In general, there do not seem to be significant differences with regard to the distribution of different semantic subgroups.

The most frequent COMPARATORS express that the problems in questions are different from those explicitly or implicitly mentioned in the previous discourse (e.g. <neu>, <zusätzlich>, <weiter>, or <besonder>). Such COMPARATORS sometimes precede a SORTAL which denotes a social issue such as <sozial>, <finanziel>, <wirtschaftlich>, <rechtlich> or <ethisch>. The adjective <problematisch> expresses a similar meaning like COMPARATORS in collocations created with <ähnlich>, <genauso>, <genauso… wie> and <so… wie>.

Apart from modifiers, the nouns <Problem> and <Schwierigkeit> can also be preceded by a determiner or the indefinite <kein>. The determiners that we encounter here are definite and indefinite articles used in plural and singular and the demonstrative <dies>. The item <kein> expresses that an action did not cause problems and as such it has a similar function as <nicht> when this item precedes verbs.

Finally, the noun <Problem> is sometimes followed by post-modifiers. These post-modifiers consist of a definite article used in the genitive case and a noun. <Problem> occurs typically in singular and together with the definite article <das> constitutes the head of a noun phrase. Semantically, the items that modify the head perform the same function like SORTALS which are part of compound nouns. Thus, (26) can be transformed into (27) and the meaning of the clause would not change.

26. Bei der Rekonstruktion **entsteht das Problem der Abgrenzung**…
27. Rekonstruktion **entsteht das Abgrenzungsproblem**…

The above description accounts for the occurrence of typical items that modify the meaning of <Problem> and <Schwierigkeit>. Below, the focus will be on items which modify the meaning of verbal elements. We will begin with modal verbs.

The following modals occur with all lexical items except with <Ursache GEN|für Problem|Schwierigkeit>: <können>, <dürfen>, <sollen>, <werden> and <müssen> These modal verbs serve to express speakers' judgement of probabilities. The first two verbs express low modality, the next two ones medium modality and the last verb expresses high modality. The first two verbs correspond roughly to <can>, <could>, <may> and <might> in English, the counterpart of <sollen> is <should> and of <must> and <have to> it is <müssen>. The modal verb <dürfen> co-occurs almost always with <nicht> forming thus a translation correspondent for <must not> and <should not>. The item <werden> has two correspondences in English. When used in the indicative form it corresponds to <will> and when used in the form of Subjunctive II (as <würden>) it corresponds to <would>. Like its English counterpart, the modal <sollen> occurs mostly in negated (28) and conditional clauses

(29). Negated clauses usually contain intransitive verbs or the verb <bereiten> which is less often. Modal verbs will be coded as PROBABILITY_OPERATORS.

28. Der Reisende sollte sich mit seiner Erkrankung gut auskennen, die Selbstmessung des Blut- oder Harnzuckers **sollte kein Problem darstellen.**
29. Die Betreuer stehen rund um die Uhr zur Verfügung und bieten so Sicherheit, **wenn** gesundheitliche **Probleme auftreten sollten.**

The collocations formed with adverbial expressions can be grouped into two classes which, following the labels used in the previous section, will be coded as USUALITY and DURATION. As far as the USUALITY items are concerned, we can distinguish between three types of meaning. The first type contains the items that express that something problems occur frequently such as <oft>, <oftmals> <immer wieder>, <in der Regel>, <meistens> and <normalerweise>. The second type which expresses the opposite meaning has only one member <nie>. Finally, <gelegentlich> and <manchmal> belong to the third type and denote that problems occur occasionally. Among the DURATION items we can draw a distinction between those that signify strong (e.g. <immer noch>) and those that signify weak continuity (e.g. <nicht mehr>).

Above, we saw that two transitive verbs occurred with the indirect object with the function to refer to entities that receive an action. The same function can also be performed by the prepositional phrase <für+NP>. There are some formal differences between the two. First, the prepositional phrase colligates with a larger number of lexical items. Second, the indirect object is mainly realised through the pronouns *uns, mir* and *ihnen* used in dative or the nouns referring to a group of people such as <Mensch>, <Frau>, <Kind>, <Familie> or <Gesellschaft>. The range of nouns that occur in the prepositional phrase is very wide as opposed to those that occur as indirect object. These nouns mostly denote a community or group of people as the following frequent expressions illustrate: *für die Betroffenen*, *für viele Menschen, für den Patienten, für die Schüler, für alle Beteiligten.* Finally, the prepositional phrase does not have a fixed position whereas the indirect object typically occurs between verbal and nominal elements. The prepositional phrase can follow the nouns <Problem> and

<Schwierigkeit> or the adjective <problematisch> but it can also follow or precede verbal elements. Items that occur in the indirect object position and in the prepositional phrase <für+NP> will be coded as RECIPIENTS.

In addition to the above prepositional phrase the nouns <Problem> and <Schwierigkeit> can also be followed by <bei NP>. This prepositional phrase serves as a circumstantial adjunct and provides information about the conditions under which problems arise. This is illustrated in (30) below where the problems related to ticket inspection in public transport arise due to overcrowding on trains. These lexical items will be coded as CIRCUMSTANTIALS.

30. Mit den gewaltig steigenden Zahlen im Personenverkehr **ergaben sich zusätzlich Probleme bei den Fahrkartenkontrollen in den vollbesetzten Zügen.**

Finally, German lexical items occurs also in the context of adversative conjunctions <jedoch>, <aber> and <allerdings> that denote opposition or contrast between the information in two clauses. These conjunctions typically occur in negated clauses with the indefinite item <kein> such as in (31) to stress that despite negative expectations problems do or did not arise.

31. Aufladung erfolgt über den Rechner, dies **stellt aber keine Probleme dar.**

Lexical items that occur in the subject position with transitive verbs are semantically as heterogeneous as was the case with the English lexical items. One common feature is that they denote activities, events or processes. This is why they will be coded as THING.

5.3.1.2 Conclusion

The above section provides a description of typical contexts in which the German lexical items from the present domain occur. These typical contexts are interpreted in terms of local

grammar classes and are displayed below in the form of grammatical structures. Table 5.20 displays only the general order of these local grammar classes in main clauses and neglects the flexible word order of German. It indicates the main local grammar rules that characterise the distribution of German lexical items from the present TLD.

THING [PROBABILITY_OPERATOR] [CONTINUITY|USUALITY] BEREITEN^{TR} [RECIPIENT] [INTENSIFIER|QUANTIFIER|SORTAL|COMPARATOR] PROBLEM [RECIPIENT][CIRCUMSTANTIAL]
BEREITEN^{EX} [PROBABILITY_OPERATOR] [CONTINUITY|USUALITY] [INTENSIFIER|QUANTIFIER|SORTAL|COMPARATOR] PROBLEM
INTENSIFIER|QUANTIFIER|SORTAL|COMPARATOR] PROBLEM [PROBABILITY_OPERATOR] BEREITEN^{NTR}
THING LINK_VB [AMPLIFIER] BEREITEN^{ADJ} [RECIPIENT]
THING LINK_VB BEREITEN^{NP}

Table 5.20: Local grammar structures for the lexical items from the TLD {PROBLEM BEREITEN}

In the following section the distribution of lexical items from these local grammar classes will be studied in more detail.

5.3.2 An intralinguistic analysis of the items from the TLD {PROBLEM BEREITEN}

5.3.2.1 General distributional differences

The frequency of use of the German lexical items from the TLD {PROBLEM BEREITEN} is displayed in the following table.

Lexical items	General frequency
<Problem|Schwierigkeit bereiten>	7820
<Problem|Schwierigkeit darstellen>	6049
<problematisch sein>	4600
<Problem|Schwierigkeit führen>	4524
<Problem|Schwierigkeit sich ergeben>	4521
<Problem|Schwierigkeit auftreten>	3573
<Ursache GEN|für Problem|Schwierigkeit>	2843
<Problem|Schwierigkeit entstehen>	2781
<Problem|Schwierigkeit bringen>	2188
<Problem|Schwierigkeit verursachen>	1595
<Problem|Schwierigkeit schaffen>	1552
<Problem|Schwierigkeit aufwerfen	1526

Table 5.21: Frequency of the lexical items from the TLD {PROBLEM BEREITEN} according to deWaC

According to the results of the chi-square test, it can be concluded that we deal with significant differences here. This is because the p-value is far below the predefined threshold of 0.05 and because the chi value is larger than the critical value of 3.940 of 10 degrees of freedom for the given probability.

Data: general frequency

χ^2 = 11636.27, df = 10, p-value < 2.2e-16.

This difference is also illustrated by the fact that the first two most frequent lexical items are four or more times as frequent as the three most infrequent items. There are no significant relations between the frequency differences and the internal structure of lexical items. The expressions with transitive verbs form, for example, both the most and least frequent items.

Similarly, <problematisch sein>, <Problem|Schwierigkeit führen> and <Problem|Schwierigkeit sich ergeben> have different internal structures but are close in raw frequency.

The expressions created with <Problem> are more frequent than those formed with <Schwierigkeit> (Table 5.22). The expressions formed with <Problem> make up between 80% and 95% of total occurrence. Two exceptions are expressions in which <bringen> and <bereiten> occur which have lower figures. Especially rare are collocations that consist of <darstellen>, <aufwerfen> and <Ursache> and the noun <Schwierigkeit>.

Lexical items	**Co-occurrences with <Problem>**	**Co-occurrences with <Schwierigkeit>**
<darstellen>	96%	4%
<aufwerfen>	94%	6%
<Ursache GEN|für>	96%	4%
<entstehen>	84%	16%
<schaffen>	83%	17%
<verursachen>	83%	17%
<führen>	82%	18%
<auftreten>	80%	20%
<ergeben>	78%	22%
<bringen>	62%	38%
<bereiten>	59%	41%

Table 5.22: Co-occurrence of verbal elements with <Problem> and <Schwierigkeit>

In all but one case collocates occur more frequently with the plural form of the two nouns than with the singular form (Table 5.23). This exception is the verb <darstellen> which forms stronger collocations with the singular form. The combinations that consist of <bereiten> or <verursachen>, on the one hand, and the plural form of the two nouns, on the other, occur with extremely high frequencies. The frequency of other collocations depends on specific word forms but there is a tendency that the collocations formed with the plural form of <Schwierigkeit> are more typical than those formed with the plural form of <Problem>. Thus, <bringen> occurs 96% of the time with *Schwierigkeiten* but 74% of the time with *Probleme*. Similarly, <schaffen> occurs in 92% of the cases with *Schwierigkeiten* and 63% of the time with Probleme. It is also interesting that identical items occur with the singular form of the two nouns above the average level (30% with the expressions with *Problem* and 13% with

Schwierigkeit). The verbs occurring with both nouns above the average level are <sich ergeben> and <entstehen>. In addition, <schaffen> occurs in this way with *Problem*.

Lexical items	**Co-occurrences with *Problem***	**Co-occurrences with *Probleme*|*Problemen***	**Co-occurrences with *Schwierigkeit***	**Co-occurrences with *Schwierigkeiten***
<auftreten>	28%	72%	6%	94%
<aufwerfen>	28%	72%	9%	91%
<bereiten>	5%	95%	2%	98%
<bringen>	26%	74%	4%	96%
<darstellen>	87%	13%	51%	49%
<entstehen>	36%	64%	16%	84%
<ergeben>	40%	60%	22%	78%
<führen>	21%	79%	6%	94%
<schaffen>	37%	63%	8%	92%
<Ursache GEN|für>	13%	87%	9%	91%
<verursachen>	12%	88%	1%	99%

Table 5.23: Co-occurrence of verbal elements with the plural and singular form of the nouns <Problem> and <Schwierigkeit>

The analysis shows that the lexical items formed with the two nouns are not equally probable.

5.3.2.2 Co-occurrence with modifiers of verbal elements

In this section the occurrence with modal verbs and adverbial expressions will be examined. Figure 5.7 displays the results of the distribution of lexical items with modal verbs. As can be seen, the expressions without modal verbs are more common. There is no strong relation between occurrences with modal verbs and the type of transitivity. The exceptions are the existential <es Problem|Schwierigkeit geben>, the adjective phrase <problematisch sein> and the noun phrase <Ursache GEN|für Problem|Schwierigkeit> which are very infrequent in this context. Three lexical items that are most common here are <Problem|Schwierigkeit schaffen>, <Problem|Schwierigkeit auftreten> and <Problem|Schwierigkeit bringen>. The following four items also occur above the median level: <Problem|Schwierigkeit führen>,

<Problem|Schwierigkeit verursachen>, <Problem|Schwierigkeit darstellen> and <Problem|Schwierigkeit entstehen>. Below this level remain <Problem|Schwierigkeit bereiten>, <Problem|Schwierigkeit sich ergeben> and <Problem|Schwierigkeit aufwerfen>.

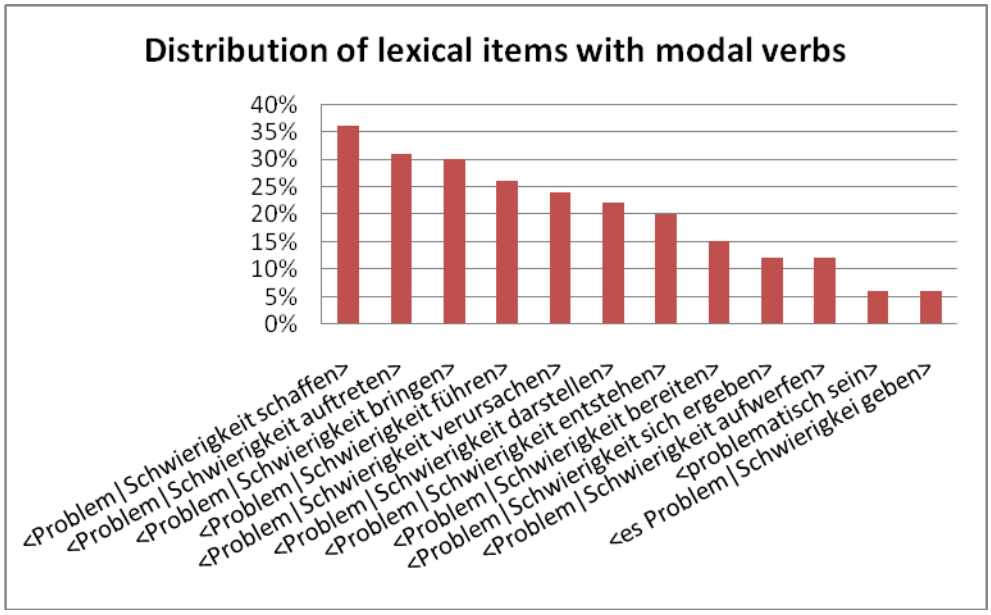


Figure 5.7: Co-occurrence of lexical items from the TLD {PROBLEM BEREITEN} with modal verbs

The most frequent modal verb with all lexical units is <können> when it is used in indicative mood. With approximately equal probability follow <könnten> and <würden>. Differences regarding these and other modal verbs generally follow the tendency displayed in the above figure. In other words, the verbs <schaffen>, <auftreten> and <bringen> occur most frequently with these three modal verbs. <zu Problem|Schwierigkeit führen> and <Problem|Schwierigkeit bereiten> are slightly more typical with the indicative form of <werden>. The verb <darstellen> is more common with <sollen>, <dürfen> and <kein> than other verbs. These expressions have similar meaning. Finally, <sollen> forms conditional expressions most frequently with the non-transitive verbs <entstehen> and <auftreten>. The verb <müssen> is equally infrequent with all lexical items.

The most frequent USUALITY item is <immer> and it most typically occurs with <problematisch sein>. This adjectival construction also most typically selects for <manchmal>. The collocation <immer wieder> prefers the occurrence with the intransitive <Problem|Schwierigkeiten auftreten> and <Problem|Schwierigkeiten sich ergeben> whereas

<oft> has the strongest association with <Problem|Schwierigkeit führen> and <Problem|Schwierigkeit bringen>. Other USUALITY items are less frequent and no significant differences are observed in their behaviour. The DURATION items are in general infrequent but when they do occur they are most typical with <problematisch sein>.

As we have seen above, in the present corpus the indirect object occurs only with <bereiten> and <bringen>. A comparison of distribution with the two verbs indicates that it is slightly more frequent and more typical with the former verb. The prepositional phrase <für+NP> is most typical with <problematisch sein> or with the collocation *Problem* <darstellen>. Significant occurrences are also found with the word form *Probleme* and the verbs <sich ergeben>, <bringen> and <schaffen>.

The prepositional phrase <bei NP> is most frequently used with the intransitive verbs <auftreten>, <sich ergeben> and <entstehen> and the transitive verb <bereiten>.

5.3.2.3 Co-occurrence with the modifiers of nominal elements

This section is concerned with the comparison of the distribution of the nouns <Problem> and <Schwierigkeit> with noun modifiers. At the beginning the occurrences with and without modifiers will be examined in a general manner.

Figure 5.8 summarises the distribution of the verbal elements that do not select for modifiers when they co-occur with the singular or plural form of the noun <Problem>. In general, modifiers are slightly more common with the plural than with the singular. The median value for the occurrence with the former is 36% and for the latter 31%. In addition, the percentage of the co-occurrence with modifiers and the plural form is above the average value for six verbs as opposed to four verbs with the singular form. The correlation coefficient for the relationship between the occurrence with the two word forms is r=0.41 which does not indicate any strong pattern here. Modifiers most typically occur with <aufwerfen> and <auftreten>. These are the only two verbs for which the value of association strength is above average when they collocate with both the singular and plural form of the noun <Problem>. The former verb is also the only one that in more than half of the cases occurs with modifiers.

The following four items occur also above the average value here: <bringen>, <sich ergeben>, <schaffen> and <verursachen>. The item with the weakest association in the present context is <es geben>.

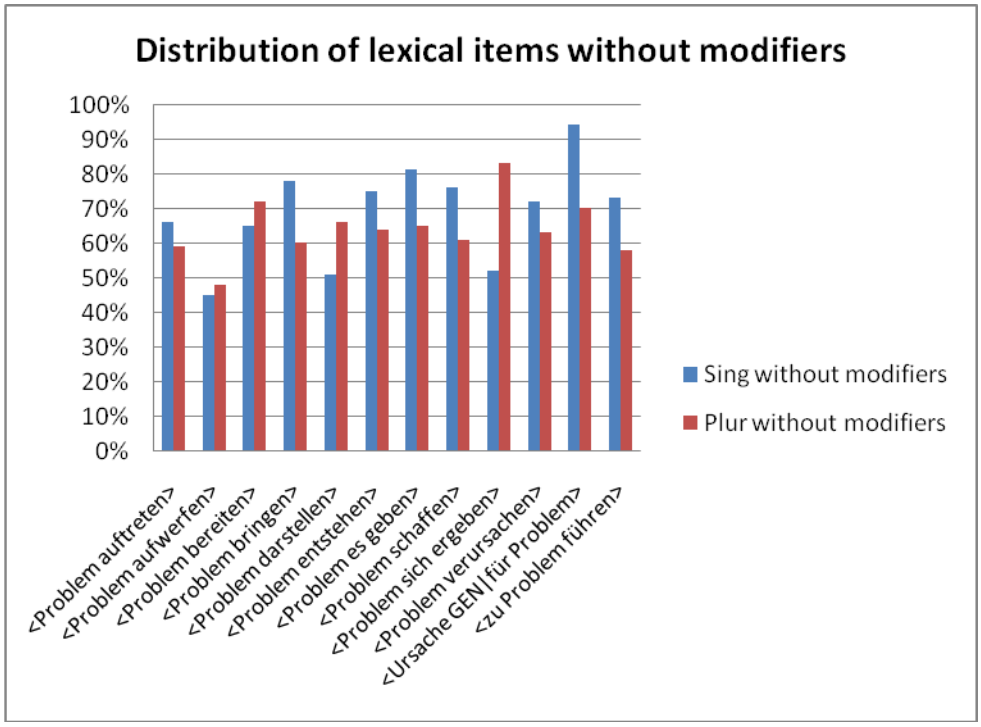


Figure 5.8: Distribution of lexical items that occur with the noun <Problem> without modifiers

The occurrence with modifiers and the two word forms of the noun <Schwierigkeit> are slightly more common than with the previously examined noun (Figure 5.9). Interestingly, the median value is the same for both word forms (44%). The correlation test for the occurrence with *Schwierigkeit* and *Schwierigkeiten* does not indicate strong relation (r=0.47). Only <darstellen> occurs above the average level with both word forms. The verbs <aufwerfen>, <sich ergeben> and <führen zu> forms common combinations with the plural form and occur with modifiers here more than 50% of the time. <es geben> and <bringen> are extremely rare. The following three items collocate with *Schwierigkeit* and have a frequency above the median: <bereiten>, <es geben> and <schaffen>. The verbs <bringen> and <entstehen> are atypical with modifiers here whereas, as have seen above, <aufwerfen>, <auftreten>, <führen zu> and <verursachen> do not occur with *Schwierigkeit* at all. It must be emphasised that the

combinations of *Schwierigkeit* and <es geben> and <schaffen> are very infrequent and the respective data should be interpreted with caution.

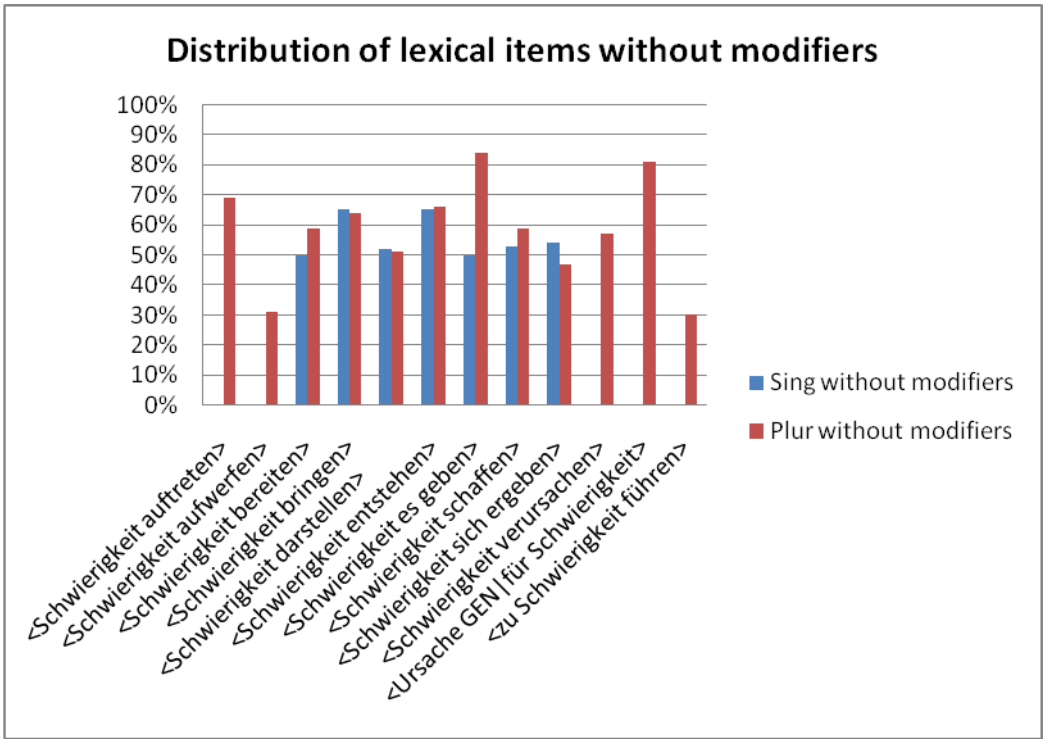


Figure 5.9: Distribution of lexical items that occur with the noun <Schwierigkeit> without modifiers

There are only minor overlaps between the data examined above. The first overlap is concerned with a weak association of the verb <es geben> and modifiers when it occurs with the two word forms of <Problem> and with the plural form of <Schwierigkeit>. The second one is due to the strong association of <aufwerfen> with <Problem> and *Schwierigkeiten* on the one hand and of <sich ergeben> with *Probleme* and *Schwierigkeiten* on the other. Similar to what has been observed in the English data the distribution of modifiers in the current context seems to be the subject to specific combinations between verbal elements and the singular and plural form of the nominal elements.

Let us now look at the occurrences with determiners and the indefinite <kein>. Two types of determiners can be found with the word form *Problem*: the definite determiners realised as the article *das* or the demonstrative pronoun *dieses* and the indefinite determiner realised as the article *ein*. With <Ursache GEN|für> the definite determiners occur in the

genitive case. Expressions with the definites are in general more common. They are proportionately most typical with <Ursache GEN|für> and the verbs <auftreten>, <aufwerfen> and <verursachen>. The indefinite article is most common with <bereiten> and <es geben>. The verb <bringen> occurs very infrequently with the indefinite article, whereas <es geben> and <darstellen> do not tend to select for the definite article at all. It has been seen above that the noun <Problem> can be followed by postmodifiers when it occurs with the definite article. With the word form *Problem* we most typically find the non-transitive verbs <sich ergeben>, <auftreten> and <entstehen>. The most common transitive verb in this context is <bringen>. There are no strong tendencies with regard to the type of SORTALS which occurs in the postmodifier position of a noun phrase. The most typical and the most frequent verb that occurs with *kein Problem* is <darstellen> which is used 43% of the time in this way. We find also very high values of association strength with <es geben> and <führen zu>. Other verbs are infrequent here. For example, <bringen> collocates with *Problem* 448 times but it occurs only 15 times with <kein> which makes about 3% of all occurrences of this verb. Similar or lower figures are observed for other verbs.

The definites that occur with the plural of <Problem> are of the same type as above. The meaning of indefiniteness, on the other hand, is realised through the zero article. In contrast to the occurrence with the word form *Problem* indefinite constructions are here more common. The lexical item <Ursache GEN|für> in both cases forms the strongest collocationional associations. The following items occur also typically occur with definite determiners: <es geben>, <verursachen> and <schaffen>. With the zero article we also commonly find <führen zu>, <sich ergeben>, <bereiten> and <verursachen>. The combinations formed with postmodifiers are less common with in the current context. Those that do occur are formed most typically with <auftreten> and <bringen>. As far as the occurrence of <kein> is concerned, it typically selects for the verbs <bereiten> and <es geben>.

The lexical item <Ursache GEN|für> remains the most typical collocate of the definite determiners with the word form *Schwierigkeit* as well. The indefinites are less frequent with

Schwierigkeit and apart from <darstellen> other items occur here marginally. <darstellen> and <bereiten> are the most typical collocates of <kein> in this context.

Definite determiners are less common with the plural form of <Schwierigkeit> than indefinites. Only <Ursache GEN|für>, <verursachen> and <darstellen> form significant collocations here. The indefinite article is typically found with <es geben>, <führen zu>, <Ursache GEN|für> and <entstehen>. Finally, <kein> does not occur frequently with *Schwierigkeiten* in the present context. The only exception is <bereiten>.

The above observation reveals the following individual differences between the lexical items from the TLD {PROBLEM BEREITEN}. The colligation of <Ursache GEN|für> and the definite determiners is the preferred option in the present corpus. Similarly, the zero article and the plural form of the nouns <Problem> and <Schwierigkeit> is most typical of <führen zu> and <es geben>, whereas <verursachen> has a strong tendency to occur with the definite article when it colligates with the plural form of the two nouns. Finally, <bereiten> appears to associate more strongly with <kein> than other lexical items.

Now, we will examine the patterns that characterise the occurrence of lexical items with modifiers in the context of the singular and plural form of the nouns <Problem> and <Schwierigkeit>. First the modifiers that occur with the plural of <Problem> will be considered as they constitute the largest set. The raw frequency of the items with modifiers and the number of modifiers are displayed in Table 5.24. The numbers in the parenthesis indicate the typical span within which modifiers occur.

Lexical items	Frequency with modifiers	Number of modifiers
{1,3} *Probleme* <bereiten>	1154	126
{1,3} *Probleme* <auftreten>	579	85
{1,3} *Probleme* <sich ergeben>	552	66
{1,3} *Probleme* <schaffen>	510	59
{1,3} *Probleme* <bringen>	474	74
{1,3} *Probleme* <entstehen>	473	74
zu {1,3} *Problemen* <führen>	441	45
{1,3} *Probleme* <aufwerfen>	430	57
{1,3} *Probleme* <es geben>	327	38
{1,3} *Probleme* <verursachen>	290	37
<Ursache GEN|für {0,3}> *Probleme*	248	35
{1,3} *Probleme* <darstellen>	92	24

Table 5.24: Frequency and the number of modifiers for lexical items that collocate with the word form *Probleme*

The frequency of lexical items strongly correlates with their number of modifiers (r=0.94). This means that the more frequent a lexical item is, the larger the number of modifiers it has. There are of course some exceptions. For example, {1,3} *Probleme* <bereiten> is almost twice as frequent as {1,3} *Probleme* <auftreten> but the number of modifiers that collocate with the former is one and a half times higher with it than with the latter item. Similarly, zu {1,3} <*Problem*en führen> is more frequent than {1,3} *Probleme* <es geben> and {1,3} *Probleme* <verursachen> but occurs with a smaller number of modifiers. As we can see below, the chi-square test indicates that there are significant differences between the values of the two observed variables. In both cases the p-value is far below the pre-defined threshold and the chi value is far above the critical value of 3.940 for the degrees of freedom 10.

data: **frequency with modifiers**
χ^2= 1460.132, df = 10, p-value < 2.2e-16

data: **number of modifiers**
χ^2 = 152.7052, df = 10, p-value < 2.2e-16

An analysis was carried out to examine tendencies regarding the distribution of shared modifiers in the present translation domain. Figure 5.10 displays the graphs for the two most

frequent and the two least frequent lexical items. The graphs that contain the data for other lexical items are included in Figure B3 in Appendix B. The blue coloured bar compares the frequency of the lexical items that occur with shared modifiers whereas the red bar displays the degree of overlap between these items. All figures are represented in percentage terms.

In the English data it was observed that degree of overlap decreased as the difference in frequency of lexical items dropped. The same tendency can be observed in the current data. For example, the values of degree of overlap range from 68% (with <bereiten>) over around 60% (with <auftreten> and <sich ergeben>) and 50% (with <schaffen> and <bringen>) to 43% (with <entstehen>). The most frequent lexical item {1,3} *Probleme* <bereiten> selects for between 58% and 75% of all modifiers that collocate with other lexical items, whereas the less frequent item {1,3} *Probleme* <schaffen> shares between 45% and 65% of modifiers with other even less frequent items. By the same token, the penultimate item in the list <Ursache GEN|für> {0,3} *Probleme* occurs with only 18% of modifiers that collocate with the least frequent {1,3} *Probleme* <darstellen>.

The tendency that degree of overlap increases as the frequency of items decreases was observed in the data containing English items and it characterises German lexical items as well. However, this tendency is less consistent in German. The correlation between frequency and degree of overlap was moderate to strong with English items in seven out of eight cases. With German lexical items such correlation values are found in six out of nine cases. The correlation degree is especially low for the expressions formed with the verbs <auftreten>, <bringen> and <führen zu>. In addition, two items that in this respect significantly deviate from the main pattern are <Ursache GEN|für> {0,3} *Probleme* and {1,3} *Probleme* <darstellen>. Both items show lower correlation than would be expected from their frequency.

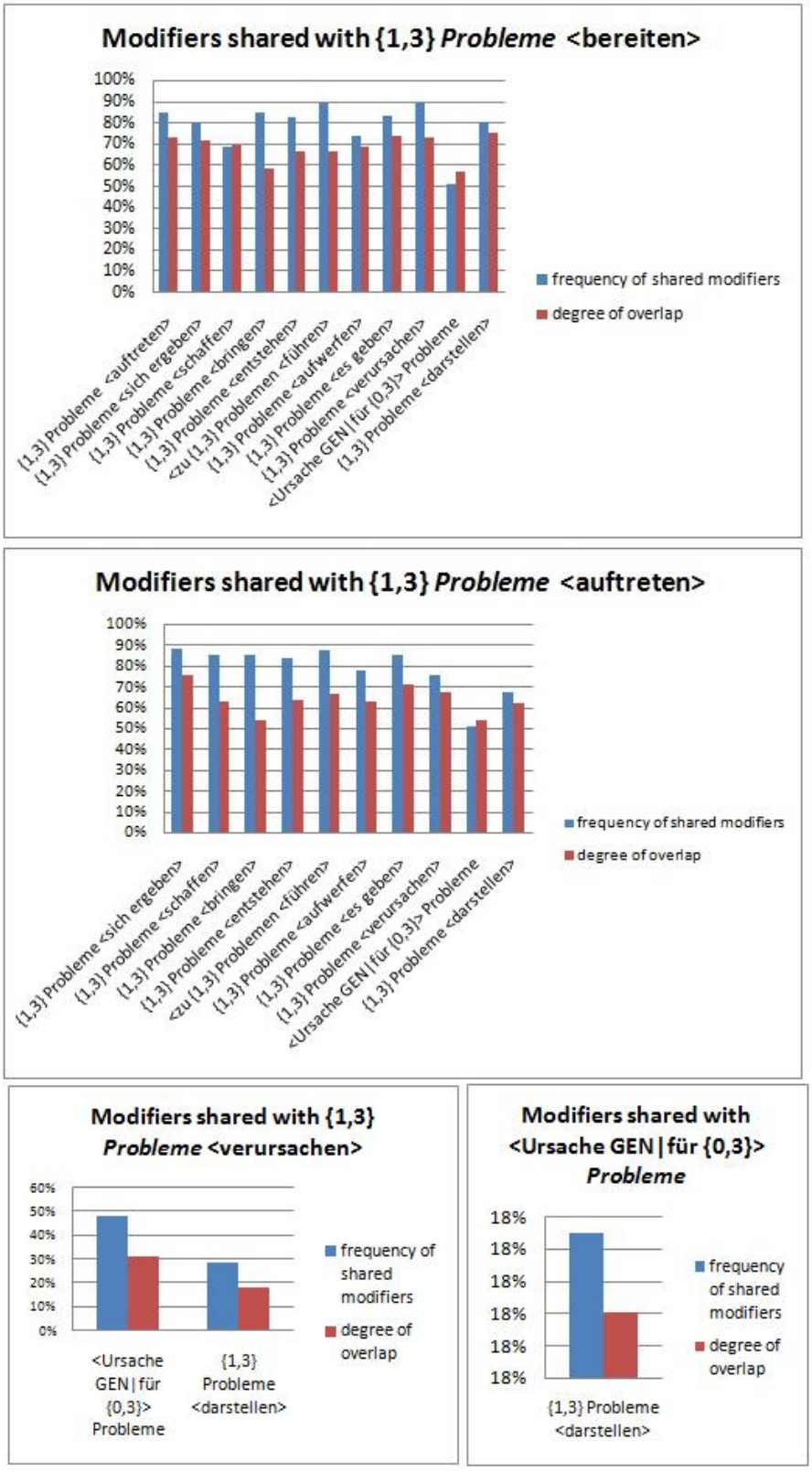


Figure 5.10: Frequency and degree of overlap of modifiers that occur with *Problem*|*Problemen* in the TLD {PROBLEM BEREITEN}

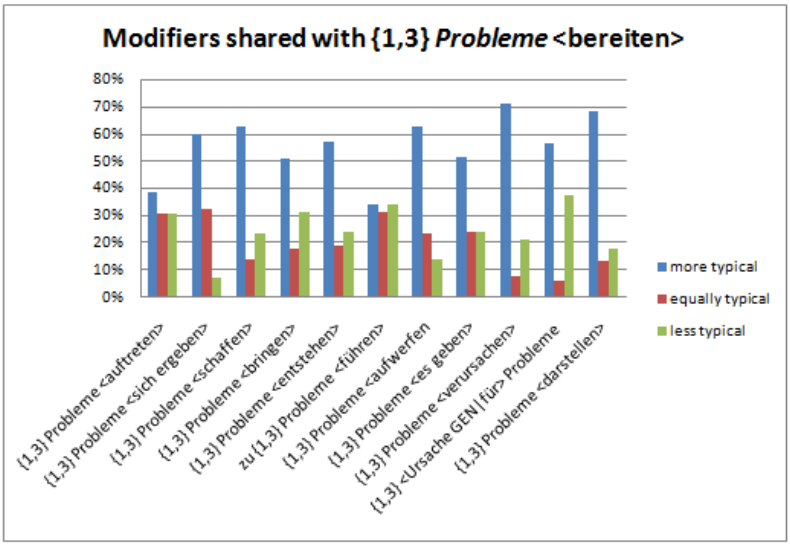
The vast majority of shared modifiers occur with very high frequency. The blue bars in the above graphs reflect this pattern. The total frequency of shared modifiers is for all but two items (<Ursache GEN|für> {0,3} *Probleme* and {1,3} *Probleme* <darstellen>) always above 50% of the total frequency of the combinations formed with modifiers. It means that the non-shared modifiers are in more instances infrequent lexical items. For example, of 59 modifiers that occur with {1,3} *Probleme* <schaffen> 18 do not collocate with {1,3} *Probleme* <bereiten> and of these only six occur more than twice. Similarly, of 15 modifiers that occur with <zu> {1,3} *Problemen* <führen> but not with {1,3} *Probleme* <auftreten> only six occur more than two times. In general, about 20% of all modifiers are shared with all lexical items.

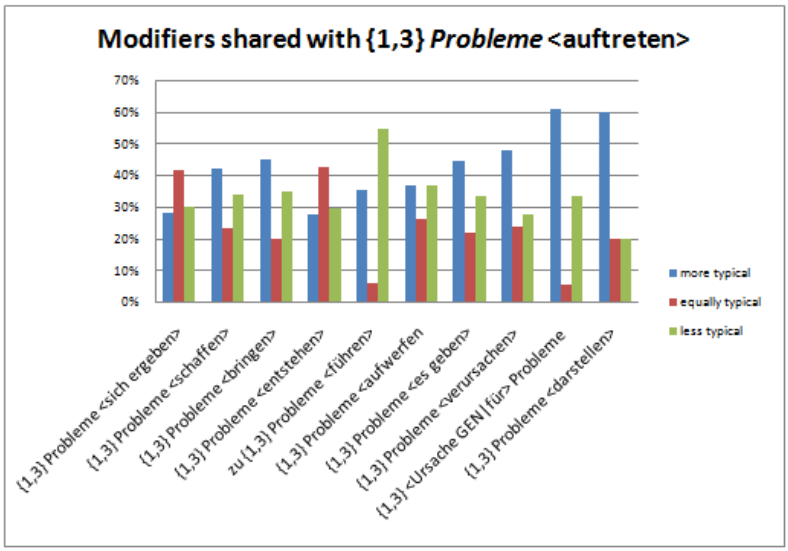
As said above, modifiers are realised either as adjectives or compound nouns and they can be classified into the local grammar classes. Adjectives are more frequent across all classes. The most frequent is the class QUANTIFIERS such as <viel>, <zahlreich>, <eine [ganze] Reihe> and <einige>. More numerous but slightly less frequent are the COMPARATORS <neu>, <zusätzlich>, <ander>, <weiter>, <folgend> and <besonder> and the INTENSIFIERS <enrsthaft>, <erheblich>, <groß> and *Hauptprobleme*. The most frequent SORTALS are the adjectives <technisch>, <gesundheitlich>, <psychisch>, <sozial>, <rechtlich> and the compounds *Sicherheitsprobleme*, *Umweltprobleme*, *Gesundheitsprobleme* and *Verkehrsprobleme.* The most frequent multi-word modifiers in the current context are the expressions that consist of an INTENSIFIER and a SORTAL such as <ernsthaft|erheblich gesundheitlich>, <groß wirtschaftlich>, <erheblich technisch> and <erheblich rechtlich>.

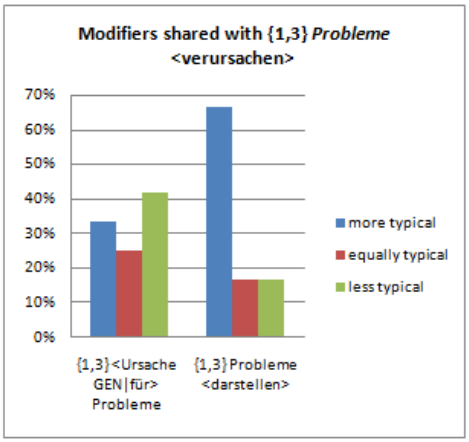
Now, we will examine the distribution of shared modifiers. In English data we observed the tendency that more frequent lexical items occurred in a larger number of strong collocations. Figure 5.11 displays results for the items with the highest and lowest frequency and the graphs that contain results for other items are provided in Figure B4 in Appendix B. The graphs compare the values of association strength between the item with the highest frequency and other less frequent items. In this way all lexical items from the current lexical domain are compared. Blue bars indicate the cases when collocations are more typical with the most frequent lexical item which is referred to in the title of graphs. Green bars indicate

stronger collocations formed with the less frequent items. Finally, the colour red points out cases when collocations are equally typical with the two categories of compared items.

The tendencies that characterised the distribution of English lexical items can be observed here as well. Thus, more frequent lexical items tend to occur more frequently with shared modifiers. It follows that the more frequent lexical items can substitute for the less frequent ones. But there are also some exceptions and they seem to be more numerous than in German data. For example, there are more equally typical collocations in German than in English. This is especially true for the modifiers shared between {1,3} *Probleme* <entstehen> on the one hand and {1,3} *Probleme* <aufwerfen>, {1,3} *Probleme* <es geben> and {1,3} *Probleme* <verursachen> on the other. In addition, the behaviour of the following three items contradict the general pattern; zu {1,3} *Problem*en <führen> and {1,3} <Ursache GEN|für> *Probleme* create in several cases stronger collocations than more frequent lexical items. On the other hand, {1,3} *Probleme* <schaffen> tends to occur in weaker collocations than less frequent items. Finally, {1,3} *Probleme* <aufwerfen> shows similar behaviour as <zu> {1,3} *Problem*en <führen> or forms more typical collocations than would be expected from its frequency.







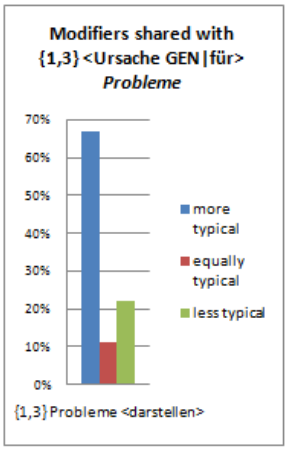


Figure 5.11: Association strength values for the expressions made up of verbal elements, shared modifiers and *Probleme*

The above investigation of the occurrence of modifiers with the plural form of the noun <Problem> points to the distinguishing features which are of the same kind as those existing in English data. These features are:

- More frequent lexical items tend to occur with a larger number of modifiers.
- The number of shared modifiers tends to increase as the frequency of lexical items drops.
- Lexical items from the German TLD tend to share frequent modifiers.
- The expressions formed with shared collocates tend to have a higher association strength if they contain more frequent lexical items.

The exceptions to these patterns are of interest as well because they constitute individual distinguishing features for the specific lexical items. One such case is {1,3} *Probleme* <darstellen> which with some other items ({1,3} *Probleme* <shaffen>, {1,3} *Probleme* <bringen>, zu {1,3} *Problem*en <führen> and ({1,3} *Probleme* es geben>) share fewer modifiers than we would expect from its frequency. Similarly, <zu {1,3} *Problem*en führen>, {1,3} <Ursache GEN|für> *Probleme* and {1,3} *Probleme* aufwerfen> create stronger and {1,3} *Probleme* schaffen> weaker collocations than would be expected.

Now the occurrence of lexical items and modifiers with the word form *Problem* will be considered. Frequency and the number of modifiers are displayed in Table 5.25 below.

Lexical items	Frequency with modifiers	Number of modifiers
{0,3} *Problem* <darstellen>	1384	151
{0,3} *Problem* <sich ergeben>	380	43
{0,3} *Problem* <entstehen>	201	31
zu {0,3} *Problem* <führen>	160	30
{0,3} *Problem* <auftreten>	82	11
{0,3} *Problem* <es geben>	79	20
{0,3} *Problem* <aufwerfen>	70	15
{0,3} *Problem* <schaffen>	59	15
{0,3} *Problem* <bringen>	53	11
{0,3} *Problem* <bereiten>	43	7
{0,3} *Problem* <verursachen>	11	4
{1,3} <Ursache GEN|für> *Problem*	10	4

Table 5.25: The number of modifiers and frequency of lexical items that collocate with the word form *Problem*

The collocations formed with the singular form of the noun <Problem> are in general less frequent and the number of modifiers is also lower. Nevertheless, the number of modifiers strongly correlates with the frequency of lexical items (r=0.99). As can be seen below, differences between the values in relation to both variables are statistically significant. P-value below the predetermined significance level and the chi-square value is far beyond the threshold of 3.940 for the given degrees of freedom (10).

data: frequency with modifiers

χ^2 = 6610.34, df = 10, p-value < 2.2e-16

data: number of modifiers

χ^2 = 489.4531, df = 10, p-value < 2.2e-16

Frequent modifiers are again usually shared by more than one lexical item. The highest degree of overlap is observed with the most frequent {0,3} *Problem* <darstellen>. This combination covers between 65% and 100% of all modifiers that occur with other items. The value of degree of overlap is lower for less frequent items. The overlap, for example, between {0,3} *Problem* <auftreten> and other less frequent items ranges between 16% and 33%. The overlap also tends to increase for the items with higher differences in frequency. Almost all

modifiers that occur with the three most infrequent items also collocate with other linguistic items. One exception is {0,3} *Problem* <aufwerfen>. More frequent lexical items share fewer modifiers with this item than we would expect from its frequency. As was the case previous studies association strength tends to be higher with more frequent lexical items. There are no exceptions to this tendency.

The most frequent modifier with almost all items is the INTENSIFIER <groß>. Among the most frequent collocates are also <ernst>, <erheblich>, <ernsthaft>, <grundsätzlich>, <gravierend> and the COMPARATORS denoting difference <weiter>, <zusätzlich>, <besonder>, <folgend>, <ander>, <neue> or those denoting similarities <gleich> and <ähnlich>. The most frequent SORTALS are <sozial>, <ethisch>, <technisch>, <politisch> and <gesellschaftlich>.

Now the distribution of modifiers with the lexical items from the TLD {PROBLEM BEREITEN} will be examined. I will start the word form *Schwierigkeiten.* Frequency and the number of modifiers are displayed in Table 5.26. The correlation between the two variables is in this case also very strong (r=-0.97) and it follows the previously observed trend that the number of modifiers decreases with frequency.

Lexical items	**Frequency with modifiers**	**Number of modifiers**
{0,3} *Schwierigkeiten* <bereiten>	812	95
{0,3} *Schwierigkeiten* <bringen>	379	60
zu {0,3} *Schwierigkeiten* <führen>	104	27
{0,3} *Schwierigkeiten* <auftreten>	84	26
{0,3} *Schwierigkeiten* <sich ergeben>	62	18
{0,3} *Schwierigkeiten* <es geben>	48	16
{0,3} *Schwierigkeiten* <verursachen>	42	13
{0,3} *Schwierigkeiten* <schaffen>	37	9
{0,3} *Schwierigkeiten* <entstehen>	34	10
{1,3} <Ursache GEN|für> *Schwierigkeiten*	24	6
{0,3} *Schwierigkeiten* <aufwerfen>	15	5
{0,3} *Schwierigkeiten* <darstellen>	4	2

Table 5.26: The number of modifiers and frequency of lexical items that collocate with the word form *Schwierigkeiten*

Differences between both variables are significant as the p-value and the chi values indicate below. Differences are less significant for the items positioned in the middle of the list.

```
data: frequency with modifiers
χ² = 5444.586, df = 10, p-value < 2.2e-16
data: number of modifiers
χ² = 628.1424, df = 10, p-value < 2.2e-16
```

The relationship between frequency and degree of overlap was observed here as well. The results show that the most frequent expression {0,3} *Schwierigkeiten* <bereiten> share between 57% and 100% of modifiers with less frequent lexical items. The less frequent <zu> {0,3} *Schwierigkeiten* <führen>, on the other hand, collocates with between 32% and 100% of modifiers found with other items. The modifiers that occur with the four least frequent items are almost all shared by the three most frequent lexical items.

The number of strong collocations also tends to correlate with the frequency of lexical items. Therefore, if two lexical items collocate with the same modifier it is the more frequent one that will probably form stronger collocations. The exceptions are the collocations formed with {0,3} *Schwierigkeiten* <bereiten> and {0,3} *Schwierigkeiten* <bringen>. In this case, the majority of modifiers (15) are equally typical. In addition, the two items are used in almost identical number of collocates (the former with nine and the latter with ten) despite the fact that they occur with different frequency.

Shared modifiers are those with high frequency. The frequency of shared modifiers is never lower than 50% of the total frequency of the occurrence with modifiers. The most frequent modifiers are almost identical as those encountered in the above studies.

The modifiers that occur with the singular form of the nouns <Schwierigkeit> are the least frequent of all four word forms. As we can see in Table 5.27 the verbs <verursachen>, <es geben>, <führen zu> and {1,3} <Ursache GEN|für> do not collocate here with modifiers at all. Even the most frequent {0,3} *Schwierigkeit* <sich ergeben> occurs only 55 times and collocates with 14 modifiers. Such low values were observed in the above data with the least frequent lexical items.

Lexical items	Frequency with modifiers	Number of modifiers
{0,3} Schwierigkeit <sich ergeben>	47	7
{0,3} Schwierigkeit <auftreten>	41	11
{0,3} Schwierigkeit <entstehen>	18	6
{0,3} Schwierigkeit <bereiten>	15	3
{0,3} Schwierigkeit <darstellen>	15	3
{0,3} Schwierigkeit <bringen>	6	3
{0,3} Schwierigkeit <schaffen>	4	2

Table 5.27: The number of modifiers and frequency of lexical items that collocate with the word form *Schwierigkeit*

Differences between the frequency of lexical items in the present context are statistically significant although less so than in previous cases. On the other hand, the p-value for the variable the number of modifiers is not below the pre-defined threshold of 0.05 and the chi-square is below the significance level of 12.59 for the degree of freedom 6 which means that we cannot reject the hypothesis that the differences are due to random variation.

data: **frequency with modifiers**
χ^2 = 80.1096, df = 6, p-value = 1.427e-13
data: **number of modifiers**
χ^2 = 12.4, df = 6, p-value = 0.05362

In spite of this, there is a correlation between the two variables (r=0.79) suggesting that the more frequent items in the present data also tend to occur with a larger number of modifiers. Because of small number of modifiers no clear tendency regarding degree of overlap, association strength or the type of modifiers can be identified. There are only individual differences. Thus, the adjective <groß> is most typical with the verb <bereiten>, <gewiß> is strongly associated with <bringen> whereas <ander>, <folgend> and <weiter> form strong collocations with <sich ergeben>. The verb <auftreten> is the only verb which collocates with in the current context the modifiers <technisch>, <gesundheitlich>, <wirtschaftlich>, <neu>, and <gravierend>.

5.3.2.4 Unique collocates

In this section the unique expressions formed with the lexical items from the TLD {PROBLEM BEREITEN} will be examined.

The lexical item <problematisch sein> is associated with two recurrent patterns that we do not find with other items. First is the expression *deshalb* <problematisch sein> *weil* such as in (31) below. The expression serves to explain why something is problematic. The same function performs as the combination *insofern* <problematisch sein>, *als* (32).

31. Die Änderung **ist** auch **deshalb problematisch weil** sie keinerlei Übergangsregelungen vorsieht.
32. In Tibet , überall in China **sind** Trekkings **insofern problematisch, als** man sie nur mit chinesischer Organisation unternehmen kann und es dort so gut wie keine Müllvermeidung oder -Entsorgung gibt.
33. Der Krieg hat **mehr Probleme geschaffen** als er gelöst hat.

The verb <schaffen>, like its most frequent English translation correspondence <create>, produces the expression <mehr> *Probleme*|*Schwierigkeiten* <schaffen als NP lösen>. Similar to its English cognates the construction denotes that something creates more problems than it solves, as exemplified in (33). The construction is more frequent with the noun <Problem> than <Schwierigkeit>. In my corpus there are also some combinations with the verbs <bereiten> and <verursachen> but they are very infrequent.

The lexical item <Ursache GEN|für> is found with three relatively frequent unique modifiers when it collocates with *Probleme*. These modifiers specify that the problems caused are relevant at the present time. This collocation is formed with the adjective <derzeitig> which is at the same time among the most frequent collocates of the given noun phrase. Two other modifiers that are used in this context are <gegenwärtig> and <bestehend>.

Among other unique modifiers we have <existentiell> that occurs with the verb <führen>, <riesengroß> and <elektronisch> that occurs with <bereiten>, <kontrovers> which is found with <aufwerfen> and <Verhandlungsproblem>, <Umsetzungsproblem>,

<Randproblem> and <Quellenproblem> which selects for <darstellen>. However, these combinations have low frequency in my corpus as they occur only two or three times.

5.3.2.5 Conclusion

The above sections in which the distribution of the German lexical items from the TLD {PROBLEM BEREITEN} was investigated showed that the distinguishing features can be identified by applying the distributional analysis. The study, therefore, confirms the validity of the distributional model.

First, general local contexts in which German items occur were examined which helped to define the general structures and the local grammar classes for the lexical items in question. Against this background, we distinguished between the following three types of linguistic units: items that consist of a nominal and a verbal element, one noun phrase and one adjectival construction. After that, within the first group of items we identified expressions that contain a transitive verb, those which contain an intransitive verb and one existential construction. A comparative analysis of the elements that belong to the local grammar classes revealed distinguishers. First the distinguishers related to the distribution of verbal elements were identified and after that those that are related to the distribution of nominal elements. Two types of distinguishers were observed: specific differences and general trends. Specific differences show how particular lexical items differ from each other. General trends show the behaviour of lexical items with regard to their frequency, co-occurrence with modifiers and collocation strength. Thus, we saw that the number of modifiers and collocational strength tended to correlate with the frequency of lexical items.

The distinguishing features observed are summarised in Tables 5.28-5.30. The first table provides general information regarding the grammar patterns, transitivity, passive voice, occurrence with RECIPIENTS and modal verbs. It makes possible to compare general distinguishers in relation to all lexical items. For example, by comparing the occurrence of the verb <bereiten> and <bringen> and we can observe that the two items share the same

general grammar pattern, the use of transitivity and co-occurrence with RECIPIENTS. On the other hand, only <bereiten> occurs significantly in the passive and <bringen> is more typically used with modal verbs. The table provides a slightly simplified picture because the type of modals are specified or nor does contain information regarding textual meaning. Nevertheless, it can serve as a useful summary of general distinguishing features.

Lexical items	**Grammar pattern**	**Transitivity**	**Passive**	**RECIPIENT**	**Typicality of occurrences with modal verbs**
<Problem|Schwierigkeit es geben>	NP+V+NP	EX			11
<Problem|Schwierigkeit bereiten>	NP+V+NP	TR/DIT	√	INDIR/<für>+NP	8
<zu Problem|Schwierigkeit führen>	NP+V+NP	TR			4
<Problem|Schwierigkeit darstellen>	NP+V+NP	TR	√	<für>+NP	6
<Problem|Schwierigkeit bringen>	NP+V+NP	TR/DIT		INDIR/<für>+NP	3
<Problem|Schwierigkeit schaffen>	NP+V+NP	TR		<für>+NP	1
<Problem|Schwierigkeit verursachen>	NP+V+NP	TR	√		5
<Problem|Schwierigkeit aufwerfen	NP+V+NP	TR			10
<Problem|Schwierigkeit entstehen>	NP+VP	INTR			7
<Problem|Schwierigkeit sich ergeben>	NP+VP	INTR		FÜR+NP	9
<Problem|Schwierigkeit auftreten>	NP+VP	INTR			2
<problematisch sein>	NP+LINKING V+ADJ			FÜR+NP	11
<Ursache GEN|für Problem|Schwierigkeit>	NP+<sein>+NP				

Table 5.28: Distinguishing features for lexical items from the TLD {PROBLEM BEREITEN}

Tables 5.29 and 5.30, which summarise the results of the local grammar analysis, are more detailed and they display the distribution of lexical items in relation to the context of different word forms of <Problem> and <Schwierigkeit>. These tables contain information regarding the general frequency and typicality of specific combinations. Typicality refers to the values of co-occurrence strength. The items are ordered according to the rank statistics. Thus, it can be seen that the verbs <aufwerfen> and <verursachen> and the noun phrase <Ursache GEN|für> are ranked lower in relation to the frequency parameter. The verb <darstellen> is infrequent

with the plural form of the two nouns but very frequent with the singular. The opposite is true for the verb <bereiten>. Together with <auftreten> and <es geben> this verb forms the strongest collocations with <kein> and *Probleme* and *Schwierigkeiten*. Similarly, the existential <es geben> and <darstellen> selects most typically for <kein> and *Problem*. With both definite and indefinite determiners we most commonly find <Ursache GEN|für>. In addition, <verursachen> tends to occur in the expressions with definite determiners, whereas <es geben> is strongly associated with indefinite determiners when they precede the noun <Problem>. On the other hand, <führen zu> strongly collocates with the same type of indeterminates but only when they occur with the plural form of the two nouns.

The above results are significant because they indicate differences between the behaviour of lexical items. These differences show the substitution potential of the lexical items in question. Thus, using this term one can say that the lexical units that occur with higher values have higher substitution potential when observed in relation to specific features. The substitution potential of individual lexical items varies because the association strength of collocations and colligations in which they are used also varies.

Variable Lexical items	Frequency *Problem*	Typicality <kein> + *Problem*	Typicality Definite + *Problem*	Typicality Indefinite + *Problem*	Typicality Modifiers + *Problem*	Frequency *Probleme | Problemen*	Typicality <kein> + *Probleme | Problemen*	Typicality Definite + *Probleme | Problemen*	Typicality Indefinite + *Probleme | Problemen*	Typicality Modifiers + *Probleme | Problemen*
<auftreten>	5		2	5	5	3	3	7	9	3
<aufwerfen>	8		3	7	1	9	6	11	10	1
<bereiten>	9		6	2	4	1	1	8	5	11
<bringen>	6		5	9	10	8	5	4	8	4
<darstellen>	1	1	10	3	2	12	8	5	12	9
<entstehen>	4		8	8	8	4	10	9	11	7
<es geben>	3	2	11	1	11	2	2	10	3	12
<führen zu>	12	3	12	12	7	6	9	12	2	8
<schaffen>	7		7	4	9	7	6	3	7	2
<sich ergeben>	2		9	6	3	5	11	6	4	5
<verursachen>	11		4	11	6	10	4	2	6	6
<Ursache GEN|für>	10		1	10	12	11		1	1	10

Table 5.29: Distinguishing features for the expressions formed with the word forms *Problem* and *Probleme|Problemen*

Variable Lexical items	Frequency *Schwierigkeit*	Typicality <kein> *Schwierigkeit*	Typicality Definite + *Schwierigkeit*	Typicality Indefinite + *Schwierigkeit*	Typicality Modifiers + *Schwierigkeit*	Frequency *Schwierigkeiten*	Typicality <kein> + *Schwierigkeiten*	Typicality Definite + *Schwierigkeiten*	Typicality Indefinite + *Schwierigkeiten*	Typicality Modifiers + *Schwierigkeiten*
<auftreten>						1	2	4	10	10
<aufwerfen						11		10	8	2
<bereiten>	4	2		2	1	3	1	6	6	6
<bringen>	5		3		5	2		8	11	8
<darstellen>	2	1	2	1	2	12		2	12	4
<entstehen>	3		3		5	7		7	4	9
<es geben>	7			2	1	5	3	10	5	12
<führen zu>						6		9	1	1
<schaffen>	7				3	9		10	2	6
<sich ergeben>	1				4	4		5	9	3
<verursachen>						8		3	7	5
<Ursache GEN|für>	6		1			10		1	3	11

Table 5.30: Distinguishing features for the expressions created with the word forms *Schwierigkeit and Schwierigkeiten*

5.3.3 An interlinguistic analysis of the lexical items from the TLD {PROBLEM BEREITEN}

5.3.3.1 Correspondence potential of the German translation correspondences

In this section the distribution of the German lexical items from the TLD {PROBLEM BEREITEN} in relation to their English translation correspondence will be studied. The analysis will show the correspondence potential of these German items. Correspondence potential will be examined in terms of two variables: the number of correspondence relations and the percentage of occurrences of translation correspondences. The two variables will be calculated in the same was as in the analysis of English items. It means that the percentage of occurrences will be given in terms of assigned values and will be called correspondence degree. The number of correspondence relations relies on counting the number of correspondences that a German lexical item has in English. Correspondence potential, finally, is the product of the two variables. The purpose of the analysis is to test whether the interlinguistic approach can be applied to distinguish between the uses of German lexical items as translation correspondences. To do this it will suffice to examine only one type of expressions. In the current analysis only the correspondences that contain the plural of the noun <Problem> will be considered. These expressions are selected because of their high frequency in the parallel corpus. According to predefined conditions, only the German lexical items that establish relations with at least two English items and those that are used at least three percent of the time will be taken into consideration.

Table 5.31 sums up the results obtained in the analysis. As can be seen from the fourth column, the items from the TLD {PROBLEM BEREITEN} do not have an equal number of English correspondences. These differences are, nevertheless, not statistically significant as the p-value which is 0.4 indicates. On average English lexical items correspond to between five and six German items. In three cases are the values higher than this and in four cases lower. <zu> *Problemen* <führen> has the widest usage as it is the only item which has ten translation correspondences in English.

Lexical items	Correspondence degree:total	Correspondence degree:average	Number of correspondences	CP
<zu> *Problemen* <führen>	21	2.1	10	12.1
Probleme <bringen>	13	1.4	9	10.4
Probleme <es gibt>	12	1.5	8	9.5
Probleme <verursachen>	10	1.4	7	8.4
Probleme <schaffen>	8	1.3	6	7.3
Probleme <bereiten>	8	1.3	6	7.3
Probleme <auftreten>	9	1.8	5	6.8
Probleme <entstehen>	8	1.6	5	6.6
Probleme <aufwerfen>	7	1.4	5	6.4
<problematisch sein>	8	2.7	3	5.7
Probleme <sich ergeben>	5	3	4	5.3
Probleme <darstellen>	3	1	3	4
<Ursache GEN|für> *Probleme*	2	1	2	3

Table 5.31: Distribution of translation correspondences from the TLD {PROBLEM BEREITEN}

The second and the third columns indicate to what degree are the items used as translation correspondences. From the second column we can see that the items have different distribution. This difference is statistically significant: p-value=0.0004. This is not very surprising given the fact that for example *Probleme* <darstellen> corresponds to only two English items and in both cases the correspondence degree is below 10%. On the other hand, the items such as <zu> *Problemen* <führen>, *Probleme* <es gibt> or *Probleme* <bringen> correspond to more than two English items and this correspondence is in more than two cases higher than 10%. Apart from *Probleme* <darstellen> and <Ursache GEN|für> *Probleme* all other lexical items correspond to at least one English item 10% of the time or more. This is why <problematisch sein> that is involved in equal number of correspondence relations as *Probleme* <auftreten> is ranked much higher: it corresponds to <to be problematic> more than 50% of the time.

By adding the values from the third and fourth columns we obtain the results of the correspondence potential for the German lexical items from the present domain. The highest value of correspondence potential has <zu> *Problemen* <führen>. This linguistic item corresponds to the largest number of English items and is the only item that is used in five cases as a translation correspondence more than 10% of the time. It has the highest correspondence degree with <give rise to> *problems* (20%), <lead to> *problems* (58%) and <result in> *problems* (62%).

There is a strong relationship between the second and the fourth column (r=0.88). This means that the number of correspondence relations positively correlates with the correspondence degree.

Table 5.32 contains the information about the total frequency of German lexical items and their values of correspondence potential. There is a positive correlation between the two variables (r=0.55).

Lexical items	Total frequency in deWaC	CP
Probleme <es gibt>	4030	9.5
Probleme <bereiten>	3562	7.3
<zu> *Problemen* <führen>	1754	12.1
Probleme <auftreten>	1609	6.8
Probleme <bringen>	1602	10.4
Probleme <entstehen>	1586	6.6
Probleme <schaffen>	1434	7.3
Probleme <sich ergeben>	1394	5.3
<Ursache> DET *Probleme*	988	3
Probleme <aufwerfen>	964	6.4
Probleme <verursachen>	940	8.4
Probleme <darstellen>	583	4

Table 5.32: Relationship between the distribution of lexical items from the TLD {PROBLEM BEREITEN} in the reference corpus and their correspondence potential

The correlation is stronger for the items from the top of the table. Out of six items that have correspondence potential above average four of them are among the most frequent items: *Probleme* <es gibt>, *Probleme* <bereiten>, <zu> *Problemen* <führen> and *Probleme* <bringen>. From the frequency values one would expect that the former two items have lower values than the latter two. There are other departures from the central tendency such as in the case of *Probleme* <verursachen> that has far stronger correspondence potential than would be predicted by its frequency. The opposite is true for the three constructions containing intransitive verbs. We can conclude that the prediction of correspondence potential relying on the frequency variable should be treated with caution. Therefore, it would be wrong to assume that one can automatically predict the correspondence potential of lexical items from the figures of substitution potential.

Finally, formal similarities seem to play a certain role in the current context as well. By formal similarities I refer chiefly to the grammar structures and etymological similarities between lexical items. Grammatically similar expressions tend to correspond to each other to greater extent than the items with different structures. In addition, it seems that the items with similar etymology also tend to correspond to each other. For example, the existential construction <there be> *problems* is most of the time translated as <es gibt> *Probleme*. Similarly, *problems* <arise> is translated most often with the three German non-transitive expressions. *Probleme* <verursachen> is the most preferred correspondence for <cause> *problems* and both verbs derive from the Latin noun *causa*. Similarly, <create> derives from Latin *creāre* and it was initially translated into Old English as *sciepani* which is etymologically related to the Old High German *skephen.* From this word comes <schaffen> which serves as the most frequent correspondence of <create>. Given that the TLD {CAUSE PROBLEM} contains only one intransitive expression and that in German there are three such expressions it is not very surprising that these German items have lower correspondence potential than transitive constructions. Formal similarities may also account for the higher value of correspondence potential for <zu> *Problemen* <führen> than would be predicted by its frequency. Namely, in English there are three lexical items related to the sense of moving along a specific path. Finally, the fact that English does not have a word which would be etymologically similar to the German verb <bereiten> might be the reason why *Probleme* <bereiten> is used less often as a translation correspondence that we would expect from its high frequency in the reference corpus.

5.3.3.2 Conclusion

The above distribution analysis of the lexical items from the TLD {PROBLEM BEREITEN} showed that the distributional method can help to arrive at distinguishing features between these items. This follows from both the intralingual and interlingual analysis carried out above.

The intralingual analysis pointed out specific and general distinctions between German lexical units. Specific features are related to the occurrence in specific contexts or to use the

term introduced in Chapter 3 in specific language games. General features have to do with the frequency variable and the number of collocates associated with an item. For a lexical item that occurs more typically in a given context than others it was said that it had higher substitution potential. In other words, it can replace less frequent items in that context.

The interlingual comparison of German and English lexical items showed differences between German items in terms of the number of correspondence relations and in relation to how often they were selected as translation correspondences. Here we observed a tendency that the lexical items that had fewer correspondence relations also had a lower correspondence degree. We also saw that although lexical items that occurred with high frequency in the reference corpus tended to have higher correspondence potential the variables frequency and correspondence potential were not perfectly correlated. Finally, it was suggested that the formal similarities and common etymology for the items from two languages might bear some influence on their use as translation correspondences. However, this issue needs to be explored in greater detail.

Two analyses give information regarding the substitution and correspondence potential of the lexical units. These two variables are not necessarily in accord to each other and they provide two different views on the structure of the TLD {PROBLEM BEREITEN}. The lexical items that are central from the intralinguistic point of view (high substitution potential) are not automatically central from the interlinguistic point of view (high correspondence potential).

Finally, we can conclude that both analyses confirm the feasibility of the distributional model in the study of differences between the lexical units that belong to the same translation lexical domains.

Chapter 6 Identification of the TLD {MANY COLLECTIVES} and {VIELE KOLLEKTIVA} and TLSd {MANY PROBLEMS} and {VIELE PROBLEME}

6.1 Introduction

I will begin by discussing the results of the distributional method applied to the analysis of lexical items in English and German. Thereafter, I shall discuss the purpose of investigations carried out in the present chapter and in Chapter 7.

In the first stage the distributional model was applied in Chapter 4 with the aim of answering the following question.

- Is it possible to generate cross-linguistic semantic sets by relying purely on the distribution of the corresponding lexical items from L1 and L2 in a parallel corpus?

Following Johansson (2007: 5) the lexical items from English and German corresponding to one another in the parallel corpus were treated as translation correspondences. The reasons for referring to them with this term rather with translation equivalents were discussed in 3.2.3. These translation correspondences were identified in the Europarl parallel corpus by relying on two distributional assumptions:

i) If a lexical item from L2 corresponds to a L1 item the two will occur in a similar textual context and as such constitute a cross-linguistic substitution set;

ii) If there is more than one lexical item from L2 that corresponds to the same item from L1 it will follow that all these L2 items occur in a similar textual context and therefore will belong to the same substitution set.

A case study conducted in Chapter 4 proved the validity of these assumptions. By exploring the contexts of translation correspondences the substitution sets for English and German lexical

units were established. These sets were named translation lexical domains (TLD). The English lexical domain established in this way was labelled {CAUSE PROBLEM}. The corresponding German domain was named {PROBLEM BEREITEN}. It was concluded that the findings provided a positive answer to the above questions.

The purpose of Chapter 5 was to explore whether the distributional method can deal with the following three questions:

- Is it possible to identify distinguishing features for lexical items that belong to the same TLD by focusing only on their distribution in the reference corpora, i.e. their use in textual contexts?
- Is it possible to find differences between translation correspondences from L2 by comparing their distribution in relation to the lexical units from L1?
- Is it possible to describe the structure of lexical domains by exploring distributional features of lexical items and their use as translation correspondences?

The first question was explored in sections 5.2.1 and 5.2.2 for English and in sections 5.3.1 to 5.3.2 for German. Here, the distinguishing features were explored from an intralinguistic perspective. The items from both languages were compared initially only in terms of their general grammatical features which revealed first type of differences between the items belonging to the same TLD. After that typical contexts in which they occur were studied and the lexical units from these contexts were classified into local grammar sets according to the functions that they perform. In this connection the null-hypothesis according to which the distribution of items from two TLD was not significantly different was proposed. The use of various statistical measurements helped to falsify this hypothesis. The results helped to establish the distributional distinguishing features of lexical items. The distinguishing features discovered are of two kinds. First, there are features having to do with particular collocates that occur with varying likelihood with lexical items. For example, modal verbs occur, in general, more typically with the lexical items <result in problem|difficulty>, <cause problem|difficulty> and <lead to problem|difficulty> than with <there be problem|difficulty>. These features are

summarised in Tables 5.12, 5.13 and 5.14 for the items from the TLD {CAUSE PROBLEM} and in 5.26, 5.26 and 5.28 for the items that belong to the TLD {PROBLEM BEREITEN}. The second kind of feature has to do with more general tendencies. These tendencies are related to the following variables: frequency, number of collocates and the values of collocation strength. In brief, we observed that the more frequent lexical items tended to occur with a larger number of collocates, that they shared most of the frequent collocates of the less frequent items and that the resulting collocations mostly had higher association strength. All distinguishing features apart from the general grammatical differences such as those concerned with transitivity of verbs are of probabilistic nature. The features were interpreted in terms of what was named the substitution potential of lexical items from the same domain. The analysis, therefore, showed that the distributional model can answer the second question in terms of substitution potential.

The second of the two questions was studied in the section 5.2.3 for English items and 5.3.3 for German lexical items. The purpose of this study was to investigate the number of relations between the lexical items from L1 and their translation correspondences on the one hand, and the percentage with which these correspondences were used on the other hand. As a result, we arrived at the values of correspondence potential for all translation correspondences from the two languages. The results obtained showed that the distributional model can provide an answer to the third question. In addition, using these findings it was possible to show the structure of lexical domains can be accounted for in terms of the distribution of translation correspondences. The results obtained therefore provide an answer to the third question.

All these results indicated that the distributional method can yield purely distributional distinguishing features for the lexical units that belong to the same TLD.

The first purpose of the current and the following chapter is to test the predictive power of the distributional model. It means that the general conclusions from previous studies will serve as the starting point for the analysis reported below. Four additional cases will be studied. They are, of course, insufficient to conclude that the observed phenomena are true for the

whole of the languages compared. However, if the results outlined below turn out to be positive it will at least suggest that the model might have wider potential applicability.

The second purpose is to test if the distributional model can be extended. Unlike in the previous study where a specific collocation (<give rise to problem>) was selected at the beginning of the analysis and only the lexical items that are substitutable with it were subsequently considered, here I will begin with a more general lexical item. This item consists of <many> and its colligation with plural nouns. To understand why this extension of the method might be important we need briefly to return to Chapter 4. In Table 4.2 we saw that some German lexical items corresponded to <give rise to> when it collocated with different nouns. For example, the verb <entstehen> corresponds to <give rise to> when it collocates with <Problem|Verwirrung|Schwierigkeit|Frage|Kosten> and when the English item co-occurs with <problem|confusion|difficulty|question|cost>. We also saw that the two verbs were part of the same domain with other German and English items only when they collocated with the nouns <Problem> and <Schwierigkeit>, that is, with <problem> and <difficulty>. Therefore, we could conclude that the two verbs formed the same domain when they occurred in the specific context of the two nouns but also that they share some collocates outside this domain. Similar can be said about other items studied. This suggests that there might be a higher-level domain for these items. This issue was not investigated in Chapter 4 and Chapter 5 due to space limitation and will be addressed here in a study of a new set of lexical items. First, it will be assumed that a higher-level TLD can be established by selecting at the beginning a general colligation instead of a specific collocation. After that, it will be expected that a description of a particular portion of this higher-level TLD can be obtained by focusing on a specific set of translation correspondences. This particular portion of the given TLD will present a lower-level TLD and will be called form a *translation lexical sub-domain* (TLSd). If the study confirms these assumptions it will follow that the distributional model can be applied from both a general and specific perspective. It seems that the general perspective is more attractive if an item occurs with a large number of collocates because in such cases the representations of data as reported in Table 4.2 would be inefficient. It seems to be more reasonable here to start from the top and then explore particular areas of a domain. On the other hand, if the collocations in which an

item occurs are less numerous one can start from below and explore one sub-domain at time which would eventually lead to the description of a higher-level domain.

6.2 Translation lexical domains {MANY COLLECTIVES} and {VIELE KOLLEKTIVA}

In this chapter the procedure of identifying translation correspondences in the parallel corpus introduced in Chapter 4 will be applied for the analysis of the lexical item <many> and its German correspondences. This lexical item is chosen partly due to the fact that it occurs as one of the modifiers with the lexical items from the TLD {CAUSE PROBLEM}. Thus, a comparison of the results from two studies will make it possible to draw some general conclusion in Chapter 8 regarding the relationship between different domains and sub-domains.

Contemporary dictionaries usually distinguish between the following three uses of <many> in front of plural nouns: as a pronoun, as a predeterminer and as a noun (Rundell, 2007; Sinclair, 1995). I will focus on its co-occurrence with plural nouns as in <many people>, <many years>, <many problems>, <many cases>, <many countries> or with noun phrases as in <many credit cards>, <many football teams>, <many call centres>, etc. Since these nouns refer to a group of people, things or events such nouns are called collective nouns or collectives (e.g. Sinclair 1990: 16-17) and will be coded in the present thesis as COLLECTIVES when the English linguistic units are concerned and KOLLEKTIVA when the German items are considered.

Table 6.1 displays nine German lexical units which correspond to the colligation <many COLLECTIVES> in the Europarl corpus. It is possible that some items have not been noticed in the analysis but if this is the case they occur with very low frequency in the present corpus. The units identified account for 94% of the distribution of <many COLLECTIVES>.

In the next stage, all these correspondences are translated back into English in order to find out if some other lexical items from this language belong to the same substitution set. A series of back translations was performed and repeated until no new items were generated. Finally, only the lexical units meeting the following three criteria, discussed in Chapter 4, were included in the newly established TLD:

a) an item from L2 corresponds to at least two lexical items from L1;

b) an item from L2 occurs at least twice when it corresponds to an item from L1;

c) an item from L2 corresponds at least to two percent of the occurrence of an item from L1.

Lexical items	**Frequency**
<many COLLECTIVES>	36223
<viele COLLECTIVES>	29029
<zahlreiche COLLECTIVES>	3899
<mehrere COLLECTIVES>	469
<eine Reihe ART|APPR COLLECTIVES>	162
<eine Vielzahl ART|APPR COLLECTIVES>	184
<eine große Zahl|Anzahl ART|APPR COLLECTIVES>	163
<ein großer Teil ART|APPR COLLECTIVES>	112
<eine Menge ART|APPR COLLECTIVES>	106
<eine beträchtliche Zahl|Anzahl ART|APPR COLLECTIVES>	12

Table 6.1: <many COLLECTIVES> and corresponding German lexical items from the Europarl corpus

The lexical domains established in this way were named {MANY COLLECTIVES} and {VIELE KOLLEKTIVA} according to the most frequent lexical items. The items from the two languages that are considered in the analysis carried out in the next chapter are displayed in Table A3 and Table A4 in Appendix A. At the top of the tables one can see the lexical items from the source language with their frequency after which follow the items from the target language and the frequency of their occurrence as translation correspondences. The search process was performed semi-manually. Potential corresponding items were first identified with ParaConc by using a feature called 'hot words' that highlights translations of the search term (Barlow, 2002: 22). After that, the concordance lines with similar textual contexts in which these translations occur were manually explored to make sure that all relevant items were located. It is possible that some lexical items were overlooked but I do not expect that the new items would significantly skew the results. The items thus identified comprise between about 70% and 95%

of all translation correspondences. As explained in 3.5, the grammatical tags ART and APPR are used to code the grammatical categories determiner and preposition, respectively.

In the present data there are ten lexical items from the TLD {MANY COLLECTIVES} and nine from the corresponding TLD {VIELE KOLLEKTIVA} that are identified following the aforementioned criteria. From the above tables one can assume that there are other items with similar behaviour. Some of them are specific to only one lexical item such as <a set of COLLECTIVES> or <a raft of COLLECTIVES> which corresponds only to <eine Reihe ART|APPR KOLLEKTIVA>. Some other, such as <a host of COLLECTIVES>, corresponds to more than one German lexical item but always below the threshold of two percent. The former case means that the translation correspondences are important only in relation to a specific item, whereas the latter suggests that the items are not among the preferred options. Using the terminology introduced in Chapter 5, we can assert that such items lack correspondence potential in the present TLD. In addition, the items of the first type seem to indicate that the given lexical item might also belong to other TLD. This is the case, for example, with the lexical item <mehrere KOLLEKTIVA>. More than half of the time it is translated as <several COLLECTIVES> which does not correspond to other items from the TLD {VIELE KOLLEKTIVA}. Such a high correspondence overlap between them implies that the two items belong to another domain. Or to put it into traditional lexicographic terms, <mehrere KOLLEKTIVA> is obviously a polysemous item. Membership of a lexical item to more than one TLD is certainly a significant question, but due to space restriction it will not be possible to discuss upon it in greater detail.

6.3 Translation lexical sub-domains {MANY PROBLEMS} and {VIELE PROBLEME}

The data described in the previous section accounts for the correspondence relation between the English and German lexical items from two lexical domains from a global perspective. As it has already been said, each collocation can be used as a starting point for a further analysis of particular sections. The collocation selected for study in this section is <many problems>. This

term belongs to {CAUSE PROBLEM} and the {MANY COLLECTIVES} and its analysis will make possible to consider relations between the two domains.

The same procedure as before was used to identify the German translation correspondences of <many problems> and to establish the translation lexical sub-domains (TLSd). A list of lexical items that will be considered are displayed in Table A5 and Table A6 in Appendix A. Seven English and five German lexical items that meet the pre-defined criteria are found in these sub-domains. The sub-domains are labelled according to the most frequent lexical units: TLSd {MANY PROBLEMS} and {VIELE PROBLEME}.

The lexical items from these sub-domains consist of two elements: an adjectival item and a noun. The function of the adjectival items is to modify the meaning of the nominal element by denoting large quantities. Following the terminology introduced in Chapter 5 they will be referred to as quantifiers. The second element in the above tables is coded as PROBLEMS. Apart from the noun <problem> it contains other nouns as well. Thus, in English we also find <difficulty>, <issue>, <subject>, <point>, <question> and <topic> all of which are used in the plural word forms. In German, in addition to <Problem> we also have <Schwierigkeit>, <Frage>, <Theme> and <Punkt>. All these nouns are mutually substitutable.

6.4 Conclusion

In this section the TLD {MANY COLLECTIVES} and {VIELE KOLLEKTIVA} and the TLSd {MANY PROBLEMS} and {VIELE PROBLEME} were identified applying the distributional model to the data from the Europal parallel corpus. These domains and sub-domains confirm the applicability of the distributional method to the identification of sets of corresponding items. Rather than selecting a specific collocation as a starting unit of analysis as it was the case in Chapter 4, in the present study a higher-level lexical unit served as a point of departure. This does not alter the nature of the results obtained. The results from this and the previous analysis show that the substitution sets of corresponding items can be generated from a parallel corpus regardless of the selected units of analysis.

The example of the lexical item <mehrere KOLLEKTIVA> and its most frequent correspondence <several COLLECTIVES> indicates that a model has a potential of dealing with the issue of polysemy by placing the lexical items that have more than one sense into different domains. For the sake of space, this, however, has not been explored here in detail.

The identification of the sub-domains {MANY PROBLEMS} and {VIELE PROBLEME} shows that it is possible to begin with a general domain and then investigate in more detail one of its particular parts by focusing on specific lexical items occurring in the same context. Above, only one such context was explored but there is no reason that the model would not be applicable to the study of sub-domains. The final result would be a description of all sub-domains that constitute a higher-level domain.

Relying on the criteria introduced to avoid untypical translation correspondences an inventory of the lexical items for both languages was established. These items will be further explored in relation to the second and third questions discussed in 6.2.

Chapter 7 TLD {MANY COLLECTIVES} and {VIELE KOLLEKTIVA} and TLSd {MANY PROBLEMS} and {VIELE PROBLEME}

7.1 Introduction

In the present chapter the distributional method will be applied to the study of lexical items from the TLD {MANY COLLECTIVES} and TLSd {MANY PROBLEMS} for English and from the TLD {VIELE KOLLEKTIVA} and TLSd {VIELE PROBLEME} for German. In particular, the focus will be on differences studied in terms of the variables that describe the substitution and correspondence potential of lexical items. The studies will be carried out from an intralinguistic and an interlinguistic perspective. The section 7.2 will be concerned with English and the section 7.3 with German lexical items.

From the previous intralinguistic study the following major patterns are expected to be found in the current data:

i) Correlation between the frequency of lexical items and the number of associated collocates.

ii) If *x* and *y* are two lexical items belonging to the same domain or sub-domain and the former is more frequent it is expected that this item will occur with a significant number of the collocates of *y*.

iii) Correlation between the frequency of lexical items and the number of stronger collocations in which it occurs.

In the study of correspondence potential we it will be expected that:

i) The lexical items from the same domain or sub-domain differ in relation to the correspondence degree and the number of correspondence relations.

ii) Correlation between the correspondence relation and the percentage by which translation correspondences are selected.

iii) Correlation between substitution potential and correspondence potential.

The lexical items analysed in this chapter derive from ukWaC and deWaC for English and German, respectively. The data was studied using the *Sketch Engine* tools, the *IMS Corpus Workbench* tools and the *RCQP* package. The duplicate sentences and the lines containing errors have been removed. The relations between variables were studied with the correlation test and the significance of differences with the chi-square test and the t-test. The values of association strength are based on the logDice coefficient.

Due to limitations of space it will not be possible to consider the contexts in which the lexical items occur and to explore them in terms of the local grammar classes. This investigation would no doubt indicate further differences between lexical items. For example, it is only <many COLLECTIVES> and <a lot of COLLECTIVES> that occur with a set of lexical items the function of which is to amplify meaning. Thus, with the former lexical unit we find <so> such as in (1) and with the latter <quite> and <rather> such as in (2-3).

1. The plague that killed **so many people** in Europe in the 1300s started on ships.
2. We can see from this that there are **quite a lot of things** that can affect the profit a firm makes.
3. I think there are **rather a lot of problems** with this argument.

7.2 TLD {MANY COLLECTIVES} and TLSd {MANY PROBLEMS}

This section is divided into four parts. 7.2.1 explores differences between the frequency of lexical and the number of collocates associated with quantifiers. In 7.2.2 the collective nouns will be classified into a provisional set of local grammar classes to explore further differences between quantifiers. After that a detailed contrastive analysis of shared collocates will be carried out. The unique collocates are dealt with in 7.2.4 and the subsection 7.2.5 will be concerned with the lexical items from the TLSd {MANY PROBLEMS} and the collocations consisting of quantifiers and the nouns from the local grammar set PEOPLE. Finally, 7.2.6

provides a description of the correspondence potential of the lexical items from the TLD {MANY COLLECTIVES} and TLSd {MANY PROBLEMS}.

7.2.1 Frequency and the number of collocates

There are ten lexical items in the TLD {MANY COLLECTIVES}: <many COLLECTIVES>, <a number of COLLECTIVES>, <a series of COLLECTIVES>, <numerous COLLECTIVES>, <a lot of COLLECTIVES>, <a large number of COLLECTIVES>, <a significant number of COLLECTIVES>, <a considerable number of COLLECTIVES>, <a huge number of COLLECTIVES> and <a substantial number of COLLECTIVES>. They are made up of two functionally different elements. The first element denotes large quantity. As we have seen in Chapter 5, the items that denote this meaning were called quantifiers they referred both to large and small quantities. Given the fact that the present items express only the former meaning they will be called *positive quantifiers*. In terms of grammar, positive quantifiers are either adjectives (<numerous> and <many>) or multi-word noun phrases (all other lexical units). The second element consists of plural or mass nouns. Since their function is to refer to a group of people, things or events they will be called collective nouns and will be coded as COLLECTIVES. In the current study only the direct co-occurrence of positive quantifiers and collective nouns are taken into account.

Table 7.1 displays the distribution of lexical items from the TLD {MANY COLLECTIVES} in ukWaC. The first column contains the name of the items, the second column shows the frequency of these items, and the third column indicates the number of COLLECTIVES that collocate with positive quantifiers.

Lexical items from the TLD {MANY COLLECTIVES}	Frequency of the lexical items in ukWaC	Number of the noun collocates
<many>	312912	4983
<a number of>	131472	3419
<a range of>	54663	1693
<a series of>	37868	2146
<a lot of>	32679	1818
<numerous>	31439	2131
<a large number of>	8818	984
<a significant number of>	1285	198
<a huge number of>	623	136
<a considerable number of>	565	124
<a substantial number of>	341	84

Table 7.1: Distribution of lexical items from the TLD {MANY COLLECTIVES} in ukWaC

From the previous results we would expect to find here statistically significant differences between the values for both variables. We would also expect to see a strong correlation between two variables. These two assumptions can be formulated negatively in the form of a null-hypothesis as was done in Chapter 5. This makes it possible to test the assumptions in terms of statistical significance with the help of the correlation test and significance test. As for the first assumption, the null-hypothesis claims that the lexical items from the TLD {MANY COLLECTIVES} are equally probable and that all positive quantifiers co-occur with all collective nouns. As for the second assumption, because the previous null-hypothesis claims that the values for both variables are identical the second null-hypothesis will claim that the correlation coefficient between them will be zero

Now, from the above table it seems that the first hypothesis is false. The most frequent combination of a positive quantifier and collective nouns (<many COLLECTIVES>) is 917 times more frequent that the least common combination (<a substantial number of COLLECTIVES>). Every other lexical item is between one and six times more frequent than the next one. To test whether these differences are not by chance the chi-square test will be used. The result calculated with the programming language R suggests that the frequency values differ significantly: χ^2 = 1573410, df = 9, p-value < 2.2e-16. We have a very high chi value and the p-

value is smaller than the predefined threshold of 0.05 which indicates that the frequency values differ from a chance distribution.

We also find significant differences between the values for the variable the number of noun collocates. There are almost 60 times as many collective nouns with <many> as with <a substantial number of>. Similarly, every lexical item is between one and five times more frequent than the next one. The chi value for this variable is 15121.38 and the p-value is again below the lowest value that can be registered: p value < 2.2e-16.

Thus, according to the results we can reject the first null-hypothesis and confirm our expectation that there are significant differences between the frequency of lexical items and the number of collocates.

The Pearson correlation will be employed to test the second null-hypothesis. The correlation coefficient is r=0.91 which indicates a very strong relationship between the two variables. We can be fairly certain that this is not due to chance because of the very high confidence interval (95%) and a very low p-value = 9.472e-05. This result indicates that higher the frequency of a lexical item the larger the number of noun collocates it has. The reverse is also true. The correlation is not 100%. The first exception is <a range of> which has higher frequency than <a series of>, <a lot of> and <numerous> but collocates with fewer collectives. The second exception is <numerous> which is less frequent than <a lot of> but occurs with more collocates.

Graphically, this correlation can be displayed as in Figure 7.1. The diagrams are based on the calculations of the closest values of the two variables. There are parallels in the distribution of the values in both cases. This is represented in the form of three clusters. The first cluster contains <many> and <a number of> because differences in values for these two items are smaller than between each of them and other lexical items. The other two clusters are based on the same principle. One can notice differences in the ordering of the items in the second cluster on two graphs. These differences mirror a discrepancy between frequency and the number of associated noun collocates. On the left-hand graph <a range of> is slightly detached from other three items because it is about one and a half time more frequent than these items and the difference in frequency between these three items is lower. On the right-hand graph it

can also be observed that the values of the number of collocates for <a series of> and <numerous> on the one hand and for <a range of> and <a lot of> on the other are more similar than for the variably frequency. This is why these four items constitute two sub-clusters on the right-hand graph.

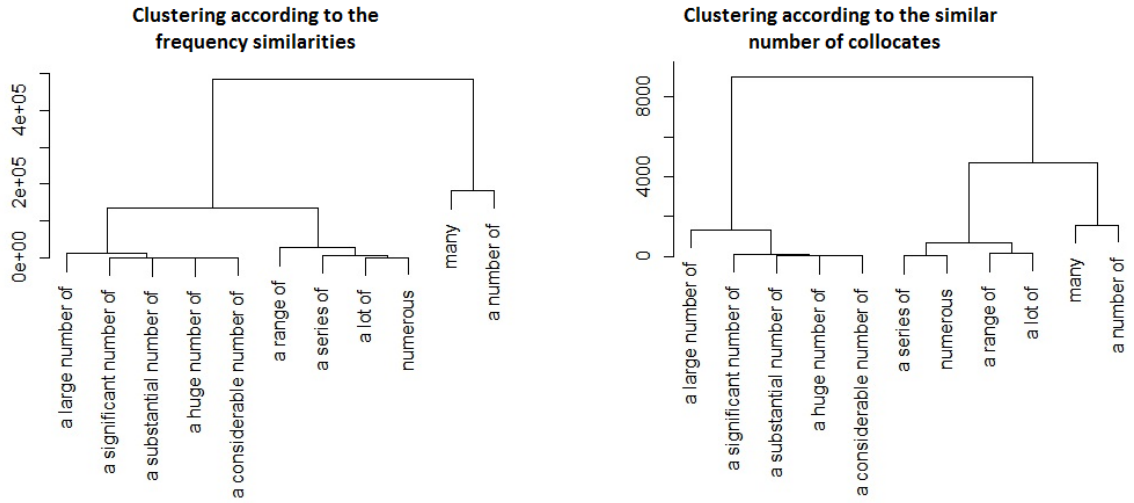


Figure 7.1: Correlation between the frequency of lexical items and the number of noun collocates

We can conclude that since the correlation coefficient has a very high value we can safely reject the second null-hypothesis.

These findings comply with the results obtained in previous studies and confirm the validity of the distributional model for use in an intralinguistic analysis of differences between lexical items from the same domain.

7.2.2 Classification of COLLECTIVES

The purpose of this section is to explore whether the collective nouns can be further classified into smaller groups according to their meaning and function. This classification should bring new insights regarding the type of collective nouns that collocate with positive quantifiers from the TLD {MANY COLLECTIVES}. One serious obstacle for this task is a very high number of items

that needs to be considered. This analysis cannot be conducted automatically. For this reason, I will focus only on the first 5% collocates. Although the selected items present only a small portion of the total number of collocates their frequency comprises between 40% and 67% of the total frequency of collocations formed with positive quantifiers and collective nouns. Therefore, these are among the most frequent collocations in the TLD {MANY COLLECTIVES}. In addition, the study will deal with a limited number of classes that will be defined in general terms. This description does not aim at completeness but is rather of the exploratory nature.

One of the most numerous sets of nouns refers to human beings. Here we can distinguish between the items with general meaning (<people>, <persons>, <folks> or <individuals>) from those that we use to talk about groups associated with specific social settings or roles (<guests>, <students>, fans>, <children>, <members> and <friends>). Both types will be coded with the term PEOPLE.

There is a set of nouns that we use to talk about locations. Here again, we can draw a distinction between the nouns such as <area>, <part>, <place>, <location>, and <site> that refer to locations in general terms and the items that refer to specific kinds of settings such as <town>, <city>, <region>, <village>. Both types will be coded as LOCATION.

Some positive quantifiers from the present TLD also occur with collective nouns that denote time. One such item is <year> which is the most frequent collocate of <many>, <a number of>, and <a considerable number of>. Other items from this class are <day>, <hour>, <month>, <century> and <week>. This class will be coded as TIME.

One type of collective nouns has to do with general terms such as: <problem>, <question>, <issue>, <reason>, <idea>, <subject>, <matter>, <issue>, and <concept>. They will be referred to as PROBLEMS.

A set of nouns coded as OPTIONS contains nouns that have to do with selecting between different options. The most representative examples are <option>, <step>, <way>, <possibility> and <type>.

The nouns coded as EVENTS refer to various types of public or educational events such as <workshops>, <events>, <seminars>, <tutorials>, <meetings>, <conferences>, <talks>, <discussions>, <exhibitions>, <concerts>, and <gigs>.

The items <publication>, <article>, <book>, <magazine>, <edition>, <report>, <journals> and <essays> are the nouns that we use to talk about different types of publication. I will refer below to this group simply as PUBLICATIONS.

There is a small set of nouns related to complaining that will be coded as REQUEST. Three most frequent items are <requests>, <enquiries> and <complaints>.

Finally, we find a group of nouns that denote opportunities and benefits. Most representative items from this set are: <sources>, <opportunity>, <benefit>and <resources>. The group will be called RESOURCE.

A comparative analysis of the distribution of linguistic items from these sets will reveal major tendencies in their use. Only the items that are most typical with positive quantifiers will be discussed. The typicality is again based on the logDice measure of association strength.

The largest number of classes is associated with <many>. This item occurs typically with the nouns from the PEOPLE class. In addition, this quantifier forms the strongest collocations with the lexical items from the LOCATION class, the class TIME and the following two items from the ISSUE set: <question> and <problem>.

With <a number of> we find typically the following nouns from the ISSUE class: <reason>, <factor>, <concern>, <topic>, <theme> and <strategy>. From the SOLUTION class we encounter <suggestion>, <proposal>, <scheme>, <improvement>, <project>, <development> and <recommendation>. This positive quantifier and <a range of> form the strongest collocations with the noun <option> and <step>.

The most typical nouns found with <a series of> belong to the set EVENTS. The most frequent ones are <workshop>, <meeting>, <seminar>, <presentation>, <meetings>, <concert>, <conference>, <talk> and <lecture>. It also collocates more usually than other quantifiers with the PUBLICATION items <essay> and <paper>.

<numerous> forms strongest collocations with the PUBLICATION and REQUEST items. The most frequent items from the first group are <article>, <book> and <publication> and from the second <complaint> and <request>. This positive quantifier forms also strong collocations with two items that do not belong to any of the above categories: <occasion> and <award>.

<a range of> forms the most typical collocations with the following items from the set RESOURCE: <resource>, <source>, <opportunity>, <service> and <facility>.

The above description indicates that the values of collocation strength vary with regard to specific types of collective nouns. In the previous studies we observed that the frequency of lexical items corresponded strongly with the values of their collocation strength. Therefore, in the present context we would expect that the values of collocation strength for the collocations that consist of positive quantifiers and collective correspond to their frequency. Thus, we would expect that the strongest collocations are always created with <many>. As Figure 7.2 below shows, this is what we indeed find in our data. The ordering of the data is identical to that displayed above on the left-hand graph in Figure 7.1.

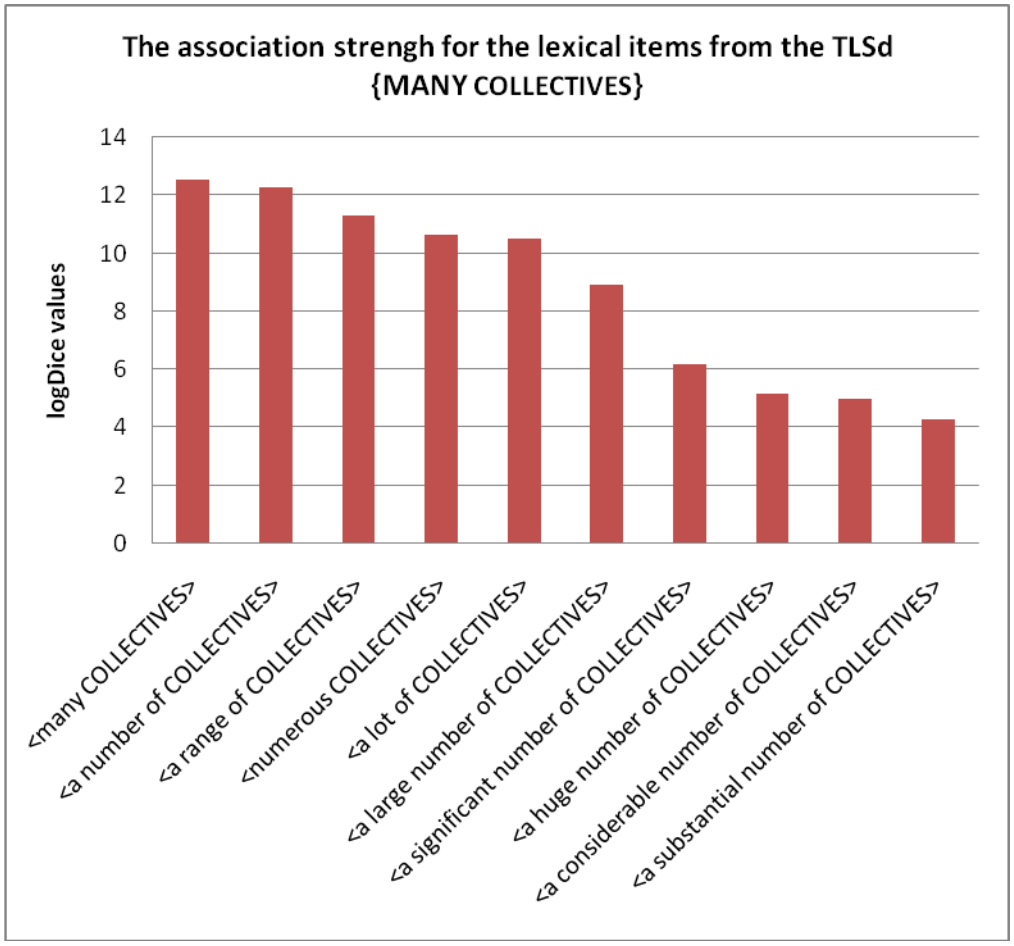


Figure 7.2: Collocation strength values for lexical items from the TLD {MANY COLLECTIVES}

However, the above investigation of specific classes of collective nouns indicates that the strength of collocations does not remain stable and that it may depend on the type of nouns with which a positive quantifier occurs. Above only the strongest collocations were discussed but in order to be able to observe how the likelihood of the collocation strengths varies we need to look at other collocations. Such an analysis was carried out for the co-occurrence of positive quantifiers with the nouns from the class PEOPLE.

The distribution of the collocations that consist of positive quantifiers and the nouns from the set PEOPLE are displayed in Figure 7.3. The following can be concluded when we compare Figure 7.2 and Figure 7.3. First, we notice that the positive quantifier <a series of> is missing from the second graph altogether. In the current corpus it does not occur with nouns from the PEOPLE set. Second, the lexical item <a lot of> replaces <a number of> as the second most typical positive quantifier that occurs with collective nouns. Similarly, the quantifier <a large number of> is both more frequent and more typical with the PEOPLE nouns than <numerous> and <a range of>. The latter two were, on the other hand, more prominent in relation to the general class of collective nouns.

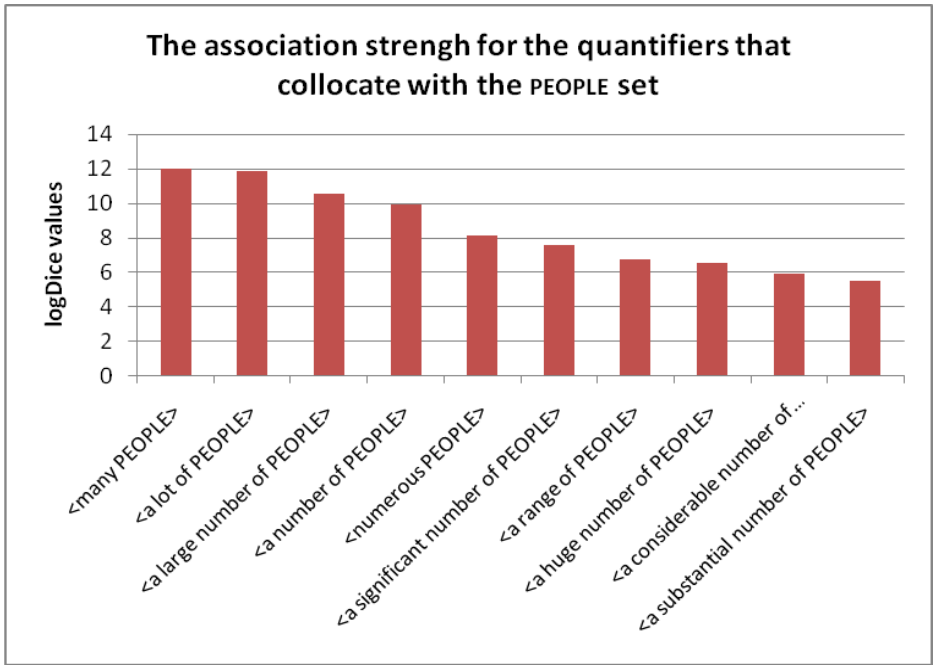


Figure 7.3: Association strength values for the collocations made up of positive quantifiers and shared collective nouns from the semantic set PEOPLE

The most typical positive quantifier with this set of nouns is <many> but very high association strength is observed also with <a lot of>. In addition, <a large number of> and <a number of> have the logDice values above the average. The comparison of differences between the logDice values indicate that the association strength of <many PEOPLE> is six times higher than that of <a lot of PEOPLE> and between 16 and 85 times as high as that of other lexical items. Similarly, the combination <a lot of PEOPLE> is between three and 70 times as strong as the collocations that consist of other positive quantifiers and the nouns from the set PEOPLE. In general, first four collocations are between seven and 85 times more frequent than other six collocations.

On this basis, the following conclusions can be drawn. First, the general set of collective nouns can be classified into smaller classes according to their meaning and function. Second, focusing on these classes we can observe how the co-occurrence of positive quantifiers with collectives varies depending on the specific set of nouns. Third, there seems to be a relation between the number of noun classes with which positive quantifiers form strongest collocations and the frequency of positive quantifiers. The quantifier <many> which, as we can see in Table 7.1, most frequently occurs with collective nouns makes at the same time the strongest collocations with four noun classes. Similarly, the next most frequent item, <a number of>, is found with three nouns and the next two items with two classes. Finally, the less frequent items do not form the strongest collocations with nouns from any of the observed classes. But, these results need to be treated with caution because <a range of> and <a lot of> in spite of their high frequency do not follow this general tendency. The comparison of the collocations created with the nouns from the set PEOPLE is another indicator that co-occurrence of positive quantifiers with noun collocates has to do with the type of nouns. This is best shown by the fact that the association strength for the collocations formed with this set of nouns does not correspond to the association strength of the combinations that contain all collective nouns. This suggests that results always need to be interpreted in relation to the context studied.

7.2.3 Shared collocates

The above analysis of the lexical items from the TLD {CAUSE PROBLEM} and {PROBLEM BEREITEN} shows that the most frequent items and those with a larger number of collocates tend to occur with a considerable portion of the collocates that combine with the less frequent items. Shared collocates occur with very high frequency. In addition, this proportion increases as the frequency difference between two items becomes bigger. The analysis also indicates that lexical items with higher frequency and a higher number of collocates were associated with a higher number of strong collocations. These findings will now be tested and further explored in a comparative study of the co-occurrence of positive quantifiers and collective nouns.

As was the case in previous analyses the results will be based on a comparison of the logDice values. The data will be obtained by using the RCQP tools that combine the CQP tools and the utilities of the programming language R.

First, we will examine the extent to which shared collocates are related to the frequency variable. This will be called *degree of overlap*. Second, it will be explored whether shared nouns occur with very high frequency. The graphs below illustrate the distribution of several lexical items from the TLD {MANY COLLECTIVES} in relation to these two issues. The complete list of graphs can be found in Appendix B in Figure B5. The values of degree of overlap are represented by red bars and the total frequency of shared collocates by blue bars. The graphs display the distribution of shared collocates in relation to more frequent positive quantifiers. Thus, the first graph compares the percentage of occurrences of collective nouns with <a number of>, <a series of>, <numerous>, <a lot of>, <a range of>, <a large number of>, <a significant number of>, <a huge number of>, <a considerable number of>, <a substantial number of> in relation to the most frequent positive quantifier <many>. Subsequently, the second graph compares the same phenomenon for <a number of> and other less frequent positive quantifiers and so on. The lexical items are ordered according to the number of collocates associated with those quantifiers.

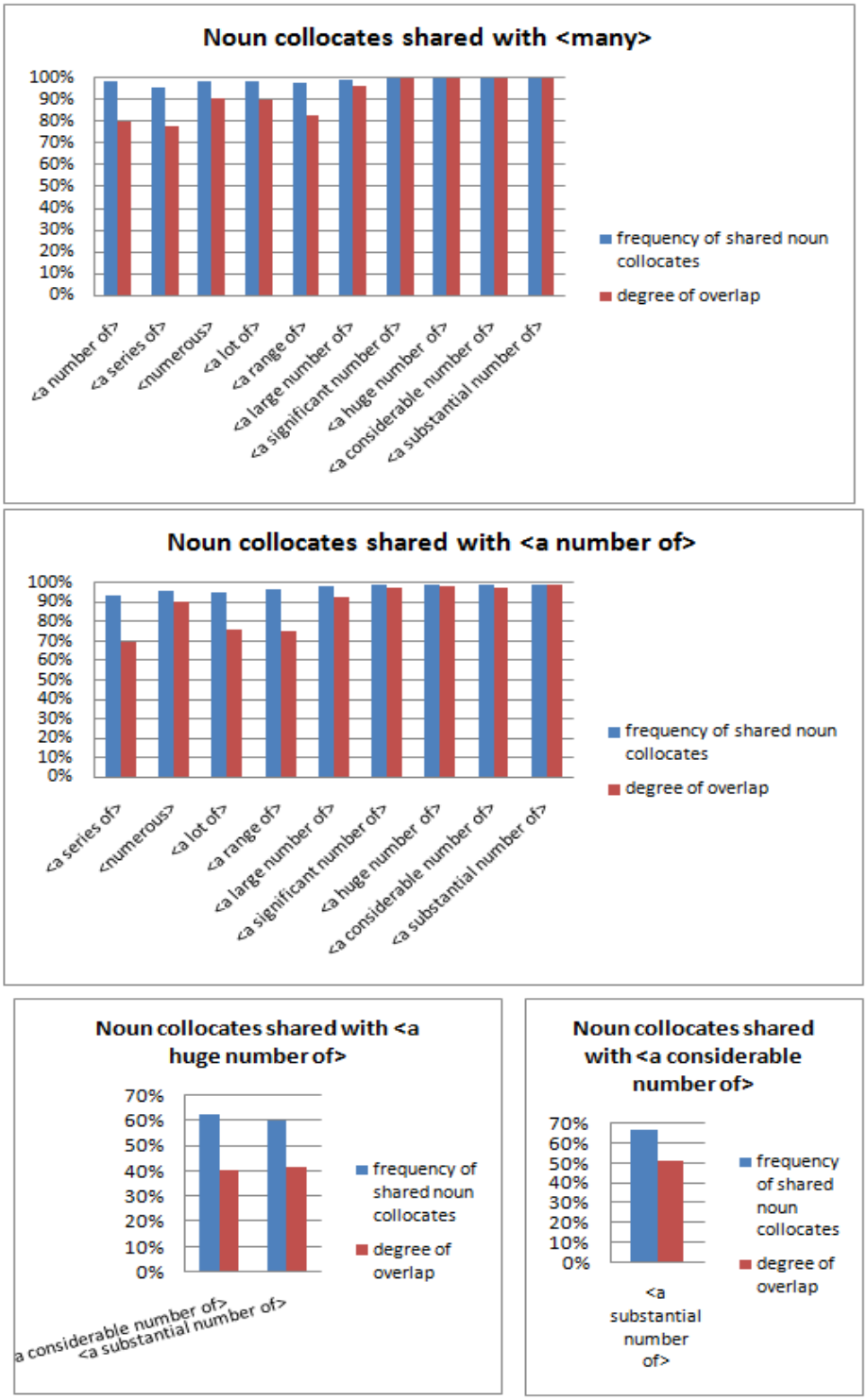


Figure 7.4: Frequency and degree of overlap of collective nouns that occur with positive quantifiers from the TLD {MANY COLLECTIVES}

The first two bars on the first graph show that 80% of the noun collocates that occur with <a number of> also occur with <many>. The total frequency here makes up 98% of the total frequency of the occurrence of <a number of> with collective nouns. The rest of the graph shows that degree of overlap for other lexical items is between 87% and 100%. This is between 98% and 100% of the total frequency of all collective nouns that occur with the less frequent quantifiers. Observing other graphs we can conclude that more frequent lexical items occur with at least 50% of collective nouns that we find with less frequent items. It was mentioned above that in the previous data degree of overlap increased as the frequency differences and differences in the number of collocates rose. The same tendency characterises the current data as well. Thus, the degree of overlap between <a number of> and <a huge number of> is higher than between the former item and <a series of>. Similarly, <numerous> and <a lot of> have very similar frequency and it is not very surprising that they have almost identical degree of overlap in relation to <a series of>. This means that the two most frequent positive quantifiers occur with almost all collocates of the four least frequent items.

There are some minor exceptions to the main tendency. <a number of> shares a greater number of collocates with <numerous> than we would expect from the frequency of the latter item and from its total number of noun collocates. Similarly, the degree of overlap for <a considerable number of> tends to be slightly lower than might be expected from its frequency and the number of collocates. This is also true for <a huge number of> and <a substantial number of> when their occurrence is compared to that <a large number of>.

The frequency bars suggest that shared collocates make up a significant proportion of all collective nouns that occur with positive quantifiers. This proportion remains high even when the number of shared collocates decreases. For example, although <a lot of> occurs with about half of the collective nouns found with <a range of> these nouns make up more than 80% of its total noun collocates. Similar figures are observed with other lexical items. For example, out of 668 collective nouns that occur with <a range of> but not with <numerous> only 14% of them occur between 10 and 171 times and as many as 260 (29%) occur only twice. The ratio is almost reverse for the noun collocates that occur with both positive quantifiers. Out of 1025 shared collocates 43% occur between 10 and 3179 and there are 162 noun collocates that occur twice

(16%). Similar results can be obtained from other lexical items. It means that shared collective nouns are among the most frequent collocates and that the majority of non-shared collocates have low frequency.

We can conclude that the patterns observed in the present data comply with the findings from previous analyses. The collocates of positive quantifiers are obviously not randomly distributed. On a general level, the current findings suggest also that that a less frequent positive quantifier in most of the cases can be substituted by a more frequent. More frequent items, therefore, have higher substitution potential.

I will now compare the values of the association strength of collocations that consist of positive quantifiers and shared collocates. The comparison will show if the patterns revealed in previous analysis can be observed in the present data. Thus, we observed before that lexical items with a larger number of collocates and/or higher frequency occurred with a higher number of stronger collocations than less frequent items. The 'and/or' suggests that this pattern can be associated with both or one of the two variables and there is no rule of precedence here. The second tendency was that the proportion of stronger collocates tended to rise with the increase in frequency differences.

Association strength is based on the logDice measurement. To compare the values of association strength I first juxtaposed the logDice values for shared collocates and then subtracted smaller values from larger values in order to find the difference in values. This difference serves as an indicator of collocation strength. For example, if a collocation AB has the value of 8.5 and CB of 8 the numerical difference of 0.5 points indicates that the former collocation is one and a half times as strong as the latter. The collocations for which the difference ranges from 0.4 to 0 are treated as equally probable.

The summary results for lexical items are given in Figure 7.5 (the complete results are shown in Figure B6 in Appendix B). The first graph shows the proportion of shared collocates between <many> and other less frequent positive quantifiers. The blue bar indicates the cases when stronger collocations are made with <many>, the red bar represents the cases when shared collocations are equally probable, and the green bar tells when a less frequent positive quantifier forms stronger collocations.

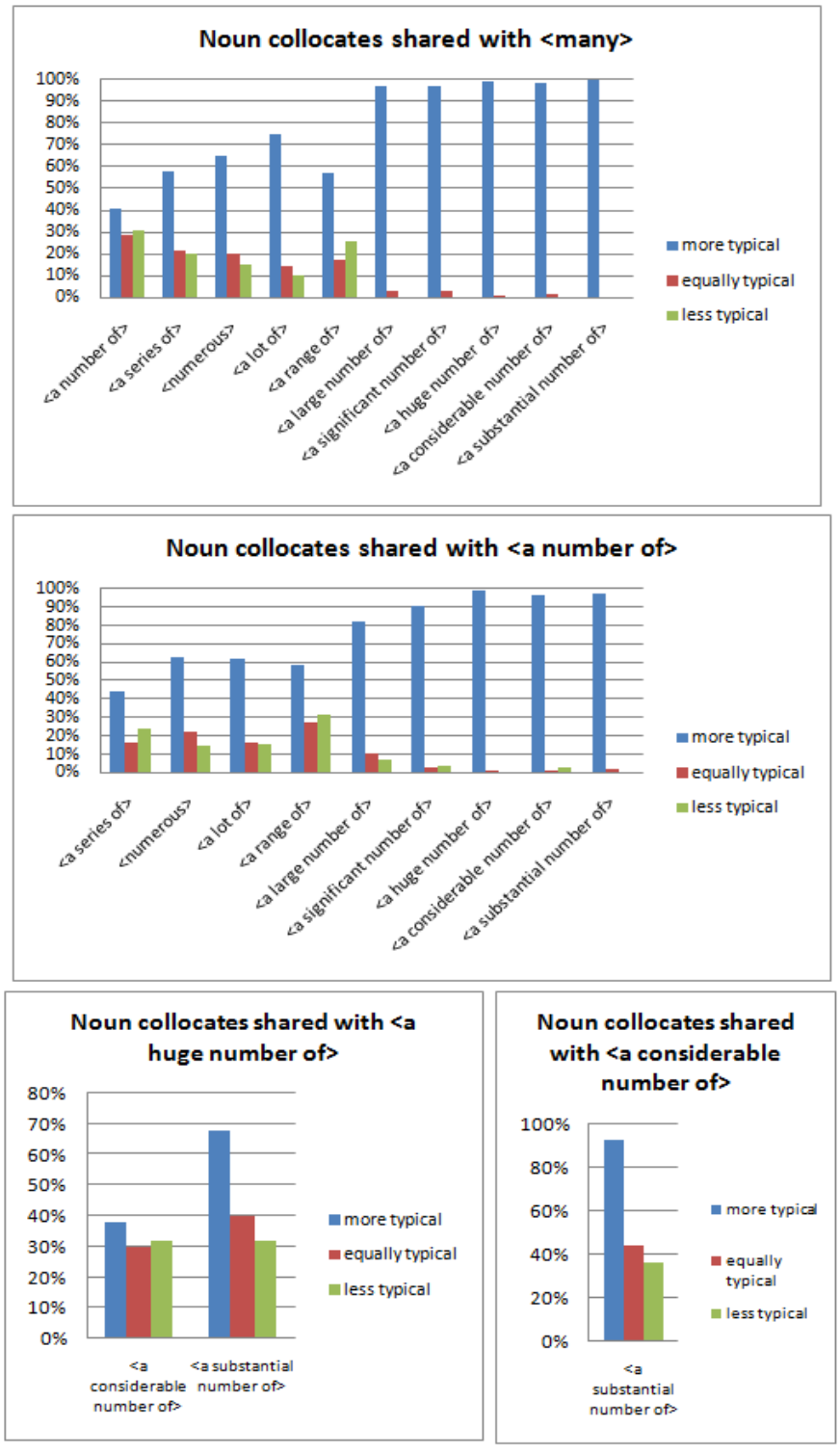


Figure 7.5: Association strength values for the collocations made up of positive quantifiers and shared collective nouns from the TLD {MANY COLLECTIVES}

In all but one case positive quantifiers that occur with a higher number of collocates tend to occur in stronger collocations. To give an example, in between 45% and 99% of instances the item <a number of> is used in stronger collocations than other less frequent positive quantifiers. Thus, out of 1500 collocates that occur both with this lexical item and <a series of> 744 combinations have stronger collocations with the former item, 402 with the latter and 354 are about equally typical with both items. An exception to this general tendency is <a range of> which participates in a higher number of stronger collocations than <a lot of> although the latter collocates with more collective nouns. Such behaviour of <a range of> is not completely surprising given its frequency. It was said above that the difference in the number of stronger collocations in previous analyses were related both to the number of total collocates and the frequency differences and the two variables were not perfectly correlated. The discrepancy was possible due to any of the two variables. In the present example, <a range of> occurs with fewer noun collocates than <a lot of> but is more frequent than it.

One can observe that blue bars in Figure B6 increase as we move from left to right. This suggests that the number of stronger collocates tends to rise as the difference in frequency and the number of associated collocates become higher. At the same time, the size of other two bars tends to decrease as this difference drops. This is observed on the third and fourth graph which demonstrates that stronger or equally likely collocations that contain <a series of>, <numerous>, <a lot of> and <ranger> are close in number. These items have also similar behaviour with regard to their frequency and co-occurrence with collective nouns. On the other hand, the number of stronger collocates increase dramatically when we compare these and the five low frequency positive quantifiers. This is in accord with the difference related to the variables discussed.

The intralinguistic analysis conducted in this section delivers positive results. The trends observed are very similar to those that characterise the distribution of lexical items studied in Chapter 5. The distribution of lexical items is not random. The frequency of lexical items correlates with the number of collocates with which they coincide. These and other findings point out the features that can be used to distinguish between lexical items from the current domain. The above study was less detailed than that in Chapter 5 because other contexts were

not considered. It can be expected that further explorations of these contexts would bring up to the surface other differences.

7.2.4 Unique collocates

Thus far only shared noun collocates were observed. In this section specific collocates will be examined because they can also serve as possible distinguishers between the positive quantifiers from the TLD {MANY COLLECTIVES}.

There is a strong relation between the frequency of lexical items and the number of unique collocates (r=0.91). The largest number of unique collocates is observed with <many> and none of these collocates occur with the three least frequent positive quantifiers. The number of unique collocates is also proportionately larger for items with higher frequency. Out of all nouns which occur in the R1 position with <many> 20% collocate only with this item. Of all those collective nouns that occur with <a number of> or <a series of> 15% are unique collocates, and for <numerous> and <a lot of> this figure is 6%, whereas for <a large number of> it is 2%. One common feature for all positive quantifies is that the majority of these collocates have very low frequency and low logDice values. Every second or third noun occurs only twice. Nevertheless, some relatively more frequent nouns are observed with several positive quantifiers. Those that co-occur with <many> can be classified into three general groups. At the top of the list are the numerals <hundreds>, <millions>, <tens> and <billions> that occur between 27 and 370 times per million words. Thereafter follow the items <tons>, <inches>, <kilometers> and <hectares> which denote measures. The next large group contains nouns that refer to individuals belonging to a social group. Here we can distinguish between those that refer to nationalities such as <Kurds>, <Poles>, <Chinese>, <Irishmen> or <Bengalis>, those that refer to the members of religion groups such as <Evangelists>, <atheists>, <heretics>, <imams>, <Puritans> or <goddesses> and those that refer to political groups such as <liberals>, <anarchists>, <libertarians>, <federalists> or <Democrats>. The unique collective nouns that occur with higher frequency with other positive quantifiers are too heterogeneous

to be considered a group. Thus, <a number of> occurs with <corpora>, <dichotomies>, <assessors>, <enclaves> and <exclusives>; <a series of> occurs with <snapshots>, <meditations>, <sluices>, <manoeuvres>, <zig zags>, <grids>; <a lot of> occurs with <guts>, <freckles>, <rappers>, <fertilizers> and <emulators>; and <numerous> occurs with <fêtes> and <retainers>.

The above results indicate that unique collocates can serve as distinguishers but since they are not very frequent they do not represent the typical behaviour of lexical items.

7.2.5 Lexical items from the TLSd {MANY PROBLEMS}

In this section the distribution of lexical items from the TLSd {MANY PROBLEMS} identified in the previous chapter will be examined. The aim of the analysis is to test the feasibility of the distributional method in the study of a sub-domain of the TLD {MANY COLLECTIVES}.

In Chapter 6 in Table 6.4 we saw that not all positive quantifiers that belong to the general TLD {MANY COLLECTIVES} occur in the sub-domain {MANY PROBLEMS}. For example, none of the low frequency items met the predefined criteria. The members of this English sub-domain consist of the following positive quantifiers: <many>, <numerous>, <a number of>, <a range of>, <a series of> and <a lot of> and the following nouns: <problem>, <difficulty>, <issue>, <subject>, <point> and <matter>. These nouns are coded as PROBLEMS according to the most frequent item.

Figure 7.6 displays the distribution of these lexical items according to the values of association strength. Unlike in the TLD {MANY COLLECTIVES} the most frequent positive quantifier is <a number of>. The second difference in the distribution of items in the domain and sub-domain is that <a series of> occurs here proportionately less often than before. In TLD {MANY COLLECTIVES} it was more frequent than <a lot of> and <numerous> whereas in the present sub-domain the reverse is the case.

The most typical collocation consists of <a number of> and the PROBLEM nouns and the least likely combinations are formed <a series of>. Only the first four lexical items have the

logDice value above average. <a number of> occurs between three and 35 times as typically as other positive quantifiers in the present context. The next most frequent item is <many> which occurs in collocations 18 times stronger than those made with other positive quantifiers. As the t-test indicates, differences between the values of association strength are statistically significant (p-value=0.00018).

There are also significant differences in the frequency of lexical items. The p-value for the frequency difference is less than 2.2e-16 and chi value is 32753.02. In addition, frequency positively correlates with the logDice values (r=0.78). There are two exceptions here. The first exception is <many> that occurs with higher frequency than <a number of> and <a range of> in the present context although it forms weaker collocations. The second exception is <a series of> which is slightly more frequent than <a large number of> but has a low logDice value when it collocates with the PEOPLE nouns than the latter term.

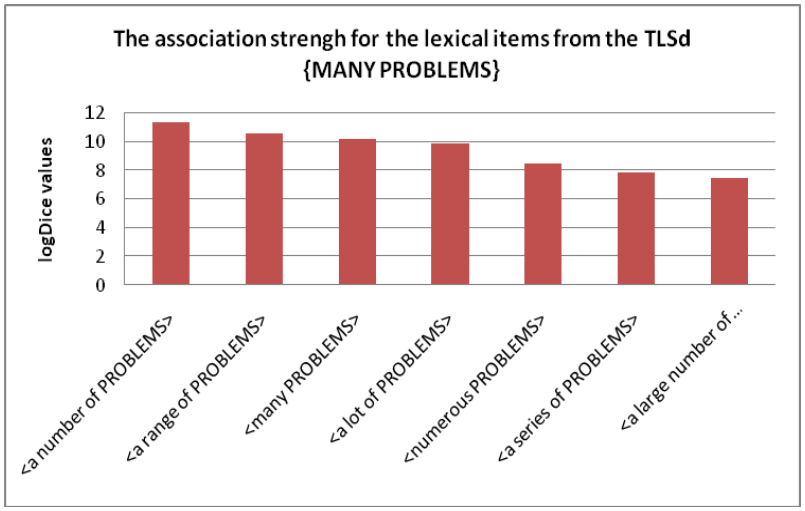


Figure 7.6: Typicality of collocations made up of lexical items from the TLSd {MANY PROBLEMS}

The present analysis shows that there are differences in the distribution of lexical items that belong to a domain and sub-domain. This suggests that the results obtained in an analysis of one context should not be overestimated. This is similar to the above observation of the nouns from the class PEOPLE. We saw there that combinations of positive quantifiers varied in their typicality with specific nouns from this set. However, it is important not to confuse the two analyses. The class of nouns from the former analysis is an analytical rather than lexical

category. It served to summarise all noun collocates in the form of several descriptive groups. The lexical items from the previous analysis are not necessarily mutually substitutable in a translation context. For example, the nouns <student>, <teacher> and <visitor> belong to the type of noun coded as PEOPLE but they are not very likely to correspond to the same lexical item in German. On the other hand, the nouns <problem>, <issue> and <question> are mutually substitutable when they occur with positive quantifiers as translation correspondences to the same lexical item from German.

7.2.6 Correspondence potential

In this section an interlinguistic study of the lexical items from TLD {MANY COLLECTIVES} and the TLSd {MANY PROBLEMS}. Following the results of the analysis of the TLD {CAUSE PROBLEM} and {PROBLEM BEREITEN} we can expect to find differences between correspondence potential. In more particular, we can expect that lexical items differ in terms of the number of correspondence relations and the percentage with which they correspond to the items from German. In addition, we can expect that the lexical items associated with more correspondence relations will have a higher degree of usage and that correspondence potential roughly correlates with the general frequency of the lexical items. Finally, we can expect that the formally similar items from two languages or items with a similar etymological origin strongly correspond to each other. I will commence by examining the distribution of lexical items from the TLD {MANY COLLECTIVES}.

Table 7.2 summarises the results for these items. The fourth column confirms our expectation that the lexical items differ in terms of the number of correspondence relations. The lexical item <many COLLECTIVES> corresponds to ten German lexical items and other items from the table establish between two and eight relations. The p-value for the difference between these values is 0.04 which is slightly below the predefined threshold. It follows that the difference is not by chance. The first four items have values above average.

Lexical items	Correspondence degree:total	Correspondence degree:average	Number of correspondence relations	CP
<many COLLECTIVES>	28	2.8	10	12.8
<a number of COLLECTIVES>	14	1.8	8	9.8
<a large number of COLLECTIVES>	12	1.7	7	8.7
<a significant number of COLLECTIVES>	3	0.4	7	7.4
<a lot of COLLECTIVES>	5	1.7	3	4.7
<a considerable number of COLLECTIVES>	5	1.7	3	4.7
<a substantial number of COLLECTIVES>	5	1.7	3	4.7
<a series of COLLECTIVES>	4	2	2	4
<a range of COLLECTIVES>	3	1.5	2	3.5
<numerous COLLECTIVES>	3	1.5	2	3.5
<a huge number of COLLECTIVES>	2	1	2	3

Table 7.2: Distribution of translation correspondences from the TLD {MANY COLLECTIVES}

As one can see in the second and third column, the percentages with which the items are used are also not identical. By comparing the sum of these values displayed in the second column we come to the conclusion that these differences are statistically significant: p-value-1.434e-12. In addition, a comparison of the second and fourth columns indicates a strong correlation between the two variables: r=0.88.

Apart from <a huge number of COLLECTIVES> all other items are found at least once with the correspondence degree higher than 10%. When it happens, these items are among the most preferred correspondences. Thus, a higher correspondence degree with <numerous COLLECTIVES> is observed when it corresponds to <zahlreiche KOLLEKTIVA>. The same holds true for <a series of COLLECTIVES> in relation to <eine Reihe DET|GEN KOLLEKTIVA> or for <a range of COLLECTIVES> in relation to <eine Vielzahl DET|GEN KOLLEKTIVA>.

The results of correspondence potential are given in the fifth column. The lexical item with the highest correspondence potential is <many COLLECTIVES> which also corresponds to the largest number of German items from the TLD {VIELE KOLLEKTIVA} and is used with the highest percentage. In seven out of ten cases it is used more than 10% of the time.

The correlation coefficient for the relation between correspondence potential and the frequency of lexical items in the reference corpus is r=0.55. It indicates that there is a relation

between these two variables but that it is not very strong. The following two lexical units adhere to this tendency more than other items: <many COLLECTIVES> and <a number of COLLECTIVES>. As for the others, following their frequency one would expect that <a range of COLLECTIVES>, <a series of COLLECTIVES> and <numerous COLLECTIVES> have higher correspondence potential and that <a significant number of COLLECTIVES>, <a considerable number of COLLECTIVES> and <a substantial number of COLLECTIVES> have lower correspondence potential.

It seems that the explanations previously suggested with regard to the formal similarities fit well here. <a significant number of COLLECTIVES>, <a considerable number of COLLECTIVES> and <a substantial number of COLLECTIVES> occur with higher percentage when they correspond to the German lexical items with a similar structure: <ein erheblich Zahl|Anzahl ART|APPR KOLLEKTIVA>, <ein beträchtlich Zahl|Anzahl ART|APPR KOLLEKTIVA> and <ein beachtlich Zahl|Anzahl ART|APPR KOLLEKTIVA>. Similarly, there are formal similarities and strong correspondences between the following items: <a series of COLLECTIVES> and <eine Reihe ART|APPR KOLLEKTIVA>; <numerous COLLECTIVES> and <zahlreiche KOLLEKTIVA>, <a range of COLLECTIVES> and <eine Vielzahl ART|APPR KOLLEKTIVA>.

Now I will report the results (Table 7.3) of the distribution of lexical items from the TLSd {MANY PROBLEMS}. They differ in terms of the number of established correspondence relations. It ranges from two to seven. However, the differences are not so high as to be statistically significant (p-value=0.32). The lexical items differ also in terms of the correspondence degree and these differences are statistically significant (p-value=8.873e-08). In addition, the data from the second and fourth column from Table 7.3 correlate positively (r=0.89). It means that the percentage with which a lexical item from the TLSd {MANY PROBLEMS} is used as a translation correspondence corresponds to the number of correspondence relations associated with it and vice versa.

Lexical items	Correspondence degree:total	Correspondence degree:average	Number of correspondence relations	CP
<a number of PROBLEMS>	18	3	6	9
<many PROBLEMS>	19	4.8	4	8.7
<a range of PROBLEMS>	6	1.2	5	6.2
<a lot of PROBLEMS>	4	1.3	3	4.3
<a large number of PROBLEMS>	4	2	2	4
<a series of PROBLEMS>	3	1.5	2	3.5
<numerous PROBLEMS>	3	1.5	2	3.5

Table 7.3: Distribution of translation correspondences from the TLD {MANY PROBLEMS}

The lexical item with the highest correspondence potential is <a number of PROBLEMS> which is followed by <many PROBLEMS>. These two items correspond to the largest number of German lexical items. They are also used with the highest percentage as translation correspondences. Each of other items is used at least once with a percentage higher than 10%, that is, as the first or second most preferred choice.

The distribution of these lexical items used as correspondences is very similar to their use in the reference corpus. The correlation between these two usages is 0.93. For example, <a number of PROBLEMS> is both the most frequent item and has the highest correspondence potential. One exception is <a large number of PROBLEMS> which has greater correspondence potential than would be predicted by its frequency.

The above results have therefore demonstrated the feasibility of the distributional method in an intralinguistic study. Lexical items differ in terms of the number of correspondence relations and the frequency of their selection. As a result, they have different correspondence potential. The values of correspondence potential correlate with the frequency of items in the reference corpus. This is true for lexical items from the TLD {MANY COLLECTIVES} and to a lesser extent to that from the TLSd {MANY PROBLEMS}.

7.3 TLD {VIELE KOLLEKTIVA} and TLSd {VIELE PROBLEME}

In this section the distribution of German lexical items from the TLD {VIELE KOLLEKTIVA} and TLSd {VIELE PROBLEME} will be considered. In 7.3.2 I will start with an intralinguistic analysis of the lexical items from the general German translation domain. Here I will focus on differences between the frequency of lexical items and the number of noun collocates with which German positive quantifiers occur. I will also consider here the distribution of positive quantifiers in relation to the local grammar classes into which German nouns can be classified. Section 7.3.3 will deal with the distribution of shared collocates and in 7.3.4 will be concerned with unique collocates. In 7.3.5 an intralinguistic analysis of the lexical items from the TLSd {VIELE PROBLEME} will be carried out. The concluding section will examine the distribution of lexical items from the given domain and sub-domain from an interlinguistic point of view. As was the case in 7.2, the purpose of the study is to check if an application of the distributional method on new data delivers the same type of results as previous studies.

7.3.1 Frequency and the number of collocates

In Chapter 6 the following lexical items were identified in my parallel corpus as members of the TLD {VIELE KOLLEKTIVA}: <viele KOLLEKTIVA>, <mehrere KOLLEKTIVA>, <zahlreiche KOLLEKTIVA>, <eine Reihe ART|APPR KOLLEKTIVA>, <eine Vielzahl ART|APPR KOLLEKTIVA>, <eine große Zahl|Anzahl ART|APPR KOLLEKTIVA>, <eine Anzahl|Zahl ART|APPR KOLLEKTIVA>, <eine erhebliche Zahl|Anzahl ART|APPR KOLLEKTIVA>, <eine beträchtliche Zahl|Anzahl ART|APPR KOLLEKTIVA> and <eine beachtliche Zahl|Anzahl ART|APPR KOLLEKTIVA>. These lexical items consist of two identical grammatical elements as English lexical items: a quantifier and a collective noun. The first element is realised grammatically either as an adjective or, more often, as a multi-word expression. The second element consists of plural or mass nouns. Following the label used for English noun collocates these nouns will be coded in German as KOLLEKTIVA.

German lexical units have a more complex structure than corresponding English constructions. The first reason for this is that German has richer inflectional morphology than English. For example, <viel> can be used with collective nouns in the following three forms *viele, vieler* and *vielen*. In addition to these morphological features, German items have a higher number of forms because <number> has two German cognates (<Zahl> or <Anzahl>) and the English expressions with <of> can be translated either as <von> or as a determiner used in genitive. The latter form will be coded as ART and APPR following the TreeTagger tag set. All these different realisations are potential sources of differences between lexical items. In order to explore all these differences it would be necessary to take into account a broader textual context or to examine in detail occurrences of expressions with <von> or determiners. Because the present section is only concerned with major tendencies of the distribution of positive quantifiers and collective nouns and due to space restrictions these particular realisations will be ignored. All lexical items will be, therefore, considered in their lemma form; the determiners and preposition that occur with the multi-word units will be treated as one item and so will be the nouns <Zahl> and <Anzahl>.

Following previous analyses we expect that positive quantifiers differ in terms of their frequency and the number of noun collocates. Table 7.4 below contains the figures for the two variables according to the reference corpus deWaC. It is clear at first sight that the data are consistent with our expectations. Observing the frequency variable we can see that the most frequent item is between four and five thousand times as frequent as less frequent lexical items. Similarly, it has between two and four hundred times more collocates than other lexical items. The chi-square test confirms that these differences are statistically significant: χ^2=3079837; p-value < 2.2e-16. Similarly, the chi-square value for the second variable (the number of noun collocates) is χ^2=52078.71 and the p-value is again lower than 2.2e-16. In also seems that the distribution of figures for the two variables is parallel. The correlation test confirms that we deal here with very strong relation: r=0.88. It follows that the more frequent a quantifier is, the larger number of collective noun collocates it has. An exception to this general tendency is <mehrere> which occurs more frequently than <zahlreiche> but has fewer noun collocates.

Lexical items	Frequency	The number of plural nouns
<viele>	548369	12771
<mehrere>	127757	8126
<zahlreiche>	116758	8431
<eine Reihe ART|APPR>	26794	2860
<eine Vielzahl ART|APPR>	15098	1714
<eine groß Zahl|Anzahl ART|APPR>	8205	1335
<eine Anzahl|Zahl ART|APPR>	1443	299
<eine erheblich Zahl|Anzahl ART|APPR>	340	66
<eine beträchtlich Zahl|Anzahl ART|APPR>	193	57
<eine beachtlich Zahl|Anzahl ART|APPR>	105	28

Table 7.4: The number of collocates and frequency of positive quantifiers from the TLD <VIELE KOLLEKTIVA>

The aforementioned parallelism between figures for the two variables is visually displayed in the dendogram graphs in Figure 7.7 below.

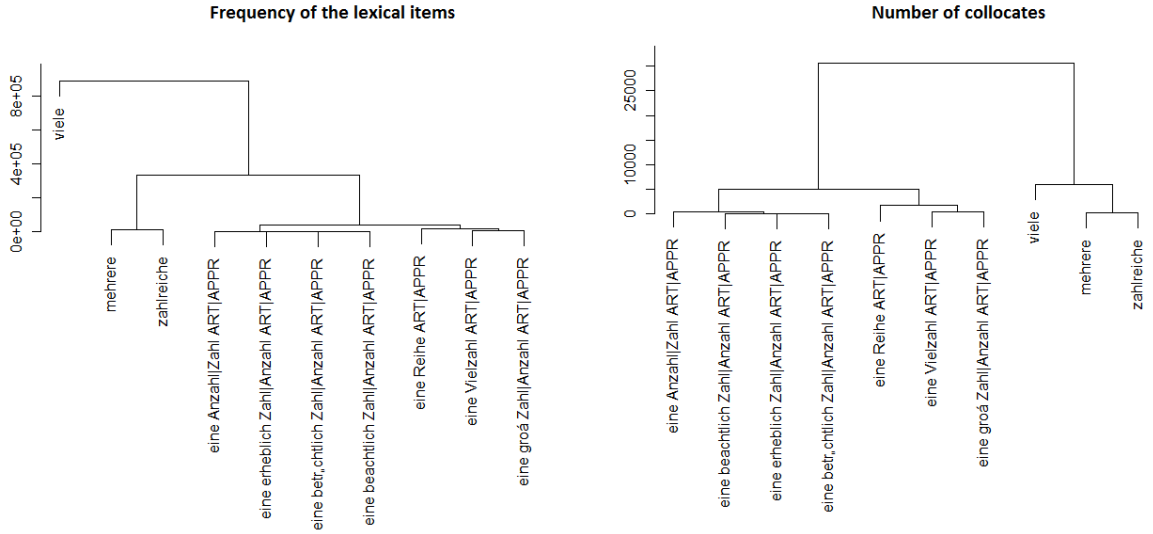


Figure 7.7: Clustering of positive quantifiers from the TLD <VIELE KOLLEKTIVA> in relation to their frequency and occurrence with noun collocates

The graphs indicate the clustering of lexical items as a result of their similar numerical values. The clustering on two graphs is very similar. Thus, one cluster consists of the four least frequent lexical. The difference in frequency and the number of collocates between these items is smaller than between them and other lexical items. Similarly, <eine Reihe ART|APPR

KOLLEKTIVA>, <eine Vielzahl ART|APPR KOLLEKTIVA> and <eine groß Zahl|Anzahl ART|APPR KOLLEKTIVA> have more in common with each other than with other positive quantifiers. The same is also true for <mehrere KOLLEKTIVA> and <zahlreiche KOLLEKTIVA> in relation to other items. The graphs indicate that the number of collocates for <viele KOLLEKTIVA> does not increase at the same pace as its frequency in relation to other items. The left-hand graph shows that it is detached from the next two most similar items, whereas in graph on the right they are clustered together. This is because <viele KOLLEKTIVA> is more than five times as frequent as the next two items but has one and a half times as many noun collocates as these items.

These findings support previous results and corroborate the suitability of the distributional model. The findings confirm that the distribution of lexical items in a lexical domain or sub-domain can help to distinguish between them. Before passing to a more detailed exploration of the observed patterns an attempt will be made to put collective nouns into subcategories. The categorisation will show whether the distribution of lexical items has to do with specific types of collective nouns.

7.3.2 Classification of KOLLEKTIVA

As was the case with English lexical items only the first 5% of noun collocates will be taken into account in this section. They present the most frequent collocates and make up between 40% and 63% of the total frequency of collective nouns. In terms of meaning and function these nouns are similar to above English nouns. In general, collective nouns in the present context denote living beings, abstract entities, physical objects, spatial and time referents. Consequently, these collective nouns will be grouped into the following classes: MENSCHEN, LAGE, PROBLEM, ZEIT, VERÖFFENTLICHING, METRIK, VERANSTALTUNG and MÖGLICHKEIT. The analysis will focus only on the collocations that have the highest association strength with one of positive quantifiers.

The living beings are mainly human beings and they are usually referred to in terms of belonging to a general or specific group. The most representative members of the former type are <Mensch>, <Leute>, <Person>, <Kind> and <Frau>. The most frequent nouns of the latter

type are <Patient>, <Besucher>, <Mitglied>, <Eltern> and <Freund>. Both types will be coded with the term MENSCHEN.

One set of items contains the nouns that speakers use to talk about regions and locations and among the most frequent are <Bereich>, <Gebiet>, <Stelle>, <Region>, <Ort>, <Stadt>, <Staat>, <Schule> and <Universität>. The former three nouns can be used both in a metaphorical way, such as when speakers talk about different areas of law, and in literal terms, when speakers refer to physical locations. For our purposes this distinction is not of much importance and will be ignored. This set of nouns will be marked as LAGE.

A set of nouns labelled as PROBLEM denote problematic abstract matters. The most frequent collective nouns from this set are <Problem>, <Frage>, <Faktor>, <Ding>, <Punkt> or <Aspekt>.

The nouns <Jahr>, <Monat>, <Woche> or <Tag> denote time relations and will be simply coded as ZEIT.

The nouns that we use to talk about various types of publications are also among the frequent collocates of positive quantifiers. Some of the most typical are <Buch>, <Artikel>, <Publikation>, <Studie> and <Veröffentlichung>. Such nouns will be called the VERÖFFENTLICHUNG nouns.

There is a small set of words referring to alternatives and options (e.g. <Möglichkeit>, <Weg>, <Funktion>, <Bedingung> and <Form>) and they will be coded as MÖGLICHKEIT.

One group of collective nouns that has to do with various types of events (e.g. <Konzert>, <Veranstaltung>, <Vortrag>, <Gespräch>) will be named VERANSTALTUNG.

Finally, some positive quantifiers collocate with the collective nouns that denote quantity such as <Meter>, <Kilometer>, <Tausend> and <Mal>. They will be labelled as METRIK.

As was the case in the study of English collective nouns, these classes are provisional and far from being complete. But, they will serve the purpose of comparing the distribution of positive quantifiers in relation to different classes of collective nouns. The results reported below are based on a comparative analysis of logDice values of collocations formed with positive quantifiers and the collective nouns from the above class.

The positive quantifier <viele> forms strongest collocations with the vast majority of the MENSCHEN nouns and the LAGE nouns. Out of seven most frequent PROBLEM nouns three collocate most typically with this positive quantifier and four with <eine Reihe ART|APPR>. The latter quantifier is also most typically associated with almost all MÖGLICHKEIT items. Apart from the noun <Jahr> that typically occurs with <viele> other four nouns from the ZEIT class form the strongest collocations with <mehrere>. This positive quantifier also forms the strongest collocation with the METRIK nouns. <zahlreiche> is the most typical collocate of the nouns that belong to the classes VERÖFFENTLICHING and VERANSTALTUNG. No collocations that have the highest association strength can be observed with other positive quantifiers.

The above description indicates that the collocations that consist of positive qualifiers and collective nouns differ in terms of the type of nouns. There is a preference for the positive qualifiers from the present domain to select collective nouns from particular classes. These findings can be used to distinguish between the lexical items in question. So far only the strongest collocations were observed. For a more detailed description of the behaviour of lexical items one should examine the distribution of positive quantifiers in relation to all types of nouns. Space restriction does not allow to provide such a comprehensive description but I will here illustrate this kind of analysis by examining the distribution of the nouns from the class MENSCHEN.

Figure 7.8 displays the collocations formed with the collective nouns from the class MENSCHEN and positive quantifiers from the present domain. These collocations are ordered according to their logDice values. The following differences can be observed here. In relation to the distribution in the whole domain the positive quantifiers <zahlreiche>, <eine Reihe ART|APPR> occur in the current context proportionately less often and less typically. On the other hand, <eine große Zahl|Anzahl ART|APPR> and <eine Vielzahl ART|APPR> are more typical. Like in 7.2.2 we can, therefore, conclude that the collocation strength observed at the general level is not automatically mirrored in a restricted set of collocations. The typicality of co-occurrence depends obviously on specific types of collocates.

The t-test indicates that differences between the values of association strength in the current data are statistically significant: p-value = 5.305e-05. These values are above average

only for the first four lexical items. The most typical collocations are formed with <viele> and they are between two and 32 times as typical as those formed with other positive quantifiers. The difference between other items ranges between two and 16 times.

There is also a very strong correlation between the values of association strength and the frequency of lexical units (r=0.94). Only the behaviour of <eine Vielezahl ART|APPR MENSCHEN> deviates from this correlation as it occurs in stronger collocations than would be expected from its frequency.

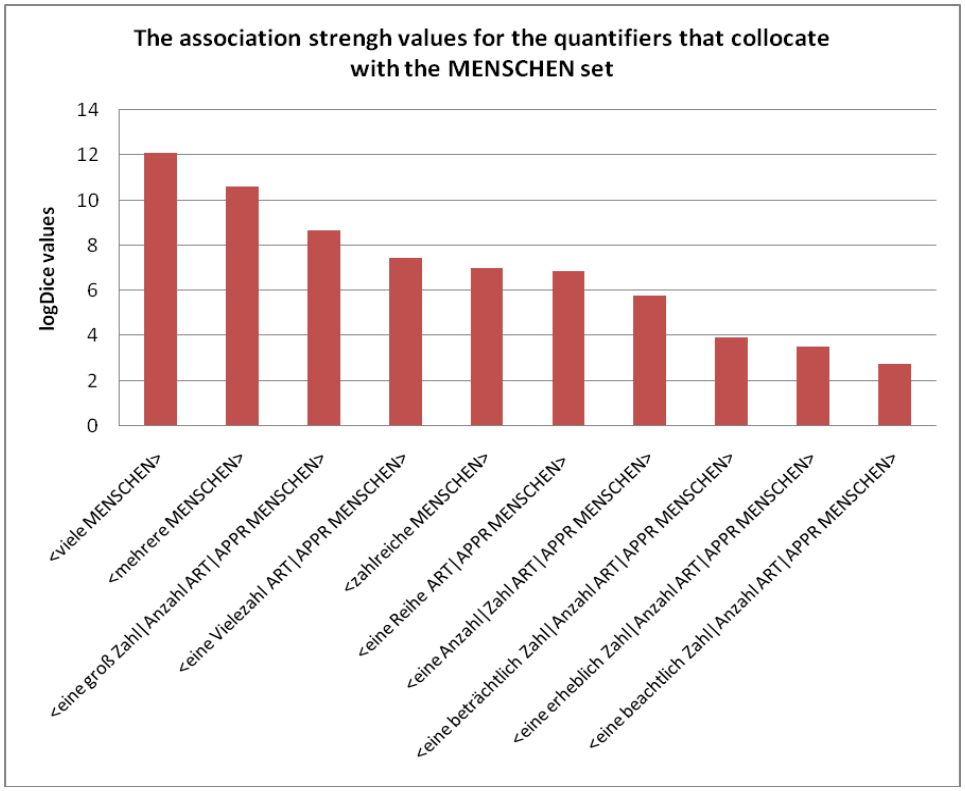


Figure 7.8: Collocations made up of positive quantifiers and the nouns from the semantic set MENSCHEN

The above results suggest that the German positive quantifiers do not occur with equal probability with collective nouns. The preference for co-occurrence is related to the type of nouns. Although above only a few classes of nouns were examined it seems that the number of classes is related to the frequency of positive quantifiers. Thus, the most frequent positive quantifier <viele> participates in the strongest collocational associations with four noun classes. The two next most frequent items are <mehrere> and <zahlreiche> and they form the most typical collocations with two noun classes and the less frequent <eine Reihe ART|APPR> co-

occurs most typically only with one group of collective nouns. Even less frequent items do not form the strongest collocations with any of the noun classes. The analysis the expressions formed with the nouns from the class MENSCHEN indicates that the distributional model can provide an accurate description of the co-occurrence of positive quantifiers and nouns from specific groups. Such a description provides purely distributional distinguishing features for the studied items.

7.3.3 Shared collocates

This section examines the distribution of the quantifiers with the shared noun collocates in greater depth. This distribution will be investigated in terms of the tendencies that have been observed in earlier analyses. Accordingly, in the analysis below it can be expected that more frequent positive quantifiers occur with the majority of frequent collocates that occur with less frequent quantifiers. In addition, we can expect that both the number of shared collocates and the difference between the frequency of positive quantifiers increase. Finally, we can expect that the number of stronger collocations will be proportional to the frequency of lexical items.

We can observe a very strong relationship between the frequency of lexical items and their occurrence with shared collocates. The more frequent a lexical item is, the higher degree of overlap it will have. In other words, if two lexical items have different frequency values the one with a higher frequency will have more collocates in common with other items than the one with a lower frequency. This is what we had in earlier analyses as well.

First, the tendency regarding the distribution of shared collocates will be studied. Here, both the degree of overlap between single quantifiers and the sum of frequency of these shared collocates will be examined. The former analysis will show to what extent is degree of overlap related to the general frequency of positive quantifiers, whereas the latter will indicate if shared collocates occur with high frequency. The data for the first two most frequent and the two least frequent lexical items are summarised below in Figure 7.9 (the results for all lexical items are given in Figure B7 in Appendix B).

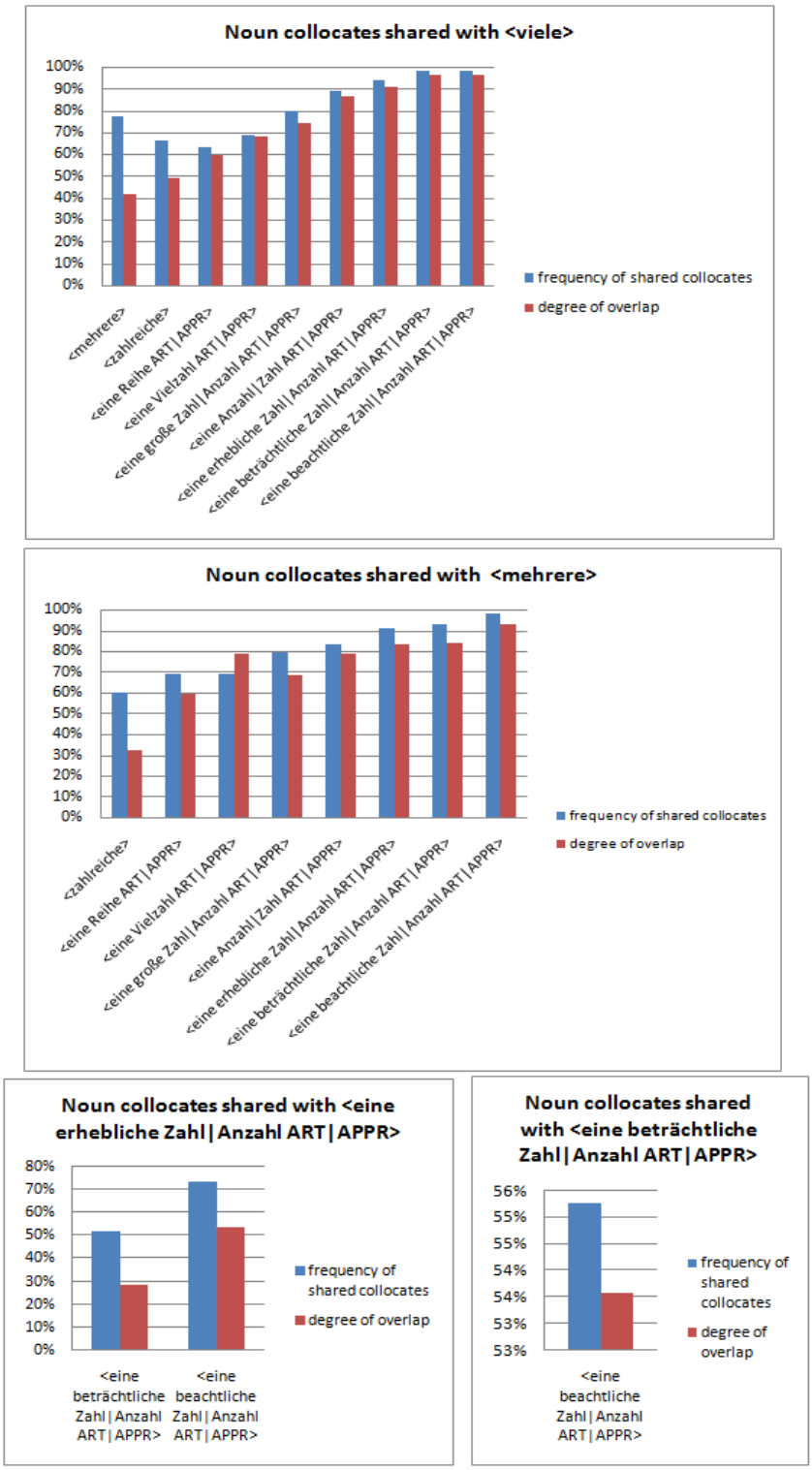


Figure 7.9: Frequency and degree of overlap for noun collocates that occur with positive quantifiers from the TLD {VIELE KOLLEKTIVA}

On the first graph, it can be observed that the degree of overlap between the quantifier <viele> and other items tends to rise from about 50% to almost 100%. Since the lexical items are ordered according to their general frequency one can conclude that this increase follows the rate with which the general frequency of positive quantifiers changes. Other graphs (Figure B7 in Appendix B) show that these two tendencies with a few exceptions remain constant in the current data. To give one more example, <zahlreiche> is about four times as frequent as <eine Reihe ART|APPR> and occurs with 50% of the nouns that occur with the latter positive quantifier. It is about seven times as frequent as <eine Vielzahl ART|APPR> and occurs with 56% of its collocates. Finally, it occurs about 1000 times more often than <eine beachtliche Zahl|Anzahl ART|APPR>. The most significant exception to this central tendency can be observed in relation to <eine beträchtliche Zahl|Anzahl ART|APPR> and three more frequent positive quantifiers (<eine Reihe ART|APPR>, <eine Vielzahl ART|APPR> and <eine große Zahl|Anzahl ART|APPR>). Here, degree of overlap is lower than one would expect from the general frequency of these lexical items.

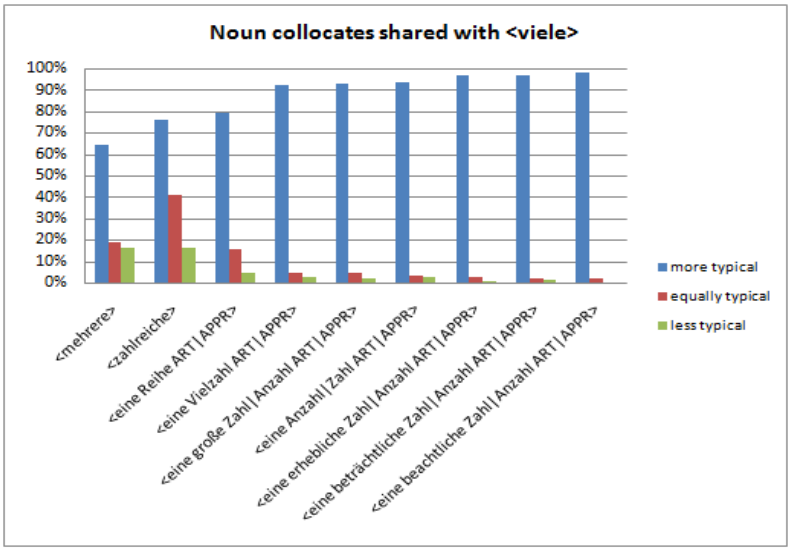
The tendency that degree of overlap decreases as the frequency of items becomes more similar is apparent on the above graphs. We can observe how the height of the red bars becomes smaller as the frequency difference between lexical items becomes closer. Thus, <viele> is 67 times as frequent as <eine große Zahl|Anzahl ART|APPR> and occurs with more than 70% of noun collocates found with this item. On the other hand, <eine Vielzahl ART|APPR> is less than twice as frequent as the latter item and occurs with 40% of its collocates. This tendency is also very clear for the three least frequent items that occur with lower frequency.

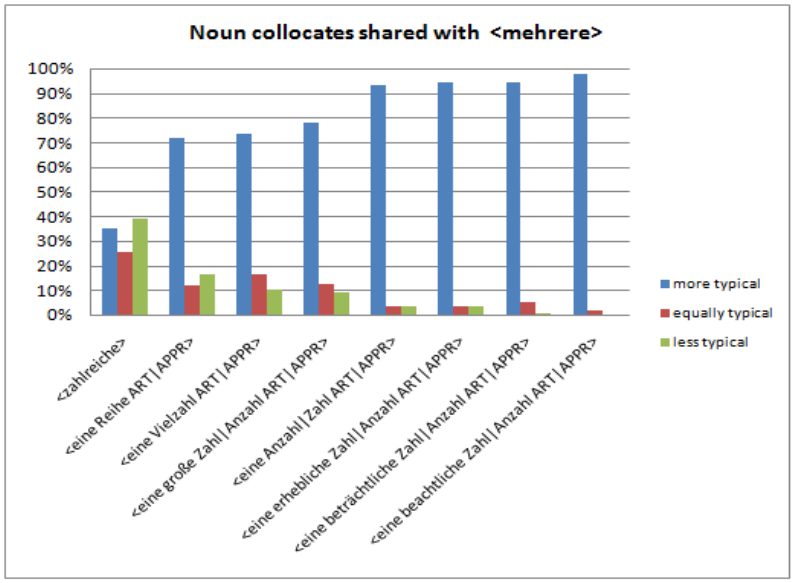
The blue bars indicate the sum of the frequency of shared collocates. It can be noticed that these values are very high in the current data as they range between 55% and 98%. From here one can conclude that common collocates occur with very high frequency. For example, although <viele> occurs with less than half of the collocates of the next two most frequent positive quantifiers, shared collocates in both cases make up more than 60% of the total frequency of collective noun collocates. It follows that non-shared items occur with lower frequency. Here are some examples. 222 out of 4724 noun collocates that co-occur with <mehrere> but not with <viele> occur between ten and 150 times. These nouns make up less

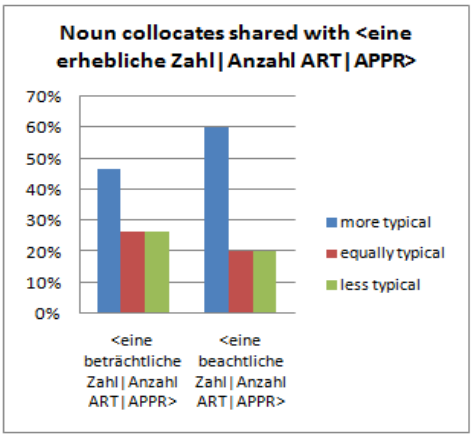
than 5% of total frequency of all nouns that combine with <mehrere>. On the other hand, out of 3402 nouns that are found with both lexical items 1127 occur between 10 and 7038 times and together make up 71% of the total frequency of the nouns that occur with <mehrere>. Similar results are observed with other items. This is even true for the items that occur with similar frequency and therefore have fewer collocates in common. For example, <eine erhebliche Zahl|Anzahl ART|APPR> occurs with 16 out of 56 nouns that collocate with <eine beträchtliche Zahl|Anzahl ART|APPR>. Although these 16 items amount for only one third of all collocates the sum of their frequency is 52% of all noun collocates. Among the other 40 nouns nine occur between three and nine times and 31 only twice.

These results confirm our expectations and we can conclude that here and in previous studies observed tendencies create an intrinsic property of the lexical items that belong to the same lexical domain.

Now, we can pass on the second part of the analysis concerned with how the quantity of stronger collocations relates to the frequency of lexical items and the number of collocates. The results for the same items as in Figure 7.9 are shown in Figure 7.10. The same coding system as in previous similar graphs is used to represent the values of association strength between. Figure B8 in Appendix B contains the complete results from the lexical items that we deal with here.







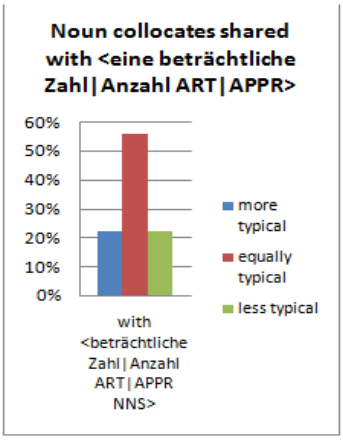


Figure 7.10: Association strength values for the collocations made up of positive quantifiers and shared collective nouns from the TLD {VIELE KOLLEKTIVA}

In the present data, stronger collocates tend to be more numerous with more frequent positive quantifiers. More frequent quantifiers also occur with a larger number of collocates. It suggests that the number of stronger collocates correlate with frequency and the total number of collocates. The exception is the lexical item <zahlreiche> which is slightly less frequent than <mehrere> but occurs in stronger collocations (Figure B8 in Appendix B).

The above graphs also show that the proportion of stronger collocations correlates with the frequency of lexical items and the total number of collocates. This proportion tends to be higher as the difference in the frequency of items increase. For example, <eine Reihe ART|APPR> is about one and a half times as frequent as <eine Vielzahl ART|APPR> when it colligates with collective nouns and more than 60% of shared collocates are more strongly associated with the former than with the latter item. On the other hand, the former positive quantifier occurs more than 200 times more often than <eine beachtliche Zahl|Anzahl ART|APPR> and nine out of ten shared collocates are more numerous with this item.

We can conclude that the results obtained correspond to the findings of earlier analyses. This is true both in relation to differences between the frequency of lexical items, the number of collocates and the correlation coefficient between these two variables and in relation to the number of shared collocates, the frequency of these collocates and the number of stronger collocations.

7.3.4 Unique collocates

In addition to the above described tendencies the lexical items from the present domain can be distinguished also by identifying their unique collocates, e.g. the collocates that occur with high frequency with only one lexical item.

The greatest number of unique collocates is observed with <viele>. About 40% of its collocates do not occur with other positive quantifiers. However, a vast majority (82%) of these collocates are very infrequent and occurs between two and ten times. The most frequent are <Emotion>, <Glied>, <Energie>, <Erkrankte>, <Feinheit>, <Gleichgesinte>, <Hausfrau>,

<Fliege|Fliegen>, <Gartenbesitzer>, <Feministin>, <Erwerbslos|Erwerbslose>, <Erzieher>, <Erfinder>, <Fernsehzuschauer>, <Ehepaar>, <Farmer>, <Heft>, <Heranwachsende> and <Halbwahrheit>.

25% of all collocates that occur with <mehrere> do not occur with other items. Similar to the previous case, most of these collocates (81%) have very low frequency. The high frequency collocates are <Stimme>, <Ton>, <Suchbegriff>, <Zeile>, <Typ> and <Verdächtige>.

The proportion of unique collocates decreases as the frequency of items drops. Thus, we find that about 38% of all collocates that occur with <zahlreich> do not occur with other positive quantifiers, whereas with <eine Reihe ART|APPR> this figure is 32% and with <ein Vielzahl ART|APPR> 28% and so on. In all these cases only a very small proportion of collective nouns occur frequently. Relatively frequent collocates that we find with <zahlreich> are <Entwurf>, <Büro>, <Einwand> and <Bürgerkrieg>, with <eine Reihe ART|APPR> we observe <Implikation>, <Kinderspiel>, <Leitlinie>, <Vorbedingung> and <Sonderbedingung>, <Handlungsoption> and with <ein Vielzahl ART|APPR> the compound nouns <Programmveranstalter>, <Druckmedium> and <Untersichungsmethode>. No significantly frequent unique collocates are observed with other positive quantifiers.

7.3.5 Lexical items from the TLSd {VIELE PROBLEME}

In the section 7.2.5 the lexical items from the TLSd {MANY PROBLEMS} were studied. Now, we will consider the German lexical units from the corresponding TLSd {VIELE PROBLEME}. There are five lexical items that constitute this sub-domain. The following positive quantifiers that occur in the general TLD {VIELE KOLLEKTIVA} do not occur here or do not meet the established criteria: <mehrere>, <eine beträchtlich Zahl|Anzahl ART|APPR>, <eine erheblich Zahl|Anzahl ART|APPR> and <eine beachtlich Zahl|Anzahl ART|APPR>. The second difference in behaviour of lexical items in this domain and sub-domain has to do with the variables frequency and the collocation strength. <eine Reihe ART|APPR> replaces <viele> as the most typical positive quantifier in the sub-domain. Also, <eine Vielezahl ART|APPR> occurs more typically with the

nouns from the set PROBLEME than <zahlreiche>, whereas the latter was strongly associated with the general class of collective nouns.

Figure 7.11 contains the information about the distribution of the lexical items from TLSd {VIELE PROBLEME}. The result of the t-test indicates that differences between the logDice values are not random (p-value = 0.0005159). The first three lexical items occur with the association strength value which is above average and the other two items are less typical. To understand better what these differences mean we can re-calculate them in terms of probability. We observe that the second and third positive quantifiers occur two times less likely with the given collective nouns than the first positive quantifier. The fourth positive quantifier is 12 times and the fifth one 38 times less likely than the first item. These are, therefore, very significant differences.

The number of items is too few to be able to measure the correspondence coefficient between frequency and logDice values.

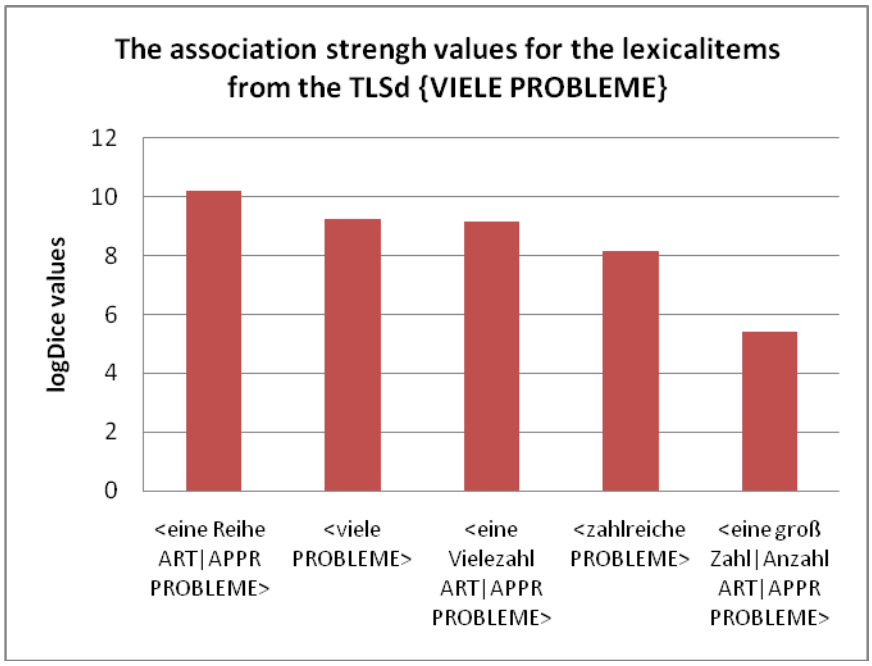


Figure 7.11: Lexical items from the TLSd {VIELE PROBLEME}

This investigation confirms that differences between lexical items at the level of a sub-domain can be found by exploring their distribution. The analysis of the distribution of positive

quantifiers at the level of the domain {VIELE KOLLEKTIVA} showed their co-occurrence with the whole set of collective nouns. The analysis conducted in the present section demonstrated that individual positive quantifiers have different preferences for the nouns belonging to the set PROBLEME. One can assume that these preferences would change in relation to other set of nouns.

7.3.6 Correspondence potential

In this section an interlinguistic investigation of the distribution of lexical items from the TLD {VIELE KOLLEKTIVA} and TLSd TLD {VIELE PROBLEME} will be carried out. This analysis will follow the procedure applied in section 7.2.6 which dealt with the same issue in relation to corresponding English lexical units. It has the same purpose as that section; to examine differences in the use of lexical items as translation correspondences. Differences will be studied in terms of correspondence potential which consists of the variable the number of correspondence relations and the variable percentage with which the lexical items from German correspond to those from English. The results will indicate the centrality of an item to the present domain, e.g. their substitutability. It is hypothesised that the higher correspondence potential a lexical item has, the higher semantic similarity it will have with other items from the same language. Following the previous results in the current data it will also be expected that:

- the lexical items with a higher number of correspondence relations are frequently used as translation correspondences and vice versa.
- there is a strong correlation between the frequency of lexical items in the reference corpus and their correspondence potential. The more frequent a lexical item is, the higher correspondence potential it has.
- formally similar and etymologically related items from the two languages tend to correspond to each other to a higher degree.

First, the lexical items from the general TLD {VIELE KOLLEKTIVA} will be investigated. The data are summarised in Table 7.5. The second and third columns show the percentage with which these lexical items are used as correspondence translations. The percentage values are converted into numbers by means of the assigned values as explained in Chapter 5. The second column shows the sum of these values and the third their average values. The number of correspondence relations is displayed in the fourth and the values of correspondence potential in the fifth column. In none of these columns are values identical for all lexical items. For example, in the second and third column we can observe that <viele KOLLEKTIVA> is used with much higher percentage as a translation correspondence than <mehrere KOLLEKTIVA> or <eine beachtliche Zahl|Anzahl ART|APPR KOLLEKTIVA>. Similarly, the former item also corresponds to a higher number of lexical items from English than the latter one. To test if these differences are statistically significant the significance test will be used. Since the values in the columns are very small instead of the chi-square test the degree of significance will be calculated by means of the t-test. The p-value for the data from the second, third, fourth and fifth columns are respectively: 0.004913, 2.098e-05, 0.0004086 and 0.0001417. All these values are lower than 0.05 and we can conclude that differences are too large to be random.

The number of correspondence relations correlates with the correspondence degree. The correlation coefficient for the second and the fourth column is 0.94. It means that lexical items with a larger number of correspondence relations tend to be used with the higher percentage.

Finally, there is a positive correlation to some extent between the frequency of lexical items as they were recorded in deWaC and their correspondence potential (r=0.64). This is especially true for the first three lexical items that occur with high frequency in the reference corpus. Two most notable exceptions are <eine große Zahl|Anzahl ART|APPR KOLLEKTIVA> that occur with higher correspondence potential than would be expected from its frequency and <mehrere KOLLEKTIVA> for which the opposite is true. It follows that the former is less central to the domain in question than one would conclude from the description of its behaviour in the reference corpus. On the other hand, <eine große Zahl|Anzahl ART|APPR KOLLEKTIVA> proves to be more central that we would expect from its frequency in the reference corpus.

Lexical items	Correspondence degree:total	Correspondence degree:average	Number of correspondence relations	CP
<viele KOLLEKTIVA>	28	3.1	9	12.1
<zahlreiche KOLLEKTIVA>	20	2.2	9	11.2
<eine Reihe ART|APPR KOLLEKTIVA>	16	2	8	10
<eine große Zahl|Anzahl ART|APPR KOLLEKTIVA>	12	2	6	8
<eine Vielzahl ART|APPR KOLLEKTIVA>	5	1.3	4	5.3
<eine erhebliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	7	2.3	3	5.3
<eine beträchtliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	7	2.3	3	5.3
<eine Anzahl|Zahl ART|APPR KOLLEKTIVA>	3	1	3	4
<mehrere KOLLEKTIVA>	2	1	2	3
<eine beachtliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	2	1	2	3

Table 7.5: Distribution of translation correspondences from the TLD {VIELE KOLLEKTIVA}

The lexical item with the highest correspondence potential is <viele KOLLEKTIVA>. Together with <zahlreiche KOLLEKTIVA> it establishes the highest number of correspondence relations. It corresponds twice to the English lexical items in more than 50% of the time and three times between 15% and 28% of the time. None of other lexical items are found with such high percentage of use. For example, <zahlreiche KOLLEKTIVA> is used once as a translation correspondence in more than 50% of the cases and three times between 11% and 22%. Other lexical units with the exception of the two items from the bottom of the table correspond at once to more than 10% of the occurrence of an English lexical item.

The assumption that formally similar or etymologically related items create stronger correspondence relations is true to some degree in the present data. Thus, <zahlreiche KOLLEKTIVA> is used as the most frequent correspondence for <numerous KOLLEKTIVA>, <eine beträchtliche Zahl|Anzahl GEN KOLLEKTIVA> for <a considerable number of KOLLEKTIVA>, and <eine Reihe ART|APPR KOLLEKTIVA> for <a series of COLLECTIVES>. There are also lexical items that do not have formally similar cognates in another language. One example is <a lot of

COLLECTIVES> which is most often translated into German as <viele KOLLEKTIVA>. But relying too strongly on formal similarities can be misleading as the following example illustrates. Although the lexical items <a number of COLLECTIVES> and <eine Anzahl|Zahl ART|APPR KOLLEKTIVA> are formally similar the latter is not the most frequent correspondence of the former.

Using the same parameters as above, the behaviour of the lexical items from the TLSd {VIELE KOLLEKTIVA} will be studied now. Table 7.6 summarises the results for these lexical items. The significance test indicates that the values in each of four columns are significantly different (the p-value for these columns is respectively 0.01355, 0.006022, 0.001401 and 0.001323). The correlation coefficient for the figures in the second and fourth columns is 0.74 which indicates positive correlation between the correspondence degree and the number of correspondence relations established with English lexical items. But, there are some exceptions as well. The distribution of <eine Reihe ART|APPR PROBLEM> deviates from this central tendency, and differences between the correspondence degree for the third, fourth and fifth item are not reflected for the variable the number of correspondence relations. The prediction of correspondence potential from the frequency of items in the reference corpus is also of limited value as indicated by the correlation coefficient of 0.53. This prediction fails for <viele PROBLEM> and <eine Vielzahl ART|APPR PROBLEM> that have lower correspondence potential than is suggested by their frequency.

Lexical items	**Correspondence degree:total**	**Correspondence degree:average**	**Number of correspondence relations**	**CP**
<eine Reihe ART|APPR PROBLEM>	18	4.5	4	8.5
<viele PROBLEM>	15	3	**5**	8
<zahlreich PROBLEM>	10	2	5	7
<eine große Zahl|Anzahl ART|APPR PROBLEM>	10	2	5	7
<eine Zahl|Anzahl ART|APPR PROBLEM>	3	1.5	2	3.5
<eine Vielzahl ART|APPR PROBLEM>	2	1	2	3

Table 7.6: Distribution of translation correspondences from the TLD {VIELE PROBLEME}

The first four lexical items in the current data have higher correspondence potential which is above the average level. The highest correspondence potential is recorded for <eine Reihe ART|APPR PROBLEM> and this is due to its very high correspondence degree when it corresponds to <a series of PROBLEMS> and <a number of PROBLEMS>. The correspondence potential of the following three items from the table does not differ significantly. All lexical items with the exception of <eine Vielzahl ART|APPR PROBLEM> are used at least one time as the first or the second most preferred option.

Since correspondence relations follow the above pattern, what was said above regarding formal similarities holds true here as well.

7.3.7 Conclusion

The present chapter was concerned with intralinguistic and interlinguistic analysis of English and German lexical items from two corresponding domains and sub-domains. The purpose of this chapter was to test the distributional model on a new set of data.

The results confirmed that in studying differences between semantically similar lexical items we can start by exploring a) their frequency and b) the number of their collocates. Further differences emerge when shared collocates of these lexical items are investigated in detail. We concluded that more frequent lexical items occurred with a significant proportion of the collocates of less frequent lexical items. We observed that the more frequent lexical units had a larger number of strong collocates than the less frequent items. Differences were also studied in terms of the distribution of lexical items as translation correspondence. Here we explored their behaviour in relation to their a) correspondence degree and b) the number of correspondence relations that they establish. The analysis showed that lexical items differ significantly in terms of both variables. In addition, we saw that the number of correspondence relations correlate positively to the percentage with which items are used as correspondences.

These are the same tendencies that were also observed in the analysis conducted in Chapter 5. These tendencies show that semantically similar lexical items can be distinguished by

exploring their distribution. As such, these tendencies confirmed the validity of the distributional model.

Chapter 8 Discussion of results and conclusion

8.1 Introduction

The present study was motivated by the following two issues:

i) How to distinguish between the lexical items from L2 that correspond to the L1 lexical items from an intralingual and interlingual perspective?

ii) How to group lexical items according to their semantic similarities?

The first question was motivated by the fact that current bilingual lexicography does not provide a reliable model for distinguishing between dictionary equivalents that correspond to the same lexical items in L1. Current bilingual dictionaries usually list translation equivalents without providing clear hints how to use them. This has to do with the approximation principle on which these dictionaries are based. This approximation principle creates the illusion that it does not make much difference which of listed equivalents will be used. The exceptions to this approach are bilingual distinctive synonym dictionaries which contain information about synonym discriminations. But these discriminations are usually methodologically and theoretically not well-founded and such dictionaries have fewer entries than conventional bilingual dictionaries. Differences between synonyms from L1 are explained in terms of corresponding lexical items from L2. In spite of being very helpful these explanations are unfortunately based on lexicographers' intuition. In addition, usually only one equivalent is given and no information about context in which it is used is provided. This is a simplistic view of relations between items from a source and target language. The issue of discrimination of synonyms has also been addressed in linguistics notably in contrastive studies based on componential analysis. As was discussed in Chapter 2, the problem with this approach is that it relies on vague assumptions regarding the nature of the components which supposedly should point to differences between semantically related terms. The present study departs from these previous approaches by tackling the problem from a perspective of the differentiation principle.

Relying on this principle a model for distinguishing between semantically similar lexical items that focuses on their distribution from an intralingual and interlingual perspective is proposed. This model is based on the language in use theory of meaning as discussed in Chapter 3. The model further involves the fusion of Zellig Harris' distributional method with corpus linguistics and the probability approach to language studies.

The second question also arises from the observation of the design of current bilingual dictionaries. These dictionaries are mainly alphabetically ordered and bilingual onomasiological dictionaries are in the minority, usually not part of huge publishing and research projects. Such dictionaries are much smaller in format and less comprehensive. However, the long tradition of bilingual lexicography speaks in favour of further development of bilingual onomasiological dictionaries. Improved bilingual onomasiological dictionaries could be used along with the existing monolingual and bilingual learners' dictionaries and remedy some of their shortcomings. For example, unlike alphabetical dictionaries they would enable users to search terms according to their meaning and not form. Similarly, unlike monolingual thesauri they would facilitate the understanding of foreign terms by providing direct translations. One serious problem with the existing monolingual or multilingual onomasiological dictionaries currently available is that the semantic groups into which terms are classified are not empirically well-founded. In this thesis an attempt is made to generate semantic classes by observing the distribution of lexical items from L1 and L2 in a parallel corpus.

In general, by exploring the above questions the present thesis tested the appropriateness of a statistically informed distributional model to the study of relations between lexical items from two languages.

To what extent the proposed model turned out to be useful will be discussed in the following section. I will first summarise the main findings in section 8.1 and then in sections 8.2 discuss their relevance in relation to practical lexicography and some general lexicological issues. In section 8.3 the potential application of the findings will be framed in the context of second language teaching and learning. Finally, the section 8.4 concludes with limitations of the present study as well as with possible further research.

8.2 Review of findings

The findings that emerged from the studies conducted indicate the feasibility of the distributional approach to deal with the first question from above.

First, the analysis of the lexical items from the TLD {CAUSE PROBLEM} and {PROBLEME BEREITEN} shows how these items differ in terms of their general grammatical features. Thus, in the analysis of the English items from the TLD {CAUSE PROBLEM} it was observed that the occurrence of the items <cause problem|difficulty> and <present problem|difficulty> with the direct object distinguish them from other items from the same domain. Similarly, we observed that the German nouns <Problem|Schwierigkeit> were used in a topicalised position when they collocated with <bereiten> but this behaviour was not found with other transitive verbs.

Second, the local grammar categories provide an accurate description of typical contexts in which lexical items occur. The lexical units from these categories are classified according to their functions. A comparison of the occurrence of lexical items in these contexts indicates that there are some additional differences between them. For example, the lexical items from the TLD {CAUSE PROBLEM} and {PROBLEM BEREITEN} are not equally likely to collocate with modal verbs.

Third, it was observed that different realisations of lexical items occur with different probabilities. Thus, the verb <cause> is much more frequently used with the plural form of the nouns <problem> and <difficulty> than with their singular forms. Similarly, the verb <schaffen> is more typical with the singular form of the nouns <Problem> and <Schwierigkeit>.

Finally, the analysis of collocates indicates that there is a relationship between the frequency of lexical items and other variables studied. The frequency of the lexical items that belong to the same domain or sub-domain tends to differ significantly. There are also statistically significant differences between lexical items in relation to the number of their collocates. In addition, these two tendencies have strong positive correlation; the variation in the number of collocates associated with an item is parallel to the variation in their frequency of occurrence. One of the tendencies observed was also that more common lexical items

shared with less common items the collocates that occurred with very high frequency. The degree of overlap here increased as the difference between frequency values decreased. Finally, we also saw that the number of stronger collocations was related to the frequency variable; e.g. the more frequent lexical items were more numerous in stronger collocations.

These findings, therefore, demonstrate that the lexical items can be differentiated purely on the distributional principle. Some of these features are displayed in Table 8.1. More detailed tables of this sort were produced in Chapter 5 and they illustrated how one can distinguish between different lexical items by relying on specific features. The table below shows that *no problem* occurs with six out of ten lexical items from the TLD {CAUSE PROBLEM} and that it most typically collocates with <there be>.

Variable Lexical items	Typicality <no> + *problem*	Typicality Modifiers + *problems*
<there be>	1	2
<cause>		1
<present>	2	6
<create>	4	3
<arise>		4
<lead to>		5
<pose>	3	7
<raise>	4	8
<result in>		9
<give rise to>		10

Table 8.1: Distinguishing features of lexical items from the TLD {CAUSE PROBLEM}

The results from above can be interpreted in terms of mutual substitutability of the items in question. Those lexical items that have higher values can substitute for the items that have lower values for the given variable. To stay with the example from above one can say that <there be> can replace all verbal expressions from the TLD {CAUSE PROBLEM} when they co-occur with *no problem* and this will not result in non-idiomatic expressions. Similarly, <cause problem> can replace all other less frequent items from the TLD {CAUSE PROBLEM} when it co-occurs with modifiers.

The above findings are relevant in so far as we are concerned with the substitutability of lexical items from a monolingual perspective; e.g. within a given domain or sub-domain. However, if we were to talk about their substitutability when they are used as translation correspondences we would need to take into account correspondence relations between items from two languages. This problem was addressed in terms of what is called correspondence potential. The calculation of correspondence potential is based on the number of the lexical units from L1 to which the items from L2 correspond and the percentage with which they are selected. These two variables tend to correlate and that the lexical items with a higher number of correspondence relations are usually among the most frequently selected translation correspondences. In some but not all cases the items with higher correspondence potential occur also with the highest frequency in the reference corpus.

Building on the idea of grouping lexical items according to their meaning the current thesis shows that such groups can be established through observation of contexts in which items from two languages occur. These classes of lexical items are named translation lexical domains. The classes are identified through an analysis of the occurrences of terms in a corpus of translation texts. The term *lexical domain* was chosen because the more commonly used term *semantic field* has become associated with the referential theory of meaning. In the initial study that dealt with the lexical items from the TLD {CAUSE PROBLEM} and {PROBLEM BEREITEN} the lexical units were grouped into corresponding sets according to their substitutability in a very specific context. In this way, we were able to observe that the verbs <schaffen>, <bereiten>, <aufwerfen> or <entstehen> corresponded to the same English items when they collocated with either <Schwierigkeit> or <Problem>. In Chapter 5 a broader context was selected at the onset. This analysis demonstrated that lexical items also formed corresponding relations when they colligated with a specific word class. A further investigation of particular collocations made it possible to explore the relationship between a whole domain and its particular sections. These sections were called sub-domains. A description of one of these sub-domains showed that the behaviour of lexical items in a specific section may differ from that observed at the general level. This is similar to the differences observed in relation to the occurrence of lexical items in Chapter 5 with different word forms of the nouns <problem>

and <difficulty> in English and <Problem> and <Schwierigkeit> in German. We concluded that the results cannot be generalised and that they rather always need to be contextualised.

8.3 Significance of findings

The significance of the above findings will be discussed in terms of their theoretical and practical contributions. In section 8.3.1 the contribution to theoretical lexicological issues will be considered. Section 8.3.2 discusses the application of these findings to practical lexicography, with the special emphasis to learners' and translation dictionaries. Section 8.3.3 reviews the potential use of translation lexical domains and sub-domains and the information obtained in analyses in teaching collocations by means of translation.

8.3.1 Contribution to lexicology

The classification of words according to their semantic similarities, at least since Bacon, has been based on the assumption that words stand for concepts or mental representations. Accordingly, the classification of these concepts into groups or *macrostructures* - to use the lexicographic term - has been considered to be equal to the representation of our knowledge of the world. Such classifications contain only a limited number of 'basic concepts' to which the whole of our knowledge can be reduced and are treated as universal for all languages (McArthur, 1986). The dictionaries that catalogue these concepts or thematic lists, as they are also known, therefore "provide their users with a view of the world" (Hüllen, 2009: 109). In this sense, such dictionaries are akin to encyclopaedias with which they are also historically related (Hüllen, 1999: 65). Bacon himself was more interested in the classification of knowledge than in lexicography or lexicology. Bacon's assumption underlies also the production of later standard works including Bathe's *Ianua linguarum*, Comenius' *Orbis* or more recent Roget's (1852) *Thesaurus*, Sanders' *Deutscher Sprachschatz,* Dornseiff's *Wortschatz nach Sachgruppen*, *Historical Thesaurus of the Oxford English Dictionary* (Kay, Roberts, Samuels and Wotherspoon,

2009), McArthur's (1981) *Longman Lexicon of Contemporary English* and *Longman Language Activator* (Summers, 1993).

The distributional model developed in the current thesis proposes a different approach. According to this model, the classification of lexical items emerges from the description of the distribution of lexical items in a parallel corpus. Therefore macrostructures are being identified rather than created as is the traditional approach. The advantage of this model is that it is less dependent on human factors. The categories constructed in the traditional model are based on intuition and are therefore inevitably subjective. They depend on authors' knowledge and discretion as to what she considers a relevant category. In spite of the belief that they represent some universal concepts "the categories of macrostructures do not mirror any reality directly, they mirror the mind which works in order to understand reality" (Hüllen, 2009: 94). Dictionaries themselves refute the assumption that universal categories exist. For example, Daniel Sanders (1873) who adapted Roget's categorisation eventually modified and reduced them from 1000 to 688 (Kühn, 1985: xxvi). On the other hand, another German lexicographer from the same period, August Schlessing, borrowed Roget's system without changes and applied it to German. But and here's the rub: authors of dictionaries usually disagree with regard to what constitutes 'basic concepts'. According to Roget, the whole English vocabulary can be reduced to six basic concepts, and in the *Historical Thesaurus* there are three such concepts, whereas McArthur's system contains 14 basic macrostructures. Even the same system can undergo changes if more authors work on it. For example, if we compare the seventh edition of Franz Dornseiff's *Der Deutsche Wortschatz nach Sachgruppen Wortschatz* with the eighth edition edited by Uwe Quasthoff we observe that the macrostructure *Society and Community* (in original *Gesellschaft und Gemeinschaft*) from the latter edition is extended in the latter edition into four new themes: *Human life*, *Food and drink, Sport and leisure* and *Society* (*Menschliches Zusammenleben, Essen und Trinken, Sport und Freizeit* and *Gesellschaft*). Moreover, such basic concepts change over time following the spirit of the times. They are always related to a specific culture or chronological context, and a specific world view such as Anglocentric or Francocentric (McArthur, 1998: 153). The concepts of *God, heaven, angels, sun, moon, earth* and *sea* which we encounter in the 11th century dictionary of Aelfric are replaced

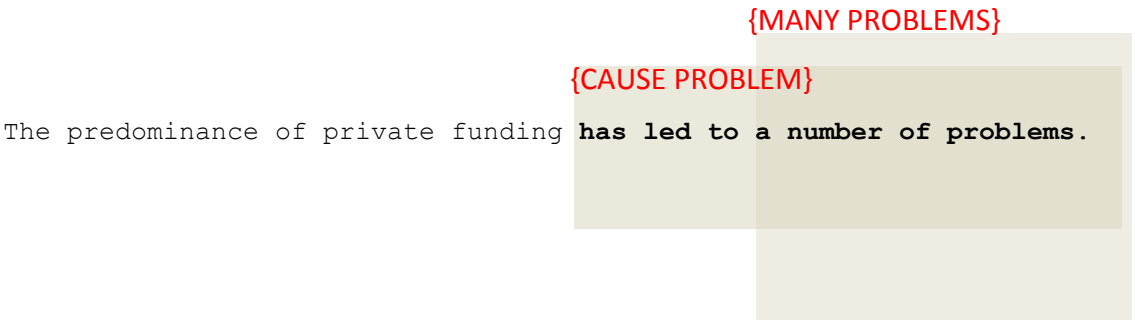
in McArthur's dictionary of 1981 with *Life* and *living things* (McArthur, 1986: 162). The former terms clearly mirror a theological viewpoint whereas the latter are products of knowledge developed under the influence of natural science. This shift reflects ideological changes in our understanding of the world where "[s]cience came to be seen as alternative to Christianity" (Harris, 2005: 34). Although no method can yield completely objective results the distributional model makes it possible to avoid the above issues. The potential flaws and shortcomings of the distributional model are transferred to apparatus-related issues. This, of course, does not mean that the problems we deal with now have become any less serious. They are simply more amenable to testing than human intuition.

Previous onomasiological dictionaries alongside basic concepts contained also some more general concepts. Such a view presupposes hierarchical relations between concepts. The hierarchical relations have been interpreted since the 19th century in terms of taxonomic representations based on botanical metaphors such as roots, stems and trees that were adapted in philology (McArthur, 1986: 142). Roget, who in his own introduction acknowledges this relation to botany and zoology (1852: xii), worked as a natural scientist and issues of categories and their interrelations were of central interest to him (Hüllen, 2004: 17). In traditional onomasiological dictionaries like in other taxonomic models all categories and sub-categories were defined in advance. Accordingly, the researcher picked out words and assigned them first a category and then ordered them according to their generality into appropriate groups and sub-groups. In the distributional model the macrostructures are not presupposed in advance and lexical items are not interpreted in terms of the existing categories. The macrostructures rather emerges as the result of the description of the occurrence of items. It is not known in advance in how many domains an item may occur and how many sub-domains a domain will consist of. The structure of domains is based on the distribution of lexical units and on their substitutability. Thus, rather than distinguishing between the items which may be more or less general in meaning we make a distinction between those which are more or less usable as substituting items. The substitutability is here determined both from a monolingual and bilingual perspective. In the former case, one item has a higher substitutability power or substitution potential than another if it occurs more often in the same context, if it occurs with

a greater number of collocates and if it is used in a larger number of strong collocations. From the latter perspective, an item from L2 has a higher correspondence potential than another item from the same language if it corresponds to a larger number of lexical units from L1 and is used with higher frequency as a translation correspondence.

Finally, lexical domains have more flexible boundaries than traditional macrostructures and semantic fields. They present open sets rather than self-contained groups. The number of members of a domain or sub-domain is provisional since it reflects the current usage of lexical items in a given corpus. This is also a potential source of problems, an issue which will be discussed below in 8.4. An item may belong to different domains which will depend not only on the context in which it occurs but also on its use as a translation correspondence. We saw in Chapter 7 on the example of the lexical item <mehrere> how an item might have very high substitution potential which still does not mean that it will be considered very central to the given domain because of its low correspondence potential.

In the distributional model the number of domains or sub-domains associated with a lexical item is not defined in advance either. This depends on the number of collocations in which an item occurs and correspondence relations that this item establishes with items from another language. Thus, in our analyses it was observed that the nouns <problem> and <difficulty> occurred both in the TLD {CAUSE PROBLEM} and TLSd {MANY PROBLEMS}. When observed in a running text the domains and sub-domains overlap when they contain the same lexical item. This can be graphically depicted in the following way:



Here, we can see how the lexical unit <a number of problems> belongs both to the TLD {CAUSE PROBLEM} and the TLSd {MANY PROBLEMS}. The role of specific lexical items may also vary

from domain to domain. Positive quantifiers that constitute one of two main elements of the lexical units from the TLD {MANY COLLECTIVES} are only an optional element in the TLD {CAUSE PROBLEM}.

8.3.2 Contribution to practical lexicography

In this section the potential application of the results obtained with the distributional model will be discussed. Practical lexicography includes a great range of topics such as the design of entries, the size of dictionaries or the type of dictionary definitions. I will here restrict my discussion only on two topics. 8.3.2.1 will explore how to prepare the obtained results for their implementation in practical lexicography. In sections 8.3.2.2 and 8.3.2.3 I will discuss the potential contribution to bilingual learners' and translation dictionaries.

8.3.2.1 Selection of options

Although the substitution and correspondence potential of the lexical items that belong to the same domain or sub-domain correlate this correlation is not perfect. There are many exceptions and one cannot automatically assume that a more frequent lexical item from a domain or sub-domain will also have higher correspondence potential than the less frequent ones. For this reason it is necessary to separate the information obtained in intralingual and interlingual studies and find a way to calculate their individual impacts. I propose here a model that relies on decision theory (e.g. Berger, 1985; Lehmann, 1950; Raiffa and Schlaifer, 1968). There is obviously no room in the present thesis to discuss this theory in detail but I will briefly introduce the aspects which are relevant for our purposes.

Decision theory was pioneered in translation studies by Levy (1967). In this paper Levy defined translating as a decision process in which a translator chooses between a set of alternatives that belong to the same "semantic paradigm of words" (Levy, 1967: 156). These alternatives are mutually exclusive since only one of them can be selected at a time. As in every

decision-related context before one can take a decision it is necessary to consider various risks and choose the option which brings maximum utility. Risks are a type of uncertainty in which insufficient information can lead to undesired outcomes. The reduction of uncertainty and risks can be interpreted positively in terms of increasing the values of utilities. Thus, by increasing utilities we simultaneously reduce risks and uncertainty. One of the advantages of the theory is that different options can be quantified and numerically represented. Therefore, one weighs the available options and after having compared their values arrives at the option with the highest numerical value. This is, therefore, the most appropriate option. The relations between the values are displayed in the form of mathematical notation. Thus, *a>b>c* means that the item *a* has a higher value than *b* which in turn has a higher value than *c*. From this also follows that *a* has the highest value in the given set. In *a=b>c* the item *a* and *b* occur with equal values and since *b* is higher than *c* it follows that *a* is also higher than *c*.

Now, it will be shown how the theory can help to implement the results obtained in earlier studies into practical lexicography. There are three sorts of uncertainty in relation to the use of the lexical items that belong to the same TLD. Lexical items, on the one hand, may occur with different probabilities even when they occur in the same context. Thus, the use of lexical items by a non-native speaker always bears the risk of selecting an option which is unidiomatic or less native-like. On the other hand, lexical items from L2 correspond with different probabilities to items from L1. Here, one risks choosing an option that might be at odds with the dominant understanding of the mutual correspondence of items. Finally, since probabilities in the two contexts do not completely correspond to each other even if one knows their values this does not automatically help in choosing the most appropriate term. If a lexical item occurs with different probabilities one can select the highest value of one of two variables and accept that the expression created is idiomatic but not a good translation or that it is not a very native-like expression but that it is an appropriate substitution for an item from another language. But there is a third solution. In decision theory, one can simply add up the values of the two variables and order the selection according to the calculated sums. The option with the highest total sum has the highest utility and lowest risk and the one with the lowest total sum is most risky.

The probability values of the first type were investigated in the present thesis in terms of association strength, whereas correspondence relations were described in terms of percentages. In order to facilitate calculation and make the two types of probabilities comparable it is better to use the same kind of measurement. Since association strength is more reliable the values of both variables will be calculated using this measurement. It means that the relation between the items from L2 and L1 will be interpreted in terms of how typically items from the same domain or sub-domain are associated with the same item from L1. The model is illustrated through two examples below.

First, we will consider a situation in which one needs to choose between several translation correspondences for the English lexical item <create problem>. We have seen that the lexical items in the TLD {CAUSE PROBLEM} occur with different probability with the plural and the singular forms of the noun <problem>. For this reason, I approached the issue of translation correspondences in two distinct analyses. In both cases the occurrence with modifiers were taken into account.

In decision theory, alternatives and compared utilities are usually presented with a help of decision tree diagrams. One such diagram has been constructed for the translation correspondences of <create> *problems* (Figure 8.1).

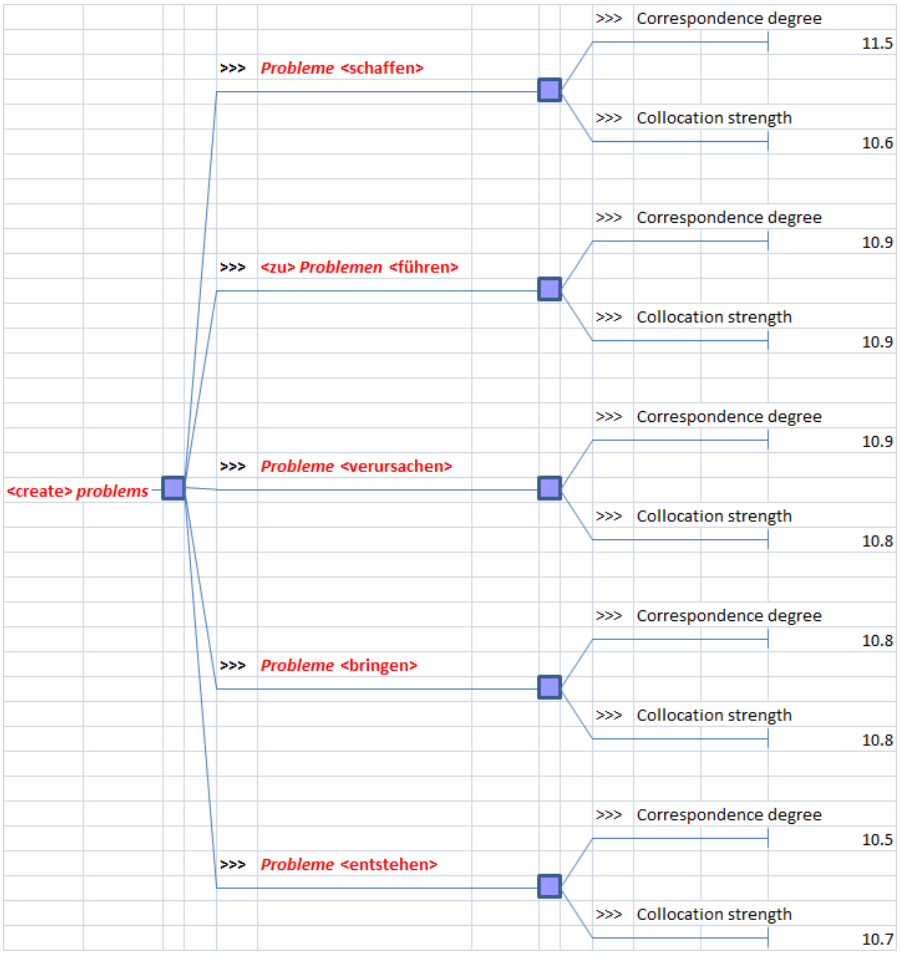


Figure 8.1: Decision tree for German translation correspondences of the lexical item <create> *problems*

The English lexical item is presented at the initial point, whereas the German options are displayed as branches. There are only five lexical items in German that meet the requirements regarding the correspondence introduced earlier in the thesis. The graph indicates the values of the correspondence degree and collocation strength for each of these alternatives. The values are based on the logDice measurement. It can be observed how these values vary with regard to the variables studied. Thus, *Probleme* <schaffen> corresponds with the highest degree to the English lexical item in question, but <zu> *Problemen* <führen> presents the most typical collocation among the alternatives. The latter is equally associated with <create> *problems* as *Probleme* <verursachen>. In order to find the option that bears minimum risks certain gains need to be sacrificed. This is done by adding up the scores of the variables and by dividing them by the number of variables, which is in our case two. This outcome represents the utility values for all items. The translation correspondence with the highest value is treated as the most

suitable option because it maximally reduces uncertainty. The results for <create> *problems* are displayed in Table 8.2.

Translation correspondences of <create> *problems*	**Utility values**
Probleme <schaffen>	11.1
<zu> *Problemen* <führen>	10.8
Probleme <verursachen>	10.7
Probleme <bringen>	10.7
Probleme <entstehen>	10.5

Table 8.2: German translation correspondences of <create> *problems*

The lexical item with the highest utility values is *Probleme* <schaffen> and therefore in a dictionary entry it would be included as the first option. Since the following three items have similar values they can be considered as equally suitable alternatives. Finally, *Probleme* <entstehen> serves as a less appropriate option here.

In a similar way we can explore the alternatives for <create> *problem*. The results are presented in Table 8.3. Here, according to the utility values the most appropriate alternative is *Problem* <darstellen>. Only slightly less risky is *Problem* <schaffen>. The difference between the two has to do with their values of the correspondence degree and collocation strength. The latter item corresponds more typically to the English lexical unit than the former (40% to 15%) but collocation strength of the former is three times higher. In addition, there are four additional options two of which in each case have equal utility values

Translation correspondences of <create> *problem*	**Utility values**
Problem <darstellen>	11
Problem <schaffen>	10.9
<zu> *Problem* <führen>	10.5
Problem <entstehen>	10.5
Problem <bereiten>	10.1
Problem <verursachen>	10.1

Table 8.3: German translation correspondences of <create> *problem*

The illustrations above suggest that results obtained through distributional analysis can be successfully implemented in practical lexicography through the framework of decision theory. Dealing with different options is one of basic questions in practical bilingual lexicography. The fact that these different choices can be quantified makes it possible to distinguish between them in terms of associated numerical values. The above examples dealt with two specific sets of collocations. The model can be equally applied to more general or to more specific expressions. Thus, one can explore the utility values for the lexical item from the TLD {CAUSE PROBLEM} which occur with a specific modifier such as <cause serious problem>. One can also select the lexical unit <zahlreiche KOLLEKTIVA>. It is, however, the lexicographer who after taking into account all other lexicographic issues decides how and in what level of detail a lexical item should be described in a dictionary. Similarly, the lexicographer can also decide whether to include the values of correspondence degree and collocation strength separately in order to inform users about specific aspects of their selection process.

As a final remark, it must be stressed that at this stage it is still not clear how significant differences between utility values are. For example, Table 8.3 indicates that the difference between the utility values of *Probleme* <verursachen> and *Probleme* <darstellen> is 0.9 and one can conclude that this is a significant difference but not how statistically significant it is. Further research is needed in this area.

8.3.2.2 Bilingual learners' dictionaries

In this section, I will discuss the advantages of using translation lexical domains and sub-domains in creating bilingual learners' dictionaries based on onomasiological principle.

One may question the need of bilingual onomasiological learners' dictionaries (henceforth BOLD) because high quality monolingual learners' dictionaries are available. I would claim that such dictionaries can supplement the existing dictionaries because their use bears several advantages. Two most relevant advantages in the context of the current thesis are economy, a higher certainty that a learner will understand the meaning of a term, and their

suitability in both text production and reception. Svartvik (1999: 287) illustrates the first two advantages with the following example:

> "To choose a very concrete word as an example: in a monolingual dictionary it takes some 50 words to define the two meanings of the English word *radiator* - and there is of course still no guarantee that the meanings will be clear to the EFL student. By contrast, a bilingual English-Swedish dictionary can achieve this by giving the two corresponding Swedish words: *varmeelement* and *radiator.*"

Therefore, by using a bilingual dictionary a language user needs less time to understand the meaning of a term and it is more likely that this understanding will be correct than by consulting a monolingual dictionary. Part of the problem with monolingual dictionaries is that provided definitions themselves might contain unknown words which lead to a situation in which an unknown term is defined through other unknown terms. By contrast in bilingual dictionaries unknown terms are explained in terms of known terms from the mother tongue. Translation lexical domains and sub-domains are useful in this setting. The ordering of lexical items in a dictionary entry according to their utility values provides dictionary users with direct access to the most typical translation correspondence of an unknown term. If a person is not interested in an accurate translation of a term but in its general meaning it will suffice to look up the name of the domain or sub-domain that contains the items from the mother tongue. Because they are labelled according to the most frequent item it will provide the user with a hint as to what the unknown item means. Thus, if someone wants to find out what <numerous> means in German she can look up the appropriate domain and see that its meaning is related to <viele> which is used in labelling the TLD {VIELE KOLLEKTIVA}. Or, for a more accurate definition one would be directed to the lexical item <zahlreiche KOLLEKTIVA>. This is illustrated in Table 8.4 which contains a simplified entry for <numerous COLLECTIVES> and corresponding German terms. In addition to the name of the German TLD, it also displays the list of translation correspondences with high utility values. The table also displays that the English item belongs

to the TLD {MANY COLLECTIVES}, which might also be useful if the learner is already familiar with the meaning of the item <many>.

<numerous COLLECTIVES>: {MANY **COLLECTIVES**}
{VIELE KOLLEKTIVA}
<zahlreiche KOLLEKTIVA>
<viele KOLLEKTIVA>
<eine Reihe KOLLEKTIVA>

Table 8.4: A dictionary entry for the lexical item <numerous COLLECTIVES> and its translation correspondences

In order to understand the potential application of TLD to dictionaries created for text production and text reception let us first consider the difference between the two. This difference can be explained in terms of what is a known item and what is an unknown item for a language user (Hannay 2003: 146-148). In the context of a foreign-language production task the "user is going from the known to the unknown" (Hannay, 2003: 146), that is, from one's own language, and she is looking up the available options in the foreign language. For the reception task, on the other hand the user "is going from the unknown to the known" (Hannay, 2003: 148). Here, the user selects a term in a dictionary from the language she knows in order to understand an unknown term from a foreign language.

The application of the TLD to the receptive tasks is illustrated above with the German adjective <zahlreich> and its translation correspondences in English. It was shown how the lexical domains and the general name of the domain create a direct link to known terms from the mother tongue.

In the production task the learner usually has an idea what she wants to say but does not easily find a suitable expression. Here she usually relies on synonym dictionaries or thesauri. The *Longman Language activator* (Summers, 1993), 'the world's first production dictionary', was created for this specific purpose. A BOLD based on the TLD can serve this function very well. Like bilingual synonymy dictionaries the TLD simultaneously provide direct access to the term from a foreign language and the list of synonymous terms. Thus, one does not first need to look up the term in a conventional bilingual dictionary to find the

corresponding term and then to consult the monolingual synonym dictionary containing a list with the available options. Both can be achieved with one step by means of the TLD. This will be illustrated by one example. Let us imagine that a learner has the lexical item <Problem schaffen> on mind and wants to express a similar idea in English. A learner needs first to look up the verb <schaffen> in a bilingual German-English dictionary. The entry for the verb <schaffen> in *Pons Collins Großwörterbuch für Experten und Universität* (Terrell, Schnorr, Morris and Breitsprecher, 1999: 696) is reproduced below:

> a) (= *hervorbringen*) to create; die schaffende Natur the creative power of nature; der schaffende Mensch the creative human being; dafür ist er wie geschaffen hes just made for it; wie ihn Gott geschaffen hatte as God made him
>
> b) pret auch schaffte [] (= *herstellen*) to make; Bedingungen, Möglichkeiten, System, Methode, Arbeitsplätze to create; (= *verursachen*) Ärger, Unruhe, Verdruss to cause, to create; Raum or Platz schaffen to make room; Probleme schaffen to create problems; Ruhe schaffen to establish order; Klarheit schaffen to provide clarification; Linderung schaffen to bring relief (für to)

From this entry, one can conclude that <schaffen> has two senses and that the collocation *Probleme schaffen* belongs to the second sense. The corresponding English term is *create problems.* Since no other options are provided the learner needs to consult an English synonymy dictionary. For this example I consulted the *Macmillan* online thesaurus in which this whole collocation does not occur. Nevertheless, one of the entries contains the following definition of the verb <create>: “to cause a situation, feeling, or problem to exist”. For this specific sense a list of similar terms with their definitions is provided:

> **catalyze** verb
> to cause something to happen, especially in way that involves a lot of change
> **bring about**
> to make something happen, especially to cause changes in a situation
> **trigger** verb

to make something happen
form verb
to make something exist or develop
invent verb
to develop a new theory, style, or method that did not exist before
develop verb
to grow something
inaugurate verb
to start or introduce something new and important
get/set/start the ball rolling
to make something start happening
make verb
to create or produce something by working
start off
to make something begin

In none of these definitions does the noun <problem> occur and it does not seem that any of the available terms can be used instead of <create problem>.

The aforementioned *Longman Language Activator* seems to be slightly more useful here. It does not contain the whole collocation and the verb <create> is classified under the section "CAUSE: to make something happen". The section contains the following ten terms: *cause, be the cause, be responsible, result in something, lead to something, give rise to, bring about, make* and *create for*. Given the fact that the noun <problem> collocates with none of these terms it is necessary to go through definitions and examples in the entry. The examples provided do not contain the collocation <create problem> either and therefore one can only rely on the definitions. According to these definitions the learner can finally conclude that in addition to <create problem> the other available options are:

be the cause: to be the particular reason for a problem or difficulty;
be responsible: to be the person or thing that causes something bad to happen;
cause: to make something happen, especially something unpleasant:

Since we know from the analysis in Chapter 5 that the lexical unit <Problem schaffen> does not occur with animate nouns the second term can be discarded. Thus, after so much effort the learner can conclude that there are three possible ways of expressing the idea in question in

English. It may be noted that some of the items that occur in the TLD {CAUSE PROBLEM} are listed in the entry, but their definitions are too general and not very helpful. From the definition "if an action or event results in something, it makes something happen" it is not straightforward that <result in> can be used with the noun <problem>.

On the other hand, by looking up a BOLD entry the learner would be provided directly with a list of corresponding terms. From, this she could conclude that in addition to <create problem> the same idea can also be expressed by using <cause problem>, <cause difficulty>, <pose problem>, <pose difficulty> or <result in problem>. Therefore, the process of searching and finding the term would become much quicker and more reliable. For example, the aforementioned *Langenscheidt Collins Großwörterbuch* contains the collocation *create problems*. In addition to <create> this dictionary lists only one more verb from the TLD {CAUSE PROBLEM} (<pose>) as a collocate of the noun <problem>. Our analysis in Chapter 5 showed that neither <create> nor <pose> were the most frequent collocates of this noun. It is, therefore, not quite clear why lexicographers decided to include collocations with these two verbs and to ignore those that, according to our study, occur with a higher frequency (such as <cause> or <lead to>). Finally, lexical domains and sub-domains are based on typical contexts in which lexical items occur and as a result collocations included in a BOLD would be more representative of these lexical items.

Second, bilingual dictionaries are rather scarce on grammatical information and description of context. For example, in the German-English dictionary cited above the only information given for the verb <verursachen> is that it is a transitive verb which can be used in past participle. Some other verbs have more accurate descriptions such as the verb <present> which is said to occur in "to present sb with sth, to present sth to sb" and in a range of other constructions. This type of information indicates only the general behaviour of lexical items. The distributional model provides a description that includes both general and specific information. Thus, in Tables 5.12 and 5.26 we summarised the general grammatical information regarding transitivity or the use with indirect object and modal verbs. In addition, by using the local grammar categories we were also able to specify the type of modifiers that occur with the lexical items from the TLD {CAUSE PROBLEM} and {PROBLEM BEREITEN} or the type of nouns

which occur in the subject position. Such information can additionally help users to understand and use terms from a foreign language.

Finally, it seems that a BOLD based on the distributional model can provide a solution for an old problem that conventional bilingual dictionaries cannot adequately deal with. This is the problem of incongruence between linguistic units from two languages. Durrell (1988: 232), for example, shows how English and German have non-congruent terms for words related to dying. Thus, the words used in German can be hierarchically represented in the following way:

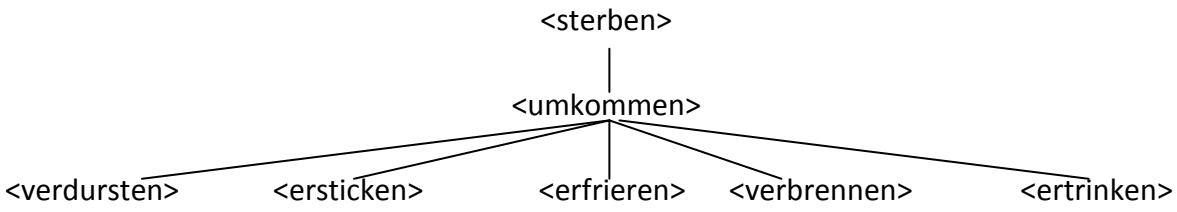


Figure 8.2: German verbs with the sense *dying* (adapted from Durrell, 1988: 232)

According to Durrell, in English there is no comparable hierarchy. There is a generic term <die> that corresponds to <sterben> but there is no item from the intermediate level. At the low level in English there is <drown> and the multi-word expressions such as <freeze to death> or <die of thirst>. Dictionary entries for such words are usually based on a componential analysis and on the description of what they imply. For example, the verb <erfrieren> would be considered to consist of the general component <die> and the more specific element <to be killed by frost>. As was said in 2.2.1, although component analysis can deal with such words it is not very helpful when it comes to more abstract terms. For example, the German adjective <weit> has two English correspondences <wide> and <broad> (Durrell, 1988: 239). The differences between two English adjectives cannot be described in terms of semantic components. On the other hand, the application of the distributional model to a parallel corpus shows that <weit> corresponds to <wide> when the latter occurs in the collocations <wide range>, <wide selection>, <wide variation> or <wide choice> and that it corresponds to <broad> when the English term collocates with <shoulder>, <support> or <term>. Thus, in a BOLD different uses of <weit> would be placed in different domains or sub-domains. Similarly, the context analysis of

the translation correspondences of <umkommen> by means of a parallel corpus would indicate how this verb is translated into English. Eventually, to acquire the term the learner would need to learn how to use it in a specific context.

8.3.2.3 Translation dictionaries

The need for special translation dictionaries has recently become an important issue (Snell-Hornby, 1996: 90-97). So far no specific onomasiological translation dictionary has been available. Especially rare are translation dictionaries based on parallel corpora (Salkie, 2008). Apart from the availability problem, the copyright issue and the quality of translation one important reason for this is also that no sufficient linguistic and lexicographic model has been developed to deal with the data. Salkie illustrates this problem by citing how a lexicographer responded to his question as to why translation corpora have not been used in compiling new dictionaries:

> "She replied that they had started to use a small journalistic corpus in the two languages, but had come up with such a huge amount of fascinating data that they had reluctantly decided to abandon it: they were spending too much time trying to work out how to handle this rich range of material." (Salkie, 2008: online)

He concludes: "It is regrettable that translation corpora have been around for about two decades but that practical and theoretical problems have prevented their use in bilingual lexicography, where their potential is vast" (Salkie, 2008: online). According to Zgusta (Wierzbicka, 1987: 1–2), if "the treatment of meaning in dictionaries is to be radically improved, preparatory work has to be done by linguists". Some preparatory work has already been done. Snell-Hornby (1984: 278) suggests that translation dictionaries should be "contrastive dictionaries of synonyms, whereby the alphabetical system gives way to arrangement in semantic fields". Similarly Atkins (1996: 8) proposes that translation dictionaries

> "should offer the skilled user the chance to make his or her own judgment on equivalences, by scanning examples of the TL items (grouped according to meaning) in various types of context, as well as - for contrastive checking purposes - examples of the relevant meaning of the SL item in a wide variety of contexts."

The translation lexical domains and sub-domains with local descriptions of the contexts in which lexical item occurs and with the data prepared by applying the decision theory seem to be a good starting point for the design of translation dictionaries. What has been said above with regard to productive and receptive tasks applies here as well. Thus, the domains and sub-domains provide quick access to foreign terms corresponding to terms from one's mother tongue. This is what recommends them for the task of translating from a mother tongue into a foreign language. With the help of TLD one can also easily find the terms for translating from a foreign into the native language. In addition, the domains and sub-domains are capable of being used as entries for synonyms. The information with regard to the correspondence degree and collocation strength, on the one hand, and to the local context, on the other, may help translators to "make his or her own judgment on equivalences" (Atkins, 1996: 8).

However, one should not conclude too quickly that translation dictionaries should be identical to the BOLD. The two would differ with regard to their content. Translators usually need the information about less frequent terms and regarding stylistic differences. Given that lexical items which constitute the TLD are based on the logDice coefficient indicating the most typical and frequent collocations it is clear that this measurement could not be used. One option might be the use of MI score which highlights peculiar and rare collocations (Church and Hanks, 1990) but this issue requires more detailed research.

Although the present research did not deal with non-textual context and stylistic differences between lexical items the distributional model seems be applicable to this issue as well. Similar to the comparative studies of textual contexts we can also explore the distribution of lexical items across different genres in order to identify dominant patterns. A provisional

analysis of the distribution of lexical items from the TLD {MANY COLLECTIVES} carried out by using Lee's (2001) Genre Classification Scheme gives reason to believe that the distributional model would be adequate for this task. Thus, the results obtained indicate that the items differ in the number of genres in which they occur. Furthermore, it seems that this difference is related to their frequency. For example, <many COLLECTIVES> which is more frequent than <a number of COLLECTIVES>, <numerous COLLECTIVES> and <a considerable number of COLLECTIVES> occurs also in a higher number of different text types. The same holds true for <a number of COLLECTIVES> and <numerous COLLECTIVES> when they are observed in relation to less frequent lexical items. A comparison of individual text sorts indicates that <numerous COLLECTIVES> occurs more typically in academic prose than <many COLLECTIVES> which, on the other hand, is more typical in classroom discourse and school essays. Although these preliminary results are encouraging from the point of view of testing the feasibility of the distributional model more detailed analyses are required.

8.3.3 Contribution to the use of translation in language teaching

The use of translation in second language learning and teaching has become less and less relevant in the last forty years after the Direct Method completely overthrew the Grammar-Translation Method (Cook, 2010: 3-15). The Grammar-Translation Method, which was the dominant way of teaching foreign languages in Europe from the later 19th to the mid 20th century, failed because it was focused on form and invented examples (Gommlich, 1997: 171-191). On the other hand, with the Direct Method teaching takes place in the target language, the focus is on meaning and on naturally- occurring language (Cook, 2010: 8-9). However, in spite of having a bad reputation translation did not disappear completely from teaching and it remained a "language teachers *forbidden friend*" (Zojer, 2009: 32, italics in original) even for the Direct Method teachers "who, in theory, totally opposed the use of translation in the classroom" (Zojer, 2009: 32-33). Nevertheless, this has started to change recently and a range of works has been published in favour of using translation in teaching (e.g. Butzkamm, 2004;

Butzkamm and Caldwell, 2009; Cook, 2007; 2010; Witte et al., 2009). According to Cook (2007: 396) translation in language teaching "should be a major topic for future applied linguistic research and discussion". In his view, bilingual teachers 'naturally' use L1 in teaching when they need to explain some more difficult terms from L2 and therefore it is wrong to train them not to use this skill but one should encourage them to do so in a systematic way.

I will discuss below the potential application of translation lexical domains and sub-domains to the use of translation in language teaching (henceforth TILT, following Cook (2007)). I will illustrate this by showing how domains and sub-domains can help learners to acquire collocations. Learning collocations is not an easy task for language learners (e.g. Lewis 1993; Nation 2001; Nesselhauf 2005). According to Hyland (2008: 31) "it is often a failure to use native-like formulaic sequences which identifies students as outsiders and there is a general consensus that formulaic sequences are difficult for L2 learners to acquire". As Walker (2008: 291) shows, a part of the problem here is that "collocation tends to be seen as something which is idiomatic, and therefore cannot be explained". Therefore, collocations are often considered as not being explainable in terms of conventional grammar rules.

There are several factors that may facilitate the acquisition of collocations. One of these factors discussed by Walker (2008: 291) is a description of contexts of lexical items which reveals their characteristic collocates. These characteristic collocates can help to distinguish between semantically closely related items and identify typical collocations. Walker demonstrated this by comparing the use of the nouns <aspect>, <factor> and <issue>. The analysis showed that although <contentious> can be used with all three items "the native speaker is much more likely to use *contentious* together with *issue* because it relates to a key feature of its meaning (or at least one of its meanings)" (Walker, 2008: 295). The results of the comparative analysis of the lexical items that belong to the same domain or sub-domain provided exactly this type of the description of contexts. The degree of typicality of collocations was expressed quantitatively in terms of the logDice coefficient. The comparison analysis of collocations created with modifiers, for example, showed that although both <create> and <lead to> collocated with *new problems, huge problems, real problems* and *particular problems* the former verb formed stronger collocations. Similarly, <zahlreiche> collocated more typically

with *Veranstaltungen, Preise, Untersuchungen, Unternehmen, Vorträge, TeilnehmerInnen, Gäste* than with <eine Reihe ART|APPR>. Besides, the investigation of the classes of collective nouns demonstrated that characteristic collocates can be identified for a whole set of collocates. I would claim that this type of information can be combined with the use of TILT. Thus, characteristic collocations can be translated into typical translation correspondences and this will help learners better understand their meaning. For example, the German learner will more easily learn the collocation <create> *new problems* if she knows it that the German cognate is *neue Probleme* <schaffen>. Similarly, the English learner would be helped in acquiring the collocation <zahlreiche> *Preise* if she knows that the corresponding item in English is <numerous> *awards* or <numerous> *prizes*.

A broader model for teaching collocations and other idiomatic expressions is proposed by Willis (1990; 2003). I will briefly illustrate how this model can be combined with translation domains and sub-domains in the use of TILT.

Following Sinclair (1991), Willis (2003) considers lexis and grammar to be inseparable. The purpose of language learning, in his view, is communication of meaning and the learning process should focus on meaning. The meaning of a term is learned through its use. Therefore, the meaning is here approached from the perspective of the language in use theory of meaning. Given the fact that lexis and grammar are interdependent, the use of lexis cannot be chaotic becuase it follows certain rules and patterns. It is these rules which can help learners to pick up a foreign language in a systematic way. In other words, in acquiring the knowledge of using a lexical item or linguistic structures the learner learns specific language games associated with given items or structures. According to Willis, there are three major stages in this process: Recognition, System building and Exploration. Although his approach is developed mainly for the use in Direct Method I will demonstrate below through a discussion of these three stages that it can be successfully use in the approach that relies on translation.

Recognition: In order to acquire the use of a lexical unit the learner needs to be acquainted with the term. For this purpose Willis proposes the use of terms from the mother tongue: “Knowing the meaning of the word and its first language equivalent or equivalents is a matter of recognition, and this provides an important starting point” (Willis, 2003: 12). This is

what we described above as the receptive task. The lexical items from TLD or TLSd can serve this purpose very well. An unknown term from a foreign language can be explained in terms of its translation correspondences from a TLD or TLSd or by the cover term. For example, the lexical item <cause problem> can help the English learner in the recognition process of the collocation <Problem verursachen>.

System building: After having recognised an item the learner develops a hypothesis or is taught how to use it and how it is related to broader language systems. According to Willis, at this stage it is useful to describe these systems via patterns and structural grammar rules. The substitutability of the lexical items that belong to the same domain and sub-domain and related local grammar categories seem to be very suitable for this task. Thus, if the learner is acquainted with the term <Problem verursachen> it can be explained to her further that the noun <Problem> can be replaced by <Schwierigkeit> and that the verb <verursachen> can be replaced by several other verbs. In addition, she can also be taught that the two nouns can be modified by a set of other items and that in the subject position only a non-animate noun can occur. This can be additionally backed up by providing the information about corresponding items in the learner's native language.

Exploration: "Foreign language learning in a natural environment involves a lot of exploration" (Willis, 2003: 13). In this phase, the learner extends her knowledge of a system which has already been built. Through the exploration of the use of lexical items she refines their knowledge and become more confident language users. At this stage, the use of characteristic collocations can help the learner to understand subtle differences between the lexical units that belong to the same system.

Studies on second language vocabulary acquisition (Milton, 2009) indicate that more frequently used words will be learnt more easily. This information can be used to systematically teach lexical items by means of TLD and TLSd relying on the three phases from above. Thus, in teaching quantifiers the teacher can start with the term <many>, which is the most frequent item in the TLD {MANY COLLECTIVES}, and gradually through the system building process extend learners' vocabulary by including other less frequent items such as <numerous>, <a range of>

or <a considerable number of>. By listing corresponding items from German one makes sure that learners understand the meaning of the terms.

8.4 Limitations of the study and further research

The focus of the present thesis was on the development of a model of analysis. This is why only a restricted number of lexical items were investigated. Although the lexical items were selected randomly and not because of their representativeness it is only a further investigation of different types of linguistic units that can show to what extent the model is generally applicable. The studies in the present theses dealt only with the constructions consisting of a verbal predicate and verb complement, or of a noun and its modifiers. It should not be ruled out prematurely that it will turn out that the proposed model is only suitable for dealing with specific types of combinations.

Due to space restrictions it was not possible to carry out a detailed investigation of the occurrence of all collocations of <give rise to> identified in the parallel corpus. Similarly, it was not possible to describe more than one lexical sub-domain which belong the TLD {MANY COLLECTIVES} and {VIELE KOLLEKTIVA}. For this reason, the conclusion regarding the structure of translation lexical domains was based partly on speculation. Further studies would provide more detailed information about the relationship between individual sub-domains, or about the complete number of sub-domains that constitute a domain.

A serious limitation is that we do not know in which language the texts that belong to the Europarl were originally composed. This is why it was not possible to explore reciprocity relations between lexical items. Some future studies based on more reliable corpora might show that some of the tendencies observed in the current thesis have to do with the selected corpus. For example, in the present thesis we observed that <cause problem|difficulty> and <Problem|Difficulty verursachen> served mutually as most frequent translation correspondences. However, a future study might lead to the conclusion that the German item is mostly translated with <cause problem|difficulty> and that the English construction is more often translated as <Problem|Schwierigkeit bereiten>.

Similarly, the Europarl corpus consists of specific text types and contains translations created by a specific translation community. We obviously need more heterogeneous parallel corpora for more representative results.

The pre-defined threshold levels with regard the inclusion of lexical items in a domain or sub-domain might need to be revised. It may turn out that it is too restrictive in some cases which may be lead to the omission of relevant information. For example, in the TLD {CAUSE PROBLEM} the lexical item <cause of problem> was excluded because it did not meet all three criteria. However, it served as the most often selected correspondence for <Ursache ART|APPR Problem> and as such should be considered as a relevant lexical item.

The present thesis was concerned only with the study of textual contexts and no information was provided regarding the distribution of lexical units across registers or genres. However, as indicated in section 8.3.2.3 the model seems suitable for this type of analysis, as well. One potential obstacle for such studies is that we still do not have comparable cross-linguistic classifications of registers and genres. For example, the classification used for the German monolingual corpus DeReKo is not comparable to Lee's classification used in the British National Corpus. Therefore, we need more research work in this area.

In the present study all work has been carried out semi-automatically which was at times very time-consuming. In order to conduct large-scale studies it is necessary to automate the research process to a larger extent. Technical improvement would not only speed up the analysis process but might also lead to further development of the model at the level of detail.

References

Abu-Samak, Z. (1996) **A Study of the Teaching Practices in Teaching Dictionary Exercises Used by Teachers of English in the Basic Stage in Amman Government Schools**. PhD Thesis, University of Jordan.

Aijmer, K. and Altenberg, B. (1996) "Introduction." In: Aijmer, K., Altenberg, B. and Johansson, M. (eds.) **Languages in Contrast. Papers from a Symposium on Text-based Cross-Linguistic Studies**. Lund: Lund University Press, pp. 11-16.

Aliseda, A. (2006) **Abductive Reasoning**. Dordrecht: Springer.

Allen, C.M. (2006) **A Local Grammar of Cause and Effect: A Corpus-driven Study**. PhD Thesis, University of Birmingham.

Altenberg, B. (1999) "Adverbial Connectors in English and Swedish: Semantic and Lexical Correspondences." **Language and Computers**, 26, pp. 249–268.

Altenberg, B. (2002) "Causative Constructions in English and Swedish." In: Altenberg, B. and Granger, S. (eds.) **Lexis in Contrast: Corpus-based Approaches**. Amsterdam: John Benjamins, pp. 97–116.

Anthony, L. (2011) **AntConc (Version 3.2.2)** Tokyo: Waseda University. Available from: <http://www.antlab.sci.waseda.ac.jp/> [Accessed 2 January 2012].

Apresjan, J. (2000) **Systematic Lexicography**. Oxford: Oxford University Press.

Atkins, B.T.S. (1996) "Bilingual Dictionaries: Past, Present and Future." In: Gellerstam, M., Jarborg, J., Malmgren, S-G., Noren, K., Rogstro, L., and Papmehl, C.R. (eds.) **Euralex'96 Proceedings I–II, Papers Submitted to the Seventh EURALEX International Congress on Lexicography in Goteborg, Sweden**. Gothenburg: Department of Swedish, Gothenburg University.

Atkins, B.T.S. and Knowles, F.E. (1990) "Interim Report on the EURALEX/AILA Research Project into Dictionary Use." In **Proceedings of BudaLex**, 88, pp. 381–392.

Atkins, B.T.S., Rundell, M. and Sato, H. (2003) "The Contribution of Framenet to Practical Lexicography." **International Journal of Lexicography**, 16 (3), pp. 333–357.

Atkins, B.T.S. and Rundell, M. (2008) **The Oxford Guide to Practical Lexicography**. Oxford: University Press.

Baayen, R.H. (2008) **Analyzing Linguistic Data**. Cambridge: Cambridge University Press.

Baker, C.F. (2012) "FrameNet, Current Collaborations and Future Goals." **Language Resources and Evaluation***,* 46 (2), pp. 269–286.

Baker, C.F., Fillmore, C.J. and Cronin, B. (2003) "The Structure of the FrameNet Database." **International Journal of Lexicography**, 16(3), pp. 281–296.

Baker, M. (1993) "Corpus Linguistics and Translation Studies: Implications and Applications." In: Baker, M., Francis G., and Tognini-Bonelli, E. (eds.) **Text and Technology: In Honour of John Sinclair**. Amsterdam: John Benjamins, pp. 233-250.

Baker, M. (1998) **Routledge Encyclopedia of Translation Studies**. London: Routledge.

Baker, G. and Hacker, P.M.S. (2009) **Wittgenstein: Rules, Grammar and Necessity: Volume 2 of an Analytical Commentary on the Philosophical Investigations, Essays and Exegesis 185-242**. Oxford: Blackwell.

Barlow, M. (2002) "ParaConc: Concordance software for multilingual parallel corpora**." Proceedings of the Third International Conference on Language Resources and Evaluation. Workshop on Language Resources in Translation Work and Research**, pp. 20-24.

Barlow, M. (2008) "Parallel Texts and Corpus-based Contrastive Analysis." In: Gómez González, M., Mackenzie, L., González Alvarez, E. (eds.) **Current Trend in Contrastive Linguistics**. Amsterdam: John Benjamins, pp. 101-121.

Barnbrook, G. (2002) **Defining Language: A Local Grammar of Definition sentences**. Amsterdam: John Benjamins.

Barnbrook, G and Sinclair, J.M. (1995) "Parsing Cobuild Entries" In: Sinclair, J.M., Hoelter, M. and Peters, C. (eds.) **The Languages of Definition: The Formalisms of Dictionary Definitions for Natural Language Processing. Studies in Machine Translation and Natural Language Processing**. Luxembourg: European Commission, pp. 13–58.

Barnbrook, G. and Sinclair, J.M. (2001) "Specialised Corpus, Local and Functional Grammars." In: Ghadessy, M., Henry, A. and Roseberry, R.L. (eds.) **Small Corpus Studies and ELT: Theory and Practice**. Amsterdam: John Benjamins, pp. 237–276

Baroni, M. and Bernardini, S. (2006) "A New Approach to the Study of Translationese: Machine-Learning the difference between original and translated Text." **Literary and Linguistic Computing**, 21(3), pp. 259–274.

Barsalou, L.W. (1992) "Frames, Concepts, and Conceptual Fields." In: Lehrer, A. and Kittay E.F. (eds.) **Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization**. London: Routledge, pp. 21-74.

Baxter, J. (1980) "The Dictionary and Vocabulary Behavior: A Single Word or a Handful?" **Tesol Quarterly**, 14, pp. 325–336.

Bendix, E.H. (1971) "The Data of Semantic Description." In: Steinberg, D.S. and Jakobovits, L.A. (eds.) **Semantics: An interdisciplinary Reader in Philosophy, Linguistics, and Psychology**. Cambridge: Cambridge: University Press, pp. 393–409.

Berger, J.O. (1985) **Statistical Decision Theory and Bayesian Analysis**. New York: Springer.

Berlin, B. and Kay, P. (1969) **Basic Color Terms: Their Universality and Evolution**. Berkley: University of California Press.

Bernardini, S. and Zanettin, F. (2004) "When is a Universal not a Universal." In: Mauranen, A. and Kujamäki, P. (eds.) **Translation Universals. Do they exist?** Amsterdam: John Benhamins Publishing, pp. 51–62

Boas, H.C. (2002) "Bilingual FrameNet Dictionaries for Machine Translation." In: González-Rodríguez, M. and Suárez Araujo, C.P. (eds.) **Proceedings of the Third International Conference on Language Resources and Evaluation**. Las Palmas, Spain, IV, pp. 1364-1371.

Boas, H.C. (2005) “Semantic Frames as Interlingual Representations for Multilingual Lexical Databases.” **International Journal of Lexicography**, 18 (4), pp. 445–478.

Braasch, A. (1994) “There’s no Accounting for Taste - Except in Dictionaries.” **Proceedings. Amsterdam: EURALEX**. Amsterdam, pp. 45-55.

Burger, H. (1998) **Phraseologie: eine Einführung am Beispiel des Deutschen**. Berlin: Schmidt.

Busse, D. (1991) **Wortbedeutung und sprachliches Handeln. Überlegungen zu den Grundlagen der Bedeutungstheorie**. Unpublished manuscript.

Butler, C.S. (1985) **Statistics in Linguistics**. Oxford: Blackwell.

Butler, C.S. (2004) “Corpus Studies and Functional Linguistic theories.” **Functions of Language**, 11(2), pp. 147–186

Butler, C.S. (2008) “Three English Adverbs and their Formal Equivalents in Romance Languages A Corpus-based Collocational Study.” **Languages in Contrast**, 8(1), pp. 107–124.

Butzkamm, W. (2004) **Lust zum Lehren, Lust zum Lernen**. Tübingen: Francke.

Butzkamm, W. and Caldwell, J.A. (2009) **The Bilingual Reform a Paradigm Shift in Foreign Language Teaching**. Tübingen: Narr.

Caillieux, M. (1974) “Bemerkungen zum Gebrauch von Regel.” In: Heringer, H.J. (ed.) **Seminar: Der Regelbegriff in der praktischen Semantik**. Frankfurt: Suhrkamp, pp. 25–47.

Carter, R. (1998) **Vocabulary**. London: Routledge.

Catford, J.C. (1965) **A Linguistic Theory of Translation: An Essay in Applied Linguistics**. **Language and Language Learning**. London: Oxford University Press.

Chomsky, N. (1957) **Syntactic Structures**. The Hague and Paris: Mouton.

Church, K.W. and Hanks, P. (1990) “Word Association Norms, Mutual Information, and Lexicography.” **Computational Linguistics**, 16(1), pp. 22–29.

Church K., Gale W., Hanks P., Hindle D., Moon R. (1994) “Lexical Substitutability.” In: Atkins B.T.S. and Zampolli A. **Computational Approaches to the Lexicon**. Oxford: Clarendon Press, pp. 153-177.

Claes, F. (1977) **Bibliographisches Verzeichnis der deutschen Vokabulare und Wörterbücher, gedruckt bis 1600**. Hindelsheim: Georg Olms Verlag.

Cook, G. (2007) “A Thing of the Future: Translation in Language Learning.” **International Journal of Applied Linguistics**, 17(3), pp. 396–401.

Cook, G. (2010) **Translation in Language Teaching: An Argument for Reassessment**. Oxford: Oxford University Press.

Coseriu, E. (1964) “Pour une sémantique diachronique structurale.” **Travaux de linguistique et de littérature**, 1, pp. 139-41.

Coseriu, E. (1968) **Les Structures lexématiques**. Wiesbaden: Franz Steiner Verlag.

Cruse, D. A. (1986) **Lexical Semantics**. Cambridge: Cambridge University Press.

Curran, J.R. (2004) **From Distributional to Semantic Similarity**. PhD thesis, University of Edinburgh.

Desgraupes, B. and Loiseau, S. (2012) “Introduction to the rcqp package.” Available from: <http://cran.r-project.org/web/packages/rcqp/vignettes/rcqp.pdf> [Accessed 25 October 2012]

Dice, L.R. (1945) “Measures of the Amount of Ecologic Association Between Species.” **Ecology**, 36(3), pp. 297–302.

Dixon, R.M.W. (1971) “A Method of Semantic Description.” In: Steinberg, D. and Jakobovits, L. (eds.) **Semantics**. Cambridge: Cambridge University, pp. 436–471.

Durrell, M. (1981) “Contrasting the Lexis of English and German.” In: Russ, C. V. J. (ed.) **Contrastive Aspects of English and German**. Heidelberg: Groos, pp. 35-54.

Durrell, M. (1988) “Some Problems of Contrastive Lexical Semantics.” In: Hüllen, W. and Schulze, R. (eds.) **Understanding the Lexicon: Meaning, Sense and World Knowledge in Lexical Semantics**. Tübingen: Niemeyer, pp. 230-241.

Durrell, M. (2000) **Using German Synonyms**. Cambridge: Cambridge University Press.

Duval, A. (1991) “L’équivalence dans le dictionnaire bilingue.” In: Hausmann, F.J., Reichmann, O., Wiegand, E. and Zgusta, L. (eds.) **Wörterbücher/Dictionaries/Dictionnaires**. **Ein internationales Handbuch zur Lexikographie/An International Encyclopedia of Lexicography/Enciclopédie internationale de lexicographie. 3.** Berlin and New York: De Gruyter, pp. 2817-2824.

Dyvik, H. (1998) “A Translational Basis for Semantics.” **Language and Computers**, 24, pp. 51–86.

Dyvik, H. (2004) “Translations as Semantic Mirrors: from Parallel Corpus to WorldNet.” In: Aijmer, K. and Altenberg, B. (eds.) **Language and Computers, Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23) Göteborg 22-26 May 2002**. Amsterdam: Rodopi, pp. 311-326.

Dyvik, H. (2005) “Translations as a Semantic Knowledge Source.” **Proceedings of the second Baltic Conference on Human Language Technologies**, Tallin, Estland, pp. 27–38.

Ervas, F. (2008) “Davidson’s Notions of Translation Equivalence.” **Journal of Language and Translation**. 9(2), pp. 7–29

Evans, V. and Green, M. (2006) **Cognitive Linguistics: An Introduction**. Edinburgh: Edinburgh University Press.

Evert, S. (2005) **The Statistics of Word Cooccurrences: Word Pairs and Collocations**. PhD Thesis, University of Stuttgart.

Evert, S. and Hardie, A. (2011) “Twenty-first Century Corpus Workbench: Updating a Query Architecture for the New Millennium.” **Corpus Linguistics Conference**. Birmingham, University of Birmgingham.

Ferraresi, A., Zanchetta, E., Baroni, M. and Bernardini, S. (2008) “Introducing and Evaluating UKWAC, a Very Large Web-derived Corpus of English.” **Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can We Beat Google**. Marrakech, Morocco, pp. 47-54.

Fillmore, C.J. (1977) “The Case for Case Reopened.” **Syntax and Semantics**, 8(1977), pp. 59–82.

Fillmore, C.J. and Atkins, B.T. (1992) “Toward a Frame-based Lexicon: The Semantics of RISK and its neighbors” In: Lehrer, A. and Kittay E.F. (eds.) **Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization**. London: Routledge, pp. 75-102.

Fillmore, C.J. and Atkins, B.T.S. (1998) “FrameNet and Lexicographic Relevance.” **Proceedings of the First International Conference on Language Resources and Evaluation**. Granada, Spain, pp. 28-30

Fillmore, C.J., Johnson, C.R. and Petruck, M.R. (2003) “Background to Framenet.” **International Journal of Lexicography**, 16(3), pp. 235–250.

Firth, J.R. (1968) “A synopsis of Linguistic Theory 1930-1955.” In: Palmer, F.R. (ed.) **Selected Papers of J.R. Firth 1952-1959**. London: Longman, pp.168-205.

Francis, G. (1993) “A Corpus-driven Approach to Grammar: Principles, Methods and Examples.” In: Baker, M., Francis, G. and Tognini-Bonelli, E. (eds.) **Text and Technology: In Honour of John Sinclair**. Amsterdam: John Benjamins, pp. 137–156.

Frege, G. (1960) “On the Foundations of Geometry.” **The Philosophical Review**, 69(1), pp. 3–17.

Gadamer, H.G. (1977) **Philosophical Hermeneutics**. Berkeley and Los Angeles: University of California Press.

Gadamer, H.G. (2004) **Truth and Method**. London: Continuum International Publishing Group.

Geeraerts, D. (2010) **Theories of Lexical Semantics**. Oxford and New York: Oxford University Press.

Geeraerts, D. (2003) “Meaning and Definition.” In: van Sterkenburg, P. (ed.) **A Practical Guide to Lexicography**. Amsterdam: Benjamins, pp. 83-93.

Goddard, C. and Thieberger, N. (1997) “Lexicographic Research on Australian Aboriginal Languages, 1969-1993” In: Tryon, D. and Walsh, M. (eds.) **Boundary Rider: Essays in Honour of Geoffrey O’Grady**. Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University, pp. 175-208.

Gómez-González, M. Á. (2001) **The Theme-Topic Interface: Evidence from English**. Amsterdam: John Benjamins Publishing.

Gommlich, K. (1997) “To Ban or Not to Ban: The Translation Syndrome in Second Language Acquisition.” In: Wotjak, G. and Schmidt, H. (eds.) **Modelle der Translation, Models of Translation: Festschrifi fur Albrecht Neubert**. Frankfurt am Main: Vervuert Verlag, pp. 171-191.

Gordon, W.T. (2003) “Semantic Theories in 20th-century America: An Overview of Approaches Outside Generative Grammar” In: Wiegand, H.E. (ed.) **History of the Language Sciences**. Berlin: de Gruyter, pp. 2213-2229.

Granger, S. (1996) “From CA to CIA and Back: An Integrated Approach to Computerized Bilingual Corpora and Learner Corpora.” In Aijmer, K., Altenberg, B. and Johansson, M. (eds.) **Languages in Contrast. Papers from a Symposium on Text-based Cross-Linguistic Studies**. Lund: Lund University Press, pp. 37-51.

Greenberg, J.H. (1966) **Language Universals: With Special Reference to Feature Hierarchies**. The Hague: Mouton.

Groom, N.W. (2007) **Phraseology and Epistemology in Humanities Writing: A Corpus-driven Study**. PhD Thesis, University of Birmingham.

Gross, M. (1993) "Local Grammars and their Representation by Finite Automata." In: Hoey, M. (ed.) **Data, Description, Discourse. Papers on the English Language in Honour of John McH Sinclair**. London: Collins, pp. 26–38.

Grosz, B.J. (1982) "Discourse Analysis." In: Kittredge, R. and Lehrberger, J. (eds.) **Sublanguage: Studies of Language in Restricted Semantic Domains**. Berlin and New York: Mouton de Gruyter, pp. 138–174.

Hahn, M. (2002) **Die Synonymenlexikographie vom 16. bis zum 20. Jahrhundert: historische Entwicklung und kommentierte Bio-Bibliographie**. Heidelberg: C. Winter.

Halliday, M.A.K. (1978) **Language as Social Semiotic: The Social Interpretation of Language and Meaning**. London: Edward Arnold.

Halliday, M.A.K. (1991) "Towards Probabilistic Interpretations." In: Ventola, E. (ed.) **Functional and Systemic Linguistics: Approaches and uses**. Berlin and New York**:** Mouton de Gruyter, pp. 39–61.

Halliday, M. A.K. (1994) **Functional Grammar**. London: Edward Arnold.

Halliday, M.A.K., Teubert, W., Yallop, C. and Čermáková, A. (eds.) **Lexicology and Corpus Linguistics**. London and New York: Continuum.

Hanks, P. (2004) "The Syntagmatics of Metaphor and Idiom." **International Journal of Lexicography**, 17(3), pp. 245–274.

Hanks, P. (2008) "Do Word Meanings Exist?" In: Fontenelle, T. (ed.) (2008) **Practical Lexicography: A Reader**, Oxford: Oxford University Press, pp. 125-135.

Hanks, P. and Pustejovsky, J. (2005) "A Pattern Dictionary for Natural Language Processing." **Revue Française de linguistique appliquée**, 10(2), pp. 63–82.

Hannay, M. (2003) "Types of Bilingual Dictionaries." In: van Sterkenburg, P. (ed.) **A Practical Guide to Lexicography**. Philadelphia: John Benjamins Publishing Company.

Harris, R. (2005) **The Semantics of Science**. London: Continuum International Publishing Group.

Harris, Z.S. (1982) "Discourse and Sublanguage." In: Kittredge, R. and Lehrberger, J. (eds.) **Sublanguage: Studies of Language in Restricted Semantic Domains**. Berlin and New York: Mouton de Gruyter, pp. 231–236.

Harris, Z.S. (1952) "Discourse Analysis." **Language**, 28(1), pp. 1–30.

Harris, Z.S. (1954) "Distributional Structure." **Word**, 10(2/3), pp. 146-162.

Harris, Z.S. (1968) **Mathematical Structures of Language**. New York: Wiley.

Harris, Z. S. (1970) **Papers in Structural and Transformational Linguistics**. Dordrecht: Reidel.

Harris, Z. S. (1988) **Language and information**. New York: Columbia University Press.

Harris, Z.S., Gottfried, M., Ryckman, T., Mattick Jr. P. Daladier, A., Harris, T.N., and Harris, S. (1989) **The Form of Information in Science: Analysis of an Immunology Sublanguage. Boston Studies in the Philosophy of Science, 104**. Dordrecht and Boston: Kluwer Academic Publishers.

Hartmann, R.R.K. (1983) “The Bilingual Learner’s Dictionary and its Uses.” **Multilingual Journal of Cross-Cultural and Interlanguage Communication**, 2(4), pp. 195–202.

Hartmann, R.R.K. (2007) **Interlingual Lexicography: Selected Essays on Translation Equivalence, Constrative Linguistics and the Bilingual Dictionary**. Berlin and Walter de Gruyter.

Hasselgard, H. (2004) “Thematic Choice in English and Norwegian.” **Functions of Language**, 11(2), pp. 187–212.

Heid, U. (1994) “Relating Lexicon and Corpus: Computational Support for Corpus-based Lexicon Building in DELIS.” In: Martin, W., Meijs, W., Moerland, M., ten Pas, E., van Sterkenburg, P. and Vossen, P. (eds.) **EURALEX ‘94 Proceedings**. Amsterdam: Vrije Universiteit, pp. 459-471.

Heid, U. (1996) “Creating a Multilingual Data Collection for Bilingual Lexicography from Parallel Monolingual Lexicons.” **Euralex'96 Proceedings**, pp. 573-590

Heringer, H.J. (1977) **Einführung in die praktische Semantik**. Heidelberg: Quelle and Meyer.

Hjelmslev, L. (1961) **Prolegomena to a Theory of Language**. Madison: Univ of Wisconsin.

Hoey, M. (1991) **Patterns of Lexis in Text**. Oxford: Oxford University Press.

Hüllen, W. (1999) **English Dictionaries, 800-1700: The Topical Tradition**. Oxford: Oxford University Press.

Hüllen, W. (2004) **A History of Roget’s Thesaurus: Origins, Development, and Design**. Oxford: Oxford University Press.

Hüllen, W. (2009) **Networks and knowledge in Roget's thesaurus**. Oxford: Oxford University Press.

Hunston, S. and Sinclair, J. M. (2000) “A Local Grammar of Evaluation.” In: Hunston S. and Thomson, G. (eds.) **Evaluation in Text: Authorial Stance and the Construction of Discourse**. Oxford: Oxford University Press, pp. 74–101.

Hyland, K. (2008) “As Can Be Seen: Lexical Bundles and Disciplinary Variation.” **English for Specific Purposes**, 27(1), pp. 4–21.

Johansson, S. (2003) “Contrastive Linguistics and Corpora.” In: Granger, S., Lerot, J. and Petch-Tyson, S. (eds.) **Corpus-based Approaches to Contrastive Linguistics and Translation Studies**. Amsterdam and New York: Rodopi, pp. 31–45.

Johansson, S. (2007) **Seeing Through Multilingual Corpora: On the Use of Corpora in Contrastive Studies**. Amsterdam: John Benjamins Publishing.

Karcher, G.L. (1979) **Kontrastive Untersuchung von Wortfeldern im Deutschen und Englischen**. Frankfurt am Main: Lang.

Katz, J.J. and Fodor, J.A. (1963) “The Structure of a Semantic Theory.” **Language**, 39(2), pp. 170–210.

Kay, C., Roberts, J., Samuels, M. and Wotherspoon, I. (2009) **Historical thesaurus of the Oxford English dictionary**. Oxford: Oxford University Press.

Keller, R. (1974) “Zum Begriff der Regel.” In: Heringer, H.J. (ed.) **Seminar: Der Regelbegriff in der praktischen Semantik**. Frankfurt: Suhrkamp, pp. 10-24.

Kilgarriff, A. and Rundell, M. (2002) “Lexical Profiling Software and its Lexicographic Applications: A Case Study.” In: Braasch, A. and Povlsen, C. (eds.) **Proceedings of the Tenth EURALEX International Congress**. Copenhagen: Center for Sprogteknologi, pp. 807-818.

Kilgarriff, A., Rychlỳ, P., Smrz, P., and Tugwell, D. (2004) “The Sketch Engine.” In: Williams, G. and Vessier, S. (eds.) **Proceedings of the Eleventh Euralex Congress**, UBS Lorient, France, pp. 105-116.

Kilgarriff, A. and Kosem, K. (2012) “Corpus Tools for Lexicographers.” In: Granger, S. and Paquot, M. (eds.) **Electronic Lexicography**. Oxford: Oxford University Press, pp. 31-57

Kittredge, R. (1982) “Variation and Homogeneity of Sublanguages.” In: Kittredge, R. and Lehrberger, J. (eds.) **Sublanguage: Studies of Language in Restricted Semantic Domains**. Berlin and New York: Mouton de Gruyter, pp. 107–137.

Kittredge, R. (1987) “The Significance of Sublanguage for Automatic Translation.” In: Nirenburg, S. (ed.) **Machine Translation: Theoretical and Methodological Issues**. Cambridge: Cambridge University Press, pp. 59-67.

Klöpfer, R. (1967) **Die Theorie der literarischen Übersetzung: Romanisch-deutscher Sprachbereich**. München: Wilhelm Fink.

Koehn, P. (2005) “A Parallel Corpus for Statistical Machine Translation.” **Proceedings of MT Summit X**. Phuket, Thailand, pp. 79-86.

Koller, W. (1978) “Äquivalenz in kontrastiver Linguistik und Übersetzungswissenschaft.” In: Grähs, L., Korlén, G. and Malmberg, B. (eds.) **Theory and Practice of Translation**. Bern, Frankfurt, Las Vegas, pp. 69–92.

Kühn, P. (1985) “Gegenwartsbezogene Synonymenwörterbücher des Deutschen: Konzept und Aufbau.” **Lexicographica**, 1, pp. 51–82.

Lee, D.Y. (2001) “Genres, Eegisters, Text Types, Domains and Styles: Clarifying the Concepts and nevigating a path through the BNC jungle.” **Language Learning and Technology**, 5(3): 37-72

Lehmann, E.L. (1950) “Some Principles of the Theory of testing hypotheses.” **The Annals of Mathematical Statistics**, 21(1), pp. 1–26.

Lehrberger, J. (1982) “Automatic Translation and the Concept of Sublanguage”, In: Kittredge, R. and Lehrberger, J. (eds.) **Sublanguage: Studies of Language in Restricted Semantic Domains**. Berlin and Walter de Gruyter, pp. 81-107.

Lehrer, A. (1974) **Semantic fields and Lexical**. Amsterdam: North-Holland.

Levy, J. (1967) “Translation as a Decision Process.” In: **To Honor Roman Jakobson, Vol. 2**. The Hague: Mouton, pp. 1171–1182.

Lewis, M. (1993) **The Lexical Approach** (Vol. 1). Hove, UK: Language Teaching Publications.

Lyons, J. (1963) **Structural Semantics. An Analysis of Part of the Vocabulary of Plato**. Basil Blackwell.

Lyons, J. (1968) **Introduction to Theoretical Linguistics**. Cambridge: Cambridge University Press.

Lyons, J. (1977) **Semantics (Vols I & II)**. Cambridge**:** Cambridge University Press.

Lyons, J. (1981) **Language and Linguistics: An Introduction**. Cambridge: Cambridge University Press.

Lyons, J. (1995) **Linguistic Semantics: An Introduction**. Cambridge: Cambridge University Press.

Mahlberg, M. (2005) **English General Nouns: A Corpus Theoretical Approach**. Amsterdam: John Benjamins Publishing.

Manning, C.D.A. and Schütze, H. (1999) **Foundations of Statistical Natural Language Processing**. Cambridge Mass: MIT Press.

Martin, S.E. (1967) "Selection and Presentation of Ready Equivalents in a Translation Dictionary." In: Householder, F.W. and Saporta, S. (eds.) **Problems in Lexicography**. Bloomington: Indiana University, pp.153-159.

Mauranen, A. and Kujamäki, P. (2004) **Translation Universals: Do They Exist?** Amsterdam: John Benjamins Publishing.

McArthur, T. (1981) **Longman Lexicon of Contemporary English**. Harlow: Longman Harlow.

McArthur, T. (1986) **Worlds of reference: Lexicography, Learning and Language from the clay tablet to the computer**. Cambridge: Cambridge University Press.

McArthur, T. (1998) **Living Words: Language, Lexicography and the Knowledge Revolution**. Exeter: University of Exeter Press.

Mendelson, J. (1979) "The Habermas-Gadamer Debate" **New German Critique**, 18: 44-73

Meyer, C.F. (2002) **English Corpus Linguistics: An Introduction**. Cambridge: Cambridge University Press.

Miller, G. A. and Johnson-Laird, P. N. (1976) **Language and Perception**. Cambridge: Mass, Belknap Press of Harvard University Press.

Milton, J. (2009) **Measuring Second Language Vocabulary Acquisition**. Bristol: Multilingual Matters.

Moon, R. (2009) "The Cobuild Project." In: Cowie, A.P. **The Oxford History of English Lexicography: Volume I: General-Purpose Dictionaries; Volume II: Specialized Dictionaries: Two-volume Set**. Oxford: Clarendon Press, 2, 436-458.

Moskovich, W. (1982) "What is a Sublanguage? The Notion of Sublanguage in Modern Soviet Linguistics." In: Kittredge, R. and Lehrberger, J. (eds.) **Sublanguage: Studies of Language in Restricted Semantic Domains**. Berlin and New York: Mouton de Gruyter, pp. 191–205.

Munday, J. (2001) **Introducing Translation Studies: Theories and Applications**. London: Routledge.

Nation, I.S.P. (2001) **Learning Vocabulary in Another Language**. Cambridge: Cambridge University Press.

Nesselhauf, N. (2005) **Collocations in a Learner Corpus**. Amsterdam: John Benjamins.

Nord, B. (2002) **Hilfsmittel beim Übersetzen: Eine empirische Studie zum Rechercheverhalten professioneller Übersetzer**. Frankfurt am Main: Lang.

Peck, R., Olsen, C. and Devore, J.L. (2008) **Introduction to Statistics and Data Analysis**. Belmont: Cengage Learning.

Penco, C. (2004) "Wittgenstein, Locality and Rules" In: Picardi, E. and Coliva, A. (eds.) **Wittgenstein Today**, Padova: Poligrafo, pp. 249-274.

Penco, C. and Vignolo, M. (2005) “Converging Towards What? Pragmatic and Semantic Competence.” In: Bouquet, P. and Serafini, L. (eds.) **Context Representation and Reasoning (Vol. 136). CEUR-WS**. Available from: <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol> [Accessed 5 June 2011].

Porzig, W. (1934) “Wesenhafte Bedeutungsbeziehungen.” **Beiträge zur Geschichte der deutschen Sprache und Literatur (PBB)**. 134(58), pp. 70–97.

Post, M. (1988) “Scenes-and-Frames Semantics as a neo-Lexical Field Theory.” In: Hüllen, W. and Schulze, R. (eds.) **Understanding the Lexicon. Meaning, Sense and World-knowledge in Lexical Semantics**. Tübingen: Niemeyer, pp. 36–47.

Raiffa, H. and Schlaifer, R. (1968) **Applied Statistical Decision Theory**. Boston: Harvard University.

Reck, E. (1997) “Frege’s Influence on Wittgenstein: Reversing Metaphysics via the Context principle.” Tait, W. W. and Linsky, L. (eds.) **Early Analytic Philosophy: Frege, Russell, Wittgenstein**. Chicago: Open Court Publishing, pp. 123–85.

Roberts, R.P. and Montgomery, C. (1996) “The Use of Corpora in Bilingual Lexicography.” In: Gellerstam, M., Järborg, J., Malmgren, S.-G., Norén, K., Rogström, L. and Papmehl, C. R. (1996) (eds.). **EURALEX’96 Proceedings**. Gothenburg: Gothenburg University, pp. 457-464.

Rundell, M. and Killgarrif, A. (2011) “Automating the Creation of Dictionaries.” In: Meunier, F. and Granger, S. (eds.) **A Taste for Corpora: In Honour of Sylviane Granger**. Amsterdam: John Benjamins, pp. 257-283.

Rychlỳ, P. (2008) “A Lexicographer-friendly Association Acore.” **Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN**, Brno, Czech Republlic, Masaryk University, pp. 6–9.

Sahlgren, M. (2008) “The Distributional Hypothesis.” **Italian Journal of Linguistics**, 20(1), 33–54.

Salkie, R. (1995) “INTERSECT: A Parallel Corpus Project at Brighton University.” **Computers and Texts**, 9, pp. 4–5.

Salkie, R. (2008) “How can Lexicographers Use a Translation Corpus.” In: Xiao, R. He, L and Yue, M. (eds.) **Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies**. Zhejiang University, Hangzhou. Available from: <http://www.lancs.ac.uk/fass/projects/Corpus/UCCTS2008Proceedings/papers/Salkie.pdf> [Accessed 12 April 2010].

Sauer, H. (2008) “Glosses, Glossaries, and Dictionaries in the Medieval Period.” In: A.P. Cowie, **The Oxford History of English Lexicography: Volume I: General-Purpose Dictionaries; Volume II: Specialized Dictionaries: Two-volume Set**. Oxford: Clarendon Press, pp. 17-41.

Schütze, H. (1998) “Automatic Word Sense Discrimination.” **Computational Linguistics**, 24(1), pp. 97–123.

Scott, M. (2008) **WordSmith Tools Version 5,** Liverpool: Lexical Analysis Software.

Siepmann, D. (2005) “Collocation, Colligation and Encoding Dictionaries. Part I: Lexicological Aspects.” **International Journal of Lexicography**, 18(4), pp. 409–443.

Sierra, G. and McNaught, J. (2000) “Extracting Semantic Clusters from MRDs for an Onomasiological Search Dictionary.” **International Journal of Lexicography**, 13(4), pp. 264–286.

Sinclair, J.M. (1966) “Beginning the Study of Lexis.” In: Bazel, C. E., Catford, J. C. and Halliday, M. A. K. (eds.) **In memory of JR Firth**. London: Longmans, pp. 410–430.

Sinclair, J.M. (ed.) (1987) **Looking up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary**. London: Collins ELT.

Sinclair, J.M. (1990) **COBUILD English Grammar**. London: HarperCollins Publishers.

Sinclair, J.M. (1991) **Corpus, Concordance, Collocation**. Oxford: Oxford University Press.

Sinclair, J.M. (1994) “Trust the Text“. In: Coulthard, M. (ed.) **Advances in Written Text Analysis**. London: Routledge, pp. 12-25.

Sinclair, J.M. (1996a) “An International Project in Multilingual Lexicography.” **International Journal of Lexicography**, 9 (3), pp. 179–196.

Sinclair, J.M. (1996b) “Corpus to Corpus: A Study of Translation Equivalence.” **International Journal of Lexicography**, 9 (3): 171–178.

Sinclair, J.M. (2004) **Trust the Text: Language, Corpus and Discourse**. London: Routledge.

Sinclair, J.M., Jones, S. and Daley, R. (1970) **English Collocation Studies: The OSTI Report**. (Krishnamurthy, R. ed. 2004) London: Continuum.

Snell-Hornby, M. (1984) “The Bilingual Dictionary: Help or Hindrance.” **LEXeter’83 Proceedings**. Tubingen: Max Niemeyer Verlag, pp. 274–281.

Snell-Hornby, M. (1987) “Towards a learners' Bilingual Dictionary.” In: Cowie, A.P. (ed.) **The dictionary and the Language learner**. Tubingen: Max Niemeyer Verlag, pp. 159-170.

Snell-Hornby, M. (1988) **Translation Studies: An Integrated Approach**. Amsterdam: John Benjamins Publishing.

Snell-Hornby, M. (1990) “Dynamics in Meaning as a Problem for Bilingual Lexicography.” In: Tomaszczyk, J. (ed.) **Meaning and Lexicography**. Amsterdam: John Benjamins, pp. 209-225.

Snell-Hornby, M. (1996) “The Translator’s Dictionary - An Academic Dream.” In: Kadric, M. (ed.) **Translation und Text. Ausgewählte Vorträge**. Wien: WUV–Universitätsverlag, pp. 90–96.

Stubbs, M. (1986) “Language Development, Lexical Competence and Nuclear Vocabulary.” In: Durkin, K. (ed.) **Language Development in the School Years**. Croom Helm, pp. 34–56.

Stubbs, M. (2001) **Words and Phrases: Corpus Studies of Lexical Semantics**. Oxford; Malden, MA: Blackwell Publishers.

Svartvik, J. (1999) “Corpora and Dictionaries.” In: Herbst, T. and Kerstin, P. (eds.) **The perfect learners’ dictionary**. Tübingen: Niemeyer, pp. 283–294.

Teubert, W. (1996) “Comparable or Parallel Corpora?” **International Journal of Lexicography**, 9(3), pp. 238–264.

Teubert, W. (2001) “Corpus Linguistics and Lexicography.” **International Journal of Corpus Linguistics**. 6, pp. 125–53.

Teubert, W. (2002) “Corpus-based Bilingual Lexicography: The Role of Parallel Corpora in Translation and Multilingual Lexicography.” In: Altenberg, B. and Granger, S. (eds.) **Lexis in Contrast**. Amstardam: John Benjamins, pp. 189-215.

Teubert, W. (2004) “Language and Corpus Linguistics.” In: Halliday, M.A.K., Teubert, W., Yallop, C. and Čermáková, A. (eds.) **Lexicology and Corpus Linguistics**. London and New York: Continuum, pp. 73–112.

Teubert, W. (2010) **Meaning, Discourse and Society**. Cambridge: Cambridge University Press.

Thomson, E.A., White, P.R. and Kitley, P. (2008) “‘Objectivity’ and ‘Hard News’ Reporting across Cultures.” **Journalism Studies**, 9(2), pp. 212–228.

Tognini-Bonelli, E. (1996) “Towards Translation Equivalence from a Corpus Linguistics Perspective.” **International Journal of Lexicography**, 9(3), pp. 197–217.

Tognini-Bonelli, E. (2001) **Corpus Linguistics at Work**. Amsterdam: John Benjamins.

Tognini-Bonelli, E. (2002) “Functionally Complete Units of Meaning Across English and Italian: Towards a Corpus-driven Approach.” In: Altenberg, B. and Granger, S. (eds.) **Lexis in Contrast. Corpus-based Approaches**. Amsterdam: John Benjamins, pp. 73–95.

Tomaszczyk, J. (1979) “Dictionaries: Users and Uses.” **Glottodidactica**, 12, pp. 103–119.

Trier, J. (1931) **Der deutsche Wortschatz im Sinnbezirk des Verstandes**. Heidelberg: C. Winter.

Van Roey, J. (1990) **French-English Contrastive Lexicology: An Introduction**. Leuven: Peeters Publishers.

Viberg, Â. (1983) “The Verbs of Perception: A Typological Study.” **Linguistics** 21(1), pp. 123-62.

Viberg, Å. (1993) “CrossLinguistic perspectives on Lexical Organization and Lexical Progression.” In: Hyltenstam, K. (ed.) **Progression and Regression in Language: Sociocultural, Neuropsychological, and Linguistic Perspectives**, Cambridge: Cambridge University Press. pp. 340-383.

Viberg, Å. (2002) “Polysemy and Disambiguation Cues Across Languages.” In: Altenberg, B. and Granger, S. (eds.) **Lexis in Contrast. Corpus-based Approaches.** Amsterdam: John Benjamins, pp. 119–50.

Viberg, Å. (2004) “Physical Contact Verbs in English and Swedish from the Perspective of CrossLinguistic Lexicology.” **Language and Computers**, 49(1), pp. 327–352.

Viberg, Å. (2005) “The Lexical Typological Profile of Swedish Mental Verbs.” **Languages in contrast**, 5(1), pp. 121–157.

Viberg, Å. (2008) “RIDING, DRIVING and TRAVELING. Swedish Verbs Describing Motion in a Vehicle in CrosssLinguistic Perspective.” In: Nivre, J., Dahllöf, M. and Megyesi, B. (eds.) **Resourceful Language Technology. Festschrift in Honor of Anna Sågvall Hein** [Acta Universitatis Upsaliensis. Studia Linguistica Upsaliensia 7] Uppsala: Uppsala University, pp. 173-201.

Viberg, Å. (2010) “Basic Verbs of Possession.” In: Lemmens, M. (ed.) **Unison in Multiplicity: Cognitive and Typological Perspectives on Grammar and Lexis**. CogniTextes. Revue de

l'Association française de linguistique cognitive, (Volume 4). Available from: <http://cogniTextes.revues.org/308>[Accessed 21 September 2010].

Waismann, F. (1965) **The Principles of Linguistic Philosophy**. London: Macmillan.

Walker, C. (2008) "Factors which Influence the Process of Collocation." In: Boers, F. and Lindstromberg, S. (eds.) **Cognitive Linguistic Approaches to Teaching Vocabulary**. Berlin and New York: Mouton de Gruyter, pp. 291-308.

Wierzbicka, A. (1987) **English Speech Act verbs: A Semantic Dictionary**. Sydney: Academic Press.

Willis, D. (1990) **The Lexical Syllabus: A New Approach to Language Learning**. London: Collins ELT.

Willis, D. (2003) **Rules, Patterns and Words**. Cambridge: Cambridge University Press.

Witte, A., Harden, T., and de Oliveira Harden, A. R. (eds.) (2009) **Translation in second Language Learning and Teaching**. Oxford and Wien: Peter Lang.

Wittgenstein, L. (1922) **Tractatus Logico-Philosophicus**. London: Kegan Paul.

Wittgenstein, L. (1953) **Philosophical Investigations**. Oxford: Basil Blackwell.

Wittgenstein, L. (1956) **Remarks on the Foundations of Mathematics**. Oxford: Basil Blackwell.

Wittgenstein, L. (1958) **The Blue and Brown Books: Preliminary Studies for the philosophical investigation**. New York: Harper Colophon Books.

Wittgenstein, L. (1967) **Zettel**. Berkley: University of California Press.

Wittgenstein, L. (1980) **Remarks on the Philosophy of Psychology**. Chicago: University of Chicago Press.

Yallop, C. (2004) "Words and Meaning." In: Halliday, M.A.K., Teubert, W., Yallop, C. and Čermáková, A. (eds.) **Lexicology and Corpus Linguistics**. London and New York: Continuum, pp. 23-71.

Yong, H. and Peng, J. (2007) **Bilingual Lexicography from a Communicative Perspective**. Amsterdam: John Benjamins Publishing.

Zgusta, L. (1971) **Manual of Lexicography**. Prague: Academia.

Zgusta, L. (2006) **Lexicography Then and Now: Selected Essays**. Lexicographica. Series maior, 129. Tübingen: Max Niemeyer.

Zojer, H. (2009) "The Methodological Potential of Translation in Second Language Acquisition: Re-evaluating Translation as a Teaching tool" In: Witte, A., Harden, T., and de Oliveira Harden, A.R. (eds.) **Translation in Second Language Learning and Teaching**. Oxford and Wien: Peter Lang, pp. 31–52.

Dictionaries used in the present thesis:

Dornseiff, F. (2004) **Der deutsche wortschatz nach sachgruppen**. Berlin and Walter de Gruyter.

Roget, P.M. (1852) **Roget's Thesaurus of English Words and Phrases**. London: Longman, Brown, Green, and Longmans.

Rundell, M. (ed.) (2007) **Macmillan English Dictionary for Advanced Learners**. Oxford: Macmillan Education.

Sanders, D. (1873) **Deutscher Sprachschatz geordnet nach Begriffen zur leichten Auffindung und Auswahl des passneden Ausdrucks**. Tübingen: Niemeyer.

Sinclair, J. M. (1995) **Collins COBUILD advanced learner's English dictionary**. London: Harper Collins.

Summers, D. (1993) **Longman Language Activator: The World's First Production Dictionary**. London: Harlow.

Terrell, P., Schnorr, V., Morris, W. V., and Breitsprecher, R. (1999) **Collins German-English, English-German Dictionary**. Stuttgart: Klett.

Oxford English Dictionary http://www.oed.com/

Corpora used in the present thesis:

British National Corpus

DeWaC German Web Corpus

English-German/German-English Europarl Parallel Corpus

UKWaC British English web Corpus

OpenSubtitles corpus

I) Appendix A

Table A1: English lexical items from the TLD {CAUSE PROBLEM} and their German translation correspondences

English as the source language		
<create problem|difficulty> 456: <Problem|Schwierigkeit schaffen> 107, <zu Problem|Schwierigkeit führen> 58, <Problem|Schwierigkeit bringen> 46, <Problem|Schwierigkeit verursachen> 41, <Problem|Schwierigkeit bereiten> 23, <Problem|Schwierigkeit entstehen> 16, <Problem|Schwierigkeit aufwerfen> 13, <Problem|Schwierigkeit auftreten 7, <problematisch sein> 7, <Problem|Schwierigkeit darstellen> 5, <es gibt <Problem|Schwierigkeit> 4, <vor Problem|Schwierigkeit stellen> 3, <Problem zur Folge haben> 3.	**<cause Problem|Schwierigkeit 569:** <Problem|Schwierigkeit verursachen> 110, <zu Problem|Schwierigkeiten führen> 79, <Problem|Schwierigkeit bereiten> 62, <Problem|Schwierigkeit bringen 61, <Problem|Schwierigkeit aufwerfen> 27, <problematisch sein>24, <Probleme darstellen> 16, <Problem|Schwierigkeit schaffen> 15, <Problem|Schwierigkeit hervorrufen> 14, <Problem|Schwierigkeit ergeben> 14, <Problem|Schwierigkeit entstehen> 10, <vor Problem|Schwierigkeit stellen> 9, <Problem|Schwierigkeit auftreten> 8, <es gibt Problem|Schwierigkeit> 7, <Ursache für die Probleme> 3, <Ursache der Probleme> 3.	**<pose problem|difficulty> 220:** <Problem|Schwierigkeit aufwerfen> 32, <Problem|Schwierigkeit darstellen> 37, <Problem|Schwierigkeit bereiten> 22, <problematisch sein> 18, <Problem|Schwierigkeit bringen 15, <vor Problem|Schwierigkeit stellen> 12, <zu Problem|Schwierigkeit führen> 12, <vor Problem|Schwierigkeit stellen> 8, <Problem| Schwierigkeit schaffen> 7, <es gibt Problem|Schwierigkeit 7>, <Problem|Schwierigkeit ergeben> 5, <Problem|Schwierigkeit stellen sich> 4, <Problem|Schwierigkeit verursachen> 3, <Problem|Schwierigkeit auftreten> 2, <stellt sich Problem> 2.
<present problem|difficulty> 125: <Problem|Schwierigkeit darstellen> 20, <vor Problem|Schwierigkeit stellen> 12, <Problem|Schwierigkeit bereiten> 12, <Problem|Schwierigkeit bringen> 12, <Problem|Schwierigkeit aufwerfen> 11, <Problem|Schwierigkeit verursachen> 8, <Problem|Schwierigkeit sich ergeben> 6, <problematisch sein> 5, <es gibt Problem|Schwierigkeit> 5, <mit Problem verbunden sein> 2, <Problem|Schwierigkeit aufweisen> 2, <stellt sich Problem> 3	**<problem|difficulty arise>688:** <Problem|Schwierigkeit entstehen> 92, <Problem|Schwierigkeit auftreten> 90, <Problem|Schwierigkeit sich ergeben> 63, <Problem|Schwierigkeit sich stellen> 46, <es gibt Problem|Schwierigkeit> 27, <Problem|Schwierigkeit auftreten> 12, <Problem|Schwierigkeit schaffen> 10, <Problem|Schwierigkeit bringen> 8, <problematisch sein> 9 , <Problem|Schwierigkeit darstellen 5, <zu Problem|Schwierigkeit führen> 3, <Problem|Schwierigkeit	**<give rise to problem|difficulty> 79:** <zu Problem|Schwierigkeiten führen> 17, <Problem|Schwierigkeit entstehen> 10, <Problem|Schwierigkeiten aufwerfen> 6, <mit <Problem|Schwierigkeiten verbunden sein> 5, <es gibt Problem|Schwierigkeit> 5, <Problem|Schwierigkeit 4, <Problem|Schwierigkeit schaffen> 3, <Problem|Schwierigkeit verursachen> 3, <Probleme|Schwierigkeit sich ergeben>3 , <Problem|Schwierigkeiten bereiten> 2,

Table A2: German lexical items from the TLD {PROBLEM BEREITEN} and their English translation correspondences

German as the source language			
<zu Problem|Schwierigkeit führen> 322: <lead to problem|difficulty> 86, <cause problem|difficulty> 60, <create problem|difficulty> 30, <give rise to problem|difficulty> 13, <pose problem|difficulty> 12, <raise problem|difficulty> 9, <there be problem|difficulty> 9, <result in problem|difficulty> 7, <present problem|difficulty> 5, <to be problematic> 5.	**<Problem|Schwierigkeit schaffen> 212:** <create problem|difficulty> 128, <cause problem|difficulty> 25 <raise problem|difficulty> 10 , <lead to problem|difficulty>7, <there be problem> 6, <give rise to problem|difficulty> 5, pose problem|difficulty> 4.	**<Problem|Schwierigkeit verursachen> 271:** <cause problem|difficulty> 135, <create problem|difficulty> 51, <present problem|difficulty> 10, <give rise to problem|difficulty> 8, <raise problem|difficulty> 7, <there be problem 4 <pose problem> 4, <lead to problem|difficulty> 3.	**<Problem|Schwierigkeit (mit sich) bringen> 196:** <cause problem|difficulty> 42, <create problem|difficulty> 34, <bring problem|difficulty> 18, <raise problem|difficulty> 15, <pose problem|difficulty> 11, <present problem|difficulty> 7, <there be problem|difficulty> 7, <lead to problem|difficulty> 6, <problem|difficulty arise> 4.
<Problem|Schwierigkeit> bereiten> 312: <cause problem|difficulty> 86, <create problem|difficulty> 39, <to be difficult> 29, <pose problem|difficulty> 21, <there be problem|difficulty> 14, <have problem|difficulty>12, <present problem|difficulty> 9, <result in problem|difficulty> 8, <raise problem|difficulty> 7, <give rise to problem|difficulty> 3, <problem arise> 3.	**< Problem|Schwierigkeit auftreten> 112:** <problem|difficulty arise> 33, <there be problem|difficulty> 22, <problem|difficulty occur> 8, <create problem|difficulty> 5, <cause problem|difficulty> 4, <present problem|difficulty> 3.	**<Problem|Schwierigkeit sich ergeben> 101:** <problem|difficulty arise>22, <there be problem|difficulty> 14, <cause problem|difficulty> 13 <raise problem|difficulty> 4, <present problem|difficulty> 4, <pose problem|difficulty> 4, <create problem|difficulty> 4, <give rise to problem|difficulty> 4, <have problem|difficulty> 3.	**<Problem|Schwierigkeit entstehen> 133:** <cause problem|difficulty> arise>51, <there be problem|difficulty> 26, <create problem|difficulty> 14, <cause problem|difficulty> 9, <give rise to problem|difficulty> 9, <lead to problem|difficulty> 3
<Problem|Schwierigkeit aufwerfen> 215: <raise problem|difficulty> 76, <pose problem|difficulty 30, <cause problem|difficulty> 27, <create	**<problematisch sein> 297:** <to be problematic> 138, <there be problem|difficulty> 45, <cause problem|difficulty> 35,	**<Problem darstellen> 397:** <to be a problem> 176, <to be an issue> 63, <present problem> 24, <pose problem> 17, <cause problem> 17,	**<es gibt Problem|Schwierigkeit > 1387** <there be problem|difficulty> 738, <have problem> 39,<cause

Table A3: English lexical items from the TLD {MANY COLLECTIVES} and their German translation correspondences

English as a source language			
Lexical items	**Frequency**	**Lexical items**	**Frequency**
<numerous COLLECTIVES>	**2106**	**<many COLLECTIVES>**	**31223**
<zahlreiche KOLLEKTIVA>	1250	<viele KOLLEKTIVA>	24118
<viele KOLLEKTIVA>	484	<zahlreiche KOLLEKTIVA>	4003
<eine Reihe ART|APPR KOLLEKTIVA>	86	<mehrere KOLLEKTIVA>	480
<eine Vielzahl ART|APPR KOLLEKTIVA>	51	<eine Reihe ART|APPR KOLLEKTIVA>	416
<mehrfache KOLLEKTIVA>	45	<eine Vielzahl ART|APPR KOLLEKTIVA>	329
<vielfaeltige KOLLEKTIVA>	26	<eine große Zahl|Anzahl ART|APPR KOLLEKTIVA>	163
<eine große Zahl|Anzahl GEN KOLLEKTIVA>	21	<ein großer Teil GEN KOLLEKTIVA>	112
		<eine Menge GEN KOLLEKTIVA>	133
		<eine beträchtliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	12
		<eine beachtlich Zahl|Anzahl ART|APPR KOLLEKTIVA>	85
		<eine erhebliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	186
<a lot of COLLECTIVES>	**2640**	**<a range of COLLECTIVES>**	**496**
<viele KOLLEKTIVA>	1610	<eine Reihe ART|APPR KOLLEKTIVA>	344
<eine große Zahl|Anzahl ART|APPR KOLLEKTIVA>	320	<viele KOLLEKTIVA>	33
<eine Menge GEN KOLLEKTIVA>	164	<eine große Zahl|Anzahl ART|APPR KOLLEKTIVA>	24
<zahlreiche KOLLEKTIVA>	133	<zahlreiche KOLLEKTIVA>	22
<eine Reihe ART|APPR KOLLEKTIVA>	56	<eine Vielzahl ART|APPR KOLLEKTIVA>	15
<eine Vielzahl ART|APPR KOLLEKTIVA>	43	<eine Palette ART|APPR KOLLEKTIVA>	15
<eine erhebliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	28	<ein Spektrum ART|APPR KOLLEKTIVA>	8
<eine beträchtliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	18	<eine Anzahl|Zahl ART|APPR KOLLEKTIVA>	4
<eine Anzahl|Zahl ART|APPR KOLLEKTIVA>	8		
<eine beachtlich Zahl|Anzahl ART|APPR KOLLEKTIVA>	2		
<a number of COLLECTIVES>	**9218**	**<a series of COLLECTIVES>**	**2103**
<eine Reihe ART|APPR KOLLEKTIVA>	3668	<eine Reihe ART|APPR KOLLEKTIVA>	1,630
<einige KOLLEKTIVA>	1544	<zahlreiche KOLLEKTIVA>	123
<mehrere KOLLEKTIVA>	933	<eine große Zahl|Anzahl ART|APPR KOLLEKTIVA>	69
<viele KOLLEKTIVA>	939	<viele KOLLEKTIVA>	58
<zahlreiche KOLLEKTIVA>	569	<eine Vielzahl ART|APPR KOLLEKTIVA>	23
<mehrfache KOLLEKTIVA>	163	<eine erhebliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	15
<eine Anzahl|Zahl ART|APPR KOLLEKTIVA>	300	<eine Anzahl|Zahl ART|APPR KOLLEKTIVA>	8
<eine Menge KOLLEKTIVA>	42	<eine beträchtliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	6
<eine Vielzahl ART|APPR KOLLEKTIVA>	45	<eine Menge ART|APPR KOLLEKTIVA>	5

Table A3: Continued

Lexical items	Frequency	Lexical items	Frequency
<a large number of COLLECTIVES>	**1511**	**<a substantial number of COLLECTIVES>**	**44**
<viele KOLLEKTIVA>	485	<eine erhebliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	12
<eine große Zahl|Anzahl ART|APPR KOLLEKTIVA>	363	<eine beträchtliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	6
<zahlreiche KOLLEKTIVA>	295	<zahlreiche KOLLEKTIVA>	5
<eine Vielzahl ART|APPR KOLLEKTIVA>	168	<eine Reihe ART|APPR KOLLEKTIVA>	3
<eine Reihe ART|APPR KOLLEKTIVA>	53	<eine Vielzahl ART|APPR KOLLEKTIVA>	3
<eine hohe Zahl|Anzahl ART|APPR KOLLEKTIVA>	51	<eine große Zahl|Anzahl ART|APPR KOLLEKTIVA>	3
<ein Großteil ART|APPR KOLLEKTIVA>	26	<eine hohe Zahl|Anzahl ART|APPR KOLLEKTIVA>	3
<eine erhebliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	18	<eine bedeutende Zahl|Anzahl ART|APPR KOLLEKTIVA>	1
<ein großer Teil ART|APPR KOLLEKTIVA>	18	<eine wesentliche Zahl ART|APPR KOLLEKTIVA>	1
<eine Menge ART|APPR KOLLEKTIVA>	16		
<eine größere Zahl|Anzahl ART|APPR KOLLEKTIVA>	9		
<eine beträchtliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	5		
<ein bedeutender Teil ART|APPR KOLLEKTIVA>	3		
<a considerable number of COLLECTIVES>	**136**	**<a significant number of COLLECTIVES>**	**128**
<eine eine beträchtliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	33	<eine erhebliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	21
<zahlreiche KOLLEKTIVA>	16	<eine beträchtliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	20
<eine große Zahl|Anzahl ART|APPR KOLLEKTIVA>	16	<zahlreiche KOLLEKTIVA>	18
<eine erhebliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	14	<bedeutend Zahl|Anzahl ART|APPR KOLLEKTIVA|Teil>	14
<viele KOLLEKTIVA>	12	<eine große Zahl|Anzahl ART|APPR KOLLEKTIVA>	13
<eine Reihe ART|APPR KOLLEKTIVA>	9	<eine beachtliche Zahl>	6
<eine beachtliche Anzahl ART|APPR KOLLEKTIVA>	6	<signifikant Zahl|Anzahl ART|APPR KOLLEKTIVA>	6
<eine Vielzahl ART|APPR KOLLEKTIVA>	5	<viele KOLLEKTIVA>	4
<eine Anzahl|Zahl ART|APPR KOLLEKTIVA>	5	<ein großer Teil ART|APPR KOLLEKTIVA>	4
<eine erhebliche Zahl|Anzahl ART|APPR KOLLEKTIVA>	14	<eine Reihe ART|APPR KOLLEKTIVA>	4
<a huge number of COLLECTIVES>	**180**		
<viele KOLLEKTIVA>	59		
<eine große Zahl|Anzahl ART|APPR KOLLEKTIVA>	39		
<zahlreiche KOLLEKTIVA>	26		
<eine Vielzahl ART|APPR KOLLEKTIVA>	20		
<eine riesige Zahl|Anzahl ART|APPR KOLLEKTIVA>	11		
<eine enorme Zahl|Anzahl ART|APPR KOLLEKTIVA>	6		
<eine hohe Zahl|Anzahl ART|APPR KOLLEKTIVA>	4		
<eine Fülle ART|APPR KOLLEKTIVA>	4		
<eine Reihe ART|APPR KOLLEKTIVA>	4		
<eine ungeheüre Zahl ART|APPR KOLLEKTIVA>	4		
<eine Menge ART|APPR KOLLEKTIVA>	3		

Table A4: German lexical items from the TLD {VIELE KOLLEKTIVA} and their English translation correspondences

German as a source language			
Lexical items	Frequency	Lexical items	Frequency
<viele KOLLEKTIVA>	**19965**	**<zahlreiche KOLLEKTIVA>**	**6680**
<many COLLECTIVES>	410	<many COLLECTIVES>	3636
<a large number of COLLECTIVES>	206	<numerous COLLECTIVES>	1316
<a range of COLLECTIVES>	154	<a number of COLLECTIVES>	585
<a variety of COLLECTIVES>	115	<a large number of COLLECTIVES>	312
<a multitude of COLLECTIVES>	66	<a lot of COLLECTIVES>	134
<numerous COLLECTIVES>	56	<a range of COLLECTIVES>	104
<a number of COLLECTIVES>	54	<a series of COLLECTIVES>	64
<a great many COLLECTIVES>	43	<a great number of COLLECTIVES>	38
<a host of COLLECTIVES>	33	<a range of COLLECTIVES>	12
<a multiplicity of COLLECTIVES>	20	<a huge number of COLLECTIVES>	25
<a huge number of COLLECTIVES>	16	<a significant number of COLLECTIVES>	18
		<a multitude of COLLECTIVES>	18
		<a considerable number of COLLECTIVES>	16
		<a good many COLLECTIVES>	14
		<a substantial number of COLLECTIVES>	6
		<a high number of COLLECTIVES>	6
		<a growing number of COLLECTIVES>	4
		<a large proportion of COLLECTIVES>	3
<eine Vielzahl ART|APPR KOLLEKTIVA>	**1634**	**<mehrere KOLLEKTIVA>**	**2496**
<many COLLECTIVES>	410	<several COLLECTIVES>	1301
<a large number of COLLECTIVES>	206	<a number of COLLECTIVES>	288
<a range of COLLECTIVES>	154	<various COLLECTIVES>	285
<a variety of COLLECTIVES>	115	<many COLLECTIVES>	186
<a host of COLLECTIVES>	58	<different COLLECTIVES>	90
<a number of COLLECTIVES>	54	<some COLLECTIVES>	60
<a multitude of COLLECTIVES>	66	<a series of COLLECTIVES>	30
<numerous COLLECTIVES>	56	<more than one COLLECTIVES>	29
<a great many COLLECTIVES>	43	<numerous COLLECTIVES>	22
<a huge number of COLLECTIVES>	16	<a large number of COLLECTIVES>	9
<a series of COLLECTIVES>	8	<a variety of COLLECTIVES>	6
<a multiplicity of COLLECTIVES>	20		6
<a great number of COLLECTIVES>	14		
<a good number of COLLECTIVES>	6		
<a considerable number of COLLECTIVES>	4		
<a significant number of COLLECTIVES>	2		
<a significant number of COLLECTIVES>	2		

Table A4: Continued

Lexical items	Frequency	Lexical items	Frequency
<eine Reihe ART|APPR KOLLEKTIVA>	**6968**	**<eine große Zahl|Anzahl KOLLEKTIVA>**	**856**
<a number of COLLECTIVES>	3624	<a large number of COLLECTIVES>	414
<a series of COLLECTIVES>	1665	<many COLLECTIVES>	104
<a range of COLLECTIVES>	596	<a huge number of COLLECTIVES>	30
<a set of COLLECTIVES>	266	<great many COLLECTIVES>	31
<many COLLECTIVES>	125	<a great number of COLLECTIVES>	25
<a raft of COLLECTIVES>	65	<numerous COLLECTIVES>	20
<a certain number of COLLECTIVES>	86	<a high number of COLLECTIVES>	16
<numerous COLLECTIVES>	84	<a considerable number of COLLECTIVES>	13
<a lot of COLLECTIVES>	65	<a significant number of COLLECTIVES>	13
<a variety of COLLECTIVES>	54	<a lot of COLLECTIVES>	10
<a large number of COLLECTIVES>	53	<a range of COLLECTIVES>	9
<a host of COLLECTIVES>	50	<a host of COLLECTIVES>	6
<a list of COLLECTIVES>	45	<a vast number of COLLECTIVES>	4
<a array of COLLECTIVES>	33	<a massive number of COLLECTIVES>	4
<a string of COLLECTIVES>	20	<maximum number of COLLECTIVES>	4
<a package of COLLECTIVES>	18	<an enormous number of COLLECTIVES>	3
<a good number of COLLECTIVES>	13	<a substantial number of COLLECTIVES>	2
<a considerable number of COLLECTIVES>	6		
<certain amount of COLLECTIVES>	6		
<a significant number of COLLECTIVES>	5		
<eine Menge ART|APPR KOLLEKTIVA>	**680**	**<eine erhebliche Zahl|Anzahl ART|APPR KOLLEKTIVA>**	**81**
<a lot of COLLECTIVES>	183	<a large number of COLLECTIVES>	16
<a great deal of COLLECTIVES>	116	<a significant number of COLLECTIVES>	24
<a number of COLLECTIVES>	45	<a large number of COLLECTIVES>	16
<many COLLECTIVES>	36	<a considerable number of COLLECTIVES>	11
<a lot of COLLECTIVES>	26	<a substantial number of COLLECTIVES>	9
<a host of>	20	<a number of COLLECTIVES>	4
<a large number of COLLECTIVES>	14		
<a great many COLLECTIVES>	14	**<eine beträchtliche Zahl|Anzahl ART|APPR KOLLEKTIVA**	**92**
<a huge amount of COLLECTIVES>	12	<a considerable number of COLLECTIVES>	**26**
<a good deal of COLLECTIVES>	13	<a significant number of COLLECTIVES>	22
<plenty of COLLECTIVES>	9	<a large number of COLLECTIVES>	6
<good many COLLECTIVES>	9	<many COLLECTIVES>	8
<considerable amount of COLLECTIVES>	6	<a substantial number of COLLECTIVES>	6
<certain amount of COLLECTIVES>	6	<a number of COLLECTIVES>	4
<large amount of COLLECTIVES>	6	<a large proportion of COLLECTIVES>	2
<numerous COLLECTIVES>	4		
<a multitude of>	3		
<a huge number of COLLECTIVES>	2		

Table A5: English lexical items from the TLSd {MANY PROBLEMS} and their German translation correspondences

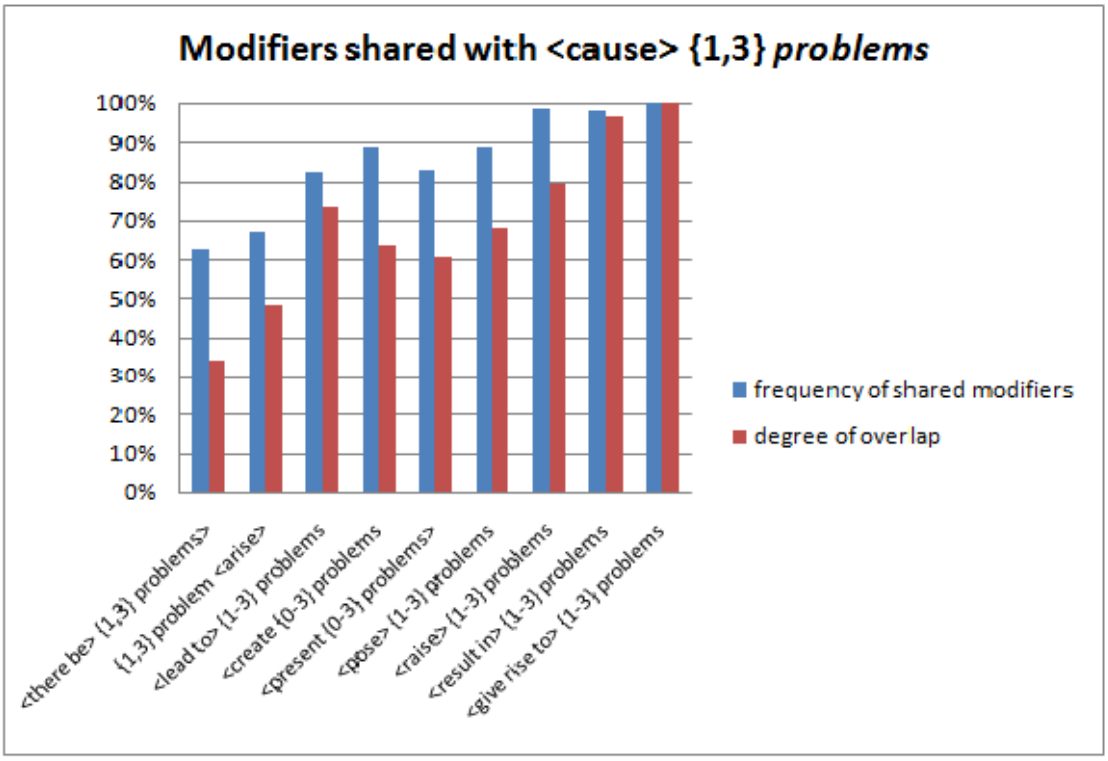
English as a source language			
Lexical items	**Frequency**	**Lexical items**	**Frequency**
<many PROBLEMS>	**1688**	**<numerous PROBLEMS>**	**160**
<viele PROBLEME>	1069	<zahlreiche PROBLEME>	63
<zahlreiche PROBLEME>	234	<viele PROBLEME>	43
<eine große Zahl|Anzahl ART|APPR PROBLEME>	165	<mehrere PROBLEME>	2
<eine Reihe ART|APPR PROBLEME>	43	<eine Reihe ART|APPR PROBLEME>	6
<eine Vielzahl ART|APPR PROBLEME>	34	<eine große Zahl|Anzahl ART|APPR PROBLEME>	21
<eine Menge GEN PROBLEME>	11		
<eine beträchtliche Zahl|Anzahl ART|APPR PROBLEME>	9		
<eine erhebliche Zahl|Anzahl ART|APPR PROBLEME>	5		
<eine Anzahl|Zahl ART|APPR PROBLEME>	5		
<mehrere PROBLEME>	5		
<a range of PROBLEMS>	**52**	**<a lot of PROBLEMS>**	**122**
<eine Reihe ART|APPR PROBLEME>	34	<viele PROBLEME>	51
<eine Vielzahl ART|APPR PROBLEME>	6	<zahlreiche PROBLEME>	20
<viele PROBLEME>	4	<eine Reihe ART|APPR PROBLEME>	10
<zahlreiche PROBLEME>	3	<eine Menge GEN PROBLEME>	9
<eine Palette ART|APPR PROBLEME>	2	<eine Vielzahl ART|APPR PROBLEME>	4
		<eine große Zahl|Anzahl ART|APPR PROBLEME>	6
<a number of PROBLEMS>	**620**	**<a large number of PROBLEMS>**	**412**
<eine Reihe ART|APPR PROBLEME>	360	<eine große Zahl|Anzahl ART|APPR PROBLEME>	326
<eine Anzahl|Zahl ART|APPR PROBLEME>	94	<viele PROBLEME>	19
<zahlreiche PROBLEME>	40	<eine erhebliche Zahl|Anzahl ART|APPR PROBLE	14
<viele PROBLEME>	39	<eine Vielzahl ART|APPR PROBLEME>	10
<mehrere PROBLEME>	36	<zahlreiche PROBLEME>	8
<eine große Zahl|Anzahl ART|APPR PROBLEME>	18	<eine beträchtliche Zahl|Anzahl ART|APPR PROE	6
<eine Menge GEN PROBLEME>	6		
<eine beträchtliche Zahl|Anzahl ART|APPR PROBLEME>	4	**<a series of PROBLEMS>**	**166**
<eine erhebliche Zahl|Anzahl ART|APPR PROBLEME>	4	<eine Reihe ART|APPR PROBLEME>	123
<eine Vielzahl ART|APPR PROBLEME>	2	<eine Anzahl|Zahl ART|APPR PROBLEME>	14
<eine Reihe ART|APPR PROBLEME>	2	<zahlreiche PROBLEME>	9
		<viele PROBLEME>	6
		<eine große Zahl|Anzahl ART|APPR PROBLEME>	2

Table A6: German lexical items from the TLSd {VIELE PROBLEME} and their English translation correspondences

German as a source language			
Lexical items	Frequency	Lexical items	Frequency
<viele PROBLEME>	**959**	**<zahlreiche PROBLEME>**	**434**
<many PROBLEMS>	686	<many PROBLEMS>	233
<a number PROBLEMS>	46	<numerous PROBLEMS>	63
<a lot of PROBLEMS>	42	<a number PROBLEMS>	51
<numerous PROBLEMS>	33	<a lot of PROBLEMS>	20
<a large number PROBLEMS>	12	<a range of PROBLEMS>	14
<a range of PROBLEMS>	9	<a series of PROBLEMS>	14
<a series of PROBLEMS>	3	<a large number PROBLEMS>	6
<a huge number of PROBLEMS>	3	<a considerable number of PROBLEMS>	2
<eine Reihe ART|APPR PROBLEME>	**620**	**<eine Vielzahl ART|APPR PROBLEME>**	**86**
<a number PROBLEMS>	362	<many PROBLEMS>	28
<a series of PROBLEMS>	142	<a number PROBLEMS>	16
<a range of PROBLEMS>	61	<a range of PROBLEMS>	16
<many PROBLEMS>	30	<a large number PROBLEMS>	6
<a lot of PROBLEMS>	6	<a lot of PROBLEMS>	2
<numerous PROBLEMS>	4	<a huge number of PROBLEMS>	2
		<eine große Zahl|Anzahl ART|APPR PROBLEME>	**34**
		<a number PROBLEMS>	14
		<a large number PROBLEMS>	8
		<many PROBLEMS>	6

II) Appendix B

Figure B1: Frequency and degree of overlap of shared modifiers that occur with the word form *problems* in the TLD {CAUSE PROBLEM} (all lexical items)



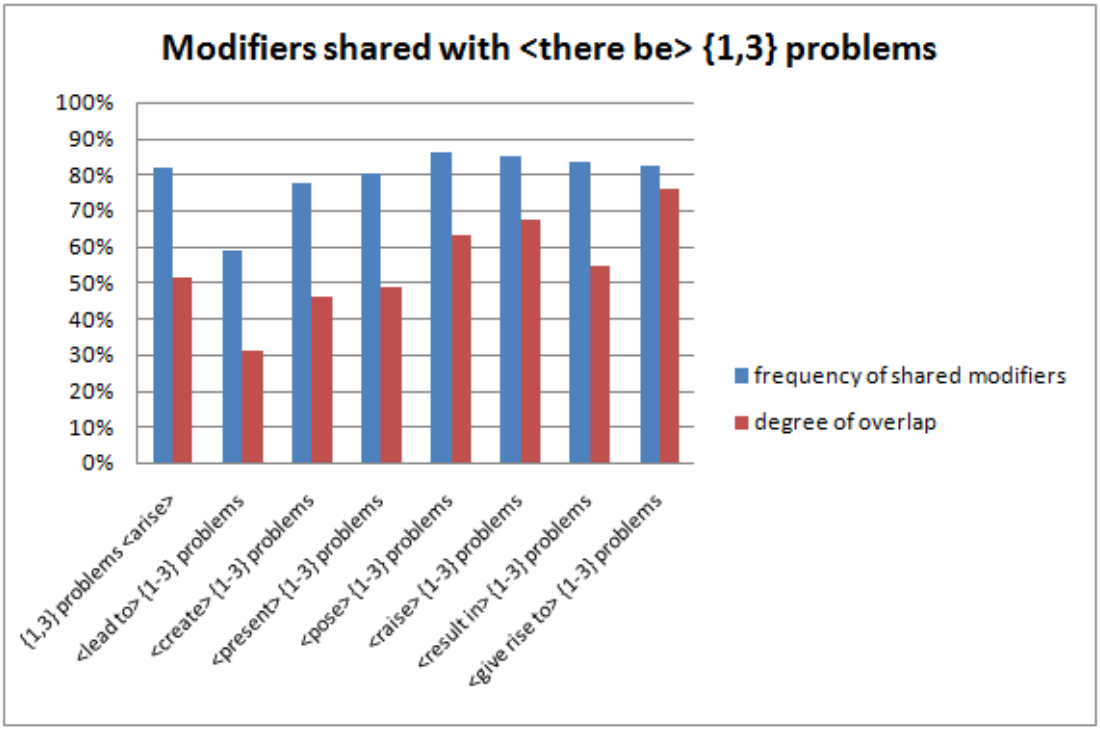
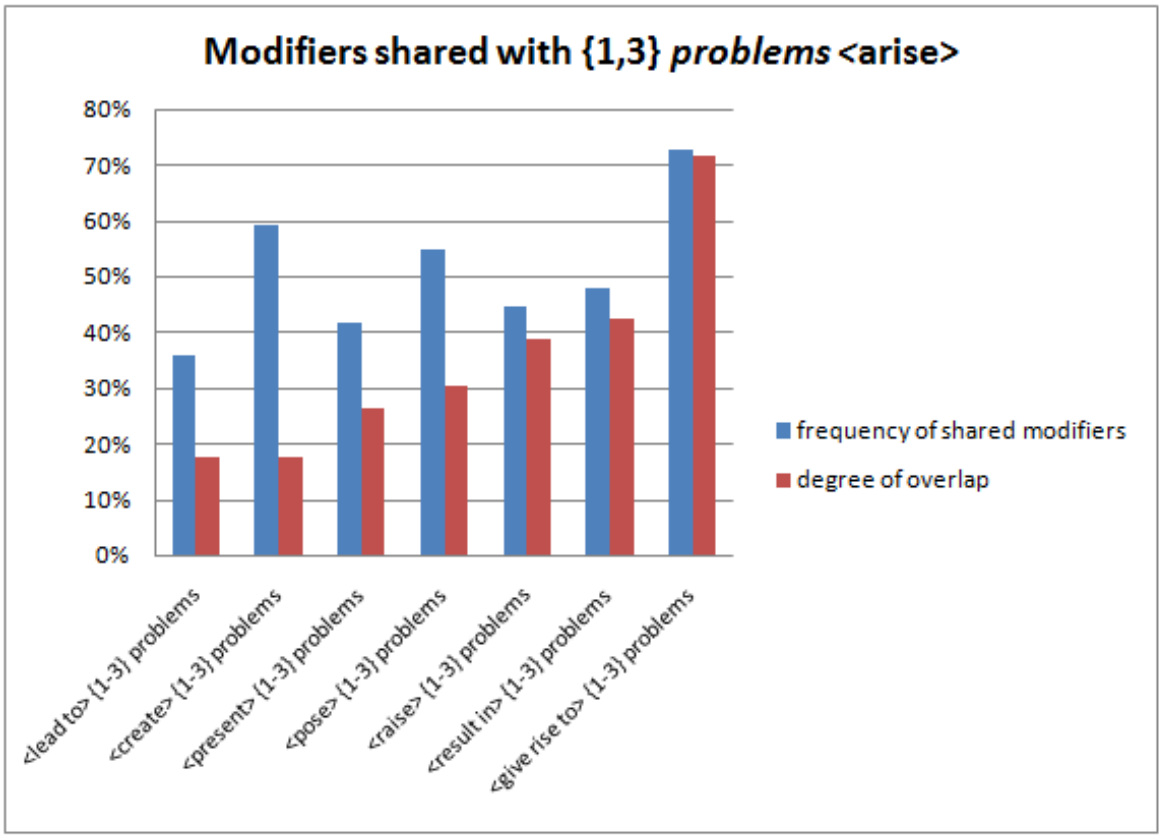


Figure B1: Continued



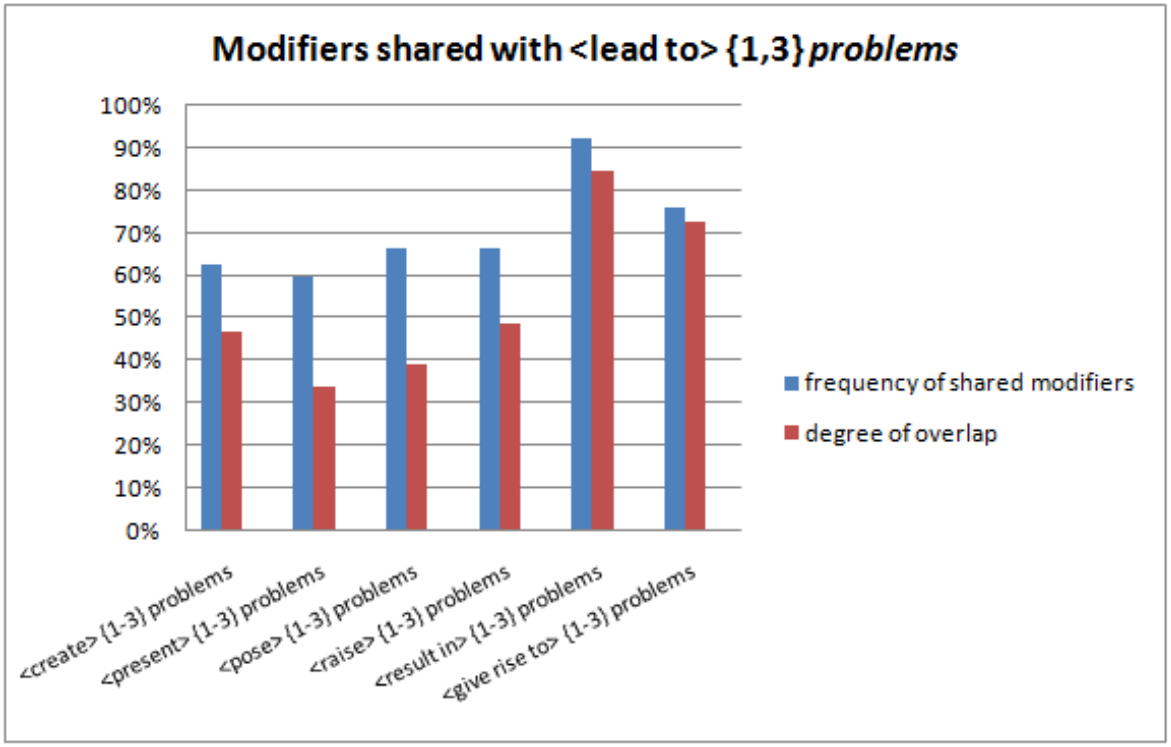
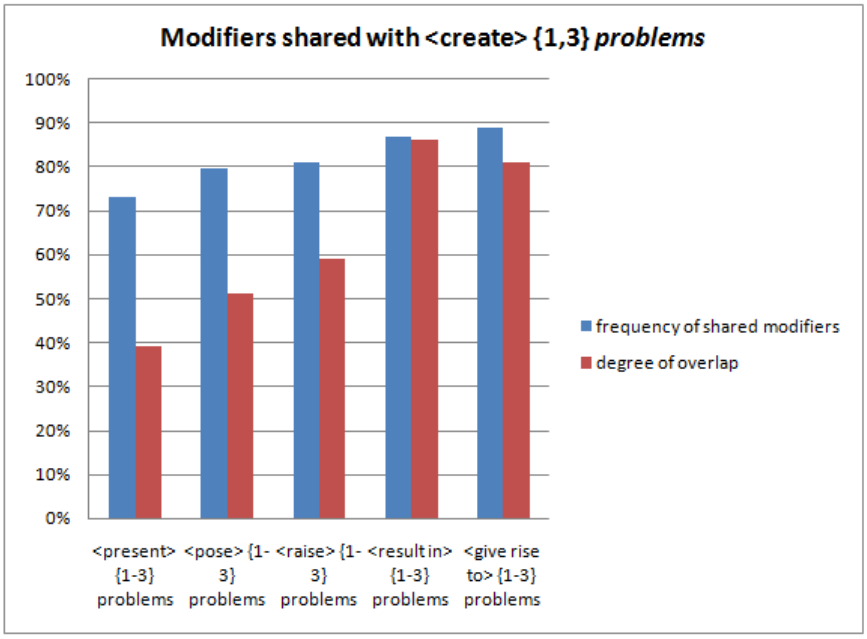
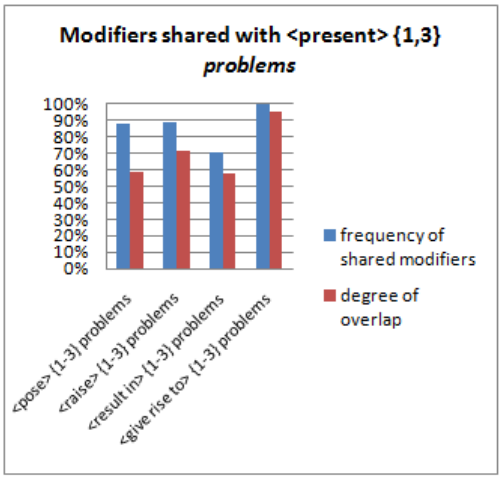
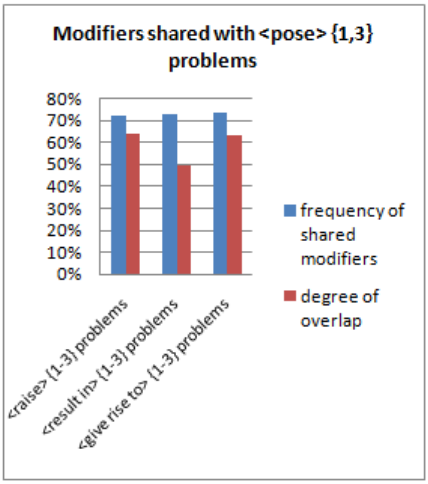
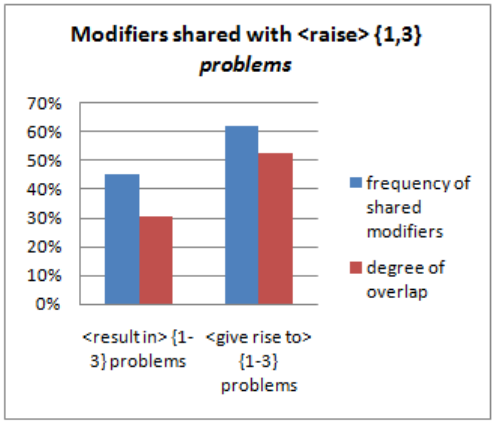


Figure B1: Continued









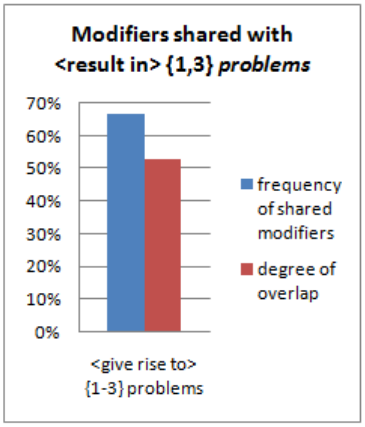
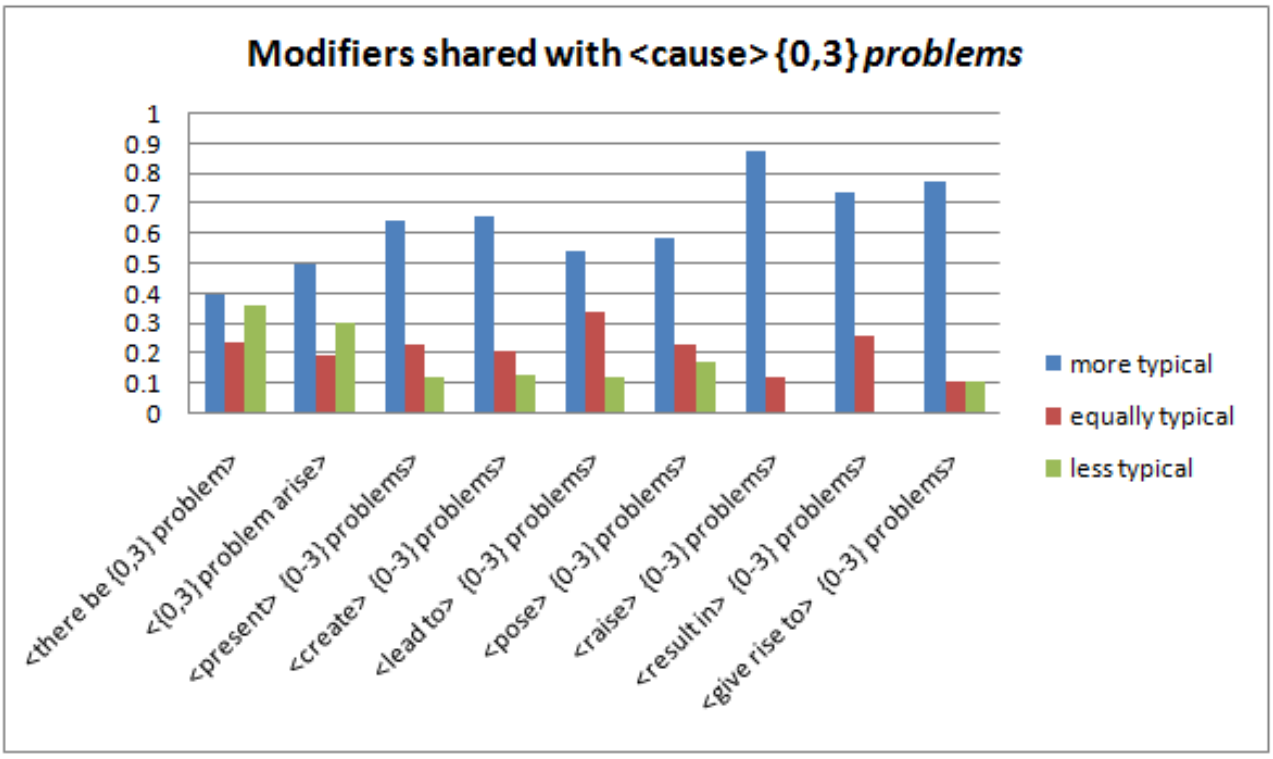


Figure B2: Association strength values for the collocations made up of verbal elements, shared modifiers and the word form *problems* (all lexical items)



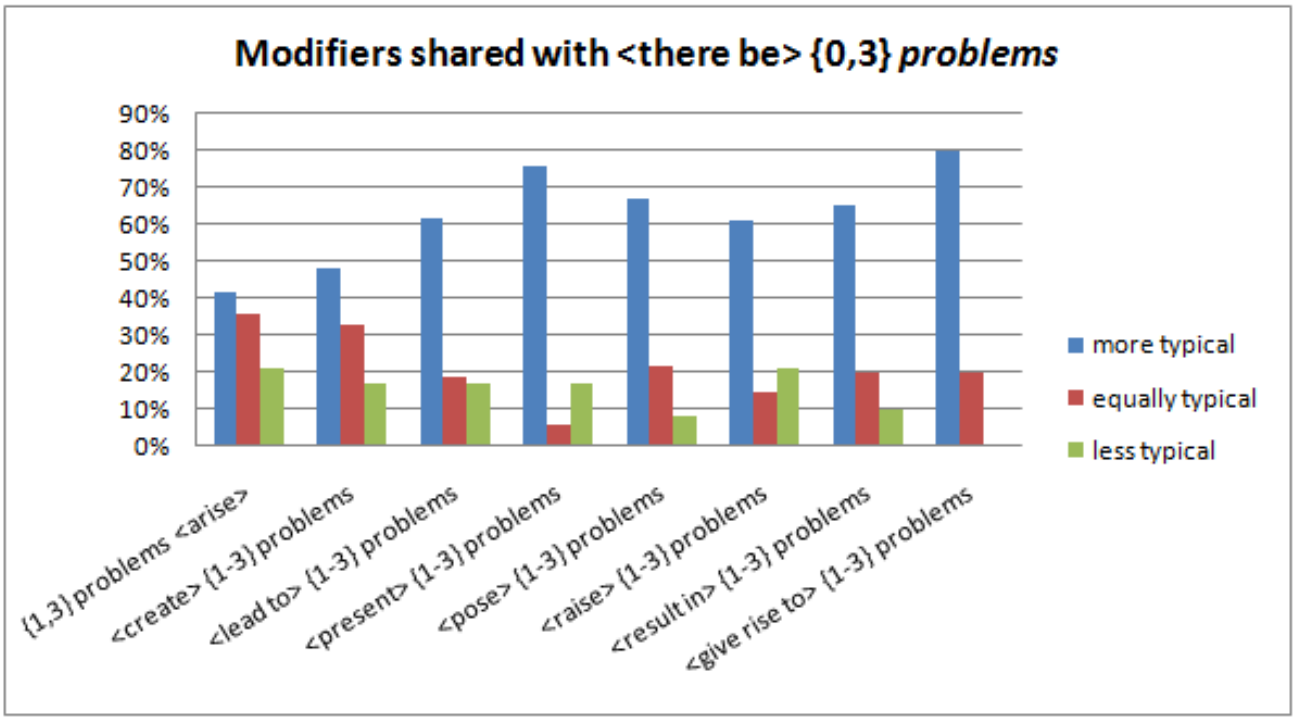
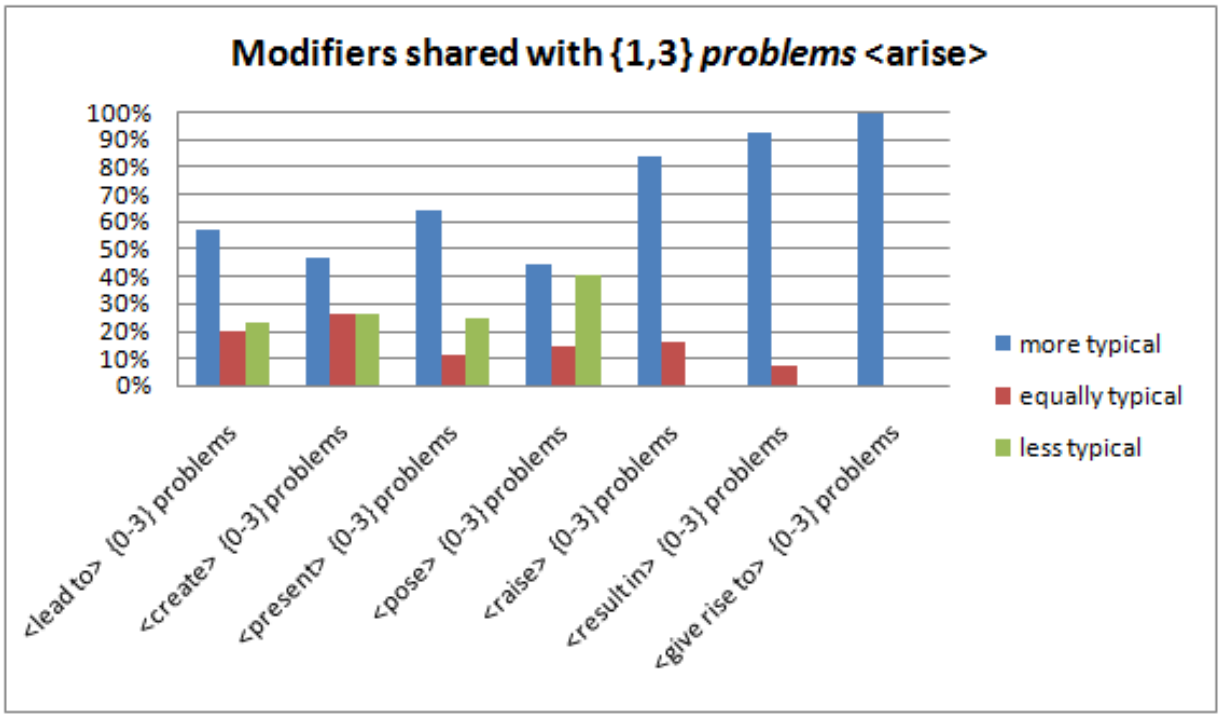


Figure B2: Continued



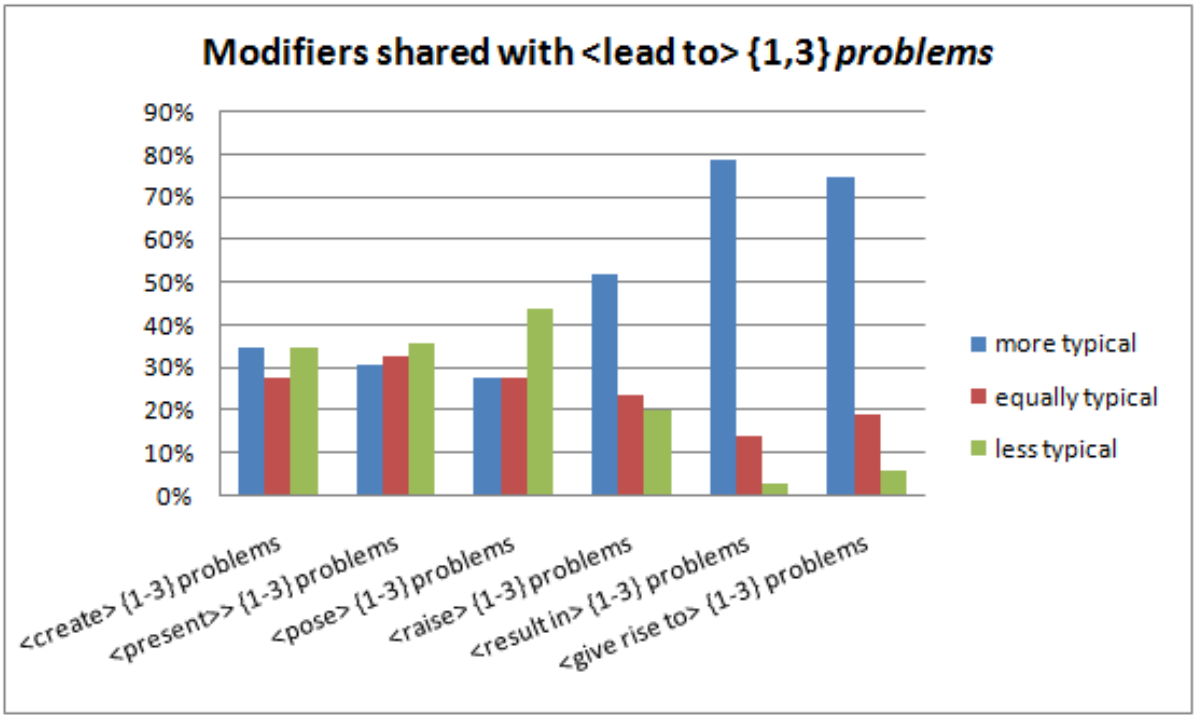
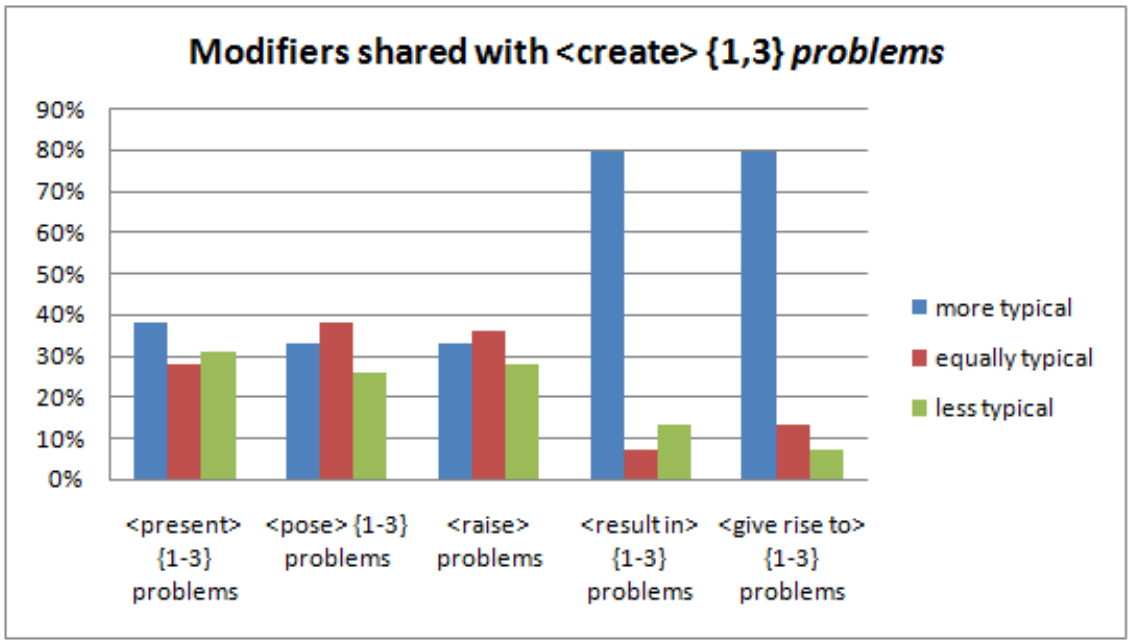


Figure B2: Continued



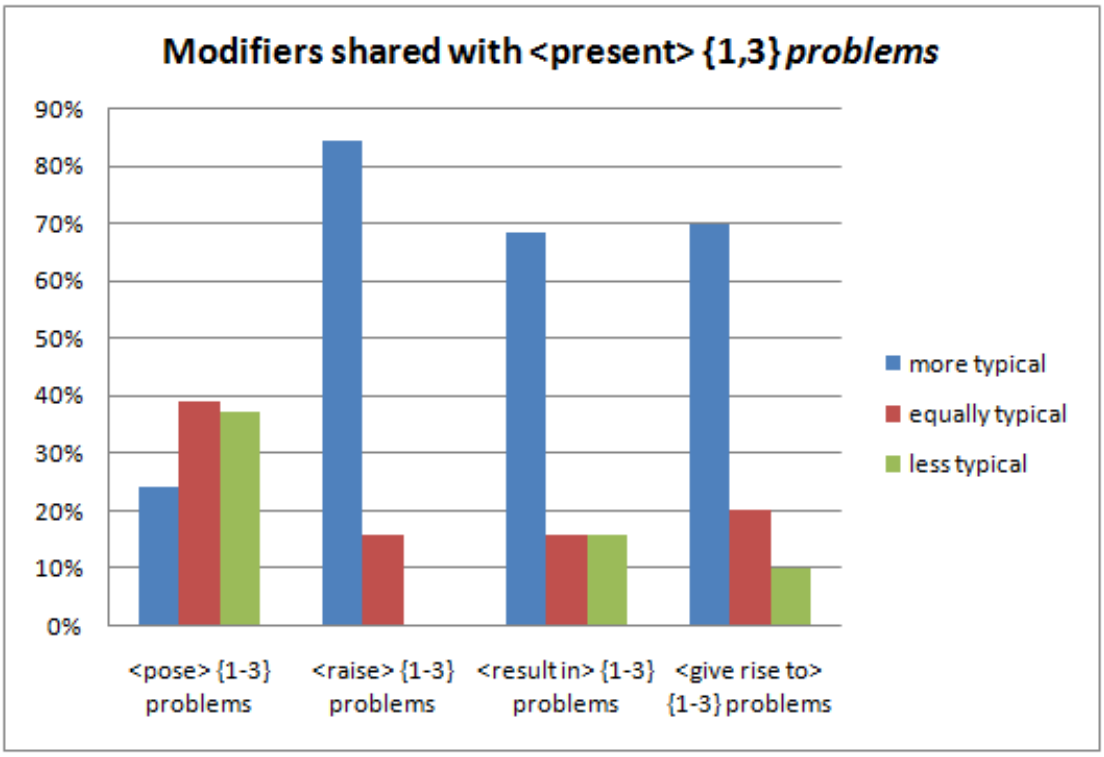
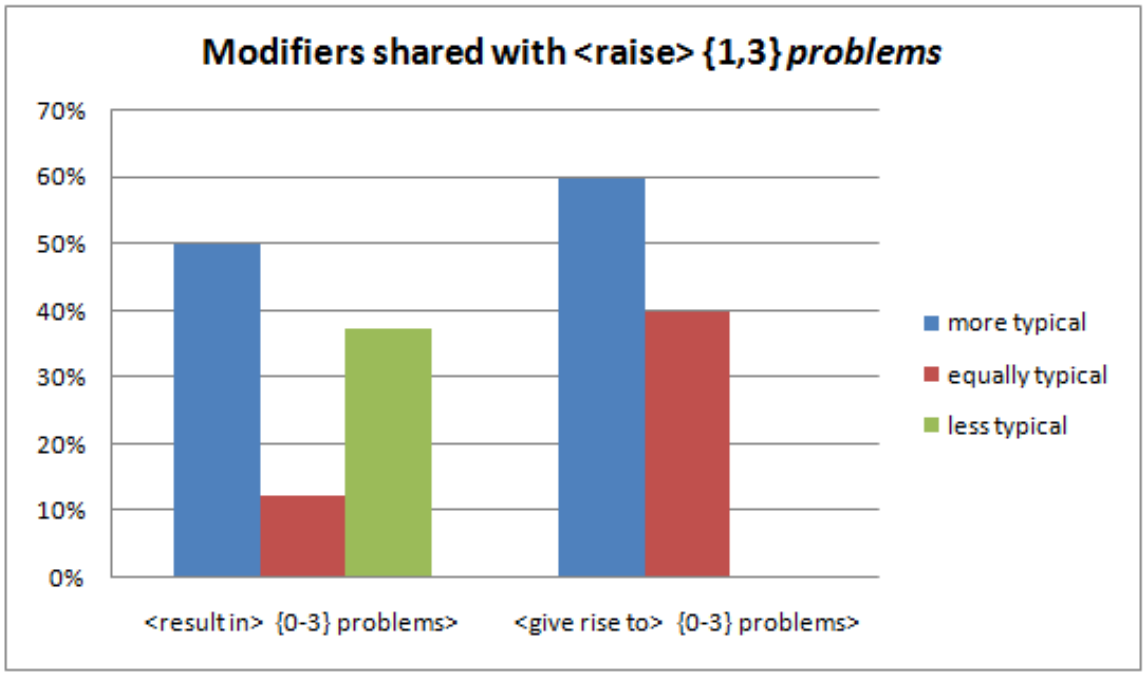
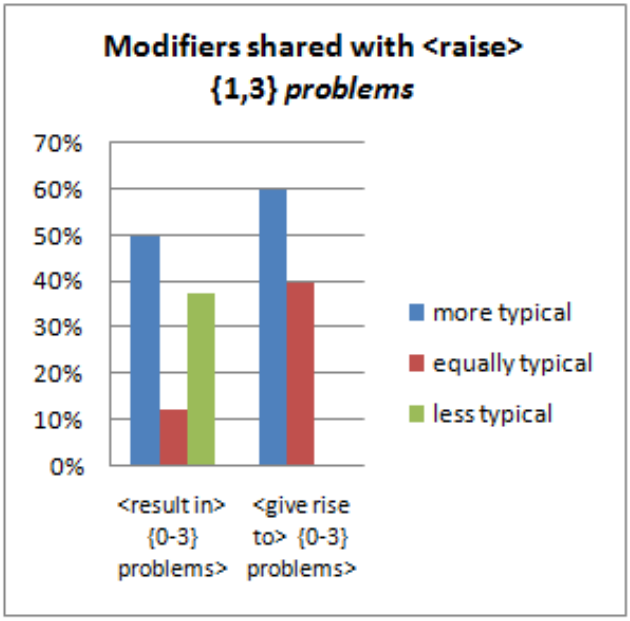


Figure B2: Continued





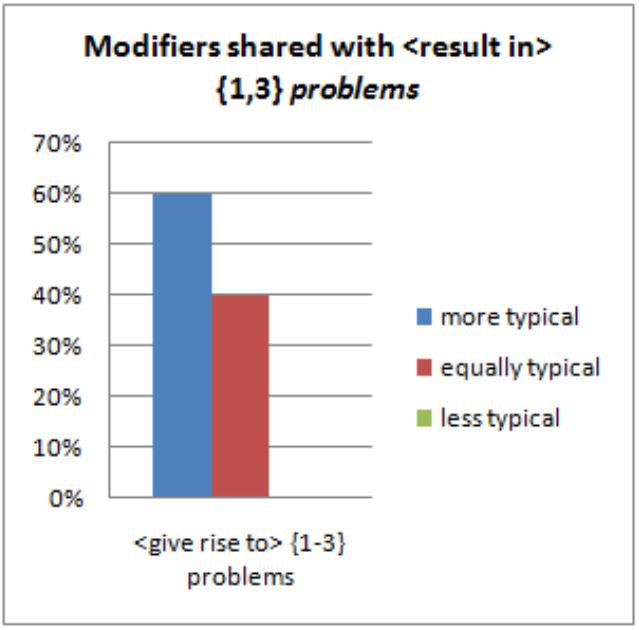
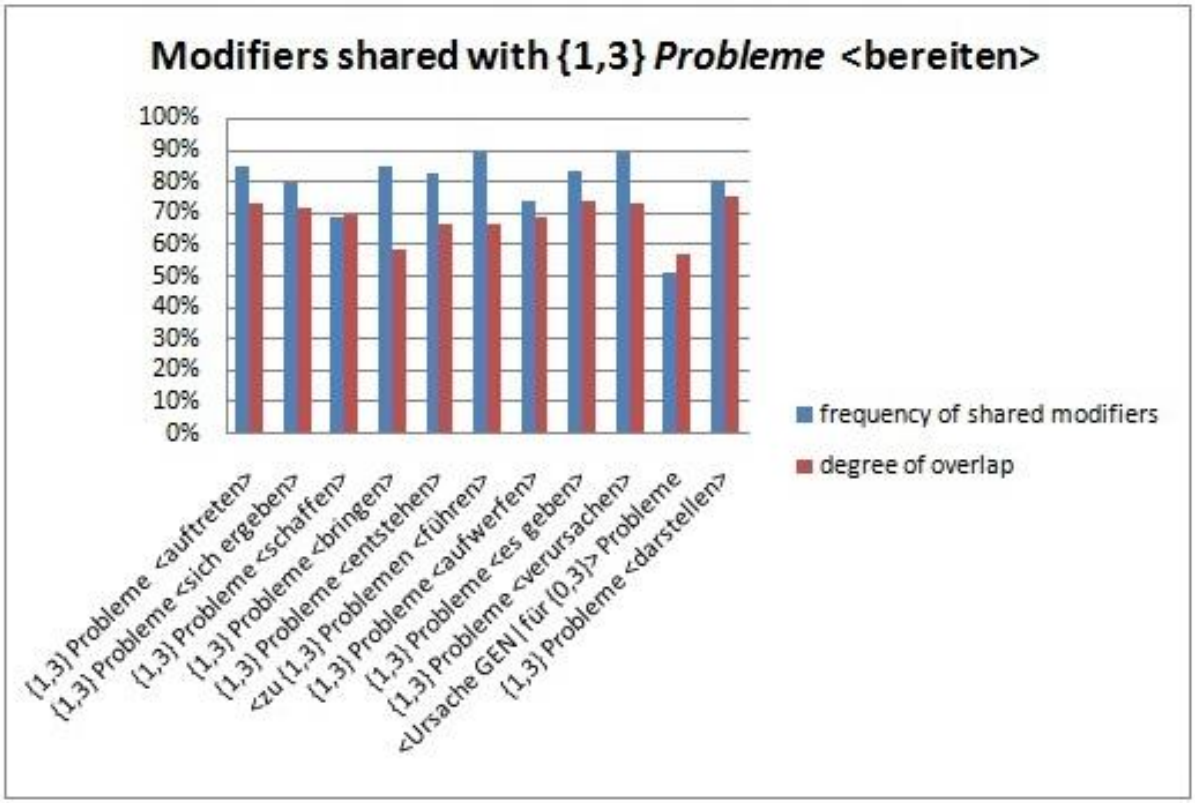


Figure B3: Frequency and degree of overlap of shared modifiers that occur with the word form *Probleme* (all lexical items)



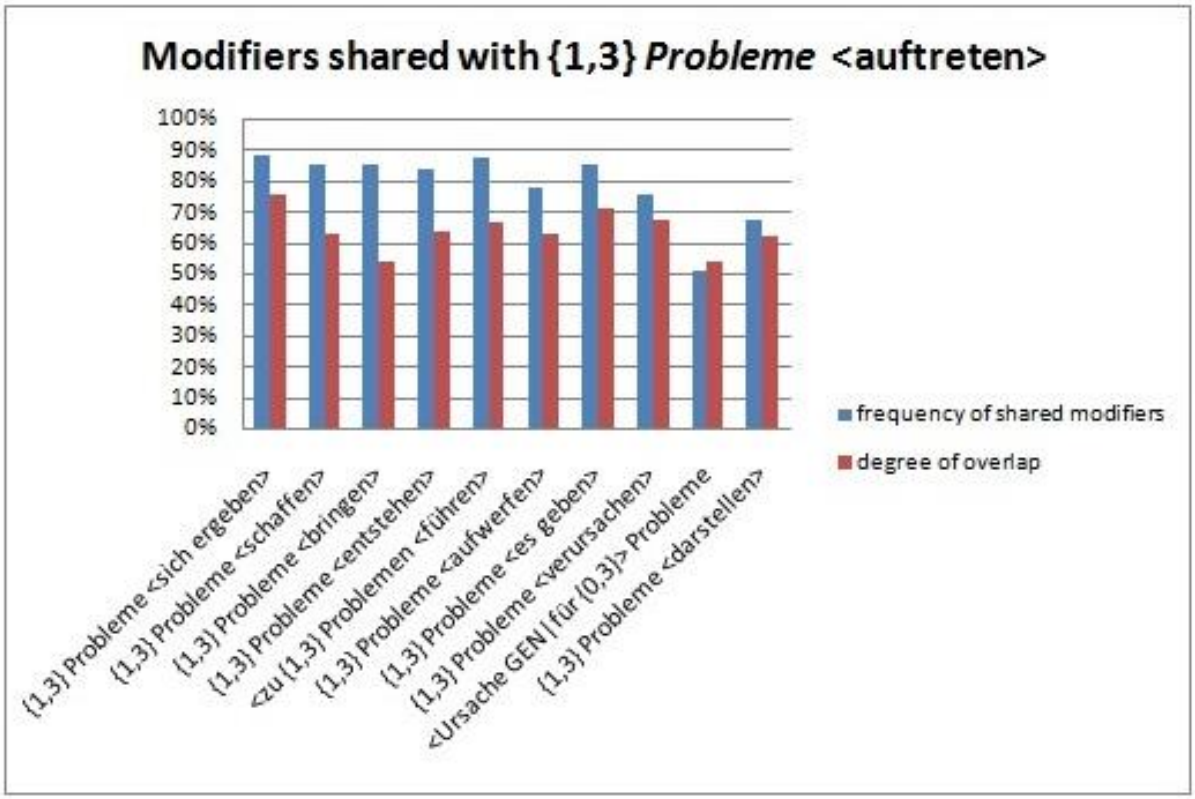
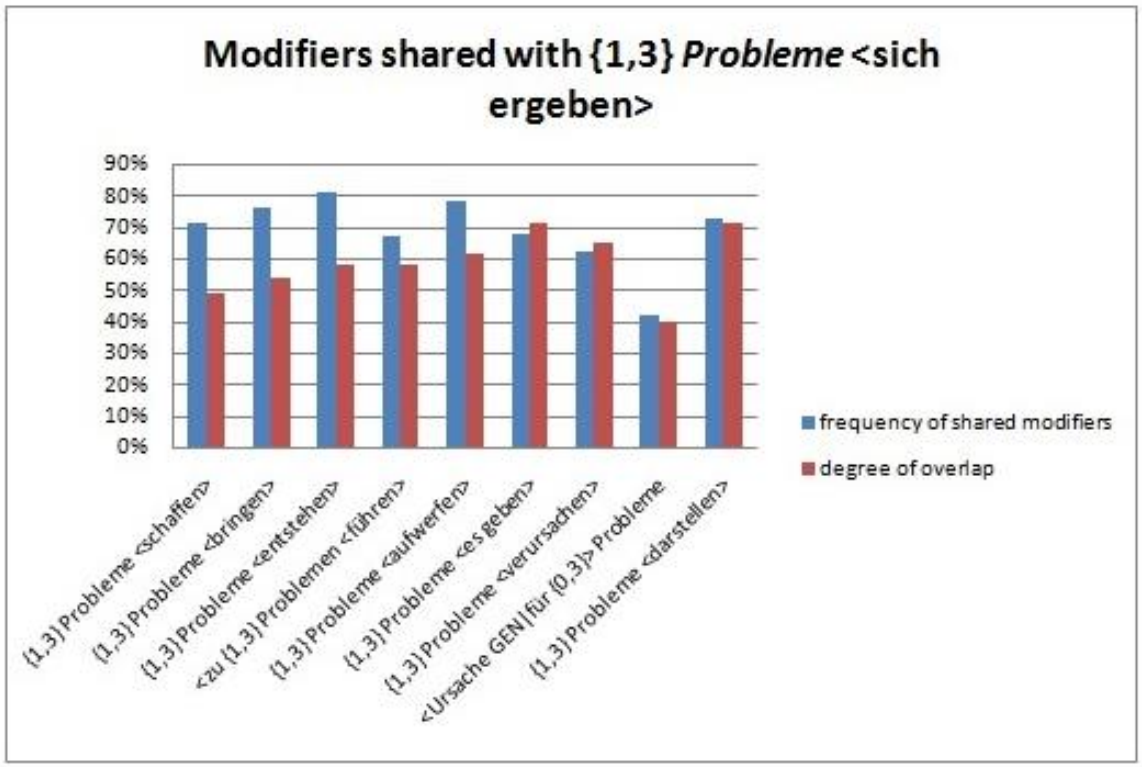


Figure B3: Continued



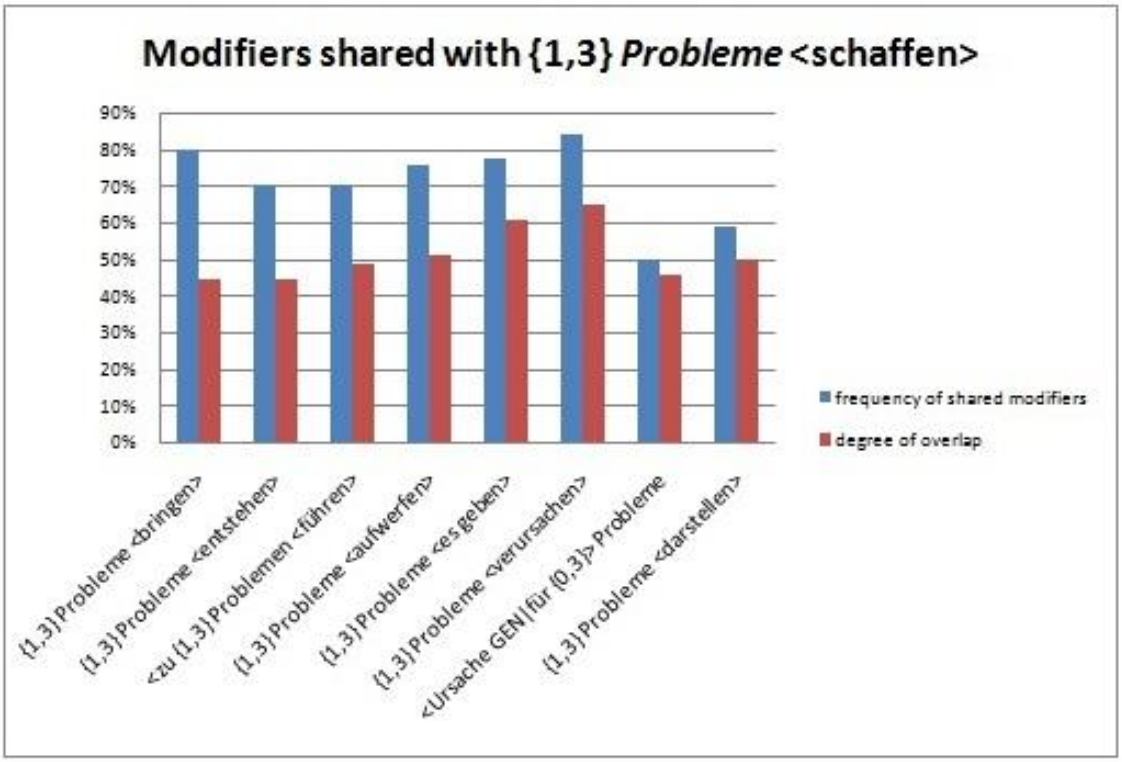
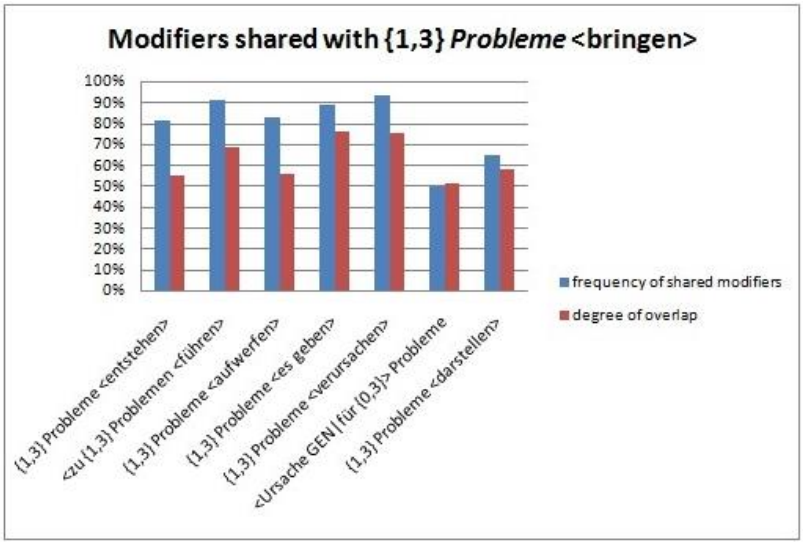
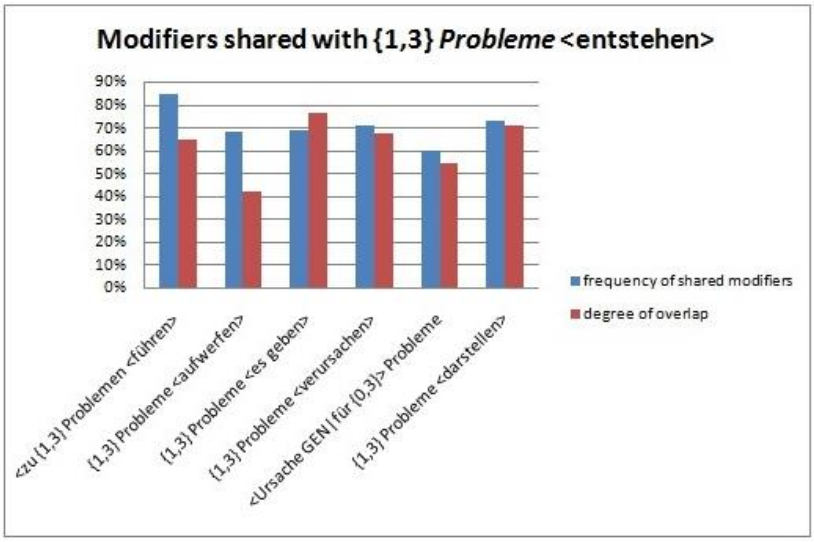


Figure B3: Continued





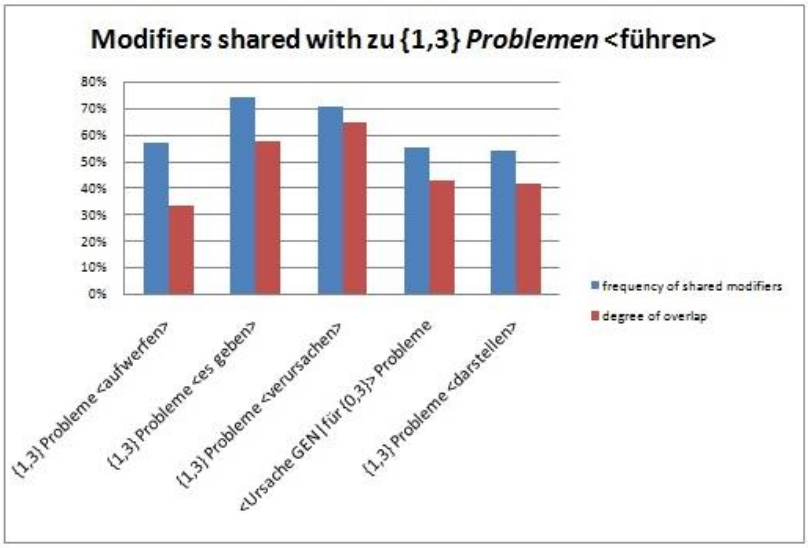
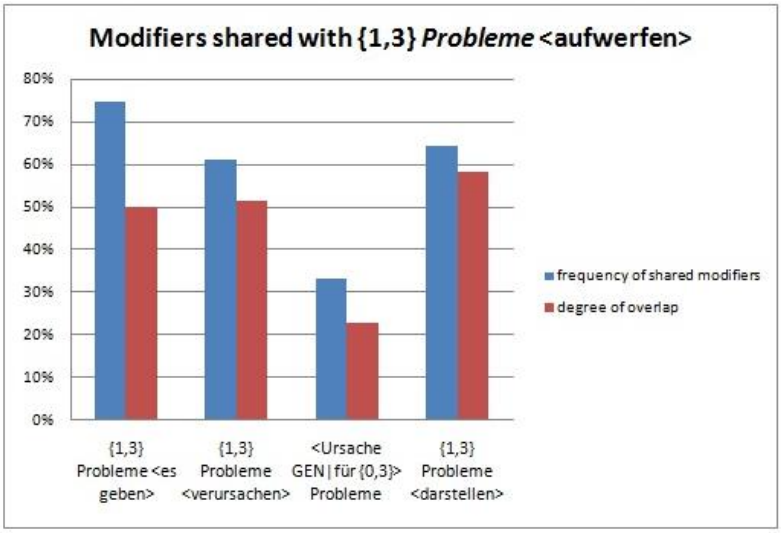
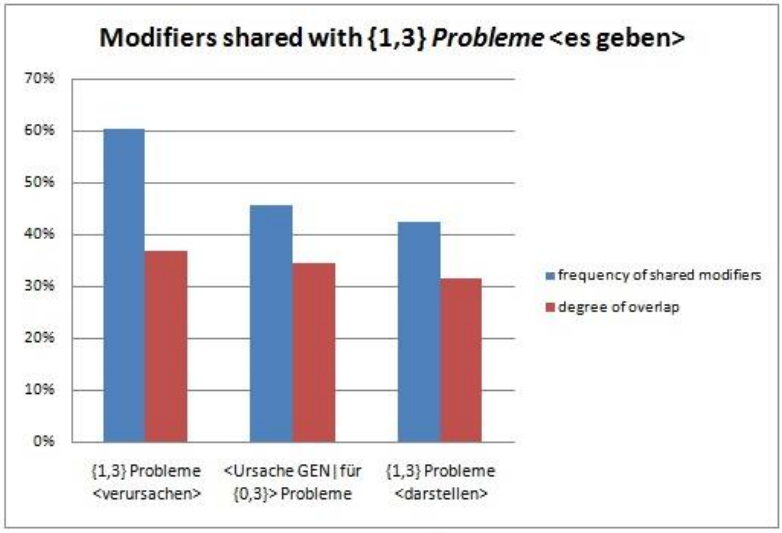
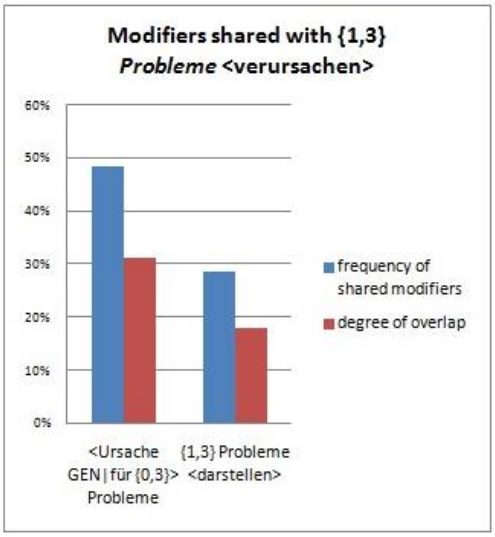


Figure B3: Continued







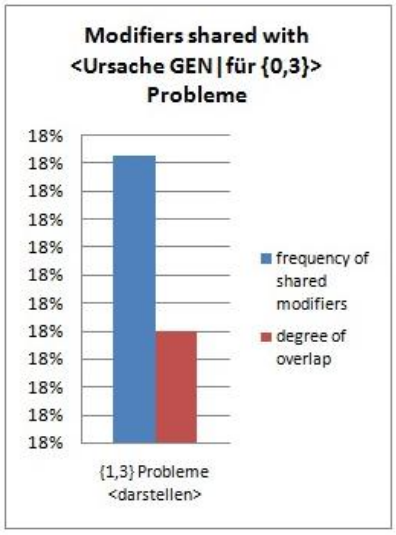
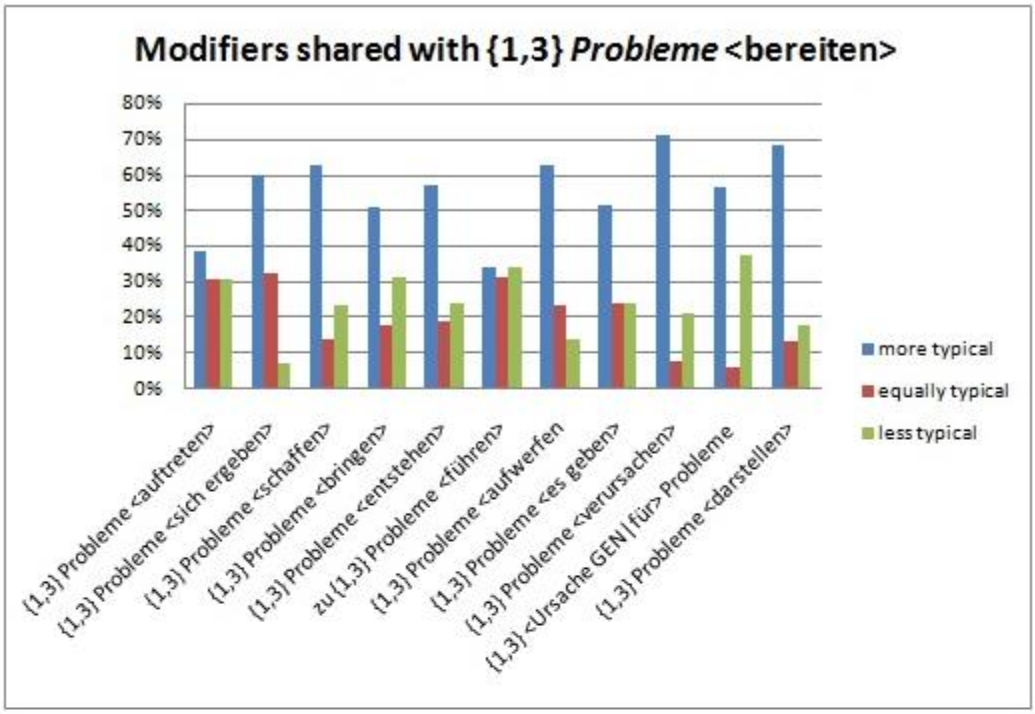


Figure B4: Association strength values for the collocations made up of verbal elements, shared modifiers and the word form *Probleme* (all lexical items)



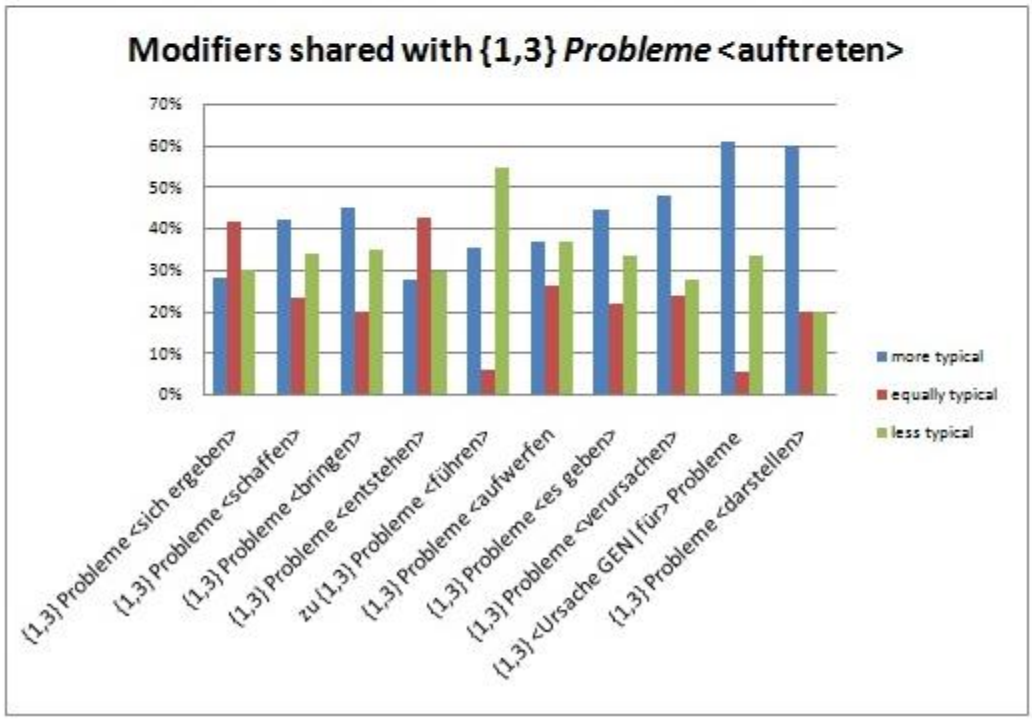
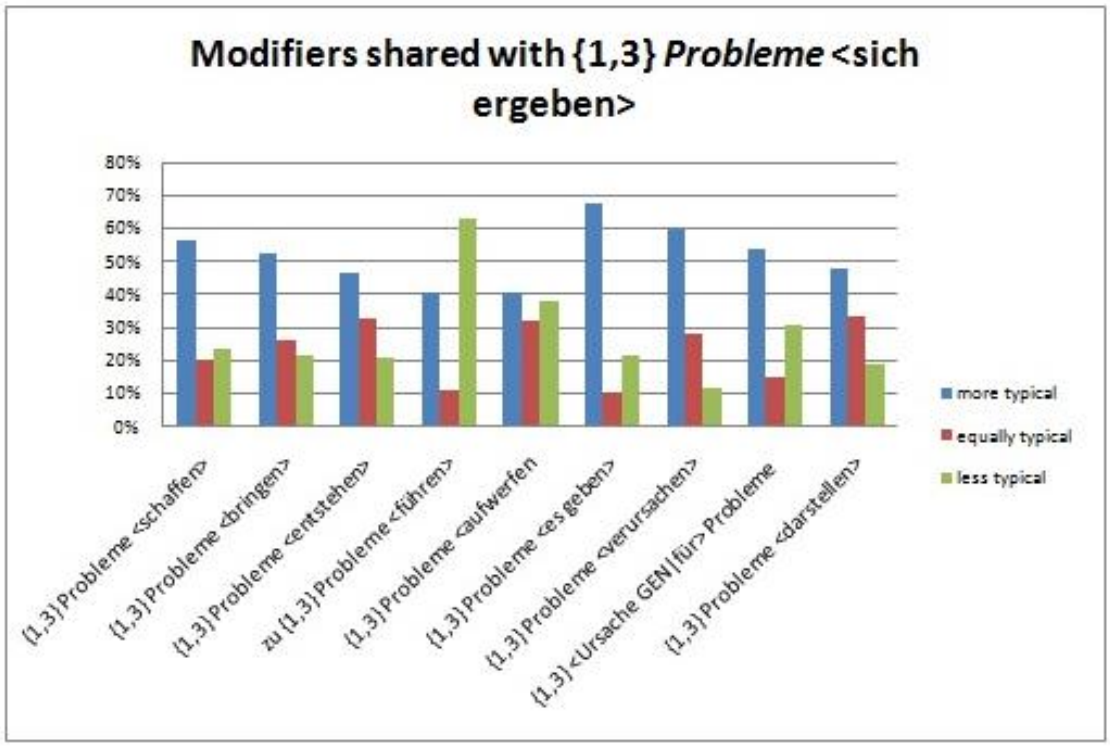


Figure B4: Continued



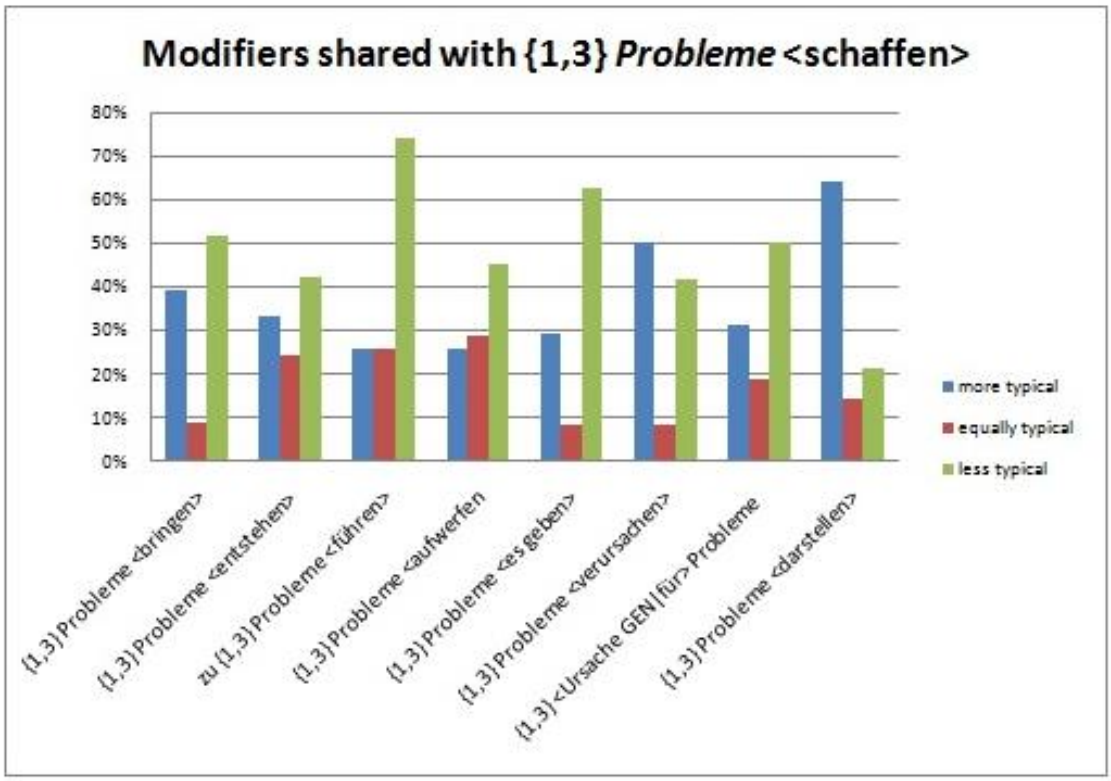
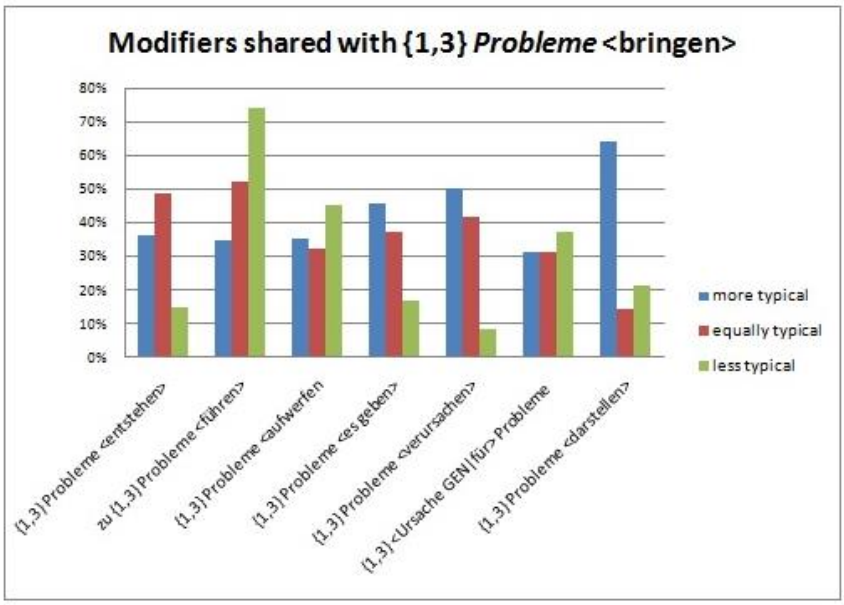
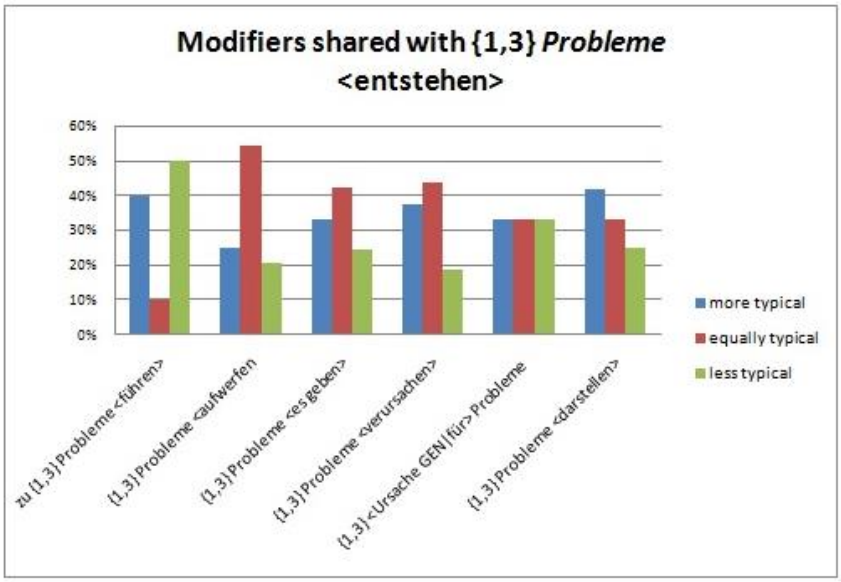


Figure B4: Continued





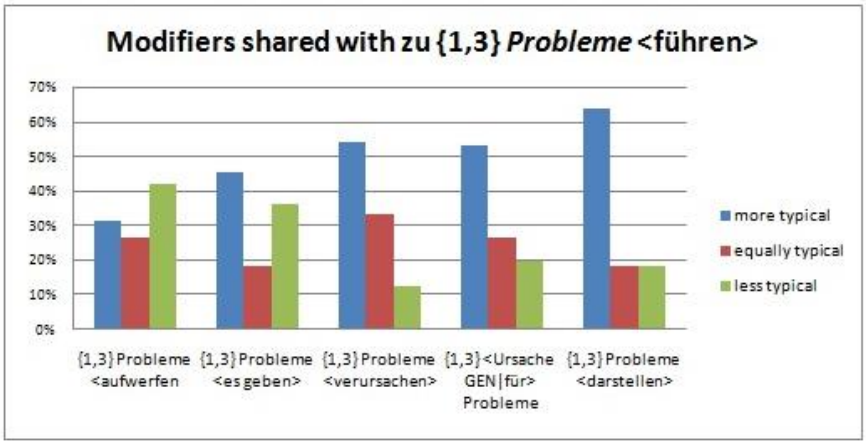
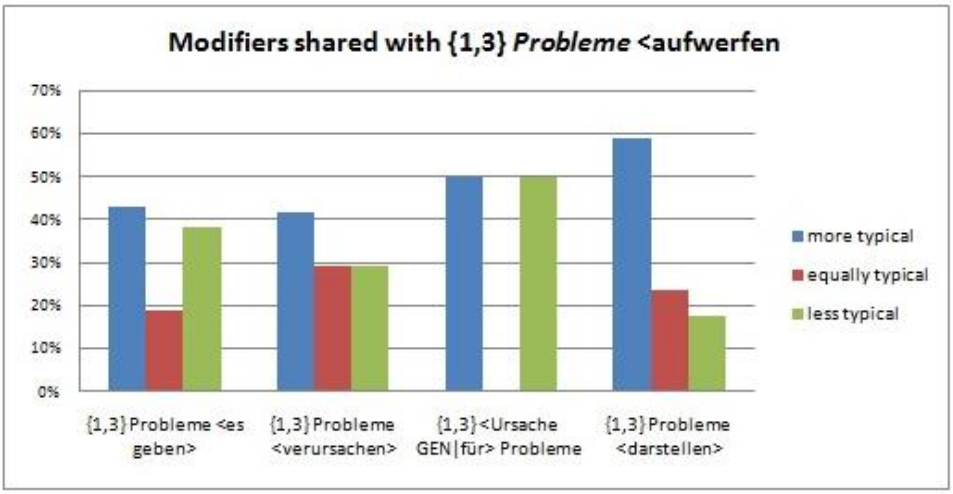
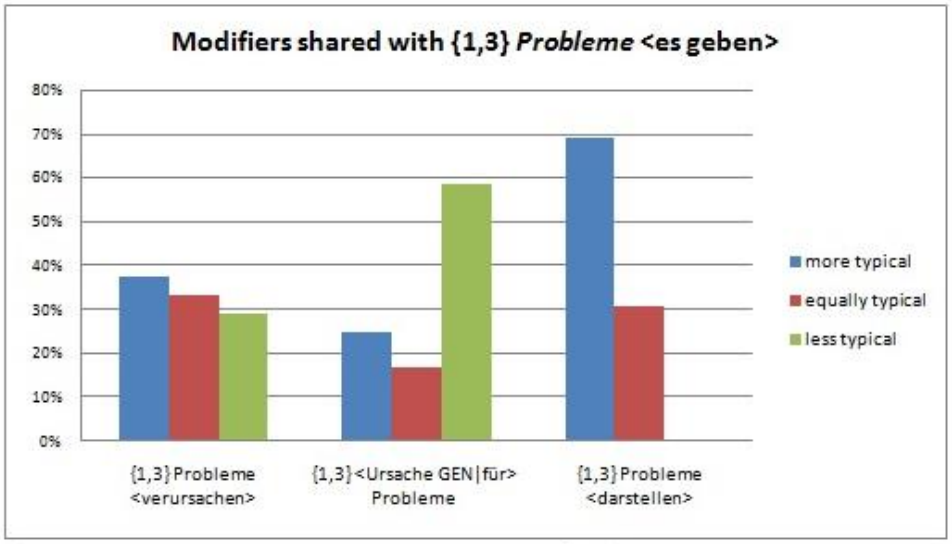
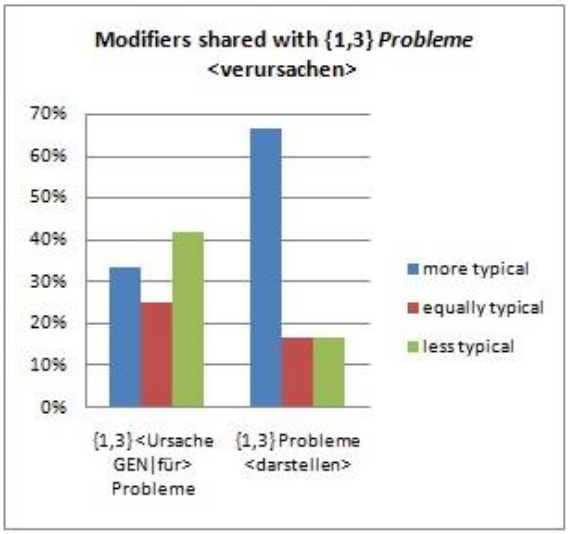


Figure B4: Continued







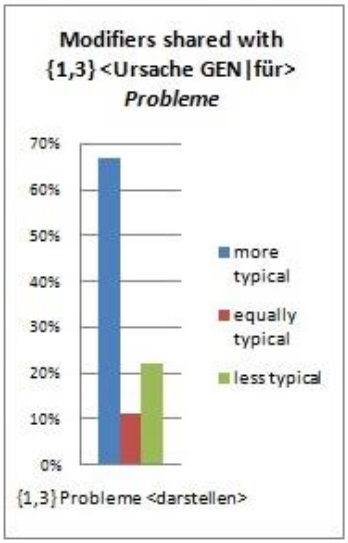
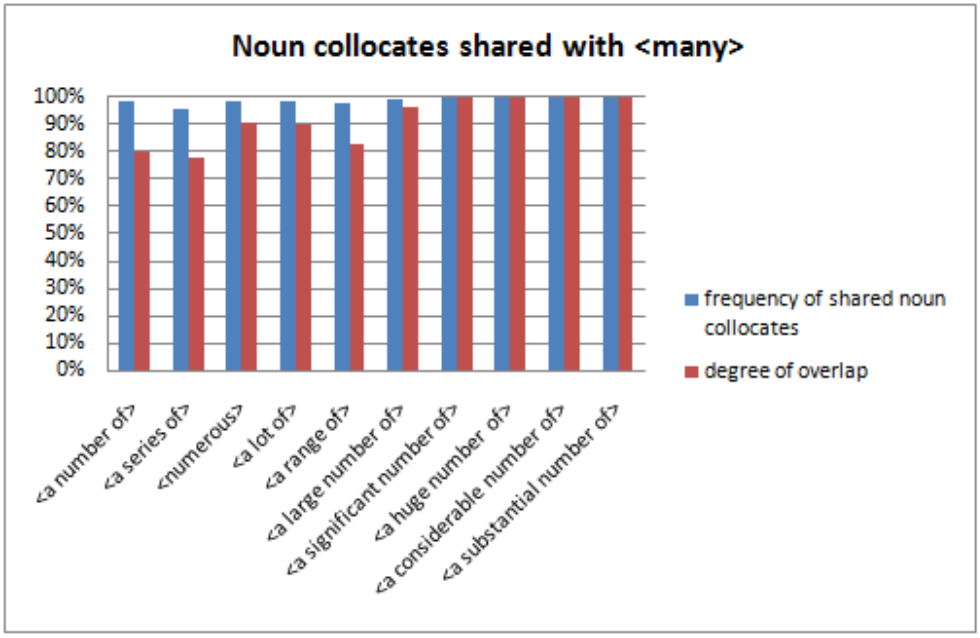


Figure B5: Frequency and degree of overlap of collective nouns that occur with positive quantifiers from the TLD {MANY COLLECTIVES} (all lexical items)



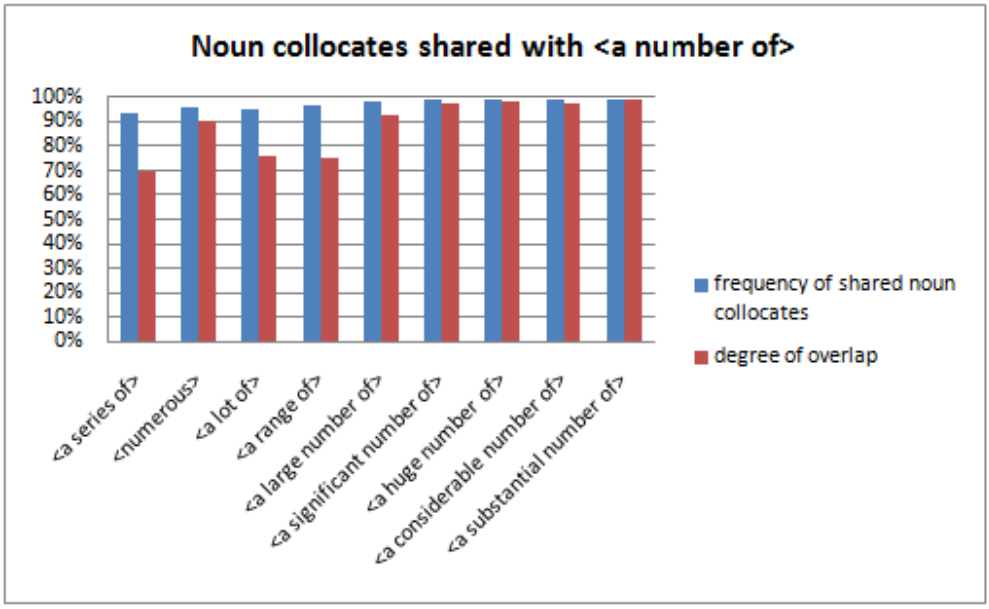
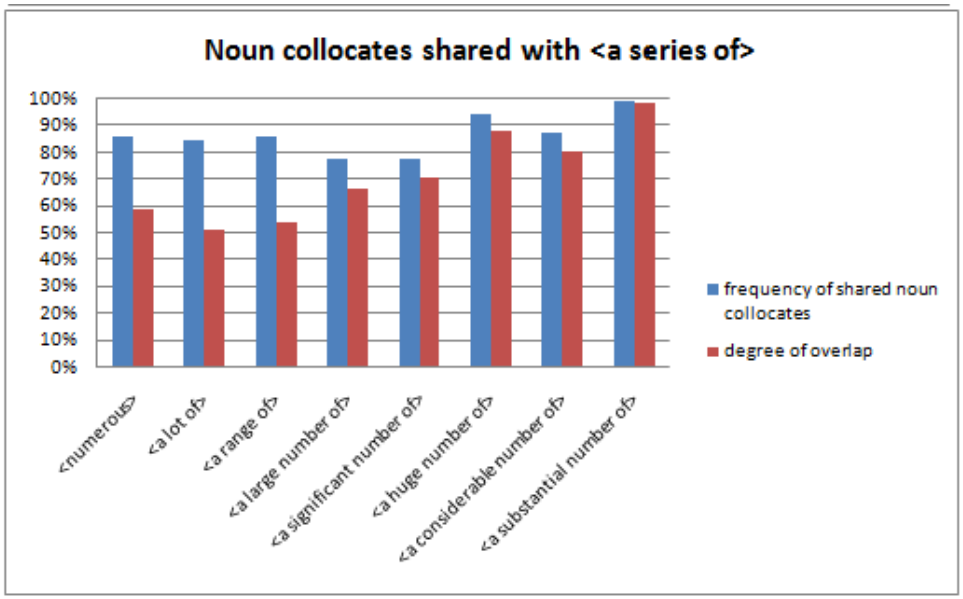


Figure B5: Continued



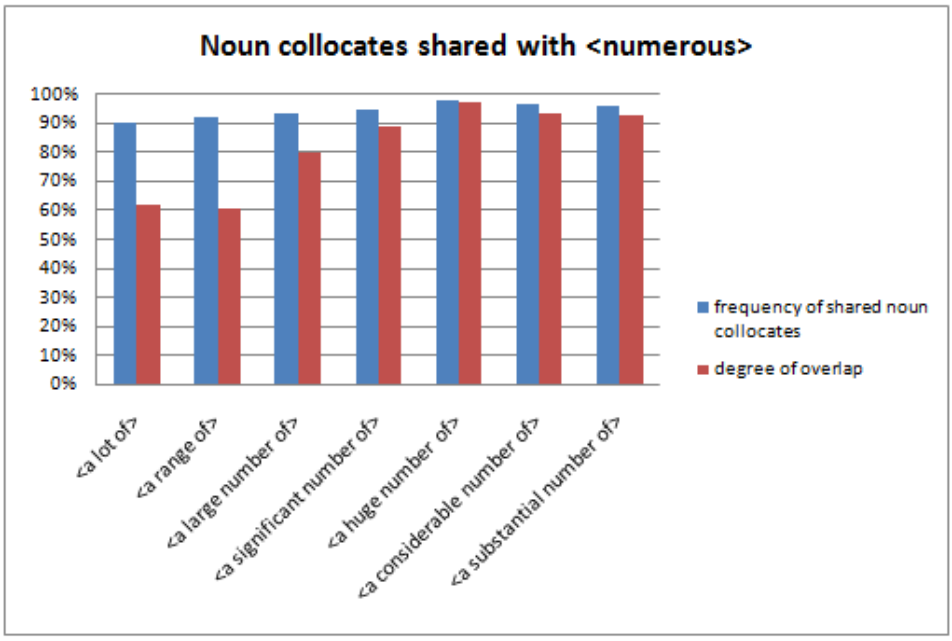
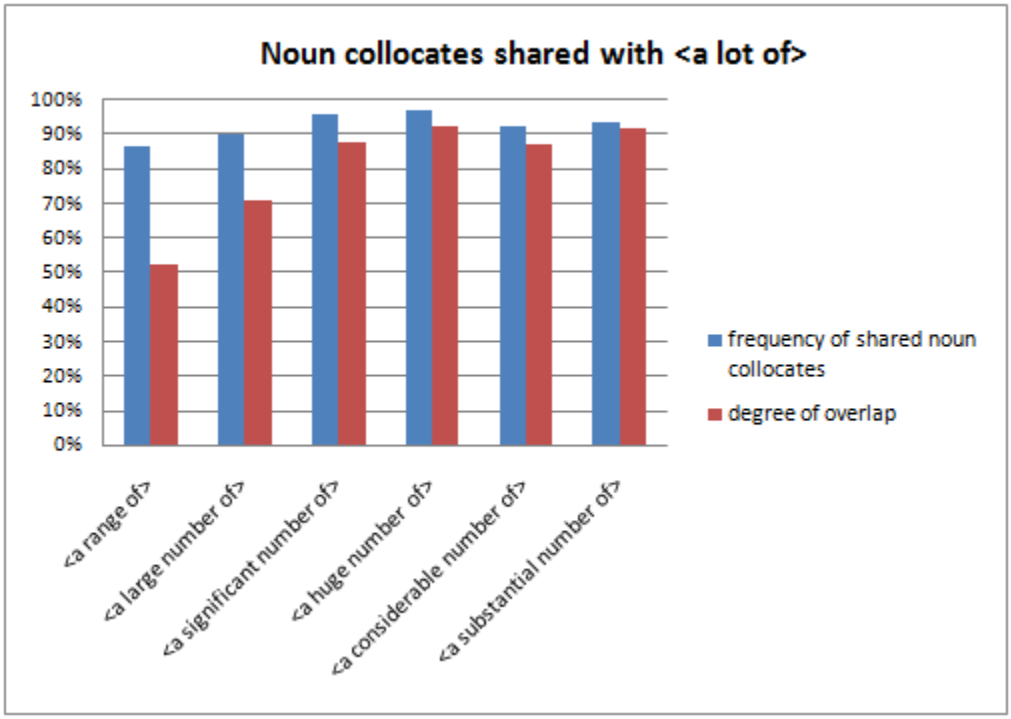


Figure B5: Continued



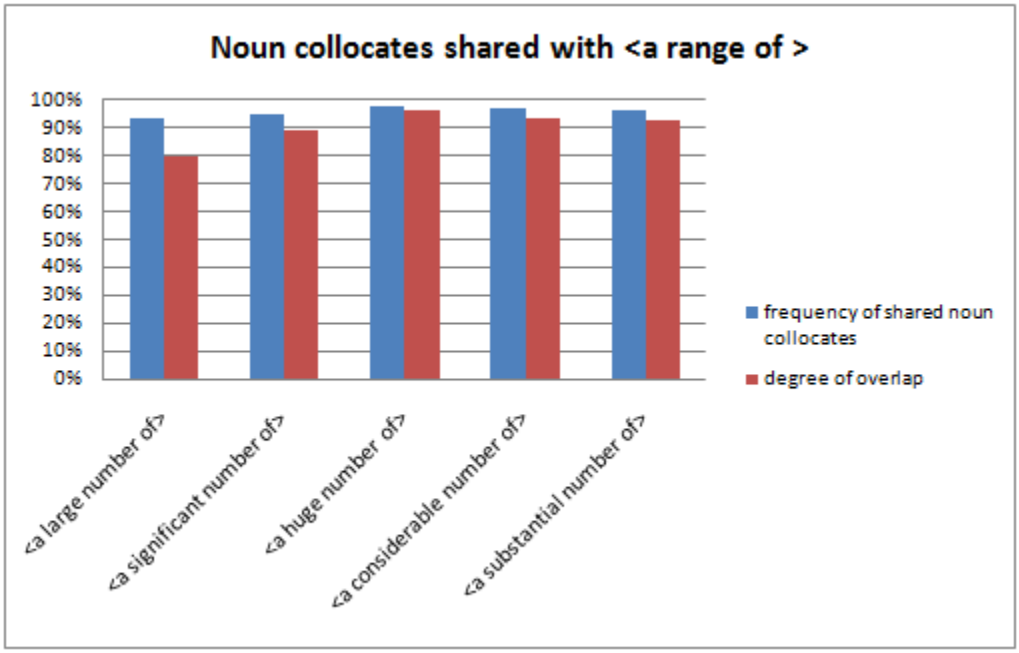


Figure B5: Continued

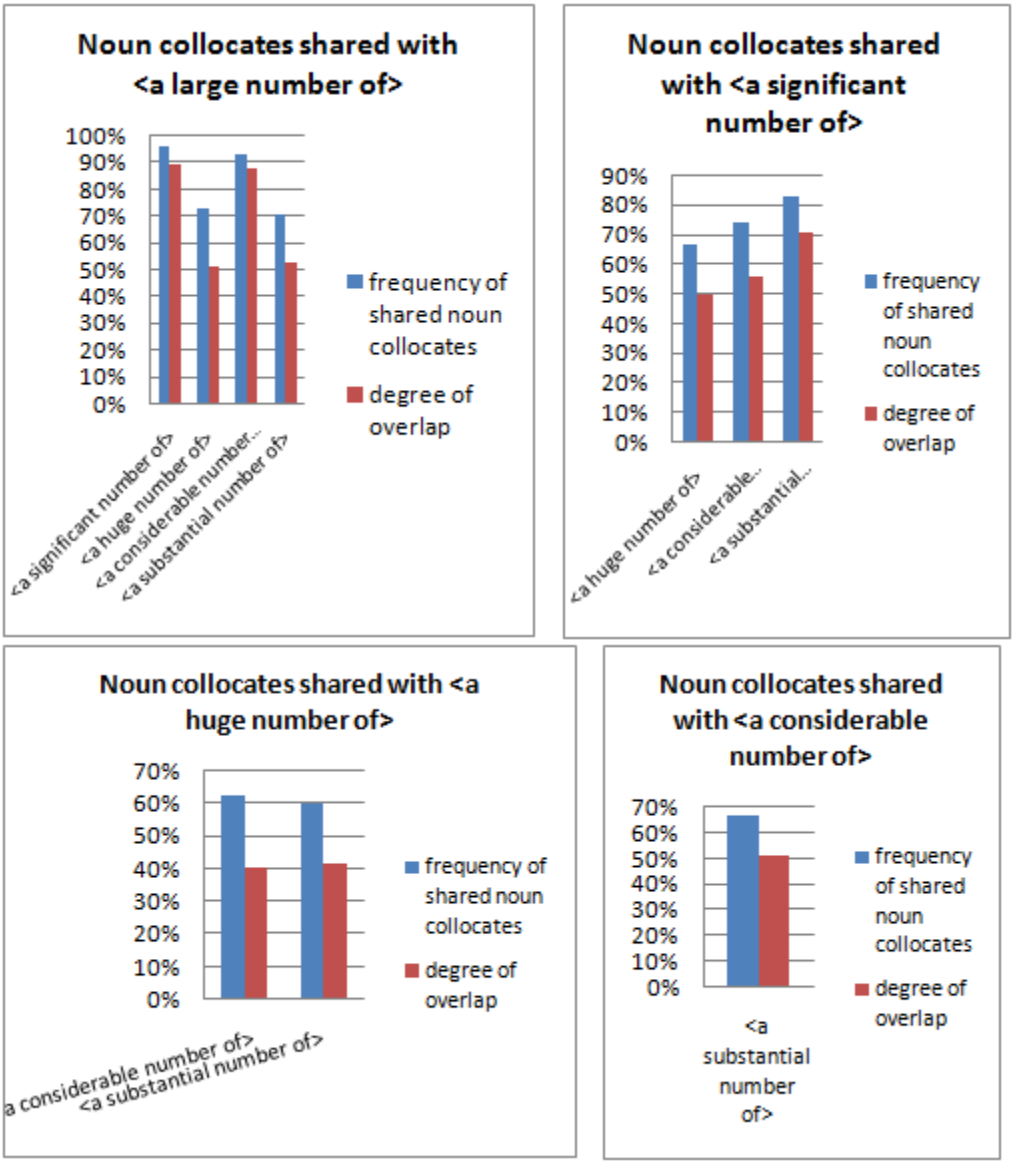
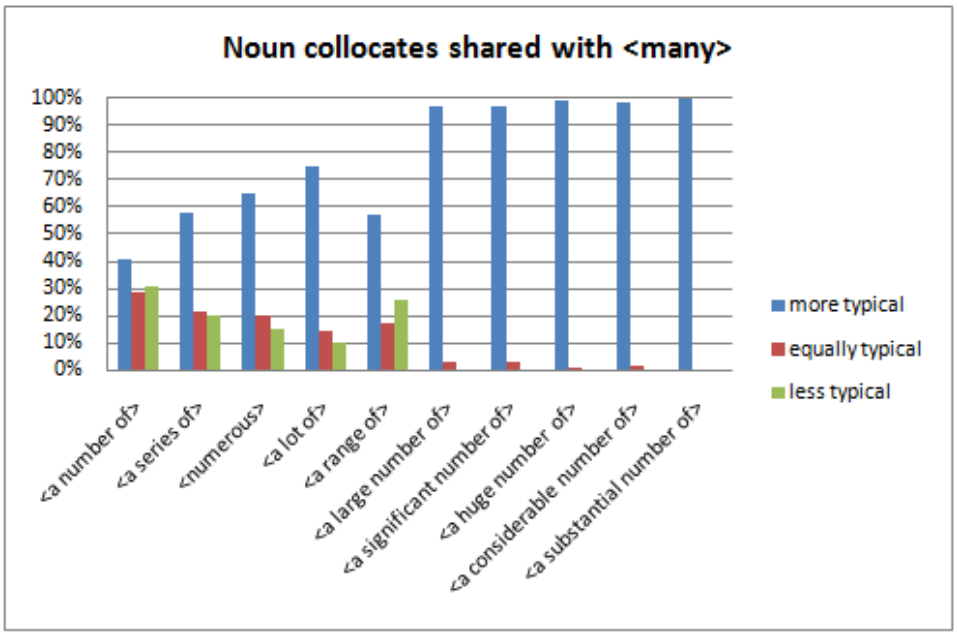


Figure B6: Association strength values for the collocations made up of positive quantifiers and shared collective nouns from the TLD {MANY COLLECTIVES} (all lexical items)



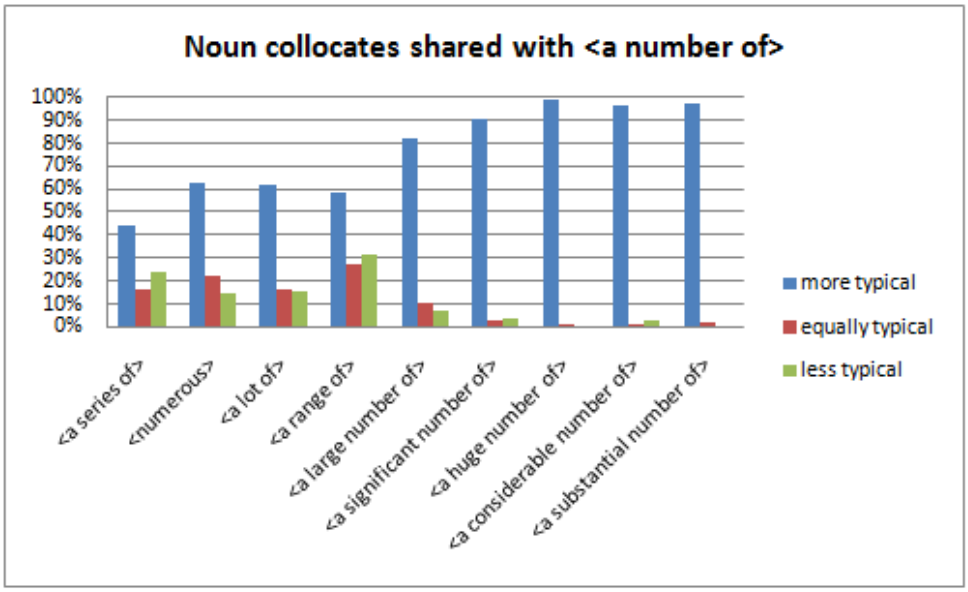
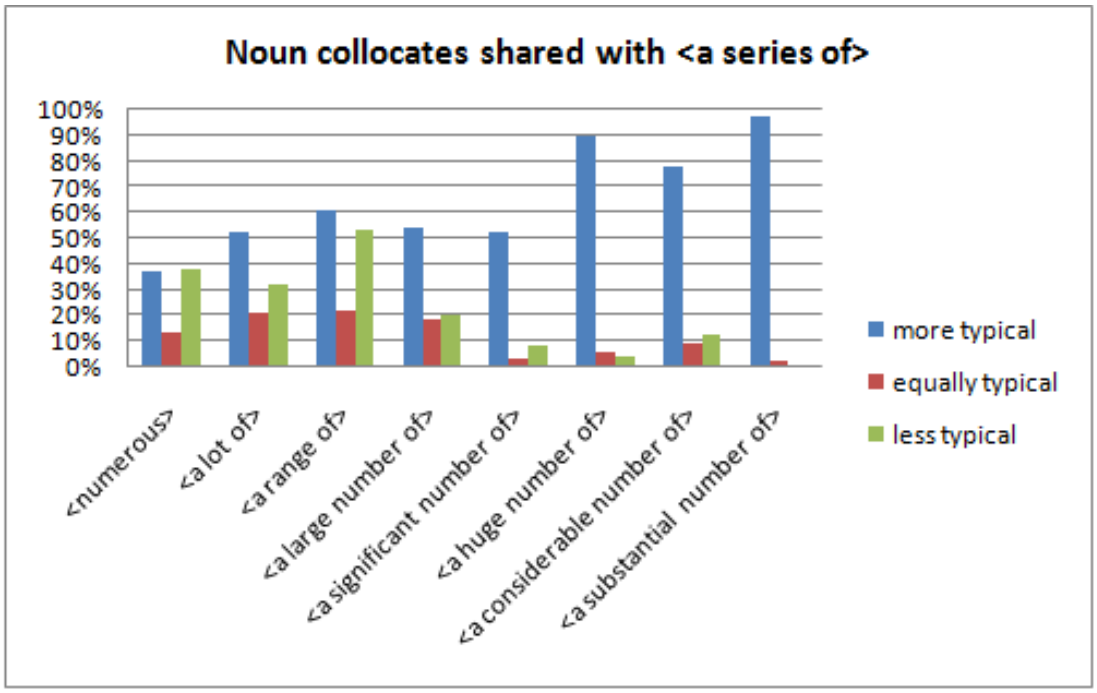


Figure B6: Continued



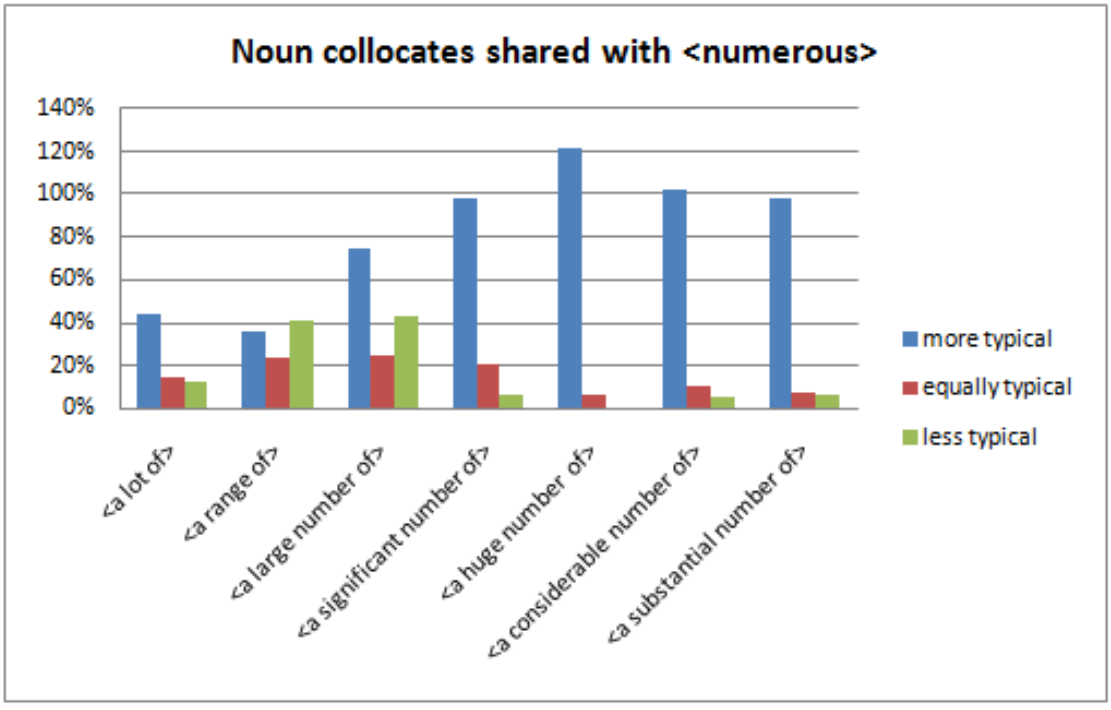
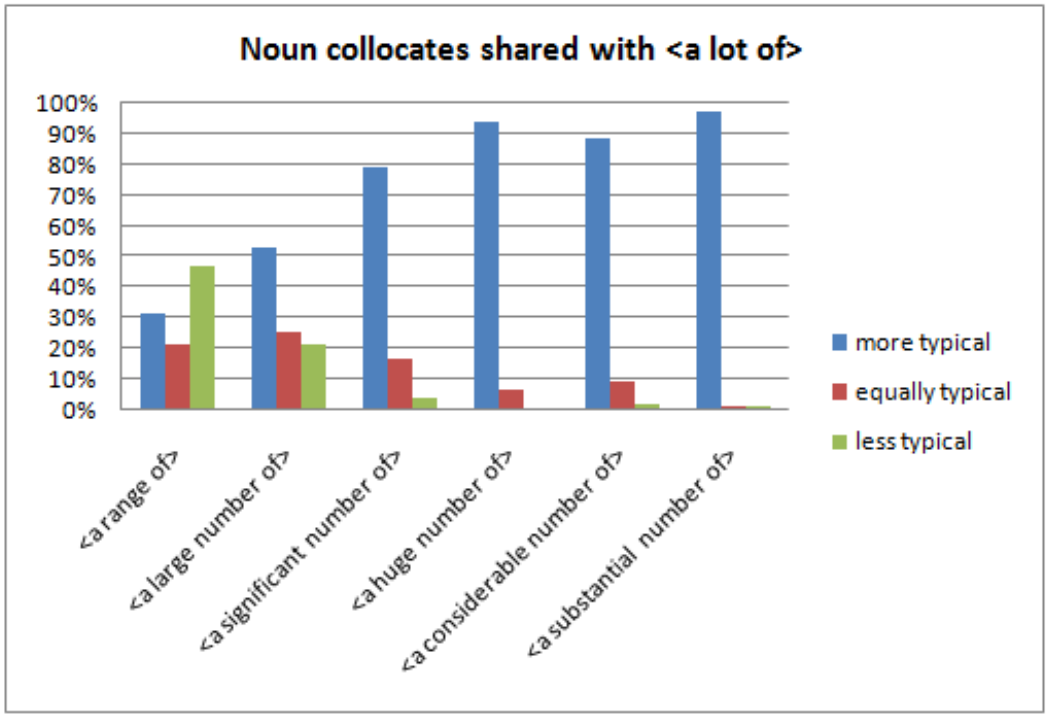


Figure B6: Continued



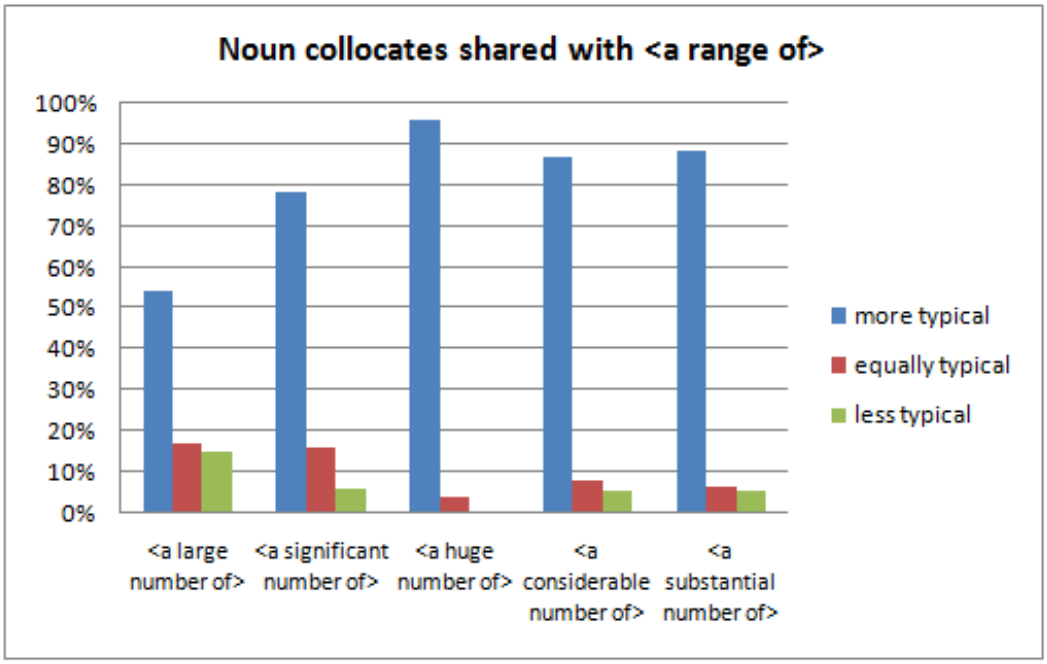


Figure B6: Continued

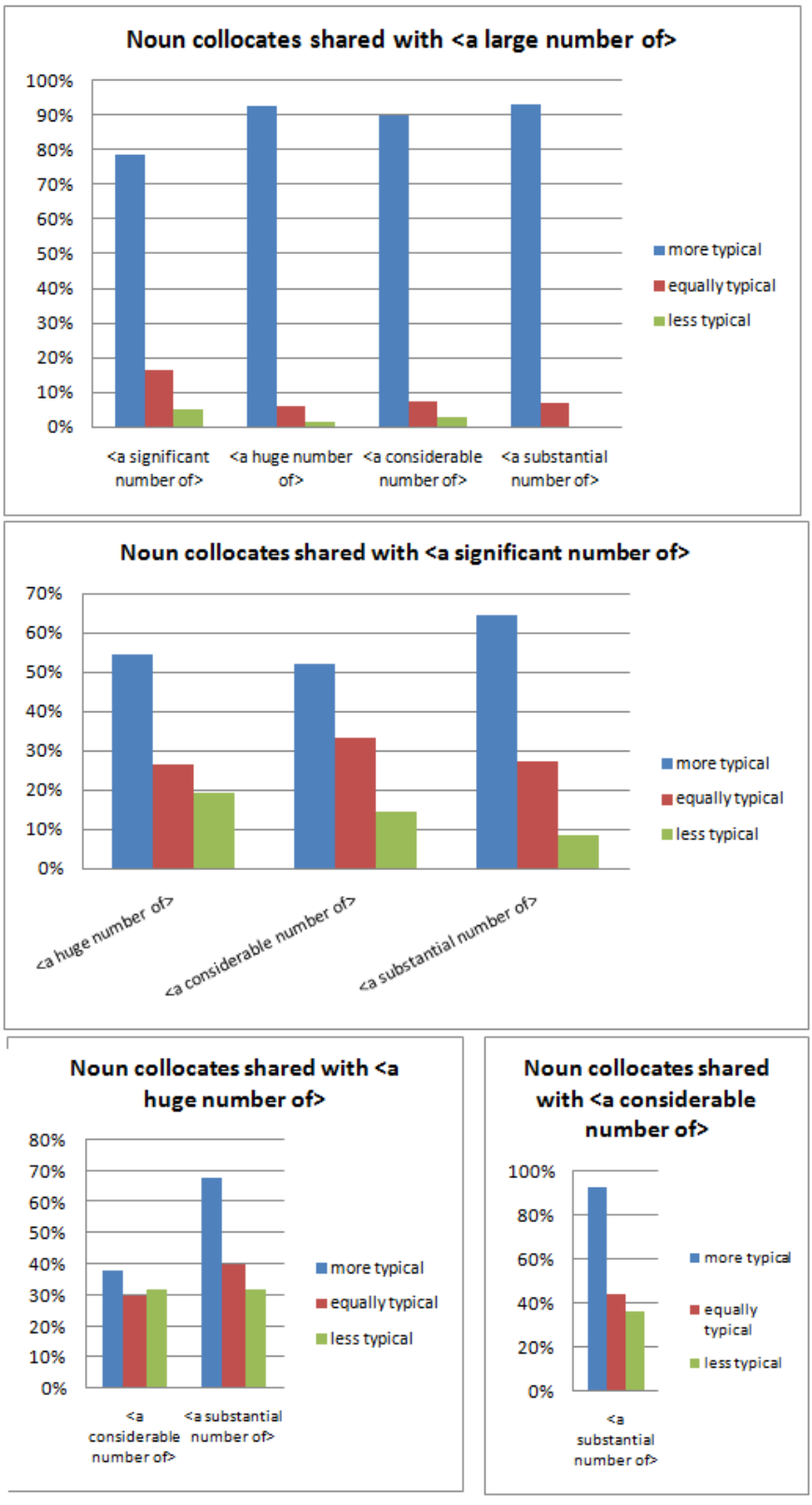
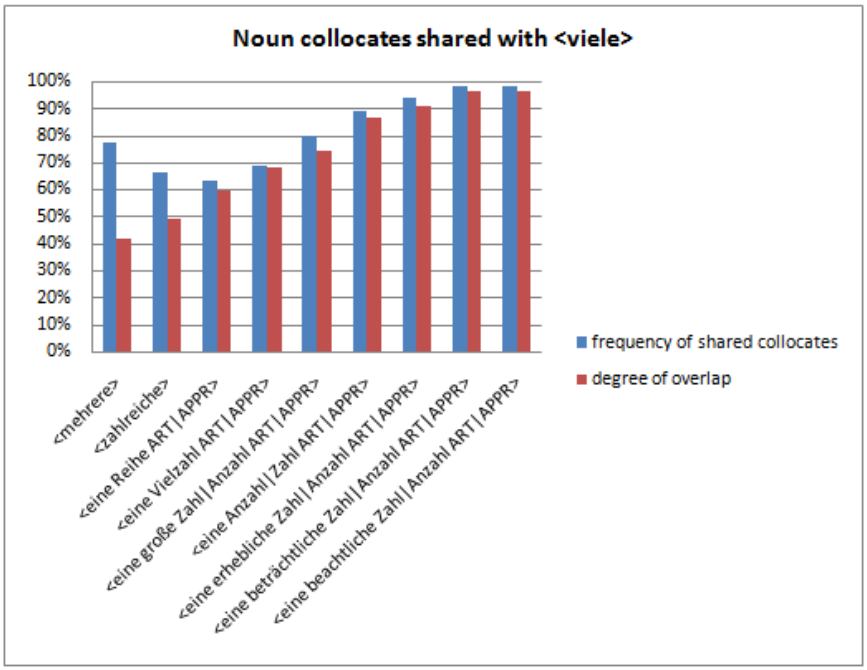


Figure B7: Frequency and degree of overlap of collective nouns that occur with positive quantifiers from the TLD {VIELE KOLLEKTIVA} (all lexical items)



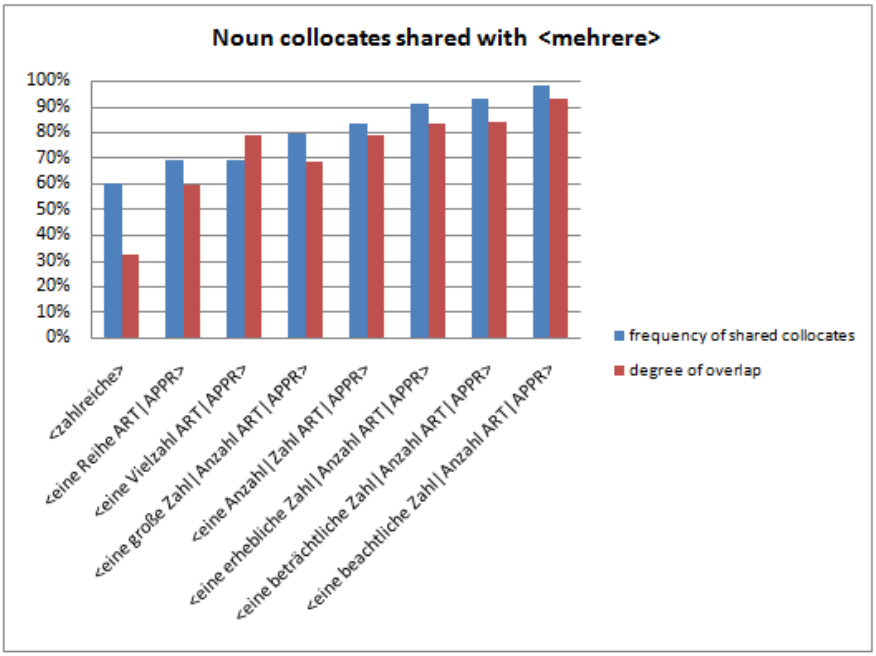
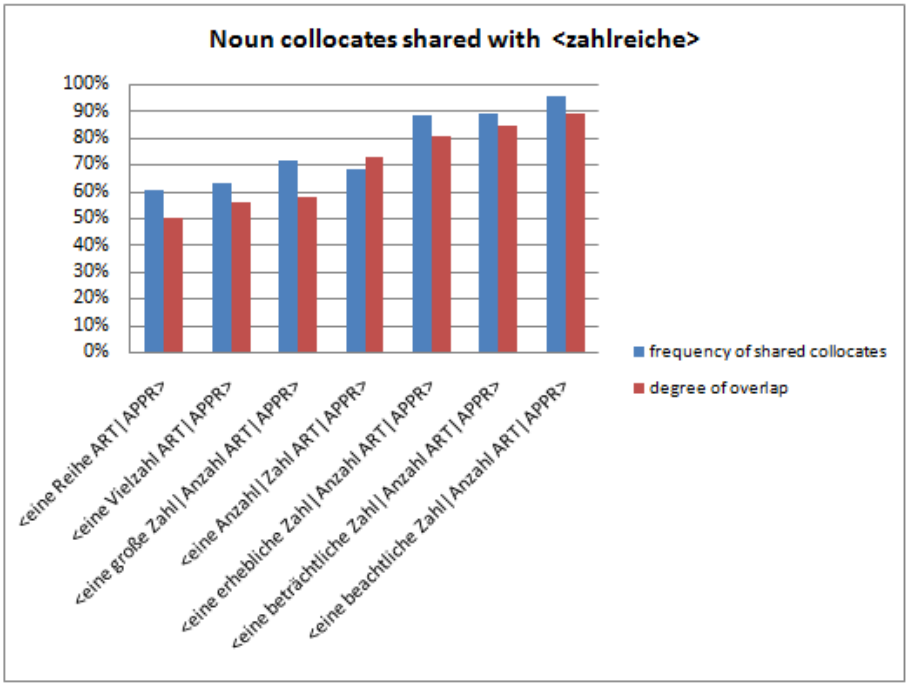


Figure B7: Continued



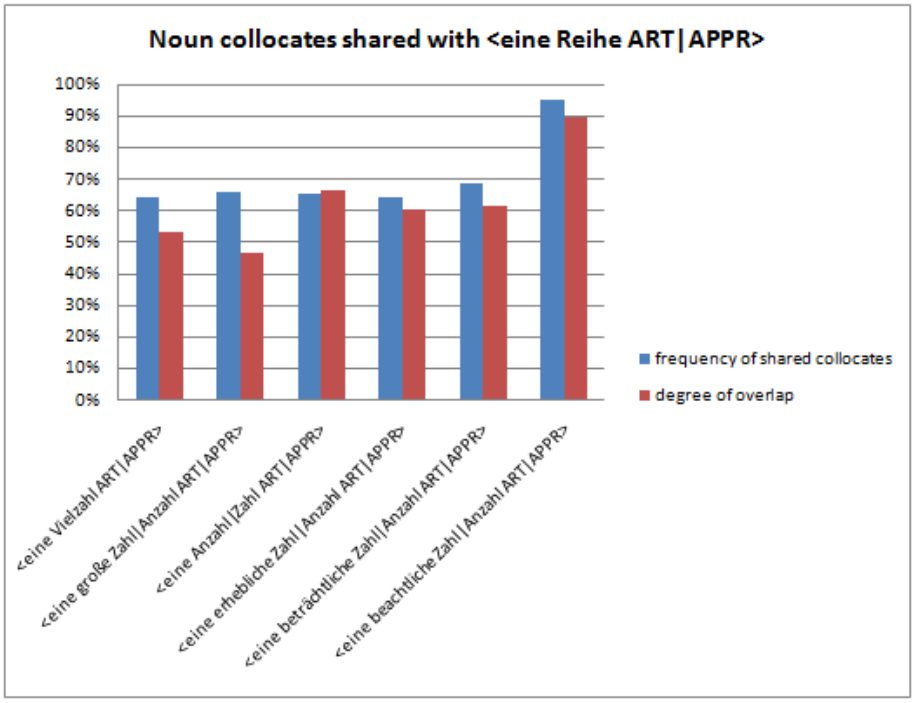
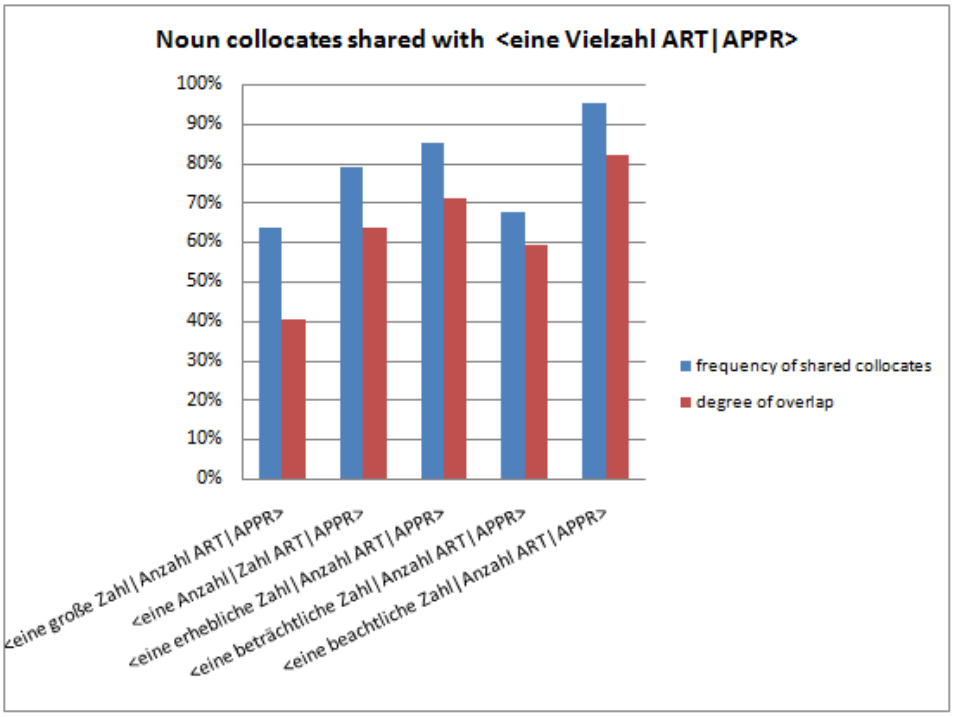


Figure B7: Continued



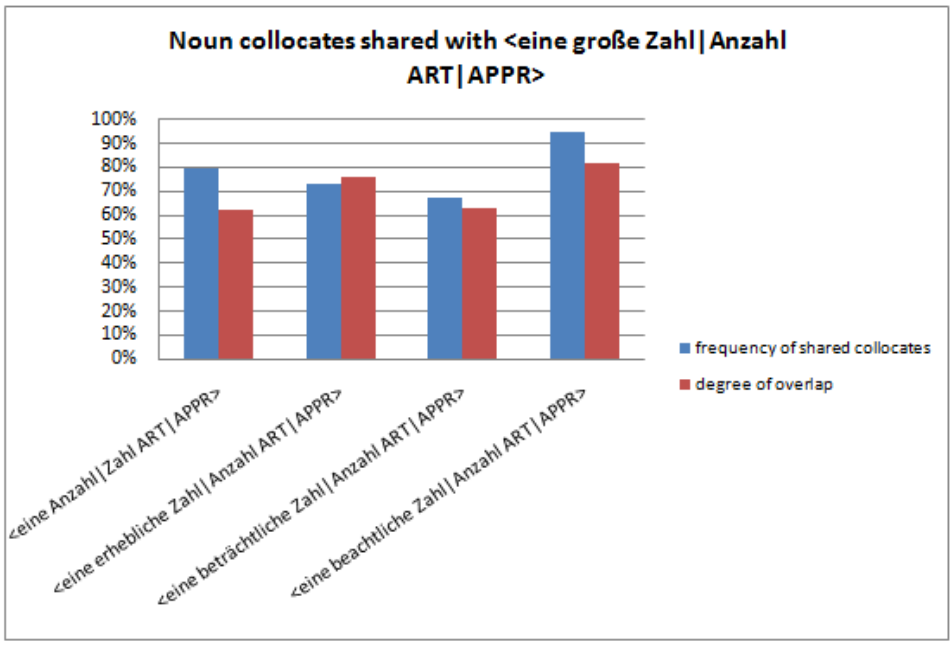
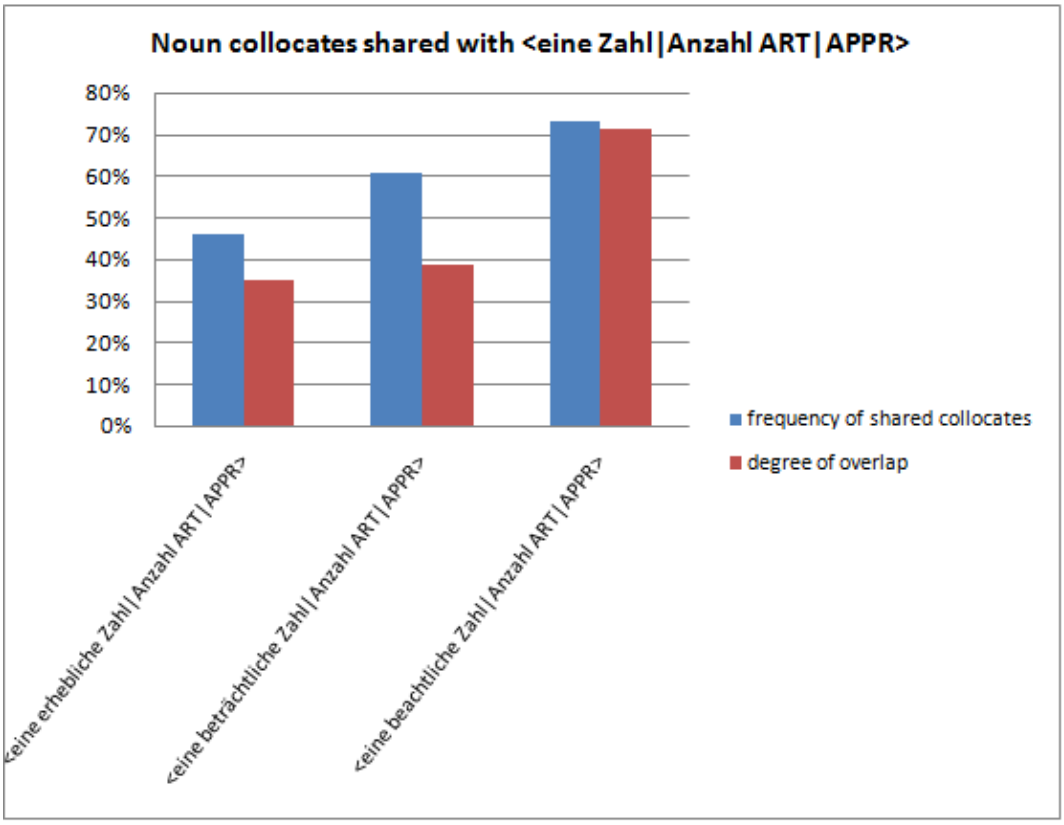
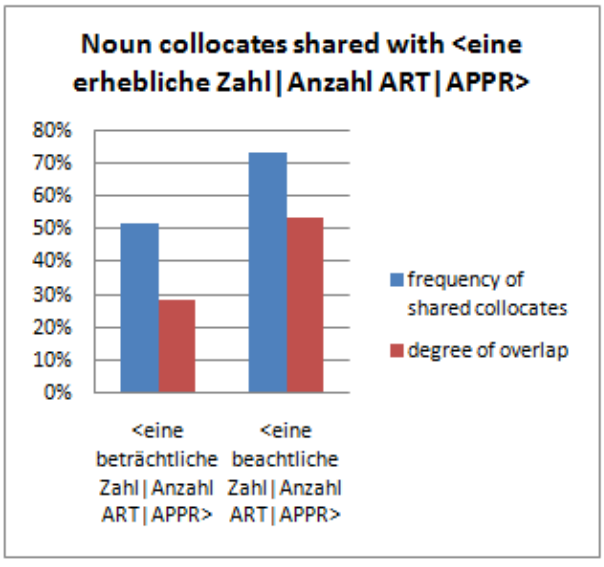


Figure B7: Continued





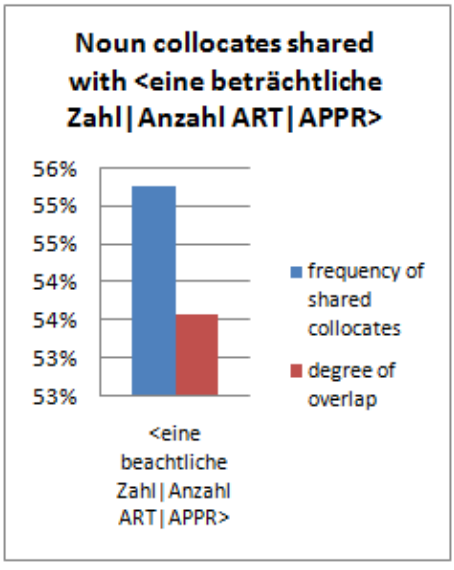
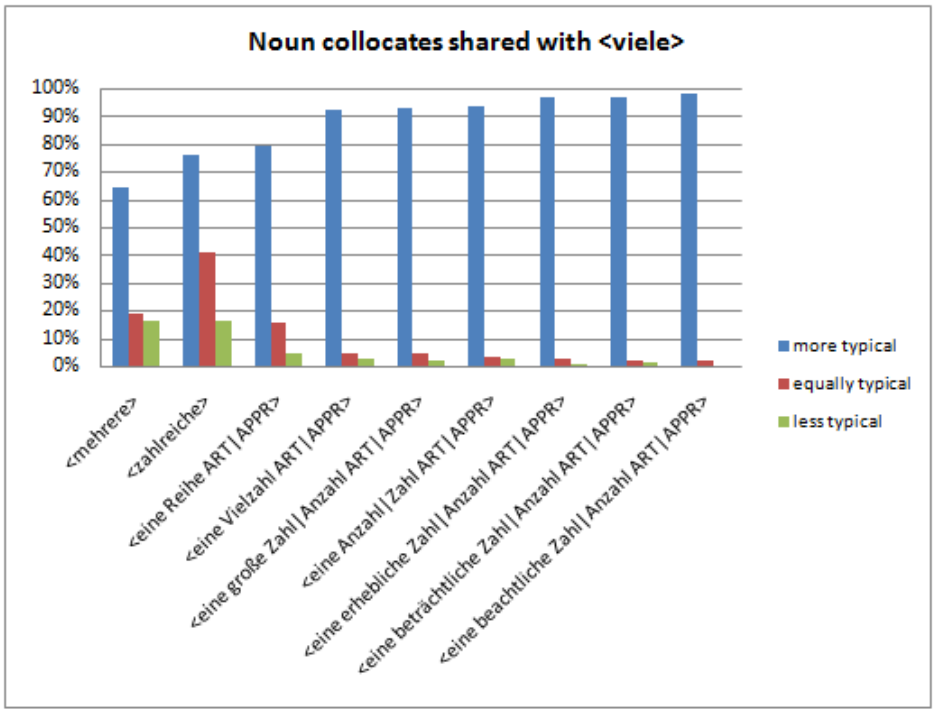


Figure B8: Association strength values for the collocations made up of positive quantifiers and collective nouns from the TLD {VIELE KOLLEKTIVA} (all lexical items)



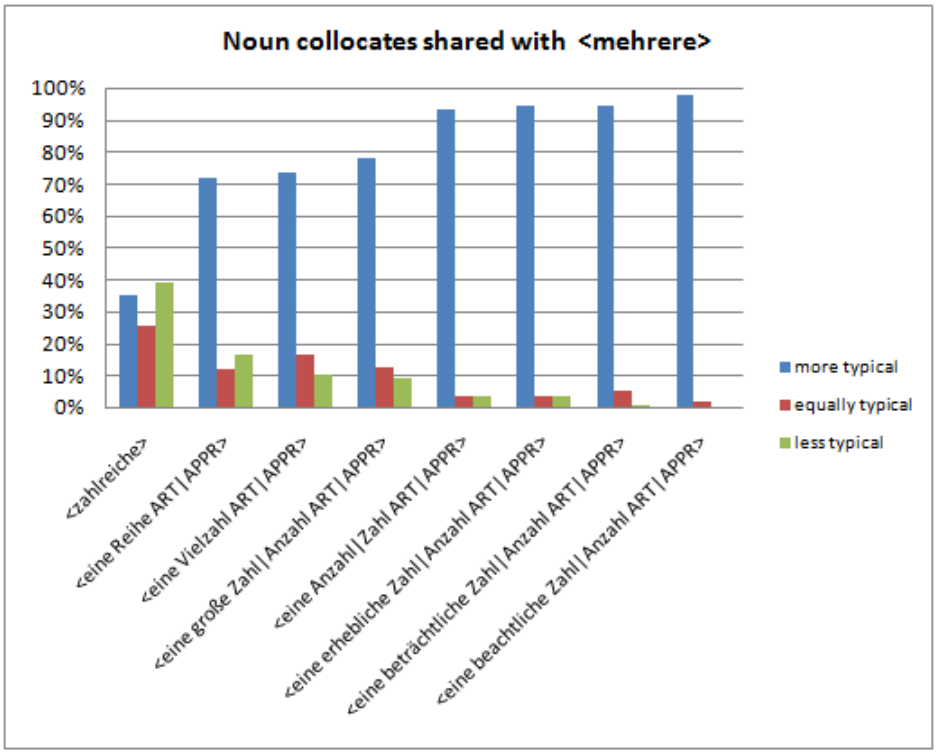
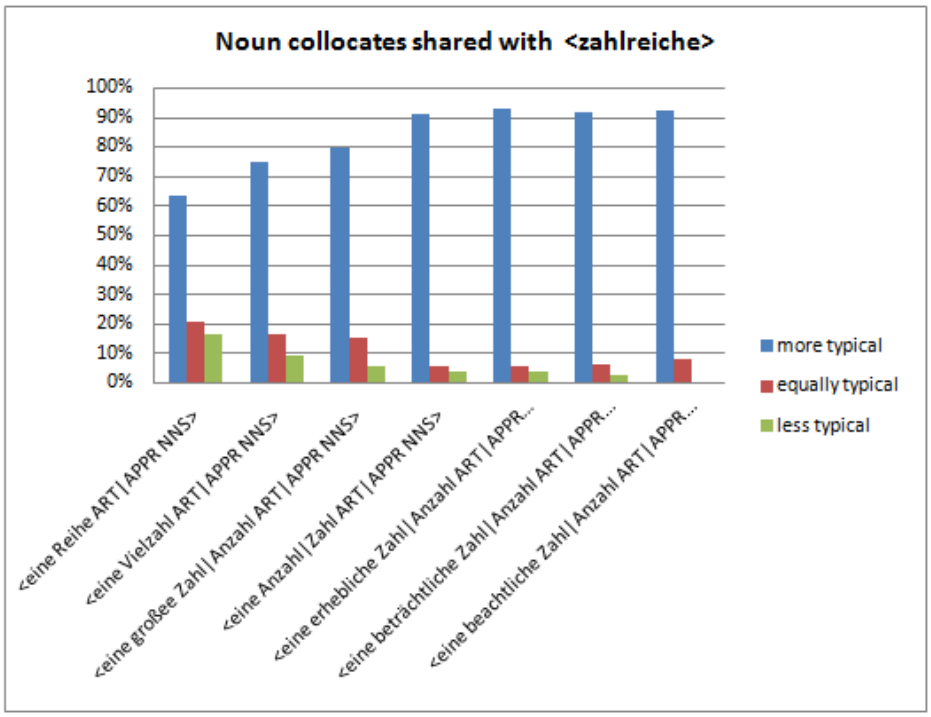


Figure B8: Continued



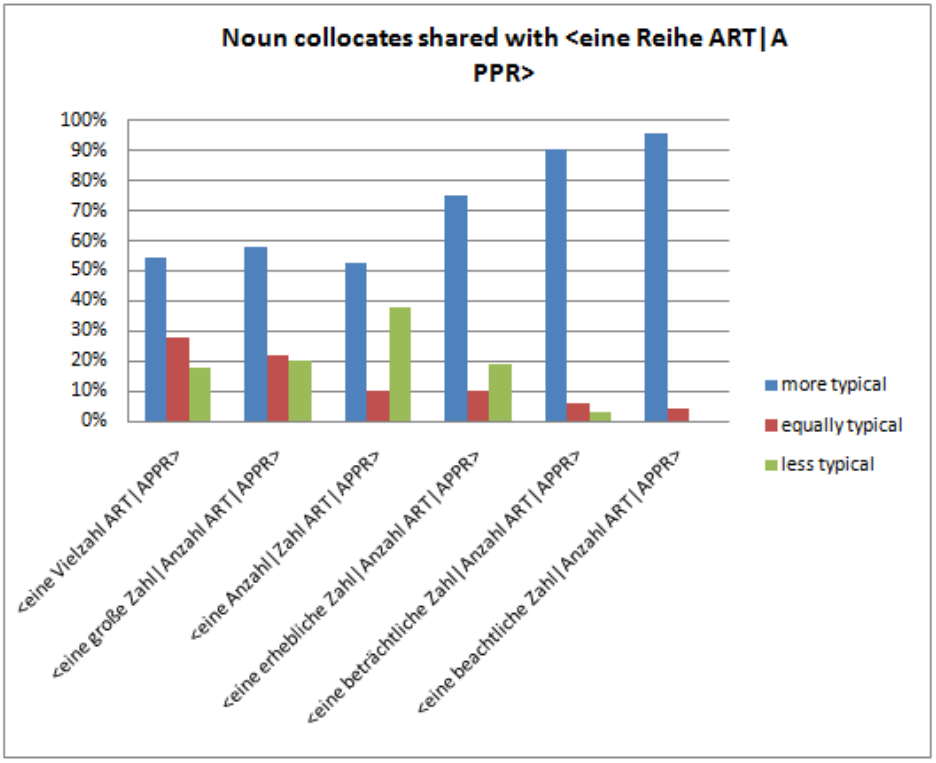
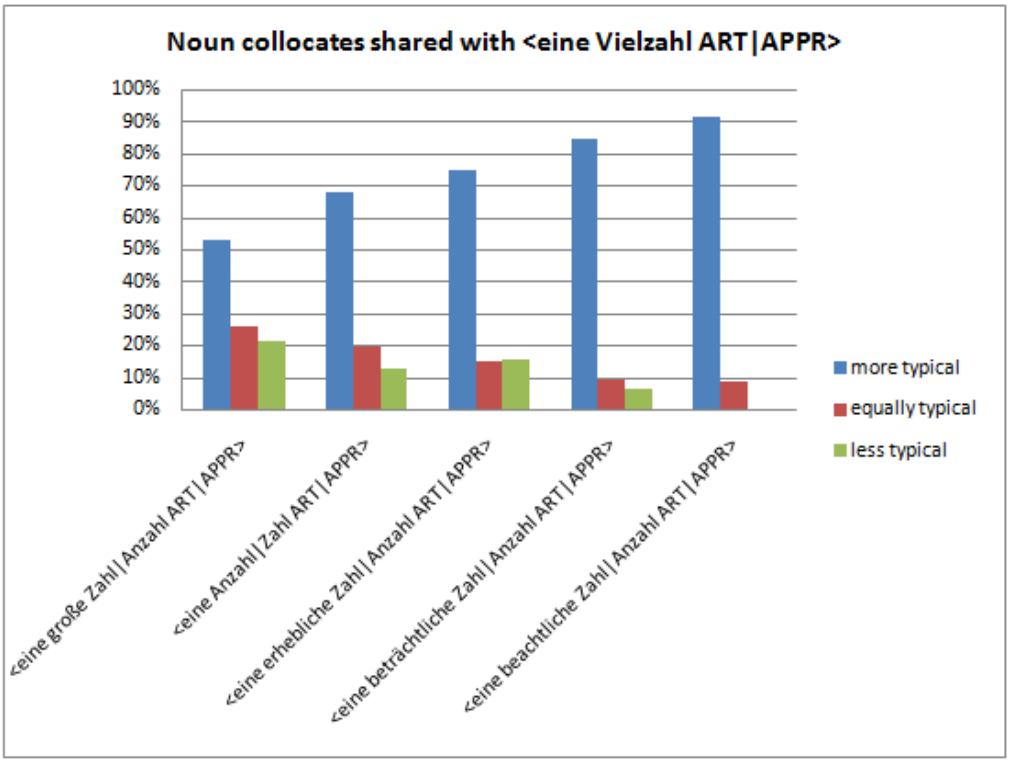


Figure B8: Continued



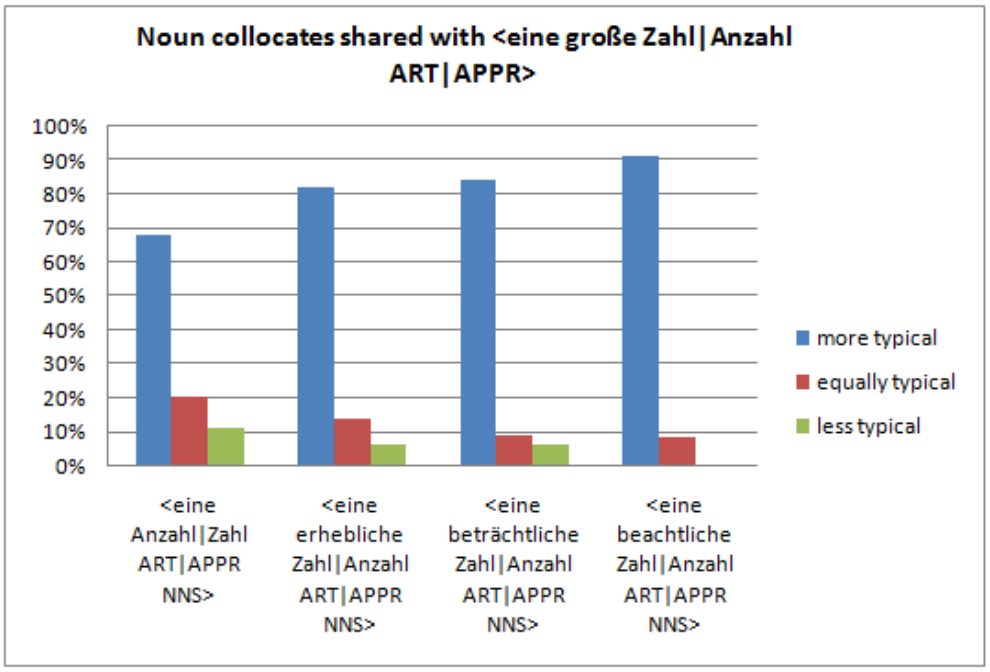


Figure B8: Continued